A bottom-up physical approach from small peptides to proteins. Methods and ab initio potentials.

Pablo Echenique

Departamento de Física Teórica
Universidad de Zaragoza

Table of Contents

Agradecimientos

Como muchas otras cosas, esta Tesis va firmada por una sola persona porque así lo exigen la tradición y la ley. Sin embargo, no habría podido ser llevada a cabo sin el cariño y la ayuda de una serie de personas a las que quiero dar las gracias aquí.

En primer lugar, gracias infinitas a mi madre y a mi padre. Además de por hacer todas las cosas por las que los padres suelen recibir gratitud y unas quinientas setenta y dos que no son tan típicas, gracias por haber añadido a la dura tarea de aguantarme, el peso de esta Tesis con tantas páginas. Gracias a mi director, José Luis Alonso, por la inquietud y la curiosidad que lo llevaron a proponerme este tema tan divertido, por todo lo que me ha enseñado (acerca de la ciencia y acerca de lo demás), por haberme tratado, desde el principio y sin ningún motivo, como a un colega, y por haber utilizado su continuo buen humor para compartir con mis padres, en horas y contextos diferentes, la ya mencionada tarea de aguantarme.

Gracias a mis profesores de la carrera de Física, por haberme enseñado a pensar, aunque digan que es su trabajo si se lo preguntan, y gracias también a mis compañeros de la carrera, Iván, Jesús, Guillermo, David, César y Diego, por lo mismo. Sin ellos, la diversión de los cinco años de formación anterior a esta Tesis habría provenido sólo de la ciencia. Especialmente, gracias a Iván, quien ha sido mi compañero de despacho estos últimos años, con quien he colaborado en dos de los artículos en los que está basada esta Tesis y sin cuyas riendas, el vicio de dar saltos me habría llevado a veces lejos de la Física, hacia el reino de la Lírica. Gracias a Jesús, en el plano académico, por ese año y pico que estuvimos trabajando y divirtiéndonos con las redes complejas; en los demás planos, otro tanto. Finalmente, el principio: gracias a Guillermo, por haberme convencido, hace tiempo, en Logroño, de que no hiciera Ingeniería, sino Física. Tenías razón.

Gracias a Yamir Moreno, codirector de esta Tesis, por el acierto de haberme animado a cambiar de tema durante un tiempo y por la diversión y el desafío de esas horas que pasamos, junto con Jesús, trabajando con las redes. Aunque, por evidentes razones de tamaño, esta Tesis está dedicada sólo a la línea principal de investigación que he seguido durante estos años, la colaboración con Yamir y Jesús ha producido, aparte de la diversión (mencionada seis o siete veces en estos agradecimientos), las siguientes publicaciones de las que quiero dejar constancia:

- J. Gómez-Gardeñes, P. Echenique and Y. Moreno, *Immunization of Real Complex Communication Networks*, European Phys. J. **49** (2006) 259.

- P. Echenique, J. Gómez-Gardeñes and Y. Moreno, *Dynamics of jamming transitions in complex networks*, Europhys. Lett. **71** (2005) 325–331.

- P. Echenique, J. Gómez-Gardeñes, Y. Moreno and A. Vázquez, *Distance-d covering problems in scale-free networks with degree correlations*, Phys. Rev. E (RC) **71** (2005) 035102.

- P. Echenique, J. Gómez-Gardeñes and Y. Moreno, *Improved routing strategies for Internet traffic delivery*, Phys. Rev. E **70** (2004) 056105.

También quiero agradecer a todas las personas del Departamento de Física Teórica de la Universidad de Zaragoza por su amabilidad, por su apoyo y por hacerlo todo siempre tan fácil. Especialmente, gracias a Alfonso Tarancón, a Fernando Falceto y a Andrés Cruz, que es a quienes más veces he ido a molestar con preguntas fáciles, y también a Pedro y a Ester, que tampoco se han salvado de mí. Entre los que estaban en el departamento cuando empecé y ahora ya no están, gracias sobre todo a Sergio y a Víctor, por hacer que el difícil comienzo no lo fuera tanto.

Gracias también a muchas de las personas de los Departamentos de Química y Bioquímica, que siempre han estado dispuestos a abrirme la puerta y resolver mis largas listas de dudas. Especialmente, gracias a Javier Sancho. La mayor parte de lo poquísimo que sé acerca de las proteínas y que es correcto proviene de él, los errores son todos míos.

De forma más puntual, algunas otras personas han contribuido, en un momento o en otro, a los trabajos presentados en esta memoria. Por eso, quiero dar las gracias también a Ramón Fernández Álvarez-Estrada, Pierpaolo Bruscolini, Marta Bueno, Isabel Campos, Gregory Chass, Nunilo Cremades, Imre Csizmadia, Jorge Estrada, Luis Antonio Fernández, Darío Ferrer, Ernesto Freire, Alfonso Jaramillo, Frank Jensen, Guillermo Losilla, Jose Onuchic, Modesto Orozco, András Perczel, Fernando Plo, Antonio Rey, David Setiadi y Tanja van Mourik.

Ya llegando al final y teniendo en mente que casi todos los grupos de personas mencionados arriba tienen intersección no nula los unos con los otros y el que viene ahora todavía más, quiero agradecer a mis amigas y amigos*, pero a los que no han sido nombrados ya, simplemente el gesto, quizás involuntario, de haber permanecido ahí todo el tiempo. No voy a decir nombres, sabéis quiénes sois.

Por último, gracias a Elena. Por ser, estar y demás infinitivos.

Zaragoza, 30 de Noviembre de 2006

Pablo Echenique Robba ha realizado esta Tesis gracias a una beca FPU del Ministerio de Educación y Ciencia de España, al grupo consolidado "Biocomputación y Física de Sistemas Complejos" del Gobierno de Aragón y al Subproyecto Zaragoza "Caracterización y análisis de fenómenos en sistemas complejos ideales y reales" del Proyecto Coordinado "Fenómenos cooperativos y emergentes en sistemas complejos entre la física y la biología", FIS2004-05073 (MEC).

* Me permito el, otras veces, deplorable desdoblamiento del masculino genérico, para dejar claro que me refiero también a ellas.

Resumen

El objetivo a largo plazo de la investigación incluida en esta Tesis Doctoral es *la correcta simulación computacional del proceso de plegamiento de las proteínas*. Por supuesto, los resultados que se presentan aquí sólo constituyen un pequeño paso en la dirección de resolver este importante problema de la biología.

En primer lugar, debido a la formación universitaria como físico teórico del candidato, una gran parte del esfuerzo invertido en esta Tesis se ha dedicado a aprender los conceptos y las herramientas necesarios para atacar el plegamiento de las proteínas. Por ello, con el convencimiento de que una de las mejores formas de organizar y fijar lo aprendido es escribirlo con la intención de que otros puedan leerlo, los dos primeros capítulos se dedican a introducir muchas de las ideas y una cierta parte del formalismo que, en el resto de la memoria, se asumen conocidos y se aplican:

- En el capítulo 1, titulado "Protein folding basics", se comentan, primero y con un inevitable contenido subjetivo, las motivaciones que pueden llevar a un científico a ocuparse del problema del plegamiento. A continuación, se realiza una introducción a la estructura de proteínas, comenzando por los componentes fundamentales: los aminoácidos, pasando por la formación del enlace peptídico y acabando con los elementos típicos de estructura secundaria. En la siguiente sección del capítulo, se describen los experimentos pioneros de Anfinsen y se define con precisión lo que entenderemos por *problema del plegamiento de las proteínas*. Finalmente, los diferentes marcos conceptuales y mecanismos de plegamiento propuestos en la literatura son discutidos, así como el formalismo mecánico-estadístico necesario para entender a grandes rasgos la estabilidad del estado nativo frente al desplegado.

- En el capítulo 2, titulado "Introduction to quantum chemistry", se realiza una introducción breve a la química cuántica, cuyas herramientas son ampliamente utilizadas en el resto de la memoria. Se comienza por la definición del Hamiltoniano molecular, que es el objeto matemático central del que emana el resto del formalismo, y se introducen las unidades atómicas, en las cuales se expresan casi todas las ecuaciones y resultados en química cuántica. Luego se discute la aproximación de Born-Oppenheimer para separar el movimiento de los núcleos del de los electrones y se demuestra el sencillo teorema variacional, el cual está en la base del método del mismo nombre para intentar encontrar buenas aproximaciones del estado fundamental del Hamiltoniano electrónico. A continuación, se describe la aproximación de Hartree como aperitivo para introducir, en la sección siguiente, la mucho más usada aproximación de Hartree-Fock, en sus variantes GHF, UHF y RHF. Luego, se escriben los orbitales moleculares de Hartree-Fock en una base finita y se derivan

las ecuaciones matriciales de Roothaan-Hall. En la siguiente sección, los diferentes tipos de orbitales atómicos que pueden formar parte de la base finita son descritos, con especial énfasis en los GTOs gaussianos y, más concretamente, los pertenecientes a las familias *split-valence* de Pople, las cuales serán las más usadas en el resto de la memoria. Finalmente, el método conocido como Møller-Plesset 2 (MP2) para incluir más correlación en los resultados es brevemente comentado.

En el apartado de investigación original de esta memoria, los trabajos han sido realizados en dos fases más o menos diferenciadas. En la primera, ciertas herramientas teóricas y computacionales fueron desarrolladas con vistas a la futura simulación de macromoléculas biológicas, así como el diseño de potenciales más precisos que sean capaces de dar cuenta de los sutiles detalles que determinan, por ejemplo, el plegamiento de las proteínas en el citoplasma celular. Los capítulos del 3 al 5 están dedicados a la introduccion de dichas herramientas:

- En el capítulo 3, titulado "A meaningful distance between potentials", se presenta un criterio estadístico (*distancia*) físicamente significativo para sistemas complejos, que permite evaluar la bondad de las aproximaciones a un cierto potencial de referencia en función de sus efectos en el comportamiento conformacional del sistema. Después de discutir las hipótesis que han de cumplir el conjunto de conformaciones de trabajo y los potenciales estudiados, se define la distancia, que es el objeto central del capítulo, se comenta su significado físico y se arguye cuáles son los valores que ha de tomar sobre una pareja de potenciales dados para que éstos sean físicamente equivalentes a temperatura T. A continuación, se discuten las posibles aplicaciones de este criterio, mencionando, aparte de la ya citada evaluación de la bondad de las aproximaciones a un cierto potencial, la medida de la robustez de una función energía potencial frente a cambios en los parámetros empíricos de los que depende o la estimación del efecto de un pequeño cambio en la naturaleza del sistema (p.ej., una mutación de un residuo en una proteína). En la sección siguiente, la distancia introducida es comparada favorablemente con otras cantidades estadísticas habitualmente usadas en la literatura, como la RMSD de la energía o el coeficiente de correlación de Pearson. Luego, algunas propiedades interesantes de la distancia son investigadas: por ejemplo, se comprueba que su cuadrado es aproximadamente aditivo y que, en determinadas condiciones bastante comunes, cumple muchas de las propiedades relevantes de una métrica, como la simetría o la desigualdad triangular. Finalmente, dos de las aplicaciones propuestas son ilustradas con ejemplos prácticos: en primer lugar, se estudia la robustez con respecto a la modificación de ciertos parámetros libres de la energía de van der Waals que hay en el campo de fuerzas de CHARMM para la proteína de 20 residuos conocida como *caja de triptófano*, demostrando que, en algunas zonas del espacio de parámetros, la robustez no es muy elevada y que, por tanto, los *fits* para parametrizar dicha parte de la energía potencial han de hacerse con sumo cuidado si se quiere que sean significativos. En segundo lugar, se comparan diferentes niveles de la teoría en el estudio ab initio del mapa de Ramachandran del dipéptido modelo HCO-L-Ala-NH_2, probando que, aunque tanto los métodos RHF como B3LYP convergen rápidamente en la base usada, las superficies de energía potencial producidas por ambos métodos

con una base de tamaño medio son físicamente disequivalentes. El trabajo descrito en este capítulo ha sido publicado como:

J. L. Alonso and P. Echenique, *A physically meaningful method for the comparison of potential energy functions*, J. Comp. Chem. **27** (2006) 238–252.

- En el capítulo 4, titulado "SASMIC internal coordinates", se introduce un conjunto sistemático de reglas que que definen un esquema de coordenadas internas, denominadas *SASMIC*, para moléculas orgánicas generales ramificadas. Estas coordenadas son modulares y separan maximalmente (utilizando solamente información acerca de la conectividad entre los átomos) los movimientos *soft* de los movimientos *hard* en dichos sistemas, lo cual permite un tratamiento más eficiente de las ligaduras relacionadas con la estructura covalente de la molécula. La modularidad del esquema SASMIC, por otro lado, hace que estas coordenadas sean muy convenientes para diseñar bases de datos de estructuras y el hecho de que su definición sea sistemática favorece su implementación en aplicaciones computacionales. Como parte del trabajo descrito en este capítulo, por ejemplo, un script de Perl ha sido desarrollado que genera las coordenadas SASMIC para péptidos a partir de la secuencia de aminoácidos. Además de proporcionar dos grupos de reglas para definir estas coordenadas, uno para polipéptidos y uno para moléculas orgánicas generales, el esquema SASMIC se usa para evaluar la frecuente aproximación que consiste en sustituir la energía libre que provendría de integrar ciertos grados de libertad irrelevantes de un sistema dado por la energía potencial constreñida a que dichos grados de libertad estén fijos en su valor de mínima energía. En este caso, se estudia la integración del ángulo χ de la cadena lateral del dipéptido modelo HCO-L-Ala-NH_2, calculando las energías con mecánica cuántica al nivel RHF/6-31+G(d) y mostrando, mediante la distancia introducida en el capítulo anterior, que la aproximación de la energía libre por la *PES* es buena hasta péptidos de alrededor de 100 residuos. El trabajo descrito en este capítulo ha sido publicado como:

  P. Echenique and J. L. Alonso, *Definition of Systematic, Approximately Separable and Modular Internal Coordinates (SASMIC) for macromolecular simulation*, J. Comp. Chem. **27** (2006) 1076–1087.

- En el capítulo 5, titulado "Explicit factorization of external coordinates in constrained statistical mechanics models", se presenta un resultado matemático que resuelve el problema de la factorización de las coordenadas externas en los determinantes de tensores métricos que aparecen en las probabilidades de equilibrio de Mecánica Estadística cuando éstas se expresan en coordenadas generalizadas o cuando se imponen ligaduras rígidas sobre el sistema. Esto permite integrar el movimiento global de las moléculas, ahorrando tiempo de computación y beneficiándose de una descripción más sencilla en función de las coordenadas internas. Además, la expresión explícita del determinante de G, el tensor métrico en el espacio total, es derivada para las coordenadas SASMIC introducidas en el capítulo anterior. El trabajo descrito en este capítulo ha sido publicado como:

  P. Echenique and I. Calvo, *Explicit factorization of external coordinates in constrained Statistical Mechanics models*, J. Comp. Chem. **27** (2006) 1748–1755.

En los dos últimos capítulos de la Tesis, las herramientas descritas arriba son utilizadas en problemas prácticos. Aunque el sistema estudiado (el dipéptido modelo HCO-L-Ala-NH_2) es uno concreto y está relacionado con el ya mencionado objetivo de simular el proceso de plegamiento de las proteínas, es necesario destacar en en este punto que las tres herramientas comentadas arriba son aplicables a moléculas generales y, en el caso de la distancia y la factorización, a cualquier sistema físico.

Dos aproximaciones destinadas a reducir el coste computacional del tratamiento cuántico del dipéptido modelo HCO-L-Ala-NH_2 son definidas, discutidas y estudiadas a continuación como comienzo de un programa más a largo plazo. Dicho programa, incluye la repetición de las investigaciones en los dipéptidos correspondientes a los 19 aminoácidos restantes, así como el análisis de la influencia de las fuerzas de largo alcance en oligopéptidos, finalizando, eventualmente, con el diseño de un potencial clásico para simular el problema del plegamiento.

- En el capítulo 6, titulado "Study of the effects of stiff and rigid constraints in the conformational equilibrium of the alanine dipeptide", se analiza la aproximación consistente en despreciar los determinantes de los tensores métricos que aparecen en la distribución de probabilidad de equilibrio y en el potencial de Fixman cuando se imponen ligaduras de tipo *stiff* o *rigid* sobre el sistema. Para ello, en primer lugar, se introduce el formalismo matemático y se derivan las expresiones de la densidad de probabilidad en ambos casos. A continuación, y después de un breve resumen acerca del uso en la literatura de las aproximaciones que van a ser estudiadas, se presentan los cálculos computacionales (al nivel MP2/6-31++G(d,p)) necesarios para obtener tanto la superficie de energía potencial del dipéptido modelo HCO-L-Ala-NH_2 en el espacio de los ángulos de Ramachandran como los diferentes determinantes de los tensores métricos en las mismas variables. En todo el trabajo, se usan las coordenadas SASMIC (introducidas en el capítulo 4) para describir el sistema, y se integran las coordenadas externas gracias a las expresiones halladas en el capítulo 5. Finalmente, se demuestra, usando la distancia definida en el capítulo 3, que algunas de las correcciones son no despreciables si uno está interesado en todo el espacio de Ramachandran, mientras que, si sólo nos centramos en la región más baja del mapa, en la que se hallan los principales elementos de estructura secundaria, todos los términos correctivos pueden ser despreciados hasta péptidos de longitud considerable. Según nuestro conocimiento, ésta es la primera vez que se calculan todos los términos correctivos en un sistema de interés biológico y con una función energía potencial realista. El trabajo descrito en este capítulo ha sido publicado como:

  P. Echenique, I. Calvo and J. L. Alonso, *Quantum mechanical calculation of the effects of stiff and rigid constraints in the conformational equilibrium of the Alanine dipeptide*, J. Comp. Chem. **27** (2006) 1733–1747.

- En el capítulo 7, titulado "Efficient model chemistries for peptides. Split-valence Gaussian basis sets and the heterolevel approximation", se realiza un estudio exhaustivo de la eficiencia de ciertas *model chemistries*, tanto homo- como heteronivel y usando los métodos RHF y MP2, para calcular el mapa de Ramachandran del dipéptido modelo HCO-L-Ala-NH_2. Los objetivos de este estudio son dos: por

un lado, estudiar las bases gaussianas y *split-valence* de Pople, identificando las funciones más eficientes en este tipo de problema, y, por otro lado, comprobar la hipótesis de heteronivel, muy utilizada en la literatura y que sostiene que es más eficiente calcular la geometría del sistema a un nivel de la teoría más bajo que el usado para la energía. Con más de 250 superficies de energía potencial y habiendo invertido un tiempo de CPU de alrededor de 9 años en el cluster bajo Linux del Instituto de Biocomputación y Física de los Sistemas Complejos (BIFI), se llega a conclusiones muy detalladas respecto de los dos objetivos mencionados. Las más importantes son, en el caso de las bases, que la inclusión de polarizaciones en los átomos pesados es muy eficiente y que existen bases pequeñas, como la 6-31G(d), que son capaces de dar cuenta de la geometría con una gran precisión y a un coste muy bajo. En el caso de la hipótesis de heteronivel, la conclusión general es que se cumple para el comportamiento conformacional de este dipéptido, permitiendo ahorrar en futuros estudios una gran cantidad de recursos computacionales. Por último, resaltar que también en este trabajo las herramientas desarrolladas en los capítulos anteriores han resultado fundamentales y que la superficie de energía potencial de referencia usada, el homonivel MP2/6-311++G(2df,2dp), es la más precisa en la literatura, según nuestro conocimiento. Este trabajo está aún en fase de realización y, por tanto, no ha sido publicado.

Finalmente, una serie de apéndices han sido incluidos para tratar algunos temas breves que no tenían fácil cabida en los capítulos. Así en el apéndice A, se comenta el significado de las densidades de probabilidad en relación con algunos de los temas de la memoria. En el apéndice B, se introduce la derivada funcional y, en el apéndice C, se describe el método de los multiplicadores de Lagrange, ambas herramientas utilizadas en el capítulo 2. En el apéndice D, se demuestra un resultado matemático que indica que lo realizado en el capítulo 5 se puede explicar geométricamente, y en el apéndice E, algunas notaciones y definiciones relativas a los dipéptidos modelo son introducidas.

Por último, aparte de los trabajos de investigación mencionados en el resto de este resumen y en los que están basados los capítulos de esta memoria, los siguientes artículos han sido también publicados como parte de un libro (el primero) o de *proceedings* de congresos (los dos últimos):

- J. L. Alonso, G. A. Chass, I. G. Csizmadia, P. Echenique and A. Tarancón, *Do theoretical physicists care about the protein-folding problem?*, In the book: *Meeting on Fundamental Physics 'Alberto Galindo'*, Alvarez-Estrada R. F. et al. (Ed.), Madrid: Aula Documental, 2004.

- J. L. Alonso and P. Echenique, *Relevant distance between two different instances of the same potential energy in protein folding*, Biophys. Chem. **115** (2005) 159–168.

- P. Echenique, J. L. Alonso and I. Calvo, *Effects of constraints in general branched molecules: A quantitative ab initio study in HCO-L-Ala-NH2*, in *From Physics to Biology. The Interface between experiment and Computation. BIFI 2006 II International Congress*, edited by J. Clemente-Gallardo, Y. Moreno, J. F. Sáenz Lorenzo and A. Velázquez-Campoy, volume **851**, pp. 108–116, AIP Conference Proceedings, Melville, New York, 2006.

Summary

The long-term objective of the research that is included in this Ph.D. dissertation is *the correct computational simulation of the protein folding process*. Of course, the results presented herein only constitute a small step towards the solution of this important problem of biology.

Let us remark first that, due to the university formation as theoretical physicist of the candidate, a great part of the effort invested in this dissertation has been devoted to learn the concepts and tools that are necessary to tackle the folding of proteins. This is why, with the belief that one of the best ways of organizing and settling the learned matter is to write it with the intention that others could read it, the first two chapters are dedicated to introduce many of the ideas and some of the formalism that are assumed to be known and that are applied in the rest of the document:

- In chapter 1, entitled "Protein folding basics", we first comment, from an unavoidably subjective point of view, the motivations that may lead a scientist to study the folding problem. Then, an introduction to protein structure is given, starting by the fundamental building-blocks: the amino acids, following with the formation of the peptide bond, and ending with the typical secondary structure elements. In the next section of this chapter, we describe the pioneering experiments by Anfinsen and we precisely define what shall be understood by *protein folding problem*. Finally, the different conceptual frameworks and mechanisms of folding proposed in the literature are discussed, as well as the statistical mechanical formalism needed to understand in broad strokes the stability of the native state with respect to the unfolded one.

- In chapter 2, entitled "Introduction to quantum chemistry", we briefly present the basics of this discipline, whose tools are persistently used in the rest of the dissertation. We start by defining the molecular Hamiltonian, which is the central mathematical object from which the rest of the formalism emanates, and we introduce the atomic units, in which most of the quantum chemistry equations and results are expressed. Then, we discuss the Born-Oppenheimer approximation to separate nuclear motion from that of lighter electrons, and we prove the simple variational theorem, which underlies the method with the same name that is frequently used to find good approximations to the fundamental state of the electronic Hamiltonian. Next, Hartree approximation is introduced as an appetizer for the much more used Hartree-Fock one in the following section, where the GHF, UHF and RHF forms are described. Then, the Hartree-Fock molecular orbitals are written in a finite basis set and the Roothaan-Hall matrix equations are derived. In the next section, the

different types of atomic orbitals that may be included in the basis set are discussed, with special emphasis on the Gaussian Type Orbitals (GTOs) and, more concretely, on those belonging to the split-valence Pople families, which are the most used ones in the rest of the dissertation. Finally, the method known as Møller-Plesset 2 (MP2) to include more correlation is briefly introduced.

In the original research part of this dissertation, the works may be considered to have been made in two relatively separate phases. In the first one, certain theoretical and computational tools have been developed with a view to future simulations of biological macromolecules, as well as the design of more accurate potentials that could account for the subtle details that determine, for example, the folding of proteins in the cellular cytoplasm. Chapters from 3 to 5 are devoted to the introduction of those tools:

- In chapter 3, entitled "A meaningful distance between potentials", we present a physically meaningful statistical criterium (*distance*) for complex systems that allows to assess the quality of the approximations to a given reference potential on the basis of their effects in the conformational behaviour of the system. After discussing the hypotheses that the working set of conformations and the potentials studied must fulfill, we define the distance, which is the central object of the chapter, we comment its physical meaning, and we argue what are the values that it must take on a given pair of potentials in order for them to be physically equivalent at temperature T. Next, we discuss the possible applications of this distance, adding to the aforementioned evaluation of the goodness of the approximations to a certain potential, the measure of the robustness of a potential energy function under changes in the free empirical parameters on which it depends or the estimation of the effect of small changes in the characteristics of a given system (e.g., a mutation of a residue in a protein). In the following section, the distance introduced is favourably compared with other statistical quantities that are commonly used in the literature, such as the energy RMSD or Pearson's correlation coefficient. Then, some interesting properties of the the distance are investigated: for example, we prove that its square is approximately additive and that, in certain rather common conditions, it satisfies some of the relevant properties of a metric, such as the symmetry or the triangle inequality. Finally, two of the proposed applications are illustrated with practical examples: first, we study the robustness of the van der Waals energy, as implemented in CHARMM, with respect to the modification of certain free parameters for the 20-residue protein known as *tryptophan-cage*, showing that, in some regions of the parameter space, the robustness is not very large and therefore the fits performed to parameterize this part of the potential energy must be done much carefully if they are intended to be meaningful. Secondly, we compare different levels of the theory in the ab initio study of the Ramachandran map of the model dipeptide HCO-L-Ala-NH_2, showing that, even though both the RHF and B3LYP methods rapidly converge in the basis set, the potential energy surfaces produced by them with the same medium-sized basis set are physically inequivalent. The work described in this chapter has been published as:

  J. L. Alonso and P. Echenique, *A physically meaningful method for the comparison of potential energy functions*, J. Comp. Chem. **27** (2006) 238–252.

- In chapter 4, entitled "SASMIC internal coordinates", we introduce a systematic set of rules that define a scheme of internal coordinates, called *SASMIC*, for general branched organic molecules. These coordinates are modular and they maximally separate (using only information about the connectivity among the atoms) the *soft* from the *hard* movements of such systems. This permits a more efficient treatment of constraints related to the covalent structure of the molecule. The modularity of the SASMIC scheme, on the other hand, renders these coordinates very convenient to be used in the design of databases of structures, and the fact that their definition is systematic favours their implementation in computer codes. As a part of the work described in this chapter, for example, a Perl script has been developed that generates the SASMIC coordinates for peptides taking the amino acid sequence as the input. In addition to providing two different set of rules for defining these coordinates, one for polypeptides and one for general organic molecules, we have used the SASMIC scheme to evaluate the frequent approximation that consists in substituting the free energy that would come from the integration of certain irrelevant degrees of freedom of a given system by the potential energy constrained to the subspace in which these degrees of freedom are fixed to their optimal value. In this case, we study the integration of the side chain angle χ of the model dipeptide HCO-L-Ala-NH_2, calculating the energy functions with quantum mechanics at RHF/6-31+G(d) and showing, with the distance introduced in the previous chapter, that the approximation of the free energy by the so-called Potential Energy Surface (PES) is good up to 100-residue peptides. The work described in this chapter has been published as:

  P. Echenique and J. L. Alonso, *Definition of Systematic, Approximately Separable and Modular Internal Coordinates (SASMIC) for macromolecular simulation*, J. Comp. Chem. **27** (2006) 1076–1087.

- In chapter 5, entitled "Explicit factorization of external coordinates in constrained statistical mechanics models", we present a mathematical result that solves the problem of the factorization of the external coordinates in the mass-metric tensors determinants that appear in the equilibrium conformational probabilities in statistical mechanics when they are expressed in general curvilinear coordinates or rigid constraints are imposed on the system. This allows to integrate the global motion of molecules, saving computational time and benefiting from a simpler description in terms of the internal coordinates. In addition, the explicit expression of the determinant of G, the whole-space mass-metric tensor, is derived for the SASMIC coordinates introduced in the preceding chapter. The work described in this chapter has been published as:

  P. Echenique and I. Calvo, *Explicit factorization of external coordinates in constrained Statistical Mechanics models*, J. Comp. Chem. **27** (2006) 1748–1755.

In the last two chapters of this dissertation, the tools and techniques described above are used in practical problems. Although the system studied (the model dipeptide HCO-L-Ala-NH_2) is a particular one and it is related to the already mentioned goal of simulating the protein folding process, it is necessary to remark that the three tools introduced in the previous chapters are applicable to general molecules and, in the case of the distance and the factorization, to any physical system.

Two approximations aimed at reducing the computational cost of the quantum treatment of the model dipeptide HCO-L-Ala-NH_2 are defined, discussed and studied as the beginning of a longer-term program, which includes the analysis of the dipeptides corresponding to the 19 remaining amino acids, as well as the evaluation of the influence of sequence long-range interaction in oligopeptides. This program will eventually end with the design of a classical potential to simulate the protein folding.

- In chapter 6, entitled "Study of the effects of stiff and rigid constraints in the conformational equilibrium of the alanine dipeptide", we analyze the approximation that consists of neglecting the mass-metric tensors determinants that appear in the equilibrium probability distribution and in Fixman's compensating potential when *stiff* or *rigid* constraints are imposed on the system. To this end, we first introduce the mathematical formalism and derive the probability density expressions in both cases. Next, and after a brief summary of the use in the literature of the approximations that are going to be studied, we present the computational calculations (at the MP2/6-31++G(d,p) level of the theory) that are needed to obtain the potential energy surface of the model dipeptide HCO-L-Ala-NH_2 in the space spanned by the Ramachandran angles, as well as the determinants of the mass-metric tensors as functions of the same variables. In the whole work, the SASMIC coordinates (introduced in chapter 4) are used to describe the system, and the external coordinates are integrated out thanks to the expressions found in chapter 5. Finally, we show, using the distance defined in chapter 3, that some of the corrections are non-negligible if one is interested in the whole Ramachandran space, whereas, if we focus in the lowest region of the map, where the main secondary structure elements are located, then we may neglect all correcting terms up to peptides of considerable length. As far as we are aware, this is the first time that all corrections are calculated in a biologically interesting molecule and with a realistic potential energy function. The work described in this chapter has been published as:

  P. Echenique, I. Calvo and J. L. Alonso, *Quantum mechanical calculation of the effects of stiff and rigid constraints in the conformational equilibrium of the Alanine dipeptide*, J. Comp. Chem. **27** (2006) 1733–1747.

- In chapter 7, entitled "Efficient model chemistries for peptides. Split-valence Gaussian basis sets and the heterolevel approximation", we present an exhaustive study of the efficiency of both homo- and heterolevel model chemistries, with the RHF and MP2 methods, for calculating the potential energy surface (PES) of the model dipeptide HCO-L-Ala-NH_2 in the space spanned by the Ramachandran angles. The aims of this study are two: on the one hand, to study Pople's families of split-valence Gaussian basis sets, identifying the most efficient functions for this type of problem, and, on the other hand, to check the heterolevel assumption, which is very frequently used in the literature and states that it is more efficient to calculate the geometry of the the system at a lower level than the one used for computing the energy. With more than 250 potential energy surfaces that have taken around 9 years of CPU time to be calculated in the Linux cluster of the Institute for Biocomputation and Physics of Complex Systems (BIFI), we arrive at very detailed conclusions with respect to the two aforementioned objectives. The most important ones are, in the case of the basis sets, that it is very efficient to include polarization

shells in heavy atoms and that there exist small basis sets, such as 6-31G(d), which can account for the geometry with high accuracy and at a low cost. Regarding the heterolevel assumption, the general conclusion is that it is indeed correct for the conformational behaviour of this dipeptide, thus allowing to save a large amount of computational resources in future studies. Finally, let us remark that, also in this work, the tools developed in the previous chapters turned out to be essential and that the reference PES, the MP2/6-311++G(2df,2pd) homolevel, is the most accurate one in the literature, as far as we are aware. This study is still in progress and, therefore, it has not been published yet.

Apart from the chapters summarized above, a number of appendices have been included to deal with some brief issues that could not be easily fit in the chapters. Hence, in appendix A, we comment the meaning of probability densities in relation to some of the topics in the main body of the dissertation. In appendix B, we introduce the functional derivative and, in appendix C, we describe the Lagrange multipliers method, both tools used in chapter 2. In appendix D, we prove a mathematical result that contributes to explain the calculations performed in chapter 5 from a more fundamental point of view. Finally, in appendix E, some notation and definitions regarding model dipeptides are introduced.

In addition to the research papers mentioned in the rest of this summary and on which the chapters in this dissertation are based, the following articles have been published as a part of a book (the first one) or in conference proceedings (the last two):

- J. L. Alonso, G. A. Chass, I. G. Csizmadia, P. Echenique and A. Tarancón, *Do theoretical physicists care about the protein-folding problem?*, In the book: *Meeting on Fundamental Physics 'Alberto Galindo'*, Alvarez-Estrada R. F. et al. (Ed.), Madrid: Aula Documental, 2004.

- J. L. Alonso and P. Echenique, *Relevant distance between two different instances of the same potential energy in protein folding*, Biophys. Chem. **115** (2005) 159–168.

- P. Echenique, J. L. Alonso and I. Calvo, *Effects of constraints in general branched molecules: A quantitative ab initio study in HCO-L-Ala-NH2*, in *From Physics to Biology. The Interface between experiment and Computation. BIFI 2006 II International Congress*, edited by J. Clemente-Gallardo, Y. Moreno, J. F. Sáenz Lorenzo and A. Velázquez-Campoy, volume **851**, pp. 108–116, AIP Conference Proceedings, Melville, New York, 2006.

Chapter 1

Protein folding basics

> I always feel like running away when any one begins to talk about proteids in my presence. In my youth I had a desire to attack these dragons, but now I am afraid of them. They are unresolved problems of chemistry; and let me add, they are likely to remain such for generations to come. Yet every one who knows anything about chemistry and physiology, knows that these proteids must be understood, before we can hope to have a clear conception of the chemical processes of the human body. [9]
>
> — Ira Ramsen, 1904

1.1 Why study proteins?

The motivation

Virtually every scientific book or article starts with a paragraph in which the writer tries to persuade the readers that the topic discussed is very important for the future of humankind. We stick to that tradition in this Ph.D. dissertation; but with the confidence that, in the case of proteins, the persuasion process will turn out to be rather easy and automatic.

Proteins are a particular type of biological molecules that can be found in every single living being on Earth. The characteristic that renders them essential for understanding life is simply their versatility. In contrast with the relatively limited structural variations present in other types of important biological molecules, such as carbohydrates, lipids or nucleic acids, proteins display a seemingly infinite capability for assuming different shapes and for producing very specific catalytic regions on their surface. As a result, proteins constitute the working force of the chemistry of living beings, performing almost every task that is complicated. Quoting the first sentence of a section (which shares this section's title) in Lesk's book [10]:

> In the drama of life on a molecular scale, proteins are where the action is.

Just to state a few examples of what is meant by 'action', in living beings, proteins

- are passive building blocks of many biological structures, such as the coats of viruses, the cellular cytoskeleton, the epidermal keratin or the collagen in bones and cartilages;
- transport and store other species, from electrons to macromolecules;
- as hormones, transmit information and signals between cells and organs;
- as antibodies, defend the organism against intruders;
- are the essential components of muscles, converting chemical energy into mechanical one, and allowing the animals to move and interact with the environment;
- control the passage of species through the membranes of cells and organelles;
- control gene expression;
- are the essential agents in the transcription of the genetic information into more proteins;
- together with some nucleic acids, form the ribosome, the large molecular organelle where proteins themselves are synthesized;
- as chaperones, protect other proteins to help them to acquire their functional three-dimensional structure.

Due to this participation in almost every task that is essential for life, protein science constitutes a support of increasing importance for the development of modern Medicine. On one side, the lack or malfunction of particular proteins is behind many pathologies; e.g., in most types of cancer, mutations are found in the tumor suppressor p53 protein [11]. Also, abnormal protein aggregation characterizes many neurodegenerative disorders, including Huntington, Alzheimer, Creutzfeld-Jakob ("mad cow"), or motor neuron diseases [12–14]. Finally, to attack the vital proteins of pathogens (HIV [15, 16], SARS [17], hepatitis [18], etc.), or to block the synthesis of proteins at the bacterial ribosome [19], are common strategies to battle infections in the frenetic field of rational drug design [20].

Apart from Medicine, the rest of human technology may also benefit from the solutions that Nature, after thousands of millions[1] of years of 'research', has found to the typical practical problems. And that solutions are often proteins: New materials of extraordinary mechanical properties could be designed from the basis of the spider silk [21, 22], elastin [23] or collagen proteins [24]. Also, some attempts are being made to integrate these new biomaterials with living organic tissues and make them respond to stimuli [25]. Even further away on the road that goes from passive structural functions to active tasks, no engineer who has ever tried to solve a difficult chemical problem can avoid to experience a feeling of almost religious inferiority when faced to the speed, efficiency and specificity with which proteins cut, bend, repair, carry, link or modify other chemical species. Hence, it is normal that we play with the idea of learning to control that power and have, as a result, nanoengines, nanogenerators, nanoscissors, nanomachines in general [26]. This Ph.D. candidate, in particular, felt a small sting of awe when he learnt about the pump and the two coupled engines of the principal energy generator in

[1] Herein, we shall use the 'British' convention for naming large numbers; in which 10^9='a thousand million', 10^{12}='a billion', 10^{15}='a thousand billion', 10^{18}='a trillion', and so on.

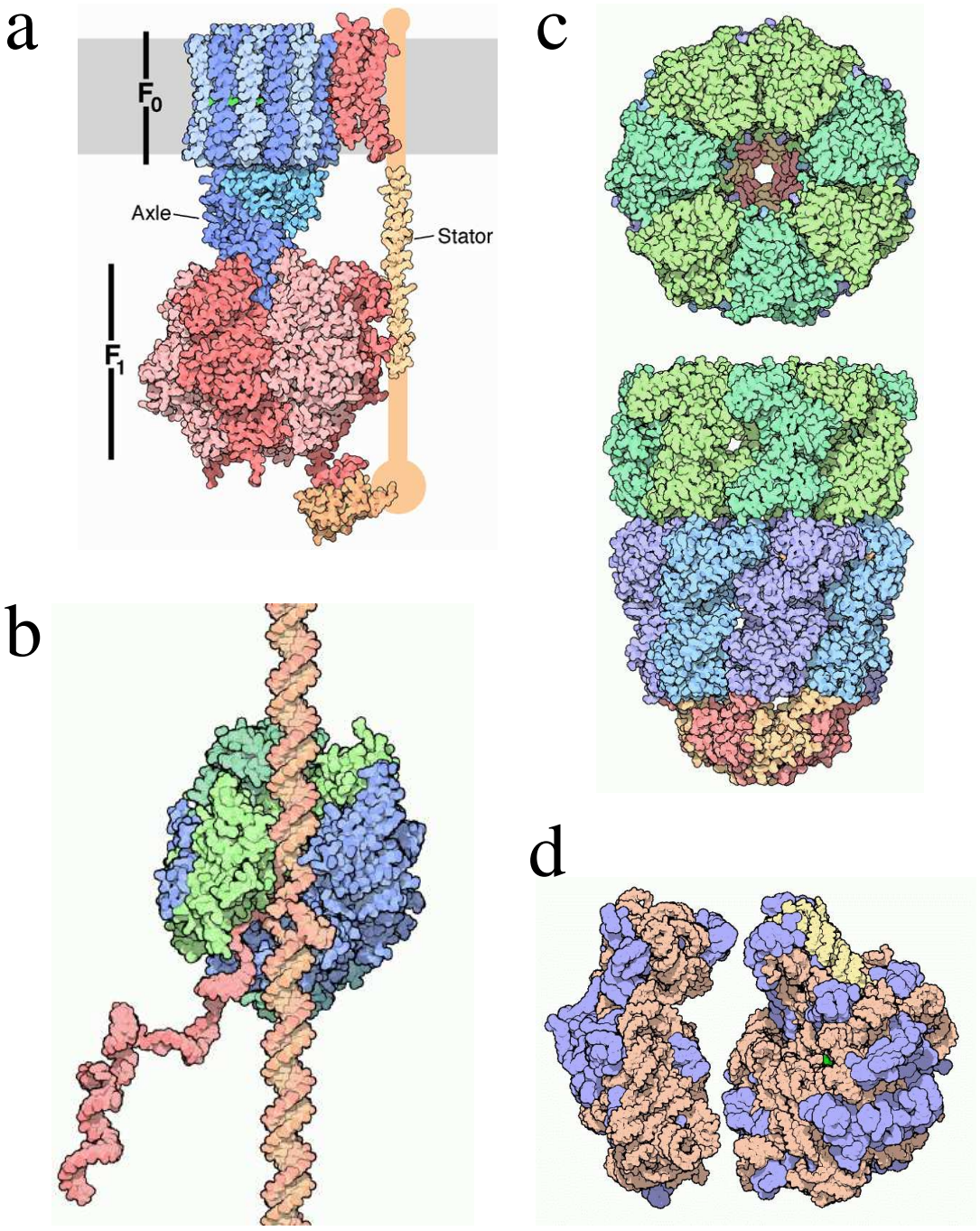


Figure 1.1: Four molecular machines formed principally by proteins. Figures taken from the *Molecule of the month* section of the RSCB Protein Data Bank (http://www.pdb.org), we thank the RSCB PDB and David S. Goodsell, from the Scripps Research Institute, for kind permission to use them. **(a)** *ATP synthase*: it acts as an energy generator when it is traversed by protons that make its two coupled engines rotate in reverse mode and the ATP molecule is produced. **(b)** *RNA polymerase*: it slides along a thread of DNA reading the base pairs and synthesizing a matching copy of RNA. **(c)** *GroEL-GroES complex*: it helps unfolded proteins to fold by sheltering them from the overcrowded cellular cytoplasm. **(d)** *Ribosome*: it polymerizes amino acids to form proteins following the instructions written in a thread of mRNA.

the cell, the *ATP synthase* (fig. 1.1a); about the genetic Xerox machine, the *RNA polymerase* (fig. 1.1b); about the hut where the proteins fold under shelter, the *GroEL-GroES* complex (fig. 1.1c); or about the macromolecular factory where proteins are created, the *ribosome* (fig. 1.1d), to quote four specially impressive examples. Agreeing again with Lesk [10]:

> Proteins are fascinating molecular devices.

From a more academic standpoint, proteins are proving to be a powerful center of interdisciplinary research, making many diverse fields and people with different formations come in contact[2]. Proteins force biologists, biochemists and chemists to learn more physics, mathematics and computation and force mathematicians, physicists and computer technicians to learn more biology, biochemistry and chemistry. This, indeed, cannot be negative.

In 2005, in a special section of the Science magazine entitled "What don't we know?" [27], a selection of the hundred most interesting yet unanswered scientific questions was presented. What indicates the role of proteins, and particularly of the protein folding problem (treated in sec. 1.3), as focuses of interdisciplinary collaboration is not the inclusion of the question *Can we predict how proteins will fold?*, which was a must, but the large number of other questions which were related to or even dependent on it, such as *Why do humans have so few genes?*, *How much can human life span be extended?*, *What is the structure of water?*, *How does a single somatic cell become a whole plant?*, *How many proteins are there in humans?*, *How do proteins find their partners?*, *How do prion diseases work?*, *How will big pictures emerge from a sea of biological data?*, *How far can we push chemical self-assembly?* or *Is an effective HIV vaccine feasible?*.

In this direction, probably the best example of the use that protein science makes of the existing human expertise, and of the positive feedback that this brings up in terms of new developments and resources, can be found in the machines that every one of us has on his/her desktops. In a first step, the enormous amount of biological data that emerges from the sequencing of the genomes of different living organisms requires computerized databases for its proper filtering. The NCBI GenBank database[3], which is one of the most exhaustive repositories of sequenced genetic material, has doubled the number of deposited DNA bases approximately every 18 months since 1982 (see fig. 1.2a) and has recently (in August 2005) exceeded the milestone of 100 Gigabases (10^{11}) from over 165,000 species.

Among them, and according to the Entrez Genome Project database[4], the sequencing of the complete genome of 366 organisms has been already achieved and there are 791 more to come in next few years. In the group of the completed ones, most are bacteria, and there are only two mammals: the poor laboratory mouse, *Mus Musculus*, and, notably [28], the *Homo Sapiens* (with $\sim 3 \cdot 10^9$ bases and a mass-media-broadcast battle between the private firm Celera and the public consortium IHGSC).

However, not all the DNA encodes proteins (not all the DNA is genes). Typically, more than 95% of the genetic material in living beings is *junk DNA*, also called *non-coding*

[2] The Institute for Biocomputation and Physics of Complex Systems, which P. Echenique and his Ph.D. advisor, J. L. Alonso are part of, is a local example of this rather new form of collaboration among scientists.

[3] http://www.ncbi.nlm.nih.gov/Genbank/

[4] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genomeprj

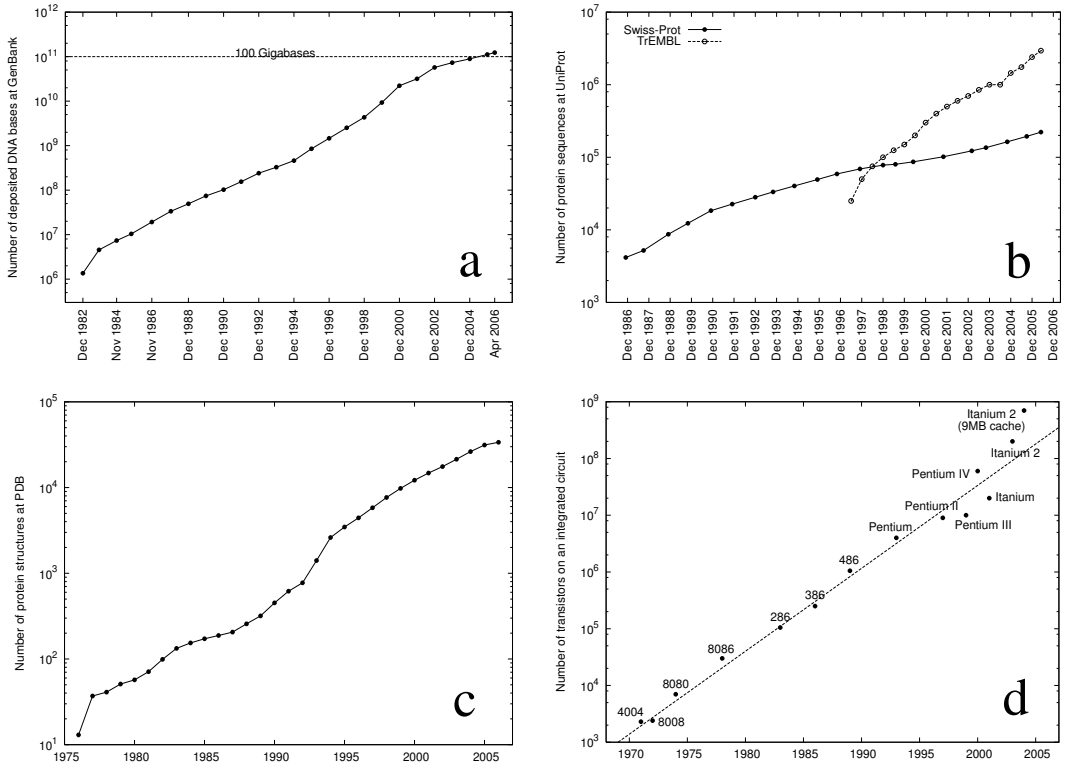


Figure 1.2: Recent exponential progress in genomics, proteomics and computer technology. **(a)** Evolution of the number of DNA bases deposited at the GenBank database. **(b)** Evolution of the number of protein sequences at the UniProt Swiss-Prot and TrEMBL databases. **(c)** Evolution of the number of protein three-dimensional structures at the Protein Data Bank. **(d)** Moore's Law: evolution of the number of transistors in the Intel CPUs.

DNA (a more neutral term which seems recommendable in the light of some recent discoveries [29–31]). So, in a second step, the coding regions must be identified and each gene translated into the amino acid sequence of a particular protein[5]. The UniProt database[6] is, probably, the most comprehensive repository of these translated protein sequences and also of others coming from a variety of sources, including direct experimental determination [34, 35]. UniProt is comprised by two different sub-databases: the Swiss-Prot Protein Knowledgebase, which contains extensively human-annotated protein sequences with low redundancy; and TrEMBL, which contains computer-annotated sequences extracted directly from the underlying nucleotide entries at databases such as GenBank and where only the most basic redundancies have been removed.

The UniProt/Swiss-Prot database contains, at the moment (on 30 May 2006), around 200,000 protein sequences from about 10,000 species, and it has experienced an exponential growth (since 1986), doubling the number of records approximately every 41 months

[5] Note that many variations [32, 33] may occur before, during and after the process of gene expression, so that the relation gene-to-protein is not one-to-one. The size of the human proteome (the number of different proteins), for example, is estimated to be an order of magnitude or two larger than the size of the genome.

[6] http://www.uniprot.org

(see fig. 1.2b). In turn, the UniProt/TrEMBL database contains almost 3 million protein sequences from more than 100,000 species, and its growth (from 1997) has also been exponential, doubling the number of records approximately every 16 months (see fig. 1.2b).

After knowing the sequence of a protein, the next step towards the understanding of biological processes is the characterization of its three-dimensional structure. Most proteins perform their function under a very specific *native* shape which involves many twists, loops and bends of the linear chain of amino acids (see sec. 1.3). This spatial structure is much more important than the sequence for biochemists to predict and understand the mechanisms of life and it can be resolved, nowadays, by fundamentally two experimental techniques: for small proteins, nuclear magnetic resonance (NMR) [36, 37] and, more commonly, for proteins of any size, x-ray crystallography [38–40]. The three-dimensional structures so obtained are deposited in a centralized public-access database called Protein Data Bank (PDB)[7] [41]. From the 13 structures deposited in 1976 to the 33,782 (from more than a thousand species) stored in June 2006, the growth of the PDB has been (guess?) exponential, doubling the number of records approximately every 3 years (see fig. 1.2c).

To summarize, in June 2006, we have sequenced partial segments of the genetic material of around 160,000 species, having completed the genomes of only 366; we know the sequences of some of the proteins of around 100,000 species and the three-dimensional structure of proteins in 1,103 species[8]. However, according to the UN Millennium Ecosystem Assessment[9], the number of species formally identified is 1.7-2 million and the estimated total number of species on Earth ranges from 5 million to 30 million [42]. Therefore, we should expect that the exponential growth of genomic and proteomic data will continue to fill the hard-disks, collapse the broadband connexions and heat the CPUs of our computers at least for the next pair of decades.

Fortunately, the improvement of silicon technology behaves in the same way: In fact, in 1965, Gordon Moore, co-founder of Intel, made the observation that the number of transistors per square inch had doubled every year since the integrated circuit was invented, and predicted that this exponential trend would continue for the foreseeable future. This has certainly happened (although the doubling time seems to be closer to 18 months) and this empirical law, which is not expected to fail in the near future, has become to be known as *Moore's Law* (see fig. 1.2d for an example involving Intel processors). So we do not have to worry about running short of computational resources!

Of course, information produces more information, and the public databases do not end at the three-dimensional structures of proteins. In the last few years, a number of more specific web-based repositories have been created in the field of molecular biology. There is the Protein Model Database (PMDB)[10] [43], where theoretical three-dimensional protein models are stored (including all models submitted to last four editions of the CASP[11] experiment [44]); the ProTherm[12] and ProNIT[13] databases [45], where a wealth of thermodynamical data is stored about protein stability and protein-nucleic acid in-

[7] http://www.rcsb.org/pdb/
[8] http://www.ebi.ac.uk/thornton-srv/databases/pdbsum/
[9] http://www.millenniumassessment.org
[10] http://www.caspur.it/PMDB/
[11] http://predictioncenter.gc.ucdavis.edu
[12] http://gibk26.bse.kyutech.ac.jp/jouhou/Protherm/protherm.html
[13] http://gibk26.bse.kyutech.ac.jp/jouhou/pronit/pronit.html

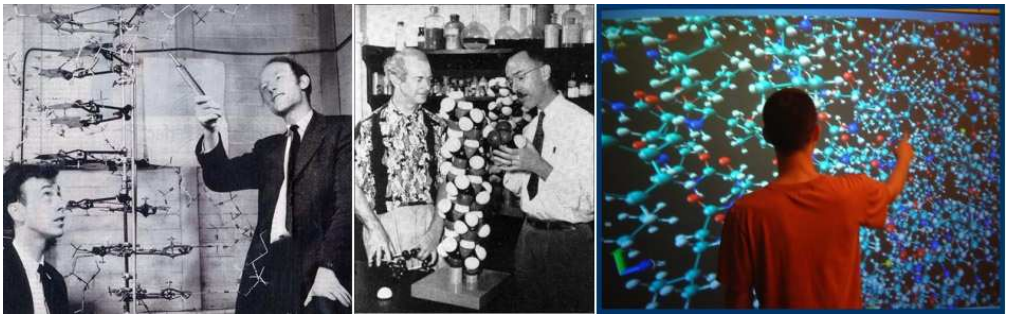


Figure 1.3: Molecular models. Compare the arduously hand-made models of Watson and Crick's DNA double helix (left) or Pauling and Corey's protein α-helix (center) with the three-dimensional structures that can be drawn in milliseconds by any of the nowadays available molecular visualization programs (right).

teractions, respectively; the dbPTM[14] database [46], that stores information on protein post-translational modifications; the PINT[15] database [47], with thermodynamical data on protein-protein interactions; and so on and so forth.

In addition to the use of computers for storage and retrieval of enormous quantities of data, the increasing numerical power of these machines is customarily used for a wide variety of applications that range from molecular visualization (see fig. 1.3), to long simulations aimed to solve the equations governing biological systems (a topic discussed more in detail in the rest of this dissertation).

Indeed, as Richard Dawkins has stated [48]:

> What is truly revolutionary about molecular biology in the post-Watson-Crick era is that it has become digital.

Finally, apart from all the convincing reasons and the appeals to authority given above, what is crystal-clear is that proteins are an unsolved and difficult enigma. And those are two irresistible qualities for any flesh and blood scientist.

1.2 Summary of protein structure

The main character of the story

In spite of their diverse biological functions, summarized in the previous section, proteins are a rather homogeneous class of molecules from the chemical point of view. They are *linear heteropolymers*, i.e., unbranched chains of different identifiable monomeric units.

Before they are assembled into proteins, these building units are called *amino acids* and can exist as standalone stable molecules. All amino acids are made up of a central *α-carbon* with four groups attached to it: an amino group (—NH_2), a carboxyl group (—COOH), a hydrogen atom and a fourth arbitrary group (—R) (see fig. 1.5). In aqueous solvent and under physiological conditions, both the amino and carboxyl groups are

[14] http://dbPTM.mbc.nctu.edu.tw
[15] http://www.bioinfodatabase.com/pint/

charged, the first accepting one proton and getting a positive charge, and the second giving one proton away and getting a negative charge (compare figs. 1.5a and 1.5c).

When the group —R is not equal to one of the other three groups attached to the α-carbon, the amino acid is *chiral*, i.e., like our hands, it may exist in two different forms, which are mirror images of one another and cannot be superimposed by rotating one of them in space (you cannot wear the left-hand glove on your right hand). In chemical jargon, one says that the α-carbon constitutes an *asymmetric center* and that the amino acid may exist as two different *enantiomers* called *L-* (fig. 1.5c) and *D-* (fig. 1.5d) forms. It is common that, when used as prefixes, the L and D letters, which come from *levorotatory* and *dextrorotatory*, are written in small capitals, as in L- and D-. This nomenclature is based on the possibility of associating the amino acids to the optically active L- and D- enantiomers of glyceraldehyde, and could be related to the +/- or to the Cahn, Ingold and Prelog's R/S [50] notations. For us, it suffices to say that the D/L nomenclature is, by far, the most used one in protein science and the one that will be used in this document. For further details, take a look at the IUPAC recommendations at http://www.chem.qmul.ac.uk/iupac/AminoAcid/.

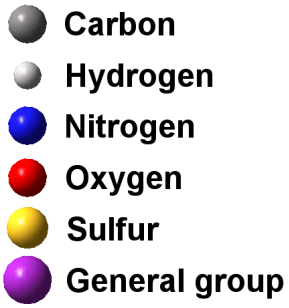


Figure 1.4: Color and size code for the atom types used in most of the figures in this section. All the figures have been made with the Gaussview graphical front-end of Gaussian03 [49] and then modified with standard graphical applications.

In principle, amino acids may be L- or D-, and the group —R may be anything provided that the resultant molecule is stable. However, for reasons that are still unclear [51], the vast majority of proteins in all living beings are made up of L-amino acids (as a rare exception, we may point out the fact that D-amino acids can be found in some proteins produced by exotic sea-dwelling organisms, such as *cone snails*) and the groups —R

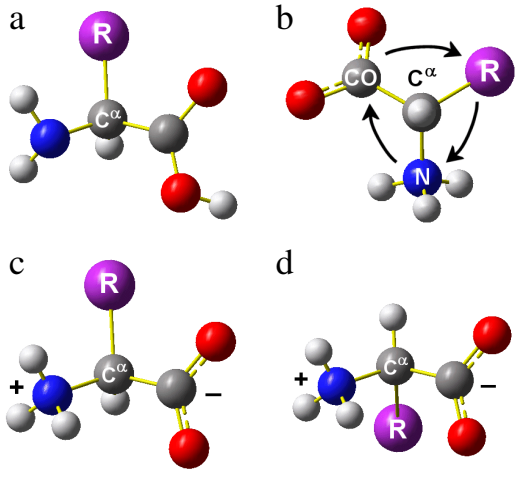


Figure 1.5: Amino acids. **(a)** Uncharged L-enantiomer. **(b)** CORN mnemotechnic rule to remember which one is the L-form. **(c)** Charged L-enantiomer (the predominant form found in living beings). **(d)** Charged D-enantiomer.

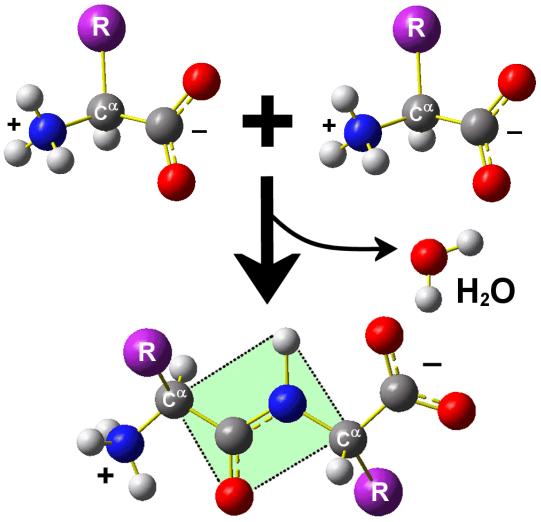


Figure 1.6: Peptide bond formation reaction. The peptide plane is indicated in green.

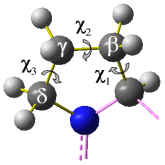
(called *side chains*) that are coded in the genetic material comprise a set of only twenty possibilities (depicted in fig. 1.7).

A frequently quoted mnemotechnic rule for remembering which one is the L-form of amino acids is the so-called *CORN rule* in fig. 1.5b. According to it, one must look from the hydrogen to the α-carbon and, if the three remaining groups are labeled as in the figure, the word *CORN* must be read in the clockwise sense of rotation. The author of this Ph.D. dissertation does not find this rule very useful, since normally he cannot recall if the sense is clockwise or counterclockwise. To know which form is the L- one, he draws the amino acid as in fig. 1.5a or 1.5c, with the α-carbon in the center, the amino group on the left and the carboxyl group on the right, all of them in the plane of the paper (which is very natural and easy to remember because it matches the normal sense of writing with the fact that, conventionally, proteins start at —NH_3^+ and end at —COO^-). Finally, he must just remember that the side chain of the L-amino acid goes out of the paper approaching the reader (which is also natural because the side chain is the relevant piece of information and we want to look at it closely).

The process through which amino acids are assembled into proteins (called *gene expression* or *protein biosynthesis*) is typically divided in two steps. In the first one, the *transcription*, the enzyme ARN polymerase (see fig. 1.1b) binds to the DNA in the cellular nucleus and makes a copy of a section –the *gene*– of the base sequence into a messenger RNA (mRNA) molecule. In the second step, called *translation*, the mRNA enters the ribosome (see fig. 1.1d) and is read stopping at each base triplet (called *codon*). Now, a specific molecule of transfer RNA (tRNA), which possesses the base triplet (called *anticodon*) that is complementary to the codon, links to the mRNA bringing with her the amino acid that is codified by the particular sequence of three bases. Each amino acid that arrives to the ribosome in this way is covalently attached to the previous one and so added to the nascent protein. In this reaction, the *peptide bond* is formed and a water molecule is released (see fig. 1.6). This process continues until a stop codon is read and the transcription is complete.

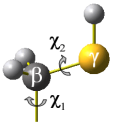
Special

G Gly Glycine

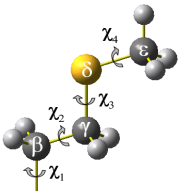


P Pro Proline

Sulfur–containing



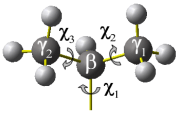
C Cys Cysteine



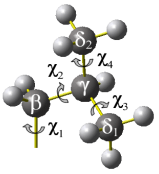
M Met Methionine

Aliphatic

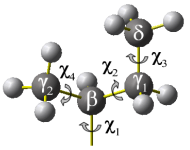
A Ala Alaline



V Val Valine

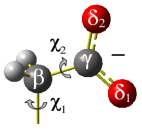


L Leu Leucine

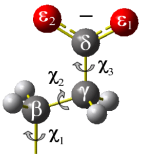


I Ile Isoleucine

Acid

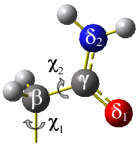


D Asp Aspartic acid

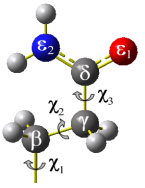


E Glu Glutamic acid

Amides

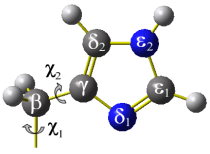


N Asn Asparagine

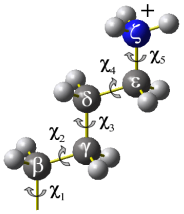


Q Gln Glutamine

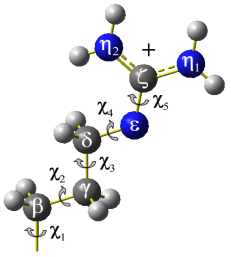
Basic



H His Histidine

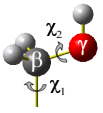


K Lys Lysine



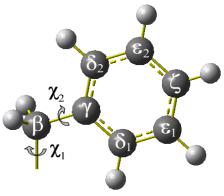
R Arg Arginine

Alcohols

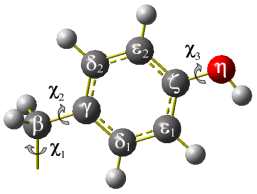


S Ser Serine

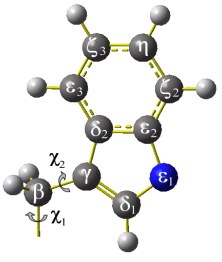
Aromatic



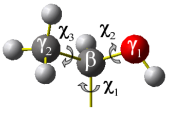
F Phe Phenylalanine



Y Tyr Tyrosine



W Trp Tryptophan



T Thr Threonine

The amino acid sequence of the resultant protein, read from the *amino terminus* to the *carboxyl terminus*, is called *primary structure*; and the amino acids included in such a polypeptide chain are normally termed *amino acid residues*, or simply *residues*, in order to distinguish them from their isolated form. The main chain formed by the repetition of α-carbons and the C' and N atoms at the peptide bond is called *backbone* and the —R groups branching out from it are called *side chains*, as it has already been mentioned.

The specificity of each protein is provided by the different properties of the twenty side chains in fig. 1.7 and their particular positions in the sequence. In textbooks, it is customary to group them in small sets according to different criteria in order to facilitate their learning. Classifications devised on the basis of the physical properties of these side chains may be sometimes overlapping (e.g., tryptophan contains polar regions as well as an aromatic ring, which, in turn, could be considered hydrophobic but is also capable of participating in, say, π-π interactions). Therefore, for a clearer presentation, we have chosen here to classify the residues according to the chemical groups contained in each side chain and discuss their physical properties individually.

Let us enumerate then the categories in fig. 1.7 and point out any special remark regarding the residues in them:

- **Special residues**:

  Glycine is the smallest of all the amino acids: its side chain contains only a hydrogen atom. So, since its α-carbon has two hydrogens attached, glycine is the only achiral natural amino acid. Its affinity for water is mainly determined by the peptide groups in the backbone; therefore, glycine is hydrophilic.

  Proline is the only residue whose side chain is covalently linked to the backbone (the backbone is indicated in purple in fig. 1.7), giving proline unique structural properties that will be discussed later. Since its side chain is entirely aliphatic, proline is hydrophobic.

- **Sulfur-containing residues**:

  Cysteine is a very important structural residue because, in a reaction catalyzed by *protein disulfide isomerases* (PDIs), it may form, with another cysteine, a very

Figure 1.7: Side chains of the twenty amino acid residues encoded in the genetic material of living beings. They have been classified according to the chemical groups they contain. The rotameric degrees of freedom χ_i are indicated with small arrows over the bonds. The name of the heavy atoms and the numbering of the branches comply with the IUPAC rules (http://www.chem.qmul.ac.uk/iupac/AminoAcid/). Below the molecular structure, the one letter code (green), the three letter code (red) and the complete name (blue) of each amino acid may be found. In the case of proline, the N and the α-carbon have been included in the scheme, and the backbone bonds have been coloured in purple. The titratable residues Asp, Glu, Lys and Arg have been represented in their charged forms, which is the most common one in aqueous solvent under physiological conditions. Histidine is shown in its neutral ε_2-tautomeric form.

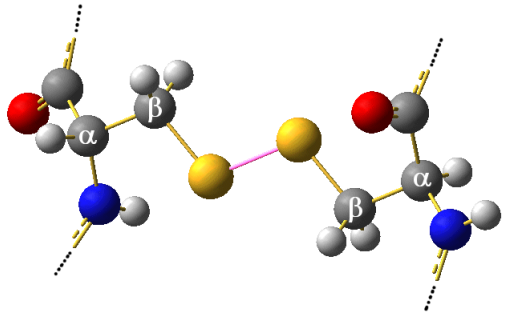


Figure 1.8: Disulfide bond between two cysteine residues.

stable covalent bond called *disulfide bond* (see fig. 1.8). Curiously, all the L-amino acids are S-enantiomers according to the Cahn, Ingold and Prelog rules [50] except for cysteine, which is R-. This is probably the reason that makes the D/L nomenclature favourite among protein scientists [33]. Cysteine is a polar residue.

Methionine is mostly aliphatic and, henceforth, apolar.

- **Aliphatic residues**:

  Alanine is the smallest chiral residue. This is the fundamental reason for using alanine models, more than any other ones, in the computationally demanding ab initio studies of peptides that are customarily performed in quantum chemistry (see chapter 7). It is hydrophobic, like all the residues in this group.

  Valine is one of the three *β-branched* residues (i.e., those that have more than one heavy atom attached to the β-carbon, apart from the α-carbon), together with isoleucine and threonine. It is hydrophobic.

  Leucine is hydrophobic.

  Isoleucine's β-carbon constitutes an asymmetric center and the only enantiomer that occurs naturally is the one depicted in the figure. Only isoleucine and threonine contain an asymmetric center in their side chain. Isoleucine is β-branched and hydrophobic.

- **Acid residues**:

  Aspartic acid is normally charged under physiological conditions. Hence, it is very hydrophilic.

  Glutamic acid is just one CH_2 larger than aspartic acid. Their properties are very similar.

- **Amides**:

  Asparagine contains a chemical group similar to the peptide bond. It is polar and can act as a hydrogen bond donor or acceptor.

  Glutamine is just one CH_2 larger than asparagine. Their physical properties are very similar.

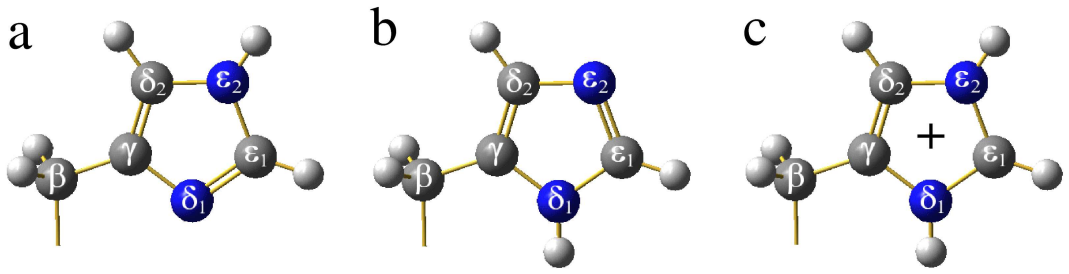


Figure 1.9: Three forms of histidine found in proteins. **(a)** Neutral ε_2-tautomer. **(b)** Neutral δ_1-tautomer. **(c)** Charged form.

- **Basic residues**:

  Histidine is a special amino acid: in its neutral form, it may exist as two different tautomers, called δ_1 and ε_2, depending on which nitrogen has an hydrogen atom attached to it. The ε_2-tautomer has been found to be slightly more stable in model dipeptides [52], although both forms are found in proteins. Histidine can readily accept a proton and get a positive charge, in fact, it is the only side chain with a pKa in the physiological range, so non-negligible proportions of both the charged and uncharged forms are typically present. Of course, histidine is hydrophilic.

  Lysine's side chain is formed by a rather long chain of CH_2 with an amino group at its end, which is nearly always positively charged. Therefore, lysine is very polar and hydrophilic.

  Arginine's properties are similar to those of lysine, although its terminal guanidinium group is a stronger basis than the amino group and it may also participate in hydrogen bonds as a donor.

- **Alcohols**:

  Serine is one of the smallest residues. It is polar due to the hydroxyl group.

  Threonine's β-carbon constitutes an asymmetric center; the enantiomer that occurs in living beings is the one shown in the figure. The physical properties of threonine are very similar to those of serine.

- **Aromatic residues**:

  Phenylalanine is the smallest aromatic residue. Its benzyl side chain is largely apolar and interacts unfavourably with water. It may also participate in specific π-stacking interactions with other aromatic groups.

  Tyrosine's properties are similar to those of phenylalanine, being only slightly more polar due to the presence of a hydroxyl group.

  Tryptophan, with 17 atoms in her side chain, is the largest residue. It is mainly hydrophobic, although it contains a small polar region and it can also participate in π-π interactions, like all the residues in this category.

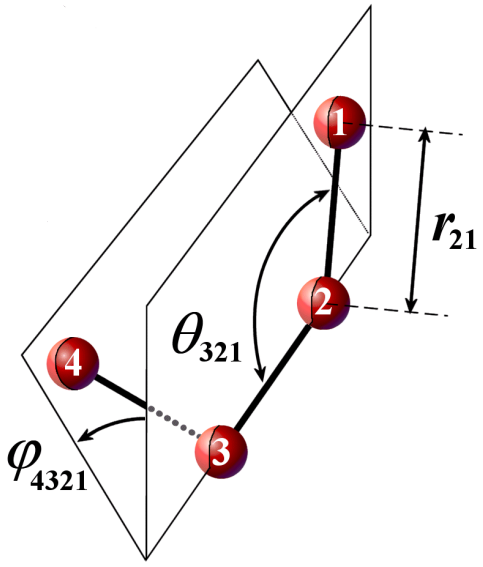


Figure 1.10: Typical definition of internal coordinates. r_{21} is the *bond length* between atoms 2 and 1. θ_{321} is the *bond angle* formed by the bonds (2,1) and (3,2), it ranges from 0 to 180°. Finally φ_{4321} is the *dihedral angle* describing the rotation around the bond (3,2); it is defined as the angle formed by the plane containing atoms 1, 2 and 3 and the plane containing atoms 2, 3 and 4; it ranges either from −180° to 180° or from 0° to 360°, depending on the convention; the positive sense of rotation for φ_{4321} is the one indicated in the figure. Also note that the definition is symmetric under a complete change in the order of the atoms, in such a way that, quite trivially, $r_{21} = r_{12}$ and $\theta_{123} = \theta_{321}$, but also, not so trivially, $\varphi_{4321} = \varphi_{1234}$.

Well then, after having introduced the building blocks of proteins, some qualifying remarks about them are worth to be done.

On one side, why amino acids encoded in DNA codons are the ones in the list or why there are exactly twenty of them are questions that are still subjects of controversy [53, 54]. In fact, although the side chains in fig. 1.7 seem to confer enough versatility to proteins in most cases, there are also rare exceptions in which other groups are needed to perform a particular function. For example, the amino acid *selenocysteine* may be incorporated into some proteins at an UGA codon (which normally indicates a stop in the transcription), or the amino acid *pyrrolysine* at an UAG codon (which is also a stop indication in typical cases). In addition, the arginine side chain may be post-translationally converted into *citrulline* by the action of a family of enzymes called *peptidylarginine deiminases* (PADs).

On the other hand, the chemical (covalent) structure of the protein chain may suffer from more complex modifications than just the inclusion of non-standard amino acid residues: A myriad of organic molecules may be covalently linked to specific points, the chain may be cleaved (cut), chemical groups may be added or removed from the N- or C-termini, disulfide bonds may be formed between cysteines, and the side chains of the residues may undergo chemical modifications just like any other molecule [52]. The vast majority of these changes either depend on the existence of some chemical agent external to the protein, or are catalyzed by an enzyme.

In this dissertation, our interest is in the folding of proteins. This problem, which will be discussed in detail in the next section, is so huge and so difficult that there is no point in worrying about details, such as the ones mentioned in the two preceding paragraphs, before the big picture is at least preliminarily understood. Therefore, when we talk about the folding of proteins in what follows, we will be thinking about single polypeptide chains, made up of L-amino acids, in water and without any other reagent present, with

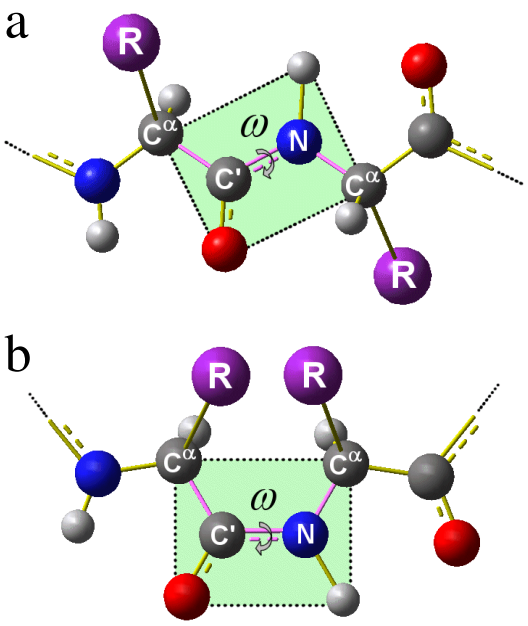


Figure 1.11: Trans and cis conformations of the peptide plane. The bonds defining the peptide bond dihedral angle ω are indicated in purple. **(a)** *Trans* conformation ($\omega \simeq \pm 180^{o}$). The most common one in proteins. **(b)** *Cis* conformation ($\omega \simeq 0^{o}$). Significantly found only in Xaa-Pro bonds.

the side chains chosen from the set in fig. 1.7, and having underwent no post-translational modifications nor any chemical change on their groups. Finally, although some simple modifications, such as the formation of disulfide bonds or the trans $\rightarrow$ cis isomerization of Xaa-Pro peptide bonds (see what follows), could be more easily included in the first approach to the problem, we shall also leave them for a later stage.

Now, with this considerations, we have fixed the covalent structure of our molecule as well as the enantiomerism of the asymmetric centers it may contain. This information is enough to specify the three-dimensional arrangement of the atoms of small rigid molecules. However, long polymers and, particularly, proteins, possess degrees of freedom (termed *soft*) that require small amounts of energy to be changed while drastically altering the relative positions of groups and atoms. In a first approximation, all bond lengths, bond angles and dihedral angles describing rotations around triple, double and partial double bonds (see fig. 1.10) may be considered to be determined by the covalent structure. Whereas dihedral angles describing rotations around single bonds may be considered to be variable and soft. The non-superimposable three dimensional arrangements of the molecule that correspond to different values of the soft degrees of freedom are called *conformations*.

In proteins, some of these soft dihedrals are located at the side chains; they are the χ_i in fig. 1.7 and, although they are important in the later stages of the folding process and must be taken into account in any ambitious model of the system, their variation only alters the conformation locally. On the contrary, a small change in the dihedral angles located at the backbone of the polypeptide chain may drastically modify the relative position of many pairs of atoms and they must be given special attention.

That is why, the special properties of the peptide bond, which is the basic building block of the backbone, are very important to understand the conformational behaviour of proteins. These properties arise from the fact that there is an electron pair delocalized between the C—N and C—O bonds (using the common chemical image of *resonance*),

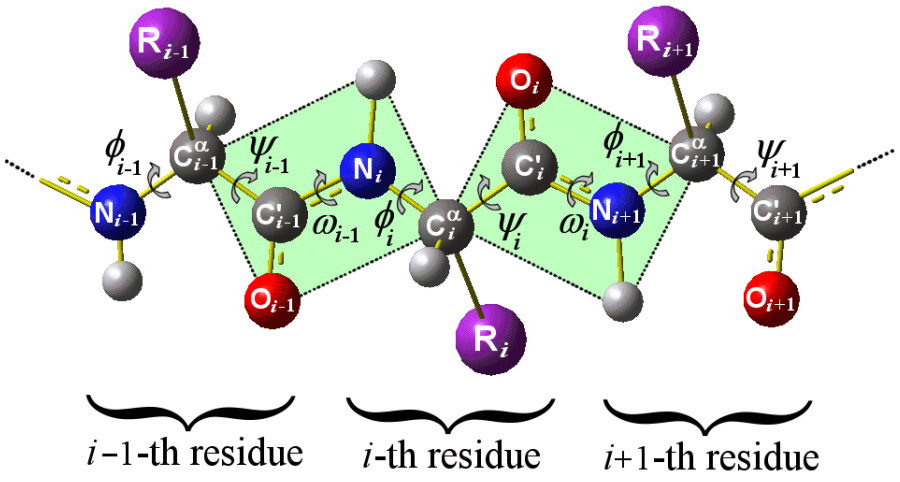


Figure 1.12: Numeration of the heavy atoms and the dihedrals angles describing rotations around backbone bonds. In agreement with IUPAC recommendations (see http://www.chem.qmul.ac.uk/iupac/AminoAcid/). The peptide planes are indicated as green rectangles.

which provokes that neither bond is single nor double, but *partial double bonds* that have a mixed character. In particular, the partial double bond character of the peptide bond is the cause that the six atoms in the green plane depicted in figs. 1.6, 1.11 and 1.12 have a strong tendency to be coplanar, forming the so-called *peptide plane*. This coplanarity allows for only two different conformations: the one called *trans* (corresponding to $\omega \simeq \pm 180^{\circ}$), in which the α-carbons lie at different sides of the line containing the C—N bond; and the one called *cis* (corresponding to $\omega \simeq 0^{\circ}$), in which they lie at the same side of that line (see fig. 1.11).

Although the quantitative details are not completely elucidated yet and the very protocol of protein structure determination by x-ray crystallography could introduce spurious effects in the structures deposited in the PDB [55], it seems clear that a great majority of the peptide bonds in proteins are in the trans conformation. Indeed, a superficial look at the two forms in fig. 1.11 suggests that the steric clashes between substituents of consecutive α-carbons will be more severe in the cis case. When the second residue is a Proline, however, the special structure of its side chain makes the probability of finding the cis conformer significantly higher: For Xaa-nonPro peptide bonds in native structures, the trans form is more common than the cis one with approximately a 3000:1 proportion; while this ratio decreases to just 15:1 if the bond is Xaa-Pro [55].

In any case, due to the aforementioned partial double bond character of the C—N bond, the rotation barrier connecting the two states is estimated to be of the order of ~ 20 kcal/mol [56], which is about 40 times larger than the thermal energy at physiological conditions, thus rendering the spontaneous trans $\rightarrow$ cis isomerization painfully slow. However, mother Nature makes use of every possibility that she has at hand and, sometimes, there are a few peptide bonds that must be cis in order for the protein to fold correctly or to function properly. Since all peptide bonds are synthesized trans at the ribosome [57], the trans $\rightarrow$ cis isomerization must be catalyzed by enzymes (called *peptidylprolyl isomerases* (PPIs)) and, in the same spirit of the post-translational modifi-

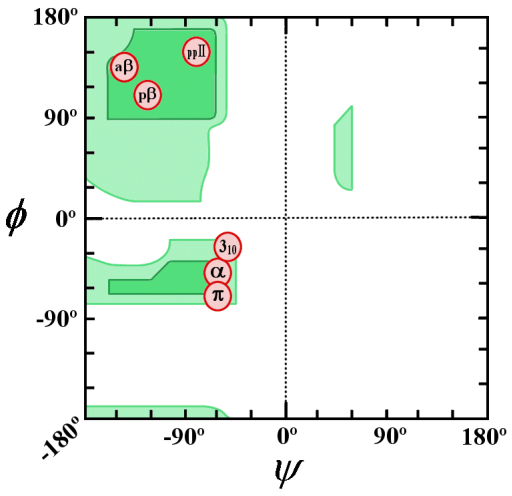


Figure 1.13: Original Ramachandran plot drawn by Ramachandran and Ramakrishnan in 1963 [58]. In dark-green, the fully allowed regions, calculated by letting the atoms approach to the average clashing distance; in light-green, the partially allowed regions, calculated by letting the atoms approach to the minimum clashing distance; in white, the disallowed regions. Some points representing secondary structure elements are shown as red circles at the ideal (ϕ, ψ)-positions in table 1.1: ($\boldsymbol{\alpha}$) α-helix. ($\boldsymbol{\pi}$) π-helix. ($\mathbf{3_{10}}$) 3_{10}-helix. (**a**$\boldsymbol{\beta}$) Antiparallel β-sheet. (**p**$\boldsymbol{\beta}$) Parallel β-sheet. (**ppII**) Polyproline II.

cations discussed before, this step may be taken into account in a later refinement of the models.

So, we shall assume in what follows that all peptide bonds (even the Xaa-Pro ones) are in the trans state and, henceforth, the conformation of the protein will be essentially determined by the values of the ϕ and ψ angles, which describe the rotation around the two single bonds next to each α-carbon (see fig. 1.12 for a definition of the dihedral angles associated to the backbone).

This assumption was introduced, as early as 1963, by Ramachandran and Ramakrishnan [58] and the ϕ and ψ coordinates are commonly named *Ramachandran angles* after the first one of them. In their famous paper [58], they additionally suppose that the bond lengths, bond angles and dihedral angles on double and partial double bonds are fixed and independent of ϕ and ψ, they define a typical distance up to which a specific pair of atoms may approach and also a minimum one (taken from statistical studies of structures) and they draw the first *Ramachandran plot* (see fig. 1.13): A depiction of the regions in the (ϕ, ψ)-space that are energetically allowed or disallowed on the basis of the local sterical clashes between atoms that are close to the α-carbon.

One of the main advantages of this type of diagrams as 'thinking tools' lies in the fact that (always in the approximation that the non-Ramachandran variables are fixed) some very common repetitive structures found in proteins may be ideally depicted as a single point in the plot. In fact, these special conformations, which are regarded as the next level of protein organization after the primary structure and are said to be elements of *secondary structure*, may be characterized exactly like that, i.e., by asking that a certain number of consecutive residues present the same values of the ϕ and ψ angles. In the book by Lesk [10], for example, one may found a table with the most common of these repetitive patterns, together with the corresponding (ϕ, ψ)-values taken from statistical investigations of experimentally resolved protein structures (see table 1.1).

However, the non-Ramachandran variables are not really constant, and the elements of secondary structure do possess a certain degree of flexibility. Moreover, the side chains may interact and exert different strains at different points of the chain, which provokes that, in the end, the secondary structure elements gain some stability by slightly altering

	ϕ	ψ
α-helix	−57	−47
3_{10}-helix	−49	−26
π-helix	−57	−70
polyproline II	−79	149
parallel β-sheet	−119	113
antiparallel β-sheet	−139	135

Table 1.1: Ramachandran angles (in degrees) of some important secondary structure elements in polypeptides. Data taken from ref. 10.

their ideal Ramachandran angles. Therefore, it is more appropriate to characterize them according to their hydrogen-bonding pattern, which, in fact, is the feature that makes these structures prevalent, providing them with more energetic stability than other repetitive conformations which are close in the Ramachandran plot.

The first element of secondary structure that was found is the *α-helix*. It is a coil-like[16] structure, with ∼ 3.6 residues per turn, in which the carbonyl group (C=O) of each i-th residue forms a hydrogen bond with the amino group (N–H) of the residue $i + 4$ (see fig. 1.14b). According to a common notation, in which x_y designates a helix with x residues per turn and y atoms in the ring closed by the hydrogen bond [59], the α-helix is also called 3.6_{13}-*helix*.

She was theoretically proposed in 1951 by Pauling, Corey and Branson [60] (see fig. 1.3b), who used precise information about the geometry of the non-Ramachandran variables, taken from crystallographic studies of small molecules, to find the structures compatible with the additional constraints that: (i) the peptide bond is planar, and (ii) every carbonyl and amino group participates in a hydrogen bond.

The experimental confirmation came from Max Perutz, who, together with Kendrew and Bragg, had proposed in 1950 (one year before Pauling's paper) a series of helices with an integer number of residues per turn [59] that are not so commonly found in native structures of proteins (see however, the discussion about the 3_{10}-helix below). Perutz read Pauling, Corey and Branson's paper one Saturday morning [61] in spring 1951 and realized immediately that their helix looked very well: free of strain and with all donor and acceptor groups participating in hydrogen bonds. So he rushed to the laboratory and put a sample of horse hair (rich in keratin, a protein that contains α-helices) in the x-ray beam, knowing that, according to diffraction theory, the regular repeat of the 'spiral staircase steps' in Pauling's structure should give rise to a strong x-ray reflection of 1.5 Å spacing from planes perpendicular to the fiber axis. The result of the experiment was positive[17]

[16] Here, we use the word "coil" to refer to the twisted shape of a telephone wire, a corkcrew or the solenoid of an electromagnet. Although this is common English usage, the same word occurs frequently in protein science to designate different (and sometimes opposed) concepts. For example, a much used ideal model of the denatured state of proteins is termed *random coil*, and a popular statistical description of helix formation is called *helix-coil theory*.

[17] Linus Pauling was awarded the Nobel prize in chemistry in 1954 "for his research into the nature of the chemical bond and its application to the elucidation of the structure of complex substances", and Max

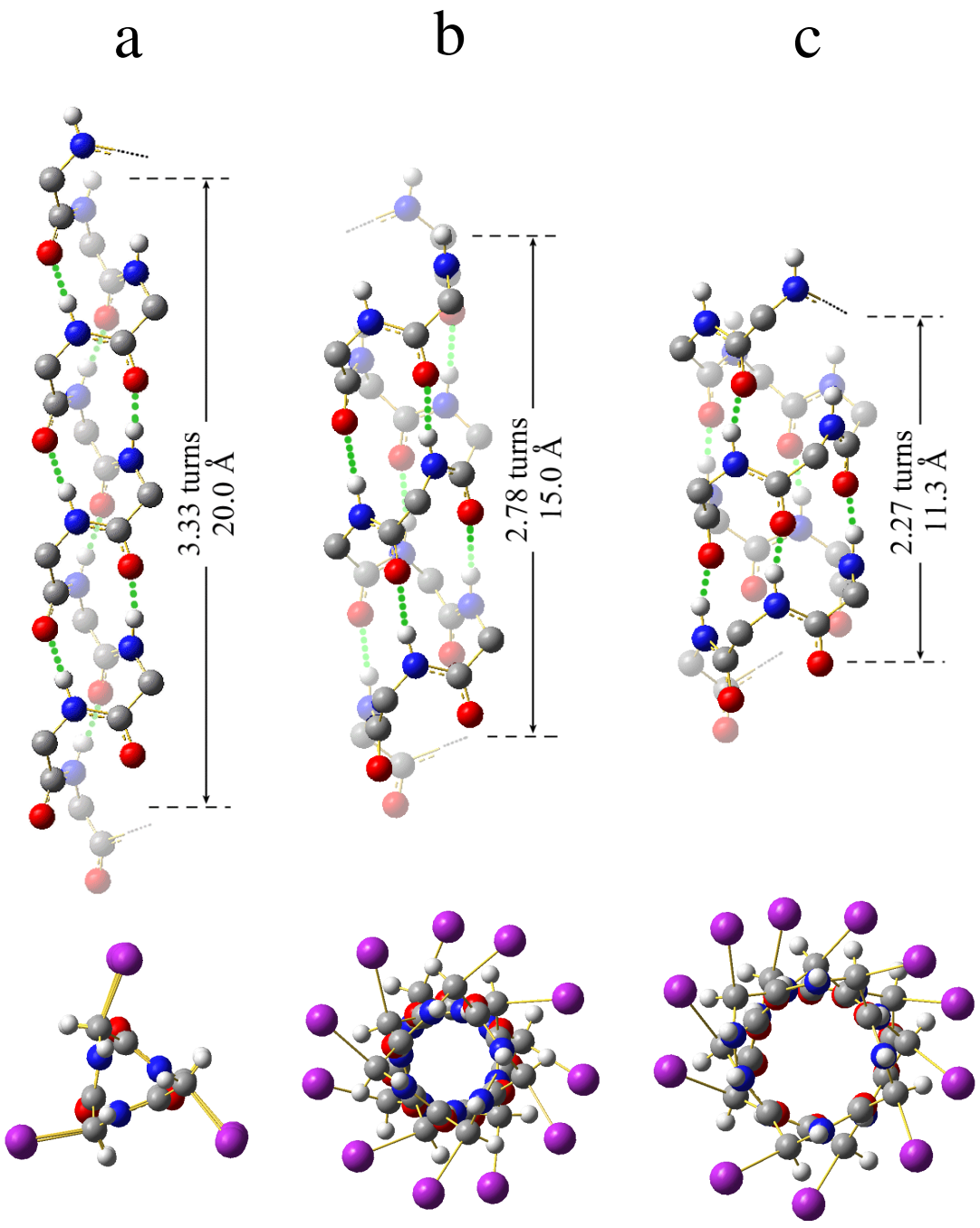


Figure 1.14: The three helices found in protein native structures. **(a)** 3_{10}-helix, **(b)** α-helix, and **(c)** π-helix. In the three cases, the helices shown are 11-residues long. In the standing views (above), the hydrogen bonds are depicted as green dotted lines and the distance and number of turns spanned by 10 residues are indicated at the right of the structures. Whereas in the standing views, the side chains and α-hydrogens have been removed for visual convenience, in the zenithal views (below), they are included.

and, in the last years of the 50s, Perutz and Kendrew saw again the same signal in myoglobin and hemoglobin, when they resolved, for the first time in history, the structure of these proteins [62, 63].

However, despite its being, by far, the most common, the α-helix is not the only coil-like structure that can be found in native proteins [64–66]. If the hydrogen bonds are formed between the carbonyl group (C=O) of each i-th residue with the amino group (N–H) of the residue $i + 3$, one obtains a 3_{10}-*helix*, which is more tightly wound and, therefore, longer than an α-helix of the same chain length (see fig. 1.14a). The 3_{10}-helix is the fourth most common conformation for a single residue after the α-helix, β-sheet and reverse turn[18] [66] but, remarkably, due to its having an integer number of residues per turn, it seemed more natural to scientists with crystallographic background and was theoretically proposed before the α-helix [59, 67].

On the other hand, if the hydrogen bonds are formed between the carbonyl group (C=O) of each i-th residue with the amino group (N–H) of the residue $i + 5$, one obtains a π-*helix* (or 4.4_{16}-*helix*), which is wider and shorter than an α-helix of the same length (see fig. 1.14c). It was originally proposed by Low and Baybutt in 1952 [68], and, although the exact fraction of each type of helix in protein native structures depends up to a considerable extent on their definition (in terms of Ramachandran angles, interatomic distances, energy of the hydrogen bonds, etc.), it seems clear that the π-helix is the less common of the three [65].

Now, it is true that, in addition to these helices that have been experimentally confirmed, some others have been proposed. For example, in the same work in which Pauling, Corey and Branson introduce the α-helix [60], they also describe another candidate: the γ-*helix* (or 5.1_{17}-*helix*). Finally, Donohue performed, in 1953, a systematic study of all possible helices and, in addition to the ones already mentioned, he proposed a 2.2_7- and a 4.3_{14}-*helix* [69]. None of them has been detected in resolved native proteins.

But not all regular local patterns are helices, there exist also a variety of repetitive conformations that do not contain strong intra-chain hydrogen bonds and that are less curled than the structures in fig. 1.14. For example, the *polyproline II* [70–72], which is thought to be important in the unfolded state of proteins, and, principally, the family of the β-*sheets*, which are, together with the α-helices, the most recognizable secondary structure elements in native states of polypeptide chains[19].

The β-sheets are rather plane structures that are typically formed by several individual β-*strands*, which align themselves to form stabilizing inter-chain hydrogen bonds with their neighbours. Two pure arrangements of these single threads may be found: the *antiparallel* β-sheets (see fig. 1.15a), in which the strands run in opposite directions (read from the amino to the carboxyl terminus); and the *parallel* β-sheets (see fig. 1.15b), in which the strands run in the same direction. In both cases, the side chains of neighbouring

Perutz shared it with John Kendrew in 1962 "for their studies of the structures of globular proteins".

[18] A conformation that some residues in proteins adopt when an acute turn in the chain is needed.

[19] It is probably more correct to define the *secondary structure* as the conformational repetition in *consecutive* residues and, from this point of view, to consider the β-strand as the proper element of secondary structure. In this sense, the assembly of β-strands, the β-sheet, together with some other simple motifs such as the coiled coils made up of two helices, the silk fibroin (made up of stacked β-sheets) or collagen (three coiled threads of a repetitive structure similar to polyproline II), may be said to be elements of *super-secondary structure*, somewhat in between the local secondary structure and the global and more complex tertiary structure (see below).

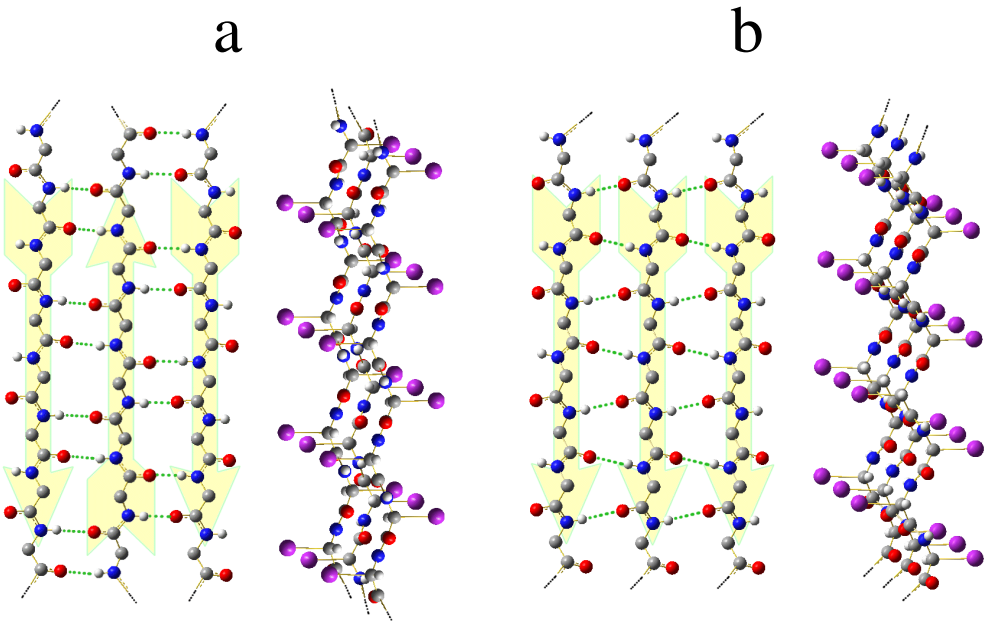


Figure 1.15: β-sheets in the pure **(a)** antiparallel, and **(b)** parallel versions. On the left, the top view is shown, with the side chains and the α-hydrogens omitted for visual convenience and the directions of the strands indicated as yellow arrows. The hydrogen bonds are represented as green dotted lines. On the right, the side view of the sheets is depicted. In this case, the side chains and the α-hydrogens are included.

residues in contiguous strands branch out to the same side of the sheet and may interact. Of course, mixed parallel-antiparallel sheets can also be found.

The next level of protein organization, produced by the assembly of the elements of secondary structure, and also of the chain segments that are devoid of regularity, into a well defined three-dimensional shape, is called *tertiary structure*. The protein folding problem (omitting relevant qualifications that have been partially made and that will be recalled and made more explicit in what follows) may be said to be *the attempt to predict the secondary and the tertiary structure from the primary structure*, and it will be discussed in the next section.

The *cuaternary structure*, which refers to the way in which protein monomers associate to form more complex systems made up of more than one individual chain (such as the ones in fig. 1.1), will not be explored in this dissertation.

1.3 The protein folding problem

The name of the game

As we have seen in the previous section and can visually check in fig. 1.1, the biologically functional *native* structure of a protein[20] is highly complex. What Kendrew saw in one of

[20] Most native states of proteins are flexible and are comprised not of only one conformation but of a set of closely related structures. This flexibility is essential if they need to perform any biological function. However, to economize words, we will use in what follows the terms *native state*, *native conformation* and

the first proteins ever resolved is essentially true for most of them [73]:

> The most striking features of the molecule were its irregularity and its total lack of symmetry.

Now, since these polypeptide chains are synthesized linearly in the ribosome (i.e., they are not manufactured in the folded conformation), in principle, one may imagine that some specific cellular machinery could be the responsible of the complicated process of folding and, in such a case, the prediction of the native structure could be a daunting task. However, in a series of experiments in the 50s, Christian B. Anfinsen ruled out this scenario and was awarded the Nobel prize for it [74].

The most famous and illuminating experiment that he and his group performed is the refolding of bovine pancreatic ribonuclease (see the scheme in fig. 1.16 for reference). They took this protein, which is 124 residues long and has all her eight cysteines forming four disulfide bonds, and added, in a first step, some reducing agent to cleave them. Then, they added urea up to a concentration of 8 M. This substance is known for being a strong denaturing agent (an 'unfolder') and produced a 'scrambled' form of the protein which is much less compact than the native structure and has no enzymatic activity. From this scrambled state, they took two different experimental paths: in the *positive* one, they removed the urea first and then added some oxidizing agent to reform the disulfide bonds; whereas, in the *negative* path, they poured the oxidizing agent first and removed the urea in a second step.

The resultant species in the two paths are very different. If one removes the urea first and then promotes the formation of disulfide bonds, an homogeneous sample is obtained that is practically indistinguishable from the starting native protein and that keeps full biological activity. The ribonuclease has been 'unscrambled'! However, if one takes the negative path and let the cysteines form disulfide bonds before removing the denaturing agent, a mixture of products is obtained containing many or all of the possible 105 isomeric disulfide bonded forms[21]. This mixture is essentially inactive, having approximately 1% the activity of the native enzyme.

One of the most clear conclusions that are commonly drawn from this experiment is that *all the information needed to reach the native state is encoded in the sequence of amino acids*. This important statement, which has stood the test of time [33, 75], allows to isolate the system under study (both theoretically and experimentally) and sharply defines the *protein folding problem*, i.e., the prediction of the three-dimensional native structure of proteins from their amino acid sequence (and the laws of physics).

It is true that we nowadays know of the existence of the so-called *molecular chaperones* (see, for example, the GroEL-GroES complex in fig. 1.1c), which help the proteins to fold in the cellular milieu [76–80]. However, according to the most accepted view [33, 75], these molecular assistants do not add any structural information to the process. Some of them simply prevent accidents related to the *cellular crowding* from happening. Indeed, in the cytoplasm there is not much room: inside a typical bacterium, for example, the total macromolecular concentration is approximately 350 mg/ml, whereas a typical

native structure as interchangeable.

[21] Take an arbitrary cysteine: she can bond to any one of the other seven. From the remaining six, take another one at random: she can bond to five different partners. Take the reasoning to its final and we have $7 \times 5 \times 3 = 105$ different possibilities.

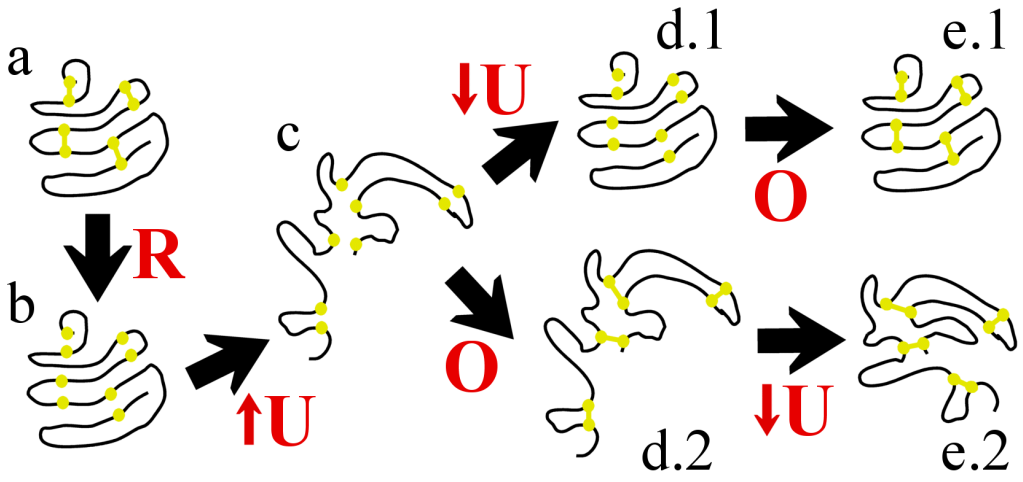


Figure 1.16: Scheme of the refolding of the bovine pancreatic ribonuclease by Anfinsen. The black arrows indicate fundamental steps of the experiment and the red labels next to them designate: **(R)** addition of reducing agent (cleavage of the disulfide bonds), **(O)** addition of oxidizing agent (reformation of the disulfide bonds), **(↑U)** and **(↓U)** increase of the urea concentration up to 8 M and decrease to 0 M respectively. The conformation of the backbone of the protein is schematically depicted by a black line, the cysteines are shown as small yellow circles and the disulfide bonds as line segments connecting them. The different states are labeled: **(a)** starting native enzyme with full activity, **(b)** non-disulfide bonded, folded form, **(c)** representant of the ensemble of inactive 'scrambled' ribonuclease, **(d.1)** non-disulfide bonded, folded form, **(e.1)** refolded ribonuclease indistinguishable from (a), **(d.2)** representant of the ensemble of the scrambled, disulfide bonded form, and, finally, **(e.2)** representant of the mixture of the 105 isomeric disulfide bonded forms.

protein crystal may contain about 600 mg/ml [75]. This crowding may hinder the correct folding of proteins, since partially folded states (of chains that are either free in the cytoplasm or being synthesized in proximate ribosomes) have more 'sticky' hydrophobic surface exposed than the native state, opening the door to aggregation. In order to avoid it, some chaperonins[22] are in charge of providing a shelter in which the proteins can fold alone. Yet another pitfall is that, when the polypeptide chain is being synthesized in the ribosome, it may start to fold incorrectly and get trapped in a non-functional conformation separated by a high energetic barrier from the native state. Again, there exist some chaperones that bind to the nascent chain to prevent this from happening.

As we have already pointed out, all this assistance to fold is seen as lacking new structural information and meant only to avoid traps which are not present *in vitro*. It seems as if molecular chaperones' aim is to make proteins believe that they are not in a messy cell but in Anfinsen's test tube!

The possibility that this state of affairs opens, the prediction of the three-dimensional native structure of proteins from the only knowledge of the amino acid sequence, is often referred to as "the second half of the genetic code" [81, 82]. The reason for such a vehement statement lays in the fact that not all proteins are accessible to the experimental

[22] A particular subset of the set of molecular chaperones.

methods of structure resolution (mostly x-ray crystallography and NMR [52, 83]) and, for those that can be studied, the process is long and expensive, thus making the databases of known structures grow much more slowly than the databases of known sequences (see fig. 1.2 and the related discussion in sec. 1.1). To solve 'the second half of the genetic code' and bridge this gap is the main objective of the hot scientific field of *protein structure prediction* [33, 84, 85].

However, the path that takes to this goal may be walked in two different ways [86, 87]: Either at a fast pragmatic pace, using whatever information we have available, increasingly refining the everything-goes prediction procedures by extensive trial-and-error tests and without any need of knowing the details of the physical processes that take place; or at a slow thoughtful pace, starting from first principles and seeking to arrive to the native structure using the same means that Nature uses: the laws of physics.

The different protocols belonging to the fast pragmatic way are commonly termed *knowledge-based*, since they take profit from the already resolved structures that are deposited in the PDB [41] or any other empirical information that may be statistically extracted from databases of experimental data. There are basically three pure forms of knowledge-based strategies [88]:

- *Homology modeling* (also called *comparative modeling*) [83, 89] is based on the observation that proteins with similar sequences frequently share similar structures [90]. Following this approach, either the whole sequence of the protein that we want to model (the *target*) or some segments of it are *aligned* to a sequence of known structure (the *template*). Then, if some reasonable measure of the *sequence similarity* [91, 92] is high enough, the structure of the template is proposed to be the one adopted by the target in the region analyzed. Using this strategy, one typically needs more than 50% sequence identity between target and template to achieve high accuracy, and the errors increase rapidly below 30% [85]. Therefore, comparative modeling cannot be used with all sequences, since some recent estimates indicate that $\sim 40\%$ of genes in newly sequenced genomes do not have significant sequence homology to proteins of known structure [93].

- *Fold recognition* (or *threading*) [84, 94] is based on the fact that, increasingly, new structures deposited in the PDB turn out to fold in shapes that have been seen before, even though conventional sequence searches fail to detect the relationship [95]. Hence, when faced to a sequence that shares low identity with the ones in the PDB, the threading user tries to fit it in each one of the structures in the databases of known folds, selecting the best choices with the help of some scoring function (which may be physics-based or not). Again, fold recognition methods are not flawless and, according to various benchmarks, they fail to select the correct fold from the databases for $\sim 50\%$ of the cases [84]. Moreover, the fold space is not completely known so, if faced with a novel fold, threading strategies are useless and they may even give false positives. Modern studies estimate that approximately one third of known protein sequences must present folds that have never been seen [96].

- *New fold* (or *de novo prediction*) methods [97, 98] must be used when the protein under study has low sequence identity with known structures and fold recognition strategies fail to fit it in a known fold (because of any of the two reasons discussed

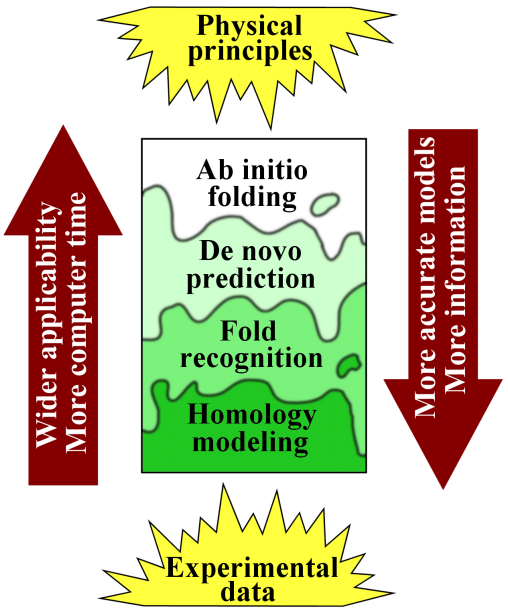


Figure 1.17: Schematic classification of protein structure prediction methods.

in the previous point). The specific strategies used in new fold methods are very heterogeneous, ranging from well-established secondary structure prediction tools or sequence-based identification of sets of possible conformations for short fragments of chains to numerical search methods, such as molecular dynamics, Monte Carlo or genetic algorithms [95].

These knowledge-based strategies may be arbitrarily combined into mixed protocols, and, although the frontiers between them may be sometimes blurry [44], it is clear that the more information available the easier to predict the native structure (see fig. 1.17). So that the three types of methods described above turn out to be written in increasing order of difficulty and they essentially coincide with the competing categories of the *CASP experiment*[23] [44, 95]. In this important meeting, held every two years and whose initials loosely stand for *Critical Assessment of techniques for protein Structure Prediction*, experimental structural biologists are asked to release the amino acid sequences of proteins (the CASP *targets*) whose structures are likely to be resolved before the contest starts. Then, the 'prediction community' gets on stage and their members submit the proposed structures (the *models*), which may be found using any chosen method. Finally, a committee of assessors, critically evaluate the predictions, and the results are published, together with some contributions by the best predictors, in a special issue of the journal *Proteins*.

Precisely, in the latest CASP meetings, the expected ordering (based on the available experimental information) of the three aforementioned categories of protein structure prediction has been observed to translate into different qualities of the proposed models (see fig. 1.17). Hence, while comparative modeling with high sequence similarity has proved to be the most reliable method to predict the native conformation of proteins (with an ac-

[23] Since CASP1, people has drifted towards knowledge-based methods and, nowadays, very few groups use pure ab initio approaches [99].

curacy comparable to low-resolution, experimentally determined structures) [85, 100], de novo modeling has been shown to remain still unreliable [44, 86] (although a special remark should be made about the increasingly good results that David Baker and his group are achieving in this field with their program Rosetta [101, 102]).

Opposed to these knowledge-based approaches, the computer simulation of the real physical process of protein folding[24] without using any empirical information and starting from first principles could be termed *ab initio protein folding* or *ab initio protein structure prediction* depending whether the emphasis is laid on the process or in the goal.

Again, the frontier between de novo modeling and ab initio protein folding is not sharply defined and some confusion might arise between the two terms. For example, the potential energy functions included in most empirical force fields such as CHARMM [104, 105] contain parameters extracted from experimental data, while molecular dynamics attempts to fold proteins using these force fields will be considered by most people (including this Ph.D. candidate) to belong to the ab initio category. As always, the limit cases are clearer, and Baker's Rosetta [101, 102], which uses statistical data taken from the PDB to bias the secondary structure conformational search, may be classified, without any doubt, as a de novo protocol; while, say, a (nowadays unfeasible) simulation of the folding process using quantum mechanics, would be deep in the ab initio region. The situation is further complicated due to the fact that score functions which are based (up to different degrees) on physical principles, are commonly used in conjunction with knowledge-based strategies to prune or refine the candidate models [85, 106]. In the end, the classification of the strategies for finding the native structure of proteins is rather continuous with wriggly, blurry frontiers (see fig. 1.17).

It is clear that, despite their obvious practical advantages and the superior results when compared to pure ab initio approaches [84], any knowledge-based features included in the prediction protocols render the assembly mechanisms physically meaningless [107]. If we want to know the real details of protein folding as it happens, for example, to properly study and attack diseases that are related to protein misfolding and aggregation [12], we must resort to pure ab initio strategies. In addition, ab initio folding does not require any experimental information about the protein, apart from its amino acid sequence. Therefore, as new fold strategies, it has a wider range of applicability than homology modeling and fold recognition, and, in contrast with the largely system-oriented protocols developed in the context of knowledge-based methods, most theoretical and computational improvements made while trying to ab initio fold proteins will be perfectly applicable to other macromolecules.

The feedback between strategies is also an important point to stress. Apart from the obvious fact that the knowledge of the whole folding process includes the capability of predicting the native conformation, and the problem of protein structure prediction would be automatically solved if ab initio folding were achieved, the design of accurate energy functions, which is a central part of ab initio strategies (see the next section), would also be very helpful to improve knowledge-based methods that make use of them (such as Rosetta [106]) or to prune and refine the candidate models on a second stage [85]. Additionally, to assign the correct conformation to those chain segments that are devoid of secondary structure (the problem known as *loop modeling*), may be considered as a 'mini protein folding problem' [85], and the understanding of the physical behaviour of polypeptide

[24] Not a new idea [103].

chains would also include a solution to this issue.

In the light of all these sweet promises, we have chosen the long ab initio path to study protein folding. What will be presented in this Ph.D. dissertation are the first steps of that journey.

However, before we delve deeper in the details, let us define clearly the playfield in which the match shall take place: Although some details of the protein folding process in vivo are under discussion [108] and many cellular processes are involved in helping and checking the arrival to the correct native structure [76]; although some proteins have been shown to fold cotranslationally [109] (i.e., during their synthesis in the ribosome) and many of them are known to be assisted by molecular chaperones (see the discussion above and references therein); although some proteins contain cis proline peptide bonds or disulfide bonded cysteines in their native structure, and must be in the presence of the respective isomerases in order to fold in a reasonable time (see sec. 1.2); although some residues may be post-translationally changed into side chains that are not included in the standard twenty that are depicted in fig. 1.7; and, although some non-peptide molecules may be covalently attached to the protein chain or some cofactor or ion may be needed to reach the native structure, we agree with the words by Alan Fersht [110]:

> We can assume that what we learn about the mechanism of folding of small, fast-folding proteins in vitro will apply to their folding in vivo and, to a large extent, to the folding of individual domains in larger proteins.

and decided to study those processes that do not include any of the aforementioned complications but that may be rightfully considered as intimately related to the process of folding in the cellular milieu and regarded as a first step on top of which to build a more detailed theory.

Henceforth, we define the *restricted protein folding problem*, which is the long-term goal towards which the steps taken in this Ph.D. dissertation are directed, as the full description of the physical behaviour, in aqueous solvent and physiological conditions, and (consequently) the prediction of the native structure, of completely synthesized proteins, made up just of the twenty genetically encoded amino acids in fig. 1.7, without any molecule covalently attached to them, and needless of molecular chaperones, cofactors, ions, disulfide bonds or cis proline peptide bonds in order to fold properly.

1.4 Folding mechanisms and energy functions

The search for the funnel

After having drawn the boundaries of the problem, we should ask the million-dollar question associated to it: *How does a protein fold into its functional native structure?* In fact, since this feat is typically achieved in a very short time, we must add: *How does a protein fold so fast?* This is the question about the *mechanisms* of protein folding, and, ever since Anfinsen's experiments, it has been asked once and again and only partially answered [74, 87, 107, 111, 112].

In order to define the theoretical framework that is relevant for the description of the folding process and also to introduce the language that is typically used in the discussions

about its mechanisms, let us start with a brief reminder of some important statistical mechanics relations. To do this, we will follow the main ideas in ref. [113], although the notation and the assumptions regarding the form of the potential energy, as well as some other minor details, will be different. The presentation will be axiomatic and we will restrict ourselves to the situation in which the macroscopic parameters, such as the temperature T or, say, the number of water molecules N_w, do not change. In these conditions, that allow us to drop any multiplicative terms in the partition functions or the probabilities, and also to forget any additive terms to the energies, we can only focus on the conformational preferences of the system (if, for example, the temperature changed, the neglected terms would be relevant and the expressions that one would need to use would be different). For further details or for the more typical point of view in physics, in which the stress is placed in the variation of the macroscopic thermodynamical parameters, see, for example, ref. 114.

The system which we will talk about is the one defined by the *restricted protein folding problem* in the previous section, i.e., *one protein surrounded by N_w water molecules*[25]; however, one must have in mind that all the subsequent reasoning and the derived expressions are exactly the same for a dilute aqueous solution of a macroscopic number of non-interacting proteins.

Now, if classical mechanics is assumed to be obeyed by our system[26], then each microscopic state is completely specified by the Euclidean[27] coordinates and momenta of the atoms that belong to the protein (denoted by x^μ and π_μ, respectively, with $\mu = 1, \ldots, N$) and those belonging to the water molecules (denoted by X^m and Π_m, with $m = N+1, \ldots, N+N_w$). The whole set of microscopic states shall be called *phase space* and denoted by $\Gamma \times \Gamma_w$, explicitly indicating that it is formed as the direct product of the protein phase space Γ and the water molecules one Γ_w.

The central physical object that determines the time behaviour of the system is the *Hamiltonian* (or *energy*) function:

$$H(x^\mu, X^m, \pi_\mu, \Pi_m) = \sum_\mu \frac{\pi_\mu^2}{2M_\mu} + \sum_m \frac{\Pi_m^2}{2M_m} + V(x^\mu, X^m)\,, \tag{1.1}$$

where M_μ and M_m denote the atomic masses and $V(x^\mu, X^m)$ is the *potential energy*.

After equilibrium has been attained at temperature T, the microscopic details about the time trajectories can be forgot and the average behaviour can be described by the laws of statistical mechanics. In the canonical ensemble, the *partition function* [114] of the system, which is the basic object from which the rest of relevant thermodynamical

[25] At this point of the discussion, the possible presence of non-zero ionic strength is considered to be a secondary issue.

[26] Although non-relativistic quantum mechanics may be considered to be a much more precise theory to study the problem, the computer simulation of the dynamics of a system with so many particles using a quantum mechanical description lies far in the future. Nevertheless, this more fundamental theory can be used to design better classical potential energy functions (which is one of the main long-term goals of the research included in this dissertation; see what follows).

[27] Sometimes, the term *Cartesian* is used instead of *Euclidean*. Here, we prefer to use the latter since it additionally implies the existence of a mass metric tensor that is proportional to the identity matrix, whereas the *Cartesian* label only asks the n-tuples in the set of coordinates to be bijective with the abstract points of the space [115].

quantities may be extracted, is given by

$$Z = \frac{1}{h^{N+N_w} N_w!} \int_{\Gamma\times\Gamma_w} \exp\left[-\beta H(x^\mu, X^m, \pi_\mu, \Pi_m)\right] \mathrm{d}x^\mu \mathrm{d}X^m \mathrm{d}\pi_\mu \mathrm{d}\Pi_m \,, \tag{1.2}$$

where h is Planck's constant, we adhere to the standard notation $\beta := 1/RT$ (per-mole energy units are used throughout this document, so R is preferred over k_B) and $N_w!$ is a combinatorial number that accounts for the quantum indistinguishability of the N_w water molecules. Also, as we have anticipated, the multiplicative factor outside the integral sign is a constant that divides out for any observable averages and represents just a change of reference in the Helmholtz free energy. Therefore, we will drop it from the previous expression and the notation Z will be kept for convenience.

Next, since the principal interest lies on the conformational behaviour of the polypeptide chain, seeking to develop clearer images and, if possible, reduce the computational demands, water coordinates and momenta are customarily *averaged* (or *integrated*) *out* [113, 116], leaving an *effective Hamiltonian* $H_{\text{eff}}(x^\mu, \pi_\mu; T)$ that depends only on the protein degrees of freedom and the temperature T, and whose potential energy (denoted by $W(x^\mu; T)$) is called *potential of mean force* or *effective potential energy*.

This effective Hamiltonian may be either empirically designed from scratch (which is the common practice in the classical force fields typically used to perform molecular dynamics simulations [104, 105, 117–126]) or obtained from the more fundamental, original Hamiltonian $H(x^\mu, X^m, \pi_\mu, \Pi_m)$ actually performing the averaging out process. In statistical mechanics, the theoretical steps that must be followed if one chooses this second option are very straightforward (at least formally):

The integration over the water momenta Π_m in eq. (1.2) yields a T-dependent factor that includes the masses M_m and that shall be dropped by the same considerations stated above. On the other hand, the integration of the water coordinates X^m is not so trivial, and, except in the case of very simple potentials, it can only be performed formally. To do this, we define the *potential of mean force* or *effective potential energy* by

$$W(x^\mu; T) := -RT \ln\left(\int \exp\left[-\beta V(x^\mu, X^m)\right] \mathrm{d}X^m\right) \,, \tag{1.3}$$

and simply rewrite Z as

$$Z = \int_\Gamma \exp\left[-\beta H_{\text{eff}}(x^\mu, \pi_\mu; T)\right] \mathrm{d}x^\mu \mathrm{d}\pi_\mu \,, \tag{1.4}$$

with the *effective Hamiltonian* being

$$H_{\text{eff}}(x^\mu, \pi_\mu; T) = \sum_\mu \frac{\pi_\mu^2}{2M_\mu} + W(x^\mu; T) \,. \tag{1.5}$$

At this point, the protein momenta π_μ may also be averaged out from the expressions. This choice, which is very commonly taken in the literature, largely simplifies the discussion about the mechanisms of protein folding and the images and metaphors typically used in the field. However, to perform this average is not completely harmless, since it brings up a number of technical and interpretation-related difficulties mostly due to the

fact that the marginal probability density in the x^μ-space in eq. (1.7) is not invariant under a change of coordinates[28] (see appendix A and chapter 6 for further details).

Bearing this in mind, the integration over π_μ produces a new T-dependent factor, which is dropped as usual, and yields a new form of the partition function, which is the one that will be used from now on in this section:

$$Z = \int_\Omega \exp\left[-\beta W(x^\mu; T)\right] \mathrm{d}x^\mu \,, \tag{1.6}$$

where Ω now denotes the coordinates part of the protein phase space Γ.

Some remarks may be done at this point. On the one hand, if one further assumes that the original potential energy $V(x^\mu, X^m)$ separates as a sum of intra-protein, intra-water and water-protein interaction terms, the effective potential energy $W(x^\mu; T)$ in the equations above may be written as a sum of two parts: a vacuum intra-protein energy and an effective solvation energy [113]. Nevertheless, this simplification is neither justified a priori, nor necessary for the subsequent reasoning about the mechanisms of protein folding; so it will not be assumed herein.

On the other hand, the (in general, non-trivial) dependence of $W(x^\mu; T)$ on the temperature T (see eq. (1.3)) and the associated fact that it contains the entropy of the water molecules, justifies its alternative denomination of *internal* or *effective free energy*, and also the suggestive notation $F(x^\mu) := W(x^\mu; T)$ used in some works [128]. In this dissertation, however, we prefer to save the name *free energy* for the one that contains some amount of protein conformational entropy and that may be assigned to finite subsets (states) of the conformational space of the chain (see eq. (1.10) and the discussion below).

Finally, we will stick to the notational practice of dropping (but remembering) the temperature T from W and H_{eff}. This is consistent with the situation of constant T that we wish to investigate and also very natural and common in the literature. In fact, most Hamiltonian functions (and their respective potentials) that are considered to be 'fundamental' actually come from the averaging out of degrees of freedom more microscopical than the ones regarded as relevant, and, as a result, the coupling 'constants' contained in them are not really constant, but dependent on the temperature T.

Now, from the *probability density function* (PDF) in the protein conformational space Ω, given by,

$$p(x^\mu) = \frac{\exp\left[-\beta W(x^\mu)\right]}{Z} \,, \tag{1.7}$$

we can tell that $W(x^\mu)$ completely determines the conformational preferences of the polypeptide chain in the thermodynamic equilibrium as a function of each point of Ω. On the opposite extreme of the details scale, we may choose to describe the macroscopic state of the system as a whole (like it is normally done in physics [114]) and define, for example, the Helmholtz free energy as $F := -RT \ln Z$, where no trace of the microscopic details of the system remains.

In protein science, it is also common practice to take a point of view somewhat in the middle of these two limit descriptions, and define *states* that are neither single points of Ω nor the whole set, but finite subsets $\Omega_i \subset \Omega$ comprising many different conformations

[28] Note that, if the momenta π_μ are kept in the integration measure, any canonical transformation leaves the probability density invariant, since its Jacobian determinant is unity [127].

that are related in some sense. These states must be precisely specified in order to be of any use, and they must fulfill some reasonable conditions, the most important of which is that they must be mutually exclusive, so that $\Omega_i \cap \Omega_j = \emptyset, \forall i \neq j$ (i.e., no point can lie in two different states at the same time).

Since the two most relevant conceptual constructions used to think about protein folding, the native ($\mathcal{N}$) and the unfolded ($\mathcal{U}$) states, as well as a great part of the language used to talk about protein stability, fit in this formalism, we will now introduce the basic equations associated to it.

To begin with, one can define the partition function of a certain state Ω_i as

$$Z_i := \int_{\Omega_i} \exp\left[-\beta W(x^\mu; T)\right] \mathrm{d}x^\mu \,, \tag{1.8}$$

so that the probability of Ω_i be given by

$$P_i := \frac{Z_i}{Z} \,. \tag{1.9}$$

The *Helmholtz free energy* F_i of this state is

$$F_i := -RT \ln Z_i \,, \tag{1.10}$$

and the following relation for the free energy differences is satisfied:

$$\Delta F_{ij} = F_j - F_i = -RT \ln \frac{Z_j}{Z_i} = -RT \ln \frac{P_j}{P_i} = -RT \ln \frac{[j]}{[i]} = -RT \ln K_{ij} \,, \tag{1.11}$$

where $[i]$ denotes the *concentration* (in chemical jargon) of the species i, and K_{ij} is the *reaction constant* (using again images borrowed from chemistry) of the $i \leftrightarrow j$ equilibrium. It is precisely this dependence on the concentrations, together with the approximate equivalence between ΔF and ΔG at physiological conditions (where the term $P\Delta V$ is negligible [113]), that renders eq. (1.11) very useful and ultimately justifies this point of view based on states, since it relates the quantity that describes protein stability and may be estimated theoretically (the folding free energy at constant temperature and constant pressure $\Delta G_{\text{fold}} := G_{\mathcal{N}} - G_{\mathcal{U}}$) with the observables that are commonly measured in the laboratory (the concentrations $[\mathcal{N}]$ and $[\mathcal{U}]$ of the native and unfolded states) [33, 52, 129].

The next step to develop this state-centered formalism is to define the *microscopic PDF in* Ω_i as the original one in eq. (1.7) conditioned to the knowledge that the conformation x^μ lies in Ω_i:

$$p_i(x^\mu) := p(x^\mu \,|\, x^\mu \in \Omega_i) = \frac{p(x^\mu)}{P_i} = \frac{\exp\left[-\beta W(x^\mu)\right]}{Z_i} \,. \tag{1.12}$$

Now, using this probability measure in Ω_i, we may calculate the *internal energy* U_i as the average potential energy in this state:

$$U_i := \langle W \rangle_i = \int_{\Omega_i} W(x^\mu) p_i(x^\mu) \, \mathrm{d}x^\mu \,, \tag{1.13}$$

and also define the *entropy* of Ω_i as

$$S_i := -R \int_{\Omega_i} p_i(x^\mu) \ln p_i(x^\mu) \, \mathrm{d}x^\mu \, . \tag{1.14}$$

Finally, ending our statistical mechanics reminder, one can show that the natural thermodynamic relation among the different state functions is recovered:

$$\Delta F_{ij} = \Delta U_{ij} - T\Delta S_{ij} \simeq \Delta G_{ij} = \Delta H_{ij} - T\Delta S_{ij} \, , \tag{1.15}$$

where H is the *enthalpy*, whose differences ΔH_{ij} may be approximated by ΔU_{ij} neglecting the term $P\Delta V$ again.

Retaking the discussion about the mechanisms of protein folding, we see (again) in eq. (1.7) that the potential of mean force $W(x^\mu)$ completely determines the conformational preferences of the polypeptide chain in the thermodynamic equilibrium. Nevertheless, in order to think about the problem, it is often useful to investigate also the underlying microscopic dynamics. The effective potential energy $W(x^\mu)$ in eq. (1.3) has been simply obtained in the previous paragraphs using the tools of statistical mechanics; the 'dynamical averaging out' of the solvent degrees of freedom in order to describe the *time evolution* of the protein subsystem, on the other hand, is a much more complicated (and certainly different) task [130–134]. However, if the relaxation of the solvent is fast compared to the motion of the polypeptide chain, the function $W(x^\mu)$ turns out to be precisely the effective 'dynamical' potential energy that determines the microscopic time evolution of the protein degrees of freedom [130]. Although this condition could be very difficult to check for real cases and it has only been studied in simplified model systems [131, 133, 134], molecular dynamics simulations with classical force fields and explicit water molecules suggest that it may be approximately fulfilled [130, 135, 136]. For the sake of brevity, in the discussion that follows, we will assume that this fast-relaxation actually occurs, so that, when reasoning about the graphical representations (commonly termed *energy landscapes*) of the effective potential energy $W(x^\mu)$, we are entitled to switch back and forth from dynamical to statistical concepts.

Now, just after noting that $F(x^\mu)$ is the central physical object needed to tackle the elucidation of the folding mechanisms, we realize that the number of degrees of freedom N in an average-length polypeptide chain is large enough for the size of the conformational space (which is exponential on N) to be astronomically astronomical. This fact was, for years, regarded as a problem, and is normally called *Levinthal's paradox* [137]. Although it belongs to the set of paradoxes that (like Zeno's or Epimenides') are called so without actually being problematic[29], thinking about it and using the language and the images related to it have dominated the views on folding mechanisms for a long time [87]. The paradox itself was first stated in a talk entitled "How to fold graciously" given by Cyrus Levinthal in 1969 [138] and it essentially says that, if, in the course of folding, a protein is required to sample all possible conformations (a hypothesis that ignores completely the laws of dynamics and statistical mechanics) and the conformation of a given residue is independent of the conformations of the rest (which is also false), then the protein will never fold to its native structure.

[29] In fact, Levinthal did not use the word "paradox" and, just after stating the problem, he proposed a possible solution to it.

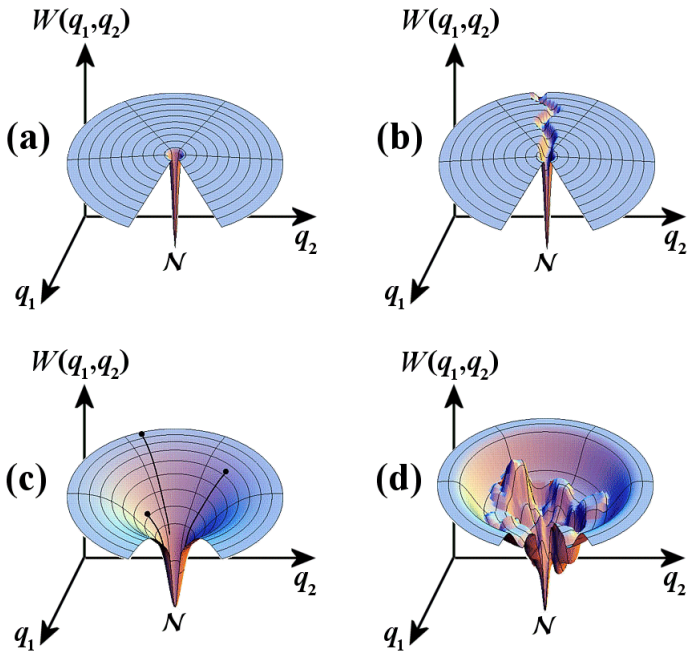


Figure 1.18: Possible energy landscapes of a protein. The conformational space is assumed to be two-dimensional, the degrees of freedom being q_1 and q_2. The degrees of freedom of the solvent have been *integrated out* (see the text), and the effective potential energy $W(q_1, q_2)$ is a function of these two variables, which are internal degrees of freedom of the molecule. $\mathcal{N}$ stands for native state and it is assumed here to be the global minimum. **(a)** *Flat golf course*: the energy landscape as it would be if Levinthal's paradox were a real problem. **(b)** *Ant trail*: the old-view pathway solution to Levinthal's paradox. **(c)** *New-view smooth funnel*. **(d)** *More realistic partially rugged funnel*. (Figures taken from ref. 128 with kind permission and somewhat modified.)

For example, let us assume that each one of the 124 residues in Anfinsen's ribonuclease (see sec. 1.3) can take up any of the six different discrete backbone conformations in table 1.1 (side chain degrees of freedom are not relevant in this qualitative discussion, since they only affect the structure locally). This makes a total of $6^{124} \simeq 10^{96}$ different conformations for the chain. If they were visited in the shortest possible time (say, $\sim 10^{-12}$ s, approximately the time required for a single molecular vibration [139]), the protein would need about 10^{76} years to sample the whole conformational space. Of course, this argument is just a *reductio ad absurdum* proof (since proteins do fold!) of the a priori evident statement that protein folding cannot be a completely random trial-and-error process (i.e., a random walk in conformational space). The *golf-course* energy landscape in fig. 1.18a represents this non-realistic, paradoxical situation: the point describing the conformation of the chain wanders aimlessly on the enormous denatured plateau until it suddenly finds the native well by pure chance.

Levinthal himself argued that a solution to his paradox could be that the folding process occurs along well-defined *pathways* that take every protein, like an ordered column of ants, from the *unfolded state* to the native structure, visiting partially folded intermediates en route [112, 140]. The *ant-trail* energy landscape in fig. 1.18b is a graphical depiction

of the pathway image.

This view, which is typically referred to as the *old view* of folding [130, 137, 141], is largely influenced by the situation in simple chemical reactions, where the barriers surrounding the minimum energy paths that connect the different local minima are very steep compared to RT, and the dynamical trajectories are, consequently, well defined. In protein folding, however, due to the fact that the principal driving forces are much weaker than those relevant for chemical reactions and comparable to RT, short-lived transient interactions may form randomly among different residues in the chain and the system describes stochastic trajectories that are never the same. Henceforth, since the native state may be reached in many ways, it is unlikely that a single minimum energy path dominates over the rest of them [130].

In the late 80s, a *new view* of folding mechanisms began to emerge based on these facts and inspired on the statistical mechanics of spin glasses [130, 137, 139, 142–144]. According to it, when a large number of identical proteins (from 10^{15} to 10^{18} [145]) are introduced in a test tube in the conditions of the *restricted protein folding problem* defined in sec. 1.3, a conformational equilibrium is attained between the native ensemble of states $\mathcal{N}$ and the ensemble made up of the rest of (non-functional) conformations (the unfolded state $\mathcal{U}$). At the same time, what is happening at the microscopic level is that each single molecule is following a partially stochastic trajectory determined by the intrinsic energetics of the system (given by $W(x^\mu)$) and subject to random fluctuations due to the thermal noise. Of course, all trajectories are different, some towards the native state and some towards the unfolded state, but, if we focus on a single molecule at an arbitrary time, the probability that she is wandering in the native basin is very high (typically more than 99%) and, in the rare case that we happen to choose a protein that is presently unfolded, we will most certainly watch a very fast race towards the native state.

In order for this to happen, we need that the energy landscape be *funneled* towards the native state, like in figs. 1.18c and 1.18d, so that any microscopic trajectory has more probability to evolve in the native direction than in the opposite one at every point of the conformational space (the 'ruggedness' of the funnel must also be small in order to avoid getting trapped in deep local minima during the course of folding). In this way, the solution to Levinthal's paradox could be said to be "funnels, not tunnels" [146], and the deterministic pathway image is changed by a statistical treatment in which folding is a heterogeneous reaction involving broad ensembles of structures [147], the kinetic intermediates that are sometimes observed experimentally being simply more or less deep wells in the walls of the funnel. Anyway, although this new view has been validated both experimentally [148] and theoretically [145], and it is widely accepted as correct by the scientific community, one must note that it is not contradictory with the old view, since the latter is only a particular case of the former in which the funnel presents a deep canyon through which most of the individual proteins roll downwards. In fact, in some studied cases, one may find a single pathway that dominates statistically [135, 145].

A marginal issue that arises both in the old and new views, is whether the native state is the global minimum of the effective potential energy $W(x^\mu)$ of the protein (in which case the folding process is said to be *thermodynamically controlled*) or it is just the lowest-lying kinetically-accessible local minimum (in which case we talk about *kinetic control*) [113]. This question was raised by Anfinsen [74], who assumed the first case to be the correct answer and called the assumption the *thermodynamic hypothesis*. Although

Levinthal pointed out a few years later that this was not necessary and that kinetic control was perfectly possible [138], and also despite some indications against it [149, 150], it is now widely accepted that the thermodynamic hypothesis is fulfilled most of the times, and almost always for small single-domain proteins [33, 75, 76, 113]. Of course, nothing fundamental changes in the overall picture if the energy landscape is funneled towards a local minimum of $W(x^\mu)$ instead of being funneled to a global one, however, from the computational point of view there is a difference: In the latter case, the prediction of the native state may be tackled both dynamically and by simple minimization[30] of the function $W(x^\mu)$ (for example, using *simulated annealing* [151, 152] or similar schemes), whereas, if the thermodynamic hypothesis is broken, the native structure may still be found performing molecular dynamics simulations, but minimization procedures could be misleading and technically problematic. This is so because, although local minima may also be found and described, the knowledge about towards which one of them the protein trajectories converge depends on kinetic information, which is absent from the typical minimization algorithms.

Now, even though a funneled energy function provides the only consistent image that accounts for all the experimental facts about protein folding, one must still explain the fact that the landscape is just like that. If one looks at a protein as if it were the first time, one sees that it is a heteropolymer made up of twenty different types of amino acid monomers (see sec. 1.2). Such a system, due to its many degrees of freedom, the constraints imposed by chain connectivity and the different affinities that the monomers show for their neighbours and for the environment, presents a large degree of *frustration*, that is, there is not a single conformation of the chain which optimizes all the interactions at the same time[31]. For the vast majority of the sequences, this would lead to a rugged energy landscape with many low-energy states, high barriers, strong traps, etc.; up to a certain degree, a landscape similar to that of spin glasses. A landscape in which fast-folding to a unique three-dimensional structure is impossible!

However, a protein is not a random heteropolymer. Its sequence has been selected and improved along thousands of millions of years by natural selection[32], and the score function that decided the contest, the fitness that drove the process, is just its ability to fold into a well-defined native structure in a biologically reasonable time[33]. Henceforth, the energy landscape of a protein is not like the majority of them, proteins are a selected

[30] See appendix A for some technical but relevant remarks about the minimization of the effective potential energy function.

[31] In order to be entitled to give such a simple definition, we need that the effective potential energy of the system separates as a sum of terms with the minima at different points (either because it is split in few-body terms, or because it is split in different 'types of interactions', such as van der Waals, Coulomb, hydrogen-bonds, etc.). This is a classical image which is rigorously wrong but approximately true (and very useful to think). If one does not want to assume the existence of 'interactions' or few-body terms that may conflict with one another, one may jump directly to the conclusion, noting that the energy landscape of a random heteropolymer is glassy but without introducing the concept of frustration.

[32] The problem of finding the protein needle in the astronomical haystack of all possible sequences and its solution are presented as another paradox, *the blind watchmaker paradox*, and inspiringly discussed by Richard Dawkins in ref. 153.

[33] One may argue that the ability to perform a catalytic function also enters the fitness criterium. While this is true, it is probably a less important factor than the folding skill, since the active site of enzymes is generally localized in a small region of the surface of the protein and it could be, in principle, assembled on top of many different folds.

minority of heteropolymers for which there exists a privileged structure (the native one) so that, in every point of the conformational space, it is more stabilizing, on average, to form 'native contacts' than to form 'non-native' ones (an image radically implemented by Gō-type models [154]). Bryngelson and Wolynes [142] have termed this fewer conflicting interactions than typically expected the *principle of minimal frustration*, and this takes us to a natural definition of a *protein* (opposed to a general *polypeptide*): a *protein* is a polypeptide chain whose sequence has been naturally selected to satisfy the principle of minimal frustration.

Now, we should note that this funneled shape emerges from a very delicate balance. Proteins are only marginally stable in solution, with an unfolding free energy ΔG_{unfold} typically in the 5 – 15 kcal/mol range. However, if we split this relatively small value into its enthalphic and entropic contributions, using eq. (1.15) and the already mentioned fact that the term $P\Delta V$ is negligible at physiological conditions [113],

$$\Delta G_{\text{unfold}} = \Delta H_{\text{unfold}} - T\Delta S_{\text{unfold}} \,, \tag{1.16}$$

we find that it is made up of the difference between two quantities (ΔH_{unfold} and $T\Delta S_{\text{unfold}}$) that are typically an order of magnitude larger than ΔG_{unfold} itself [113, 155], i.e., the native state is enthalpically favoured by hundreds of kilocalories per mole and entropically penalized by approximately the same amount.

In addition, both quantities are strongly dependent on the details of the effective potential energy $W(x^\mu)$ (see eqs. (1.13) and (1.14)), which could be imagined to be made up of the sum of thousands of non-covalent terms each one of a size comparable to ΔG_{unfold}. This very fine tuning that has been achieved after thousands of millions of years of natural selection is easily destroyed by a single-residue mutation or by slightly altering the temperature, the pH or the concentration of certain substances in the environment (parameters on which $W(x^\mu)$ implicitly depends).

For the same reasons, if the folding process is intended to be simulated theoretically, the chances of missing the native state and (what is even worse) of producing a non-funneled landscape, which is very difficult to explore using conventional molecular dynamics or minimization algorithms, are very high if poor energy functions are used [143, 156, 157]. Therefore, it is not surprising that current force fields [104, 105, 117–126], which include a number of strong assumptions (additivity of the 'interactions', mostly pairwise terms, simple functional forms, etc.), are widely recognized to be incapable of folding proteins [33, 84, 98, 100, 158–161].

The improvement of the effective potential energy functions describing polypeptides, with the long-term goal of reliable ab initio folding, is one of the main objectives towards which the steps taken during the research period that is described in this Ph.D. dissertation are directed.

Chapter 2

Introduction to quantum chemistry

> It would indeed be remarkable if Nature fortified herself against further advances in knowledge behind the analytical difficulties of the many-body problem.
>
> — Max Born, 1960

2.1 Introduction

In the previous chapter, we pointed out that the principal computational hindrances to successfully fold proteins are two: the necessity to explore a huge conformational space (either dynamically or stochastically) and the lack of accurate potential energy functions that describe their behaviour [84, 98, 100, 107, 161, 162].

The former could be alleviated by choosing a convenient set of coordinates (see chapter 4) and imposing constraints on the least important degrees of freedom of the molecule (see chapter 6). To tackle the design of accurate effective potentials, on the other hand, two essential ingredients are needed: a *reference potential* that could be regarded as accurate enough and a *physically meaningful criterium* to compare less numerically expensive approximations to this reference.

In chapter 3, we introduce such a criterium and thoroughly discuss its convenience with respect to other frequently used ones. Regarding the reference energy function, it is widely accepted [159, 160, 163–167] that the effective potential for the nuclei calculated with non-relativistic quantum mechanics in the Born-Oppenheimer approximation should be accurate enough to describe a wide range of phenomena among which the conformational behaviour of biological molecules at physiological conditions is included. The study of physico-chemical processes at this level of theoretical detail and the design of computationally efficient approximations for solving the demanding equations that appear constitute the major part of the field called *quantum chemistry* [168, 169].

In this chapter, we introduce the basic quantum chemical formalism that will be exploited in the rest of the dissertation. In sec. 2.2, we introduce the molecular Hamiltonian and a special set of units (the atomic ones) that are convenient to simplify the basic equations. In sec. 2.3, we present in an axiomatic way the concepts and expressions related to the separation of the electronic and nuclear problems in the Born-Oppenheimer scheme.

In sec. 2.4, we introduce the variational method that underlies the derivation of the basic equations of the Hartree and Hartree-Fock approximations, discussed in sec. 2.6 and 2.7 respectively. The computational implementation of the Hartree-Fock approximation is tackled in sec. 2.8, where the celebrated Roothaan-Hall equations are derived and a brief introduction to Gaussian basis sets is made in sec. 2.9. Finally, in sec. 2.10, the Møller-Plesset 2 (MP2) theory to incorporate correlation to the Hartree-Fock results is discussed.

2.2 Molecular Hamiltonian and atomic units

Getting rid of constants and powers of ten

Since 1960, the international scientific community has agreed on an 'official' set of basic units for measurements: *Le Système International d'Unités*, or SI for short (see http://www.bipm.org/en/si/ and ref. 170). The meter (m), the kilogram (kg), the second (s), the ampere (A), the kelvin (K), the mole (mol), the joule (J) and the pascal (Pa) are examples of SI units.

Sticking to the SI scheme, the non-relativistic quantum mechanical Hamiltonian operator of a molecule consisting of N_N nuclei (with atomic numbers Z_α and masses M_α, $\alpha = 1, \ldots, N_N$) and N electrons (i.e., the *molecular Hamiltonian*) is expressed as[34]:

$$\begin{aligned}\hat{H} \;=\; & -\sum_{\alpha=1}^{N_N}\frac{\hbar^2}{2M_\alpha}\nabla_\alpha^2 - \sum_{i=1}^{N}\frac{\hbar^2}{2m_e}\nabla_i^2 + \frac{1}{2}\sum_{\alpha\neq\beta}\left(\frac{e^2}{4\pi\epsilon_0}\right)\frac{Z_\alpha Z_\beta}{|\vec{R}_\beta - \vec{R}_\alpha|} \\ & - \sum_{i=1}^{N}\sum_{\alpha=1}^{N_N}\left(\frac{e^2}{4\pi\epsilon_0}\right)\frac{Z_\alpha}{|\vec{R}_\alpha - \vec{r}_i|} + \frac{1}{2}\sum_{i\neq j}\left(\frac{e^2}{4\pi\epsilon_0}\right)\frac{1}{|\vec{r}_j - \vec{r}_i|}\,,\end{aligned} \tag{2.1}$$

where $\hbar$ stands for $h/2\pi$, being h Planck's constant, m_e denotes the electron mass, e the proton charge, $\vec{r}_i$ the position of the i-th electron, $\vec{R}_\alpha$ that of the α-th nucleus, ϵ_0 the vacuum permittivity and ∇_i^2 the Laplacian operator with respect to the coordinates of the i-th particle.

Although using a common set of units presents obvious communicative advantages, when circumscribed to a particular field of science, it is common to appeal to non-SI units in order to simplify the most frequently used equations by getting rid of some constant factors that always appear grouped in the same ways and, thus, make the numerical values in any calculation of the order of unity. In the field of quantum chemistry, *atomic units* (see table 2.1), proposed in ref. 171 and named in ref. 172, are typically used. In these units, eq. (2.1) is substantially simplified to

$$\begin{aligned}\hat{H} \;=\; & -\sum_{\alpha=1}^{N_N}\frac{1}{2M_\alpha}\nabla_\alpha^2 - \sum_{i=1}^{N}\frac{1}{2}\nabla_i^2 + \frac{1}{2}\sum_{\alpha\neq\beta}\frac{Z_\alpha Z_\beta}{|\vec{R}_\beta - \vec{R}_\alpha|} \\ & - \sum_{i=1}^{N}\sum_{\alpha=1}^{N_N}\frac{Z_\alpha}{|\vec{R}_\alpha - \vec{r}_i|} + \frac{1}{2}\sum_{i\neq j}\frac{1}{|\vec{r}_j - \vec{r}_i|}\,.\end{aligned} \tag{2.2}$$

[34] Note that the non-relativistic molecular Hamiltonian does not depend on spin-like variables.

Unit of mass:	mass of the electron = $m_e = 9.1094 \cdot 10^{-31}$ kg
Unit of charge:	charge on the proton = $e = 1.6022 \cdot 10^{-19}$ C
Unit of length:	1 bohr = $a_0 = \frac{4\pi\epsilon_0\hbar^2}{m_e e^2} = 0.52918$ Å$= 5.2918 \cdot 10^{-11}$ m
Unit of energy:	1 hartree = $\frac{\hbar^2}{m_e a_0^2} = 627.51$ kcal/mol $= 4.3597 \cdot 10^{-18}$ J

Table 2.1: Atomic units up to five significant digits. Taken from the National Institute of Standards and Technology (NIST) web page at http://physics.nist.gov/cuu/Constants/.

	1 hartree	1 eV	1 kcal/mol	1 kJ/mol	1 cm^{-1}
1 hartree	1	27.211	627.51	262.54	219470
1 eV	$3.6750 \cdot 10^{-2}$	1	23.061	96.483	8065.5
1 kcal/mol	$1.5936 \cdot 10^{-3}$	$4.3363 \cdot 10^{-2}$	1	4.1838	349.75
1 kJ/mol	$3.8089 \cdot 10^{-4}$	$1.0364 \cdot 10^{-2}$	$2.3902 \cdot 10^{-1}$	1	83.595
1 cm^{-1}	$4.5560 \cdot 10^{-6}$	$1.2398 \cdot 10^{-4}$	$2.8592 \cdot 10^{-3}$	$1.1962 \cdot 10^{-2}$	1

Table 2.2: Energy units conversion factors to five significant digits. Taken from the National Institute of Standards and Technology (NIST) web page at http://physics.nist.gov/cuu/Constants/. The table must be read by rows. For example, the value 4.1838, in the third row, fourth column, indicates that 1 kcal/mol = 4.1838 kJ/mol.

Since all the relevant expressions in quantum chemistry are derived in one way or another from the molecular Hamiltonian, the simplification brought up by the use of atomic units propagates to the whole formalism. Consequently, they shall be the choice all throughout this document.

Apart from the atomic units and the SI ones, there are some other miscellaneous units that are often used in the literature: the *ångström*, which is a unit of length defined as 1 Å = 10^{-10} m, and the units of energy cm^{-1} (which reminds about the spectroscopic origins of quantum chemistry and, even, quantum mechanics), *electronvolt* (eV), *kilocalorie per mole* (kcal/mol) and *kilojoule per mole* (kJ/mol). The last two are specially used in the field of macromolecular simulations and quantify the energy of a mole of entities; for example, if one asserts that the torsional barrier height for H_2O_2 is $\sim$ 7 kcal/mol, one is really saying that, in order to make a mole of H_2O_2 (i.e., $N_A \simeq 6.0221 \cdot 10^{23}$ molecules) rotate 180º around the O–O bond, one must spend $\sim$ 7 kcal. For the conversion factors between the different energy units, see table 2.2.

Finally, to close this section, we rewrite eq. (2.2) introducing some self-explanatory notation that will be used in the subsequent discussion:

$$\hat{H} = \hat{T}_N + \hat{T}_e + \hat{V}_{NN} + \hat{V}_{eN} + \hat{V}_{ee} \,, \tag{2.3a}$$

$$\hat{T}_N := -\sum_{\alpha=1}^{N_N} \frac{1}{2M_\alpha} \nabla_\alpha^2 \,, \tag{2.3b}$$

$$\hat{T}_e := -\sum_{i=1}^{N} \frac{1}{2} \nabla_i^2 \,, \tag{2.3c}$$

$$\hat{V}_{NN} := \frac{1}{2} \sum_{\alpha \neq \beta} \frac{Z_\alpha Z_\beta}{R_{\alpha\beta}} \,, \tag{2.3d}$$

$$\hat{V}_{eN} := -\sum_{i=1}^{N} \sum_{\alpha=1}^{N_N} \frac{Z_\alpha}{R_{\alpha i}} \,, \tag{2.3e}$$

$$\hat{V}_{ee} := \frac{1}{2} \sum_{i \neq j} \frac{1}{r_{ij}} \,. \tag{2.3f}$$

2.3 The Born-Oppenheimer approximation

Separating electrons from nuclei

To think of a macromolecule as a set of quantum objects described by a *wavefunction* $\Psi(X_1, \ldots, X_{N_N}, x_1, \ldots, x_N)$ dependent on the spatial and spin[35] degrees of freedom, $x_i := (\vec{r}_i, \sigma_i)$, of the electrons and on those of the nuclei, $X_\alpha := (\vec{R}_\alpha, \Sigma_\alpha)$, would be too much for the imagination of physicists and chemists. All the language of chemistry would have to be remade and simple sentences in textbooks, such as "rotation about this single bond allows the molecule to avoid steric clashes between atoms" or even "a polymer is a long chain-like molecule composed of repeating monomer units", would have to be translated into long and counter-intuitive statements involving probability and 'quantum jargon'. Conscious or not, we think of molecules as classical objects.

More precisely, we are ready to accept that electrons are quantum (we know of the interference experiments, electrons are light, we are accustomed to draw atomic 'orbitals', etc.), however, we are reluctant to concede the same status to nuclei. Nuclei are heavier than electrons (at least $\sim$ 2000 times heavier, in the case of the single proton nucleus of hydrogen) and we picture them in our imagination as 'classical things' that move, bond to each other, rotate around bonds and are at precise points at precise times. We imagine nuclei 'slowly moving' in the field of the electrons, which, for each position of the first, immediately 'adjust their quantum state'.

The formalization of these ideas is called *Born-Oppenheimer approximation* [173, 174] and the confirmation of its being good for many relevant problems is a fact that supports our intuitions about the topic and that lies at the foundations of the vast majority

[35] One convenient way of thinking about functions that depend on spin-like variables is as an m-tuple of ordinary $\mathbb{R}^{3N}$ functions, where m is the finite number of possible values of the spin. In the case of a one-particular wavefunction describing an electron, for example, σ can take two values (say, $-1/2$ and $1/2$) in such a way that one may picture any general *spin-orbital* $\Psi_i(x)$ as a 2-tuple $(\Phi_i^{-1/2}(\vec{r}), \Phi_i^{1/2}(\vec{r}))$. Of course, another valid way of imagining $\Psi_i(x)$ is simply as a function of four variables, three real and one discrete.

of the images, the concepts and the research in quantum chemistry[36].

Like any approximation, the Born-Oppenheimer one may be either *derived* from the exact problem (in this case, the entangled behaviour of electrons and nuclei as the same quantum object) or simply *proposed* on the basis of physical intuition, and later confirmed to be good enough (or not) by comparison with the exact theory or with the experiment. Of course, if it is possible, the first way should be preferred, since it allows to develop a deeper insight about the terms we are neglecting and the specific details that we will miss. However, although in virtually every quantum chemistry book [176–180] hands-waving derivations up to different levels of detail are performed and the Born-Oppenheimer approximation is typically presented as unproblematic, it seems that the fine mathematical details on which these 'standard' approaches are based are far from clear [181–183]. This state of affairs does not imply that the final equations that will need to be solved are ill-defined or that the numerical methods based on the theory are unstable; in fact, it is just the contrary (see the discussion below), because the problems are related only to the precise relation between the concepts in the whole theory and those in its simplified version. Nevertheless, the many subtleties involved in a *derivation* of the Born-Oppenheimer approximation scheme from the exact equations and the fact that, in this dissertation, no correcting terms to it will be calculated, suggest that the second way be taken. Hence, in the following paragraphs, an *axiomatic* presentation of the main expressions, aimed mostly to fix the notation and to introduce the language, will be performed.

First of all, if we examine the Hamiltonian operator in eq. (2.2), we see that the term $\hat{V}_{Ne}$ prevents the problem from being separable in the nuclear and electronic coordinates, i.e., if we define $\underline{\mathrm{x}} := (x_1, \ldots, x_N)$ as the set of all electronic coordinates (spatial and spin-like) and do likewise with the nuclear coordinates $\underline{\mathrm{X}}$, the term $\hat{V}_{Ne}$ prevents any wavefunction $\Psi(\underline{\mathrm{X}}, \underline{\mathrm{x}})$ solution of the *time-independent Schrödinger equation*,

$$\hat{H}\,\Psi(\underline{\mathrm{X}}, \underline{\mathrm{x}}) = \left(\hat{T}_N + \hat{T}_e + \hat{V}_{NN} + \hat{V}_{eN} + \hat{V}_{ee}\right)\Psi(\underline{\mathrm{X}}, \underline{\mathrm{x}}) = E\,\Psi(\underline{\mathrm{X}}, \underline{\mathrm{x}})\,, \tag{2.4}$$

from being written as a product, $\Psi(\underline{\mathrm{X}}, \underline{\mathrm{x}}) = \Psi_N(\underline{\mathrm{X}})\Psi_e(\underline{\mathrm{x}})$, of an electronic wavefunction and a nuclear one. If this were the case, the problem would still be difficult (because of the Coulomb terms $\hat{V}_{NN}$ and $\hat{V}_{ee}$), but we would be able to focus on the electrons and on the nuclei separately.

The starting point for the Born-Oppenheimer approximation consists in assuming that a less strict separability is achieved, in such a way that, for a pair of suitably chosen $\Psi_N(\underline{\mathrm{X}})$ and $\Psi_e(\underline{\mathrm{x}}; \underline{\mathrm{X}})$, any wavefunction solution of eq. (2.4) (or at least those in which we are interested; for example, the eigenstates corresponding to the lowest lying eigenvalues) can be expressed as

$$\Psi(\underline{\mathrm{X}}, \underline{\mathrm{x}}) = \Psi_N(\underline{\mathrm{X}})\Psi_e(\underline{\mathrm{x}}; \underline{\mathrm{X}})\,, \tag{2.5}$$

where we have used a ';' to separate the two sets of variables in the electronic part of the wavefunction in order to indicate that, in what follows, it is convenient to use the image that 'from the point of view of the electrons, the nuclear degrees of freedom are fixed', so that the electronic wavefunction depends 'parametrically' on them. In other

[36] There are many phenomena, however, in which the Born-Oppenheimer approximation is broken. For example, in striking a flint to create a spark, mechanical motion of the nuclei excites electrons into a plasma that then emits light [175].

words, that the $\underline{\mathrm{X}}$ are not quantum variables in eq. (2.6) below. Of course, it is just a 'semantic' semicolon; if anyone feels uncomfortable about it, she may drop it and write a normal comma.

Notably, in ref. 184, Hunter showed that any solution of the Schrödinger equation can in fact be written exactly in the form of eq. (2.5), and that the two functions, $\Psi_N(\underline{\mathrm{X}})$ and $\Psi_e(\underline{\mathrm{x}};\underline{\mathrm{X}})$, into which $\Psi(\underline{\mathrm{X}},\underline{\mathrm{x}})$ is split may be interpreted as marginal and conditional probability amplitudes respectively. However, despite the insight that is gained from this treatment, it is of no practical value, since the knowledge of the exact solution $\Psi(\underline{\mathrm{X}},\underline{\mathrm{x}})$ is required in order to compute $\Psi_N(\underline{\mathrm{X}})$ and $\Psi_e(\underline{\mathrm{x}};\underline{\mathrm{X}})$.

In the Born-Oppenheimer scheme, an additional assumption is made in order to avoid this drawback: *the equations obeyed by the electronic and nuclear parts of the wavefunction are supposed to be known*. Hence, $\Psi_e(\underline{\mathrm{x}};\underline{\mathrm{X}})$ is assumed to be a solution of the time-independent *clamped nuclei Schrödinger equation*,

$$\left(\hat{T}_e + \hat{V}_{eN}(\underline{\mathrm{r}};\underline{\mathrm{R}}) + \hat{V}_{ee}(\underline{\mathrm{r}})\right)\Psi_e(\underline{\mathrm{x}};\underline{\mathrm{R}}) := \hat{H}_e(\underline{\mathrm{R}})\,\Psi_e(\underline{\mathrm{x}};\underline{\mathrm{R}}) = E_e(\underline{\mathrm{R}})\,\Psi_e(\underline{\mathrm{x}};\underline{\mathrm{R}})\,, \tag{2.6}$$

where the *electronic Hamiltonian operator* $\hat{H}_e(\underline{\mathrm{R}})$ and the *electronic energy* $E_e(\underline{\mathrm{R}})$ (both dependent on the nuclei positions) have been defined, and, since the nuclear spins do not enter the expression, we have explicitly indicated that Ψ_e depends parametrically on $\underline{\mathrm{R}}$ and not on $\underline{\mathrm{X}}$.

The common interpretation of the clamped nuclei equation is, as we have advanced at the beginning of the section, that the nuclei are much 'slower' than the electrons and, therefore, the latter can automatically adjust their quantum state to the instantaneous positions of the former. Physically, eq. (2.6) is just the time-independent Schrödinger equation of N particles (the electrons) of mass m_e and charge $-e$ in the *external electric field* of N_N *point charges* (the nuclei) of size eZ_α at locations $\vec{R}_\alpha$. Mathematically, it is an eigenvalue problem that has been thoroughly studied in the literature and whose properties are well-known [185–190]. In particular, it can be shown that, in the case of neutral or positively charged molecules (i.e., with $Z := \sum_\alpha Z_\alpha \geq N$), the clamped nuclei equation has an infinite number of normalizable solutions in the discrete spectrum of $\hat{H}_e(\underline{\mathrm{R}})$ (*bound-states*) for every value of $\underline{\mathrm{R}}$ [191, 192].

These solutions must be regarded as the different *electronic energy levels*, and a further approximation that is typically made consists in, not only accepting that the electrons immediately 'follow' nuclear motion, but also that, for each value of the nuclear positions $\underline{\mathrm{R}}$, they are in the fundamental electronic state[37], i.e., the one with the lower $E_e(\underline{\mathrm{R}})$.

Consequently, we define

$$E_e^{\mathrm{eff}}(\underline{\mathrm{R}}) := E_e^0(\underline{\mathrm{R}})\,. \tag{2.7}$$

to be the *effective electronic field* in which the nuclei move, in such a way that, once we have solved the problem in eq. (2.6) and know $E_e^0(\underline{\mathrm{R}})$, the time-independent *nuclear*

[37] This is customarily assumed in the literature and it is supported by the general fact that electronic degrees of freedom are typically more difficult to excite than nuclear ones. Hence, in the vast majority of the numerical implementations of the theory, only the fundamental electronic state is sought. We will see this in the forecoming sections.

Schrödinger equation obeyed by $\Psi_N(\underline{X})$ is:

$$\left(\hat{T}_N + \hat{V}_{NN}(\underline{R}) + E_e^{\text{eff}}(\underline{R})\right)\Psi_N(\underline{X}) := \hat{H}_N\Psi_N(\underline{X}) = E_N\Psi_N(\underline{X}) \,, \tag{2.8}$$

where the *effective nuclear Hamiltonian* $\hat{H}_N$ has been implicitly defined.

Now, to close the section, we gather the main expressions of the Born-Oppenheimer approximation for quick reference and we discuss them in some more detail:

$$\hat{H}_e(\underline{R})\,\Psi_e(\underline{x};\underline{R}) := \left(\hat{T}_e + \hat{V}_{eN}(\underline{R}) + \hat{V}_{ee}\right)\Psi_e(\underline{x};\underline{R}) = E_e(\underline{R})\,\Psi_e(\underline{x};\underline{R}) \,, \tag{2.9a}$$

$$E_e^{\text{eff}}(\underline{R}) := E_e^0(\underline{R}) \,, \tag{2.9b}$$

$$\hat{H}_N\Psi_N(\underline{X}) := \left(\hat{T}_N + V_{NN}(\underline{R}) + E_e^{\text{eff}}(\underline{R})\right)\Psi_N(\underline{X}) = E_N\Psi_N(\underline{X}) \,, \tag{2.9c}$$

$$\Psi(\underline{x},\underline{X}) \simeq \Psi_e^0(\underline{x};\underline{R})\,\Psi_N(\underline{X}) \,, \qquad E \simeq E_N \,. \tag{2.9d}$$

To start, note that the above equations are written in the logical order in which they are imagined and used in any numerical calculation. First, we assume the nuclei fixed at $\underline{R}$ and we (hopefully) solve the clamped nuclei electronic Schrödinger equation (eq. (2.9a)), obtaining the fundamental electronic state $\Psi_e^0(\underline{x},\underline{R})$ with its corresponding energy $E_e^0(R)$. Next, we repeat this procedure for all possible values[38] of $\underline{R}$ and end up with an hypersurface $E_e^0(\underline{R})$ in $\underline{R}$-space. Finally, we add this function to the analytical and easily computable $V_{NN}(\underline{R})$ and find the *effective potential* that determines the nuclear motion:

$$V_N^{\text{eff}}(\underline{R}) := V_{NN}(\underline{R}) + E_e^0(\underline{R}) \,. \tag{2.10}$$

It is, precisely, this effective potential that is called *Potential Energy Surface* (PES) (or, more generally, *Potential Energy Hyper-Surface* (PEHS)) in quantum chemistry and that is the central object through which scientists picture chemical reactions or conformational changes of macromolecules [193]. In fact, the concept is so appealing and the classical image so strong that, after 'going quantum', we can 'go classical' back again and think of nuclei as perfectly classical particles that move in the classical potential $V_N^{\text{eff}}(\underline{R})$. In such a case, we would not have to solve eq. (2.9c) but, instead, integrate the Newtonian equations of motion. This is the basic assumption of every typical force field used for *molecular dynamics*, such as the ones in the popular CHARMM [104, 105], AMBER [125, 194, 195] or OPLS [124] packages.

Finally, we would like to remind the reader that, despite the hands-waving character of the arguments presented, up to this point, every computational step has a clear description and eqs. (2.9a) through (2.9c) could be considered as *definitions* involving a certain degree of notational abuse. To assume that the quantities obtained through this process are close to those that proceed from a rigorous solution of the time independent Schrödinger equation (eq. (2.4)) is where the approximation really lies. Hence, the more accurate eqs. (2.9d) are, the better the Born-Oppenheimer guess is, and, like any other one, if one

[38] Of course, this cannot be done in practice. Due to the finite character of available computational resources, what is customarily done is to define a 'grid' in $\underline{R}$-space and compute $E_e^0(\underline{R})$ in a finite number of points.

does not trust in the heuristic grounds on which the final equations stand, they may be taken as axiomatic and judged a posteriori according to their results in particular cases[39].

In quantum chemistry, the Born-Oppenheimer approximation is assumed in a great fraction of the studies and it allows the central concept of potential energy surface to be well-defined, apart from considerably simplifying the calculations. The same decision is taken in this Ph.D. dissertation.

2.4 The variational method

Looking for the fundamental state

There exists a mathematically appealing way of deriving the time independent Schrödinger equation (eq. (2.4)) from an extremal principle. If we define the functional (see appendix B) that corresponds to the expected value of the energy,

$$\mathcal{F}[\Psi] := \langle\Psi|\,\hat{H}\,|\Psi\rangle\,, \tag{2.11}$$

and we restrict the search space to the normalized wavefunctions, the constrained-extremals problem that results can be solved via the *Lagrange multipliers method* (see appendix C) by constructing the *associated functional* $\widetilde{\mathcal{F}}[\Psi]$, where we introduce a *Lagrange multiplier* λ to force normalization:

$$\widetilde{\mathcal{F}}[\Psi] := \mathcal{F}[\Psi] + \lambda\Big(\langle\Psi|\Psi\rangle - 1\Big)\,. \tag{2.12}$$

If we now ask that the functional derivative of $\widetilde{\mathcal{F}}[\Psi]$ with respect to the complex conjugate Ψ^* of the wave function[40] be zero, i.e., we look for the stationary points of $\mathcal{F}[\Psi]$ conditioned by $\langle\Psi|\Psi\rangle = 1$, we obtain the eigenvalues equation for $\hat{H}$, i.e., the time-independent Schrödinger equation. Additionally, it can be shown, first, that, due to the self-adjointedness of $\hat{H}$, the equation obtained from the stationarity condition with respect to Ψ (not Ψ^*) is just the complex conjugate and adds no new information.

Moreover, one can see that the reverse implication is also true, so that, if a given normalized wavefunction Ψ is a solution of the eigenvalue problem and belongs to the discrete spectrum of $\hat{H}$, then the functional in eq. (2.12) is stationary with respect to Ψ^*:

$$\frac{\delta\widetilde{\mathcal{F}}[\Psi]}{\delta\Psi^*} = 0 \quad\Longleftrightarrow\quad \hat{H}\,\Psi = -\lambda\,\Psi := E\,\Psi \quad \text{and} \quad \langle\Psi|\Psi\rangle = 1\,. \tag{2.13}$$

This result, despite its conceptual interest, is of little practical use, because it does not indicate an operative way to solve the Schrödinger equation different from the ones that we already knew. The equivalence above simply illustrates that mathematical variational principles are over-arching theoretical statements from which the differential equations that actually contain the details of physical systems can be extracted. Nevertheless, using

[39] Until now, two approximations have been done: the non-relativistic character of the objects studied and the Born-Oppenheimer approximation. In the forecoming, many more will be done. The a priori quantification of their goodness in large molecules is a formidable task and, despite the efforts in this direction, in the end, the comparison with experimental data is the only sound method for validation.

[40] A function of a complex variable z (or, analogously, a functional on a space of complex functions) may be regarded as depending on two different sets of independent variables: either $\mathrm{Re}(z)$ and $\mathrm{Im}(z)$ or z and z^*. The choice frequently depending on technical issues.

similar ideas, we will derive another simple theorem which is indeed powerfully practical: the *variational theorem*.

Let $\{\Psi_n\}$ be a basis of eigenstates of the Hamiltonian operator $\hat{H}$ and $\{E_n\}$ their corresponding eigenvalues. Since $\hat{H}$ is self-adjoint, the eigenstates Ψ_n can be chosen to be orthonormal (i.e., $\langle\Psi_m|\Psi_n\rangle = \delta_{mn}$) and any normalized wavefunction Ψ in the Hilbert space can be written as a linear combination of them[41]:

$$|\Psi\rangle = \sum_n C_n|\Psi_n\rangle \quad \text{provided that} \quad \sum_n |C_n|^2 = 1 \,. \tag{2.14}$$

If we now denote by E_0 the lowest E_n (i.e., the energy of the *fundamental state*)[42] and calculate the expected value of the energy on an arbitrary state Ψ such as the one in eq. (2.14), we obtain

$$\begin{aligned} \langle\Psi|\,\hat{H}\,|\Psi\rangle &= \sum_{m,n} C_m^* C_n \langle\Psi_m|\,\hat{H}\,|\Psi_n\rangle = \sum_{m,n} C_m^* C_n E_n \langle\Psi_m|\Psi_n\rangle \\ &= \sum_{m,n} C_m^* C_n E_n \delta_{mn} = \sum_n |C_n|^2 E_n \geq \sum_n |C_n|^2 E_0 = E_0 \,. \end{aligned} \tag{2.15}$$

This simple relation is the *variational theorem* and it states that any wavefunction of the Hilbert space has an energy larger than the one of the fundamental state (the equality can only be achieved if $\Psi = \Psi_0$). However trivial this fact may appear, it allows a very fruitful 'everything-goes' strategy when trying to approximate the fundamental state in a difficult problem. If one has a procedure for finding a promising guess wavefunction (called *variational ansatz*), no matter how heuristic, semi-empirical or intuitive it may be, one may expect that the lower the corresponding energy, the closer to the fundamental state it is[43]. This provides a systematic strategy for improving the test wavefunction which may take a number of particular forms.

One example of the application of the variational theorem is to propose a family of normalized wavefunctions Ψ_θ parametrically dependent on a number θ and calculate the θ-dependent expected value of the energy:

$$E(\theta) := \langle\Psi_\theta|\,\hat{H}\,|\Psi_\theta\rangle \,. \tag{2.16}$$

Then, one may use the typical tools of one-variable calculus to find the minimum of $E(\theta)$ and thus make the best guess of the energy E_0 constrained to the family Ψ_θ. If the ansatz is cleverly chosen, this estimate could be rather accurate, however, for large systems that lack symmetry, it is very difficult to write a good enough form for Ψ_θ.

[41] We assume here, for the sake of simplicity and in order to highlight the relevant concepts, that $\hat{H}$ has only discrete spectrum. The ideas involved in a general derivation are the same but the technical details and the notation are more complicated [196].

[42] Its existence is not guaranteed: it depends on the particular potential in $\hat{H}$. However, for the physically relevant cases, there is indeed a minimum energy in the set $\{E_n\}$.

[43] Of course, this not necessarily so (and, in any case, it depends on the definition of 'closer'), since it could happen that the $\langle\Psi|\hat{H}|\Psi\rangle$ landscape in the constrained subset of the Hilbert space in which the search is performed be 'rugged'. In such a case, we may have very different wavefunctions (say, in the sense of the L^2-norm) with similar energies $\langle\Psi|\hat{H}|\Psi\rangle$. The only 'direction' in which one can be sure that the situation improves when using the variational procedure is the (very important) energetic one. That one is also moving towards better values of any other observable is, in general, no more than a *bona fide* assumption.

When dealing with a large number of particles, there exists another protocol based on the variational theorem that will permit us to derive the Hartree and Hartree-Fock equations for the electronic wavefunction Ψ_e (see secs. 2.6 and 2.7, respectively). The first step is to devise a restricted way to express Ψ_e in terms of *one-electron wavefunctions*, also called *orbitals* and denoted by $\{\psi_a(x)\}$[44], thus reducing the search space to a (typically small) subset of the whole Hilbert space:

$$\Psi_e(x_1, \ldots, x_N) = f\Big(\{\psi_a(x_i)\}\Big) \,. \tag{2.17}$$

The second step consists in establishing a (possibly infinite) number of constraints on the one-electron functions[45],

$$L_k\Big(\{\psi_a(x_i)\}\Big) = 0 \,. \tag{2.18}$$

With these two ingredients, we can now write the Lagrange functional that describes the constrained problem in terms of the orbitals ψ_a (see eq. (2.12)):

$$\widetilde{\mathcal{F}}\,[\{\psi_a\}] = \Big\langle f(\{\psi_a(x_i)\}) \Big| \hat{H} \Big| f(\{\psi_a(x_i)\}) \Big\rangle + \sum_k \lambda_k L_k\Big(\{\psi_a(x_i)\}\Big) \,. \tag{2.19}$$

Finally, we take the derivatives of $\widetilde{\mathcal{F}}[\{\psi_a\}]$ with respect to every $\psi_a(x)$ (normally, with respect to the complex conjugate $\psi_a^*(x)$, see footnote 40) and we ask each one to be zero (see appendix B). This produces the final equations that must be solved in order to find the stationary one-electron orbitals.

Of course, these final equations may have multiple solutions. In the cases treated in this dissertation, there exist procedures to check that a particular solution (found computationally) is, not only stationary, but also minimal [197]. However, to assure that it is, not only locally minimal, but also globally (i.e., that is *optimal*), could be, in general, as difficult as for any other multi-dimensional optimization problem [151, 152, 198]. In the Hartree and Hartree-Fock cases, discussed in secs. 2.6 and 2.7 respectively, the aufbau principle and a clever choice of the starting guess are particular techniques intended to alleviate this problem.

2.5 Statement of the problem

Solve the electronic Hamiltonian

Assuming the Born-Oppenheimer approximation (see sec. 2.3 and eqs. (2.9)), the central problem that one must solve in quantum chemistry is *to find the fundamental state of the*

[44] In principle, there could be more orbitals than electrons, however, in both the Hartree and Hartree-Fock applications of this formalism, the index a runs, just like i, from 1 to N.

[45] Actually, both restrictions (the one at the level of the total wavefunction in eq. (2.17) and the one involving the one-particle ones in eq. (2.18)) are simply constraints (see appendix C). The distinction is not fundamental but operative, and it also helps us to devise variational ansatzs separating the two conceptual playgrounds.

electronic Hamiltonian for a fixed position $\underline{\mathrm{R}}$ *of the nuclei*[46]:

$$\hat{H} := \hat{T} + \hat{V}_{eN} + \hat{V}_{ee} := -\sum_{i=1}^{N} \frac{1}{2}\nabla_i^2 - \sum_{i=1}^{N}\sum_{\alpha=1}^{N_N} \frac{Z_\alpha}{R_{\alpha i}} + \frac{1}{2}\sum_{i\neq j} \frac{1}{r_{ij}} \,. \tag{2.20}$$

As already remarked in sec. 2.3, this problem is well posed for neutral and positively charged molecules, and, in the same way in which the term $\hat{V}_{eN}$ prevented the total wavefunction to be a product of an electronic and a nuclear part, the term $\hat{V}_{ee}$ in the expression above breaks the separability in the one-electron variables x_i of the electronic time-independent Schrödinger equation associated to $\hat{H}$. Hence, a general solution $\Psi(\underline{\mathrm{x}})$ cannot be a product of orbitals and the search must be a priori performed in the whole Hilbert space. However, this is a much too big place to look for $\Psi(\underline{\mathrm{x}})$, since the computational requirements to solve the Schrödinger equation grow exponentially on the number of electrons.

Partially recognizing this situation, in the first days of quantum mechanics, Dirac wrote that,

> The underlying physical laws necessary for the mathematical theory of a large part of physics and the whole of chemistry are thus completely known, and the difficulty is only that the exact application of these equations leads to equations much too complicated to be soluble. It therefore becomes desirable that approximate practical methods of applying quantum mechanics should be developed, which can lead to an explanation of the main features of complex atomic systems without too much computation [196].

The description of the most popular approximate methods, which the great physicist envisaged to be necessary, will be the objective of the following sections. Two basic points responsible of the relative success of such an enterprise are the severe reduction of the space in where the fundamental state is sought (which, of course, leads to only an approximation of it) and the availability of computers unimaginably faster than anything that could be foreseen in times of Dirac.

2.6 The Hartree approximation

Distinguishable spinless independent electrons

One of the first and most simple approximations aimed to solve the problem posed in the previous section is due to Hartree in 1927 [171] (although the way in which the Hartree equations will be derived here, using the variational theorem, is due to Slater [199]). In this approximation, the total wavefunction is constrained to be a product (typically referred to as *Hartree product*) of N one-electron orbitals (see eq. (2.17)), where the spin

[46] Since, from now on, we will only be dealing with the 'electronic problem', the notation has been made simpler by dropping superfluous subindices e where there is no possible ambiguity. As a consequence, for example, the electronic Hamiltonian is now denoted by $\hat{H}$, the electronic kinetic energy by $\hat{T}$ and the electronic wavefunction by $\Psi(\underline{\mathrm{x}})$ (dropping the parametric dependence on $\underline{\mathrm{R}}$ in the same spirit).

of the electrons and the antisymmetry (i.e., the Pauli exclusion principle) are not taken into account[47]:

$$\Phi(\vec{r}_1,\ldots,\vec{r}_N) = \prod_{i=1}^{N} \phi_i(\vec{r}_i)\,, \tag{2.21}$$

where the a index in the orbitals has been substituted by i due to the fact that each function is paired to a specific set of electron coordinates, consequently being the same number of both of them.

Also, the additional requirement that the one-particle wavefunctions be normalized is imposed (see eq. (2.18)):

$$\langle\phi_i|\phi_i\rangle = 1\,, \qquad i = 1,\ldots,N\,. \tag{2.22}$$

With these two ingredients, we can construct the auxiliary functional whose zero-derivative condition produces the solution of the constrained stationary points problem (see eq. (2.19)). To this effect, we introduce N Lagrange multipliers λ_i that force the normalization constraints[48]:

$$\widetilde{\mathcal{F}}\left[\{\phi_i\}\right] = \left\langle \prod_{i=1}^{N}\phi_i(\vec{r}_i)\middle|\,\hat{H}\,\middle|\prod_{i=1}^{N}\phi_i(\vec{r}_i)\right\rangle + \sum_{i=1}^{N}\lambda_i\Big(\langle\phi_i|\phi_i\rangle - 1\Big)\,. \tag{2.23}$$

This functional may be considered to depend on $2N$ independent functions: the N one-electron ϕ_i and their N complex conjugates (see footnote 40). The *Hartree equations* are obtained by imposing that the functional derivative of $\widetilde{\mathcal{F}}$ with respect to ϕ_k^* be zero for $k = 1,\ldots,N$. In order to obtain them and as an appetizer for the slightly more complicated process in the more used Hartree-Fock approximation, the functional derivative will be here computed in detail following the steps indicated in appendix B.

First, we write out[49] the first term in the right-hand side of eq. (2.23):

$$\begin{aligned}
&\left\langle \prod_i \phi_i(\vec{r}_i)\middle|\,\hat{H}\,\middle|\prod_i \phi_i(\vec{r}_i)\right\rangle = \\
&\quad -\frac{1}{2}\sum_i\left(\prod_{j\neq i}\langle\phi_j|\phi_j\rangle\right)\int \phi_i^*(\vec{r})\nabla^2\phi_i(\vec{r})\mathrm{d}\vec{r} \\
&\quad -\sum_i\left(\prod_{j\neq i}\langle\phi_j|\phi_j\rangle\right)\int \phi_i^*(\vec{r})\,\phi_i(\vec{r})\left(\sum_{A=1}^{N_N}\frac{Z_A}{|\vec{r}-\vec{R}_A|}\right)\mathrm{d}\vec{r} \\
&\quad +\frac{1}{2}\sum_i\sum_{j\neq i}\left(\prod_{k\neq i,j}\langle\phi_k|\phi_k\rangle\right)\iint \frac{\phi_i^*(\vec{r})\,\phi_i(\vec{r})\,\phi_j^*(\vec{r}\,')\,\phi_j(\vec{r}\,')}{|\vec{r}-\vec{r}\,'|}\,\mathrm{d}\vec{r}\,'\mathrm{d}\vec{r}\,,
\end{aligned} \tag{2.24}$$

[47] We shall denote with capital Greek letters the wavefunctions depending on all the electronic variables, and with lowercase Greek letters the one-electron orbitals. In addition, by Ψ (or ψ), we shall indicate wavefunctions containing spin part (called *spin-orbitals*) and, by Φ (or ϕ), those that depend only on spatial variables.

[48] Note that the normalization of the total wavefunction is a consequence of the normalization of the one-electron ones and needs not to be explicitly asked.

[49] The limits in sums and products are dropped if there is no possible ambiguity.

where $\mathrm{d}\vec{r}$ denotes the Euclidean $\mathbb{R}^3$ volume element $\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z$.

Now, we realize that the products outside the integrals can be dropped using the constraints in eq. (2.22) (see the last paragraphs of appendix C for a justification that this can be done before taking the derivative). Then, using the previous expression and conveniently rearranging the order of the integrals and sums, we write out the first term in the numerator of the left-hand side of eq. (B.1) that corresponds to an infinitesimal variation of the function ϕ_k^*:

$$
\begin{aligned}
&\widetilde{\mathcal{F}}\left[\phi_k^* + \varepsilon\delta\phi_k^*\right] := \\
&\widetilde{\mathcal{F}}\left[\phi_1, \phi_1^*, \ldots, \phi_k, \phi_k^* + \varepsilon\delta\phi_k^*, \ldots, \phi_N, \phi_N^*\right] = \\
&\quad -\frac{1}{2}\sum_i \int \phi_i^*(\vec{r})\nabla^2\phi_i(\vec{r})\,\mathrm{d}\vec{r} - \sum_i \int \left|\phi_i(\vec{r})\right|^2 \left(\sum_{A=1}^{N_N} \frac{Z_A}{|\vec{r} - \vec{R}_A|}\right)\mathrm{d}\vec{r} \\
&\quad + \frac{1}{2}\sum_i \int \left|\phi_i(\vec{r})\right|^2 \left(\int \frac{\sum_{j\neq i}\left|\phi_j(\vec{r}\,')\right|^2}{|\vec{r} - \vec{r}\,'|}\,\mathrm{d}\vec{r}\,'\right)\mathrm{d}\vec{r} + \sum_i \lambda_i \left(\int \left|\phi_i(\vec{r})\right|^2 \mathrm{d}\vec{r} - 1\right) \\
&\quad - \frac{1}{2}\varepsilon \int \delta\phi_k^*(\vec{r})\nabla^2\phi_k(\vec{r})\,\mathrm{d}\vec{r} - \varepsilon \int \delta\phi_k^*(\vec{r})\,\phi_k(\vec{r}) \left(\sum_{A=1}^{N_N} \frac{Z_A}{|\vec{r} - \vec{R}_A|}\right)\mathrm{d}\vec{r} \\
&\quad + \varepsilon \int \delta\phi_k^*(\vec{r})\,\phi_k(\vec{r}) \left(\int \frac{\sum_{i\neq k}\left|\phi_i(\vec{r}\,')\right|^2}{|\vec{r} - \vec{r}\,'|}\mathrm{d}\vec{r}\,'\right)\mathrm{d}\vec{r} + \varepsilon\lambda_k \int \delta\phi_k^*(\vec{r})\,\phi_k(\vec{r})\,\mathrm{d}\vec{r}\,. \qquad (2.25)
\end{aligned}
$$

We subtract from this expression the quantity $\widetilde{\mathcal{F}}\left[\{\phi_i(\vec{r}_i)\}\right]$, so that the first four terms cancel, and we can write

$$
\begin{aligned}
&\lim_{\varepsilon\to 0} \frac{\widetilde{\mathcal{F}}\left[\phi_k^* + \varepsilon\delta\phi_k^*\right] - \widetilde{\mathcal{F}}\left[\phi_k^*\right]}{\varepsilon} = \\
&\qquad \int \left[-\frac{1}{2}\nabla^2\phi_k(\vec{r}) - \left(\sum_{A=1}^{N_N} \frac{Z_A}{|\vec{r} - \vec{R}_A|}\right)\phi_k(\vec{r}) \right. \\
&\qquad \left. + \left(\int \frac{\sum_{i\neq k}|\phi_i(\vec{r}\,')|^2}{|\vec{r} - \vec{r}\,'|}\,\mathrm{d}\vec{r}\,'\right)\phi_k(\vec{r}) + \lambda_k\phi_k(\vec{r})\right]\delta\phi_k^*(\vec{r})\,\mathrm{d}\vec{r}\,. \qquad (2.26)
\end{aligned}
$$

Now, by simple inspection of the right-hand side, we see that the functional derivative (see eq. (B.1)) is the part enclosed by square brackets:

$$
\frac{\delta\widetilde{\mathcal{F}}\left[\{\phi_i\}\right]}{\delta\phi_k^*} = \left(-\frac{1}{2}\nabla^2 + \hat{V}_e(\vec{r}) + \hat{V}_e^k(\vec{r}) + \lambda_k\right)\phi_k(\vec{r})\,, \qquad (2.27)
$$

where the *nuclear potential energy* and the *electronic potential energy* have been respectively defined as[50]

[50] Compare the notation with the one in eqs. (2.3), here a subindex e has been dropped to distinguish the new objects defined.

$$\hat{V}_N(\vec{r}) := -\sum_{A=1}^{N_N} \frac{Z_A}{|\vec{r} - \vec{R}_A|} \,, \tag{2.28a}$$

$$\hat{V}_e^k(\vec{r}) := \int \frac{\sum_{i\neq k} |\phi_i(\vec{r}\,')|^2}{|\vec{r} - \vec{r}\,'|} \, \mathrm{d}\vec{r}\,' \,. \tag{2.28b}$$

Finally, if we ask the functional derivative to be zero for $k = 1, \ldots, N$ and we define $\varepsilon_k := -\lambda_k$, we arrive to the equations that the stationary points must satisfy, the *Hartree equations*:

$$\hat{\mathcal{H}}_k[\phi]\, \phi_k(\vec{r}) := \left(-\frac{1}{2}\nabla^2 + \hat{V}_N(\vec{r}) + \hat{V}_e^k(\vec{r}) \right) \phi_k(\vec{r}) = \varepsilon_k\, \phi_k(\vec{r}) \,, \quad k = 1, \ldots, N \,. \tag{2.29}$$

Let us note that, despite the fact that the object $\hat{\mathcal{H}}_k[\phi]$ defined above is not a operator strictly speaking, since, as the notation emphasizes, it depends on the orbitals $\phi_{i\neq k}$, we will stick to the name *Hartree operator* for it, in order to be consistent with most of the literature.

Now, some remarks related to the Hartree equations are worth making. First, it can be shown that, if the variational ansatz in eq. (2.21) included the spin degrees of freedom of the electrons, all the expressions above would be kept, simply changing the orbitals $\phi_i(\vec{r}_i)$ by the spin-orbitals $\psi_i(\vec{r}_i, \sigma_i)$.

Secondly, and moving into more conceptual playgrounds, we note that the special structure of $\hat{V}_e^k(\vec{r})$ in eq. (2.28b) makes it mandatory to interpret the Hartree scheme as one in which each electron 'feels' only the average effect of the rest. In fact, if the *quantum charge density* $\rho_i(\vec{r}) := |\phi_i(\vec{r})|^2$ is regarded for a moment as a classical continuum distribution, then the potential produced by all the electrons but the k-th is precisely the one in eq. (2.28b). Supporting this image, note also the fact that, if we write the joint probability density of electron 1 being at the point $\vec{r}_1$, electron 2 being at the point $\vec{r}_2$ and so on (simply squaring eq. (2.21)),

$$\rho(\vec{r}_1, \ldots, \vec{r}_N) := \left|\Phi(\vec{r}_1, \ldots, \vec{r}_N)\right|^2 = \prod_{i=1}^{N} \left|\phi_i(\vec{r}_i)\right|^2 = \prod_{i=1}^{N} \rho_i(\vec{r}_i) \,, \tag{2.30}$$

we see that, *in a probabilistic sense*, the electrons are independent (they could not be independent in a physical, complete sense, since we have already said that they 'see' each other in an average way).

Anyway, despite these appealing images and also despite the fact that, disguised under the misleading (albeit common) notation, these equations seem 'one-particle', they are rather complicated from a mathematical point of view. On the one hand, it is true that, whereas the original electronic Schrödinger equation in (2.9a) depended on $3N$ spatial variables, the expressions above only depend on 3. This is what we have gained from drastically reducing the search space to the set of Hartree products in eq. (2.21) and what renders the approximation tractable. On the other hand, however, we have paid the price of greatly increasing the mathematical complexity of the expressions, so that, while the electronic Schrödinger equation was one linear differential equation, the Hartree ones in (2.29) are N coupled non-linear integro-differential equations [200].

This complexity precludes any analytical approach to the problem and forces us to look for the solutions using less reliable iterative methods. Typically, in computational studies, one proposes a *starting guess* for the set of N orbitals $\{\phi_k^0\}$; with them, the Hartree operator $\hat{\mathcal{H}}_k[\phi^0]$ in the left-hand side of eq. (2.29) is constructed for every k and the N equations are solved as simple eigenvalue problems. For each k, the ϕ_k^1 that corresponds to the lowest ε_k^1 is selected and a new Hartree operator $\hat{\mathcal{H}}_k[\phi^1]$ is constructed with the $\{\phi_k^1\}$. The process is iterated until (hopefully) the n-th set of solutions $\{\phi_k^n\}$ differ from the $(n-1)$-th one $\{\phi_k^{n-1}\}$ less than a reasonably small amount.

Many technical issues exist that raise doubts about the possible success of such an approach. The most important ones being related to the fact that a proper definition of the *Hartree problem* should be: *find the global minimum of the energy functional* $\langle\Phi|\hat{H}|\Phi\rangle$ *under the constraint that the wavefunction* Φ *be a Hartree product* and not: *solve the Hartree equations* (2.29), whose solutions indeed include the global minimum sought but also all the rest of stationary points.

While the possibility that a found solution be a maximum or a saddle point can be typically ruled out [197, 200], as we remarked in sec. 2.4 and due to the fact that there are an infinite number of solutions to the Hartree equations [201], to be sure that any found minimum is the global one is, in a general case, impossible. There exists, however, one way, related to a theorem by Simon and Lieb [202, 203], of hopefully biasing a particular found solution of the Hartree equations to be the global minimum that we are looking for. They showed that, first, for neutral or positively charged molecules ($Z \geq N$), the Hartree global minimization problem has a solution (its uniqueness is not established yet [200]) and, second, that the minimizing orbitals $\{\phi_k\}$ correspond to the lowest eigenvalues of the $\hat{\mathcal{H}}_k[\phi]$ operators self-consistently constructed with them[51]. Now, although *the reverse of the second part of the theorem is not true in general* (i.e., from the fact that a particular set of orbitals are the eigenstates corresponding to the lowest eigenvalues of the associated Hartree operators, does not necessarily follow that they are the optimal ones) [200], in practice, the insight provided by Lieb and Simon's result is invoked to build each successive state in the iterative procedure described above, choosing the lowest lying eigenstates each time. In this way, although one cannot be sure that the global minimum has been reached, the fact that the found one has a property that the former also presents is regarded as a strong hint that it must be so (see also the discussion for the Hartree-Fock case in the next section).

This drawback and all the problems arising from the fact that an iterative procedure such as the one described above could converge to a fixed point, oscillate eternally or even diverge, are circumvented in practice by a clever choice of the starting guess orbitals $\{\phi_k^0\}$. If they are extracted, for example, from a slightly less accurate theory, one may expect that they could be 'in the basin of attraction' of the true Hartree minimum (so that the stationary point found will be the correct one) and close to it (so that the iterative pro-

[51] In quantum chemistry, where the number of electrons considered is typically small, the version of the Hartree equations that is used is the one derived here, with the Hartree operators depending on the index k in a non-trivial way. However, if the number of electrons is large enough (such as in condensed matter applications), is customary to add to the effective electronic repulsion in eq. (2.28b) the self-interaction of electron k with himself. In such a case, the Hartree operator is independent of k so that, after having achieved self-consistency, the orbitals ϕ_k turn out to be eigenstates corresponding to different eigenvalues of the same Hermitian operator, $\hat{\mathcal{H}}[\phi]$, and, therefore, mutually orthogonal.

cedure will converge). This kind of wishful thinking combined with large amounts of heuristic protocols born from many decades of trial-and-error-derived knowledge pervade and make possible the whole quantum chemistry discipline.

2.7 The Hartree-Fock approximation

Indistinguishable quasi-independent electrons with spin

The Hartree theory discussed in the previous section is not much used in quantum chemistry and many textbooks on the subject do not even mention it. Although it contains the seed of almost every concept underlying the *Hartree-Fock approximation* discussed in this section, it lacks an ingredient that turns out to be essential to correctly describe the behaviour of molecular species: *the indistinguishability of the electrons*. This was noticed independently by Fock [204] and Slater [199] in 1930, and it was corrected by proposing a variational ansatz for the total wavefunction that takes the form of a so-called *Slater determinant* (see eq. (2.32) below).

The most important mathematical consequence of the indistinguishability among a set of N quantum objects of the same type is the requirement that the total N-particle wavefunction must either remain unchanged (*symmetric*) or change sign (*antisymmetric*) when any pair of coordinates, x_i and x_j, are swapped. In the first case, the particles are called *bosons* and must have integer spin, while in the second case, they are called *fermions* and have semi-integer spin. Electrons are fermions, so the total wavefunction must be antisymmetric under the exchange of any pair of one-electron coordinates. This is a property that is certainly not met by the single Hartree product in eq. (2.21) but that can be easily implemented by forming linear combinations of many of them. The trick is to add all the possible Hartree products that are obtained from eq. (2.21) changing the order of the orbitals labels while keeping the order of the coordinates ones[52], and assigning to each term the sign of the permutation p needed to go from the natural order $1,\dots,N$ to the corresponding one $p(1),\dots,p(N)$. The sign of a permutation p is 1 if p can be written as a composition of an even number of two-element transpositions, and it is -1 if the number of transpositions needed is odd. Therefore, we define $\mathcal{T}(p)$ as the minimum number[53] of transpositions needed to perform the permutation p, and we write the sign of p as $(-1)^{\mathcal{T}(p)}$.

Using this, an antisymmetric wavefunction constructed from Hartree products of N different orbitals may be written as

$$\Psi(x_1,\dots,x_N) = \frac{1}{\sqrt{N!}} \sum_{p\in S_N} (-1)^{\mathcal{T}(p)} \psi_{p(1)}(x_1)\cdots\psi_{p(N)}(x_N)\,, \tag{2.31}$$

where the factor $1/\sqrt{N!}$ enforces normalization of the total wavefunction Ψ (if we use the constraints in eq. (2.33)) and S_N denotes the *symmetric group* of order N, i.e., the set of all permutations of N elements (with a certain multiplication rule).

The above expression is more convenient to perform the calculations that lead to the Hartree-Fock equations, however, there is also a compact way of rewriting eq. (2.31)

[52] It is immaterial whether the orbitals labels are kept and the coordinates ones changed or vice versa.

[53] It can be shown that the parity of all decompositions of p into products of elementary transpositions is the same. We have chosen the minimum only for $\mathcal{T}(p)$ to be well defined.

which is commonly found in the literature and that is useful to illustrate some particular properties of the problem. It is the *Slater determinant*:

$$\Psi(x_1,\ldots,x_N) = \frac{1}{\sqrt{N!}} \begin{vmatrix} \psi_1(x_1) & \psi_2(x_1) & \cdots & \psi_N(x_1) \\ \psi_1(x_2) & \psi_2(x_2) & \cdots & \psi_N(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ \psi_1(x_N) & \psi_2(x_N) & \cdots & \psi_N(x_N) \end{vmatrix} . \tag{2.32}$$

Now, having established the constraints on the form of the total wavefunction, we ask the Hartree-Fock one-electron orbitals to be, not only normalized, like we did in the Hartree case, but also mutually orthogonal:

$$\langle\psi_i|\psi_j\rangle = \delta_{ij} , \qquad i,j = 1,\ldots,N . \tag{2.33}$$

Note also that, contrarily to what we did in the previous section, we have now used one-electron wavefunctions ψ_i dependent also on the spin σ (i.e., spin-orbitals) to construct the variational ansatz. A *general* spin-orbital[54] may be written as (see also footnote 35)

$$\psi(x) = \phi^{\alpha}(\vec{r})\,\alpha(\sigma) + \phi^{\beta}(\vec{r})\,\beta(\sigma) , \tag{2.34}$$

where the functions α and β correspond to the *spin-up* and *spin-down* eigenstates of the operator associated to the z-component of the one-electron spin. They are defined as

$$\begin{aligned} \alpha(-1/2) &= 0 \qquad \beta(-1/2) = 1 \\ \alpha(1/2) &= 1 \qquad \beta(1/2) = 0 . \end{aligned} \tag{2.35}$$

Next, to calculate the expected value of the energy in a state such as the one in eqs. (2.31) and (2.32), let us denote the one-particle part (that operates on the i-th coordinates) of the total electronic Hamiltonian $\hat{H}$ in eq. (2.20) by

$$\hat{h}_i := -\frac{\nabla_i^2}{2} - \sum_{\alpha=1}^{N_N} \frac{Z_\alpha}{|\vec{R}_\alpha - \vec{r}_i|} , \tag{2.36}$$

in such a way that,

$$\langle\Psi|\,\hat{H}\,|\Psi\rangle = \sum_i \langle\Psi|\,\hat{h}_i\,|\Psi\rangle + \frac{1}{2}\sum_{i\neq j}\langle\Psi|\,\frac{1}{r_{ij}}\,|\Psi\rangle , \tag{2.37}$$

where $r_{ij} := |\vec{r}_j - \vec{r}_i|$.

We shall compute separately each one of the sums in the expression above. Let us start now with the first one: For a given i in the sum, the expected value $\langle\Psi|\,\hat{h}_i\,|\Psi\rangle$ is a sum of $(N!)^2$ terms of the form

$$\frac{1}{N!}\,(-1)^{\mathcal{T}(p)+\mathcal{T}(p')}\langle\psi_{p(1)}(x_1)\cdots\psi_{p(N)}(x_N)\,\big|\,\hat{h}_i\,\big|\,\psi_{p'(1)}(x_1)\cdots\psi_{p'(N)}(x_N)\rangle , \tag{2.38}$$

[54] Note that, if we had not included the spin degrees of freedom, the search space would have been half as large, since, where we now have $2N$ functions of $\vec{r}$ (i.e., $\phi_i^{\alpha}(\vec{r})$ and $\phi_i^{\beta}(\vec{r})$, with $i = 1,\ldots,N$), we would have had just N (the $\phi_i(\vec{r})$).

but, since $\hat{h}_i$ operates on the x_i and due to the orthogonality of the spin-orbitals with different indices, we have that the only non-zero terms are those with $p = p'$. Taking this into account, all permutations are still present, and we see that every orbital ψ_j appears depending on each coordinates x_i (in the terms for which $p(i) = j$). In such a case, the term above reads

$$\frac{1}{N!}\left(\prod_{k\neq j}\langle\psi_k|\psi_k\rangle\right)\langle\psi_j|\,\hat{h}\,|\psi_j\rangle\,, \tag{2.39}$$

where we have used that $(-1)^{2\mathcal{T}(p)} = 1$, and we have dropped the index i from $\hat{h}_i$ noticing that the integration variables in $\langle\psi_j(x_i)|\,\hat{h}_i\,|\psi_j(x_i)\rangle$ are actually dummy.

Next, we use again the one-electron wavefunctions constraints in eq. (2.33) to remove the product of norms in brackets, and we realize that, for each j, there as many terms like the one in the expression above as permutations of the remaining $N-1$ orbital indices (i.e., $(N-1)!$). In addition, we recall that all j must occur and we perform the first sum in eq. (2.37), yielding

$$\sum_i\langle\Psi|\,\hat{h}_i\,|\Psi\rangle = \sum_i\frac{(N-1)!}{N!}\sum_j\langle\psi_j|\,\hat{h}\,|\psi_j\rangle = \sum_j\langle\psi_j|\,\hat{h}\,|\psi_j\rangle\,. \tag{2.40}$$

The next step is to calculate the second sum in eq. (2.37). Again, we have that, for each pair (i,j), $\langle\Psi|\,1/r_{ij}\,|\Psi\rangle$ is a sum of $(N!)^2$ terms like

$$\frac{1}{N!}(-1)^{\mathcal{T}(p)+\mathcal{T}(p')}\langle\psi_{p(1)}(x_1)\cdots\psi_{p(N)}(x_N)\,\Big|\,\frac{1}{r_{ij}}\,\Big|\,\psi_{p'(1)}(x_1)\cdots\psi_{p'(N)}(x_N)\rangle\,. \tag{2.41}$$

For this expected value, contrarily to the case of $\hat{h}_i$ and due to the two-body nature of the operator $1/r_{ij}$, not only do the terms with $p = p'$ survive, but also those in which p and p' only differ on i and j, i.e., those for which $p(i) = p'(j)$, $p(j) = p'(i)$ and $p(k) = p'(k), \forall k \neq i, j$.

Using that $1/r_{ij}$ operates only on x_i and x_j, the orthonormality conditions in eq. (2.33) and the fact that $(-1)^{2\mathcal{T}(p)} = 1$, we have that, when ψ_k depends on x_i and ψ_l depends on x_j, the $p = p'$ part of the term in eq. (2.41) reads

$$\frac{1}{N!}\langle\psi_k\psi_l|\,\frac{1}{r}\,|\psi_k\psi_l\rangle\,, \tag{2.42}$$

where we have defined

$$\langle\psi_i\psi_j|\,\frac{1}{r}\,|\psi_k\psi_l\rangle := \sum_{\sigma,\sigma'}\iint\frac{\psi_i^*(x)\,\psi_j(x')\,\psi_k^*(x)\,\psi_l(x')}{|\vec{r}-\vec{r}'|}\,\mathrm{d}\vec{r}\,\mathrm{d}\vec{r}'\,. \tag{2.43}$$

Next, we see that, for each pair (k,l), there are $(N-2)!$ permutations among the $N-2$ indices of the orbitals on which $1/r_{ij}$ does not operate, therefore producing $(N-2)!$ identical terms like the one above. In addition, if we perform the sum on i and j in eq. (2.37) and remark that the term in eq. (2.42) does not depend on the pair (i,j) (which

is obvious from the suggestive notation above), we have that the $p = p'$ part of the second sum in eq. (2.37), which is typically called *Coulomb energy*, reads

$$\frac{1}{2}\sum_{i\neq j}\frac{(N-2)!}{N!}\sum_{k\neq l}\langle\psi_k\psi_l|\,\frac{1}{r}\,|\psi_k\psi_l\rangle = \frac{1}{2}\sum_{k\neq l}\langle\psi_k\psi_l|\,\frac{1}{r}\,|\psi_k\psi_l\rangle\,. \tag{2.44}$$

On the other hand, in the case in which p and p' only differ in that the indices of the orbitals that depend on x_i and x_j are swapped, all the derivation above applies except for the facts that, first, $(-1)^{\mathcal{T}(p)+\mathcal{T}(p')} = -1$ and, second, the indices k and l must be *exchanged* in eq. (2.44) (it is immaterial if they are exchanged in the bra or in the ket, since the indices are summed over and are dummy). Henceforth, the remaining part of the second sum in eq. (2.37), typically termed *exchange energy*, may be written as

$$-\frac{1}{2}\sum_{k\neq l}\langle\psi_k\psi_l|\,\frac{1}{r}\,|\psi_l\psi_k\rangle\,, \tag{2.45}$$

so that the expected value of the energy in the Hartree-Fock variational state Ψ turns out to be

$$E := \langle\Psi|\,\hat{H}\,|\Psi\rangle = \sum_i \underbrace{\langle\psi_i|\,\hat{h}\,|\psi_i\rangle}_{h_i} + \frac{1}{2}\sum_{i,j}\left(\underbrace{\langle\psi_i\psi_j|\,\frac{1}{r}\,|\psi_i\psi_j\rangle}_{J_{ij}} - \underbrace{\langle\psi_i\psi_j|\,\frac{1}{r}\,|\psi_j\psi_i\rangle}_{K_{ij}}\right), \tag{2.46}$$

where the *one-electron integrals* h_i have been defined together with the *two-electron integrals*, J_{ij} and K_{ij}, and the fact that $J_{ii} = K_{ii}, \forall i$ has been used to include the diagonal terms in the second sum.

Now, the energy functional above is what we want to minimize under the orthonormality constraints in eq. (2.33). So we are prepared to write the auxiliary functional $\widetilde{\mathcal{F}}$, introducing N^2 Lagrange multipliers λ_{ij} (see eq. (2.19) and compare with the Hartree example in the previous section):

$$\widetilde{\mathcal{F}}\left[\{\psi_i\}\right] = \sum_i h_i + \frac{1}{2}\sum_{i,j}(J_{ij} - K_{ij}) + \sum_{i,j}\lambda_{ij}\left(\langle\psi_i|\psi_j\rangle - \delta_{ij}\right). \tag{2.47}$$

In order to get to the *Hartree-Fock* equations that the stationary orbitals ψ_k must satisfy, we impose that the functional derivative of $\widetilde{\mathcal{F}}\left[\{\psi_i\}\right]$ with respect to ψ_k^* be zero. To calculate $\delta\widetilde{\mathcal{F}}/\delta\psi_k^*$, we follow the procedure described in appendix B, using the same notation as in eq. (2.25):

$$\begin{aligned}
&\lim_{\varepsilon\to 0}\frac{\widetilde{\mathcal{F}}\left[\psi_k^* + \varepsilon\delta\psi_k^*\right] - \widetilde{\mathcal{F}}\left[\psi_k^*\right]}{\varepsilon} = \\
&\langle\delta\psi_k|\,\hat{h}\,|\psi_k\rangle + \sum_j\left(\langle\delta\psi_k\psi_j|\,\frac{1}{r}\,|\psi_k\psi_j\rangle - \langle\delta\psi_k\psi_j|\,\frac{1}{r}\,|\psi_j\psi_k\rangle\right) + \sum_j\lambda_{kj}\langle\delta\psi_k|\psi_j\rangle = \\
&\int\left[\hat{h}\,\psi_k(x) + \sum_j\left(\psi_k(x)\int\frac{|\psi_j(x')|^2}{|\vec{r}-\vec{r}\,'|}\,\mathrm{d}x' - \psi_j(x)\int\frac{\psi_j^*(x')\,\psi_k(x')}{|\vec{r}-\vec{r}\,'|}\,\mathrm{d}x'\right)\right. \\
&\left. + \sum_j\lambda_{kj}\,\psi_j(x)\right]\delta\psi_k^*(x)\,\mathrm{d}x\,,
\end{aligned} \tag{2.48}$$

where we have used the more compact notation $\int \mathrm{d}x$ instead of $\sum_\sigma \int \mathrm{d}\vec{r}$.

Now, like in the previous section, by simple inspection of the right-hand side, we see that the functional derivative is the part enclosed by square brackets (see eq. (B.1)):

$$\frac{\delta \widetilde{\mathcal{F}}\,[\{\psi_i\}]}{\delta \psi_k^*} = \left[\hat{h} + \sum_j \left(\hat{J}_j[\psi] - \hat{K}_j[\psi] + \lambda_{kj}\right)\right] \psi_k(\vec{x})\,, \tag{2.49}$$

where the *Coulomb* and *exchange operators* are respectively defined by their action on an arbitrary function $\varphi(x)$ as follows[55]:

$$\hat{J}_j[\psi]\,\varphi(x) := \left(\int \frac{|\psi_j(x')|^2}{|\vec{r} - \vec{r}\,'|}\,\mathrm{d}x'\right)\varphi(x)\,, \tag{2.50a}$$

$$\hat{K}_j[\psi]\,\varphi(x) := \left(\int \frac{\psi_j^*(x')\,\varphi(x')}{|\vec{r} - \vec{r}\,'|}\,\mathrm{d}x'\right)\psi_j(x)\,. \tag{2.50b}$$

Therefore, if we define the *Fock operator* as

$$\hat{F}[\psi] := \hat{h} + \sum_j \left(\hat{J}_j[\psi] - \hat{K}_j[\psi]\right), \tag{2.51}$$

and denote $\eta_{kj} := -\lambda_{kj}$, we may arrive to a first version of the *Hartree-Fock equations* by asking that the functional derivative in eq. (2.49) be zero:

$$\hat{F}[\psi]\,\psi_k(x) = \sum_j \eta_{kj}\psi_j(x)\,, \qquad k = 1, \ldots, N\,. \tag{2.52}$$

Now, in order to obtain a simpler version of them, we shall take profit for the fact that the whole problem is invariant under a unitary transformation among the one-electron orbitals.

If we repeat the calculation in eq. (2.48) but varying ψ_k this time, instead of ψ_k^*, we find that

$$\lim_{\varepsilon\to 0} \frac{\widetilde{\mathcal{F}}\,[\psi_k + \varepsilon\delta\psi_k] - \widetilde{\mathcal{F}}\,[\psi_k]}{\varepsilon} = \\ \langle\psi_k|\,\hat{h}\,|\delta\psi_k\rangle + \sum_j \left(\langle\psi_k\psi_j|\,\frac{1}{r}\,|\delta\psi_k\psi_j\rangle - \langle\psi_j\psi_k|\,\frac{1}{r}\,|\delta\psi_k\psi_j\rangle\right) + \sum_j \lambda_{jk}\langle\psi_j|\delta\psi_k\rangle\,, \tag{2.53}$$

where the order of the indices in the exchange part has been conveniently arranged (using that they are dummy) in order to facilitate the next part of the calculations.

If we use the following relations:

$$\langle\psi_i|\psi_j\rangle^* = \langle\psi_j|\psi_i\rangle\,, \tag{2.54a}$$

$$\langle\psi_i|\,\hat{h}\,|\psi_j\rangle^* = \langle\psi_j|\,\hat{h}\,|\psi_i\rangle\,, \tag{2.54b}$$

$$\langle\psi_i\psi_j|\,\frac{1}{r}\,|\psi_k\psi_l\rangle^* = \langle\psi_k\psi_l|\,\frac{1}{r}\,|\psi_i\psi_j\rangle\,, \tag{2.54c}$$

[55] Like in the Hartree case in the previous section, the word *operator* is a common notational abuse if they act upon the very ψ_i on which they depend. This is again made explicit in the notation.

we may subtract the complex conjugate of eq. (2.53) from eq. (2.48) (both must be zero when evaluated on solutions of the Hartree-Fock equations) yielding

$$\sum_j (\lambda_{kj} - \lambda^*_{jk}) \langle \delta\psi_k | \psi_j \rangle = 0 \,, \qquad k = 1, \ldots, N \,, \tag{2.55}$$

so that, since the variations $\delta\psi_k$ are arbitrary, we have that $\lambda_{kj} = \lambda^*_{jk}$ and the $N \times N$ matrix $\lambda := (\lambda_{ij})$ of Lagrange multipliers is Hermitian (i.e., $\lambda = \lambda^+$). Of course, the same is true about the matrix $\eta := (\eta_{ij})$ in eq. (2.52), and, consequently, a unitary matrix U exists that *diagonalizes* η; in the sense that $\varepsilon := U^{-1}\eta U = U^+\eta U$ is a diagonal matrix, i.e., $\varepsilon_{ij} = \delta_{ij}\varepsilon_i$.

Using that unitary matrix U or any other one, we can transform the set of orbitals $\{\psi_i\}$ into a new one $\{\psi'_i\}$:

$$\psi_k(x) = \sum_j U_{kj}\, \psi'_j(x) \,. \tag{2.56}$$

This transformation is physically legitimate since it only changes the N-electron wavefunction Ψ in an unmeasurable phase $e^{i\phi}$. To see this, let us denote by S_{ij} the (ij)-element of the matrix inside the Slater determinant in eq. (2.32), i.e., $S_{ij} := \psi_j(x_i)$. Then, after using the expression above, the (ij)-element of the new matrix S' can be related to the old ones via $S_{kj} = \sum_j U_{kj} S'_{ij}$, in such a way that $S = S'U^T$ and the desired result follows:

$$\Psi(\{\psi_i\}) = \frac{\det S}{\sqrt{N!}} = \frac{\det(S'U^T)}{\sqrt{N!}} = \frac{\det S' \det U^T}{\sqrt{N!}} = \frac{e^{i\phi}\det S'}{\sqrt{N!}} = e^{i\phi}\Psi(\{\psi'_i\}) \,. \tag{2.57}$$

Now, we insert eq. (2.56) into the first version of the Hartree-Fock equations in (2.52):

$$\hat{F}[U\psi'] \left(\sum_j U_{kj}\psi'_j(x) \right) = \sum_{i,j} \eta_{ki} U_{ij} \psi'_j(x) \,, \qquad k = 1, \ldots, N \,. \tag{2.58}$$

Next, we multiply by U^{-1}_{lk} each one of the N expressions and sum in k:

$$\hat{F}[U\psi'] \left(\sum_{j,k} \underbrace{U^{-1}_{lk} U_{kj}}_{\delta_{lj}} \psi'_j(x) \right) = \sum_{i,j,k} \underbrace{U^{-1}_{lk} \eta_{ki} U_{ij}}_{\varepsilon_{lj} = \delta_{lj}\varepsilon_j} \psi'_j(x)$$

$$\implies \quad \hat{F}[U\psi']\,\psi'_l(x) = \varepsilon_l \psi'_l(x) \,, \qquad l = 1, \ldots, N \,. \tag{2.59}$$

Although this new version of the Hartree-Fock equations can be readily seen as a *pseudo-eigenvalue problem* and solved by the customary iterative methods, we can go a step further and show that, like the N-particle wavefunction Ψ (see eq. (2.57)), the Fock operator $\hat{F}[\psi]$, as a function of the one-electron orbitals, is invariant under a unitary transformation such as the one in eq. (2.56). In fact, this is true for each one of the sums of Coulomb and exchange operators in eq. (2.51) separately:

$$\sum_j \hat{J}_j[U\psi']\,\varphi(x) =$$

$$\sum_j \left(\int \frac{\left|\sum_k U_{jk}\,\psi'_k(x')\right|^2}{|\vec{r}-\vec{r}\,'|}\,\mathrm{d}x' \right)\varphi(x) = \sum_j \left(\int \frac{\sum_{k,l} \overbrace{U^*_{jk}}^{U^{-1}_{kj}} U_{jl}\,\psi'^*_k(x')\,\psi'_l(x')}{|\vec{r}-\vec{r}\,'|}\,\mathrm{d}x' \right)\varphi(x) =$$

$$\left(\int \frac{\sum_{j,k,l} U^{-1}_{kj} U_{jl}\,\psi'^*_k(x')\,\psi'_l(x')}{|\vec{r}-\vec{r}\,'|}\,\mathrm{d}x' \right)\varphi(x) =$$

$$\sum_k \left(\int \frac{|\psi'_k(x')|^2}{|\vec{r}-\vec{r}\,'|}\,\mathrm{d}x' \right)\varphi(x) = \sum_j \hat{J}_j[\psi']\,\varphi(x)\,, \qquad \forall\varphi(x)\,, \tag{2.60}$$

where, in the step before the last, we have summed on j and l, using that $\sum_j U^{-1}_{kj} U_{jl} = \delta_{kl}$. Performing very similar calculations, one can show that

$$\sum_j \hat{K}_j[U\psi']\,\varphi(x) = \sum_j \hat{K}_j[\psi']\,\varphi(x)\,, \qquad \forall\varphi(x)\,, \tag{2.61}$$

and therefore, that $F[U\psi'] = F[\psi']$. In such a way that any unitary transformation on a set of orbitals that constitute a solution of the Hartree-Fock equations in (2.52) yields a different set that is also a solution of *the same* equations. For computational and conceptual reasons (see, for example, Koopmans' theorem below), it turns out to be convenient to use this freedom and choose the matrix U in such a way that the Lagrange multipliers matrix is diagonalized (see eq. (2.55) and the paragraph below it). The particular set of one-electron orbitals $\{\psi'_i\}$ obtained with this U are called *canonical orbitals* and their use is so prevalent that we will circumscribe the forecoming discussion to them and drop the prime from the notation.

Using the canonical orbitals, the *Hartree-Fock equations* can be written as

$$\hat{F}[\psi]\,\psi_i(x) = \varepsilon_i \psi_i(x)\,, \qquad i = 1, \ldots, N\,. \tag{2.62}$$

Many of the remarks related to these equations are similar to those made about the Hartree ones in (2.29), although there exist important differences due to the inclusion of the indistinguishability of the electrons in the variational ansatz. This is clearly illustrated if we calculate the joint probability density associated to a wavefunction like the one in eq. (2.31) of the coordinates with label 1 taking the value x_1, the coordinates with label 2 taking the value x_2, and so on:

$$\rho(x_1, \ldots, x_N) = \left|\Psi(x_1, \ldots, x_N)\right|^2 =$$

$$\frac{1}{N!} \sum_{p,p' \in S_N} (-1)^{\mathcal{T}(p)+\mathcal{T}(p')} \psi^*_{p(1)}(x_1) \cdots \psi^*_{p(N)}(x_N)\,\psi_{p'(1)}(x_1) \cdots \psi_{p'(N)}(x_N)\,. \tag{2.63}$$

If we compare this expression with eq. (2.30), we see that the antisymmetry of Ψ has completely spoiled the statistical independence among the one-electron coordinates. However, there is a weaker quasi-independence that may be recovered: If, using the same

reasoning about permutations that took us to the one-electron part $\sum_i \langle\Psi|\,\hat{h}_i\,|\Psi\rangle$ of the energy functional in page 53, we calculate the marginal probability density of the i-th coordinates taking the value x_i, we find

$$\rho_i(x_i) := \int \left(\prod_{k\neq i} \mathrm{d}x_k\right) \rho(x_1, \ldots, x_N) = \frac{1}{N} \sum_j \left|\psi_j(x_i)\right|^2 . \tag{2.64}$$

Now, since the coordinates indices are just immaterial labels, the actual probability density of finding *any* electron with coordinates x is given by

$$\rho(x) := \sum_i \rho_i(x) = \sum_i \left|\psi_i(x)\right|^2 , \tag{2.65}$$

which can be interpreted as a *charge density* (except for the sign), as, in atomic units, the charge of the electron is $e = -1$. The picture being consistent with the fact that $\rho(x)$ is normalized to the number of electrons N:

$$\int \rho(x)\,\mathrm{d}x = N . \tag{2.66}$$

Additionally, if we perform the same type of calculations that allowed to calculate the two-electron part of the energy functional in page 54, we have that the two-body probability density of the i-th coordinates taking the value x_i and of the j-th coordinates taking the value x_j reads

$$\rho_{ij}(x_i, x_j) := \int \left(\prod_{k\neq i,j} \mathrm{d}x_k\right) \rho(x_1, \ldots, x_N) = \\ \frac{1}{N(N-1)} \left(\sum_k \left|\psi_k(x_i)\right|^2 \sum_l \left|\psi_l(x_j)\right|^2 - \sum_{k,l} \psi_k^*(x_i)\,\psi_l^*(x_j)\,\psi_l(x_i)\,\psi_k(x_j) \right) , \tag{2.67}$$

and, if we reason in the same way as in the case of $\rho_i(x_i)$, in order to get to the probability density of finding *any* electron with coordinates x at the same time that *any other* electron has coordinates x', we must multiply the function above by $N(N-1)/2$, which is the number of immaterial (i, j)-labelings, taking into account that the distinction between x and x' is also irrelevant:

$$\rho(x, x') := \frac{N(N-1)}{2}\,\rho_{ij}(x, x') = \\ \frac{1}{2} \left(\sum_k \left|\psi_k(x)\right|^2 \sum_l \left|\psi_l(x')\right|^2 - \sum_{k,l} \psi_k^*(x)\,\psi_l^*(x')\,\psi_l(x)\,\psi_k(x') \right) . \tag{2.68}$$

Finally, taking eq. (2.65) to this one, we have

$$\rho(x, x') = \frac{1}{2} \left(\rho(x)\,\rho(x') - \sum_{k,l} \psi_k^*(x)\,\psi_l^*(x')\,\psi_l(x)\,\psi_k(x') \right) , \tag{2.69}$$

where the first term corresponds to independent electrons and the second one could be interpreted as an exchange correction.

Although, in general, this is the furthest one may go, when additional constraints are imposed on the spin part of the one-electron wavefunctions (see the discussion about Restricted Hartree-Fock in the following pages, for example), the exchange correction in eq. (2.69) above vanishes for electrons of opposite spin, i.e., electrons of opposite spin turn out to be pairwise independent. However, whereas it is true that more correlation could be added to the Hartree-Fock results by going to higher levels of the theory (see, for example, sec. 2.10) and, in this sense, Hartree-Fock could be considered the first step in the 'correlation ladder', one should not regard it as an 'uncorrelated' approximation, since, even in the simplest case of RHF (see below), Hartree-Fock electrons (of the same spin) are statistically correlated.

Let us now point out that, like in the Hartree case, the left-hand side of the Hartree-Fock equations in (2.62) is a complicated, non-linear function of the orbitals $\{\psi_i\}$ and the notation chosen is intended only to emphasize the nature of the iterative protocol that is typically used to solve the problem. However, note that, while the Hartree operator $\hat{\mathcal{H}}_k[\phi]$ depended on the index of the orbital ϕ_k on which it acted, the Fock operator in eq. (2.62) is the same for all the spin-orbitals ψ_i. This is due to the inclusion of the $i = j$ terms in the sum of the Coulomb and exchange two-electron integrals in eq. (2.46) and it allows to perform the iterative procedure solving only one eigenvalue problem at each step, instead of N of them like in the Hartree case.

The one-particle appearance of eqs. (2.62) is again strong and, whereas the 'eigenvalues' ε_i are not the energies associated to individual orbitals (or electrons) as it may seem, they have some physical meaning via the well-known *Koopmans' theorem* [205].

To introduce it, let us multiply eq. (2.62) from the left by $\psi_i(x)$, for a given i, and then integrate over x. Using the definition of the Fock operator in eq. (2.51) together with the Coulomb and exchange ones in eqs. (2.50), we obtain

$$\langle\psi_i|\,\hat{F}\,|\psi_i\rangle = h_i + \sum_j \Big(J_{ij} - K_{ij}\Big) = \varepsilon_i\,, \qquad i = 1, \ldots, N\,, \tag{2.70}$$

where we have used the same notation as in eq. (2.46) and the fact that the one-electron orbitals are normalized.

If we next sum on i and compare the result with the expression in eq. (2.46), we found that the relation of the eigenvalues ε_i with the actual Hartree-Fock energy is given by

$$E = \sum_i \varepsilon_i - \frac{1}{2}\sum_{i,j}\Big(J_{ij} - K_{ij}\Big)\,. \tag{2.71}$$

Finally, if we assume that upon 'removal of an electron from the k-th orbital' the rest of the orbitals will remain unmodified, we can calculate the *ionization energy* using the expression in (2.46) together with the equations above:

$$\begin{aligned}\Delta E := E_{N-1} - E_N = \sum_{i\neq k} h_i - \sum_i h_i + \frac{1}{2}\sum_{i,j\neq k}\Big(J_{ij} - K_{ij}\Big) - \frac{1}{2}\sum_{i,j}\Big(J_{ij} - K_{ij}\Big) = \\ -h_k - \sum_j \Big(J_{kj} - K_{kj}\Big) = -\varepsilon_k\,,\end{aligned} \tag{2.72}$$

and this is Koopmans' theorem, namely, that *the k-th ionization energy in the frozen-orbitals approximation is* ε_k (of course, the ambiguity in the sign depends on whether we define the ionization energy as the one lost by the system or as the one that we must provide from the outside to make the process happen).

Turning now to the issue about the solution of the Hartree-Fock equations in (2.62), we must remark that the necessity of using the relatively unreliable iterative approach to tackle them stems again from their complicated mathematical form. Like in the Hartree case, we have managed to largely reduce the dimension of the space on which the basic equations are defined: from $3N$ in the electronic Schrödinger equation in (2.9a) to 3 in the Hartree-Fock ones. However, to have this, we have payed the price of dramatically increasing their complexity [200], since, while the electronic Schrödinger equation was one linear differential equation, the Hartree-Fock ones in (2.62) are N coupled non-linear integro-differential equations, thus precluding any analytical approach to their solution.

A typical iterative procedure[56] begins by proposing a *starting guess* for the set of N spin-orbitals $\{\psi_i^0\}$. With them, the Fock operator $\hat{F}[\psi^0]$ in the left-hand side of eq. (2.62) is constructed and the set of N equations is solved as one simple eigenvalue problem. Then, the $\{\psi_i^1\}$ that correspond to the N lowest eigenvalues ε_i^1 are selected (see the discussion of the aufbau principle below) and a new Fock operator $\hat{F}[\psi^1]$ is constructed with them. The process is iterated until (hopefully) the n-th set of solutions $\{\psi_i^n\}$ differs from the $(n-1)$-th one $\{\psi_i^{n-1}\}$ less than a reasonably small amount (defining the distance among solutions in some suitable way typically combined with a convergence criterium related to the associated energy change). When this occurs, the procedure is said to have converged and the solution orbitals are called *self-consistent*; also, a calculation of this kind is commonly termed *self-consistent field* (SCF).

Again, like in the Hartree case, many issues exist that raise doubts about the possible success of such an approach. The most important ones are related to the fact that a proper definition of the *Hartree-Fock problem* should be: *find the global minimum of the energy functional* $\langle\Psi|\hat{H}|\Psi\rangle$ *under the constraint that the wavefunction* Ψ *be a Slater determinant of one-electron spin-orbitals*, and not: *solve the Hartree-Fock equations* (2.62). The solutions of the latter are all the stationary points of the constrained energy functional, while we are interested only in the particular one that is the global minimum. Even ruling out the possibility that a found solution may be a maximum or a saddle point (which can be done [197, 200]), one can never be sure that it is the global minimum and not a local one.

There exists, however, one way, related to the Hartree-Fock version of the theorem by Simon and Lieb [202, 203] mentioned in the previous section, of hopefully biasing a particular found solution of eqs. (2.62) to be the global minimum that we are looking for. They showed, first, that for neutral or positively charged molecules ($Z \geq N$), the Hartree-Fock global minimization problem has a solution (its uniqueness is not established yet [200]) and, second, that the minimizing orbitals $\{\psi_i\}$ correspond to the N lowest eigenvalues of the Fock operator $\hat{F}[\psi]$ that is self-consistently constructed with them. Therefore, although *the reverse is not true in general* [200] (i.e., from the fact that a particular set of orbitals are the eigenstates corresponding to the lowest eigenvalues of the associated

[56] The process described in this paragraph must be taken only as an outline of the one that is performed in practice. It is impossible to deal in a computer with a general function as it is (a non-countable infinite set of numbers), and the problem must be discretized in some way. The truncation of the one-electron Hilbert space using a finite basis set, described in secs. 2.8 and 2.9, is the most common way of doing this.

Fock operator, does not necessarily follow that they are the optimal ones), the information contained in Simon and Lieb's result is typically invoked to build each successive state in the iterative procedure described above by keeping only the N orbitals that correspond to the N lowest eigenvalues ε_i. Indeed, by doing that, one is effectively constraining the solutions to have a property that the true solution does have, so that, in the worst case, the space in which one is searching is of the same size as the original one, and, in the best case (even playing with the possibility that the reverse of Simon and Lieb's theorem be true, though not proved), the space of solutions is reduced to the correct global minimum alone. This wishful-thinking way of proceeding is termed the *aufbau principle* [200], and, together with a clever choice of the starting-guess set of orbitals [206] (typically extracted from a slightly less accurate theory, so that one may expect that it could be 'in the basin of attraction' of the true Hartree-Fock minimum), constitute one of the many heuristic strategies that make possible that the aforementioned drawbacks (and also those related to the convergence of iterative procedures) be circumvented in real cases, so that, in practice, most of SCF calculations performed in the field of quantum chemistry do converge to the true solution of eqs. (2.62) in spite of the theoretical notes of caution.

Now, to close this section, let us discuss some points regarding the imposition of constraints as a justification for subsequently introducing two commonly used forms of the Hartree-Fock theory that involve additional restrictions on the variational ansatz (apart from the ones in eqs. (2.31) and (2.33)).

In principle, the target systems in which we are interested and to which the theory developed in this chapter will be applied are rather complex. They have many degrees of freedom and the different interactions that drive their behaviour typically compete with one another, thus producing complicated, 'frustrated' energy landscapes (see chapter 1, but note, however, that we do not need to think about macromolecules; a small molecule like CO_2 already has 22 electrons). This state of affairs renders the a priori assessment of the accuracy of any approximation to the exact equations an impossible task. As researchers calculate more and more properties of molecular species using quantum chemistry and the results are compared to higher-level theories or to experimental data, much empirical knowledge about 'how good is theory A for calculating property X' is being gathered. However, if the characterization of a completely new molecule that is not closely related to any one that has been previously studied is tackled with, say, the Hartree-Fock approximation, it would be very unwise not to 'ask for a second opinion'.

All of this also applies, word by word, to the choice of the constraints on the wavefunction in variational approaches like the one discussed in this section: For example, it is impossible to know a priori what will be the loss of accuracy due to the requirement that the N-particle wavefunction Ψ be a Slater determinant as in eq. (2.31). However, in the context of the Hartree-Fock approximation, there exists a way of proceeding, again, partly based on wishful thinking and partly confirmed by actual calculations in particular cases, that is almost unanimously used to choose additional constraints which are expected to yield more *efficient* theories. It consists of imposing constraints to the variational wavefunction that are *properties that the exact solution to the problem does have*. In such a way that the obvious loss of accuracy due to the reduction of the search space is expected to be minimized, while the decrease in computational cost could be considerable.

This way of thinking is clearly illustrated by the question of whether or not one should allow that the one-electron spin-orbitals ψ_i (and therefore the total wavefunction Ψ) be

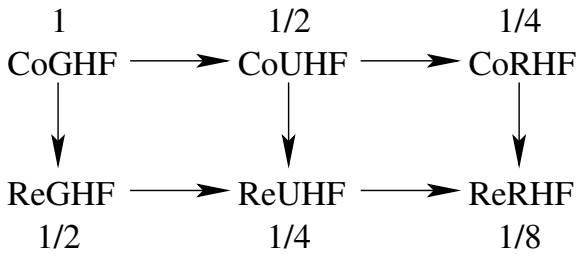


Figure 2.1: Schematic relation map among the six types of Hartree-Fock methods discussed in the text: *General Hartree-Fock* (GHF), *Unrestricted Hartree-Fock* (UHF) and *Restricted Hartree-Fock* (RHF), in both their *complex* (Co) and *real* (Re) versions. The arrows indicate imposition of constraints; horizontally, in the spin part of the orbitals, and, vertically, from complex- to real-valued wavefunctions. Next to each method, the size of the search space relative to that in CoGHF is shown.

complex valued. Indeed, due to the fact that the electronic Hamiltonian in eq. (2.20) is linear, the real and imaginary parts of any complex eigenfunction solution of the time independent Schrödinger equation in (2.9a) are also solutions of it [200]. Therefore, the fundamental state, which is the exact solution of the problem that we are trying to solve, may be chosen to be real valued. Nevertheless, the exact minimum will not be achieved, in general, in the smaller space defined by the Hartree-Fock constraints in eqs. (2.31) and (2.33), so that there is no a priori reason to believe that allowing the Hartree-Fock wavefunction to take complex values would not improve the results by finding a lower minimum. In fact, in some cases, this happens [206]. Nevertheless, if one constrains the search to real orbitals, the computational cost is reduced by a factor two, and, after all, 'some constraints must be imposed'.

Additionally, apart from these 'complex vs. real' considerations, there exist two further restrictions that are commonly found in the literature and that affect the spin part of the one-electron orbitals ψ_i. The N-electron wavefunction Ψ of the *General Hartree-Fock* (GHF) approximation (which is the one discussed up to now) is not an eigenstate of the total-spin operator, $\hat{S}^2$, nor of the z-component of it, $\hat{S}_z$ [206]. However, since both of them commute with the electronic Hamiltonian in eq. (2.20), the true fundamental state of the exact problem can be chosen to be an eigenstate of both operators simultaneously. So two additional constraints on the spin part of the GHF wavefunction in (2.31) are typically made that force the variational ansatz to satisfy these fundamental-state properties and that should be seen in the light of the above discussion, i.e., as reducing the search space, thus yielding an intrinsically less accurate theory, but also as being good candidates to hope that the computational savings will pay for this.

The first approximation to GHF (in a logical sense) is called *Unrestricted Hartree-Fock* (UHF) and it consists of asking the orbitals ψ_i to be a product of a part $\phi_i(\vec{r})$ depending on the positions $\vec{r}$ times a spin eigenstate of the one-electron $\hat{s}_z$ operator, i.e., either $\alpha(\sigma)$ or $\beta(\sigma)$ (see eq. (2.35)). This is denoted by $\psi_i(x) := \phi_i(\vec{r})\gamma_i(\sigma)$, where γ_i is either the α or the β function. Now, if we call N_α and N_β the number of spin-orbitals of each type, we have that, differently from the GHF one, the UHF N-particle wavefunction Ψ is an eigenstate of the $\hat{S}_z$ operator with eigenvalue $(1/2)(N_\alpha - N_\beta)$ (in atomic units, see sec. 2.2). However, it is not an eigenstate of $\hat{S}^2$ (the UHF wavefunction can be projected

into pure $\hat{S}^2$-states, however, the result is multideterminantal [206] and will not be considered here). Regarding the computational cost of the UHF approximation, it is certainly lower than that of GHF, since the search space is half as large: In the latter case, we had to consider $2N$ (complex or real) functions of $\mathbb{R}^3$ (the $\phi_i^\alpha(\vec{r})$ and the $\phi_i^\beta(\vec{r})$, see eq. (2.34)), while in UHF we only have to deal with N of them: the $\phi_i(\vec{r})$.

Now, we may perform a derivation analogous to the one performed for the GHF case, using the UHF orbitals, and get to the *Unrestricted version of the Hartree-Fock equations*[57]:

$$\hat{F}_i^{\mathrm{UHF}}[\phi]\,\phi_i(\vec{r}) := \left(\hat{h} + \sum_j^N \hat{J}_j[\phi] - \sum_j^N \delta_{\gamma_i\gamma_j}\hat{K}_j[\phi]\right)\phi_i(\vec{r}) = \varepsilon_i\phi_i(\vec{r})\,, \quad i = 1,\ldots,N\,, \tag{2.73}$$

where the Coulomb and exchange operators dependent on the spatial orbitals ϕ_i are defined by their action on an arbitrary function $\varphi(\vec{r})$ as follows:

$$\hat{J}_j[\phi]\,\varphi(\vec{r}) := \left(\int \frac{|\phi_j(\vec{r}\,')|^2}{|\vec{r}-\vec{r}\,'|}\,\mathrm{d}\vec{r}\,'\right)\varphi(\vec{r})\,, \tag{2.74a}$$

$$\hat{K}_j[\phi]\,\varphi(\vec{r}) := \left(\int \frac{\phi_j^*(\vec{r}\,')\,\varphi(\vec{r}\,')}{|\vec{r}-\vec{r}\,'|}\,\mathrm{d}\vec{r}\,'\right)\phi_j(\vec{r})\,, \tag{2.74b}$$

and one must note that, differently from the GHF case, due to the fact that the exchange interaction only takes place between orbitals 'of the same spin', the *UHF Fock operator* $\hat{F}_i^{\mathrm{UHF}}[\phi]$ depends on the index i.

If we now introduce the special form of the UHF orbitals into the general expression in (2.68), we can calculate the two-body probability density of finding *any* electron with coordinates x at the same time that *any other* electron has coordinates x':

$$\begin{aligned}\rho^{\mathrm{UHF}}(x,x') \;=\; \frac{1}{2}\Bigg(&\sum_{k,l}\left|\phi_k(\vec{r})\right|^2\left|\phi_l(\vec{r}\,')\right|^2\gamma_k(\sigma)\,\gamma_l(\sigma') \\ &- \sum_{k,l}\phi_k^*(\vec{r})\,\phi_l^*(\vec{r}\,')\,\phi_l(\vec{r})\,\phi_k(\vec{r}\,')\,\gamma_k(\sigma)\,\gamma_l(\sigma')\,\gamma_k(\sigma')\,\gamma_l(\sigma)\Bigg)\,.\end{aligned} \tag{2.75}$$

If we compute this probability density for 'electrons of the same spin', i.e., for $\sigma = \sigma'$, we obtain[58]

[57] Placing *functions* as arguments of the Kronecker's delta $\delta_{\gamma_i\gamma_j}$ is a bit unorthodox mathematically, but it constitutes an intuitive (and common) notation.

[58] Placing *a function and a coordinate* as arguments of the Kronecker's delta $\delta_{\gamma_k\sigma}$ is even more unorthodox mathematically than placing two functions (in fact, $\delta_{\gamma_k\sigma}$ is exactly the same as $\gamma_k(\sigma)$), however, the intuitive character of the notation compensates again for this.

$$\rho^{\mathrm{UHF}}(\vec{r},\vec{r}\,';\sigma=\sigma') = \frac{1}{2}\sum_{k,l\in I_\sigma}\left(\left|\phi_k(\vec{r})\right|^2\left|\phi_l(\vec{r}\,')\right|^2 - \phi_k^*(\vec{r})\,\phi_l^*(\vec{r}\,')\,\phi_l(\vec{r})\,\phi_k(\vec{r}\,')\right) = \frac{1}{2}\left(\rho(\vec{r},\sigma)\,\rho(\vec{r}\,',\sigma) - \sum_{k,l}\delta_{\gamma_k\sigma}\delta_{\gamma_l\sigma}\phi_k^*(\vec{r})\,\phi_l^*(\vec{r}\,')\,\phi_l(\vec{r})\,\phi_k(\vec{r}\,')\right), \tag{2.76}$$

where the following expression for the one-electron charge density has been used:

$$\rho^{\mathrm{UHF}}(\vec{r},\sigma) = \sum_{k\in I_\sigma}\left|\phi_k(\vec{r})\right|^2 . \tag{2.77}$$

At this point, note that eq. (2.76) contains, like in the GHF case, the exchange correction to the first (independent electrons) term. Nevertheless, as we advanced, if we calculate the two-body ρ^{UHF} for 'electrons of opposite spin', i.e., for $\sigma \neq \sigma'$, we have that

$$\rho^{\mathrm{UHF}}(\vec{r},\vec{r}\,';\sigma\neq\sigma') = \frac{1}{2}\,\rho(\vec{r},\sigma)\,\rho(\vec{r}\,',\sigma') , \tag{2.78}$$

i.e., that *UHF electrons of opposite spin are statistically pairwise independent*.

The other common approximation to GHF is more restrictive than UHF and is called *Restricted Hartree-Fock* (RHF). Apart from asking the orbitals ψ_i to be a product of a spatial part times a spin eigenstate of the one-electron $\hat{s}_z$ operator like in the UHF case, in RHF, the number of 'spin-up' and 'spin-down' orbitals is the same, $N_\alpha = N_\beta$ (note that this means that RHF may only be used with molecules containing an even number of electrons), and each spatial wavefunction occurs twice: once multiplied by $\alpha(\sigma)$ and the other time by $\beta(\sigma)$. This is typically referred to as a *closed-shell* situation and we shall denote it by writing $\psi_i(x) := \phi_i(\vec{r})\,\alpha(\sigma)$ if $i \leq N/2$, and $\psi_i(x) := \phi_{i-N/2}(\vec{r})\,\beta(\sigma)$ if $i > N/2$; in such a way that there are $N/2$ different spatial orbitals denoted by $\phi_I(\vec{r})$, with, $I = 1,\ldots,N/2$. Due to these additional restrictions, we have that, differently from the GHF one, the RHF N-particle wavefunction Ψ is an eigenstate of both the $\hat{S}^2$ and the $\hat{S}_z$ operators, with zero eigenvalue in both cases [180, 206], just like the fundamental state of the exact problem. Regarding the computational cost of the RHF approximation, it is even lower than that of UHF, since the size of the search space has been reduced to one quarter that of GHF: In the latter case, we had to consider $2N$ (complex or real) functions of $\mathbb{R}^3$ (the $\phi_i^\alpha(\vec{r})$ and the $\phi_i^\beta(\vec{r})$, see eq. (2.34)), while in RHF we only have to deal with $N/2$ of them: the $\phi_I(\vec{r})$ (see fig. 2.1).

Again, we may perform a derivation analogous to the one performed for the GHF case and get to the *Restricted version of the Hartree-Fock equations*:

$$\hat{F}^{\mathrm{RHF}}[\phi]\,\phi_I(\vec{r}) := \left[\hat{h} + \sum_J^{N/2}\left(2\hat{J}_J[\phi] - \hat{K}_J[\phi]\right)\right]\phi_I(\vec{r}) = \varepsilon_I\phi_I(\vec{r}) , \quad I = 1,\ldots,N/2 , \tag{2.79}$$

where the *RHF Fock operator* $\hat{F}^{\mathrm{RHF}}[\phi]$ has been defined in terms of the Coulomb and exchange operators in eq. (2.74) and, differently from the UHF case, it does not depend on the index I of the orbital on which it operates.

If we follow the same steps as in the GHF case in page 60, we can relate the *RHF energy* to the eigenvalues ε_I and the two-electron spatial integrals:

$$E = 2\sum_{I}^{N/2}\varepsilon_I - \sum_{I,J}^{N/2}\left(2\langle\phi_I\phi_J|\,\frac{1}{r}\,|\phi_I\phi_J\rangle - \langle\phi_I\phi_J|\,\frac{1}{r}\,|\phi_J\phi_I\rangle\right). \tag{2.80}$$

Now, if we introduce the special form of the RHF orbitals into the general expression in (2.68), we can calculate the RHF two-body probability density of finding *any* electron with coordinates x at the same time that *any other* electron has coordinates x':

$$\begin{aligned}\rho^{\mathrm{RHF}}(x,x') &= \frac{1}{2}\left(\sum_{K,L}^{N/2}\left|\phi_K(\vec{r})\right|^2\left|\phi_L(\vec{r}\,')\right|^2 - \delta_{\sigma\sigma'}\sum_{K,L}^{N/2}\phi_K^*(\vec{r})\,\phi_L^*(\vec{r}\,')\,\phi_L(\vec{r})\,\phi_K(\vec{r}\,')\right) = \\ &\frac{1}{2}\left(\rho(\vec{r},\sigma)\,\rho(\vec{r}\,',\sigma') - \delta_{\sigma\sigma'}\sum_{K,L}^{N/2}\phi_K^*(\vec{r})\,\phi_L^*(\vec{r}\,')\,\phi_L(\vec{r})\,\phi_K(\vec{r}\,')\right).\end{aligned} \tag{2.81}$$

where the following expression for the RHF one-electron charge density has been used:

$$\rho^{\mathrm{RHF}}(\vec{r},\sigma) = \sum_{K}^{N/2}\left|\phi_K(\vec{r})\right|^2. \tag{2.82}$$

We notice that the situation is the same as in the UHF case: For RHF electrons with equal spin, there exists an exchange term in $\rho^{\mathrm{RHF}}(x,x')$ that corrects the 'independent' part, whereas *RHF electrons of opposite spin are statistically pairwise independent*.

Finally, let us stress that the RHF approximation (in its real-valued version) is the most used one in the literature [163, 207–215] (for molecules with an even number of electrons) and that all the issues related to the existence of solutions, as well as those regarding the iterative methods used to solve the equations, are essentially the same for RHF as for GHF (see the discussion in the previous pages).

In this dissertation, the only molecule treated with ab initio quantum chemistry is the model dipeptide HCO-L-Ala-NH_2 (see appendix E), which has an even number of electrons (62), so that real-RHF is the chosen form of the Hartree-Fock method to perform the calculations.

2.8 The Roothaan-Hall equations

Let's discretize

The Hartree-Fock equations in the RHF form in expression (2.79) are a set of $N/2$ of coupled integro-differential equations. As such, they can be tackled by finite-differences methods and solved on a discrete grid; this is known as *numerical Hartree-Fock* [216], and, given the present power of computers, it is only applicable to very small molecules.

In order to deal with larger systems, such as biological macromolecules, independently proposed by Roothaan [217] and Hall [218] in 1951, a different kind of discretization must be performed, not in $\mathbb{R}^3$ but in the Hilbert space $\mathcal{H}$ of the one-electron orbitals. Hence,

although the actual dimension of $\mathcal{H}$ is infinite, we shall approximate any function in it by a finite linear combination of M different functions χ_a[59]. In particular, the one-electron orbitals that make up the RHF wavefunction, shall be approximated by

$$\phi_I(\vec{r}) \simeq \sum_a^M c_{aI}\,\chi_a(\vec{r})\,, \qquad I = 1,\ldots,N/2\,, \quad M \geq \frac{N}{2}\,. \tag{2.83}$$

In both cases, numerical Hartree-Fock and discretization of the function space, the correct result can be only be reached asymptotically; when the grid is very fine, for the former, and when $M \to \infty$, for the latter. This exact result, which, in the case of small systems, can be calculated up to several significant digits, is known as the *Hartree-Fock limit* [219].

In practical cases, however, M is finite (often, only about an order of magnitude larger than $N/2$) and the set $\{\chi_a\}_{a=1}^M$ in the expression above is called the *basis set*. We shall devote the next section to discuss its special characteristics, but, for now, it suffices to say that, in typical applications, the functions χ_a are *atom-centered*, i.e., each one of them has non-negligible value only in the vicinity of a particular nucleus. Therefore, like all the electronic wavefunctions we have dealt with in the last sections, they parametrically depend on the positions $\underline{\mathrm{R}}$ of the nuclei (see sec. 2.3). This is why, sometimes, the functions χ_a are called *atomic orbitals*[60] (AO) (since they are localized at individual atoms), the ϕ_I are referred to as *molecular orbitals* (MO) (since they typically have non-negligible value in the whole space occupied by the molecule), and the approximation in eq. (2.83) is called *linear combination of atomic orbitals* (LCAO). In addition, since we voluntarily circumscribe to real-RHF, we assume that both the coefficients c_{aI} and the functions χ_a in the above expression are real.

Now, if we introduce the linear combination in eq. (2.83) into the Hartree-Fock equations in (2.79), multiply the result by χ_b (for a general value of b) and integrate on $\vec{r}$, we obtain

$$\sum_a F_{ba}c_{aI} = \varepsilon_I \sum_a S_{ba}c_{aI}\,, \qquad I = 1,\ldots,N/2\,,\ b = 1,\ldots,M\,, \tag{2.84}$$

where we denote by F_{ba} the (b,a)-element of the *Fock matrix*[61], and by S_{ba} the one of the *overlap matrix*, defined as

$$F_{ba} := \langle\chi_b|\,\hat{F}[\phi]\,|\chi_a\rangle \qquad \text{and} \qquad S_{ba} := \langle\chi_b|\chi_a\rangle\,, \tag{2.85}$$

respectively.

[59] In all this section and the forecoming ones, the indices belonging to the first letters of the alphabet, a, b, c, d, etc., run from 1 to M (the number of functions in the finite basis set); whereas those named with capital letters from I towards the end of the alphabet, I, J, K, L, etc., run from 1 to $N/2$ (the number of spatial wavefunctions ϕ_I, also termed the number of *occupied orbitals*).

[60] Some authors [177] suggest that, being strict, the term *atomic orbitals* should be reserved for the one-electron wavefunctions ϕ_I that are the solution of the Hartree-Fock problem (or even to the exact Schrödinger equation of the isolated atom), and that the elements χ_a in the basis set should be termed simply *localized functions*. However, it is very common in the literature not to follow this recommendation and choose the designation that appear in the text [177, 217, 220]. We shall do the same for simplicity.

[61] Note that the RHF superindex has been dropped from F.

Note that we do not ask the χ_a in the basis set to be mutually orthogonal, so that the overlap matrix is not diagonal in general.

Next, if we define the $M \times M$ matrices $F[c] := (F_{ab})$ and $S := (S_{ab})$, together with the (column) M-vector $c_I := (c_{aI})$, we can write eq. (2.84) in matricial form:

$$F[c]c_I = \varepsilon_I S c_I \, . \tag{2.86}$$

Hence, using the LCAO approximation, we have traded the $N/2$ coupled integro-differential Hartree-Fock equations in (2.79) for this system of $N/2$ algebraic equations for the $N/2$ orbital energies ε_I and the $M \cdot N/2$ coefficients c_{aI}, which are called *Roothaan-Hall equations* [217, 218] and which are manageable in a computer.

Now, if we forget for a moment that the Fock matrix depends on the coefficients c_{aI} (as stressed by the notation $F[c]$) and also that we are only looking for $N/2$ vectors c_I while the matrices F and S are $M \times M$, we may regard the above expression as a M-dimensional *generalized eigenvalue problem*. Many properties are shared between this kind of problem and a classical eigenvalue problem (i.e., one in which $S_{ab} = \delta_{ab}$) [217], being the most important one that, due to the hermiticity of $F[c]$, one can find an orthonormal set of M vectors c_a corresponding to real eigenvalues ε_a (where, of course, some eigenvalue could be repeated).

In fact, it is using this formalism how actual Hartree-Fock computations are performed, the general outline of the iterative procedure being essentially the same as the one discussed in sec. 2.7: Choose a *starting-guess* for the coefficients c_{aI} (let us denote it by c^0_{aI}), construct the corresponding Fock matrix $F[c^0]$[62] and solve the generalized eigenvalue problem in eq. (2.86). By virtue of the aufbau principle discussed in the previous section, from the M eigenvectors c_a, keep the $N/2$ ones c^1_I that correspond to the $N/2$ lowest eigenvalues ε^1_I, construct the new Fock matrix $F[c^1]$ and iterate (by convention, the eigenvalues ε^n_a, for all n, are ordered from the lowest to the largest as a runs from 1 to M). This procedure ends when the n-th solution is close enough (in a suitable defined way) to the $(n-1)$-th one. Also, note that, after convergence has been achieved, we end up with M orthogonal vectors c_a. Only the $N/2$ ones that correspond to the lowest eigenvalues represent real one-electron solutions and they are called *occupied orbitals*; the $M - N/2$ remaining ones do not enter in the N-electron wavefunction (although they are relevant for calculating corrections to the Hartree-Fock results, see sec. 2.10) and they are called *virtual orbitals*.

Regarding the mathematical foundations of this procedure, let us stress, however, that, whereas in the finite-dimensional GHF and UHF cases it has been proved that the analogous of Lieb and Simon's theorem (see the previous section) is satisfied, i.e., that the global minimum of the original optimization problem corresponds to the lowest eigenvalues of the self-consistent Fock operator, in the RHF case, contrarily, no proof seems to exist [200]. Of course, in practical applications, the positive result is assumed to hold.

Finally, if we expand F_{ab} in eq. (2.85), using the shorthand $|a\rangle$ for $|\chi_a\rangle$, we have

$$F_{ab} = \langle a|\hat{h}|b\rangle + \sum_{c,d} \underbrace{\left(\sum_J c_{cJ} c_{dJ}\right)}_{D_{cd}[c]} \underbrace{\left(2\langle ac|\frac{1}{r}|bd\rangle - \langle ac|\frac{1}{r}|db\rangle\right)}_{G^{cd}_{ab}} , \tag{2.87}$$

[62] Note (in eq. (2.87), for example) that the Fock matrix only depends on the vectors c_a with $a \leq N/2$.

where we have introduced the *density matrix* $D_{cd}[c]$, and also the matrix G^{cd}_{ab}, made up by the four-center integrals $\langle ac|\, 1/r\, |bd\rangle$ defined by

$$\langle ac|\, \frac{1}{r}\, |bd\rangle := \iint \frac{\chi_a(\vec{r})\chi_b(\vec{r}\,')\chi_c(\vec{r})\chi_d(\vec{r}\,')}{|\vec{r}-\vec{r}\,'|}\, \mathrm{d}\vec{r}\, \mathrm{d}\vec{r}\,' \ . \tag{2.88}$$

After convergence has been achieved, the *RHF energy* in the finite-dimensional case can be computed using the discretized version of eq. (2.80):

$$\begin{aligned} E &= 2\sum_{I}^{N/2} \varepsilon_I - \sum_{I,J}^{N/2} \sum_{a,b,c,d} \left(2c_{aI}c_{bJ}c_{cI}c_{dJ}\langle ab|\, \frac{1}{r}\, |cd\rangle - c_{aI}c_{bJ}c_{cJ}c_{dI}\langle ab|\, \frac{1}{r}\, |dc\rangle \right) = \\ &2\sum_{I}^{N/2} \varepsilon_I - \sum_{I,J}^{N/2} \sum_{a,b,c,d} c_{aI}c_{bJ}c_{cI}c_{dJ}\langle ab|\, \frac{1}{r}\, |cd\rangle = \\ &2\sum_{i}^{N/2} \varepsilon_i - \sum_{a,b,c,d} D_{ac}[c]D_{bd}[c]\langle ab|\, \frac{1}{r}\, |cd\rangle \ , \end{aligned} \tag{2.89}$$

where a convenient rearrangement of the indices in the two sums has been performed from the first to the second line.

2.9 Introduction to Gaussian basis sets

Flora and fauna

In principle, arbitrary functions may be chosen as the χ_a to solve the Roothaan-Hall equations in the previous section, however, in eq. (2.87), we see that one of the main numerical bottlenecks in SCF calculations arises from the necessity of calculating the $\sim M^4$ four-center integrals $\langle ab|\, \frac{1}{r}\, |cd\rangle$ (since the solution of the generalized eigenvalue problem in eq. (2.86) typically scales only like M^3, and there are $\sim M^2$ two-center $\langle a|\, \hat{h}\, |b\rangle$ integrals).

Either if these integrals are calculated at each iterative step and directly taken from RAM memory (*direct SCF*) or if they are calculated at the first step, written to disk, and then read from there when needed (*conventional SCF*), an appropriate choice of the functions χ_a in the finite basis set is essential if accurate results are sought, M is intended to be kept as small as possible and the integrals are wanted to be computed rapidly. When one moves into higher-level theoretical descriptions (such as the MP2 method discussed in the next section) and the numerical complexity scales with M even more unpleasantly, the importance of this choice greatly increases.

In chapter 7 of this dissertation, a large number of basis sets (belonging to a particular 'family') are compared in ab initio calculations of the PES of the model dipeptide HCO-L-Ala-NH_2 (see appendix E). In this section, in order to support that study, we shall introduce some of the concepts involved in the interesting field of basis-set design. For further details not covered here, the reader may want to check refs. 177, 180, 221, 222.

The only analytically solvable molecular problem in non-relativistic quantum mechanics is the *hydrogen-like atom*, i.e., the system formed by a nucleus of charge Z and only one electron (H, He^+, Li^{2+}, etc.). Therefore, it is not strange that all the thinking about

atomic-centered basis sets in quantum chemistry is much influenced by the particular solution to this problem.

The spatial eigenfunctions of the Hamiltonian operator of an hydrogen-like atom, in atomic units and spherical coordinates, read[63]

$$\phi_{nlm}(r,\theta,\varphi) = \sqrt{\left(\frac{2Z}{n}\right)^3 \frac{(n-l-1)!}{2n[(n+l)!]^3}} \left(\frac{2Z}{n}r\right)^l L_{n-l-1}^{2l+1}\left(\frac{2Z}{n}r\right) e^{-Zr/n} Y_{lm}(\theta,\varphi)\,, \tag{2.90}$$

where n, l and m are the energy, total angular momentum and z-angular momentum quantum numbers, respectively. Their ranges of variation are coupled: all being integers, n runs from 1 to ∞, l from 0 to $n-1$ and m from $-l$ to l. The function L_{n-l-1}^{2l+1} is a *generalized Laguerre polynomial* [223], for which it suffices to say here that it is of order $n-l-1$ (thus having, in general, $n-l-1$ zeros), and the function $Y_{lm}(\theta,\varphi)$ is a *spherical harmonic*, which is a simultaneous eigenfunction of the total angular momentum operator $\hat{l}^2$ (with eigenvalue $l(l+1)$) and of its z-component $\hat{l}_z$ (with eigenvalue m).

The hope that the one-electron orbitals that are the solutions of the Hartree-Fock problem in many-electron atoms could not be very different from the ϕ_{nlm} above[64], together with the powerful chemical intuition that states that 'atoms-in-molecules are not very different from atoms-alone', is what mainly drives the choice of the functions χ_a in the basis set, and, in the end, the variational procedure that will be followed is expected to fix the largest failures coming from these too-simplistic assumptions.

Hence, it is customary to choose functions that are centered at atomic nuclei and that partially resemble the exact solutions for hydrogen-like atoms. In this spirit, the first type of AOs to be tried [220] were the *Slater-type orbitals* (STOs), proposed by Slater [224] and Zener [225] in 1930:

$$\chi_a^{\mathrm{STO}}(\vec{r};\vec{R}_{\alpha_a}) := \mathcal{N}_a^{\mathrm{STO}} \widetilde{Y}_{l_a m_a}^{c,s}(\theta_{\alpha_a},\varphi_{\alpha_a}) \left|\vec{r}-\vec{R}_{\alpha_a}\right|^{n_a-1} \exp\left(-\zeta_a \left|\vec{r}-\vec{R}_{\alpha_a}\right|\right), \tag{2.91}$$

where $\mathcal{N}_a^{\mathrm{STO}}$ is a normalization constant and ζ_a is an adjustable parameter. The index α_a is that of the nucleus at which the function is centered, and, of course, in the majority of cases, there will be several χ_a^{STO} corresponding to different values of a centered at the same nucleus. The integers l_a and m_a can be considered quantum numbers, since, due to the fact that the only angular dependence is in $\widetilde{Y}_{l_a m_a}^{c,s}$ (see below for a definition), the STO defined above is still a simultaneous eigenstate of the one-electron angular momentum operators $\hat{l}^2$ and $\hat{l}_z$ (with the origin placed at $\vec{R}_{\alpha_a}$). The parameter n_a, however, should be regarded as a 'principal (or energy) quantum number' only by analogy, since, on the one hand, it does not exist a 'one-atom Hamiltonian' whose exact eigenfunctions

[63] For consistency with the rest of the text, the Born-Oppenheimer approximation has been also assumed here. So that the *reduced mass* $\mu := m_e M_N/(m_e+M_N)$ that should enter the expression is considered to be the mass of the electron $\mu \simeq m_e$ (recall that, in atomic units, $m_e = 1$ and $M_N \gtrsim 2000$).

[64] Note that the N-electron wavefunction of the exact fundamental state of a non-hydrogen-like atom depends on $3N$ spatial variables in a way that cannot be written, in general, as a Slater determinant of one-electron functions. The image of single electrons occupying definite orbitals, together with the possibility of comparing them with the one-particle eigenfunctions of the Hamiltonian of hydrogen-like atoms, vanishes completely outside the Hartree-Fock formalism.

it could label and, on the other hand, only the leading term of the Laguerre polynomial in eq. (2.90) has been kept in the STO[65].

Additionally, in the above notation, the fact that χ_a^{STO} parametrically depends on the position of a certain α_a-th nucleus has been stressed, and the functions $\widetilde{Y}^{c,s}_{l_a m_a}$, which are called *real spherical harmonics* [226] (remember that we want to do real RHF), are defined in terms of the classical *spherical harmonics* $Y_{l_a m_a}$ by

$$\widetilde{Y}^{c}_{l_a m_a}(\theta_{\alpha_a},\varphi_{\alpha_a}) := \frac{Y_{l_a m_a}(\theta_{\alpha_a},\varphi_{\alpha_a}) + Y^*_{l_a m_a}(\theta_{\alpha_a},\varphi_{\alpha_a})}{\sqrt{2}} \propto P_{l_a}^{m_a}(\cos\theta_{\alpha_a})\cos(m_a\varphi_{\alpha_a})\,, \tag{2.92a}$$

$$\widetilde{Y}^{s}_{l_a m_a}(\theta_{\alpha_a},\varphi_{\alpha_a}) := -i\frac{Y_{l_a m_a}(\theta_{\alpha_a},\varphi_{\alpha_a}) - Y^*_{l_a m_a}(\theta_{\alpha_a},\varphi_{\alpha_a})}{\sqrt{2}} \propto P_{l_a}^{m_a}(\cos\theta_{\alpha_a})\sin(m_a\varphi_{\alpha_a})\,, \tag{2.92b}$$

where c stands for *cosine*, s for *sine*, the functions $P_{l_a}^{m_a}$ are the *associated Legendre polynomials* [223], and the spherical coordinates θ_{α_a} and φ_{α_a} also carry the α_a-label to remind that the origin of coordinates in terms of which they are defined is located at $\vec{R}_{\alpha_a}$. Also note that, using that $\widetilde{Y}^{c}_{l_a 0} = \widetilde{Y}^{s}_{l_a 0}$, there is the same number of real spherical harmonics as of classical ones.

These χ_a^{STO} have some good physical properties. Among them, we shall mention that, for $|\vec{r} - \vec{R}_{\alpha_a}| \to 0$, they present a *cusp* (a discontinuity in the radial derivative), as required by Kato's theorem [227]; and also that they decay at an exponential rate when $|\vec{r} - \vec{R}_{\alpha_a}| \to \infty$, which is consistent with the image that, an electron that is taken apart from the vicinity of the nucleus must 'see', at large distances, an unstructured point-like charge (see, for example, the STO in fig. 2.2). Finally, the fact that they do not present radial nodes (due to the aforementioned absence of the non-leading terms of the Laguerre polynomial in eq. (2.90)) can be solved by making linear combinations of functions with different values of ζ_a[66].

Now, despite their being good theoretical candidates to expand the MO ϕ_I that make up the N-particle solution of the Hartree-Fock problem, these STOs have serious computational drawbacks: Whereas the two-center integrals (such as $\langle a|\,\hat{h}\,|b\rangle$ in eq. (2.87)) can be calculated analytically, the four-center integrals $\langle ac|\,1/r\,|bd\rangle$ can not [177, 220] if functions like the ones in eq. (2.91) are used. This fact, which was known as "the nightmare of the integrals" in the first days of computational quantum chemistry [220], precludes the use of STOs in practical ab initio calculations of large molecules.

A major step to overcome these difficulties that has revolutioned the whole field of quantum chemistry [200, 220] was the introduction of *Cartesian Gaussian-type orbitals* (cGTO):

$$\chi_a^{\text{cGTO}}(\vec{r};\vec{R}_{\alpha_a}) := \mathcal{N}_a^{\text{cGTO}}\left(r^1 - R^1_{\alpha_a}\right)^{l_a^x}\left(r^2 - R^2_{\alpha_a}\right)^{l_a^y}\left(r^3 - R^3_{\alpha_a}\right)^{l_a^z}\exp\left(-\zeta_a\,|\vec{r} - \vec{R}_{\alpha_a}|^2\right), \tag{2.93}$$

[65] If we notice that, within the set of all possible STOs (as defined in eq. (2.91)), every hydrogen-like energy eigenfunction (see eq. (2.90)) can be formed as a linear combination, we easily see that the STOs constitute a complete basis set. This is important to ensure that the Hartree-Fock limit could be actually approached by increasing M.

[66] This way of proceeding renders the choice of the exponent carried by the $|\vec{r} - \vec{R}_{\alpha_a}|$ part ($n_a - 1$ in the case of the STO in eq. (2.91)) a rather arbitrary one. As a consequence, different definitions may be found in the literature and the particular exponent chosen in actual calculations turns out to be mostly a matter of computational convenience.

where the r^p and the $R^p_{\alpha_a}$, with $p = 1, 2, 3$, are the Euclidean coordinates of the electron and the α_a-th nucleus respectively, and the integers l^x_a, l^y_a and l^z_a, which take values from 0 to ∞, are called *orbital quantum numbers*[67].

Although these GTOs do not have the good physical properties of the STOs (compare, for example, the STO and the GTO in fig. 2.2), in 1950, Boys [228] showed that all the integrals appearing in SCF theory could be calculated analytically if the χ_a had the form in eq. (2.93). The enormous computational advantage that this entails makes possible to use a much larger number of functions to expand the one-electron orbitals ϕ_i if GTOs are used, partially overcoming their bad short- and long-range behaviour and making the Gaussian-type orbitals the universally preferred choice in SCF calculations [177].

To remedy the fact that the angular behaviour of the Cartesian GTOs in eq. (2.93) is somewhat hidden, they may be linearly combined to form *Spherical Gaussian-type orbitals* (sGTO):

$$\chi^{\text{sGTO}}_a(\vec{r}\,;\vec{R}_{\alpha_a}) := \mathcal{N}^{\text{sGTO}}_a \widetilde{Y}^{c,s}_{l_a m_a}(\theta_{\alpha_a}, \varphi_{\alpha_a}) \left|\vec{r} - \vec{R}_{\alpha_a}\right|^{l_a} \exp\left(- \zeta_a \left|\vec{r} - \vec{R}_{\alpha_a}\right|^2 \right), \tag{2.94}$$

which are proportional to the real spherical harmonic $\widetilde{Y}^{c,s}_{l_a m_a}(\theta_{\alpha_a}, \varphi_{\alpha_a})$, and to which the same remarks made in footnote 66 for the STOs, regarding the exponent in the $|\vec{r} - \vec{R}_{\alpha_a}|$ part, may be applied.

The fine mathematical details about the linear combination that relates the Cartesian GTOs to the spherical ones are beyond the scope of this introduction. We refer the reader to refs. 229 and 226 for further information and remark here some points that will have interest in the subsequent discussion.

First, the cGTOs that are combined to make up a sGTO must have all the same value of $l_a := l^x_a + l^y_a + l^z_a$ and, consequently, this sum of the three orbital quantum numbers l^x_a, l^y_a and l^z_a in a particular Cartesian GTO is typically (albeit dangerously) referred to as the *angular momentum* of the function. In addition, apart from the numerical value of l_a, the spectroscopic notation is commonly used in the literature, so that cGTOs with $l_a = 0, 1, 2, 3, 4, 5, \ldots$ are called $s, p, d, f, g, h, \ldots$, respectively. Where the first four come from the archaic words *sharp*, *principal*, *diffuse* and *fundamental*, while the subsequent ones proceed in alphabetical order.

Second, for a given $l_a > 1$, there are more Cartesian GTOs ($(l_a + 1)(l_a + 2)/2$) than spherical ones ($2l_a + 1$), in such a way that, from the $(l_a + 1)(l_a + 2)/2$ functionally independent linear combinations that can be formed using the cGTOs of angular momentum l_a, only the angular part of $2l_a + 1$ of them turns out to be proportional to a real spherical harmonic $\widetilde{Y}^{c,s}_{l_a m_a}(\theta_{\alpha_a}, \varphi_{\alpha_a})$; the rest of them are proportional to real spherical harmonic functions with a different value of the angular momentum quantum number. For example, from the six different d-Cartesian GTOs, whose polynomial parts are x^2, y^2, z^2, xy, xz and yz (using an evident, compact notation), only five different spherical GTOs can be constructed: the ones with polynomial parts proportional to $2z^2 - x^2 - y^2$, xz, yz, $x^2 - y^2$ and xy [229]. Among these new sGTOs, which, in turn, are proportional (neglecting also powers of r, see footnote 66) to the real spherical harmonics $\widetilde{Y}_{20}$, $\widetilde{Y}^c_{21}$, $\widetilde{Y}^s_{21}$, $\widetilde{Y}^c_{22}$ and $\widetilde{Y}^s_{22}$, the

[67] Since the harmonic-oscillator energy eigenfunctions can be constructed as linear combinations of Cartesian GTOs, we have that the latter constitute a complete basis set and, like in the case of the STOs, we may expect that the Hartree-Fock limit is approached as M is increased.

linear combination $x^2 + y^2 + z^2$ is missing, since it presents the angular behaviour of an s-orbital (proportional to $\widetilde{Y}_{00}$).

Finally, let us remark that, whereas Cartesian GTOs in eq. (2.93) are easier to be coded in computer applications than sGTOs[229], it is commonly accepted that these spurious spherical orbitals of lower angular momentum that appear when cGTOs are used do not constitute efficient choices to be included in a basis set [226] (after all, if we wanted an additional s-function, why include it in such an indirect and clumsy way instead of just designing a specific one that suits our particular needs?). Consequently, the most common practice in the field is to use Cartesian GTOs removing from the basis sets the linear combinations such as the $x^2 + y^2 + z^2$ above. This has also been the choice in chapter 7 of this dissertation, where a large number of basis sets are compared and their efficiency assessed.

Now, even if the integrals involving cGTOs can be computed analytically, there are still $\sim M^4$ of them in a SCF calculation. For example, in the model dipeptide HCO-L-Ala-NH_2 (see appendix E), which is extensively studied in this dissertation, there are 62 electrons and henceforth 31 RHF spatial orbitals ϕ_I. If a basis set with only 31 functions is used (this is a lower bound that will be rarely reached in practical calculations due to symmetry issues, see below), near a million of four-center $\langle ac|\, 1/r \,|bd\rangle$ integrals must be computed. This is why, one must use the freedom that remains once the decision of sticking to cGTOs has been taken (namely, the choice of the *exponents* ζ_a and the *angular momentum* l_a) to design basis sets that account for the relevant behaviour of the systems studied while keeping M below the 'pain threshold'.

The work by Nobel Prize John Pople's group has been a major reference in this discipline, and their STO-nG family [230], together with the split-valence Gaussian basis sets, 3-21G, 4-31G, 6-31G, etc. [231–238], which are extensively studied in chapter 7 of this dissertation, shall be used here to exemplify some relevant issues.

To begin with, let us recall that the short- and long-range behaviour of the Slater-type orbitals in eq. (2.91) is better than that of the more computationally efficient GTOs. In order to improve the physical properties of the latter, it is customary to linearly combine M_a Cartesian GTOs, denoted now by ξ_a^μ ($\mu = 1, \ldots, M_a$), and termed *primitive Gaussian-type orbitals* (PGTO), having the same atomic center $\vec{R}_{\alpha_a}$, the same set of orbital quantum numbers, l_a^x, l_a^y and l_a^z, but different exponents ζ_a^μ, to make up a *contracted Gaussian-type orbitals* (CGTO), defined by

$$\chi_a(\vec{r}\,;\vec{R}_{\alpha_a}) := \sum_\mu^{M_a} g_a^\mu \xi_a^\mu(\vec{r}\,;\vec{R}_{\alpha_a}) = \\ \left(r^1 - R^1_{\alpha_a}\right)^{l_a^x}\left(r^2 - R^2_{\alpha_a}\right)^{l_a^y}\left(r^3 - R^3_{\alpha_a}\right)^{l_a^z} \sum_\mu^{M_a} g_a^\mu \mathcal{N}_a^\mu \exp\left(-\zeta_a^\mu\, |\vec{r} - \vec{R}_{\alpha_a}|^2\right), \qquad (2.95)$$

where the normalization constants $\mathcal{N}_a^\mu$ have been kept inside the sum because they typically depend on ζ_a^μ. Also, we denote now by M_C the number of contracted GTOs and, by $M_P := \sum_a M_a$, the number of primitive ones.

In the STO-nG family of basis sets, for example, n primitive GTOs are used for each contracted one, fitting the coefficients g_a^μ and the exponents ζ_a^μ to resemble the radial behavior of Slater-type orbitals [230]. In fig. 2.2, the 1s-contracted GTO (see the discussion

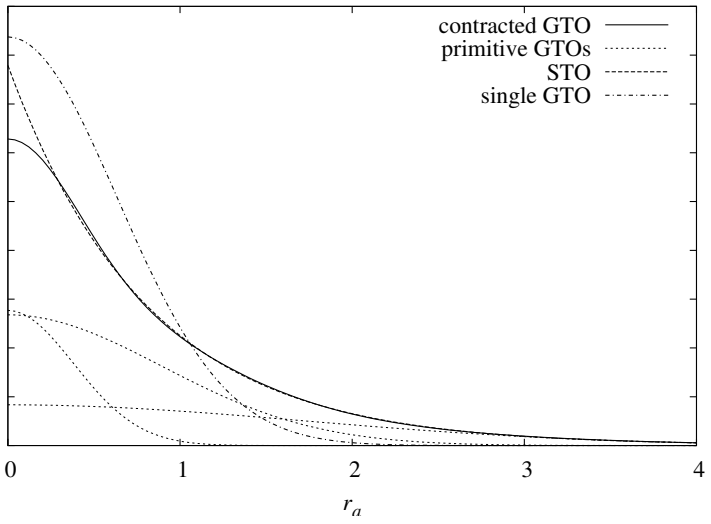


Figure 2.2: Radial behaviour of the 1s-contracted GTO of the hydrogen atom in the STO-3G basis set [230], the three primitive GTOs that form it, the STO that is meant to be approximated and a single GTO with the same norm and exponent as the STO. The notation r_a is shorthand for the distance to the α_a-th nucleus $|\vec{r} - \vec{R}_{\alpha_a}|$.

below) of the hydrogen atom in the STO-3G basis set is depicted, together with the three primitive GTOs that form it and the STO that is meant to be approximated[68]. We can see that the contracted GTO has a very similar behavior to the STO in a wide range of distances, while the single GTO that is also shown in the figure (with the same norm and exponent as the STO) has not.

Typically, the fitting procedure that leads to contracted GTOs is performed on isolated atoms and, then, the already mentioned chemical intuition that states that 'atoms-in-molecules are not very different from atoms-alone' is invoked to keep the linear combinations fixed from there on. Obviously, better results would be obtained if the contraction coefficients were allowed to vary. Moreover, the number of four-center integrals that need to be calculated depends on the number of *primitive* GTOs (like $\sim M_P^4$), so that we have not gained anything on this point by contracting. However, the size of the variational space is M_C (i.e., the number of *contracted* GTOs), in such a way that, once the integrals $\langle ac|\, 1/r\, |bd\rangle$ are calculated (for non-direct SCF), all subsequent steps in the iterative self-consistent procedure scale as powers of M_C. Also, the disk storage (again, for non-direct schemes) depends on number of *contracted* GTOs and, frequently, it is the disk storage and not the CPU time the limiting factor of a calculation.

An additional chemical concept that is usually defined in this context and that is

[68] Basis sets were obtained from the Extensible Computational Chemistry Environment Basis Set Database at http://www.emsl.pnl.gov/forms/basisform.html, Version 02/25/04, as developed and distributed by the Molecular Science Computing Facility, Environmental and Molecular Sciences Laboratory which is part of the Pacific Northwest Laboratory, P.O. Box 999, Richland, Washington 99352, USA, and funded by the U.S. Department of Energy. The Pacific Northwest Laboratory is a multi-program laboratory operated by Battelle Memorial Institute for the U.S. Department of Energy under contract DE-AC06-76RLO 1830. Contact Karen Schuchardt for further information.

1s-shell		2sp-shell		
ζ_a^μ	g_a^μ	ζ_a^μ	g_a^μ (p)	g_a^μ (s)
71.6168370	0.15432897	2.9412494	-0.09996723	0.15591627
13.0450960	0.53532814	0.6834831	0.39951283	0.60768372
3.5305122	0.44463454	0.2222899	0.70115470	0.39195739

Table 2.3: Exponents ζ_a^μ and contraction coefficients g_a^μ of the primitive Gaussian shells that make up the three different contracted ones in the STO-3G basis set for carbon (see ref. 230 and footnote 68). The exponents of the 2s- and 2p-shells are constrained to be the same.

needed to continue with the discussion is that of *shell*: Atomic shells in quantum chemistry are defined analogously to those of the hydrogen atom, so that each electron is regarded as 'filling' the multi-electron atom 'orbitals' according to Hund's rules [239]. Hence, the *occupied shells* of carbon, for example, are defined to be 1s, 2s and 2p, whereas those of, say, silicon, would be 1s, 2s, 2p, 3s an 3p. Each shell may contain $2(2l+1)$ electrons if complete, where $2l+1$ accounts for the orbital angular momentum multiplicity and the 2 factor for that of electron spin.

Thus, using these definitions, all the basis sets in the aforementioned STO-nG family are *minimal*; in the sense that they are made up of only $2l+1$ contracted GTOs for each completely or partially occupied shell, so that the STO-nG basis sets for carbon, for example, contain two s-type contracted GTOs (one for the 1s- and the other for the 2s-shell) and three p-type ones (belonging to the 2p-shell). Moreover, due to rotational-symmetry arguments in the isolated atoms, all the $2l+1$ functions in a given shell are chosen to have the same exponents and the same contraction coefficients, differing only on the polynomial that multiplies the Gaussian part. Such $2l+1$ CGTOs shall be said to constitute a *Gaussian shell* (GS), and we shall also distinguish between the *primitive* (PGS) and *contracted* (CGS) versions.

In table 2.3, the exponents ζ_a^μ and the contraction coefficients g_a^μ of the primitive GTOs that make up the three different shells in the STO-3G basis set for carbon are presented (see ref. 230 and footnote 68). The fact that the exponents ζ_a^μ in the 2s- and 2p-shells are constrained to be the same is a particularity of some basis sets (like this one) which saves some computational effort and deserves no further attention.

Next, let us introduce a common notation that is used to describe the contraction scheme: It reads (*primitive shells*) / [*contracted shells*], or alternatively (*primitive shells*) → [*contracted shells*]. According to it, the STO-3G basis set for carbon, for example, is denoted as (6s,3p) → [2s,1p], or (6,3) → [2,1]. Moreover, since for organic molecules it is frequent to have only hydrogens and the 1st-row atoms C, N and O[69] (whose occupied shells are identical), the notation is typically extended and the two groups of shells are separated by a slash; as in (6s,3p/3s) → [2s,1p/1s] for STO-3G.

The first improvement that can be implemented on a minimal basis set such as the ones in the STO-nG is the *splitting*, which consists in including more than one Gaussian shell

[69] In proteins, one may also have sulphur in cysteine and methionine residues (see sec. 1.2).

1s-shell		2sp-shell		
ζ_a^μ	g_a^μ	ζ_a^μ	g_a^μ (p)	g_a^μ (s)
3047.52490	0.0018347	7.8682724	-0.1193324	0.0689991
457.36951	0.0140373	1.8812885	-0.1608542	0.3164240
103.94869	0.0688426	0.5442493	1.1434564	0.7443083
29.21015	0.2321844			
9.286663	0.4679413	0.1687144	1.0000000	1.0000000
3.163927	0.3623120			

Table 2.4: Exponents ζ_a^μ and contraction coefficients g_a^μ of the primitive Gaussian shells that make up the three different constrained ones in the 6-31G basis set for carbon (see ref. 232 and footnote 68). In the 1s-shell, there is only one contracted Gaussian shell made by six primitive ones, whereas, in the 2s- and 2p-valence shells, there are two CGSs, one of them made by three PGSs and the other one only by a single PGS. The exponents of the 2s- and 2p-shells are constrained to be the same.

for each occupied one. If the splitting is evenly performed, i.e., each shell has the same number of GSs, then the basis set is called *double zeta* (DZ), *triple zeta* (TZ), *quadruple zeta* (QZ), *quintuple zeta* (5Z), *sextuple zeta* (6Z), and so on; where the word *zeta* comes from the Greek letter ζ used for the exponents. A hypothetical TZ basis set in which each CGTO is made by, say, four primitive GTOs, would read (24s,12p/12s) → [6s,3p/3s] in the aforementioned notation.

At this point, the already familiar intuition that says that 'atoms-in-molecules are not very different from atoms-alone' must be refined with another bit of chemical experience and qualified by noticing that 'core electrons are less affected by the molecular environment and the formation of bonds than valence electrons'[70]. In this spirit, the above evenness among different shells is typically broken, and distinct basis elements are used for the energetically lowest lying (*core*) shells than for the highest lying (*valence*) ones.

On one side, the contraction scheme may be different. In which case, the notation used up to now becomes ambiguous, since, for example, the designation (6s,3p/3s) → [2s,1p/1s], that was said to correspond to STO-3G, would be identical for a *different* basis set in which the 1s-Gaussian shell of heavy atoms be formed by 4 PGSs and the 2s-Gaussian shell by 2 PGSs (in 1st-row atoms, the 2s- and 2p-shells are defined as valence and the 1s-one as core, while in hydrogen atoms, the 1s-shell is a valence one). This problem can be solved by explicitly indicating how many primitive GSs form each contracted one, so that, for example, the STO-3G basis set is denoted by (33,3/3) → [2,1/1], while the other one mentioned would be (42,3/3) → [2,1/1] (we have chosen to omit the angular momentum labels this time).

The other point at which the core and valence Gaussian shells may differ is in their respective 'zeta quality', i.e., the basis set may contain a different number of contracted Gaussian shells in each case. For example, it is very common to use a single CGS for the

[70] Recall that, for the very concept of 'core' or 'valence electrons' (actually for any label applied to a single electron) to have any sense, we must be in the Hartree-Fock formalism (see footnote 64).

core shells and a multiple splitting for the valence ones. These type of basis sets are called *split-valence* and the way of naming their quality is the same as before, except for the fact that a capital V, standing for *valence*, is added either at the beginning or at the end of the acronyms DZ, TZ, QZ, etc., thus becoming VDZ, VTZ, VQZ, etc. or DZV, TZV, QZV, etc.

Pople's 3-21G [236], 4-31G [231], 6-31G [232] and 6-311G [235] are well-known examples of split-valence basis sets that are commonly used for SCF calculations in organic molecules and that present the two characteristics discussed above. Their names indicate the contraction scheme, in such a way that the number before the dash represents how many primitive GSs form the single contracted GSs that is used for core shells, and the numbers after the dash how the valence shells are contracted, in much the same way as the notation in the previous paragraphs. For example, the 6-31G basis set (see table 2.4), contains one CGS made up of six primitive GSs in the 1s-core shell of heavy atoms (the 6 before the dash) and two CGS, formed by three and one PGSs respectively, in the 2s- and 2p-valence shells of heavy atoms and in the 1s-shell of hydrogens (the 31 after the dash). The 6-311G basis set, in turn, is just the same but with an additional single-primitive Gaussian shell of functions in the valence region. Finally, to fix the concepts discussed, let us mention that, using the notation introduced above, these two basis sets may be written as (631,31/31) → [3,2/2] and (6311,311/311) → [4,3/3], respectively.

Two further improvements that are typically used and that may also be incorporated to Pople's split-valence basis sets are the addition of *polarization* [233, 234] or *diffuse functions* [234, 237, 238]. We shall discuss them both to close this section.

Up to now, neither the contraction nor the splitting involved GTOs of larger angular momentum than the largest one among the occupied shells. However, the molecular environment is highly anisotropic and, for most practical applications, it turns out to be convenient to add these *polarization* (large angular momentum) Gaussian shells to the basis set, since they present lower symmetry than the GSs discussed in the preceding paragraphs. Typically, the polarization shells are single-primitive GSs and they are denoted by adding a capital P to the end of the previous acronyms, resulting into, for example, DZP, TZP, VQZP, etc., or, say, DZ2P, TZ3P, VQZ4P if more than one polarization shell is added. In the case of Pople's basis sets [233, 234], these improvements are denoted by specifying, in brackets and after the letter G, the number and type of the polarization shells separating heavy atoms and hydrogens by a comma[71]. For example, the basis set 6-31G(2df,p) contains the same Gaussian shells as the original 6-31G plus two d-type shells and one f-type shell centered at the heavy atoms, as well as one p-type shell centered at the hydrogens.

Finally, for calculations in charged species (specially anions), where the charge density extends further in space and the tails of the distribution are more important to account for the relevant behaviour of the system, it is common to augment the basis sets with *diffuse functions*, i.e., single-primitive Gaussian shells of the same angular momentum as some preexisting one but with a smaller exponent ζ than the smallest one in the shell. In general, this improvement is commonly denoted by adding the prefix *aug-* to the name of the basis set. In the case of Pople's basis sets, on the other hand, the insertion of a plus sign '+' between the contraction scheme and the letter G denotes that the set contains one diffuse function in the 2s- and 2p-valence shells of heavy atoms. A second + indicates that

[71] There also exists an old notation for the addition of a single polarization shell per atom that reads 6-31G** and that is equivalent to 6-31G(d,p). It will not be used in this dissertation

there is another one in the 1s-shell of hydrogens. For example, one may have the doubly augmented (and doubly polarized) 6-31++G(2d,2p) basis set.

2.10 Møller-Plesset 2

Despite the simplicity of the variational ansatz in eq. (2.31) and the severe truncation of the N-electron space that it implies, Hartree-Fock calculations typically account for a large part of the energy of molecular systems [177, 180], and they are known to yield surprisingly accurate results in practical cases [207, 240–243]. Therefore, it is not strange that the Hartree-Fock method is chosen as a first step in many works found in the literature [163, 208, 244, 245].

However, the interactions that drive the conformational behaviour of organic molecules are weak and subtle [246, 247], so that, sometimes, higher levels of the theory than HF are required in order to achieve the target accuracy (see for example refs. 208, 209 and the study in chapter 7). As we discussed in the previous sections, the error in the Hartree-Fock results may come from two separate sources: the fact that we are using only one Slater determinant (see sec. 2.7) and the fact that, in actual calculations, the basis set that spans the one-electron space is finite (see sec. 2.9). Increasing the basis set size leads to the so-called *Hartree-Fock limit*, i.e., the exact HF result. On the other hand, any improvement in the description of the N-electron space is said to add *correlation* to the problem, since, due to the good behaviour mentioned above, Hartree-Fock is almost ever used as a reference to improve the formalism, and, as we remarked in sec. 2.7, the popular RHF and UHF versions of Hartree-Fock present some interesting properties of statistical independence between pairs of electrons. Apart from that, the word *correlation* is not much more than a convenient way of referring to *the difference between the exact fundamental state of the electronic Schrödinger equation and that provided by the solution of the Hartree-Fock problem in the infinite basis set limit*. In this spirit, the *correlation energy* is defined as

$$E_{\text{corr}} := E - E_{\text{HF}} \,, \tag{2.96}$$

where E is the exact energy and E_{HF} is the Hartree-Fock one (in the HF limit).

One may improve the description of the N-electron wavefunction in different ways that yield methods of distinct computational and physical properties (see refs. 177, 180, 248, 249 for detailed accounts). Here, we shall circumscribe to the particular family of methods known as *Møller-Plesset*, after the physicists that introduced them in 1934 [250], and, we shall give explicit expressions for the *2nd order Møller-Plesset* (MP2) variant, which is the only correlated method used in this dissertation (see chapters 4, 6 and 7). MP2 is typically regarded as accurate in the literature and it is considered as the reasonable starting point to include correlation [166, 241, 244, 246, 251, 252]. It is also commonly used as a reference calculation to evaluate or parameterize less demanding methods [163, 164, 253–255].

The basic formalism on which the Møller-Plesset methods are based is the *Rayleigh-Schrödinger perturbation theory* [180]. According to it, one must separate, if possible, the exact Hamiltonian operator $\hat{H}$ of the problem into a part $\hat{H}_0$ which is easy to solve and another one $\hat{H}'$ which is 'much smaller' than $\hat{H}_0$ and that can, therefore, be regarded as a

perturbation:

$$\hat{H} = \hat{H}_0 + \hat{H}' \ . \tag{2.97}$$

Then, a dimensionless parameter λ is introduced,

$$\hat{H}(\lambda) = \hat{H}_0 + \lambda \hat{H}' \ , \tag{2.98}$$

so that, for $\lambda = 0$, we have the *unperturbed Hamiltonian* (i.e., $\hat{H}(0) = \hat{H}_0$), and, for $\lambda = 1$, we have the exact one (i.e., $\hat{H}(1) = \hat{H}$).

The only utility of this parameter λ is to formally expand any exact eigenfunction Ψ_n of $\hat{H}$ and the corresponding exact eigenvalue E_n in powers of it,

$$\Psi_n(\underline{\mathrm{x}}) = \Psi_n^{(0)}(\underline{\mathrm{x}}) + \lambda\, \Psi_n^{(1)}(\underline{\mathrm{x}}) + \lambda^2 \Psi_n^{(2)}(\underline{\mathrm{x}}) + \lambda^3 \Psi_n^{(3)}(\underline{\mathrm{x}}) + o(\lambda^4) \ , \tag{2.99a}$$

$$E_n = E_n^{(0)} + \lambda\, E_n^{(1)} + \lambda^2 E_n^{(2)} + \lambda^3 E_n^{(3)} + o(\lambda^4) \ , \tag{2.99b}$$

and then truncate the expansion at low orders invoking the aforementioned 'relative smallness' of $\hat{H}'$[72].

Now, to calculate the different terms in the expressions above, we take them, together with eq. (2.98), to the Schrödinger equation $\hat{H}\Psi_n(\underline{\mathrm{x}}) = E_n \Psi_n(\underline{\mathrm{x}})$, obtaining a series of coupled relations that result from equating the factors that multiply each power of λ:

$$\hat{H}_0 \Psi_n^{(0)}(\underline{\mathrm{x}}) = E_n^{(0)} \Psi_n^{(0)}(\underline{\mathrm{x}}) \ , \tag{2.100a}$$

$$\hat{H}_0 \Psi_n^{(1)}(\underline{\mathrm{x}}) + \hat{H}' \Psi_n^{(0)}(\underline{\mathrm{x}}) = E_n^{(0)} \Psi_n^{(1)}(\underline{\mathrm{x}}) + E_n^{(1)} \Psi_n^{(0)}(\underline{\mathrm{x}}) \ , \tag{2.100b}$$

$$\hat{H}_0 \Psi_n^{(2)}(\underline{\mathrm{x}}) + \hat{H}' \Psi_n^{(1)}(\underline{\mathrm{x}}) = E_n^{(0)} \Psi_n^{(2)}(\underline{\mathrm{x}}) + E_n^{(1)} \Psi_n^{(1)}(\underline{\mathrm{x}}) + E_n^{(2)} \Psi_n^{(0)}(\underline{\mathrm{x}}) \ , \tag{2.100c}$$

$$\cdots$$

$$\hat{H}_0 \Psi_n^{(k)}(\underline{\mathrm{x}}) + \hat{H}' \Psi_n^{(k-1)}(\underline{\mathrm{x}}) = \sum_{i=0}^{k} E_n^{(i)} \Psi_n^{(k-i)}(\underline{\mathrm{x}}) \ . \tag{2.100d}$$

Eq. (2.100a) indicates that the zeroth order energies $E_n^{(0)}$ and wavefunctions $\Psi_n^{(0)}(\underline{\mathrm{x}})$ are actually the eigenvalues and eigenvectors of the unperturbed Hamiltonian $\hat{H}_0$. Then, we denote $|n\rangle := |\Psi_n^{(0)}\rangle$ and require that $\langle n|\Psi_n\rangle = 1$, $\forall n$. This condition, which is called *intermediate normalization*, can always be achieved, if $|\Psi_n\rangle$ and $|n\rangle$ are not orthogonal, by appropriately choosing the norm of $|\Psi_n\rangle$. Next, if we 'multiply' eq. (2.99a) from the left by $\langle n|$ and use this property, we have that

$$\langle n|\Psi_n\rangle = \langle n|n\rangle + \lambda \langle n|\Psi_n^{(1)}\rangle + \lambda^2 \langle n|\Psi_n^{(2)}\rangle + \lambda^3 \langle n|\Psi_n^{(3)}\rangle + o(\lambda^4) = 1 \ . \tag{2.101}$$

Since this identity must hold for every value of λ, we see that, due to the normalization chosen, the zeroth order eigenfunctions $|n\rangle$ are orthogonal to all their higher order corrections, i.e., that $\langle n|\Psi_n^{(k)}\rangle = 0$, $\forall k > 0$. Next, if we use this property and multiply all

[72] The approach presented here is common in quantum chemistry books [177, 180] but different from the one usually taken in physics [256, 257], where the perturbation is not $\hat{H}'$ but $\lambda\hat{H}'$, since λ is considered to be small and $\hat{H}'$ of the same 'size' as $\hat{H}_0$. Nevertheless, the derivation and the concepts involved are very similar and, of course, the final operative equations are identical.

eqs. (2.100) from the left by $\langle n|$, we obtain very simple expressions for the different orders of the energy:

$$E_n^{(0)} = \langle n|\,\hat{H}_0\,|n\rangle\,, \tag{2.102a}$$
$$E_n^{(1)} = \langle n|\,\hat{H}'\,|n\rangle\,, \tag{2.102b}$$
$$E_n^{(2)} = \langle n|\,\hat{H}'\,|\Psi_n^{(1)}\rangle\,, \tag{2.102c}$$
$$\dots$$
$$E_n^{(k)} = \langle n|\,\hat{H}'\,|\Psi_n^{(k-1)}\rangle\,. \tag{2.102d}$$

The expressions for the zeroth and first orders of the energy E_n, in eqs. (2.102a) and (2.102b) respectively, do not require the knowledge of any correction to the wavefunction. The second order energy $E_n^{(2)}$, however, involves the function $\Psi_n^{(1)}$, which is a priori unknown.

In order to arrive to an equation for $E_n^{(2)}$ that is operative, we write each correction $\Psi_n^{(k>0)}$ to the n-th eigenstate (and particularly $\Psi_n^{(1)}$) as a linear combination of the eigenstates of $\hat{H}_0$ themselves (which we assume to form a complete basis of the N-electron Hilbert space)[73]:

$$|\Psi_n^{(k)}\rangle = \sum_{m\neq n} c_{nm}^{(k)}|m\rangle = \sum_{m\neq n}\langle m|\Psi_n^{(k)}\rangle\,|m\rangle\,, \qquad \forall k>0\,, \tag{2.103}$$

where the $m = n$ term has been omitted because, from the orthogonality conditions mentioned in the preceding paragraphs, it follows that $c_{nn}^{(k)} = 0$, $\forall k > 0$.

Now, we take this to eq. (2.102c) and obtain

$$E_n^{(2)} = \sum_{m\neq n}\langle n|\,\hat{H}'\,|m\rangle\langle m|\Psi_n^{(1)}\rangle\,. \tag{2.104}$$

Next, we rearrange eq. (2.100b):

$$(E_n^{(0)} - \hat{H}_0)\,|\Psi_n^{(1)}\rangle = (\hat{H}' - E_n^{(1)})\,|n\rangle = (\hat{H}' - \langle n|\,\hat{H}'\,|n\rangle)\,|n\rangle\,, \tag{2.105}$$

where, in the last step, eq. (2.102b) has been used.

Recalling that $\langle m|n\rangle = 0$, we multiply the identity above from the left by an arbitrary eigenvector $\langle m| \neq \langle n|$ of $\hat{H}_0$:

$$(E_n^{(0)} - E_m^{(0)})\,\langle m|\Psi_n^{(1)}\rangle = \langle m|\,\hat{H}'\,|n\rangle \quad\Longrightarrow\quad \langle m|\Psi_n^{(1)}\rangle = \frac{\langle m|\,\hat{H}'\,|n\rangle}{E_n^{(0)} - E_m^{(0)}}\,. \tag{2.106}$$

Finally, we take this to eq. (2.104) to write an operative expression for the *second order correction* to the energy:

$$E_n^{(2)} = \sum_{m\neq n}\frac{|\langle m|\,\hat{H}'\,|n\rangle|^2}{E_n^{(0)} - E_m^{(0)}}\,. \tag{2.107}$$

[73] Like we did in sec. 2.4, we assume that the Hamiltonian operator $\hat{H}_0$ has only discrete spectrum in order to render the derivation herein more clear. Anyway, in the only application of the Rayleigh-Schrödinger perturbation theory that is performed in this dissertation, the Møller-Plesset family of methods in computational quantum chemistry, the Hilbert space is finite-dimensional (since the basis sets is finite) and we are in the conditions assumed.

The higher order corrections, which can be derived performing similar calculations, shall not be obtained here. Also, note that, although the equation above is valid for all n, we intend to apply it to the fundamental state, in such a way that $n = 0$ and, since $E_0^{(0)} < E_m^{(0)}$, $\forall m \neq 0$, all the terms in the sum are *negative*, and $E_0^{(2)} < 0$.

So far, we have introduced the Rayleigh-Schrödinger perturbation theory in a completely general manner that may be applied to any quantum system. The practical use of all this formalism in quantum chemistry is what is known as the *Møller-Plesset* approach. According to it, one must first perform a SCF Hartree-Fock calculation (which shall be assumed here to be of the RHF class) using a finite basis set of size M (see secs. 2.8 and 2.9 for reference). As a result, a set of M orthogonal orbitals ϕ_a (also specified by M-tuples c_{ba}) are produced, among which the $N/2$ ones with the lowest eigenvalues ε_a of the self-consistent Fock operator $\hat{F}[\phi]$ are said to be *occupied* and denoted with capital indices starting in the middle part of the alphabet ($I, J, K, L, \ldots$). The remaining $M - N/2$ are called *virtual orbitals* and shall be indicated using indices beginning by $r, s, t, \ldots$

The *unperturbed Hamiltonian* $\hat{H}_0$ used in the general expressions above is defined to be here the sum of the self-consistent one-electron Fock operators in eq. (2.79), acting each one of them on one of the N coordinates $\vec{r}_i$ (just add an i-label to the $\vec{r}$ coordinates in the definition of the Coulomb and exchange operators in in eq. (2.74) and simply use eq. (2.36) for $\hat{h}_i$):

$$\hat{H}_0 := \sum_i^N \hat{F}_i[\phi] = \sum_i^N \left[\hat{h}_i + \sum_J^{N/2} \left(2\hat{J}_J^i[\phi] - \hat{K}_J^i[\phi] \right) \right] . \tag{2.108}$$

The *perturbation* $\hat{H}'$ is the difference

$$\hat{H}' := \hat{H} - \hat{H}_0 = \frac{1}{2} \sum_{i \neq j} \frac{1}{r_{ij}} - \sum_J^{N/2} \left(2\hat{J}_J^i[\phi] - \hat{K}_J^i[\phi] \right) , \tag{2.109}$$

where $\hat{H}$ is the total electronic Hamiltonian in eq. (2.20) and one must note that there is no reason to believe that $\hat{H}'$ is 'small'. As a result, the Møller-Plesset series may be ill-behaved or even divergent in some situations [177] and, like in the Hartree-Fock case, the quality of theoretical predictions must be ultimately assessed by comparison to experimental data or to more accurate methods.

Now, if, despite this lack of certainty, we are brave enough to keep on walking the Rayleigh-Schrödinger path, we must first note that, although it is easy to show that the N-particle Slater determinant constructed with the $N/2$ Hartree-Fock optimized orbitals is an eigenstate of $\hat{H}_0$, we need to know the rest of the eigenstates if the second order expression for the energy in eq. (2.107) is to be used.

The whole N-electron Hilbert space in which these eigenstates 'live' has infinite dimension, however, as a part of the Møller-Plesset strategy, there is a convenient way of *extrapolating the LCAO approximation in sec. 2.8 to the N-particle space*. This strategy is not only used in MP2 but also in other post-HF methods [177, 180, 248, 249] and it allows us to construct a natural finite basis of eigenfunctions of $\hat{H}_0$. The trick is simple: Due to the fact that, except for the case of minimal basis sets and complete-shell atoms (noble gases), the total number of canonical orbitals ϕ_a obtained from a SCF calculation (M) is larger than that of the occupied ones ($N/2$), we may take any ϕ_I from this

second group and substitute it by any ϕ_r among the virtual orbitals in order to form a new Slater determinant. This is called a *single substitution*, and the determinants obtained from all possible single, double, triple,..., $M-1$, and M substitutions constitute a basis set of N-particle eigenfunctions of $\hat{H}_0$ that is complete in the (antisymmetrized) tensorial product of the one-particle Hilbert spaces spanned by the M molecular orbitals ϕ_a (i.e., in the *full-Configuration Interaction space* [177, 180, 248, 249]).

Any one of these N-electron states is completely specified, in the RHF case, by two *unordered* set of integers: on the one hand, $\mathcal{I}_\alpha := \{a_1, a_2, \ldots, a_{N/2}\}$, which contains the indices of the spatial orbitals that appear multiplied by α spin functions, and, on the other, $\mathcal{I}_\beta := \{b_1, b_2, \ldots, b_{N/2}\}$, which corresponds to the β electrons. In order for the Slater determinant defined by a given pair $\mathcal{I} := (\mathcal{I}_\alpha, \mathcal{I}_\beta)$ not to be zero, all the indices within each set must be different, however, any of them may appear in both sets at the same time. The number of possible $\mathcal{I}_\alpha$ (or $\mathcal{I}_\beta$) sets is, therefore,

$$\binom{M}{N/2} := \frac{M!}{(M-N/2)!(N/2)!} ,$$

and, consequently, the number of different RHF Slater determinants constructed following this procedure (the number of elements in $\mathcal{I}$) is

$$\binom{M}{N/2}^2 := \left(\frac{M!}{(M-N/2)!(N/2)!}\right)^2 ,$$

which, already for small molecules and small basis sets, is a daunting number. For example, in the case of the model dipeptide HCO-L-Ala-NH_2 investigated in this dissertation (see appendix E) and the 6-31G(d) basis set (see sec. 2.9), we have that $N/2 = 31$ and $M = 128$, so that there exist more than 10^{59} different RHF Slater determinants that can be constructed with the Hartree-Fock molecular orbitals.

Due to the fact that any given orbital ϕ_a is an eigenfunction of the Fock operator $\hat{F}[\phi]$ with eigenvalue ε_a, one can show that, if $\Psi_{\mathcal{I}}$ is the Slater determinant that corresponds to the set of indices $\mathcal{I} := (\{a_1, \ldots, a_{N/2}\}, \{b_1, \ldots, b_{N/2}\})$, then, it is an eigenstate of $\hat{H}_0$ with eigenvalue

$$\sqrt{N!}\left(\sum_I^{N/2} \varepsilon_{a_I} + \sum_I^{N/2} \varepsilon_{b_I}\right) . \tag{2.110}$$

Henceforth, it is clear that the Slater determinant of lowest $\hat{H}_0$ eigenvalue, i.e., the zeroth order approximation $|0\rangle$ to the fundamental state, is the one that corresponds to the particular set of orbital indices $\mathcal{I}_0 := \{1, 1, 2, 2, \ldots, N/2, N/2\}$. This is precisely the optimal Hartree-Fock wavefunction, and the *zeroth order approximation to the energy of the fundamental state* in eq. (2.102a) is

$$E_0^{(0)} = 2\sum_I^{N/2} \varepsilon_I . \tag{2.111}$$

Next, in order to find the first and second order corrections, we need to calculate, in the basis of N-electron functions introduced above, the matrix elements $\langle m|\hat{H}'|n\rangle$ which are associated to the perturbation.

First, if we are interested in a diagonal matrix element $\langle\Psi_I|\,\hat{H}'\,|\Psi_I\rangle$, corresponding to the eigenstate $|\Psi_I\rangle := |m\rangle = |n\rangle$, the calculations performed in sec. 2.7 can be recalled to show that, in the RHF case, we have that

$$\begin{aligned}\langle\Psi_I|\,\hat{H}'\,|\Psi_I\rangle = \\ -\sum_{I,J}^{N/2}\Bigg[\langle\phi_{a_I}\phi_{b_J}|\,\frac{1}{r}\,|\phi_{a_I}\phi_{b_J}\rangle + \frac{1}{2}\Bigg(\langle\phi_{a_I}\phi_{a_J}|\,\frac{1}{r}\,|\phi_{a_I}\phi_{a_J}\rangle + \langle\phi_{b_I}\phi_{b_J}|\,\frac{1}{r}\,|\phi_{b_I}\phi_{b_J}\rangle\Bigg) \\ -\frac{1}{2}\Bigg(\langle\phi_{a_I}\phi_{a_J}|\,\frac{1}{r}\,|\phi_{a_J}\phi_{a_I}\rangle + \langle\phi_{b_I}\phi_{b_J}|\,\frac{1}{r}\,|\phi_{b_J}\phi_{b_I}\rangle\Bigg)\Bigg]\,.\end{aligned} \tag{2.112}$$

If we now particularize this equation to the case of the Hartree-Fock wavefunction $|\Psi_{I_0}\rangle = |0\rangle$, an operative expression for the *first order energy correction* in eq. (2.102b) is obtained:

$$E_0^{(1)} = \langle 0|\,\hat{H}'\,|0\rangle = -\sum_{I,J}^{N/2}\Bigg(2\langle\phi_I\phi_J|\,\frac{1}{r}\,|\phi_I\phi_J\rangle - \langle\phi_I\phi_J|\,\frac{1}{r}\,|\phi_J\phi_I\rangle\Bigg)\,. \tag{2.113}$$

Using this relation together with eq. (2.111), we see that the *first order-corrected energy*, i.e., the *MP1* result,

$$E_0^{\text{MP1}} := E_0^{(0)} + E_0^{(1)} = 2\sum_{I}^{N/2}\varepsilon_I - \sum_{I,J}^{N/2}\Bigg(2\langle\phi_I\phi_J|\,\frac{1}{r}\,|\phi_I\phi_J\rangle - \langle\phi_I\phi_J|\,\frac{1}{r}\,|\phi_J\phi_I\rangle\Bigg)\,, \tag{2.114}$$

exactly matches the RHF energy in eq. (2.80), so that the differences between Hartree-Fock and Møller-Plesset perturbation theory appear at second order for the first time.

In order to calculate $E_0^{(2)}$, we first remark that it can be easily shown that, since $\hat{H}'$ is made up of (at most) two-body operators, the matrix element $\langle m|\,\hat{H}'\,|n\rangle$ will be zero if $|n\rangle$ and $|m\rangle$ differ in *more than two* ϕ_a orbitals (due to the orthogonality of the latter).

On the other hand, Brillouin's theorem [177, 180] states that the matrix elements $\langle m|\,\hat{H}\,|n\rangle$ of the exact electronic Hamiltonian $\hat{H}$ between states that differ in *only one* orbital are zero, and, using again the orthogonality of the ϕ_a, one can additionally show that the same happens for matrix elements $\langle m|\,\hat{H}_0\,|n\rangle$ of the unperturbed Hamiltonian. Since $\hat{H}'$ is just the difference $\hat{H} - \hat{H}_0$, it follows that $\langle m|\,\hat{H}'\,|n\rangle$ is zero if $|n\rangle$ and $|m\rangle$ differ in *only one* ϕ_a orbital.

Putting all together, the only terms that survive in the sum in eq. (2.107) (for $n = 0$) are those corresponding to matrix elements between the Hartree-Fock fundamental state $|0\rangle$ and its doubly-substituted Slater determinants. Consequently, and after some calculations, it can be shown that the *second order correction to the energy of the fundamental state* in terms of the SCF orbitals $|\phi_a\rangle$ (denoted again as $|a\rangle$) is given by

$$E_0^{(2)} = \sum_{I,J}^{N/2}\sum_{r,s=N/2+1}^{M}\frac{2\left|\langle IJ|\,\frac{1}{r}\,|rs\rangle\right|^2 - \langle IJ|\,\frac{1}{r}\,|rs\rangle\langle IJ|\,\frac{1}{r}\,|sr\rangle}{\varepsilon_I + \varepsilon_J - \varepsilon_r - \varepsilon_s}\,. \tag{2.115}$$

The computational cost of calculating $E_0^{(2)}$ scales like M^5 [177], and the *MP2 energy* is

$$E_0^{\mathrm{MP2}} := E_0^{(0)} + E_0^{(1)} + E_0^{(2)} = E_0^{\mathrm{RHF}} + E_0^{(2)} \,. \tag{2.116}$$

Finally, let us note that, in agreement with the already mentioned chemical intuition (in sec. 2.9) that states that 'core electrons are less affected by the molecular environment and the formation of bonds than valence electrons', it is common to remove the core shells from the calculation of the second order correction to the energy in eq. (2.115). This is called the *frozen core* approximation and it is used in the study performed in chapter 7.

Chapter 3

A physically meaningful distance between potentials

This chapter is based on the article:
J. L. Alonso and Pablo Echenique, *A physically meaningful method for the comparison of potential energy functions*, J. Comp. Chem. **27** (2006) 238–252.

> I often say that, when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely in your thoughts advanced to the state of *Science*, whatever the matter may be. [258]
>
> — William Thompson, Lord Kelvin, 1891

3.1 Introduction

As we have seen in sec. 1.4, the effective potential energy $V(\vec{q})$ of a system[74], as a function of the coordinates $\vec{q}$, completely determines its physical behaviour. In the long path that leads to the design of proper energy functions that can successfully fold proteins (which is one of the central topics of this dissertation), researchers often tackle the problem of designing new algorithms to calculate the potential energy of these complex systems, looking for the improvement of the relation between accuracy and computational demands [259–270]. At a more pragmatic level, when the aim simply is to study and describe a particular phenomenon, the dilemma of choosing among many different ways of calculating a conceptually unique potential V always shows up [207, 271–276].

[74] In this chapter, we will notationally forget the fact that the potential energy comes from the averaging out of more microscopical degrees of freedom than the relevant coordinates $\vec{q}$, and we will use the more common designation V, instead of W.

The energy V that we are talking about may be the total potential energy of the system or any of the terms in which it is traditionally factorized[75] and the different ways of calculating it may stem from distinct origins, namely, that different algorithms or approximations A are used, that the potential energy function depends on a number of free parameters $\vec{P}$ or, finally, that it is computed on different but somehow related systems S (for a proper definition of this, see sec. 3.6).

Changes in these inputs produce different *instances* of the same physical potential energy, which we shall denote by subscripting V. For example, if it is calculated on the same system S, with the algorithm and approximations A held constant but with two different set of parameters $\vec{P}_1$ and $\vec{P}_2$, we shall write[76]:

$$V_1(\vec{q}) := V(A, \vec{P}_1, \vec{q}, S) \qquad \text{and} \qquad V_2(\vec{q}) := V(A, \vec{P}_2, \vec{q}, S)\,. \tag{3.1}$$

For each practical application of these two potential energy functions, there is a limit on how different V_1 could be from V_2 to preserve the relevant features of the system under scrutiny. Clearly, if V_1 is 'too distant' from V_2, the key characteristics of the system behaviour will be lost when going from one function to the other.

In the literature, a number of different methods are used to quantify this distance. Among them, the Pearson's correlation coefficient r does not have units and its meaning is only semi-quantitative. Some others, such as the root mean square deviation (RMSD), the mean error of the energies (ER), the standard deviation of the error (SDER) or the mean absolute error (AER), do have units of energy and their values can be compared to the physically relevant scale in each problem. However, they tend to overestimate the sought distance even in the interesting situations in which the potential energy functions under study are physically proximate. The aim of this chapter is to define, justify and describe a new meaningful measure $d(V_1, V_2)$ of the distance between two instances of the same potential energy that overcomes the aforementioned difficulties, and that allows to make precise statistical statements about the way in which the energy differences change when going from V_1 to V_2[77]. The introduction of this distance is the natural and necessary first step to carry out the program that begins with this dissertation.

In sec. 3.2, the hypotheses made on V_1 and V_2 to accomplish this are outlined and, in sec. 3.3, the central object, $d(V_1, V_2)$, is defined. The statistical meaning of the distance herein introduced is discussed in sec. 3.4 and 3.5 and some of its possible applications to practical situations are proposed in sec. 3.6. A comparison to other commonly used criteria is made and illustrated with a numerical example in sec. 3.7. The important issue of the additivity of $d(V_1, V_2)$, when the potentials studied are only a part of the total energy, is investigated in sec. 3.8 and, in sec. 3.9, the metric properties of our distance are discussed. The robustness of the van der Waals potential energy (as implemented in CHARMM [104, 105]) under a change in the free parameters and the ab initio Ramachandran plots of HCO-L-Ala-NH_2 at different levels of the theory are studied in sec. 3.10 as

[75] For example, in the case of proteins [2, 33, 137], some of the terms in which the total potential energy is typically factorized are the hydrogen-bonds energy, the van der Waals interaction, the excluded volume repulsion, the Coulomb energy and the solvation energy.

[76] Analogous definitions may be made if different algorithms or approximations, A_1 and A_2, are used or if V is computed on two related systems, S_1 and S_2.

[77] The convenience of this approach has been remarked in ref. 1. Note, however, that in this chapter a different distance is defined.

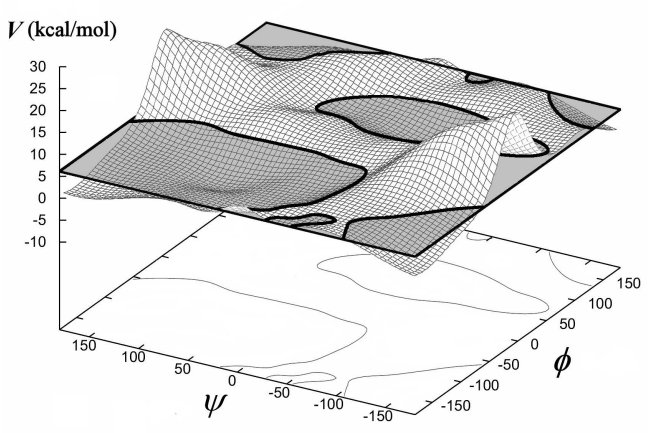


Figure 3.1: Space of constant potential energy V in a simple system with only two degrees of freedom: the alanine dipeptide HCO-L-Ala-NH_2 in vacuo (see appendix E). Potential energy surface (PES) calculated ab initio at the B3LYP/6-311++G(d,p) level of the theory in ref. 207.

examples of applications of the distance. Finally, sec. 3.11 is devoted to the conclusions and to a useful summary of the steps that must be followed to use the distance in a practical case.

3.2 Hypotheses

In some cases traditionally studied in physics, the dependence of the potential energy V on the parameters is simple enough to allow a closed functional dependence $V_2(V_1)$ to be found[78]. However, in the study of complex systems, such as proteins, this dependence is often much more complicated, due to the high dimensionality of the conformational space $\mathcal{I}$ and to the fact that the energy landscape lacks any evident symmetry. The set $C(V_1)$ of the conformations with a particular value of the potential energy V_1 typically spans large regions of $\mathcal{I}$ containing structurally very different conformations (see fig. 3.1). When the system is slightly modified, from S_1 to S_2, or an approximation is performed (or the algorithm is changed), from A_1 to A_2, or the free parameters are shifted, from $\vec{P}_1$ to $\vec{P}_2$, each conformation $\vec{q}$ in $C(V_1)$ is affected in a different way and its potential energy is modified, from $V_1(\vec{q})$ to $V_2(\vec{q})$, in a manner that does not depend trivially on the particular region of the conformational space which the conformation $\vec{q}$ belongs to. In such a case, a simple functional relation $V_2(V_1)$ is no longer possible to be found: for each value of V_1, there corresponds now a whole distribution of values of V_2 associated with conformations which share the same value of V_1 but which are far apart in the conformational space. Moreover, the projection of this high-dimensional $\vec{q}$-space into the 1-dimensional V_1-space makes it possible to treat V_2 as a random variable parametrically dependent on V_1 (see fig. 3.2), in the already suggested sense that, if one chooses at random a particular conformation $\vec{q}_i \in C(V_1)$, the 'outcome' of the quantity $V_2(\vec{q}_i)$ is basically unpredictable[79].

[78] For example, if the recovering force constant of a harmonic oscillator is changed from k_1 to k_2, the potential energy functions satisfy the linear relation $V_2(\vec{q}) = (k_1/k_2)V_1(\vec{q})$ for all the conformations of the system; if the the atomic charges are expressed as αQ_i and they are rescaled by changing α from α_1 to α_2, the free energies of solvation calculated via the Poisson equation satisfy the linear relation $V_2(\vec{q}) = (\alpha_1/\alpha_2)^2 V_1(\vec{q})$, etc.

[79] The same may be said in the case that the conformations belong to $C(V_2)$ and the random variable is, in turn, V_1. The role of the two instances of V is interchangeable in the whole following reasoning, however, for the sake of clarity, this fact will be made explicit in some cases and will be tacitly assumed in others.

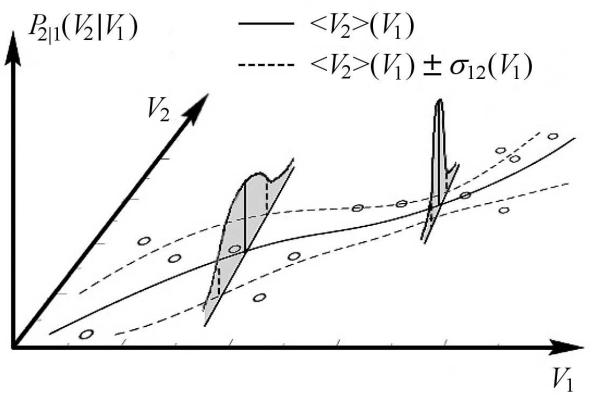


Figure 3.2: Illustration of the conditional-probability density function $P_{2|1}(V_2|V_1)$ at different V_1-positions. The points represent single conformations of the system. The V_1-conditioned mean of V_2, $\langle V_2\rangle(V_1)$, is depicted as a solid line and, although in the hypotheses it is assumed to depend linearly on V_1, here it is shown as a more general function to better illustrate the concepts involved. Analogously, the V_1-conditioned standard deviation of V_2, $\sigma_{12}(V_1)$ (which is assumed to be constant in the hypotheses) is added to $\langle V_2\rangle(V_1)$ (and subtracted from it) and the result is depicted as broken lines.

In this context, the hypotheses to be done about the two instances of V are, first, that, in our *working set* of conformations $\{\vec{q}_i\}_{i=1}^N$, the pair of values $(V_1(\vec{q}_i), V_2(\vec{q}_i))$ is independent of $(V_1(\vec{q}_j), V_2(\vec{q}_j))$ if $i \neq j$ and, second, that the probability distribution of V_2 conditioned by V_1 is *normal* with mean $b_{12}V_1 + a_{12}$ and standard deviation σ_{12} and that, conversely, the probability distribution of V_1 conditioned by V_2 (i.e., the distribution of the random variable V_1 in the space $C(V_2)$, analogous to $C(V_1)$) is *normal* with mean $b_{21}V_2 + a_{21}$ and standard deviation σ_{21}. Where a_{12}, b_{12}, σ_{12}, a_{21}, b_{21} and σ_{21} are constants not dependent on V_1 or V_2. This can be summarized in the following expressions for the conditional-probability density functions:

$$P_{2|1}(V_2|V_1) = \frac{1}{\sqrt{2\pi}\sigma_{12}} \exp\left[-\frac{(V_2 - (b_{12}V_1 + a_{12}))^2}{2\sigma_{12}^2}\right], \tag{3.2a}$$

$$P_{1|2}(V_1|V_2) = \frac{1}{\sqrt{2\pi}\sigma_{21}} \exp\left[-\frac{(V_1 - (b_{21}V_2 + a_{21}))^2}{2\sigma_{21}^2}\right]. \tag{3.2b}$$

At first sight, albeit symmetric, the above treatment, in which one of the instances is seen as an independent parameter and the other one as a random variable, may seem artificial and partially asymmetric. In fact, one could reason about the whole conformational space of the system and regard each randomly selected conformation $\vec{q}_i$ as a single *numerical experiment* to which the value of two random variables, $V_1(\vec{q}_i)$ and $V_2(\vec{q}_i)$, can be assigned. However, no hypotheses need to be made about the joint probability density function $P_{12}(V_1, V_2)$[80]. For the distance herein introduced to be meaningful, it suffices to

[80] The hypothesis that $P_{12}(V_1, V_2)$ is bivariate normal, for example, is stronger than the assumptions in eqs. (3.2). The latter can be derived from the former.

assume eqs. (3.2) and the partially asymmetric treatment performed above turns out to be less restrictive.

Regarding the question of whether in a typical case these hypotheses are fulfilled or not, some remarks should be made. First, the satisfaction of the independence hypothesis depends mainly on the process through which the *working set of conformations* $\{\vec{q}_i\}_{i=1}^N$ is generated. For example, if the conformations are extracted from a single molecular dynamics trajectory letting only a short simulation time pass between any pair of them, their energies will be obviously correlated and the independence will be broken. If, on the contrary, each conformation $\vec{q}_i$ is taken from a different trajectory (see the first application of the distance in sec. 3.10), one may reasonably expect this assumption to be fulfilled, i.e., the independence hypothesis is normally under researcher's control.

The normality hypothesis, however, is of a different nature. That the distribution of V_2 be normal for a particular value of V_1 may be thought as a consequence of the large number of degrees of freedom the system possesses, of the usual pairwise additivity of the forces involved and of the Central Limit Theorem (this, in fact, can be proved in some simple cases. Nevertheless, that the V_1-conditioned mean of V_2, $\langle V_2\rangle(V_1)$, is linear in V_1 and that the V_1-conditioned standard deviation of V_2, $\sigma_{12}(V_1)$, is a constant must be regarded as a zeroth order approximation that may be fulfilled for small energy ranges and that should be checked in each particular case (see fig. 3.2).

Finally, it is worth pointing out, that, for the commonly used statistical quantities r, RMSD, etc. to be useful, these two assumptions must be equally made (see sec. 3.7) and also that the normality hypothesis has been found to be approximately fulfilled in several cases studied (see, for example, ref. 1 and fig. 3.3b).

3.3 Definition

For the aforementioned cases in which the dependence of the potential energy on the parameters is simple enough (see footnote 78), one can describe $V_2(V_1)$ by a closed analytical formula and exactly compute a_{12}, b_{12} and σ_{12} (this last quantity being equal to zero in such a situation). However, in a general case, the parameters entering eqs. (3.2) cannot be calculated analytically, and one may at most have a finite collection of N conformations $\{\vec{q}_i\}_{i=1}^N$ (the *working set*) together with the respective values $V_1(\vec{q}_i)$ and $V_2(\vec{q}_i)$ for each one of them.

From this finite knowledge about the system, one may statistically estimate the values of a_{12}, b_{12} and σ_{12}. Under the hypothesis assumed in the previous section, the least-squares estimators [277, 278] of these quantities are optimal in the precise statistical sense that they are maximum-likelihood and have minimum variance in the class of linear and unbiased estimators[81] [279].

If we denote $V_1^i := V_1(\vec{q}_i)$ and $V_2^i := V_2(\vec{q}_i)$, the least-squares maximum-likelihood estimators[82] of a_{12}, b_{12} and σ_{12} are given by the following expressions [277, 278]:

[81] The same letters are used for the ideal parameters a_{12}, b_{12}, and σ_{12} and for their least-squares best estimators, because the only knowledge that one may have about the former comes from the calculation of the latter.

[82] The maximum-likelihood estimator for σ_{12} (with N in the denominator) is preferred to the unbiased one (with $N-2$ in the denominator) for consistency. Anyway, for the values of N typically used, the difference between them is negligible.

$$b_{12} = \frac{\mathrm{Cov}(V_1, V_2)}{\sigma_1^2} \,, \tag{3.3a}$$

$$a_{12} = \mu_2 - b_{12}\mu_1 \,, \tag{3.3b}$$

$$\sigma_{12} = \left[\frac{\sum_{i=1}^{N} (V_2^i - (b_{12} V_1^i + a_{12}))^2}{N} \right]^{1/2} \,, \tag{3.3c}$$

where

$$\mu_1 := \frac{1}{N} \sum_{i=1}^{N} V_1^i \,, \tag{3.4a}$$

$$\mu_2 := \frac{1}{N} \sum_{i=1}^{N} V_2^i \,, \tag{3.4b}$$

$$\sigma_1 := \left[\frac{1}{N} \sum_{i=1}^{N} (V_1^i - \mu_1)^2 \right]^{1/2} \,, \tag{3.4c}$$

$$\mathrm{Cov}(V_1, V_2) := \frac{1}{N} \sum_{i=1}^{N} (V_1^i - \mu_1)(V_2^i - \mu_2) \,, \tag{3.4d}$$

and the quantities with 21 subscripts are found by changing $1 \leftrightarrow 2$

Now, the central object of this chapter, the *distance* $d(V_1, V_2)$ between two different instances of the same potential energy V is defined by

$$d(V_1, V_2) := \left(\sigma_{12}^2 + \sigma_{21}^2 \right)^{1/2} \,. \tag{3.5}$$

It must be stressed here that the measured distance between any given V_1 and V_2 depends on the working set, $\{\vec{q}_i\}_{i=1}^N$, of conformations chosen. More precisely, it depends on the occurrence probability of an arbitrary conformation $\vec{q}$ in the set. This probability must be decided a priori from considerations regarding which regions of the conformational space $\mathcal{I}$ are more relevant to answer the questions posed and up to what extent. For example, if one believes the system under study to be in thermodynamical equilibrium, then it would be reasonable to generate a working set in which the probability that $\vec{q}$ occurs is proportional to its Boltzmann weight ($e^{-\beta V}$). If, on the contrary, one doubts whether or not the system is really sampling the whole conformational space (like in the case of proteins) or one simply wants to study in detail the dynamical trajectories out of equilibrium, then, all the conformations should be weighted equally and the probability should be flat.

3.4 Meaning

Under the hypotheses made in sec. 3.2 (independence and eqs. (3.2)), a simple expression may be written for the probability density function of the V_2-energy differences ΔV_2 conditioned by the knowledge of the V_1-energy differences ΔV_1:

$$P_{\Delta 2|\Delta 1}(\Delta V_2 | \Delta V_1) = \frac{1}{\sqrt{2\pi} d_{12}} \exp\left[-\frac{(\Delta V_2 - b_{12} \Delta V_1)^2}{2 d_{12}^2} \right] \,, \tag{3.6}$$

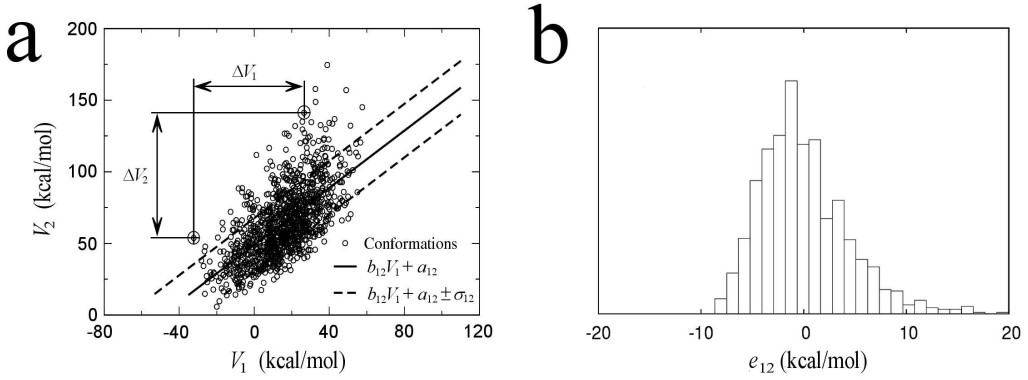


Figure 3.3: **(a)** V_1- and V_2-energies of the set of 1100 conformations of the Trp-Cage protein used in the first example of sec. 3.10. Both potentials are the van der Waals energy as implemented in CHARMM; V_1 corresponds to $R_C = 1.275$ Å and V_2 to $R_C = 3.275$ Å, ε_C is fixed to -0.020 kcal/mol. The values of ΔV_1 and ΔV_2 for a selected pair of conformations are also depicted. The solid line represents the least-squares fit and the region where the probability of finding a conformation is largest is enclosed by broken lines. **(b)** Histogram of the residues $e_{12}(\vec{q}_i) := V_2(\vec{q}_i) - [b_{12}V_1(\vec{q}_i) + a_{12}]$ associated to the least-squares fit in fig. (a).

where the quantity d_{12} appearing in this equation is defined as

$$d_{12} := \sqrt{2}\sigma_{12} \,, \tag{3.7}$$

it is related to the distance defined in eq. (3.5) via

$$d(V_1, V_2) = \left(\frac{d_{12}^2 + d_{21}^2}{2} \right)^{1/2} \,, \tag{3.8}$$

and it encodes the *loss of information* involved in the transit from V_1 to V_2 through the following important properties:

1. The addition of an energy reference shift a_{12} between V_1 and V_2 has neither an implication in the physical behaviour of the system nor in the numerical value of d_{12}.

2. One of the novel features of the distance herein defined is that no loss of information is considered to occur (i.e., $d_{12} = 0$) if there is only a constant rescaling b_{12} between the two potential energy functions studied. Although such a transformation does have physical implications and would change the transition rates in a molecular dynamics simulation and alter the effective temperature in any typical Monte Carlo algorithm, V_1 can be easily recovered from V_2, if pertinent, upon division of V_2 by b_{12}. For example, if the two potential energy functions are on equal footing (e.g., they correspond to different values of the free parameters (see sec. 3.6)), there is no 'correct' energy scale defined and the division by b_{12} is optional. However, in the case that the distance is used to compare an approximate potential to a more ab initio one or even to experimental data, the 'correct' energy scale must be considered to be that of the latter and the rescaling b_{12} may be safely removed as indicated above.

In such a case, note that the quantity d_{12} changes, when this removal is performed, from d_{12} to $d_{12}/|b_{12}|$[83] and it is the second value which must be considered as the relevant one.

3. Directly from its very definition in eq. (3.7), one has that $d_{12} = 0$ is equivalent to V_2 being exactly a linear transformation of V_1, i.e., $d_{12} = 0$ is equivalent to $V_2(\vec{q}_i) = b_{12}V_1(\vec{q}_i) + a_{12}, \forall \vec{q}_i \in \{\vec{q}_i\}_{i=1}^N$.

4. According to the previous points, $b_{12} \neq 1$ and/or $a_{12} \neq 0$ must be regarded as two different types of systematic errors easily removable and not involving any loss of information when one changes V_1 by V_2. In the general case, however, the *energy differences* associated to each potential energy function (which are the relevant physical quantities that govern the system behavior) present an additional random error which is intrinsic to the discrepancies between the potentials and cannot be removed. As can be seen in eq. (3.6), in this situation, d_{12} is the standard deviation of the random variable ΔV_2 and, as its value decreases, the distribution becomes sharper around the average $b_{12}\Delta V_1$. Moreover, because the distribution is normal, the probability of ΔV_2 being in the interval $(b_{12}\Delta V_1 - Kd_{12}, b_{12}\Delta V_1 + Kd_{12})$ is $\sim 38\%$ for $K = 1/2$, $\sim 68\%$ for $K = 1$, $\sim 95\%$ for $K = 2$, etc. Hence, d_{12} quantifies the random error between the trivially transformed potential $b_{12}V_1 + a_{12}$ and V_2, i.e., the unavoidable and fundamentally statistical part of the difference between V_1 and V_2 which stems from the complex character of the system.

5. To gain some additional insight about the meaning of d_{12}, the following *gedanken experiment* may be performed: if a Gaussian 'noise' with zero mean and variance equal to s^2 were independently added to the linearly transformed V_1-energy, $b_{12}V_1(\vec{q}) + a_{12}$, of each conformation and the resulting potential were denoted by V_2, then one would have that the hypothesis in eq. (3.2a) is fulfilled and that $d_{12} = \sqrt{2}s$. In such a case, d_{12} may be regarded (except for a harmless factor $\sqrt{2}$) as the size of the Gaussian noise arising in the whole energy landscape when one changes $b_{12}V_1 + a_{12}$ by V_2.

6. Closely related to the properties in the two preceding points, an illuminating statistical statement about the *energetic ordering* of the conformations can be derived from eq. (3.6). The probability that the energetic order of two randomly selected conformations is maintained when going from V_1 to V_2 (more precisely, that $\text{sign}(\Delta V_2) = \text{sign}(b_{12}\Delta V_1)$), conditioned by the knowledge of ΔV_1, can be easily shown to be

$$P_{\text{ord}}\left(\frac{|b_{12}\Delta V_1|}{d_{12}}\right) = \frac{1}{2} + \frac{1}{\sqrt{2\pi}}\int_0^{|b_{12}\Delta V_1|/d_{12}} \exp\left[-\frac{x^2}{2}\right] \mathrm{d}x \,. \tag{3.9}$$

The intuitive meaning of this expression is that d_{12} is the V_1-energy difference at which two randomly selected conformations can be typically 'resolved' using V_2 after the removal of the harmless rescaling b_{12} (see point 4). Certainly, if one has

[83] To see this, take the analogous for σ_2 of eq. (3.4c) and change V_2^i by V_2^i/b_{12}, then, take the result to eq. (3.12a) noting that r_{12} is invariant under the change

that $|b_{12}\Delta V_1| \ll d_{12}$, then $P_{\text{ord}} = 1/2$, reflecting a total lack of knowledge about the sign of ΔV_2 and, consequently, V_2 can not be used to resolve V_1-energy differences. If, on the contrary, $|b_{12}\Delta V_1| \gg d_{12}$, then $P_{\text{ord}} = 1$ and the conformations ordering is exactly conserved. At any intermediate point, P_{ord} is a rapidly increasing function of $|b_{12}\Delta V_1|/d_{12}$ that reaches a 'reasonable' value ($\sim 84\%$) when its argument equals 1, i.e., when $|b_{12}\Delta V_1| = d_{12}$. Some other interesting points are $P_{\text{ord}}(1/2) \simeq 69\%$ or $P_{\text{ord}}(2) \simeq 98\%$.

7. Finally, some clarifying properties of the distance which are associated to its relation to the Pearson's correlation coefficient will be investigated in sec. 3.7 (see specially eq. (3.12a)).

Now, we should note that the same considerations may be done about d_{21} regarding the transit from V_2 to V_1, and, as can be seen in eq. (3.8), the square of $d(V_1, V_2)$ is the mean of the squares of d_{12} and d_{21}. Therefore, the symmetrized distance $d(V_1, V_2)$ quantifies the average size of the uncertainty in the energy differences of the system that arises from changing one of the potentials studied by the other. If the comparison is performed between potential energy functions that stand on the same footing (see, for example, the second possible application in sec. 3.6), the symmetric quantity $d(V_1, V_2)$ should be used as a summarizing measure of the loss of information involved in the transit from V_1 to V_2 and vice versa. However, if one of the potentials is a priori considered to be more ab initio or more accurate (say, V_1) and it is compared to a less reliable instance (V_2), one may use simply d_{12} as the measure of the discrepancies between them[84].

Hence, although both the discussion regarding the relevant values of the distance in the following section and the investigation of its mathematical properties in secs. 3.8 and 3.9 are referred to $d(V_1, V_2)$ for generality, they may be equally applied to d_{12}. Conversely, the comparison between d_{12} and the quantities commonly used in the literature done in sec. (3.7) may be extended to $d(V_1, V_2)$ upon symmetrization of the mean error of the energy, ER, which is the only asymmetrical one.

3.5 Relevant values of the distance

Regarding the value of $d(V_1, V_2)$ in a practical case, some remarks must be made. To begin with, one may expect two special values of the distance to exist: d_{min} and d_{max}. In such a way that, if $d(V_1, V_2) < d_{\text{min}}$, one potential energy function may be substituted by the other without altering the key characteristics of the system behavior, and that, if $d(V_1, V_2) > d_{\text{max}}$, then, the substitution is not acceptable. These limit values must be set depending on the particularities of the system studied and on the questions sought to be answered, and it may even be the case that some special features of the energy landscape are the main responsible of the behaviour under scrutiny. For example, we are not going to establish any strict limit on the accuracy required for a potential energy function to successfully predict the folding of proteins [2, 137]. We consider this question a difficult theoretical issue, whose solution probably requires a much deeper knowledge of the protein folding problem itself than the one that exists at present, and we believe that it may be possible a priori that some special features of the energy landscapes of proteins (such

[84] Note, from eq. (3.8), that, if $d_{12} = d_{21}$, then $d(V_1, V_2) = d_{12} = d_{21}$.

as a funnel-like shape [130, 280], see sec. 1.4) are the main responsible of the high efficiency and cooperativity of the folding process [2, 137]. If this were the case, a different procedure for measuring the distance between potential energy functions could be devised for this situation [281–283], as any change of V_1 by V_2 which did not significantly alter these special features could be valid even if the value of $d(V_1, V_2)$ were very large. Our definition of $d(V_1, V_2)$, being based in characteristics shared by many complex systems and statistically referred to the whole energy landscape, is of more general application but cannot detect such particular characteristics as the ones mentioned.

However, thanks to the laws of statistical mechanics, a rather stringent but general value for $d_{\min}$ can be used to a priori assess the interchangeability of V_1 and V_2. As can be seen in the thermodynamical equilibrium Boltzmann distribution, in which the probability $p(\vec{q}_i)$ of a conformation $\vec{q}_i$ is proportional to $\exp(-V(\vec{q}_i)/RT)$[85], the order of the physical uncertainty in the potential energies of a system in contact with a thermal reservoir at temperature T is given by the quantity RT. This typical energy sets the scale of the thermal fluctuations and it also determines the transition probability, $\min[1, \exp(-(V(\vec{q}_{i+1}) - V(\vec{q}_i))/RT)]$, in the Metropolis Monte Carlo scheme and the spread of the stochastic term in the Langevin equation [284–286]. Consequently, in this dissertation, the quantity RT (which equals ~ 0.6 kcal/mol at room temperature) will be used as a general lower bound for $d_{\min}$. The results will be presented in units of RT and any two instances V_1 and V_2 of the same potential energy function whose distance $d(V_1, V_2)$ be smaller than RT will be regarded as physically equivalent[86].

Regarding d_{max}, no estimations of its value can a priori be made without referring to the particular potential energy functions compared and the relevant behaviour studied. The fact that eq. (3.12b) has an absolute maximum when $r_{12} = 0$ sets only the worst possible upper bound and is only of mathematical interest.

3.6 Possible applications

There are at least three basic situations in which the distance defined in this chapter may be used to quantify the discrepancies between two different instances, V_1 and V_2, of the same potential energy:

- If the difference between V_1 and V_2 arises from the use of two distinct algorithms or approximations, $d(V_1, V_2)$ (or d_{12}, see the final lines of sec. 3.4) may help us decide whether the less numerically complex instance could be used or not. For example, one may compare the electrostatic part of the solvation energy calculated via solving the Poisson equation [288–291] with the instance of the same energy calculated using one of the many implementations of the Generalized Born model [260, 263–266, 268, 292, 293], which are much less computationally demanding and more suitable for simulating macromolecules. If the distance between them is

[85] We are assuming here either that the coordinates $\vec{q}$ are Euclidean or that the Jacobian determinant of the change from the Euclidean coordinates to the $\vec{q}$ is slowly varying and can be neglected in this qualitative discussion. See appendix A and chapter 6 for further details regarding this issue.

[86] This discussion is closely related to the common use of the concept of *chemical accuracy*, typically defined in the field of ab initio quantum chemistry as predicting bond-breaking energies to 1 kcal/mol [287].

small enough for the behaviour under study not to be much modified (see sec. 3.5), then the latter could be used.

Among the three possible applications proposed in this section, this one is, not only the most commonly found one in the literature, but also the most used one in the rest of the chapters of this dissertation; showing that the distance defined here is the natural and necessary first step in the long path towards the design of better energy functions that we have decided to walk.

With an illustrative purpose, in the second example of sec. 3.10, several levels of quantum chemistry theory are compared in the model dipeptide HCO-L-Ala-NH_2 (see appendix E) using the distance, and in chapter 7, a much more exhaustive study of the same kind is performed. Finally, the study in chapter 6 about the possibility of neglecting the mass metric tensor determinants that show up when constraints are imposed has largely benefited from the properties and physical meaning of the distance described in this chapter. All these applications are examples of this point.

- If the algorithm and the approximations A are fixed and only one system S is studied, any reasonable functional form used to account for V will be a simplified model of physical reality and it will contain a number of free parameters $\vec{P}$. These parameters, which, in most of the cases, are not physically observable, must be fit against experimental or more ab initio results before using the function for practical purposes. For any fit to yield statistically significant values of the parameters, the particular region of the parameter space in which the final result lies must have the property of *robustness*, i.e., it must occur that, if the found set of parameters values is slightly changed, then, the relevant characteristics of the potential energy function which depends on them are also approximately kept unmodified. If this were not the case, a new fit, performed using a different set of experimental (or more ab initio) points, could produce a very distant potential. If V_1 and V_2 come from the same family of potential energy functions and they correspond to different values of the free parameters $\vec{P}$, the distance $d(V_1, V_2)$ between them may help us to assess the robustness of the potentials under changes on the parameters.

  In the first example of sec. 3.10, the robustness of the van der Waals potential energy implemented in the well-known molecular dynamics program CHARMM [104, 105] is quantified as an example of this.

- The last application of the distance is fundamentally different from the ones in the previous points but, although the reasoning throughout this chapter is intentionally biased, for the sake of clarity, toward the study of potential energies of the same system, one may appeal to the same underlying assumptions to compare two different systems, S_1 and S_2, provided that a meaningful mapping can be established between both conformational spaces[87]. For example, if the conformations of a particular protein are described only by their backbone angles, one can define an unambiguous mapping between the conformations of, say, the wild-type sequence and any mutated form, in such a way that $V_1(\vec{q})$ would represent the energies of the

[87] In short, for the distance criterion to be applied, one needs to be able to assign two energies, $V_1(\vec{q})$ and $V_2(\vec{q})$, to each conformation $\vec{q}$. This is done trivially in the first two points but it requires a mapping between the conformational spaces of S_1 and S_2 in the third case.

former and $V_2(\vec{q})$ those of the latter. The distance $d(V_1, V_2)$, in this case, quantifies how different the energy landscapes of the two systems are and, depending on the features under study, how sensitive the behaviour of the protein is to mutations.

The comparison of a potential, V_1, to another one, V_2, which comes from the first via integrating-out certain degrees of freedom (see sec. 1.4), may be considered to be another example of this type of application. The study performed in sec. 4.4 of chapter 4 is a clear example of this.

3.7 Relation to other statistical quantities

In the literature, some comparisons between potentials[88] are performed a posteriori, i.e., not directly studying the energies but computing some derived quantities, such as the pK_a of titratable groups [265, 266], investigating molecular dynamics trajectories [260, 262, 271, 275], comparing the ability of the different instances of V to select the correct native state of a protein from a set of decoys [262], etc.

For the a priori comparison of two ways of calculating the same potential energy, one may investigate the whole energy landscape visually if the system has no more than two degrees of freedom [269, 270], but, if the object of study is a protein or another complex system, the vastness of the conformational space (see the discussion about Levinthal's paradox in sec. 1.4) and its lack of symmetries require the utilization of statistical quantities calculated from the energies of a finite set of conformations. Among the most common such measures, one may find the *root mean square deviation* (RMSD) [260, 263, 270, 271, 273], the *mean energy error* (ER) [259, 272, 273], the *standard deviation of the energy error* (SDER) [272], the *mean of the absolute energy error (AER)* [264], all of which have units of energy, and the *Pearson's correlation coefficient r* [207, 259, 261, 268, 271, 272, 274], which does not have units. Finally, in ref. 263, a *root mean square of the difference in the relative energies* (REL) (see eq. (3.10e) for a clarification) which has several points in common with d_{12} (see eq. (3.7)) is defined, however, it has not been found to be used in any other work.

If we use the same notation as in eqs. (3.3) and we define $\Delta V_{12}^i := V_2^i - V_1^i$, the statistical quantities mentioned in the preceding paragraph (except r, which will be discussed later) are given by the expressions:

$$\mathrm{RMSD}(V_1, V_2) := \left[\frac{1}{N} \sum_{i=1}^{N} (\Delta V_{12}^i)^2 \right]^{1/2} , \tag{3.10a}$$

$$\mathrm{ER}(V_1, V_2) := \frac{1}{N} \sum_{i=1}^{N} \Delta V_{12}^i , \tag{3.10b}$$

[88] It must be pointed out that we have only found in the literature examples of the comparison between two potentials corresponding to the first case described in sec. 3.6, which is associated to different algorithms or approximations. No examples of robustness studies have been found and, regarding the third case, in which the differences arise from a slight change in the system, such as a mutation in a protein, only articles investigating the total free energy of folding have been found [294, 295].

$$\mathrm{SDER}(V_1, V_2) := \left[\frac{1}{N}\sum_{i=1}^{N}(\Delta V_{12}^i - \mathrm{ER}(V_1, V_2))^2\right]^{1/2}, \tag{3.10c}$$

$$\mathrm{AER}(V_1, V_2) := \frac{1}{N}\sum_{i=1}^{N}|\Delta V_{12}^i|\,, \tag{3.10d}$$

$$\mathrm{REL}(V_1, V_2) := \left[\frac{2}{N(N-1)}\sum_{i=1}^{N}\sum_{j=i+1}^{N}(\Delta V_{12}^j - \Delta V_{12}^i)^2\right]^{1/2}. \tag{3.10e}$$

In the following points, these measures of the difference between potential energy functions are individually compared to the distance defined in sec. 3.3 and their limitations with respect to d_{12} are pointed out[89]:

- The first one, the RMSD, which is one of the most commonly used, presents the major flaw of overestimating the importance of an energy reference shift between V_1 and V_2. This transformation, which has no physical implications in the conformational behaviour of the system, must not influence the assessment of the difference between potentials. This fact is, for example, detected in some of the comparisons performed in ref. 271 and recognized to be conceptually erroneous in ref. 273, where the shift is removed by minimizing the RMSD with respect to it. In addition, the RMSD also overestimates the effect of a slope $b_{12} \neq 1$ between the two potentials, a fact that, as it has been remarked in sec. 3.4, is not desirable (for a practical case in which the loss of information is small but $b_{12} \neq 1$ see ref. 259; for a numerical example see fig. 3.4 and the discussion at the end of this section). It can be proved that, if $b_{12} = 1$ and $a_{12} = 0$, then $\mathrm{RMSD}(V_1, V_2) = d_{12}/\sqrt{2}$.

- ER, in turn, only accounts for a systematic error between the two potentials (an offset). The relation $\mathrm{ER}(V_1, V_2) = \mu_2 - \mu_1$ holds and ER equals the energy-reference shift a_{12} if $b_{12} = 1$ (see eq. (3.3b)). Thus, the changes in the conformational behaviour of the system are not reflected by this quantity.

- In SDER, which is the standard deviation associated to ER, the reference shift is removed by subtracting ER from each difference ΔV_{12}^i. However, this quantity still overestimates the effect of a slope $b_{12} \neq 1$, in fact, only if $b_{12} = 1$, one has that $\mathrm{SDER}(V_1, V_2) = d_{12}/\sqrt{2}$.

- To establish precise relations between AER and d_{12} is difficult because of the modulus function that enters this quantity. Nevertheless, it is clear from its definition that AER, like ER, overestimates both the effect of an energy reference shift a_{12} and of a slope $b_{12} \neq 1$. For a numerical confirmation of this fact, see table. 3.1.

- Finally, the measure REL, introduced in ref. 263, has much of the spirit of the distance defined in this work. On one hand, it focuses on the energy differences, which are indeed the relevant physical quantities to study the conformational behaviour of the system, on the other hand, it correctly removes the effect of an energy reference

[89] The quantity ER is not symmetric. This is why all the measures in eq. (3.10) are compared to d_{12} and not to its symmetrized version $d(V_1, V_2)$ (see sec. 3.9 and the final part of sec. 3.4).

shift a_{12}. However, it still overestimates the importance of a slope $b_{12} \neq 1$ and one only has that $\mathrm{REL}(V_1, V_2) = d_{12}$ if $b_{12} = 1$.

Apart from the ones defined above, there is yet another quantity commonly used for measuring the differences between two potentials: the Pearson's correlation coefficient (denoted by r_{12} in the following):

$$r_{12} := \frac{\mathrm{Cov}(V_1, V_2)}{\sigma_1 \sigma_2} \,. \tag{3.11}$$

This statistical measure differs from the rest in several points. On one hand, it has no units; a fact that renders difficult to extract from its value relevant information about the energies studied. While some statistical statements can be made about the real value of r_{12} (the value in an infinite sample, denoted by ρ_{12}), to do this, the sampling distribution of r_{12} must be known. Without making stringent assumptions about the joint probability density $P_{12}(V_1, V_2)$ (see sec. 3.2) only the null hypothesis of ρ_{12} being equal to 0 can be rejected from the knowledge of r_{12} in a finite sample [296]. This is clearly insufficient, because, in the vast majority of the cases, the researcher *knows* that the two potentials are correlated, i.e., the null hypothesis can be easily rejected from a priori considerations. If, in turn, one assumes $P_{12}(V_1, V_2)$ to be bivariate Gaussian, the Fisher transformation can be used to make inferences about ρ_{12} which are more general than $\rho_{12} \neq 0$ [296]. In any case, unfortunately, these type of statements are not directly translated into statements regarding the energies; a fact that undermines much of the physical meaning in r_{12}.

On the other hand and despite the disadvantages remarked in the preceding lines, r_{12} behaves satisfactorily when an energy reference a_{12} is added or when a rescaling $b_{12} \neq 1$ is introduced between V_1 and V_2. Like d_{12}, and in contrast to RMSD, ER, SDER, AER and REL, the Pearson's correlation coefficient does not overestimate such transformations, in fact, r_{12} is completely insensitive to them. Therefore, it is not surprising that a simple general relation can be written between r_{12} and d_{12}:

$$d_{12} = \sqrt{2}\sigma_2 (1 - r_{12}^2)^{1/2} \,, \tag{3.12a}$$

$$d(V_1, V_2) = [(\sigma_1^2 + \sigma_2^2)(1 - r_{12}^2)]^{1/2} \,. \tag{3.12b}$$

In these expressions, it can be observed that the distance herein introduced depends on two factors: on one side, the width of the probability distributions associated to the potentials (σ_1 and σ_2), which set the physical scale and give energy units to $d(V_1, V_2)$, on the other, the quantity $1 - r_{12}^2$, which measures the degree of correlation between V_1 and V_2. The second factor is completely insensitive to a change in the energy reference shift or in the slope (due to the properties of r_{12}); the part that depends on the width of the distributions, in turn, makes the distance sensitive to a change in the slope (remaining insensitive to a change in the reference), through σ_2 if the rescaling is performed on V_2 ($\sigma_2 \to \sigma_2/|b_{12}|$). However, contrarily to the case of the quantities in eqs. (3.10), the implications of such a transformation are not overestimated. In the case of our distance, the sensitivity to a rescaling arises only from the dilatation of the random errors, whereas the other quantities take erroneously into account the fact that the best fit line is not necessarily parallel to the line $V_2 = V_1$ (see below).

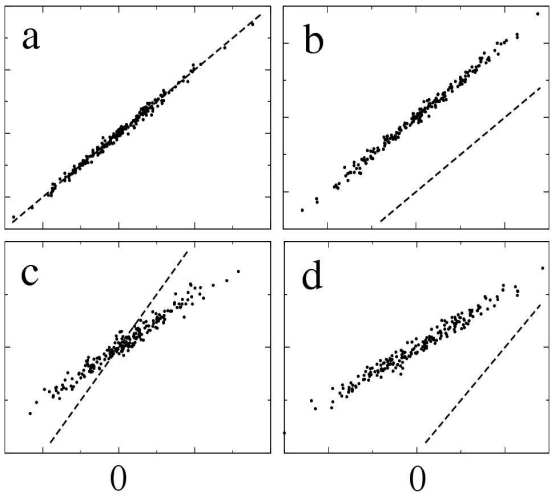


Figure 3.4: Numerical examples of the possible situations found in practical problems. 200 conformations are depicted for each case with the values of V_1 in the x-axis and the ones of V_2 in the y-axis (both in arbitrary energy units). The broken line corresponds to the line $V_2 = V_1$. **(a)** $b_{12} \simeq 1$ and $a_{12} \simeq 0$. **(b)** $b_{12} \simeq 1$ and $a_{12} \simeq 200$. **(c)** $b_{12} \simeq 1/2$ and $a_{12} \simeq 0$. **(d)** $b_{12} \simeq 1/2$ and $a_{12} \simeq 200$.

In short, *the distance defined in this chapter consists of a physically meaningful way of giving energy units to the Pearson's correlation coefficient*. A property that is very convenient if one wants to validate potential energy functions.

To close this section, a numerical example is presented that summarizes the situations that may be found in practical examples (see ref. 259 for a real case of the issues raised) and that makes explicit the aforementioned disadvantages of the commonly used statistical quantities. In fig. 3.4, four samples of 200 conformations are depicted with the values of V_1 in the x-axis and the ones of V_2 in the y-axis (both in arbitrary energy units). The different situations correspond to all generic cases in which $a_{12} = 0$ or $a_{12} \neq 0$ and in which $b_{12} = 1$ or $b_{12} \neq 1$. All the quantities discussed in this section, including d_{12}, have been computed in each case and their values are presented in table. 3.1.

From these data and the preceding discussion, some conclusions may be extracted. First, among the quantities with energy units, SDER and REL are the most proximate

		RMSD	ER	SDER	AER	REL	r_{12}	d_{12}
$b_{12} \simeq 1$	$a_{12} \simeq 0$	9.6	-0.7	9.6	7.7	13.6	0.995	13.5
$b_{12} \simeq 1$	$a_{12} \simeq 200$	199.8	199.6	9.4	199.6	13.3	0.996	13.2
$b_{12} \simeq 1/2$	$a_{12} \simeq 0$	52.7	-8.4	52.0	41.9	73.8	0.980	14.4
$b_{12} \simeq 1/2$	$a_{12} \simeq 200$	205.0	198.0	52.9	198.0	74.9	0.985	13.2
Overestimates $a_{12} \neq 0$		Yes	Yes	No	Yes	No	No	No
Overestimates $b_{12} \neq 1$		Yes	Yes	Yes	Yes	Yes	No	No
Has units of energy		Yes	Yes	Yes	Yes	Yes	No	Yes

Table 3.1: Values of the statistical quantities RMSD, ER, SDER, AER, REL, r_{12} and d_{12} computed in the situations depicted in fig. 3.4. All the values are in arbitrary energy units except the ones of r_{12} which have no units. A summary of the properties of each quantity is presented in the bottom part of the table.

to the distance d_{12}, although they will overestimate the difference between potentials in situations in which there is a constant rescaling $b_{12} \neq 1$ between them. In fig. 3.4c, for example, the contribution of the points that lie further apart from the origin of coordinates is overestimated by all the quantities in eqs. (3.10) for the sole fact that the best fit-line and the line $V_2 = V_1$ are not parallel (note that $a_{12} = 0$ and that the random noise associated to these points is not particularly large compared to the one that corresponds to the points in the central region of the figure). This is due to the fact that all quantities in eqs. (3.10) are based on ΔV_{12}^i, which measures the distance of each point to the line $V_2 = V_1$. A disadvantage that is not shared by d_{12}, which measures the differences with respect to the best-fit line.

Second, the Pearson's correlation coefficient r_{12} has good properties, although no physically relevant statements can be extracted from its value due, among other reasons, to the fact that it does not have units. In all cases in table. 3.1, for example, the value of r_{12} is close enough to 1 to be considered as a sound sign of correlation, however, the value of d_{12} (if we pretend it to be in kcal/mol, which could be the case) tells us that the typical indetermination in the energy differences, when substituting V_1 by V_2, is around 13 kcal/mol, a value an order of magnitude larger than RT. As it is explained in sec. 3.5, this suggests that the relevant behaviour of the system may be essentially modified.

Finally, it is worth stressing that all the considerations made in this section and throughout the chapter are valid when the physical quantities compared are potential energy functions of the same system or closely related systems (see sec. 3.6). When other quantities, such as the pK_a, charges, dipoles, Born radii, etc. or energies of distinct systems are the object of the comparison, the assessment of the discrepancies rests on different theoretical basis and, frequently, only semi-quantitative statements can be made. Acknowledging this limitation, the use of any of the quantities studied in this section, *including* d_{12}, may be fully justified in such cases. Note, in addition, that the numerical effort needed for the calculation of d_{12} is both low and very similar to the one required to compute any of the other quantities (see sec. 3.11).

3.8 Additivity

Frequently, the potentials compared are instances of only a part of the total potential energy of the system. If the conclusions extracted, via $d(V_1, V_2)$, in such a case are pretended to be meaningfully transferred to the total energy, this measure of the difference between potentials must obey some reasonable additivity rules. Here, we will see that, for some relevant cases in which certain independence hypotheses are fulfilled, our distance is approximately additive, although, in other relevant situations, it is not.

For the sake of brevity, the notation will be much relaxed in this section and we will assume that we are working with six different potentials, x, y, p, r, q and s, that satisfy the following relations:

$$x = p + q \qquad \text{and} \qquad y = r + s\,. \tag{3.13}$$

Conceptually, x and y must be regarded as instances of the same potential energy and the same can be said about the pair p and r and the pair q and s. Hence, the study of the additivity of our distance rests on finding a way of expressing $d(x, y)$ as a function of

$d(p, r)$ and $d(q, s)$. If one assumes that p is independent from q and that r is independent from s (see the discussion at the end of this section for the implications of such an hypothesis), one has that $r_{pq} = 0$, $r_{rs} = 0$, $\sigma_x^2 = \sigma_p^2 + \sigma_q^2$ and $\sigma_y^2 = \sigma_r^2 + \sigma_s^2$. In such a case, the following additivity relation can be written:

$$d^2(x, y) = d^2(p, r) + d^2(q, s) + \Delta d \,, \tag{3.14}$$

where

$$\Delta d := (\sigma_p^2 + \sigma_r^2)(r_{pr}^2 - r_{xy}^2) + (\sigma_q^2 + \sigma_s^2)(r_{qs}^2 - r_{xy}^2) \,, \tag{3.15}$$

and the correlation coefficient r_{xy} can be expressed in terms of quantities associated to p, r, q and s in the following way (note that r_{xy} is indeed not additive):

$$r_{xy} = \frac{\sigma_p \sigma_r r_{pr} + \sigma_q \sigma_s r_{qs}}{(\sigma_p^2 + \sigma_q^2)^{1/2}(\sigma_r^2 + \sigma_s^2)^{1/2}} \,. \tag{3.16}$$

Now, one can see in eq. (3.14) that, if Δd were zero, the square of the distance would be exactly additive in the aforementioned sense, making it possible to assert, for example, that, if p is proximate to r and q is proximate to s, then $x = p + q$ is proximate to $y = r + s$. Unfortunately, this is not the case. It can be shown that $\Delta d \geq 0$ (the distance is *over-additive*) and, without imposing any restriction on the potentials studied, nothing satisfactory can be said in addition to that. For example, a particularly undesirable, albeit also uncommon, situation is that for which $\mathrm{Cov}(p, r) = -\mathrm{Cov}(q, s)$. Such a relation, makes zero the numerator in eq. (3.16) and, consequently, r_{xy}. Substituting $r_{xy} = 0$ in eq. (3.15) and taking Δd to eq. (3.14), one has that, for every allowed value of r_{pr} and r_{qs},

$$\mathrm{Cov}(p, r) = -\mathrm{Cov}(q, s) \Rightarrow d^2(x, y) = \sigma_p^2 + \sigma_q^2 + \sigma_r^2 + \sigma_s^2 = \sigma_x^2 + \sigma_y^2 \,, \tag{3.17}$$

which is the worst possible value of $d(x, y)$.

However, there exists a particular class of situations than can be argued to be proximate to the situations found in typical cases and for which the additivity is approximately accomplished. These special situations are characterized for the satisfaction of the following relation:

$$\sigma_p/\sigma_r = \sigma_q/\sigma_s := k \,. \tag{3.18}$$

When this equality is satisfied, it can be proved that the following quotient:

$$\Delta d_{\mathrm{rel}} := \Delta d/(d^2(p, r) + d^2(q, s)) \,, \tag{3.19}$$

which measures the relative deviation from the exact additivity, does not depend on k and can be expressed as a function of only σ_r, σ_s, r_{pr} and r_{qs}. If, in addition, we define c through $\sigma_s = c\,\sigma_r$, without loss of generality, we can write Δd_{rel} as a function of only r_{pr}, r_{qs}, and c as follows:

$$\Delta d_{\mathrm{rel}} = \frac{c^2(r_{pr} - r_{qs})^2}{(1 + c^2)(1 - r_{pr}^2 + c^2(1 - r_{qs}^2))} \,. \tag{3.20}$$

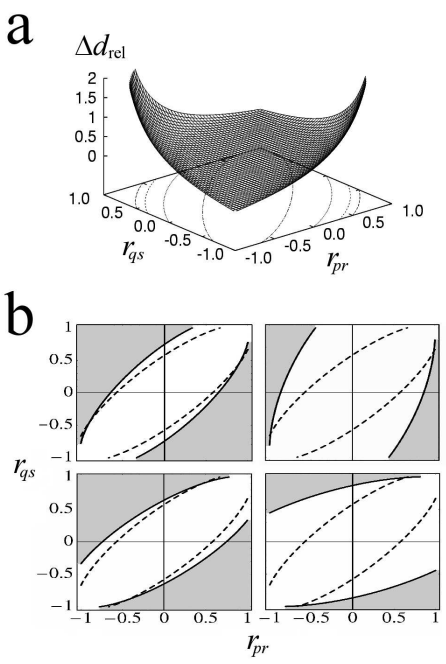


Figure 3.5: Graphical study of the additivity of the distance. **(a)** $\Delta d_{\rm rel}$ as a function of r_{pr} and r_{qs} for $c = 1$ (see eq. (3.20)). Contour lines are plotted at the levels $\Delta d_{\rm rel} = 10\%, 50\%, 100\%, 150\%$. **(b)** In white, the regions in (r_{pr}, r_{qs})-space with $\Delta d_{\rm rel} < 10\%$ for different values of c. From left to right and from top to bottom, each figure corresponds to $c = 1/2, 1/5, 2, 5$. In each case, the borders of the $\Delta d_{\rm rel} < 10\%$ region for $c = 1$ are shown with broken lines for comparison.

Representing this equation as a three-dimensional surface (see fig. 3.5a), one can see a 'valley' whose lowest region lies in the line $r_{pr} = r_{qs}$ and has zero height, i.e., $\Delta d_{\rm rel}(r_{pr} = r_{qs}) = 0$. The slopes of the valley are curved and ascend as one moves away from the minimum height line, eventually reaching arbitrarily large values of $\Delta d_{\rm rel}$ when $(r_{pr}, r_{qs}) \to (1, -1)$ or $(r_{pr}, r_{qs}) \to (-1, 1)$.

Numerically, the region for which the value of $\Delta d_{\rm rel}$ is acceptable is rather large. In fig. 3.5b, the contour lines corresponding to $\Delta d_{\rm rel} = 10\%$ are depicted for some values of c that may be found in practical cases. It can be seen that, as one departs from $c = 1$, the region for which $\Delta d_{\rm rel} < 10\%$ gets larger, occupying, in any case, the majority of the (r_{pr}, r_{qs})-space. Therefore, we conclude that, for the cases in which eq. (3.18) is satisfied, the square of the distance introduced in this chapter is approximately additive in the relevant situations in which the correlations between p and r and between q and s are similar. Of course, for continuity arguments, one has that, in the case that eq. (3.18) were only approximately satisfied, the situation would be proximate to the one described above.

Additionally, let us point out that, if the calculations above are repeated for the unsymmetric version d_{xy} of the distance, the same expression for $\Delta d_{\rm rel}$ is obtained, and, therefore, all the conclusions drawn in this section about $d(x, y)$ are applicable to d_{xy}.

Finally, some remarks must be made about the assumption of independence between p and q and between r and s. At first sight, one would say that this hypothesis, like the independence hypothesis in sec. 3.2, is under researcher's control. In the case of a generic complex system (a spin glass, a random heteropolymer, etc.), this is indeed the case, however, if the object of study is a protein, one must be cautious. It is widely believed that the sequences of proteins are the result of a million-years-long selection process whose driving force is the search for the ability to fold rapidly and robustly (see sec. 1.4). Regarding the interactions responsible of the folding process, this means that they have been optimized in the sequence space to be *minimally frustrated*, i.e., maximally cooperative. In such a case, the correlations between different parts of the total potential

energy may be large and the study of the additivity done in this section should be regarded only as a privileged reference situation.

3.9 Metric properties

For completeness, and because, in the case of our distance, it is illustrative to do so, we will investigate, in this section, in which situations (which will turn out to be rather common) the behaviour of $d(V_1, V_2)$ approaches that of a traditional mathematical *distance*. Nevertheless, it must be stressed that the measure introduced in this chapter was never intended to be such an object. Its meaning is encoded in the statistical statements derived from its value (see sec. 3.4) and the name 'distance' must be used in a more relaxed manner than the one traditionally found in mathematics.

The object $\mathcal{D}(x, y)$ is said to be a *distance* (also a *metric*) in mathematics if it satisfies the following properties:

1. $\mathcal{D}(x, y) = 0 \Leftrightarrow x = y$
2. Positivity: $\mathcal{D}(x, y) \geq 0$
3. Symmetry: $\mathcal{D}(x, y) = \mathcal{D}(y, x)$
4. Triangle inequality: $\mathcal{D}(x, z) \leq \mathcal{D}(x, y) + \mathcal{D}(y, z)$

Whereas, in the case of $d(V_1, V_2)$:

1. The first property is not fulfilled. One certainly has the implication to the left, but the direct implication is false in general. As it has been stated in point 3, in sec. 3.4, the analogous property that $d(V_1, V_2)$ satisfies is that $d(V_1, V_2) = 0$ is equivalent to V_2 being a linear transformation of V_1 and vice versa, i.e., to $V_2(\vec{q}_i) = b_{12}V_1(\vec{q}_i)+a_{12}$ and $V_1(\vec{q}_i) = b_{21}V_2(\vec{q}_i) + a_{21}$, $\forall \vec{q}_i \in \{\vec{q}_i\}_{i=1}^N$, where, additionally, one has that $b_{21} = 1/b_{12}$ and $a_{21} = -a_{12}/b_{12}$. The fact that this property of a mathematical distance is not satisfied by $d(V_1, V_2)$ must be considered an advantage, because, as it has been remarked in previous sections, it is reasonable to regard as equivalent two potentials if there is only a linear transformation between them.
2. $d(V_1, V_2) \geq 0$ for every V_1 and V_2.
3. Directly from its definition in eq. (3.5) or from eq. (3.8), it is evident that $d(V_1, V_2)$ is symmetrical under change of V_1 by V_2. On the other hand, although this property is not fulfilled by the quantity d_{12}, the situation in which it is reasonable to use it (the comparison of a particular instance of a potential energy V to a less accurate one) is also intrinsically asymmetrical (see the final part of sec. 3.4).
4. The triangle inequality, in this context, is a relation that must be expressed as a function of the statistical quantities related to three different potentials, V_1, V_2 and V_3, as follows:

$$\sqrt{\sigma_1^2 + \sigma_3^2}\sqrt{1 - r_{13}^2} \leq \sqrt{\sigma_1^2 + \sigma_2^2}\sqrt{1 - r_{12}^2} + \sqrt{\sigma_2^2 + \sigma_3^2}\sqrt{1 - r_{23}^2}\,. \tag{3.21}$$

This relation is not fulfilled for every triplet (V_1, V_2, V_3), i.e., the distance introduced in this chapter does not satisfy, in general, the triangle inequality. A simple counterexample is found if one makes σ_3 grow, keeping the rest of the quantities in eq. (3.21) constant. For σ_3 large enough, the relation above may be approximated by

$$\sigma_3 \left[\sqrt{1 - r_{13}^2} - \sqrt{1 - r_{23}^2} \right] \leq \sqrt{\sigma_1^2 + \sigma_2^2} \sqrt{1 - r_{12}^2} \, . \tag{3.22}$$

Then, if $r_{13}^2 < r_{23}^2$, one may make σ_3 even larger and eventually break the inequality (in the case that it were not already broken for the value of σ_3 for which eq. (3.22) is a good approximation).

As a final remark, it is worth pointing out that, despite the general mathematical facts stated above, there is a relevant situation in which the distance has been found to satisfy the triangle inequality. If one has that $\sigma_1 = \sigma_2 = \sigma_3$ (something that is expected to be approximately true in the case that the three potentials are proximate), eq. (3.21) turns into a relation involving only the correlation coefficients:

$$\sqrt{1 - r_{13}^2} \leq \sqrt{1 - r_{12}^2} + \sqrt{1 - r_{23}^2} \, . \tag{3.23}$$

In addition, assuming the hypotheses discussed in sec. 3.2, the following inequalities can be proved [296] without any further assumptions about the potentials:

$$r_{13} \geq r_{12} r_{23} - \sqrt{1 - r_{12}^2} \sqrt{1 - r_{23}^2} \, , \tag{3.24a}$$

$$r_{13} \leq r_{12} r_{23} + \sqrt{1 - r_{12}^2} \sqrt{1 - r_{23}^2} \, . \tag{3.24b}$$

We have numerically found that, if the relations in eqs. (3.24) are satisfied, so is the one in eq. (3.23). Hence, if $\sigma_1 = \sigma_2 = \sigma_3$, then, for all values of r_{12}, r_{23} and r_{13}, the distance satisfies the triangle inequality. Clearly, for continuity, if $\sigma_1 = \sigma_2 = \sigma_3$ is not exactly but approximately satisfied, then, although the triangle inequality may be broken, it will broken by a small relative amount.

3.10 Practical examples

To illustrate one of the possible practical applications of the distance, we first study the robustness of the van der Waals energy, as implemented in the CHARMM molecular dynamics program [104, 105], in a particular system: the de novo designed 20-residue protein known as *Trp-Cage* [297] (PDB code: 1L2Y).

The program CHARMM itself was used as a conformation generator. From the native conformation stored in the Protein Data Bank [41], a 10 ps heating dynamics[90] was performed on the system, from an initial temperature $T_\mathrm{i} = 0$ K to eleven different final temperatures (from $T_\mathrm{f} = 500$ K to $T_\mathrm{f} = 1000$ K in steps of 50 K). This was repeated 100

[90] The `c27b4` version of the CHARMM program was used. The molecular dynamics were performed using the *Leap Frog* algorithm therein implemented and the `param22` parameter set, which is optimized for proteins and nucleic acids. The water was taken into account implicitly with a specific version of the *Generalized Born Model* that is built into the program [260].

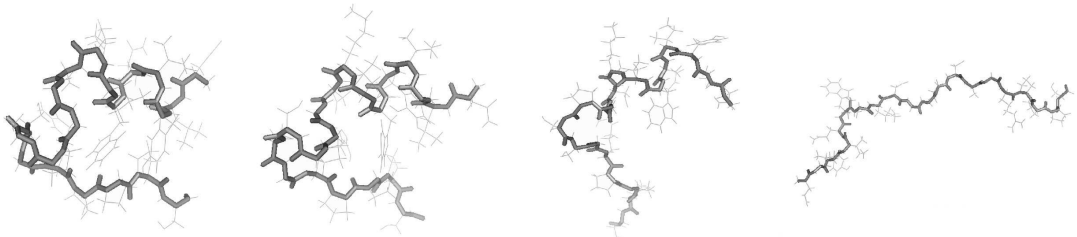


Figure 3.6: Native conformation of the Trp-Cage protein together with arbitrarily chosen structures from three particular subsets of the working set. The average radius of gyration $\langle R_G \rangle$ and the average RMSD with respect to the native structure is presented for each set. Both quantities have been computed taking into account only the α-carbons. Pictures generated with PyMOL (DeLano, W. L., 2002, http://www.pymol.org).

times for each final temperature with a different seed for the random numbers generator each time. The overall result of the process was the production of a working set of 1100 different conformations of the protein, whose structures range from 'close to native' (the T_f = 500 K set) to 'completely unfolded' (the T_f = 1000 K set) (see fig. 3.6). It is worth remarking that the short time in which the system was heated (10 ps) and the fact that there was no equilibration after this process cause the final temperatures to be only *labels* for the eleven aforementioned sets of conformations. They are, by no means, the thermodynamical temperatures of any equilibrium state from which the structures are taken. These sets of conformations are only meant to reasonably sample the most representative regions of the conformational space.

In fig. 3.6, arbitrarily chosen structures from three particular sets are shown together with the native conformation. The average radius of gyration $\langle R_G \rangle$ of each set, depicted in the same figure, must be compared to the radius of gyration of the native state[91]. The average RMSD of the structures in each set with respect to the native structure, calculated via the quaternion-based method described in ref. 298, is also presented[92].

The van der Waals energy implemented in CHARMM may be expressed as follows:

$$V := \sum_{i<j} \left[(\varepsilon_i \varepsilon_j)^{1/2} \left(\left(\frac{R_i + R_j}{r_{ij}} \right)^{12} - 2 \left(\frac{R_i + R_j}{r_{ij}} \right)^{6} \right) \right] , \tag{3.25}$$

where the sum is extended to all the pairs of atoms and the free parameters ε_i and R_i only depend on the type of atom (i.e., two atoms i and j of the same type have the same parameters assigned).

Using the working set of conformations of the Trp-Cage protein described above, the

[91] Both R_G and the RMSD have been computed taking into account only the α-carbons.

[92] The notation for this quantity, which is the root mean square deviation of the atomic coordinates of two structures after optimal superposition [298], is the same as the one used for the RMSD of the energies in sec. 3.7. This choice has been made for consistency with the literature, in which this ambiguity is also very common.

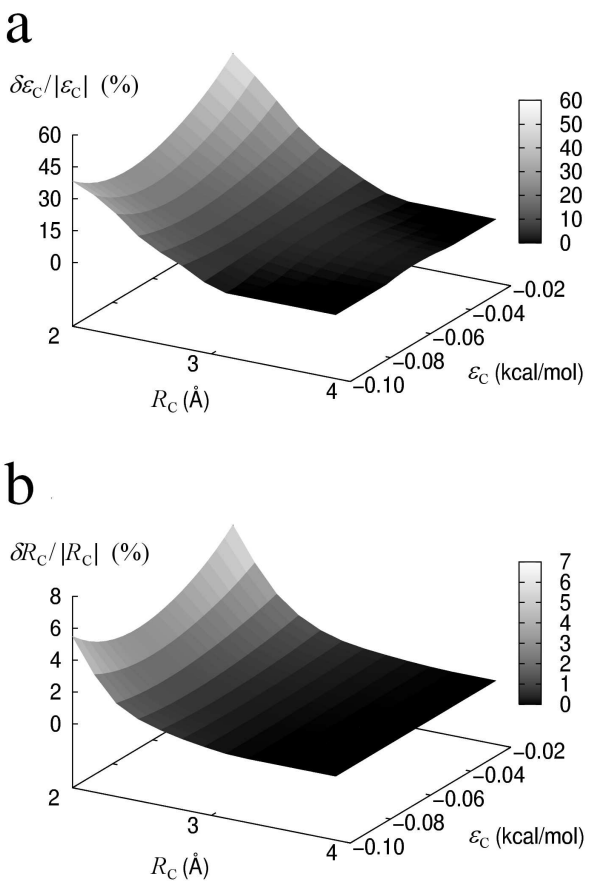


Figure 3.7: Robustness of the van der Waals energy in CHARMM with respect to changes in some free parameters. Relative indetermination in ε_C **(a)** and in R_C **(b)** associated to $d(V_1, V_2) = RT$ as a function of the central point in the parameter space. Larger values of the relative indetermination correspond to greater robustness.

robustness of this potential energy function with respect to changes in the free parameters ε_C and R_C, associated to the aliphatic sp3 carbon CH (denoted by CT1 in CHARMM), is investigated. To do this, a finite grid-like set of points (ε_C^k, R_C^k) is chosen in the bi-dimensional parameter space, with ε_C^k ranging from -0.10 kcal/mol to -0.02 kcal/mol and R_C^k ranging from 2 Å to 4 Å. Then, for each point in this set, different values $\delta\varepsilon_C$ are added to and subtracted from ε_C^k, or different values δR_C are added to and subtracted from R_C^k independently. The potential that corresponds to $\varepsilon_C = \varepsilon_C^k - \delta\varepsilon_C$ is denoted by V_1, the one that corresponds to $\varepsilon_C = \varepsilon_C^k + \delta\varepsilon_C$ is denoted by V_2 (analogously with R_C) and the distance $d(V_1, V_2)$ between the two instances is computed in each case (i.e., for each *central point* (ε_C^k, R_C^k) and for each $\delta\varepsilon_C$ (or δR_C))[93].

This procedure allows us to study the dependence of the distance between V_1 and V_2 on the size of the corresponding difference, $\delta\varepsilon_C$ or δR_C, between these two potentials in the parameter space, and to do that for each central point (ε_C^k, R_C^k). This relation may be regarded as one between indetermination in the values of the free parameters and its influence on the conformational behaviour of the system. From this point of view, the difference $\delta\varepsilon_C$ (or δR_C) for which the distance associated equals RT (see sec. 3.5) must be considered the maximum amount of indetermination in the parameters that does not involve relevant physical changes in the system. Therefore, if the parameters are known

[93] It can be proved that, in this particular case, the normality hypothesis in eq. (3.2) is approximately fulfilled.

to a precision equal or greater than the one associated to these particular values of $\delta\varepsilon_C$ or δR_C, the statistical indetermination of the parameters in a hypothetical fit process may be regarded as harmless. The values of this differences (as a function of the central point (ε_C^k, R_C^k)) computed for the system studied in this section are depicted in fig. 3.7.

Although this study only pretends to be an illustration of the concepts introduced in the previous sections and more features of the van der Waals energy should be investigated elsewhere, some interesting remarks may be made about the results herein presented. One one hand, directly from fig. 3.7, one can see that the precision needed in R_C is much greater than the one needed in ε_C, i.e., the van der Waals energy is more sensitive to changes in R_C than in ε_C. This is reasonable because V depends on R_C raised to the 12th and 6th power whereas ε_C only enters the expression raised to $1/2$ (see eq. (3.25)). On the other hand, the allowed indetermination in the parameters grows, in both cases, as R_C diminishes (the dependence on ε_C is much weaker). The reason for this being probably that, when the van der Waals radius R_C is large enough, the atoms begin to clash, i.e., the 12th power in eq. (3.25), associated to the steric repulsion, begins to dominate over the 6th power term, associated to the attractive dispersion forces.

Finally, note that, for the values $\varepsilon_C = -0.02$ kcal/mol and $R_C = 2.275$ Å, which are the ones used in the CHARMM `param22` parameter file, the allowed indeterminations in the parameters are $\delta\varepsilon_C/|\varepsilon_C| = 35$ % and $\delta R_C/R_C = 3$ %, in the region of relatively lower required precision (i.e., the relatively more favourable region). However, the indetermination for R_C corresponds to ~ 0.07 Å, which is a rather demanding accuracy, suggesting that, if the van der Waals radii set is changed, the behaviour of the system may be significantly modified.

Now, as a second brief example of the possible applications of the distance, we present an exploratory comparison of different levels of the theory in the quantum mechanical ab initio study of the *Potential Energy Surface* (PES) associated with the Ramachandran angles of the model dipeptide HCO-L-Ala-NH_2 (see appendix E). This comparison is an example of the first point discussed in sec. 3.6.

In ref. 207, the PES of HCO-L-Ala-NH_2 is calculated with two electronic structure methods, RHF and B3LYP, using, for each one, three different basis sets, 3-21G, 6-31+G(d) and 6-311++G(d,p) (see chapter 2). To do this, the Ramachandran space is divided in a 12×12 grid and, fixing the values of the ϕ and ψ torsional angles, a geometrical optimization of the structure is performed at each point. This process produces the values of six different instances of the same potential energy on a working set of 144 conformations of the system.

In fig. 3.8, each one of the six levels of the theory is compared to the other five (using the data kindly provided by A. Perczel) and some relevant numerical measures are presented. The distance d_{12} is given in units of RT (at 300 K), the energy reference shift a_{12} and slope b_{12} that result from the fit are also shown and the only quantity that requires further explanation is N_{res} (see eq. (3.27) below).

One of the interests in studying PESs of peptide models lies on the possibility of using the results for modeling short oligo-eptides or even proteins (see chapter 7). If we imagine that we use the PES of HCO-L-Ala-NH_2 to construct a potential that describes the behaviour of a peptide made up of N alanine residues, the first naive attempt would be to simply add N times the potential energy surface of the individual HCO-L-Ala-NH_2 (making each term suitably depend on different pairs of Ramachandran angles). We may

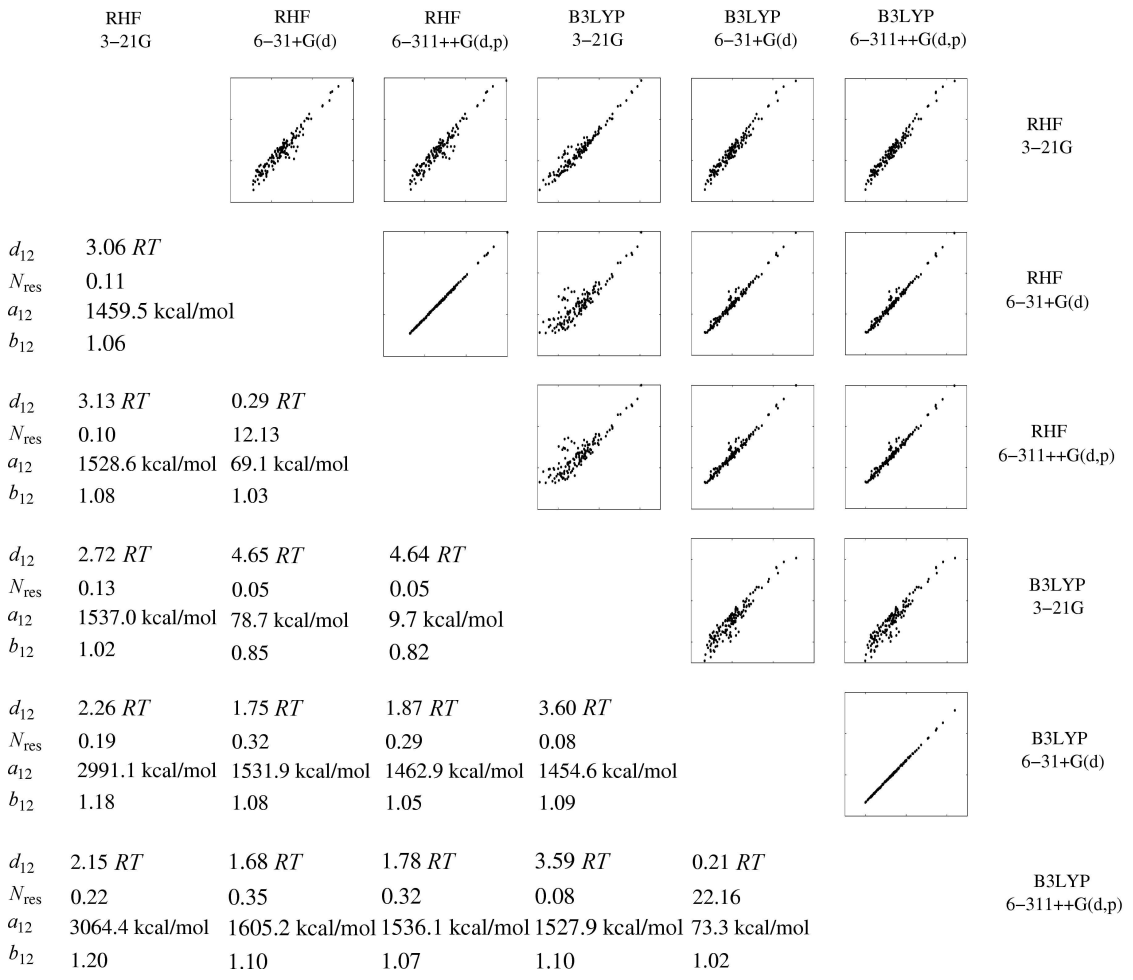


Figure 3.8: Comparison between different levels of the theory in the quantum mechanical ab initio study of the *Potential Energy Surface* (PES) associated with the Ramachandran angles of the model dipeptide HCO-L-Ala-NH_2 (see appendix E). The figure must be read as follows: **(i)** Any numerical set of measures is associated to the comparison between the level of the theory in the corresponding *row* (denoted by V_1) and *column* (denoted by V_2). **(ii)** The conformations scatter plot that belongs to a particular set of measures is the one that lies in the position which is obtained via reflection (of the set) with respect to the blank diagonal.

now ask whether the distance between two different instances of the N-residue peptide potential can be related to the distance between the corresponding mono-residue ones. It can be proved, appropriately choosing the working set of conformations of the larger system and using the relations presented in sec. 3.8, that the following relation holds:

$$d_{12}(N) = \sqrt{N} d_{12}(1) \,, \tag{3.26}$$

where we have denoted by $d_{12}(N)$ the distance between the V_1 and V_2 potentials of the N-residue peptide constructed as indicated above.

Hence, we define N_{res} as the N for which $d_{12}(N) = RT$, representing the maximum

number of residues up to which the criterium given in sec. 3.5 will be satisfied:

$$d_{12}(N_{\text{res}}) := RT \quad \Longrightarrow \quad N_{\text{res}} = \left(\frac{RT}{d_{12}(1)} \right)^2 . \tag{3.27}$$

Although a much more exhaustive study is carried out in chapter 7, let us extract some meaningful conclusions from the data in fig. 3.8 to close this section. Note, first, that the only two cases for which $d_{12} < RT$ are RHF/6-31+G(d) vs. RHF/6-311++G(d,p) and B3LYP/6-31+G(d) vs. B3LYP/6-311++G(d,p). This means that the convergence in basis sets is achieved for both methods somewhere between 6-31+G(d) and 6-311++G(d,p) and it suggests (for HCO-L-Ala-NH_2) that there is no need in going beyond 6-31+G(d). Of course, the fact that $N_{\text{res}} \simeq 22$, in the B3LYP case, and $N_{\text{res}} \simeq 12$, in the RHF case, places a limit on the size of the system for which the similarity of the two levels should be considered as sufficient. Finally, note that the distance between RHF/6-311++G(d,p) and B3LYP/6-311++G(d,p) is 1.78 RT, which means that the convergence in electronic structure methods has not been achieved and some more accurate method should be studied.

3.11 Summary and conclusions

In this chapter, a measure $d(V_1, V_2)$ of the differences between two instances of the same potential energy has been defined and the following points about it have been discussed:

- It rests on hypotheses whose validity stems from general characteristics shared by many complex systems and from the statistical laws of large numbers. We believe that, without knowing specific details of the system, the statistical approach is unavoidable and, among the many criteria, our distance is the most meaningful way of quantifying the differences between potentials.
- It allows to make physically meaningful statements about the way in which the energy differences between conformations change (or how the energetic ordering of the conformations is altered) upon substitution of one potential by the other.
- It may be applied to at least three practical situations characterized by the origin of the differences between the potentials:
  - Different algorithms or approximations are used (potential design).
  - The potential energy function depends on free parameters and the two instances correspond to different values of them (robustness).
  - Slightly different systems are compared (mutational studies, effective potentials).
- It presents advantages over the commonly used quantities RMSD, ER, SDER and AER that consist mainly of not overestimating irrelevant transformations on the potentials, such as adding an energy reference or rescaling one of them. Regarding the Pearson's correlation coefficient r, our distance may be considered as a physically meaningful way of giving him energy units. Finally, the numerical complexity involved in the calculation of $d(V_1, V_2)$ (see below) is similar to the one associated to any of the other quantities.

- It is approximately additive for most of the interesting situations encountered in practical cases.

In addition, a first practical example, which consists in the study of the robustness to changes in the free parameters of the van der Waals energy in CHARMM, and a second one, in which the ab initio PESs of the HCO-L-Ala-NH_2 molecule calculated at different levels of the theory are compared, have been presented to illustrate the concepts discussed.

Finally, we summarize the steps that must be followed to compute the distance in a practical case. Although all that follows has already been said, we believe that a brief 'recipe' could be useful for quick reference:

1. Generate a working set of independent conformations $\{\vec{q}_i\}_{i=1}^N$ (see sec. 3.2 and the last paragraph of sec. 3.3).

2. Denote $V_1^i := V_1(\vec{q}_i)$, $V_2^i := V_2(\vec{q}_i)$ and compute the statistical quantities μ_1, μ_2, σ_1 and $\mathrm{Cov}(V_1, V_2)$ in eq. (3.4).

3. With them, calculate the mean-square estimators using eqs. (3.3). First b_{12}, then a_{12} and, finally σ_{12}.

4. If comparing a potential energy function to a more accurate instance, use the relation $d_{12} = \sqrt{2}\sigma_{12}$ to find the asymmetrical version of the distance between them, and rescale V_2 dividing it by b_{12} if desired. Otherwise, repeat the steps 2 and 3 changing $1 \leftrightarrow 2$ in all the expressions to compute σ_{21} and use eq. (3.5) to finally arrive to $d(V_1, V_2)$.

5. If $d(V_1, V_2) < RT$ (or $d_{12} < RT$, depending on the case), the two potentials may be considered physically equivalent. If not, the behaviour described by them could be significantly different.

Chapter 4

SASMIC internal coordinates

This chapter is based on the article:
Pablo Echenique and J. L. Alonso, *Definition of Systematic, Approximately Separable and Modular Internal Coordinates (SASMIC) for macromolecular simulation*, J. Comp. Chem. **27** (2006) 1076–1087.

> A child of five would understand this. Send someone to fetch a child of five.
>
> — Groucho Marx, *Duck Soup*, 1933

4.1 Introduction

Apart from the accuracy of the potential energy functions discussed in other chapters, the choice of the coordinates used to describe proteins is also an important issue if computational considerations are to be taken into account and the efficiency of the simulations is pursued. This choice also affects the coding of applications: If cumbersomely defined coordinates are used, an unnecessary complexity may be added to the design of Monte Carlo movements, the construction and pruning of a database of structures [299, 300] or the programming of molecular visualization and manipulation tools.

Suitable coordinates frequently used to describe arbitrary conformations of molecules are the so-called *internal* or *valence-type* coordinates [176] (see fig. 1.10). Their adequacy stems from a number of characteristics: first, they are closely related to chemically meaningful structural parameters, such as bond lengths or bond angles; second, they are local, in the sense that each one of them involves only a small number of (normally close) atoms in its definition; and finally, there are only $3n-6$ of them (where n is the number of atoms in the molecule), in such a way that the overall rotation and translation have been naturally removed (see chapter 5 for relevant qualifications to this issue).

There also exists a family of coordinates [301–304], extensively used in the inner calculations of many quantum chemistry packages (such as Gaussian [49] or GAMESS [305]) and based on the *natural internal coordinates* originally proposed by Pulay and coworkers [306–308], which are defined through linear combinations of the original internals. These coordinates are specially designed to describe normal-mode vibrations in

the immediate neighbourhood of energy minima and represent the best choice for accelerating the convergence of geometry optimizations in a particular basin of attraction, via diagonal estimation of the Hessian matrix [302]. Accordingly, they maximally separate hard and soft movements in these conditions. However, if the conformation of the molecule is far from a minimum, this type of coordinates lose great part of their meaning and they introduce many computational difficulties without increasing the efficiency. In addition, some of the definitions are *redundant* [303, 306–309], i.e., they use a number of linear combinations of internals larger than the number of degrees of freedom, significantly reducing human comfort and readability. In this chapter, we will only discuss coordinates, such as internals or Euclideans, that may be conveniently used to specify an *arbitrary* conformation of the system and that can be directly related to simple geometrical variables.

The numerical complexity of the methods used to predict the physical behaviour of macromolecules, such as proteins, via computer simulations, often scales harshly with the number of degrees of freedom of the system. Therefore, it is also advisable that the set of coordinates chosen allows for a direct implementation of physically meaningful constraints that reduce the dimensionality of the conformational space considered. Most of the expressions used in statistical mechanics or in molecular dynamics are best written in Euclidean coordinates, however, the implementation of naturally appearing constraints is far from being straightforward in these coordinates. In internal coordinates, on the contrary, the approximate separation of hard and soft movements of the system allows to easily constrain the molecule [310–312] by setting the hard coordinates (those that require a considerable amount of energy to change noticeably) to constant values or to particular functions of the soft coordinates. Moreover, in internal coordinates (and appealing to some reasonable approximations), the statistical mechanics formulae for the constrained system may be written in convenient closed form [313, 314] (see chapter 6 for further information on this topic).

Still, although the bond lengths and bond angles are customarily regarded as hard and their definition is unproblematic, the same is not true for dihedral angles. Some definitions of dihedrals may lead to difficulties or to worse separation of hard and soft modes in branched molecules. Let us exemplify this with a particular case:

Consider the definition of Z-matrix-like [315, 316] internal coordinates for the HCO-L-Ala-NH_2 molecule in fig. 4.8. Imagine that we 'position' (i.e., we write the corresponding Z-matrix row) every atom up to the hydrogen denoted by H_9 and that we are now prepared to position the hydrogens in the side chain (H_{10}, H_{11} and H_{12}) via one bond length, one bond angle and one dihedral angle for each one of them. We will denote by (i, j) the bond length between atoms i and j; by (i, j, k), the bond angle between the vectors $\vec{r}_{jk}$ and $\vec{r}_{ji}$; and by (i, j, k, l) the dihedral angle between the plane defined by the atoms i, j and k and the one defined by j, k and l.

A choice to position the atoms that is frequently seen in the literature [209, 299, 300, 317, 318] is the one shown in table 4.1.

If we now perform the *gedanken experiment* that consists of taking a typical conformation of the molecule and slightly moving each internal coordinate at a time while keeping the rest constant, we find that any one of the three dihedrals in the previous definition is a hard coordinate, since moving one of them while keeping the other two constant distorts the internal structure of the methyl group. Hence, in these coordinates, the soft rotameric

Atom name	Bond length	Bond angle	Dihedral angle
H_{10}	(10,8)	(10,8,5)	γ_1 :=(10,8,5,3)
H_{11}	(11,8)	(11,8,5)	γ_2 :=(11,8,5,3)
H_{12}	(12,8)	(12,8,5)	γ_3 :=(12,8,5,3)

Table 4.1: A part of the internal coordinates, in Z-matrix form, of the protected dipeptide HCO-L-Ala-NH_2 (see also appendix E), as frequently defined in the literature.

degree of freedom χ, which we know, for chemical arguments, that must exist[94], is ill-represented. In fact, it must be described as a *concerted* movement of the three dihedrals. In ref. 319 this is clearly explained; in refs. 299 and 300, the problem is recognized and the concept of *related dihedrals* is introduced, however, no action is taken to change the definition of the coordinates.

In this chapter, using the ideas of R. Abagyan and coworkers [310–312], we introduce a set of rules to uniquely and systematically number the groups, the atoms and define the internal coordinates of polypeptides[95], and also a modified set of rules to do the same for general organic molecules. The main difference with other Z-matrix-like coordinates normally used in the literature [209, 299, 300, 317, 318] is that, instead of positioning each atom with a bond length, a bond angle and a dihedral angle, we use normal dihedral angles (called, from now on, *principal dihedrals*) only to fix the orientation of whole chemical groups and a different type of dihedrals, termed *phase dihedrals* by R. Abagyan and coworkers [310–312] (see fig. 4.1), to describe the covalent structure inside a group[96]. This allows to *approximately* separate soft and hard conformational movements of the molecule using only topological information (i.e., not knowing the exact form of the potential) and to easily implement constraints by forcing the coordinates that correspond to hard movements to take constant values or ones that depend on the soft coordinates (see chapter 6 for a practical example of this)[97].

In addition, the coordinates herein defined, are straightforwardly cast into Z-matrix form and can be directly implemented in any quantum chemistry package, such as Gaussian [49] or GAMESS [305]. This is due to the fact that, although they involve atoms whose covalent structure is different, the mathematical construction of the two types of

[94] According to our calculations, at the RHF/6-31+G(d) level of the theory, the barrier for crossing from one of the three equivalent minima to any of the other two ranges from 3.1 to 6.8 kcal/mol, depending on the values of the Ramachandran angles ϕ and ψ. Compare with the barriers in ϕ or ψ which may be as large as 20 kcal/mol depending on the region of the Ramachandran map explored.

[95] IUPAC conventions only define a numeration system for the groups, for the branches and for some selected dihedral angles. They focus on functional considerations and not in computational problems. For related documents and references, see `http://www.chem.qmul.ac.uk/iupac/jcbn/`.

[96] Another option may be to use, as a third internal coordinate for each atom, another bond angle. This is rather awkward, however, since two bond angles and a bond length do not specify the position of a point in space. Any values of these three coordinates (except for irrelevant degenerate cases) are compatible with two different symmetrical positions and a fourth number must be provided to break the ambiguity. Also, out-of-plane angles may be used. In ref. 319, different options are described.

[97] In ref. 210 they correctly take this approach into account using out-of-plane angles instead of phase dihedrals, however, they do not describe any rules for a general definition and their numeration of the atoms is non-modular, as it proceeds first through the backbone (see sec. 4.2).

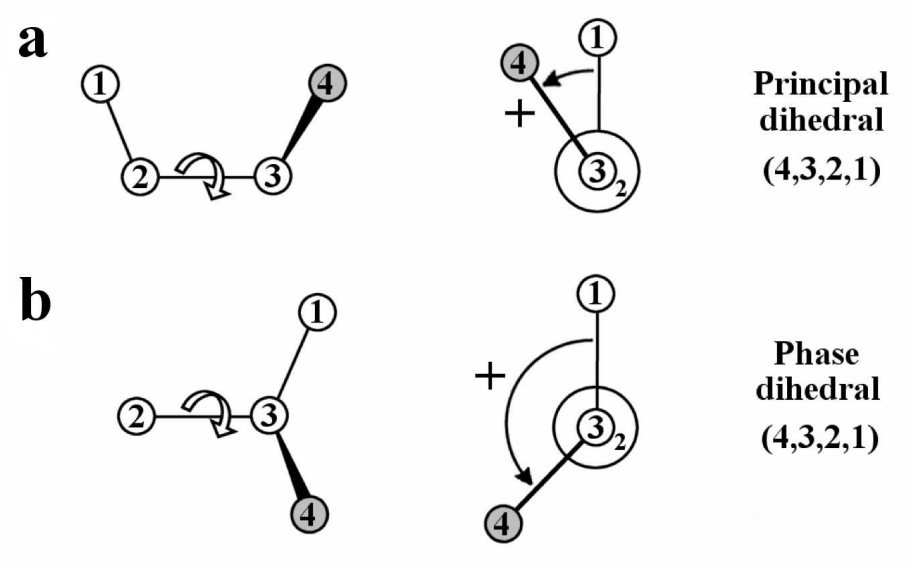


Figure 4.1: Two types of dihedral angles. **(a)** *Principal dihedral*. Used to describe the rotation of whole groups around bonds. **(b)** *Phase dihedral*. Used to describe the internal covalent structure of groups. The positive sense of rotation is indicated.

angles in fig. 4.1 is exactly the same, and the phase dihedrals are treated like principal ones without any problem by the applications.

Taking profit from this fact, a number of Perl scripts have been coded (and are publicly available at `http://neptuno.unizar.es/files/public/gen_sasmic/`) that number the atoms and generate the coordinates herein defined for polypeptide chains. The applications read a sequence file in which the different ionization states of the titratable side chains, the tautomeric forms of histidine (see sec. 1.2) and several terminal groups may be specified. Then, an output file is generated with the symbolic definition of the Z-matrix of the molecule which may be directly pasted into the input files of Gaussian [49] or GAMESS [305] (and, upon slight modifications, of any quantum chemistry package that is capable of reading Z-matrix format).

Now, if we redo the example in table 4.1 using phase dihedrals, we must write the rows of the Z-matrix for the hydrogens in the side chain as shown in table 4.2, where the angle **(10,8,5,3)** is now the principal dihedral χ describing the relative rotation of the methyl group around the bond (8,5) and the other two are phase dihedrals that describe

Atom name	Bond length	Bond angle	Dihedral angle
H_{10}	(10,8)	(10,8,5)	χ :=**(10,8,5,3)**
H_{11}	(11,8)	(11,8,5)	α_1 :=(11,8,5,10)
H_{12}	(12,8)	(12,8,5)	α_2 :=(12,8,5,10)

Table 4.2: A part of the internal coordinates, in Z-matrix form, of the protected dipeptide HCO-L-Ala-NH_2 (see also appendix E), as defined by the rules given in sec. 4.2.

the internal structure of the group and that are 'pure' hard coordinates (as far as can be told only from topological information). Note, however, that, although all bond lengths, bond angles and phase dihedrals may be regarded as hard coordinates, not all the principal dihedrals will be soft. Examples of hard principal dihedrals are the ones that describe the rotation around a double bond (or a triple one) or some of the principal dihedrals in cyclic parts of molecules.

The *physical approach* to design internal coordinates described in this section and exemplified by the simple case above, is generalized in this chapter and embodied in a set of rules for polypeptide chains in sec. 4.2, while a slightly different prescription for general organic molecules is described in sec 4.3. The systematic numeration introduced facilitates the computational treatment of this type of systems and the rules given for polypeptide chains ensure modularity [2, 299], i.e., allows to add any residue with minimal modification of the already existing notation and to easily construct databases of structures or of potential energy surfaces.

All the characteristics and advantages mentioned in the preceding paragraphs have led us to term the coordinates herein defined *Systematic, Approximately Separable and Modular Internal Coordinates* (SASMIC), and we will use them in many of the rest of the chapters of this Ph.D. dissertation.

Also note that, although in this chapter, we will only deal with the numeration of one isolated molecule according to the SASMIC scheme, the procedure described may be easily generalized (and will be in future research) to systems of many molecules (an important example being a macromolecular solute in a bath of solvent molecules). This could be achieved using *ghost atoms* in a similar manner to what is done in ref. [313], to position the center of mass of the system, and in refs. [310–312], to actually define the coordinates of a system of molecules.

In sec. 4.4, we use the new coordinates and ab initio quantum mechanical calculations in order to evaluate the approximation of the effective potential energy (obtained from integrating out the rotameric degree of freedom χ) with the typical PES in the protected dipeptide HCO-L-Ala-NH_2 (see appendix E). We also present a small part of the Hessian matrix in two different sets of coordinates to illustrate the approximate separation of soft and hard movements when the SASMIC scheme defined in this chapter is used. Finally, sec. 4.5 is devoted to the conclusions.

4.2 Numeration rules for polypeptides

4.2.1 Definitions

First, we realize that any molecule may be formally divided in groups such as those in fig. 4.2. We will call *centers* the shaded atoms in the figure and *vertices* the white ones. In general, there may exist groups with more than four vertices, however, in proteins, only groups with four or less vertices occur. Examples of tetrahedral groups are the one whose center is the C_α in the backbone or the C_β in the side chain of alanine, triangular groups occur, for example, at the N or the C' in the backbone, finally, linear groups may be found at the O in the side chain of tyrosine or at the S in methionine (see figs. 1.7 and 4.12).

A particular atom may be vertex of different groups but may only be center of one group. There exist atoms that are only vertices but there do not exist atoms that are only

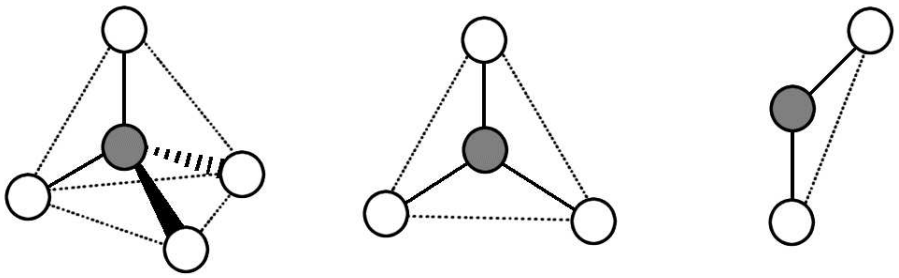


Figure 4.2: Schematic representation of the groups found in proteins (the angle in the linear group might as well be 180^o). From left to right: tetrahedral, triangular and linear.

centers, except in the case of molecules with only one group. In the trivial case of diatomic molecules (in which the only internal coordinate is a bond length), neither of the previous definitions are possible, since we cannot identify a group.

Atoms that are covalently bonded to more than one atom will be called *internal atoms* and are indicated as shaded circles in fig. 4.3. Atoms that are covalently attached to only one internal atom will be called *external atoms* and are indicated as white-filled circles in fig. 4.3. In proteins, only H and O may be external atoms.

In most macromolecular models (such as the Born-Oppenheimer approximation used in sec. 4.4 and described in chapter 2), nuclei are considered point-like particles. Hence, rotation around bonds joining external and internal atoms (termed *external bonds* or *non-dihedral bonds*) is neglected, i.e., there are no internal coordinates associated to this movement. On the other hand, rotation around bonds joining two internal atoms (called *internal bonds* or *dihedral bonds* and indicated with curved arrows in fig. 4.3) is relevant and there may exist internal coordinates describing it.

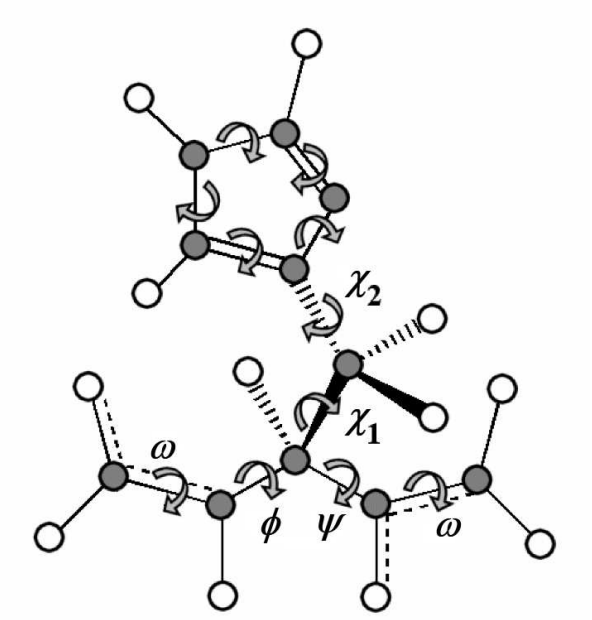


Figure 4.3: Schematic representation of the HCO-L-His-NH_2 model dipeptide (with the side chain in its uncharged ε_2-tautomeric form). Internal atoms are shown as gray-filled circles, external ones as white-filled circles. Internal bonds are indicated with curved arrows. Typical biochemical definitions of some principal dihedrals are also shown.

In order to conform with the *physical approach* stated in the introduction, only one *golden rule* must be followed when defining the internal coordinates:

> One principal dihedral, at most[98], must be defined on each internal bond.

The rest of the rules that will be given are mere tidy conventions and systematics.

4.2.2 Rules for numbering the groups

First of all, we will divide the peptide in groups and number them. To do this we proceed 'by branches', i.e., we choose the next group following a linear sequence of covalently attached groups until there is no possible next one, in which case, we either have finished the numeration process or we start another branch. Every group is numbered once and it cannot be renumbered as the process continues. This numeration is done for completeness and as a support for the numeration of atoms and coordinates. In fig. 4.4, we have implemented these rules in a protected histidine dipeptide.

The rules are as follows:

i) We select as the *first group* (and number it $j = 1$):

- The *amino group* at the N-terminus (either protonated or not) if the polypeptide is not N-protected.
- The *formyl group* at the N-terminus if the polypeptide is formyl-N-protected.
- The *methyl group* at the N-terminus if the polypeptide is acetyl-N-protected.

These three cases are the most frequent. If a different species is used to N-protect the polypeptide chain, a convention must be sought that also starts at the N-terminus. This choice takes into account that the primary structure of a polypeptide is normally presented from the N- to the C-terminus (see sec 1.2).

ii) If there is only one unnumbered group linked to group j, we number it as $j + 1$, set $j = j + 1$ and go to (ii).

iii) If there are two or more unnumbered groups linked to group j, we choose the next one as *the one with the greatest mass* (the mass of a group is defined as the sum of the atomic masses of its constituents). If two or more neighbouring unnumbered groups have the same mass, we add the mass of their first neighbours to break the tie. If this does not lead to a decision, we proceed to the second neighbours and so on. If we run out of neighbours and there is still a tie, we choose a group arbitrarily among the ones that have been selected via this process and we indicate the convention. We number the group chosen as $j + 1$, set $j = j + 1$ and go to point (ii).

Exception: When we must choose the next group to the one whose center is a C_α in the backbone, instead of applying the rule of greatest mass, which would yield the group at the C' as the next one, we choose *the first group in the side chain* (for residues that are different from glycine). Then, we number the group chosen as

[98] It is not possible to define principal dihedrals for each internal bond for structures containing rings due to the well known limitation of Z-matrix internal coordinates.

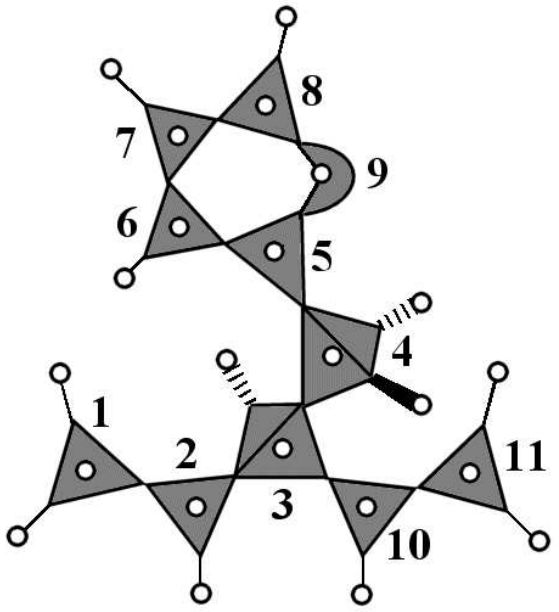


Figure 4.4: Group identification and numeration in the protected dipeptide HCO-L-His-NH_2 (with the side chain in its uncharged ε_2-tautomeric form). The different types of groups are shown as gray-filled polyhedra.

$j+1$, set $j = j+1$ and go to (ii). This is done in order to ensure modularity, since, otherwise, the backbone would be always numbered first and the whole numeration would have to be modified if we added a new residue to the chain.

iv) If there are no unnumbered groups linked to group j, we prepare to start another branch and have two choices: For modularity reasons, we want to completely number the side chain before proceeding into the backbone. Hence, if we are numbering side chain groups and there are still unnumbered groups in the same side chain, we set j to *the number of the lowest numbered group that has unnumbered neighbours and that belongs to the side chain of the residue whose groups we are numbering.*. If we are not numbering side chain groups or we are numbering side chain groups but there is no unnumbered groups in the same side chain left, we set j to *the number of the lowest numbered group that has unnumbered neighbours in the whole peptide*. Then, we go to (ii).

This process terminates when all the groups are numbered.

4.2.3 Rules for numbering the atoms

The atoms will be numbered in the order that they will be positioned via internal coordinates in the Z-matrix. In fig. 4.5, the rules given in this section are exemplified in a protected histidine dipeptide.

The rules are as follows:

i) The first atom ($k = 1$), is chosen as *the heaviest of the external atoms in the first group*. If there are two or more candidates with the same mass, we choose arbitrarily and indicate the convention.

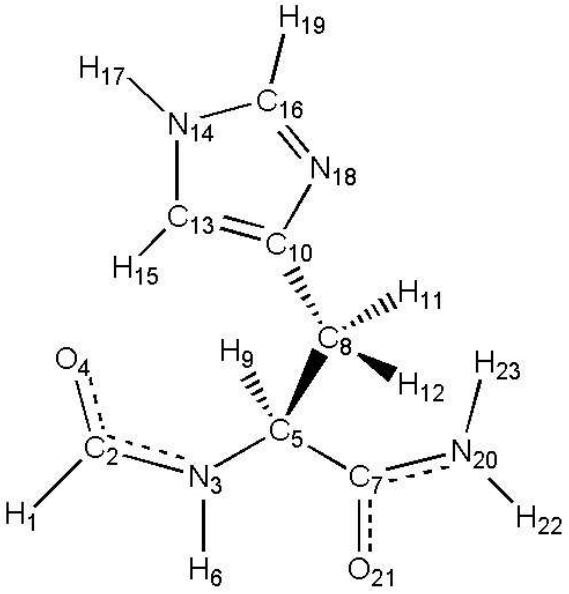


Figure 4.5: Atom numeration of the protected dipeptide HCO-L-His-NH_2 (with the side chain in its uncharged ε_2-tautomeric form).

Exception: If the polypeptide is formyl-N-protected, instead of applying the rule, which would yield the oxygen at the formyl group, we choose *the hydrogen at the formyl group*

ii) The second atom ($k = 2$) is the center of the first group and we set $j = 1$ (the index of the group).

iii) If group $j + 1$ exists and is covalently attached to group j, we number the unnumbered vertices of group j *starting by the center of group* $j + 1$ *and, then, in order of decreasing mass*. If, otherwise, group $j + 1$ does not exist or it is not covalently attached to group j, *we simply number the unnumbered vertices of group* j *in order of decreasing mass*. If, at any point, there are two or more candidates with the same mass, we choose arbitrarily and indicate the convention.

Exception 1: If groups j and $j+1$ belong to the same cyclic part of the molecule, the vertices of j that are centers of groups (other than $j+1$) belonging to the same cycle *must not be numbered at this step* (for an example of this rule, see the numeration of C_{13} and N_{18} in fig. 4.5).

Exception 2: If the polypeptide is amide-C-protected, instead of applying the above rule and arbitrarily choosing one of the hydrogens in the terminal amide group before the other, we number *the trans hydrogen* before the other (see fig. 4.5).

Exception 3: Due to the rules for the numeration of groups given in the previous section, the next group to the one at the C_α is the first one in the side chain. If we applied the general rule for numbering the vertices of the C_α-group, we would number first the center of the first group at the side chain and, then, the C' in the backbone. This would make the only principal dihedral defined on bond (C_α, N) different from the conventional Ramachandran angle ϕ (see sec 1.2). In order to avoid this, at this point, we number the C' first among the unnumbered vertices of the C_α-group and, then, resume the usual numeration process (see fig. 4.5).

iv) If group $j+1$ does not exist, we have finished. Otherwise, we set $j = j+1$ and go back to (iii).

The exception to rule (i) and the exceptions 2 and 3 to rule (iii) are introduced in order that the principal dihedrals that are to be defined after numbering the atoms conform to the biochemical IUPAC conventions for the dihedrals ϕ, ψ and ω in the backbone. At the termini, we have ensured that the atom where the C_α of the hypothetical residue 0 or $N+1$ would occur is used to define the principal dihedrals.

See fig. 4.12 for the numeration of the twenty naturally occurring amino acids with formyl-N- and amide-C-protection.

4.2.4 Rules for defining the internal coordinates

Using the numeration for the atoms given in the previous section, we give now a set of rules for defining the internal coordinates that conform with the *physical approach* discussed in the introduction of this chapter. The coordinates are written in Z-matrix form (see table 4.3) for convenience and the rules are applied to the protected dipeptide HCO-L-His-NH_2 (with the side chain in its uncharged ε_2-tautomeric form) using the numeration given in fig. 4.5.

The rules are as follows:

i) The positioning of *the first three atoms* is special. The corresponding rows of the Z-matrix are *always* as the ones in table 4.3 (except, of course, for the chemical symbol in the first column, which may change).

ii) The positioning of the remaining vertices of group number 1 (if there is any) is also special, their rows in the Z-matrix are:

$$\mathrm{T}_i \quad (i,2) \quad (i,2,1) \quad (i,2,1,3)$$

Where T is the chemical symbol of the i-th atom, and $(i,2,1,3)$ is a phase dihedral.

iii) We set i to *the number that follows that of the last vertex of the first group*.

iv) We choose j as *the lowest numbered atom that is covalently linked to* i.

v) We choose k as *the lowest numbered atom that is covalently linked to* j.

vi) If no principal dihedral has been defined on the bond (j,k) (we say that a principal dihedral (i,j,k,l) *is* 'on the bond (j,k)'), we choose l as *the lowest numbered atom that is covalently linked to* k. Otherwise, we choose l as *the second lowest numbered atom that is covalently linked to* j (i.e., the lowest numbered atom that is covalently linked to j and that is different from k, or, equivalently, the atom that was used to define the only principal dihedral on the bond (j,k)).

vii) The row of the Z-matrix that corresponds to atom i is:

$$\mathrm{T}_i \quad (i,j) \quad (i,j,k) \quad (i,j,k,l)$$

Atom name	Bond length	Bond angle	Dihedral angle
H_1			
C_2	(2,1)		
N_3	(3,2)	(3,2,1)	
O_4	(4,2)	(4,2,1)	(4,2,1,3)
C_5	(5,3)	(5,3,2)	**(5,3,2,1)**
H_6	(6,3)	(6,3,2)	(6,3,2,5)
C_7	(7,5)	(7,5,3)	**(7,5,3,2)**
C_8	(8,5)	(8,5,3)	(8,5,3,7)
H_9	(9,5)	(9,5,3)	(9,5,3,7)
C_{10}	(10,8)	(10,8,5)	**(10,8,5,3)**
H_{11}	(11,8)	(11,8,5)	(11,8,5,10)
H_{12}	(12,8)	(12,8,5)	(12,8,5,10)
C_{13}	(13,10)	(13,10,8)	**(13,10,8,5)**
N_{14}	(14,13)	(14,13,10)	**(14,13,10,8)**
H_{15}	(15,13)	(15,13,10)	(15,13,10,14)
C_{16}	(16,14)	(16,14,13)	**(16,14,13,10)**
H_{17}	(17,14)	(17,14,13)	(17,14,13,16)
N_{18}	(18,16)	(18,16,14)	**(18,16,14,13)**
H_{19}	(19,16)	(19,16,14)	(19,16,14,18)
N_{20}	(20,7)	(20,7,5)	**(20,7,5,3)**
O_{21}	(21,7)	(21,7,5)	(21,7,5,20)
H_{22}	(22,20)	(22,20,7)	**(22,20,7,5)**
H_{23}	(23,20)	(23,20,7)	(23,20,7,22)

Table 4.3: Internal coordinates in Z-matrix form of the protected dipeptide HCO-L-His-NH_2 (with the side chain in its uncharged ε_2-tautomeric form), following the rules given in sec. 4.2.4. Principal dihedrals are indicated in bold face.

Where T is the chemical symbol of atom i, (i, j) is a bond length, (i, j, k) is a bond angle and (i, j, k, l) is a principal dihedral if the first case in point (vi) has occurred or a phase dihedral otherwise.

viii) If $i + 1$ does not exist, we have finished. Otherwise, we set $i = i + 1$ and go to (iv).

4.3 Numeration rules for general organic molecules

4.3.1 Definitions

In order to introduce the modified version of the SASMIC rules for general organic molecules, the definitions found in sec. 4.2.1 are kept. The only changes affect the rules for numbering the groups and the atoms, since, when this is achieved, the rules for defining the internal coordinates are the same as the ones in sec. 4.2.4.

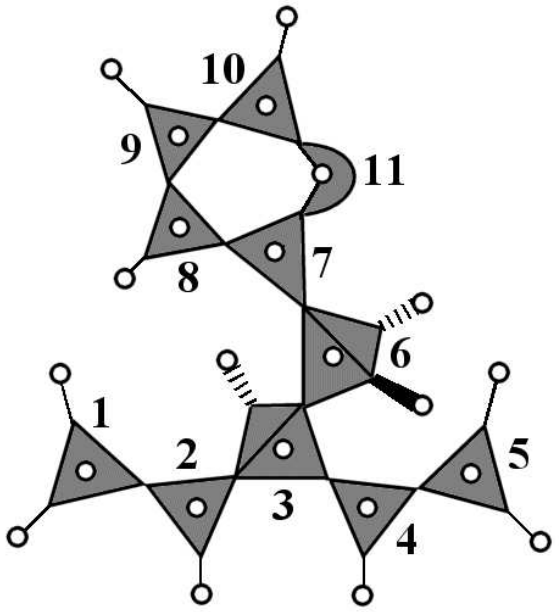


Figure 4.6: Group identification and numeration in the protected dipeptide HCO-L-His-NH_2 (with the side chain in its uncharged ε_2-tautomeric form), following the rules for general organic molecules. The different types of groups are shown as gray-filled polyhedra.

4.3.2 Rules for numbering the groups

Like we have done for peptides, first of all, we will divide the molecule in groups and number them. To do this we proceed 'by branches', i.e., we choose the next group following a linear sequence of covalently attached groups until there is no possible next one, in which case, we either have finished the numeration process or we start another branch. Every group is numbered one time and it cannot be renumbered as the process continues.

In fig. 4.6, we have implemented these rules for general organic molecules in a protected histidine dipeptide.

The rules are as follows:

i) The first group ($j = 1$), is chosen, among those that are linked to the molecule via only one internal bond (termed *terminal groups*), as *the one that has the greater mass* (the mass of a group is defined as the sum of the atomic masses of its constituents). If two or more terminal groups have the same mass, we add the mass of their first neighbours to break the tie. If this does not lead to a decision, we proceed to the second neighbours and so on. If we run out of neighbours and there is still a tie, we choose a group arbitrarily among the ones that have been selected via this process and we indicate the convention. If there are no terminal groups, we perform this selection process among those groups that have *at least one external atom*[99].

ii) If there is only one unnumbered group linked to group j, we number it as $j + 1$, set $j = j + 1$ and go to (ii).

iii) If there are two or more unnumbered groups linked to group j, we choose *the one with the greater mass* like in point (i), we number it as $j + 1$, set $j = j + 1$ and go to point (ii).

[99] The rare case in which there are neither terminal groups nor external atoms (such as C_{60} fullerene) will not be treated here, although it would require only a small number of adjustments to the rules.

iv) If there are no unnumbered groups linked to group j but there are still unnumbered groups in the molecule, we set j to *the number of the lowest numbered group that has unnumbered neighbours* (we prepare to start another branch) and we go to (ii).

This process terminates when all the groups are numbered.

4.3.3 Rules for numbering the atoms

The atoms will be numbered in the order that they will be positioned via internal coordinates in the Z-matrix. Like in the previous section, in fig. 4.7, these rules for a general organic molecule are exemplified in a protected histidine dipeptide.

The rules are as follows:

i) The first atom ($k = 1$), is chosen as *the heaviest of the external atoms in the first group*. If there are two or more candidates with the same mass, we choose arbitrarily and indicate the convention.

ii) The second atom ($k = 2$) is the center of the first group and we set $j = 1$ (the index of the group).

iii) If group $j + 1$ exists and is covalently attached to group j, we number the unnumbered vertices of group j *starting by the center of group* $j + 1$ *and, then, in order of decreasing mass*. If, otherwise, group $j + 1$ does not exist or it is not covalently attached to group j, *we simply number the unnumbered vertices of group* j *in order of decreasing mass*. If, at any point, there are two or more candidates with the same mass, we choose arbitrarily and indicate the convention.

Exception: If groups j and $j+1$ belong to the same cyclic part of the molecule, the vertices of j that are centers of groups (other than $j+1$) belonging to the same cycle

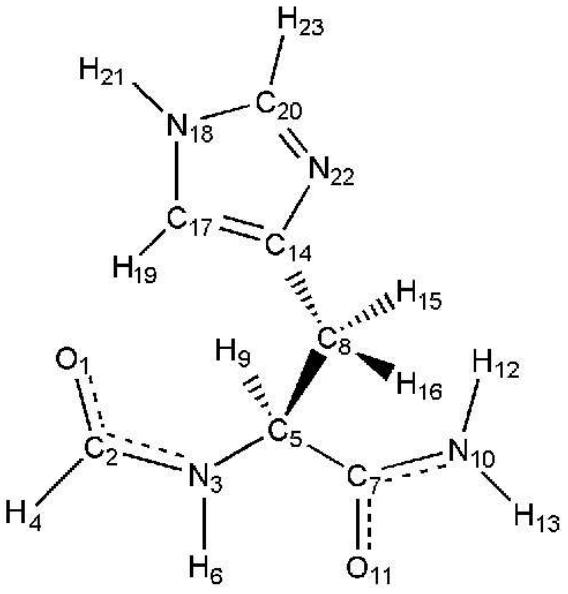


Figure 4.7: Atom numeration of the protected dipeptide HCO-L-His-NH_2 (with the side chain in its uncharged ε_2-tautomeric form), following the rules for general organic molecules.

must not be numbered at this step (for an example of this rule, see the numeration of C_{17} and N_{22} in fig. 4.7).

iv) If group $j+1$ does not exist, we have finished. Otherwise, we set $j = j+1$ and go back to (iii).

4.3.4 Rules for defining the internal coordinates

Using the numeration for the atoms given in the previous section, the rules for defining the SASMIC internal coordinates that conform with the *physical approach* discussed in the introduction are the same as the ones given in the sec. 4.2.4.

The coordinates, written in Z-matrix form, of the protected dipeptide HCO-L-His-NH_2 (with the side chain in its uncharged ε_2-tautomeric form) using the numeration given in fig. 4.7 are given in table 4.4.

Atom name	Bond length	Bond angle	Dihedral angle
O_1			
C_2	(2,1)		
N_3	(3,2)	(3,2,1)	
H_4	(4,2)	(4,2,1)	(4,2,1,3)
C_5	(5,3)	(5,3,2)	**(5,3,2,1)**
H_6	(6,3)	(6,3,2)	(6,3,2,5)
C_7	(7,5)	(7,5,3)	**(7,5,3,2)**
C_8	(8,5)	(8,5,3)	(8,5,3,7)
H_9	(9,5)	(9,5,3)	(9,5,3,7)
N_{10}	(10,7)	(10,7,5)	**(10,7,5,3)**
O_{11}	(11,7)	(11,7,5)	(11,7,5,10)
H_{12}	(12,10)	(12,10,7)	**(12,10,7,5)**
H_{13}	(13,10)	(13,10,7)	(13,10,7,12)
C_{14}	(14,8)	(14,8,5)	**(14,8,5,3)**
H_{15}	(15,8)	(15,8,5)	(15,8,5,14)
H_{16}	(16,8)	(16,8,5)	(16,8,5,14)
C_{17}	(17,14)	(17,14,8)	**(17,14,8,5)**
N_{18}	(18,17)	(18,17,14)	**(18,17,14,8)**
H_{19}	(19,17)	(19,17,14)	(19,17,14,18)
C_{20}	(20,18)	(20,18,17)	**(20,18,17,14)**
H_{21}	(21,18)	(21,18,17)	(21,18,17,20)
N_{22}	(22,20)	(22,20,18)	**(22,20,18,17)**
H_{23}	(23,20)	(23,20,18)	(23,20,18,22)

Table 4.4: Internal coordinates in Z-matrix form of the protected dipeptide HCO-L-His-NH_2 (with the side chain in its uncharged ε_2-tautomeric form), following the rules for general molecules. Principal dihedrals are indicated in bold face.

4.4 Practical example

4.4.1 Theory

When a number of degrees of freedom are removed from the description of the conformations of a physical system via their integrating out in the partition function, the potential energy function that remains, which describes the behaviour of the system only in terms of the rest of the degrees of freedom, is termed *effective potential energy* (see sec. 1.4). It depends on the temperature and contains the entropy of the information that has been averaged out as well as the enthalpy (which is similar to the average internal energy at physiological conditions).

On the other hand, it is frequent, when studying the conformational preferences of model dipeptides in order to use the information for designing effective potentials of polypeptides [104, 105, 117, 118, 123, 159, 195], that the energy of these molecules be approximated by the *Potential Energy Surface* (PES) in the bidimensional space spanned by the Ramachandran angles ϕ and ψ [159, 207, 208, 210]. If we recognize that the potential energy of the system in the Born-Oppenheimer approximation (denoted by V_{3n-6}) depends on the $3n-6$ internal coordinates, this surface (denoted by V_2) may be defined as:

$$V_2(\phi,\psi) := \min_{q^I} V_{3n-6}(\phi,\psi,q^I)\,, \tag{4.1}$$

where q^I denotes the rest of the internal coordinates.

The use of this surface, instead of the effective energy function with the q^I degrees of freedom integrated out, is partially justified in the approximation that these internal coordinates are hard and that they are comparably much more difficult to excite at room temperature than ϕ and ψ. If we assume that this is correct, these hard degrees of freedom may be easily eliminated (see chapter 6) and the partition function of the system may be written as follows:

$$Z = \int \exp\left[-\beta V_{3n-6}(\phi,\psi,q^I)\right] \mathrm{d}\phi\,\mathrm{d}\psi\,\mathrm{d}q^I \simeq \int \exp\left[-\beta V_2(\phi,\psi)\right] \mathrm{d}\phi\,\mathrm{d}\psi\,, \tag{4.2}$$

where the momenta have been previously integrated out, and a number of (maybe different) T-dependent multiplicative factors have been omitted as usual.

Note however that, precisely due to the averaging over momenta, in the *stiff* picture for the constraints, this expression is correct only if we assume that the Jacobian determinant of the change of coordinates from Euclideans to (ϕ,ψ,q^I) and the determinant of the potential second derivatives matrix with respect to the hard coordinates q^I, both evaluated at the equilibrium values, do not depend on ϕ and ψ. If, alternatively, we accept the *rigid* picture for the constraints, we must ask that the determinant of the reduced mass-metric tensor in the constrained hypersurface do not depend on ϕ and ψ. If these approximations (which will be carefully examined in chapter 6) do not hold but the hardness of the q^I degrees of freedom is still assumed, the expressions in eq. (4.2) must be modified by adding some correction terms to $V_2(\phi,\psi)$.

Now, in eq. (4.2) for the partition function, one also may see that the use of the PES $V_2(\phi,\psi)$ as the fundamental energy function of the system is justified because it plays

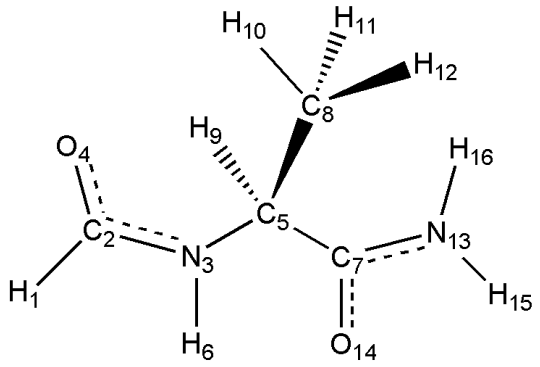


Figure 4.8: Atom numeration of the protected dipeptide HCO-L-Ala-NH_2 (see also appendix E).

the same role as the whole potential energy of the system in the first integral. However, although the hardness of the bond lengths, the bond angles and even the dihedral ω in the peptide bond may be assumed, this is not a good approximation for the rotameric degrees of freedom in the side chains of residues. In the frequently studied [208, 210] example of HCO-L-Ala-NH_2 (see fig. 4.8), as it has already been said in footnote 94, the side chain degree of freedom χ must be regarded as *soft*. Still, although it is a more complex task and one is not a priori entitled to write eq. (4.2), a soft degree of freedom may also be averaged out if it is considered convenient.

In this section, we will assume that the energy of the formyl-L-alanine-amide dipeptide may be correctly approximated by a *Potential Energy Hypersurface* (PEHS) (denoted by V_3) that depends on the Ramachandran angles ϕ and ψ but also on the principal dihedral χ that describes the rotation of the methyl group in the side chain. Analogously to eq. (4.1), its definition in terms of the whole energy of the system is

$$V_3(\phi,\psi,\chi) := \min_{q'^I} V_{3n-6}(\phi,\psi,\chi,q'^I) \,, \tag{4.3}$$

where q'^I represents the internal coordinates that are neither ϕ, ψ nor χ.

Note, in addition, that the two definitions are related by the following expression:

$$V_2(\phi,\psi) = \min_{\chi} V_3(\phi,\psi,\chi) \,. \tag{4.4}$$

We will also assume for $V_3(\phi,\psi,\chi)$ the aforementioned approximations that led to eq. (4.2) (in this case, the hardness of the q'^I and the indepedence from ϕ, ψ and χ of any of the determinants associated to the imposition of constraints on them), in such a way that we can write

$$Z \simeq \int \exp\left[-\beta V_3(\phi,\psi,\chi)\right] \mathrm{d}\phi\,\mathrm{d}\psi\,\mathrm{d}\chi = \int \exp\left[-\beta W(\phi,\psi)\right] \mathrm{d}\phi\,\mathrm{d}\psi \,, \tag{4.5}$$

where we have defined (following the analogous prescription to eq. (1.3)) the *effective potential energy* as

$$W(\phi,\psi) := -RT \ln \int \exp\left[-\beta V_3(\phi,\psi,\chi)\right] \mathrm{d}\chi \,. \tag{4.6}$$

This, what is the same as the integration out of the water degrees of freedom performed in sec. 1.4, is what must be done in general when a soft degree of freedom is needed to

be averaged out in statistical mechanics [113] and the hardness hypothesis behind the approximations studied in chapter 6 cannot be assumed.

We must remark at this point that, to integrate out the side chain angle χ could be reasonable if one's aim is to use the ab initio obtained information from a single dipeptide to include it in an effective potential for simulating polypeptides. With this objective in mind, there is no point in integrating out the Ramachandran angles ϕ and ψ, since the conformation of the larger system will depend crucially on their particular values, because they lie in the backbone of the molecule and there are as many pairs (ϕ, ψ) as residues in the chain. The side chain angle χ, on the contrary, will only influence its immediate surroundings and its importance could be of different magnitude depending on the treatment that the side chains are given in the model for the polypeptide.

In this context, if we wanted to use an energy function that does not depend on χ (in some circumstances, a computational must), we would have to perform the integral in the last term of eq. (4.6) and use $W(\phi, \psi)$ instead of $V_2(\phi, \psi)$, since, as it has already been remarked, χ is not a hard coordinate and the hypotheses needed to write eq. (4.2) do not hold. Therefore, if we compare the last term in eq. (4.5) with the last term in eq. (4.2), we see that, apart from additive constants that do not depend on ϕ and ψ and that come from the multiplicative constants omitted, the PES $V_2(\phi, \psi)$ must be understood as a candidate for *approximating* the more realistic $W(\phi, \psi)$ and saving much computational effort.

In the following subsections, the validity of this approximation will be assessed in the particular case of formyl-L-alanine-amide with ab initio quantum mechanics calculations.

4.4.2 Methods

The ab initio quantum mechanical calculations have been done with the package GAMESS [305] under Linux. The set of coordinates used for the HCO-L-Ala-NH_2 dipeptide in the GAMESS input files and the ones used to 'move' the molecule in the the automatic Perl scripts that generated the input files are the SASMIC defined in sec. 4.2. They are presented in table 4.5 indicating the name of the conventional dihedral angles (see also fig. 4.8 for reference). In the energy optimizations, on the contrary, they have been converted to *delocalized coordinates* [301] to accelerate convergence.

As a first step to perform the assessement described in the previous section, we have calculated the typical PES $V_2(\phi, \psi)$ defined in eq. (4.1) in a regular 12×12 grid, with both ϕ and ψ ranging from -165^o to 165^o in steps of 30^o. This has been done by running energy optimizations at the RHF/6-31+G(d) level of the theory, freezing the two Ramachandran angles at each value of the grid, starting from geometries previously optimized at a lower level of the theory and setting the gradient convergence criterium to `OPTTOL`=0.0001 and the self-consistent Hartree-Fock convergence criterium to `CONV`=0.00001.

Then, at each grid point, we have defined another one-dimensional grid in the coordinate χ that ranges from $\chi_0(\phi, \psi) - 50^o$ to $\chi_0(\phi, \psi) + 60^o$ in steps of 10^o, where $\chi_0(\phi, \psi)$ is one of the three equivalent equilibrium values (selected arbitrarily) of this degree of freedom at each point of the original PES. This partition in 12 points spans one third of the χ-space, but it is enough for computing the integrals because the surface $V_3(\phi, \psi, \chi)$ has exact three-fold symmetry in χ (note, for example, that the value of V_3 at $\chi_0(\phi, \psi) - 60^o$ would be equal to the one at $\chi_0(\phi, \psi) + 60^o$). Next, we have run energy optimizations, with the same parameters described above and at the same level of theory, at each point

Atom name	Bond length	Bond angle	Dihedral angle
H_1			
C_2	(2,1)		
N_3	(3,2)	(3,2,1)	
O_4	(4,2)	(4,2,1)	(4,2,1,3)
C_5	(5,3)	(5,3,2)	ω_0 :=**(5,3,2,1)**
H_6	(6,3)	(6,3,2)	(6,3,2,5)
C_7	(7,5)	(7,5,3)	ϕ :=**(7,5,3,2)**
C_8	(8,5)	(8,5,3)	(8,5,3,7)
H_9	(9,5)	(9,5,3)	(9,5,3,7)
H_{10}	(10,8)	(10,8,5)	χ :=**(10,8,5,3)**
H_{11}	(11,8)	(11,8,5)	(11,8,5,10)
H_{12}	(12,8)	(12,8,5)	(12,8,5,10)
N_{13}	(13,7)	(13,7,5)	ψ :=**(13,7,5,3)**
O_{14}	(14,7)	(14,7,5)	(14,7,5,13)
H_{15}	(15,13)	(15,13,7)	ω_1 :=**(15,13,7,5)**
H_{16}	(16,13)	(16,13,7)	(16,13,7,15)

Table 4.5: SASMIC internal coordinates in Z-matrix form of the protected dipeptide HCO-L-Ala-NH_2 (see also appendix E). Principal dihedrals are indicated in bold face and their typical biochemical name is also given.

of the χ-grid for every grid-value of the PES (i.e., freezing the three angles). The starting geometries have been automatically generated via Perl scripts taking the final geometries in the (ϕ,ψ)-grid and systematically changing χ. Note that this amounts to only changing the principal dihedral (10,8,5,3) in the Z-matrix in table 4.5; with poorly designed coordinates that did not separate the hard modes from the soft ones, this process would have been more difficult and rather unnatural.

After all the optimizations ($\sim$ 54 days of CPU time in 3.20 GHz PIV machines), we have 12×12×12=1728 points with grid coordinates (ϕ_i,ψ_j,χ_k), $i,j,k = 1\ldots12$ of the function $V_3(\phi,\psi,\chi)$ and we may approximate the integral defining $W(\phi,\psi)$ in eq. (4.6) by a finite sum:

$$W(\phi_i,\psi_j) := -RT\ \ln\left(\sum_k \exp\left[-\beta V_3(\phi_i,\psi_j,\chi_k)\right]\right) =$$
$$= -RT\ \ln\left(\sum_k \exp\left[-\beta\Big(V_3(\phi_i,\psi_j,\chi_k) - \langle V_3\rangle(\phi_i,\psi_j)\Big)\right]\right) + \langle V_3\rangle(\phi_i,\psi_j)\,, \qquad (4.7)$$

where the additive constants produced by the three-fold symmetry in the coordinate χ have been omitted, and the quantity $\langle V_3\rangle(\phi,\psi)$, defined as

$$\langle V_3\rangle(\phi_i,\psi_j) := \frac{1}{12}\sum_k V_3(\phi_i,\psi_j,\chi_k)\,, \qquad (4.8)$$

has been introduced in order for the values of the exponential function to be in the precision range of the computer.

Now, in order to gain further insight about the influence of the removed degree of freedom χ, we recall the possibility, mentioned in sec. 1.4 of interpreting the effective potential energy W as a free energy and define the finite-sum approximation of the associated average energy by

$$U(\phi_i,\psi_j) := \frac{\sum_k V_3(\phi_i,\psi_j,\chi_k)\ \exp\left[-\beta V_3(\phi_i,\psi_j,\chi_k)\right]}{\sum_k \exp\left[-\beta V_3(\phi_i,\psi_j,\chi_k)\right]} = \\ = \frac{\sum_k V_3(\phi_i,\psi_j,\chi_k)\ \exp\Big[-\beta(V_3(\phi_i,\psi_j,\chi_k)-\langle V_3\rangle(\phi_i,\psi_j))\Big]}{\sum_k \exp\Big[-\beta(V_3(\phi_i,\psi_j,\chi_k)-\langle V_3\rangle(\phi_i,\psi_j))\Big]}\ . \tag{4.9}$$

Finally, using the same image, we may extract the entropy related to the loss of information due to the integration of χ from the following expression:

$$W(\phi_i,\psi_j) = U(\phi_i,\psi_j) - TS(\phi_i,\psi_j)\ . \tag{4.10}$$

Additionally, apart from the calculations needed to integrate out χ, we have also performed an unconstrained geometry optimization in the basin of attraction of the local minima of the PES normally known as γ_L or $C7_{eq}$, depending on the author [207]. This calculation has been done at the MP2/6-31++G(d,p) level of the theory and with the same values of the variables `OPTTOL` and `CONV` than the ones used in the PES case. The starting geometry has been the final structure corresponding to the point $(-75^o, 75^o)$ of the PES calculations at the lower level of the theory described in the preceding paragraphs.

In the local minimum found in this way, we have computed the Hessian matrix (also at MP2/6-31++G(d,p)) in two different sets of coordinates: the properly defined SASMIC shown in table 4.5 and an ill-defined set in which the lines corresponding to the hydrogens H_{10}, H_{11} and H_{12} in the side chain have been substituted by those in table 4.1. This has been done to numerically illustrate the better separation of the hard and soft modes achieved by the internal coordinates defined in this chapter with respect to other Z-matrix-like coordinates.

4.4.3 Results

In order to assess whether $V_2(\phi,\psi)$ could be considered a good approximation of $W(\phi,\psi)$, we have used the physically meaningful distance introduced in chapter 3. If we measure it between $W(\phi,\psi)$ and $V_2(\phi,\psi)$, using the 144 points in the (ϕ,ψ)-grid, we obtain

$$d(W,V_2) = 0.098\ RT\ . \tag{4.11}$$

We have argued in sec. 3.5 that, if the distance between two different approximations of the energy of the same system is less than RT, one may safely substitute one by the

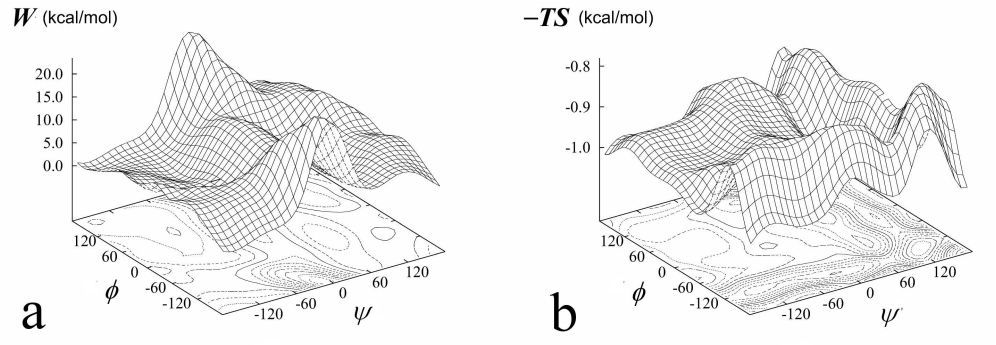


Figure 4.9: Ramachandran plots of **(a)** the effective potential energy $W(\phi,\psi)$ and **(b)** $-TS(\phi,\psi)$ in the model dipeptide HCO-L-Ala-NH_2.

other without altering the relevant physical properties. In this case, this criterium is widely satisfied.

Moreover, if one assumes that the effective energy studied will be used to construct a polypeptide potential and that the latter will be designed as simply the sum of mono-residue ones (making each term suitably depend on different pairs of Ramachandran angles), then, the number N_{res} of residues up to which one may go keeping the distance between the two approximations of the N-residue potential below RT is (see again chapter 3):

$$N_{res}(W,V_2) = \left(\frac{RT}{d(W,V_2)}\right)^2 \simeq 104\,. \tag{4.12}$$

The goodness of the approximation in this case is much due to the simplicity and small size of the side chain of the alanine residue and also to the fact that the dipeptide is isolated. For bulkier residues included in polypeptides, we expect the difference between $W(\phi,\psi)$ and $V_2(\phi,\psi)$ to be more important.

Now, although the essential result is the one stated in the previous paragraphs, we wanted to look in more detail at the origin of the differences between $W(\phi,\psi)$ and $V_2(\phi,\psi)$. For this, we have first subtracted from $W(\phi,\psi)$, $U(\phi,\psi)$ and $V_2(\phi,\psi)$ the same constant reference (min $W(\phi,\psi)$)[100] in order to render the numerical values more manageable and to minimize the statistical error of the y-intercept in the linear fits [277, 278] that will be made in the following.

Then, fitting $U(\phi,\psi)$ against $V_2(\phi,\psi)$, we have found that they are more correlated than $W(\phi,\psi)$ and $V_2(\phi,\psi)$ (compare the Pearson's correlation coefficient, $r(U,V_2) = 0.999999$ vs. $r(W,V_2) = 0.999954$, and the aforementioned distance, $d(U,V_2) = 0.015\ RT$ vs. $d(F,V_2) = 0.098\ RT$), and that they are separated by an almost constant offset: $V_2(\phi,\psi)$ is ~ 0.3 kcal/mol lower that $U(\phi,\psi)$ (on the other hand, $V_2(\phi,\psi)$ is ~ 0.6 kcal/mol higher than $W(\phi,\psi)$). Hence, the three Ramachandran surfaces $W(\phi,\psi)$, $U(\phi,\psi)$ and $V_2(\phi,\psi)$ are very similar, except for an offset. In fig. 4.9a, $W(\phi,\psi)$ is depicted graphically and, in fig 4.10, the relative offsets among the three energies are schematically shown.

[100] At the level of the theory used in the calculations, the minimum of $W(\phi,\psi)$ in the grid is −414.7985507934 hartree.

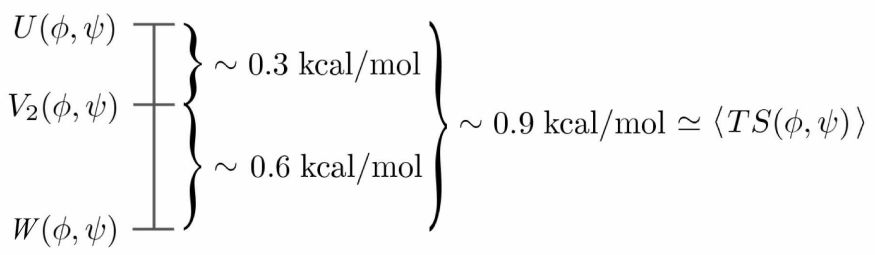


Figure 4.10: Relative offsets among the thermodynamical surfaces involved in the study.

Contrarily, the entropy (we use $TS(\phi,\psi)$ in order to deal with quantities that have units of energy), which may be found in fig. 4.9b, and whose average magnitude is ~ 0.9 kcal/mol, is almost uncorrelated with $W(\phi,\psi)$, $U(\phi,\psi)$ and $V_2(\phi,\psi)$, being the correlation coefficients $r(TS,W) = 0.382$, $r(TS,U) = 0.379$ and $r(TS,V_2) = 0.381$, respectively. Hence, given that $d(U,V_2)$ is almost an order of magnitude lower than $d(W,V_2)$, it is reasonable to conclude that the greatest part of the (little) noise between $W(\phi,\psi)$ and $V_2(\phi,\psi)$ comes from the entropic term $-TS(\phi,\psi)$. This is supported by the fact that the difference $W(\phi,\psi) - V_2(\phi,\psi)$ is highly correlated with $TS(\phi,\psi)$, being the correlation coefficient $r(W - V_2, TS) = 0.998$.

Finally, and in order to illustrate the better separation of the hard and soft modes achieved by the internal coordinates defined in this chapter, we have calculated the Hessian matrix in the minimum γ_L (also $C7_{eq}$) in two different sets of coordinates. They are described at the end of sec. 4.4.2 and they correspond to the SASMIC set, defined according to the rules given in sec. 4.2, and a set in which the coordinates that position the hydrogens in the side chain have been ill-defined.

In fig. 4.11, we present the sub-boxes of the two Hessian matrices corresponding to the coordinates in tables 4.2 and 4.1.

Properly defined coordinates

	χ	α_1	α_2
χ	15.74	1.40	8.71
α_1	1.40	110.98	−54.23
α_2	8.71	−54.23	115.37

Ill-defined coordinates

	γ_1	γ_2	γ_3
γ_1	113.49	−55.55	−52.60
γ_2	−55.55	110.98	−54.23
γ_3	−52.60	−54.23	115.37

Figure 4.11: Sub-boxes of the Hessian matrix in the minimum γ_L (also $C7_{eq}$) corresponding to the coordinates defined in tables 4.2 and 4.1. The quantities are expressed in kcal/mol · rad^{-2}. See the text for more details.

From the values shown, one can conclude that, in the 'properly defined coordinates', some convenient characteristics are present: on one side, the relatively low values of the elements $H_{\chi\alpha_1}$ and $H_{\chi\alpha_2}$ (and their symmetric ones) indicate that the soft degree of freedom χ and the hard ones α_1 and α_2, which describe the internal structure of the methyl group, are uncoupled to a reasonable extent; on the other side, the relatively low value of $H_{\chi\chi}$ compared to $H_{\alpha_1\alpha_1}$ and $H_{\alpha_2\alpha_2}$ (a difference of almost an order of magnitude) proves that χ

may be regarded as soft when compared to the hard degrees of freedom α_1 and α_2.

On the contrary, in the 'ill-defined coordinates', the three dihedrals are hard, considerably coupled and equivalent.

4.5 Conclusions

Extending the approach of refs. 310–312 and the ideas stated in refs. 2, 299, 300, we have defined in this chapter a systematic numeration of the groups, the atoms and the internal coordinates (termed SASMIC) of polypeptide chains. The advantages of the rules herein presented are many-fold:

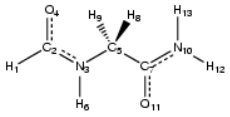
- The internal coordinates may be easily cast into conventional Z-matrix form and they can be directly implemented into quantum chemical packages.
- The algorithm for numbering allows for automatizing and facilitates the coding of computer applications.
- The modularity of the numeration system in the case of polypeptides permits the addition of new residues without essentially changing the already numbered items. This is convenient if databases of peptide structures need to be designed.
- The set of internal coordinates defined reasonably separate the hard and soft movements of polypeptides for arbitrary conformations using only topological information.

A number of Perl scripts that automatically generate these coordinates for polypeptide chains have been developed and may be found at `http://neptuno.unizar.es/files/public/gen_sasmic/`. Also, a slightly modified set of rules is provided that allows to number the groups, atoms and define the internal coordinates in general organic molecules.

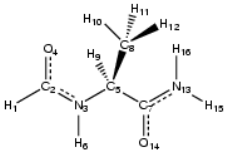
Finally, we have used the coordinates herein defined and ab initio quantum mechanics to assess the approximation via the conventional PES of the effective potential energy obtained from averaging out the rotameric degree of freedom χ in the protected dipeptide HCO-L-Ala-NH_2. Applying the criterium in chapter. 3, we have found that approximating $W(\phi,\psi)$ by $V_2(\phi,\psi)$ is justified up to polypeptides of medium length ($\sim$ 100 residues) and much computational effort may be saved using the PES instead of the more realistic effective potential energy. However, the small size of the side chain of the alanine residue and the fact that the dipeptide is isolated do not allow to extrapolate this result. For bulkier residues included in polypeptides, we expect the difference between $W(\phi,\psi)$ and $V_2(\phi,\psi)$ to be more important.

Figure 4.12: Numeration of the dipeptides HCO-L-Xxx-NH_2 (see appendix E), where Xxx runs on the twenty naturally occurring amino acids. Uncharged side chains are displayed and histidine is shown in its ε_2-tautomeric form. See also fig. 1.7 for reference.

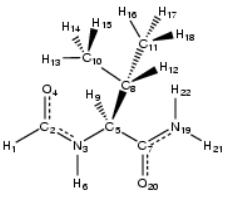
Aliphatic



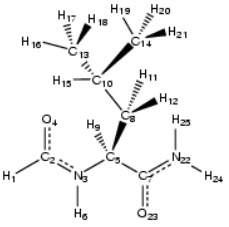
Glycine - G - Gly



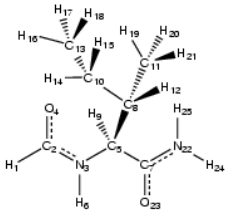
Alanine - A - Ala



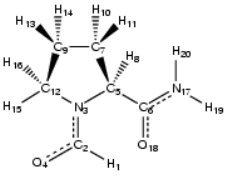
Valine - V - Val



Leucine - L - Leu

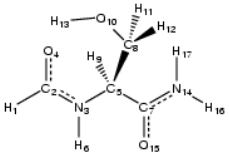


Isoleucine - I - Ile

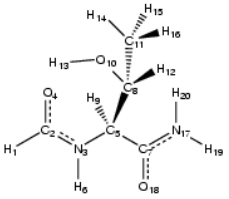


Proline - P - Pro

Alcohols

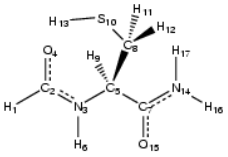


Serine - S - Ser

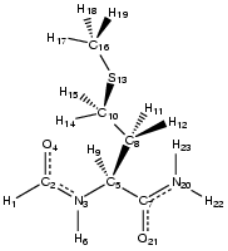


Threonine - T - Thr

Sulphur-containing

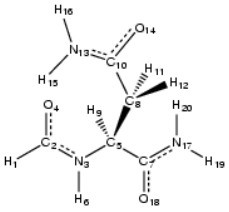


Cysteine - C - Cys

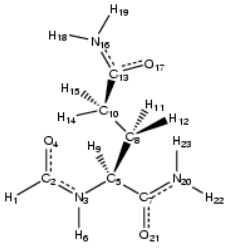


Methionine - M - Met

Amides

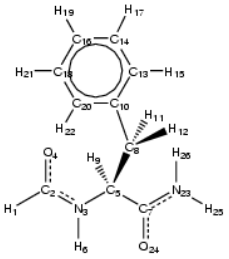


Asparagine - N - Asn

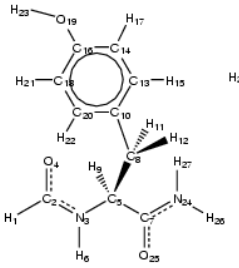


Glutamine - Q - Gln

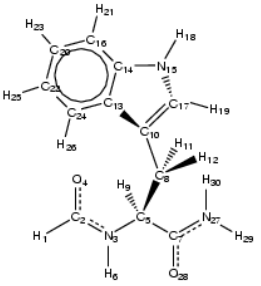
Aromatic



Phenylalanine - F - Phe

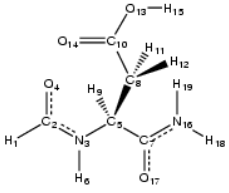


Tyrosine - Y - Tyr



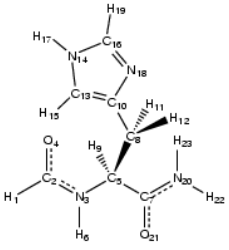
Tryptophan - W - Trp

Acids

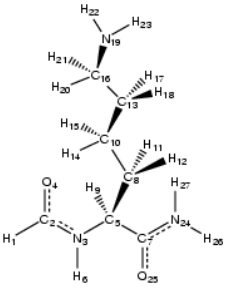


Aspartic Acid - D - Asp

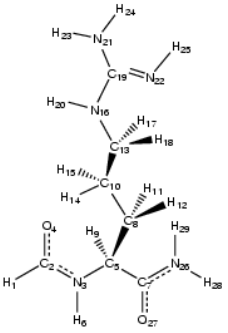
Bases



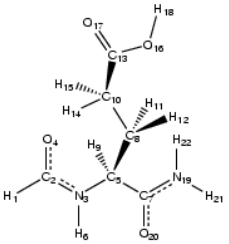
Histidine - H - His



Lysine - K - Lys



Arginine - R - Arg



Glutamic Acid - E - Glu

Chapter 5

Explicit factorization of external coordinates in constrained statistical mechanics models*

This chapter is based on the article:
Pablo Echenique and Iván Calvo, *Explicit factorization of external coordinates in constrained Statistical Mechanics models*, J. Comp. Chem. **27** (2006) 1748–1755.

Man muß immer generalisieren.
(*One should always generalize.*) [320]

— Carl Jacobi

5.1 Introduction

Monte Carlo simulations are among the most useful tools for studying the behaviour of macromolecules in thermal equilibrium [321–328]. Typically, the simulations are carried out in the *coordinate space* Ω, i.e., the momenta are averaged out and Monte Carlo movements that only change the coordinates of the system are designed (see sec. 1.4 and appendix A).

Moreover, the most interesting properties of macromolecules depend only on conformational transitions in the *internal subspace* (or *conformational space*) $\mathcal{I}$ of the whole coordinate space. The protein folding problem discussed in sec. 1.3, the docking of ligands to proteins [329], or proteins to proteins [330], the prediction of Raman [331, 332], IR [333, 334], CD [335], VCD [336, 337], NMR [36, 37] spectra, etc. are tasks that require knowledge of the probability density in the conformational space only, i.e., having averaged out the external coordinates that describe overall translations and rotations of the system.

* In this chapter, a series of mathematical results are proved about the factorization of certain determinants that appear in the equilibrium conformational probability of constrained systems. The more practice-oriented readers may want to directly jump to chapter 6, where the expressions herein derived are briefly recalled and applied to a practical case.

If Euclidean coordinates are used, the integration over the momenta produces a constant factor (which depends on the temperature T but does not depend on the coordinates[101]) and the marginal probability density in the coordinate space Ω resembles the common Boltzmann weight $e^{-\beta H}$ but using the potential energy $V(x^\mu)$ instead of the whole energy (see eq. (1.7)):

$$p_{\mathrm{c}}(x^\mu) = \frac{\exp\left[-\beta V(x^\mu)\right]}{\displaystyle\int_\Omega \exp\left[-\beta V(x^\nu)\right] \mathrm{d}x^\nu} \,. \tag{5.1}$$

In the absence of external fields, the potential energy does not change under global translations and rotations of the system. In addition, as we have already mentioned, one is normally not interested in averages of observables that depend on these degrees of freedom. Hence, it would be convenient to average them out from eq. (5.1). However, this cannot be done in Euclidean coordinates: one must use a set of coordinates adapted to overall translations and rotations.

In the simulation of macromolecules, it is customary [176, 299, 310–313, 321] to define a set of curvilinear coordinates q^μ in which the first six ones, denoted by q^A, are called *external coordinates* and parameterize the *external subspace* $\mathcal{E}$, i.e., the system overall translation, specifying the position of a selected point (normally an atom), and rotation, via three Euler angles (see sec. 5.2). The remaining $3n-6$ coordinates (where n is the number of mass points or *atoms*) are called *internal coordinates* and will be denoted herein by q^a (for concreteness, we can imagine that these q^a are the SASMIC coordinates introduced in the previous chapter).

The change of coordinates from Euclideans to internals modifies the mass-metric tensor in the kinetic energy. Thus, when the momenta are averaged out and the marginal probability density in the whole coordinate space Ω is considered, the square root of the determinant of the mass-metric tensor (which now does depend on the coordinates; see also appendix A) shows up in the probability density function[102]:

$$p_{\mathrm{w}}(q^\mu) = \frac{\det{}^{\frac{1}{2}} G(q^A, q^a)\, \exp\left[-\beta V(q^a)\right]}{\displaystyle\int_\Omega \det{}^{\frac{1}{2}} G(q^B, q^b)\, \exp\left[-\beta V(q^b)\right] \mathrm{d}q^B \mathrm{d}q^b} \,. \tag{5.2}$$

More interestingly, if holonomic constraints are imposed on the system (the so-called *classical rigid* model [338, 339]), the reduced mass-metric tensor on the *constrained hypersurface* $\mathcal{E} \times \Sigma$, where Σ denotes the constrained *internal* hypersurface, appears in the kinetic energy. Hence, when the momenta are integrated out from the joint probability

[101] If no confusion is possible, the word *coordinates*, which is more common in biochemical context, will be used in this dissertation to mean the 'non-momenta part of the phase space', i.e., the x^μ in this case and the q^μ later on. When some ambiguity is present or we want to explicitly stress the difference between the q^μ and the momenta, the more precise term *positions* will be used for the former.

[102] In chapter 6, a careful derivation of this expression and of the rest of statistical mechanics formulae in this section is presented. In this introduction, we give a brief advance for convenience.

density in the phase space, the square root of its determinant occurs:

$$p_{\mathrm{r}}(q^u) = \frac{\det{}^{\frac{1}{2}} g(q^A, q^i)\, \exp\left[-\beta V_\Sigma(q^i)\right]}{\displaystyle\int_{\mathcal{E}\times\Sigma} \det{}^{\frac{1}{2}} g(q^B, q^j)\, \exp\left[-\beta V_\Sigma(q^j)\right] \mathrm{d}q^B \mathrm{d}q^j} \,, \tag{5.3}$$

where V_Σ stands for the potential energy in Σ, $q^u \equiv (q^A, q^i)$ denotes the *soft coordinates*, among which the external ones q^A are included, and q^i denotes the *soft internal coordinates*.

If, on the other hand, the constraints are imposed via a steep potential that energetically penalizes the conformations that leave the constrained hypersurface (the so-called *classical stiff* model [313, 338, 339]), the probability density is the same as in eq. (5.2) except for the determinant of the Hessian matrix $\mathcal{H}$ of the constraining part of the potential V that appears when the *hard coordinates* q^I are averaged out and for the fact that all the functions are evaluated on the constrained hypersurface $\mathcal{E} \times \Sigma$, consequently depending only on the soft coordinates q^u:

$$p_{\mathrm{s}}(q^u) = \frac{\det{}^{\frac{1}{2}} G(q^A, q^i) \det{}^{-\frac{1}{2}} \mathcal{H}(q^i)\, \exp\left[-\beta V_\Sigma(q^i)\right]}{\displaystyle\int_{\mathcal{E}\times\Sigma} \det{}^{\frac{1}{2}} G(q^B, q^j) \det{}^{-\frac{1}{2}} \mathcal{H}(q^j)\, \exp\left[-\beta V_\Sigma(q^j)\right] \mathrm{d}q^B \mathrm{d}q^j} \,. \tag{5.4}$$

Finally, the *Fixman's compensating potential* [338–340], denoted by V_{F} and which is customarily used to reproduce the stiff equilibrium distribution using rigid molecular dynamics simulations [133, 321], is also expressed as a function of these determinants:

$$V_{\mathrm{F}}(q^u) := \frac{RT}{2} \ln\left[\frac{\det{}^{\frac{1}{2}} g(q^A, q^i) \det{}^{\frac{1}{2}} \mathcal{H}(q^i)}{\det{}^{\frac{1}{2}} G(q^A, q^i)}\right] . \tag{5.5}$$

Now, in the three physical models, given by eqs. (5.2), (5.3) and (5.4), and in the Fixman's potential written above, neither the potential energy nor the Hessian matrix of the constraining part of the potential depend on the external coordinates q^A. Therefore, it would be very convenient to integrate them out in order to obtain a simpler probability density depending only on the internal coordinates. Such an improvement may render the coding of Monte Carlo computer simulations and also the visualization of molecules easier, since all the movements in the molecule may be performed fixing an atom in space and keeping the orientation of the system with respect to a set of axes fixed in space constant [341].

One may argue that it is not clear that much computational effort will be saved for macromolecules, since the gain expected when reducing the degrees of freedom from $M + 6$ to M (where M is the number of soft internal coordinates) is only appreciable if M is small and the difficulties arising from the use of curvilinear coordinates may well be more important. However, in cases where the use of curvilinear coordinates is a must, such as the simulation of constrained systems, the calculations in this work allow to save a time (which will depend on the size of the system) that, otherwise, would be wasted in movements of the external coordinates. If, for example, the molecule treated contains a few tens of atoms and most of the degrees of freedom are constrained, as it is common in ab initio quantum mechanical calculations in model peptides [207, 210–213], the relevance of omitting the externals may be considerable.

While in the second practical example in sec. 3.10 and in chapter 7, all the determinants arising from the imposition of constraints are not taken into account, and in the practical example in sec. 4.4 they are assumed to be negligible, in chapter 6 they are calculated and their influence is assessed. There, the model dipeptide HCO-L-Ala-NH_2 (see appendix E) is studied and the soft internals are two: the Ramachandran angles ϕ and ψ; hence, $M = 2$ and the formulae in this chapter have permitted to reduce the number of degrees of freedom from 8 to 2.

Now, for the resulting expressions to be manageable and the integration out process be straightforward, the determinant of the mass-metric tensor G, in p_w and p_s (see eqs. (5.2) and (5.4) respectively), and the determinant of g, in the rigid case (in eq. (5.3)), should *factorize* as a product of a function that depends only on the external coordinates and another function that depends only on the internal ones. Then, the function depending on the external coordinates, could be integrated out in the probability densities p_w, p_r and p_s or taken out of the logarithm in V_F[103].

For some simple cases, it has already been proven in the literature that this factorization actually happens. In ref. 313, for example, the determinant of G is shown to factorize for a serial polymer in a particular set of curvilinear coordinates. In ref. 342, the determinant of g is shown to factorize for the same system, in similar coordinates, with frozen bond lengths and bond angles.

In this chapter, we *generalize* these results, showing that they hold in *arbitrary* internal coordinates (for general branched molecules) and with *arbitrary* constraints. Perhaps more importantly, we provide explicit expression for the functions involved in the factorization. It is worth remarking that, although the calculations herein have been performed thinking in macromolecules as target system, they are completely general and applicable to any classical system composed by discrete mass points.

In sec. 5.2, we present the notation and conventions that will be used throughout the chapter and also in chapter 6. In sec. 5.3, we explicitly factorize the determinant of the reduced mass-metric tensor g as a product of a function that depends only on the external coordinates and another function that depends on arbitrary internal coordinates. In sec. 5.4, we perform the analogous calculations for the determinant of the mass-metric tensor G. In sec. 5.5, the determinant of the mass-metric tensor G is computed in the SASMIC set of curvilinear coordinates introduced in chapter 4, which, also in this point, turn out to be convenient for dealing with general branched molecules. Moreover, we show that the classical formula for serial polymers [313] is actually valid for any macromolecule. Finally, sec. 5.6 is devoted to the conclusions, and in appendix D, a general mathematical argument underlying the results in this chapter is given.

5.2 General set-up and definitions

This section is devoted to introduce certain notational conventions that will be used extensively in the rest of this chapter and also in chapter 6.

- The system under scrutiny will be a set of n mass points termed *atoms*. The Euclidean coordinates of the atom α in a set of axes fixed in space are denoted by $\vec{x}_\alpha$.

[103] What really happens, (see secs. 5.3 and 5.4) is that the factor that depends on the external coordinates is the same for $\det G$ and $\det g$. Hence, it divides out in eq. (5.5) (see sec. 5.6).

The subscript α runs from 1 to n.

- The curvilinear coordinates suitable to describe the system will be denoted by q^μ, $\mu = 1, \ldots, 3n$ and the set of Euclidean coordinates by x^μ when no explicit reference to the atoms index needs to be made. We shall often use $N := 3n$ for the total number of degrees of freedom, and the whole *coordinate space*, parameterized by either the x^μ or the q^μ, will be denoted by Ω.

- We choose the coordinates q^μ so that the first six are *external coordinates*. They are denoted by q^A and their ordering is $q^A \equiv (X, Y, Z, \phi, \theta, \psi)$. The first three ones, $\vec{X}^T := (X, Y, Z)$[104], describe the overall position of the system. The three angles (ϕ, θ, ψ) are related to its overall orientation. More concretely, they give the orientation of a frame fixed in the system with respect to the frame fixed in space, and they parameterize the *external subspace* $\mathcal{E}$.

- To define the set of axes *fixed in the system*, we select three atoms (denoted by 1, 2 and 3) in such a way that $\vec{X}$ is the position of atom 1 (i.e., $\vec{x}_1 = \vec{X}$). The orientation of the fixed axes (x', y', z') is chosen such that atom 2 lies in the positive half of the z'-axis and atom 3 is contained in the (x', z')-plane, in the $x' > 0$ semiplane (see fig. 5.1). The position of atom α in these axes is denoted by $\vec{x}'_\alpha$.

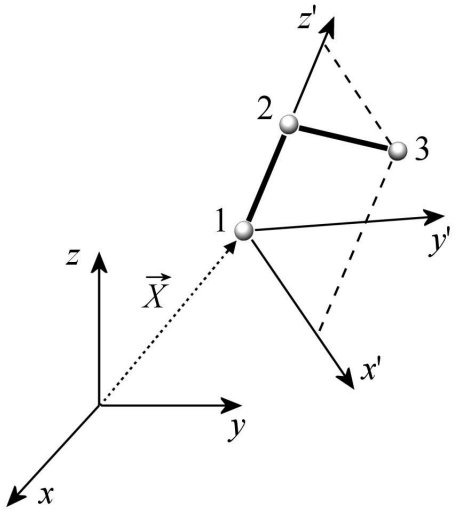


Figure 5.1: Definition of the axes fixed in the system.

- Let $E(\phi, \theta, \psi)$ be the Euler rotation matrix (in the ZYZ convention [343]) that takes a vector of primed components $\vec{a}'$ to the frame fixed in space, i.e., $\vec{a} = E(\phi, \theta, \psi)\,\vec{a}'$. Its explicit expression is the following:

$$E(\phi,\theta,\psi) = \underbrace{\begin{pmatrix} \cos\phi & -\sin\phi & 0 \\ \sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{pmatrix}}_{\Phi(\phi)} \underbrace{\begin{pmatrix} -\cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & -\cos\theta \end{pmatrix}}_{\Theta(\theta)} \underbrace{\begin{pmatrix} \cos\psi & -\sin\psi & 0 \\ \sin\psi & \cos\psi & 0 \\ 0 & 0 & 1 \end{pmatrix}}_{\Psi(\psi)} . \tag{5.6}$$

[104] The superindex T indicates matrix transposition. By $\vec{a}^T$ we shall understand the row vector (a^1, a^2, a^3).

The unusual minus signs of the cosines in the diagonal of matrix $\Theta(\theta)$ come from the fact that, due to frequent biochemical conventions, the rotation with respect to the y-axis is of angle $\tilde{\theta} := \pi - \theta$.

- The coordinates q^μ are split into (q^A, q^a), $a = 7, \ldots, N$. The coordinates q^a are said *internal coordinates* and determine the positions of the atoms in the frame fixed in the system. The transformation from the Euclidean coordinates $\vec{x}_\alpha$ to the curvilinear coordinates q^μ may be written as follows:

$$\vec{x}_\alpha = \vec{X} + E(\phi, \theta, \psi)\, \vec{x}'_\alpha(q^a) \qquad \alpha = 1, \ldots, n\,. \tag{5.7}$$

- The coordinates q^a parameterize what we shall call the *internal subspace*, denoted by $\mathcal{I}$ (so that $\Omega = \mathcal{E} \times \mathcal{I}$). Assume that L independent constraints are imposed on $\mathcal{I}$, so that only points on a hypersurface $\Sigma \subset \mathcal{I}$ of dimension $M := N - L - 6$ are allowed. Then, we choose a splitting $q^a \equiv (q^i, q^I)$, with $i = 7, \ldots, M + 6$ and $I = M + 7, \ldots, N$, where q^i (the *internal soft coordinates*) parameterize Σ, and q^I (the *hard coordinates*) are functions of the soft coordinates:

$$q^I = f^I(q^i) \qquad I = M + 7, \ldots, N\,. \tag{5.8}$$

Now, if these constraints are used, together with eq. (5.7), the Euclidean position of any atom may be parameterized with the set of *all soft coordinates*, denoted by $q^u \equiv (q^A, q^i)$, with $u = 1, \ldots, M + 6$, as follows:

$$\vec{x}_\alpha = \vec{X} + E(\phi, \theta, \psi)\, \vec{x}'_\alpha(q^i, f^I(q^i)) \qquad \alpha = 1, \ldots, n\,. \tag{5.9}$$

- In table 5.1, a summary of the indices used is given.

Indices	Range	Number	Description
$\alpha, \beta, \gamma, \ldots$	$1, \ldots, n$	n	Atoms
$p, q, r, s, \ldots$	$1, 2, 3$	3	Components of trivectors
$\mu, \nu, \rho, \ldots$	$1, \ldots, N$	$N = 3n$	All coordinates
$A, B, C, \ldots$	$1, \ldots, 6$	6	External coordinates
$a, b, c, \ldots$	$7, \ldots, N$	$N - 6$	Internal coordinates
$i, j, k, \ldots$	$7, \ldots, M + 6$	M	Soft internal coordinates
$I, J, K, \ldots$	$M + 7, \ldots, N$	$L = N - M - 6$	Hard internal coordinates
$u, v, w, \ldots$	$1, \ldots, M + 6$	$M + 6$	All soft coordinates

Table 5.1: Definition of the indices used.

5.3 Constrained case

The *reduced mass-metric tensor*, in the constrained hypersurface Σ plus the external subspace spanned by the q^A, may be written as follows:

$$g_{vw}(q^u) := \sum_{\mu=1}^{N} \frac{\partial x^\mu(q^u)}{\partial q^v} m_\mu \frac{\partial x^\mu(q^u)}{\partial q^w} \,. \tag{5.10}$$

In matrix notation, this is written as

$$g = J_c^T M J_c \,, \tag{5.11}$$

where c stands for *constrained* and M is the diagonal $N \times N$ *mass matrix* given by

$$M := \begin{pmatrix} m_1^{(3)} & & 0 \\ & \ddots & \\ 0 & & m_n^{(3)} \end{pmatrix}, \quad \text{with} \quad m_\alpha^{(3)} := m_\alpha \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}}_{I^{(3)}} . \tag{5.12}$$

Using eq. (5.9) and noting that the derivatives with respect to the externals, q^A, only affect the $\vec{X}$ vector and the Euler rotation matrix E, while differentiation with respect to soft internals, q^i, only act on the $\vec{x}'_\alpha$, we have that J_c is the $N \times (M+6)$ matrix

$$J_c = \left(\begin{array}{cccc|c} I^{(3)} & & 0 & & 0 \\ I^{(3)} & \frac{\partial E}{\partial \phi}\vec{x}'_2 & \frac{\partial E}{\partial \theta}\vec{x}'_2 & \frac{\partial E}{\partial \psi}\vec{x}'_2 & \cdots\ E\frac{\partial \vec{x}'_2}{\partial q^j}\ \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ I^{(3)} & \frac{\partial E}{\partial \phi}\vec{x}'_\alpha & \frac{\partial E}{\partial \theta}\vec{x}'_\alpha & \frac{\partial E}{\partial \psi}\vec{x}'_\alpha & \cdots\ E\frac{\partial \vec{x}'_\alpha}{\partial q^j}\ \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{array}\right) . \tag{5.13}$$

Now, if we perform the matrix multiplications in eq. (5.11), we obtain

$$g = \left(\begin{array}{cc|c} I^{(3)} \sum\limits_\alpha m_\alpha & \sum\limits_\alpha m_\alpha \partial E_\alpha & \cdots\ \sum\limits_\alpha m_\alpha E \frac{\partial \vec{x}'_\alpha}{\partial q^j}\ \cdots \\ \sum\limits_\alpha m_\alpha \partial E_\alpha^T & \sum\limits_\alpha m_\alpha \partial E_\alpha^T \partial E_\alpha & \cdots \sum\limits_\alpha m_\alpha \partial E_\alpha^T E \frac{\partial \vec{x}'_\alpha}{\partial q^j} \cdots \\ \hline \vdots & \vdots & \vdots \\ \sum\limits_\alpha m_\alpha \frac{\partial \vec{x}'^{\,T}_\alpha}{\partial q^i} E^T & \sum\limits_\alpha m_\alpha \frac{\partial \vec{x}'^{\,T}_\alpha}{\partial q^i} E^T \partial E_\alpha & \cdots \sum\limits_\alpha m_\alpha \frac{\partial \vec{x}'^{\,T}_\alpha}{\partial q^i} \frac{\partial \vec{x}'_\alpha}{\partial q^j} \cdots \\ \vdots & \vdots & \vdots \end{array}\right) , \tag{5.14}$$

where all the sums in α can be understood as ranging from 1 to n if we note that $\vec{x}'_1 = \vec{0}$. Also, in the bottom right block, the fact that E is an orthogonal matrix (i.e., that $E^T E = I^{(3)}$) has been used, and we have defined the 3×3 block as

$$\partial E_\alpha := \left(\frac{\partial E}{\partial \phi}\vec{x}'_\alpha \quad \frac{\partial E}{\partial \theta}\vec{x}'_\alpha \quad \frac{\partial E}{\partial \psi}\vec{x}'_\alpha \right) . \tag{5.15}$$

We can write g as

$$g := \begin{pmatrix} E & 0 \\ 0 & I^{(M+3)} \end{pmatrix} g_1 \begin{pmatrix} E^T & 0 \\ 0 & I^{(M+3)} \end{pmatrix}, \tag{5.16}$$

where $I^{(M+3)}$ is the $(M+3)\times(M+3)$ identity matrix and g_1 is defined as

$$g_1 := \left(\begin{array}{cc|c} I^{(3)} \sum_\alpha m_\alpha & \sum_\alpha m_\alpha E^T \partial E_\alpha & \cdots \quad \sum_\alpha m_\alpha \frac{\partial \vec{x}'_\alpha}{\partial q^j} \quad \cdots \\ \sum_\alpha m_\alpha \partial E_\alpha^T E & \sum_\alpha m_\alpha \partial E_\alpha^T E E^T \partial E_\alpha & \cdots \sum_\alpha m_\alpha \partial E_\alpha^T E \frac{\partial \vec{x}'_\alpha}{\partial q^j} \cdots \\ \hline \vdots & \vdots & \vdots \\ \sum_\alpha m_\alpha \frac{\partial \vec{x}'^T_\alpha}{\partial q^i} & \sum_\alpha m_\alpha \frac{\partial \vec{x}'^T_\alpha}{\partial q^i} E^T \partial E_\alpha & \cdots \sum_\alpha m_\alpha \frac{\partial \vec{x}'^T_\alpha}{\partial q^i} \frac{\partial \vec{x}'_\alpha}{\partial q^j} \cdots \\ \vdots & \vdots & \vdots \end{array}\right). \tag{5.17}$$

Note that $I^{(3)} = EE^T$ has been introduced in the bottom right 3×3 submatrix of the top left block.

Next, we introduce some simplifying notation for the matrices $E^T \partial E_\alpha$:

$$E^T \partial E_\alpha = \left(E^T \frac{\partial E}{\partial \phi} \vec{x}'_\alpha \;\; E^T \frac{\partial E}{\partial \theta} \vec{x}'_\alpha \;\; E^T \frac{\partial E}{\partial \psi} \vec{x}'_\alpha \right) =: \left(\vec{y}^1_\alpha \; \vec{y}^2_\alpha \; \vec{y}^3_\alpha \right), \tag{5.18}$$

and, defining

$$g_1^\alpha := \left(\begin{array}{cccc|c} I^{(3)} & \vec{y}^1_\alpha & \vec{y}^2_\alpha & \vec{y}^3_\alpha & \cdots \quad \frac{\partial \vec{x}'_\alpha}{\partial q^j} \quad \cdots \\ \vec{y}^{1\,T}_\alpha & \vec{y}^{1\,T}_\alpha \vec{y}^1_\alpha & \vec{y}^{1\,T}_\alpha \vec{y}^2_\alpha & \vec{y}^{1\,T}_\alpha \vec{y}^3_\alpha & \cdots \; \vec{y}^{1\,T}_\alpha \frac{\partial \vec{x}'_\alpha}{\partial q^j} \cdots \\ \vec{y}^{2\,T}_\alpha & \vec{y}^{2\,T}_\alpha \vec{y}^1_\alpha & \vec{y}^{2\,T}_\alpha \vec{y}^2_\alpha & \vec{y}^{2\,T}_\alpha \vec{y}^3_\alpha & \cdots \; \vec{y}^{2\,T}_\alpha \frac{\partial \vec{x}'_\alpha}{\partial q^j} \cdots \\ \vec{y}^{3\,T}_\alpha & \vec{y}^{3\,T}_\alpha \vec{y}^1_\alpha & \vec{y}^{3\,T}_\alpha \vec{y}^2_\alpha & \vec{y}^{3\,T}_\alpha \vec{y}^3_\alpha & \cdots \; \vec{y}^{3\,T}_\alpha \frac{\partial \vec{x}'_\alpha}{\partial q^j} \cdots \\ \hline \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{\partial \vec{x}'^T_\alpha}{\partial q^i} & \frac{\partial \vec{x}'^T_\alpha}{\partial q^i} \vec{y}^1_\alpha & \frac{\partial \vec{x}'^T_\alpha}{\partial q^i} \vec{y}^2_\alpha & \frac{\partial \vec{x}'^T_\alpha}{\partial q^i} \vec{y}^3_\alpha & \cdots \; \frac{\partial \vec{x}'^T_\alpha}{\partial q^i} \frac{\partial \vec{x}'_\alpha}{\partial q^j} \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{array}\right), \tag{5.19}$$

we have that

$$g_1 = \sum_\alpha m_\alpha g_1^\alpha \,. \tag{5.20}$$

Now, the vectors $\vec{y}_\alpha^{\,p}$ may be extracted from g_1^α as follows:

$$g_1^\alpha = Y_\alpha^T \left(\begin{array}{cc|c|c} I^{(3)} & I^{(3)} & \cdots & \dfrac{\partial \vec{x}'_\alpha}{\partial q^j} \quad \cdots \\ I^{(3)} & I^{(3)} & \cdots & \dfrac{\partial \vec{x}'_\alpha}{\partial q^j} \quad \cdots \\ \hline \vdots & \vdots & & \vdots \\ \dfrac{\partial \vec{x}'^{\,T}_\alpha}{\partial q^i} & \dfrac{\partial \vec{x}'^{\,T}_\alpha}{\partial q^i} & \cdots & \dfrac{\partial \vec{x}'^{\,T}_\alpha}{\partial q^i}\dfrac{\partial \vec{x}'_\alpha}{\partial q^j} \cdots \\ \vdots & \vdots & & \vdots \end{array}\right) Y_\alpha \,, \tag{5.21}$$

where

$$Y_\alpha := \begin{pmatrix} I^{(3)} & 0 & 0 \\ 0 & \vec{y}_\alpha^{\,1}\, \vec{y}_\alpha^{\,2}\, \vec{y}_\alpha^{\,3} & 0 \\ 0 & 0 & I^{(M)} \end{pmatrix}, \tag{5.22}$$

and the central matrix in eq. (5.21) only depends on the soft internal coordinates. After some lengthy calculations, one shows that

$$\left(\vec{y}_\alpha^{\,1}\, \vec{y}_\alpha^{\,2}\, \vec{y}_\alpha^{\,3}\right) = \underbrace{\begin{pmatrix} 0 & -x_\alpha'^{\,3} & x_\alpha'^{\,2} \\ x_\alpha'^{\,3} & 0 & -x_\alpha'^{\,1} \\ -x_\alpha'^{\,2} & x_\alpha'^{\,1} & 0 \end{pmatrix}}_{v(\vec{x}'_\alpha)} \underbrace{\begin{pmatrix} \sin\theta\cos\psi & \sin\psi & 0 \\ -\sin\theta\sin\psi & \cos\psi & 0 \\ \cos\theta & 0 & -1 \end{pmatrix}}_{W(\theta,\psi)} . \tag{5.23}$$

Thus, the matrix Y_α in eq. (5.22) may be written as

$$Y_\alpha = \begin{pmatrix} I^{(3)} & 0 & 0 \\ 0 & v(\vec{x}'_\alpha) & 0 \\ 0 & 0 & I^{(M)} \end{pmatrix} \begin{pmatrix} I^{(3)} & 0 & 0 \\ 0 & W(\theta,\psi) & 0 \\ 0 & 0 & I^{(M)} \end{pmatrix} . \tag{5.24}$$

If we now take this expression to eq. (5.21) and use that, for any pair of vectors $\vec{a}$ and $\vec{b}$, $\vec{a}^{\,T} v(\vec{b}\,) = (\vec{a}\times\vec{b}\,)^T$ and $v^T(\vec{b}\,)\,\vec{a} = \vec{a}\times\vec{b}$, where $\times$ denotes the usual vector cross product, we may rewrite eq. (5.21) as follows:

$$g_1^\alpha = \begin{pmatrix} I^{(3)} & 0 & 0 \\ 0 & W^T(\theta,\psi) & 0 \\ 0 & 0 & I^{(M)} \end{pmatrix} g_2^\alpha \begin{pmatrix} I^{(3)} & 0 & 0 \\ 0 & W(\theta,\psi) & 0 \\ 0 & 0 & I^{(M)} \end{pmatrix} , \tag{5.25}$$

where g_2^α is defined as

$$g_2^\alpha = \left(\begin{array}{cc|c|c} I^{(3)} & v(\vec{x}'_\alpha) & \cdots & \dfrac{\partial \vec{x}'_\alpha}{\partial q^j} \quad \cdots \\ v^T(\vec{x}'_\alpha) & v^T(\vec{x}'_\alpha)v(\vec{x}'_\alpha) & \cdots & \dfrac{\partial \vec{x}'_\alpha}{\partial q^j} \times \vec{x}'_\alpha \cdots \\ \hline \vdots & \vdots & & \vdots \\ \dfrac{\partial \vec{x}'^{\,T}_\alpha}{\partial q^i} & \left(\dfrac{\partial \vec{x}'_\alpha}{\partial q^i} \times \vec{x}'_\alpha\right)^T & \cdots & \dfrac{\partial \vec{x}'^{\,T}_\alpha}{\partial q^i}\dfrac{\partial \vec{x}'_\alpha}{\partial q^j} \cdots \\ \vdots & \vdots & & \vdots \end{array}\right) . \tag{5.26}$$

At this point, we insert eq. (5.25) in eq. (5.20) and take the W matrices out of the sum. Since $\det W(\theta,\psi) = \det W^T(\theta,\psi) = -\sin\theta$, we obtain

$$\det g_1 = \sin^2\theta \, \det \underbrace{\left(\sum_\alpha m_\alpha g_2^\alpha \right)}_{g_2} . \tag{5.27}$$

Recalling that $\det E = \det E^T = 1$, from eq. (5.16), we have that

$$\det g(q^A, q^i) = \det g_1(q^A, q^i) = \sin^2\theta \, \det g_2(q^i) , \tag{5.28}$$

and the factorization of the external coordinates has been finally accomplished, since g_2 is the following matrix, which depends only on the soft internal coordinates q^i:

$$g_2 = \left(\begin{array}{cc|c} m_{tot} I^{(3)} & m_{tot}\, v(\vec{R}) & \cdots \quad m_{tot} \dfrac{\partial \vec{R}}{\partial q^j} \quad \cdots \\ m_{tot}\, v^T(\vec{R}) & \mathcal{J} & \cdots \sum\limits_\alpha m_\alpha \dfrac{\partial \vec{x}'_\alpha}{\partial q^j} \times \vec{x}'_\alpha \cdots \\ \hline \vdots & \vdots & \vdots \\ m_{tot} \dfrac{\partial \vec{R}}{\partial q^i} & \sum\limits_\alpha m_\alpha \left(\dfrac{\partial \vec{x}'_\alpha}{\partial q^i} \times \vec{x}'_\alpha \right)^T & \cdots \sum\limits_\alpha m_\alpha \dfrac{\partial \vec{x}'^T_\alpha}{\partial q^i} \dfrac{\partial \vec{x}'_\alpha}{\partial q^j} \cdots \\ \vdots & \vdots & \vdots \end{array} \right) , \tag{5.29}$$

where we have defined the *total mass* of the system $m_{tot} := \sum_\alpha m_\alpha$, the position of the *center of mass* of the system in the primed reference frame $\vec{R} := m_{tot}^{-1} \sum_\alpha m_\alpha \vec{x}'_\alpha$ and the *inertia tensor* of the system, also in the primed reference frame:

$$\mathcal{J} := \begin{pmatrix} \sum_\alpha m_\alpha ((x'^2_\alpha)^2 + (x'^3_\alpha)^2) & -\sum_\alpha m_\alpha x'^1_\alpha x'^2_\alpha & -\sum_\alpha m_\alpha x'^1_\alpha x'^3_\alpha \\ -\sum_\alpha m_\alpha x'^1_\alpha x'^2_\alpha & \sum_\alpha m_\alpha ((x'^1_\alpha)^2 + (x'^3_\alpha)^2) & -\sum_\alpha m_\alpha x'^2_\alpha x'^3_\alpha \\ -\sum_\alpha m_\alpha x'^1_\alpha x'^3_\alpha & -\sum_\alpha m_\alpha x'^2_\alpha x'^3_\alpha & \sum_\alpha m_\alpha ((x'^1_\alpha)^2 + (x'^2_\alpha)^2) \end{pmatrix} . \tag{5.30}$$

5.4 Unconstrained case

If no constraints are assumed and the system lives in the whole internal space $\mathcal{I}$ plus the external subspace spanned by the q^A, the Euclidean coordinates of the n atoms must be expressed using eq. (5.7), instead of eq. (5.9).

We now wish to calculate the determinant of the *whole-space mass-metric tensor* in the coordinates q^μ:

$$G_{\nu\rho}(q^\mu) := \sum_{\sigma=1}^{N} \frac{\partial x^\sigma(q^\mu)}{\partial q^\nu} m_\sigma \frac{\partial x^\sigma(q^\mu)}{\partial q^\rho} , \tag{5.31}$$

which, in matrix form, reads

$$G = J^T M J . \tag{5.32}$$

The only difference with eq. (5.11) is that, instead of the rectangular matrix J_c (see eq. (5.13)), in the above expression the full *Jacobian matrix* of the change of coordinates from Cartesian to curvilinear coordinates appears:

$$J^{\sigma}_{\ \rho}(q^{\mu}) := \frac{\partial x^{\sigma}(q^{\mu})}{\partial q^{\rho}} \ . \tag{5.33}$$

Obviously, one can deduce the factorization of $\det G$ as a particular case of the results of sec. 5.3 with $L = 0$, so that the indices i, j now run over all internal coordinates q^a. Explicitly,

$$\det G(q^A, q^a) = \sin^2\theta \, \det G_2(q^a) \ , \tag{5.34}$$

with

$$G_2 := \left(\begin{array}{cc|c} m_{tot} I^{(3)} & m_{tot}\, v(\vec{R}) & \cdots \quad m_{tot}\dfrac{\partial \vec{R}}{\partial q^b} \quad \cdots \\ m_{tot}\, v^T(\vec{R}) & \mathcal{J} & \cdots \ \sum\limits_{\alpha} m_\alpha \dfrac{\partial \vec{x}'_\alpha}{\partial q^b} \times \vec{x}'_\alpha \ \cdots \\ \hline \vdots & \vdots & \vdots \\ m_{tot}\dfrac{\partial \vec{R}}{\partial q^a} & \sum\limits_{\alpha} m_\alpha \left(\dfrac{\partial \vec{x}'_\alpha}{\partial q^a} \times \vec{x}'_\alpha\right)^T & \cdots \ \sum\limits_{\alpha} m_\alpha \dfrac{\partial \vec{x}'^{\,T}_\alpha}{\partial q^a}\dfrac{\partial \vec{x}'_\alpha}{\partial q^b} \cdots \\ \vdots & \vdots & \vdots \end{array}\right) . \tag{5.35}$$

However, in this section we would like to benefit from the special structure of eq. (5.32), where, differently from the constrained case, only $N \times N$ matrices occur, and find an expression simpler than eq. (5.34).

If we take determinants on both sides of eq. (5.32), we obtain

$$\det G = \left(\prod_{\alpha=1}^{n} m_\alpha^3\right) \det^2 J \ , \tag{5.36}$$

where, similarly to eq. (5.13), J may be written as follows:

$$J = \left(\begin{array}{ccccc|c} I^{(3)} & & 0 & & & 0 \\ \hline I^{(3)} & \dfrac{\partial E}{\partial \phi}\vec{x}'_2 & \dfrac{\partial E}{\partial \theta}\vec{x}'_2 & \dfrac{\partial E}{\partial \psi}\vec{x}'_2 & & \cdots \ E\dfrac{\partial \vec{x}'_2}{\partial q^b} \cdots \\ I^{(3)} & \dfrac{\partial E}{\partial \phi}\vec{x}'_3 & \dfrac{\partial E}{\partial \theta}\vec{x}'_3 & \dfrac{\partial E}{\partial \psi}\vec{x}'_3 & & \cdots \ E\dfrac{\partial \vec{x}'_3}{\partial q^b} \cdots \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ I^{(3)} & \dfrac{\partial E}{\partial \phi}\vec{x}'_\alpha & \dfrac{\partial E}{\partial \theta}\vec{x}'_\alpha & \dfrac{\partial E}{\partial \psi}\vec{x}'_\alpha & & \cdots \ E\dfrac{\partial \vec{x}'_\alpha}{\partial q^b} \cdots \\ \vdots & \vdots & \vdots & \vdots & & \vdots \end{array}\right) . \tag{5.37}$$

Now, the following identity is useful:

$$J = \begin{pmatrix} I^{(3)} & & & 0 \\ & E & & \\ & & \ddots & \\ 0 & & & E \end{pmatrix} J_1 \,, \tag{5.38}$$

where we have defined

$$J_1 := \left(\begin{array}{cccc|c} I^{(3)} & & 0 & & 0 \\ \hline I^{(3)} & E^T \frac{\partial E}{\partial \phi}\vec{x}_2' & E^T \frac{\partial E}{\partial \theta}\vec{x}_2' & E^T \frac{\partial E}{\partial \psi}\vec{x}_2' & \cdots \frac{\partial \vec{x}_2'}{\partial q^b} \cdots \\ I^{(3)} & E^T \frac{\partial E}{\partial \phi}\vec{x}_3' & E^T \frac{\partial E}{\partial \theta}\vec{x}_3' & E^T \frac{\partial E}{\partial \psi}\vec{x}_3' & \cdots \frac{\partial \vec{x}_3'}{\partial q^b} \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ I^{(3)} & E^T \frac{\partial E}{\partial \phi}\vec{x}_\alpha' & E^T \frac{\partial E}{\partial \theta}\vec{x}_\alpha' & E^T \frac{\partial E}{\partial \psi}\vec{x}_\alpha' & \cdots \frac{\partial \vec{x}_\alpha'}{\partial q^b} \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{array}\right) , \tag{5.39}$$

and we have that only the determinant of J_1 needs to be computed, since $\det J = \det^{n-1} E \, \det J_1 = \det J_1$.

Next, we note that, according to the definition of the primed reference frame in sec. 5.2, some of the components of the vectors $\vec{x}_2'$ and $\vec{x}_3'$ are zero, namely, we have that

$$\vec{x}_2' = \begin{pmatrix} 0 \\ 0 \\ x_2'^3 \end{pmatrix} \quad \text{and} \quad \vec{x}_3' = \begin{pmatrix} x_3'^1 \\ 0 \\ x_3'^3 \end{pmatrix} . \tag{5.40}$$

Hence, the derivatives with respect to q^b of the zero components are also zero, rendering three zero rows in the bottom right block of eq. (5.39). Performing two row permutations so that the zero rows are the top-most ones, we obtain a matrix J_2 whose determinant is the same as the one of J_1:

$$J_2 = \begin{pmatrix} I^{(3)} & & 0 \\ & J_2^{\mathcal{E}} & \\ \mathcal{X} & & J_2^{\mathcal{I}} \end{pmatrix} , \tag{5.41}$$

where the blocks in the diagonal have been defined as

$$J_2^{\mathcal{E}} = \begin{pmatrix} \left(E^T \frac{\partial E}{\partial \phi}\vec{x}_2'\right)^1 & \left(E^T \frac{\partial E}{\partial \theta}\vec{x}_2'\right)^1 & \left(E^T \frac{\partial E}{\partial \psi}\vec{x}_2'\right)^1 \\ \left(E^T \frac{\partial E}{\partial \phi}\vec{x}_2'\right)^2 & \left(E^T \frac{\partial E}{\partial \theta}\vec{x}_2'\right)^2 & \left(E^T \frac{\partial E}{\partial \psi}\vec{x}_2'\right)^2 \\ \left(E^T \frac{\partial E}{\partial \phi}\vec{x}_3'\right)^2 & \left(E^T \frac{\partial E}{\partial \theta}\vec{x}_3'\right)^2 & \left(E^T \frac{\partial E}{\partial \psi}\vec{x}_3'\right)^2 \end{pmatrix} , \tag{5.42}$$

the superindices standing for vector components, and

$$J_2^{\mathcal{I}} = \begin{pmatrix} \cdots & \dfrac{\partial x_2'^{\,3}}{\partial q^b} & \cdots \\ \cdots & \dfrac{\partial x_3'^{\,1}}{\partial q^b} & \cdots \\ \cdots & \dfrac{\partial x_3'^{\,3}}{\partial q^b} & \cdots \\ \cdots & \dfrac{\partial \vec{x}_4'}{\partial q^b} & \cdots \\ & \vdots & \\ \cdots & \dfrac{\partial \vec{x}_\alpha'}{\partial q^b} & \cdots \\ & \vdots & \end{pmatrix} . \tag{5.43}$$

The concrete form of the submatrix $\mathcal{X}$ in eq. (5.41) is irrelevant for our purposes, since

$$\det J = \det J_2 = \det J_2^{\mathcal{E}} \det J_2^{\mathcal{I}} . \tag{5.44}$$

Now, an explicit computation of $J_2^{\mathcal{E}}$ yields

$$J_2^{\mathcal{E}} = \begin{pmatrix} x_2'^{\,3} \sin\theta \sin\psi & -x_2'^{\,3} \cos\psi & 0 \\ x_2'^{\,3} \sin\theta \cos\psi & x_2'^{\,3} \sin\psi & 0 \\ -x_3'^{\,1} \cos\theta + x_3'^{\,3} \sin\theta \cos\psi & x_3'^{\,3} \sin\psi & x_3'^{\,1} \end{pmatrix} , \tag{5.45}$$

with determinant $\det J_2^{\mathcal{E}} = \sin\theta\, x_3'^{\,1} (x_2'^{\,3})^2$.
Finally, using eqs. (5.44) and (5.36), we obtain

$$\det G(q^A, q^a) = \sin^2\theta (x_3'^{\,1}(q^a))^2 (x_2'^{\,3}(q^a))^4 \det J_2^{\mathcal{I}}(q^a) \left(\prod_{\alpha=1}^{n} m_\alpha^3 \right) , \tag{5.46}$$

where the factorization has been achieved, since the only factor that depends on the external coordinates is $\sin^2\theta$.

5.5 Determinant of G in SASMIC coordinates

The SASMIC scheme, introduced in chapter 4, is a set of rules to define particular Z-matrix coordinates [315, 316] of general branched molecules, with convenient properties of modularity and approximate separability of soft and hard modes.

According to the rules, to each atom α, one uniquely assigns three atoms $\beta(\alpha)$, $\gamma(\alpha)$ and $\delta(\alpha)$ in such a way that the three Z-matrix internal coordinates that position atom α are

$$\begin{aligned} r_\alpha &:= (\alpha, \beta(\alpha)) \\ \theta_\alpha &:= (\alpha, \beta(\alpha), \gamma(\alpha)) \\ \phi_\alpha &:= (\alpha, \beta(\alpha), \gamma(\alpha), \delta(\alpha)) , \end{aligned} \tag{5.47}$$

being r_α a bond length, θ_α a bond angle and ϕ_α a dihedral angle.

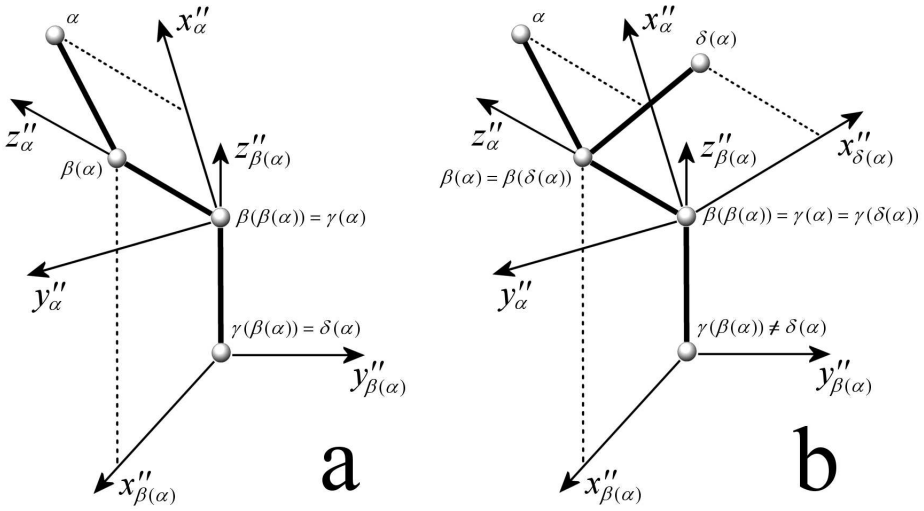


Figure 5.2: Local reference frames associated to atoms α and $\beta(\alpha)$ (see text) in the cases that **(a)** ϕ_α is a principal dihedral or **(b)** ϕ_α is a phase dihedral.

The procedure that will be followed in order to express the position $\vec{x}'_\alpha$ of atom α in the primed reference frame in fig. 5.1 as a function of the SASMIC internal coordinates starts by expressing the vector that goes from $\beta(\alpha)$ to α in a set of axes $(x''_\alpha, y''_\alpha, z''_\alpha)$ associated to α. This local reference frame is defined such that the z''_α-axis lies along the directional bond $(\gamma(\alpha), \beta(\alpha))$ and the x''_α-axis lies along the projection of $(\beta(\alpha), \alpha)$ onto the plane orthogonal to $(\gamma(\alpha), \beta(\alpha))$ (see fig. 5.2).

In these axes, the components of the vector $(\beta(\alpha), \alpha)$ are

$$\vec{x}''^{\,T}_\alpha := (r_\alpha \sin\theta_\alpha, 0, -r_\alpha \cos\theta_\alpha) \,. \tag{5.48}$$

Now, if the atom $\delta(\alpha)$ that is used to define ϕ_α is bonded to atom $\gamma(\alpha)$ (fig. 5.2a), ϕ_α is called a *principal dihedral* and we have that

$$\underbrace{\begin{pmatrix} -\cos\theta_{\beta(\alpha)} & 0 & \sin\theta_{\beta(\alpha)} \\ 0 & 1 & 0 \\ -\sin\theta_{\beta(\alpha)} & 0 & -\cos\theta_{\beta(\alpha)} \end{pmatrix}}_{\Theta(\theta_{\beta(\alpha)})} \underbrace{\begin{pmatrix} \cos\phi_\alpha & -\sin\phi_\alpha & 0 \\ \sin\phi_\alpha & \cos\phi_\alpha & 0 \\ 0 & 0 & 1 \end{pmatrix}}_{\Phi(\phi_\alpha)} \vec{x}''_\alpha \tag{5.49}$$

are the components of the vector $(\beta(\alpha), \alpha)$ in the local reference frame $(x''_{\beta(\alpha)}, y''_{\beta(\alpha)}, z''_{\beta(\alpha)})$ associated to atom $\beta(\alpha)$.

On the other hand, if we are at a branching point and the atom $\delta(\alpha)$ that is used to define ϕ_α is bonded to atom $\beta(\alpha)$ (fig. 5.2b), ϕ_α is called a *phase dihedral* and we have to change first to the local reference frame associated to $\delta(\alpha)$. In this case, the components of the vector $(\beta(\alpha), \alpha)$ in the local reference frame $(x''_{\beta(\alpha)}, y''_{\beta(\alpha)}, z''_{\beta(\alpha)})$ are

$$\Theta(\theta_{\beta(\alpha)})\Phi(\phi_{\delta(\alpha)})\Phi(\phi_\alpha)\, \vec{x}''_\alpha \,. \tag{5.50}$$

If we now iterate the procedure, by changing the axes to the ones associated to the atom $\beta(\beta(\alpha))$, i.e., the β atom that corresponds to $\beta(\alpha)$ according to the SASMIC scheme, an so on, we will eventually arrive to the set of axis (x_3'', y_3'', z_3'') (since, in the SASMIC scheme (see chapter 4), we have that $\beta(\alpha) < \alpha$). Note however that, according to the definition of the local reference frame given in the preceding paragraphs, the one associated to atom 3 is *exactly* the primed reference frame in fig. 5.1.

Hence, let us define, for each atom α, a matrix R_α as the product of the matrices obtained using eqs. (5.49) and (5.50) and successively applying the function $\beta(\alpha)$. Then, R_α takes the vector $(\beta(\alpha), \alpha)$ in eq. (5.48) to the primed reference frame.

Let the superindex on β denote composition of functions, let us define $\beta^0(\alpha) := \alpha$ and $\mathcal{N}_\alpha$ as the integer such that $\beta^{\mathcal{N}_\alpha+1}(\alpha) = 3$. Adding all the vectors corresponding to $(\beta^{p+1}(\alpha), \beta^p(\alpha))$ in the primed reference frame, with $p = 0, \ldots, \mathcal{N}_\alpha$, to $\vec{x}_3'$ yields the position of atom α in the primed reference frame as a function of the internal coordinates:

$$\vec{x}_\alpha' = \vec{x}_3' + \sum_{p=\mathcal{N}_\alpha}^{0} R_{\beta^p(\alpha)} \vec{x}_{\beta^p(\alpha)}'' \,. \tag{5.51}$$

Now, ordering the internal coordinates as $(r_2, r_3, \theta_3, r_4, \theta_4, \phi_4, \ldots, r_n, \theta_n, \phi_n)$ and using the already mentioned fact that $\beta(\alpha) < \alpha$, and also that $\delta(\alpha) < \alpha$, we have that the matrix $J_2^{\mathcal{I}}$ in eq. (5.43) is

$$J_2^{\mathcal{I}} = \begin{pmatrix} A_0 & & & 0 \\ & A_4 & & \\ & & \ddots & \\ \mathcal{X} & & & A_n \end{pmatrix} \quad \text{and} \quad \det J_2^{\mathcal{I}} = \det A_0 \prod_{\alpha=4}^{n} \det A_\alpha \,. \tag{5.52}$$

Using that

$$\vec{x}_2' = \begin{pmatrix} 0 \\ 0 \\ r_2 \end{pmatrix} \quad \text{and} \quad \vec{x}_3' = \begin{pmatrix} r_3 \sin\theta_3 \\ 0 \\ r_2 - r_3\cos\theta_3 \end{pmatrix}, \tag{5.53}$$

we have

$$A_0 := \begin{pmatrix} \frac{\partial x_2'^3}{\partial r_2} & \frac{\partial x_2'^3}{\partial r_3} & \frac{\partial x_2'^3}{\partial \theta_3} \\ \frac{\partial x_3'^1}{\partial r_2} & \frac{\partial x_3'^1}{\partial r_3} & \frac{\partial x_3'^1}{\partial \theta_3} \\ \frac{\partial x_3'^3}{\partial r_2} & \frac{\partial x_3'^3}{\partial r_3} & \frac{\partial x_3'^3}{\partial \theta_3} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \sin\theta_3 & r_3\cos\theta_3 \\ 1 & -\cos\theta_3 & r_3\sin\theta_3 \end{pmatrix}. \tag{5.54}$$

Now, we note that the matrix $\Phi(\phi_\alpha)$ occurs always at the right-most place in R_α and that the derivatives in the blocks A_α, with $\alpha > 4$, kill all the terms in eq. (5.51) except for the one corresponding to $p = 0$. Hence, if we define $R_\alpha =: R'_\alpha \Phi(\phi_\alpha)$, the block A_α may be

expressed as follows:

$$A_\alpha := R'_\alpha \left(\frac{\partial \Phi(\phi_\alpha)\vec{x}''_\alpha}{\partial r_\alpha} \quad \frac{\partial \Phi(\phi_\alpha)\vec{x}''_\alpha}{\partial \theta_\alpha} \quad \frac{\partial \Phi(\phi_\alpha)\vec{x}''_\alpha}{\partial \phi_\alpha} \right) = \begin{pmatrix} \sin\theta_\alpha \cos\phi_\alpha & r_\alpha \cos\theta_\alpha \cos\phi_\alpha & -r_\alpha \sin\theta_\alpha \sin\phi_\alpha \\ \sin\theta_\alpha \sin\phi_\alpha & r_\alpha \cos\theta_\alpha \sin\phi_\alpha & r_\alpha \sin\theta_\alpha \cos\phi_\alpha \\ -\cos\theta_\alpha & r_\alpha \sin\theta_\alpha & 0 \end{pmatrix} . \tag{5.55}$$

Finally, using eq. (5.52), noting that $\det A_0 = r_3$ and $\det A_\alpha = -r_\alpha^2 \sin\theta_\alpha$, and calculating the remaining terms of eq. (5.46) with eq. (5.53), we obtain the desired result:

$$\det{}^{\frac{1}{2}} G(q^A, q^a) = \left(\prod_{\alpha=1}^{n} m_\alpha^{3/2} \right) |\sin\theta| \left(\prod_{\alpha=2}^{n} r_\alpha^2 \right) \left(\prod_{\alpha=3}^{n} |\sin\theta_\alpha| \right) . \tag{5.56}$$

It is worth remarking at this point that the previous expression does not explicitly depend on the dihedral angles ϕ_α and that it is the same result as the one found in ref. 313 for serial polymers.

5.6 Conclusions

In this chapter, we have calculated explicit expressions in which the determinant of the mass-metric tensor G (eqs. (5.46) and (5.34)) and the determinant of the reduced mass-metric tensor g (eq. (5.28)), occurring in classical statistical mechanics in the coordinate space, are written as a product of two functions; one depending only on the external coordinates that describe the overall translation and rotation of the system, and the other only on the internal coordinates. This has been done for any molecule, general internal coordinates and arbitrary constraints, extending the work in refs. 313 and 342.

This factorization allows to integrate out the external coordinates and perform Monte Carlo simulations in the internal conformational space, gaining insight of the problem, simplicity in the description of the system and, for small molecules, some computational effort. Also, our results show that, in general, the Fixman's compensating potential [338–340], customarily used to reproduce the stiff equilibrium distribution using rigid molecular dynamics simulations, does not depend on the external variables.

In appendix D, we give a general mathematical argument showing that the factorization is a consequence of the symmetries of the metric tensors involved and, in sec. 5.5, the determinant of the mass-metric tensor G is computed explicitly in the SASMIC set of curvilinear coordinates for general branched molecules introduced in the previous chapter (see eq. (5.56)) showing that the classical formula for serial polymers [313] is actually valid for any macromolecule.

Finally, we would like to note that all the expressions derived in the present chapter have been directly applied to a real case in chapter 6.

Chapter 6

Study of the effects of stiff and rigid constraints in the conformational equilibrium of the alanine dipeptide

This chapter is based on the article:
Pablo Echenique, Iván Calvo and J. L. Alonso, *Quantum mechanical calculation of the effects of stiff and rigid constraints in the conformational equilibrium of the alanine dipeptide*, J. Comp. Chem. **27** (2006) 1733–1747.

> The most misleading assumptions are the ones you don't even know you're making. [344]
>
> — Douglas N. Adams, 1990

6.1 Introduction

As we have already mentioned, in computer simulations of large complex systems, such as macromolecules and, specially, proteins [2, 76, 126, 137, 311, 345], one of the main bottlenecks to design efficient algorithms is the necessity to sample an astronomically large conformational space [137, 138]. In addition, being the typical timescales of the different movements in a wide range, demandingly small timesteps must be used in molecular dynamics simulations in order to properly account for the fastest modes, which lie in the femtosecond range. However, most of the biological interesting behaviour (allosteric transitions, protein folding, enzymatic catalysis) is related to the slowest conformational changes, which occur in the timescale of milliseconds or even seconds [134, 345–348]. Fortunately, the fastest modes are also the most energetic ones and are rarely activated at room temperature. Therefore, in order to alleviate the computational problems and also simplify the images used to think about these elusive systems, one may naturally consider the reduction of the number of degrees of freedom describing macromolecules via the imposition of constraints [349].

How to study the conformational equilibrium of these constrained systems has been an object of much debate [131, 313, 340, 350, 351]. Two different classical models exist in

the literature which are conceptually [131, 313, 338–340, 351] and practically [126, 132, 133, 342, 351–353] inequivalent. In the *classical rigid* model, the constraints are assumed to be *exact* and all the *velocities* (not the momenta) that are orthogonal to the hypersurface defined by them vanish. In the *classical stiff* model, on the other hand, the constraints are assumed to be *approximate* and they are implemented by a steep potential that drives the system to the constrained hypersurface. In this case, the orthogonal velocities are activated and may act as 'heat containers'.

In this dissertation, we do not address the question of which model is a better approximation of physical reality. Although, in the literature, it is commonly assumed (often implicitly) that the classical stiff model should be taken as a reference [126, 132, 134, 338, 340, 342, 354], we believe that this opinion is much influenced by the use of popular classical force fields [104, 105, 117–126] (which are stiff by construction) and by the goal of reproducing their results at a lower computational cost, i.e., using rigid molecular dynamics simulations and adding the appropriate correcting terms to the potential energy in the constrained subspace [131–134, 311, 338, 345, 346, 352, 354–360]. In our opinion, the question whether the rigid or the stiff model should be used to approximate the real quantum mechanical statistics of an arbitrary organic molecule has not been satisfactorily answered yet (for discussions about the topic, see refs. 131, 313, 339, 350, 351, 361–363). Therefore, in this chapter, we adopt the cautious position that any of the two models may be useful in certain cases or for certain purposes and we study them both on equal footing. Our concern is, then, to investigate the effects that either way of imposing constraints causes in the conformational equilibrium of macromolecules.

In this type of systems, the natural constraints are those derived from the relative rigidity of the internal covalent structure of groups of atoms that share a common center (and also from the rigidity of rotation around double or triple bonds) compared to the energetically 'cheaper' rotation around single bonds.In internal coordinates, these chemical constraints may be directly implemented by asking that some conveniently selected *hard* coordinates (normally, bond lengths, bond angles and some dihedrals) have constant values or values that depend on the remaining *soft* coordinates (see ref. 313 for a definition). In Euclidean coordinates, on the other hand, the expression of the constraints is more cumbersome and complicated procedures [354, 356, 360, 364–366] must be used at each timestep to implement them in molecular dynamics simulations. This is why, in the classical stiff model, as well as in the rigid one, it is common to use internal coordinates and they are also the choice throughout this chapter.

As it has been already advanced, in the equilibrium statistical mechanics of both the stiff and rigid models, the marginal probability density in the coordinate part of the phase space in these internal coordinates is not proportional to the naive $\exp[-\beta V_\Sigma(q^i)]$, where $V_\Sigma(q^i)$ denotes the potential energy on the constrained hypersurface depending on the soft internal coordinates q^i. Instead, some correcting terms that come from different sources must be added to the potential energy $V_\Sigma(q^i)$ [313, 338, 339, 351, 357, 367, 368]. These terms involve determinants of two different mass-metric tensors and also of the Hessian matrix of the constraining part of the potential (see sec. 6.2). If Monte Carlo simulations in the coordinate space are to be performed [311, 321, 369–372] and the probability densities that correspond to any of these two models sampled, the corrections should be included or, otherwise, showed to be negligible.

Additionally, the three different correcting terms are involved in the definition of the

so-called *Fixman's compensating potential* [340], which is frequently used to reproduce the stiff equilibrium distribution using rigid molecular dynamics simulations [131–134, 338, 352, 357–359, 367].

Customarily in the literature, some of these corrections to the potential energy are assumed to be independent of the conformation and thus dropped from the basic expressions. Also, subtly entangled to the assumptions underlying many classical results, a second type of approximation is made that consists of assuming that the equilibrium values of the hard coordinates do not depend on the soft coordinates.

In this chapter, using all the analytical and computational techniques developed in the previous ones, we measure the conformational dependence of all correcting terms and of the Fixman's compensating potential in the model dipeptide HCO-L-Ala-NH_2 (see appendix E) without any simplifying assumption. The potential energy function is considered to be the effective Born-Oppenheimer potential for the nuclei derived from ab initio quantum mechanical calculations including electron correlation at the MP2 level. We also repeat the calculations, with the same basis set (6-31++G(d,p)) and at the Hartree-Fock level of the theory in order to investigate if this less demanding method without electron correlation may be used in further studies. It is the first time, as far as we are aware, that this type of study is performed in a relevant biomolecule with a realistic potential energy function.

In sec. 6.2 of this chapter, we derive the statistical mechanics formulae of the rigid and stiff models in the general case. In sec. 6.4, we describe the computational methods used and we make a brief reminder of the factorization of the external coordinates presented in chapter 5. In sec. 6.3, we discuss the use of the different approximations in the literature and we give a precise definition of *exactly* and *approximately separable hard and soft coordinates* which will shed some light on the relation between the different types of simplifications aforementioned. Sec. 6.5 is devoted to the presentation and discussion of the assessment of the approximation that consists of neglecting the different corrections to the potential energy in the model dipeptide HCO-L-Ala-NH_2 (see appendix E), which is the central aim of this chapter. Finally, the conclusions are summarized in sec. 6.6.

6.2 Theory

First of all, let us note that the notational conventions that will be used in the following sections are those introduced in sec. 5.2, and the form of the constraints is that of eq. (5.8) (see also table 5.1 for quick reference about the different indices used).

The *general set-up* of the problem may be described as follows: Instead of us being interested on the conformational equilibrium of the system in the external subspace $\mathcal{E}$ plus the whole internal subspace $\mathcal{I}$ (i.e., the whole space, denoted by $\Omega = \mathcal{E} \times \mathcal{I}$), we wish to find the probability density on a constrained hypersurface $\Sigma \subset \mathcal{I}$ of dimension M (plus the external subspace $\mathcal{E}$), i.e., on $\mathcal{E} \times \Sigma$.

6.2.1 Classical stiff model

In the classical stiff model, the constraints in eq. (5.8) are implemented by imposing an strong energy penalization when the internal conformation of the system, described by q^a,

departs from the constrained hypersurface Σ. To ensure this, we must have that the potential energy function in $\mathcal{I}$ satisfies certain conditions. First, we write the potential $V(q^a)$ as follows[105]:

$$V(q^i, q^I) = \underbrace{V(q^i, f^I(q^i))}_{V_\Sigma(q^i)} + \underbrace{\left[V(q^i, q^I) - V(q^i, f^I(q^i))\right]}_{V_c(q^i, q^I)} . \tag{6.1}$$

Next, we impose the following conditions on the *constraining potential* $V_c(q^i, q^I)$ defined above:

(i) That $V_c(q^i, f^I(q^i)) \leq V_c(q^i, q^I) \quad \forall q^i, q^I$, i.e., that Σ be the global minimum of V_c (and, henceforth, a local one too) with respect to variations of the hard coordinates.

(ii) That, for small variations Δq^I on the hard coordinates (i.e., for changes Δq^I considered as physically irrelevant), the associated changes in $V_c(q^i, q^I)$ are much larger than the thermal energy RT.

The advantages of this formulation, much similar to that on ref. 313, are many. First, it sets a convenient framework for the derivation of the statistical mechanics formulae of the classical stiff model relating it to the flexible[106] model in the whole space $\mathcal{E} \times \mathcal{I}$. Second, it clearly separates the potential energy on Σ from the part that is responsible of implementing the constraints. Third, contrarily to the formulation based on delta functions [367], it allows to clearly understand the necessity of including the correcting term associated to the determinant of the Hessian of V_c (see the derivation that follows). Finally, and more importantly for us, it provides a direct prescription for calculating $V_\Sigma(q^i)$ and Σ (the Potential Energy Surface (PES), frequently used in quantum chemistry calculations [193, 207, 210–213]) via geometry optimization at fixed values of the soft coordinates.

We also remark that, in order to satisfy point (ii) above and to allow the derivation of the different correcting terms that follows and the validity of the final expressions, the hard coordinates q^I must be indeed hard, however, the *soft coordinates* q^i do not have to be soft (in the sense that they produce energetic changes much smaller than RT when varied). They may be interesting for some other reason and hence voluntarily picked to describe the system studied, without altering the formulae presented in this section. Despite these qualifications, the terms *soft* and *hard* shall be used in this dissertation for consistence with most of the existing literature [313, 339, 368, 373, 374], although, in some cases, the labels *important* and *unimportant* (for q^i and q^I respectively), proposed by Karplus and Kushick [314], may be more appropriate.

In the case of the model dipeptide HCO-L-Ala-NH_2 investigated in this chapter, for example (see also appendix E), the barriers in the Ramachandran angles ϕ and ψ may be as large as $\sim 40\ RT$, however, the study of small dipeptides is normally aimed to the design of effective potentials for polypeptides [159, 163, 164], where long-range interactions in the sequence may compensate these local energy penalizations. This and the fact that the Ramachandran angles are the relevant degrees of freedom to describe the conformation

[105] Note that we have simply added and subtracted from the total potential energy $V(q^i, q^I) \equiv V(q^a)$ of the system the same quantity, $V(q^i, f^I(q^i))$.

[106] Some authors use the word *flexible* to refer to this model [132, 133, 313, 360]. We, however, prefer to term it *stiff* [339] and keep the name *flexible* to refer to the case in which no constraints are imposed.

of the backbone of these systems, make it convenient to choose them as *soft coordinates* q^i despite the fact that they may be energetically hard in the case of the isolated dipeptide HCO-L-Ala-NH_2. As remarked above, this does not affect the calculations.

Now, due to condition (ii) above, the statistical weights of the conformations which lie far away from the constrained hypersurface Σ are negligible and, therefore, it suffices to describe the system in the vicinity of the equilibrium values of the q^I. In this region, for each value of the internal soft coordinates q^i, we may expand $V_c(q^i, q^I)$ in eq. (6.1) up to second order in the hard coordinates around Σ (i.e., around $q^I = f^I(q^i)$) and drop the higher order terms[107]:

$$\begin{aligned} V_c(q^i, q^I) \;\simeq\; & V_c(q^i, f^I(q^i)) + \left[\frac{\partial V_c}{\partial q^J}\right]_\Sigma (q^J - f^J(q^i)) \\ & + \frac{1}{2}\underbrace{\left[\frac{\partial^2 V_c}{\partial q^J \partial q^K}\right]_\Sigma}_{\mathcal{H}_{JK}(q^i)} (q^J - f^J(q^i))(q^K - f^K(q^i)) \,, \end{aligned} \tag{6.2}$$

where the subindex Σ indicates evaluation on the constrained hypersurface and a more compact notation, $\mathcal{H}(q^i)$, has been introduced for the Hessian matrix of V_c with respect to the hard variables evaluated on Σ.

In this expression, the zeroth order term $V_c(q^i, f^I(q^i))$ is zero by definition of V_c (see eq. (6.1)) and the linear term is also zero, because of the condition (i) above. Hence, the first non-zero term of the expansion in eq. (6.2) is the second order one. Using this, together with eq. (6.1), we may write the *stiff Hamiltonian*

$$\begin{aligned} H_s(q^\mu, \pi_\mu) \;:=\; & \frac{1}{2}\pi_\nu G^{\nu\rho}(q^u, q^I)\pi_\rho + V_\Sigma(q^i) \\ & + \frac{1}{2}\mathcal{H}_{JK}(q^i)(q^J - f^J(q^i))(q^K - f^K(q^i)) \,, \end{aligned} \tag{6.3}$$

with π_μ denoting the momenta canonically conjugate to the q^μ, and the *mass-metric tensor* $G_{\nu\rho}$ being

$$G_{\nu\rho}(q^u, q^I) := \sum_{\sigma=1}^{N} \frac{\partial x^\sigma(q^\mu)}{\partial q^\nu} m_\sigma \frac{\partial x^\sigma(q^\mu)}{\partial q^\rho} \tag{6.4}$$

and $G^{\nu\rho}$ its inverse, defined by

$$G^{\nu\sigma}(q^u, q^I)\, G_{\sigma\rho}(q^u, q^I) = \delta^\nu_\rho \,, \tag{6.5}$$

where δ^ν_ρ denotes the Kronecker's delta.

Therefore, a first version of the partition function of the system is[108]

$$Z = \frac{\alpha_{QM}}{h^N} \int \exp\left[-\beta H_s(q^\mu, \pi_\mu)\right] \mathrm{d}q^\mu \mathrm{d}\pi_\mu \,, \tag{6.6}$$

[107] Einstein's sum convention is assumed on repeated indices throughout the whole document.

[108] First note that no Jacobian appears in the integral measure because q^μ and p_μ are obtained from the Euclidean coordinates via a canonical transformation [127]. Second, contrarily to the more qualitative discussion in sec. 1.4, here we will carry the T-dependent factors in order to render the derivation completely rigorous.

where α_{QM} is a combinatorial number that accounts for quantum indistinguishability and that must be specified in each particular case (e.g., for a gas of N indistinguishable particles or for the water molecules in sec. 1.4, $\alpha_{QM} = 1/N!$).

Now, using the condition (ii) again, the q^I appearing in the mass-metric tensor G in H_s (in eq. (6.6)) can be approximately evaluated at their equilibrium values $f^I(q^i)$, yielding the *stiff partition function*:

$$\begin{aligned} Z_s \;:=\; & \frac{\alpha_{QM}}{h^N}\int \exp\Big[-\beta\Big(\frac{1}{2}\pi_\nu G^{\nu\rho}(q^u, f^I(q^i))\pi_\rho + V_\Sigma(q^i) \\ & + \frac{1}{2}\mathcal{H}_{JK}(q^i)(q^J - f^J(q^i))(q^K - f^K(q^i))\Big)\Big]\,\mathrm{d}q^u\mathrm{d}q^I\mathrm{d}\pi_\mu\,. \end{aligned} \tag{6.7}$$

If we now integrate over the hard coordinates q^I, we have

$$\begin{aligned} Z_s \;=\; & \left(\frac{2\pi}{\beta}\right)^{\frac{L}{2}}\frac{\alpha_{QM}}{h^N}\int \exp\Big[-\beta\Big(\frac{1}{2}\pi_\nu G^{\nu\rho}(q^u, f^I(q^i))\pi_\rho \\ & + \; V_\Sigma(q^i) + T\frac{R}{2}\ln\Big[\det\mathcal{H}(q^i)\Big]\Big)\Big]\,\mathrm{d}q^u\mathrm{d}\pi_\mu\,. \end{aligned} \tag{6.8}$$

where the part of the result of the Gaussian integral consisting of $\det^{-1/2}\mathcal{H}$ has been taken to the exponent.

Note that the Hessian matrix $\mathcal{H}_{JK}$ involves only derivatives with respect to the hard coordinates (see eq. (6.2)), so that the minimization protocol embodied in eq. (6.1) (which is identical to the procedure followed in quantum chemistry for computing the PES along reaction coordinates) guarantees that $\mathcal{H}_{JK}$ is positive defined and, hence, $\det\mathcal{H}$ is positive, allowing to take its logarithm like in the previous expression. The fact that it is only this 'partial Hessian' that makes sense in the computation of equilibrium properties along soft (or reaction) coordinates, has been recently pointed out in ref. 375.

It is also frequent to integrate over the momenta in the partition function (see appendix A for some details related to this issue). Doing this in eq. (6.8) and taking the determinant of the mass-metric tensor that shows up[109] to the exponent, we may write the partition function as an integral only on the coordinates:

$$\begin{aligned} Z_s = \chi_s(T)\int \exp\Big[-\beta\Big(V_\Sigma(q^i) + \; T\frac{R}{2}\ln\Big[\det\mathcal{H}(q^i)\Big] - \\ -T\frac{R}{2}\ln\Big[\det G(q^u, f^I(q^i))\Big]\Big)\Big]\,\mathrm{d}q^u\,, \end{aligned} \tag{6.9}$$

where the multiplicative factor that depends on T has been defined as follows:

$$\chi_s(T) := \left(\frac{2\pi}{\beta}\right)^{\frac{N+L}{2}}\frac{\alpha_{QM}}{h^N}\,. \tag{6.10}$$

[109] Note that, by G, we denote the matrix that corresponds to the mass-metric tensor with two covariant indices $G_{\mu\nu}$. The same convention has been followed for the Hessian matrix $\mathcal{H}$ in eq. (6.8) and for the reduced mass-metric tensor g in eq. (6.20).

If we exploit again the useful image used in previous chapters and the exponent in eq. (6.9) is seen as a free energy, then, $V_\Sigma(q^i)$ may be regarded as the *internal energy* and the two conformation-dependent correcting terms that are added to it as *effective entropies* (which is compatible with their being linear in RT). The second one comes only from the desire to write the marginal probabilities in the coordinate space (i.e., averaging the momenta, see appendix A) and may be called a *kinetic entropy* [350], the first term, on the other hand, is truly an entropic term that comes from the averaging out of certain degrees of freedom and it is reminiscent of the *conformational* or *configurational entropies* appearing in quasiharmonic analysis [126, 314, 376].

In this spirit, we define[110]

$$F_{\mathrm{s}}(q^u) := V_\Sigma(q^i) - T(S^c_{\mathrm{s}}(q^i) + S^{\mathrm{k}}_{\mathrm{s}}(q^u)) \,, \tag{6.11a}$$

$$S^c_{\mathrm{s}}(q^i) := -\frac{R}{2} \ln\left[\det \mathcal{H}(q^i)\right] , \tag{6.11b}$$

$$S^{\mathrm{k}}_{\mathrm{s}}(q^u) := \frac{R}{2} \ln\left[\det G(q^u, f^I(q^i))\right] . \tag{6.11c}$$

In such a way that the *stiff equilibrium probability* in the soft subspace $\mathcal{E} \times \Sigma$ is given by

$$p_{\mathrm{s}}(q^u) = \frac{\exp\left[-\beta F_{\mathrm{s}}(q^u)\right]}{Z'_{\mathrm{s}}} \,, \quad \text{with} \quad Z'_{\mathrm{s}} := \int \exp\left[-\beta F_{\mathrm{s}}(q^u)\right] \mathrm{d}q^u \,. \tag{6.12}$$

Now, it is worth remarking that, although the kinetic entropy $S^{\mathrm{k}}_{\mathrm{s}}$ depends on the external coordinates q^A, using the results in the previous chapter, the determinant of the mass-metric tensor G may be written as a product of two functions; one depending only on the external coordinates, and the other only on the internal ones q^a. Hence the externals-dependent factor in eq. (6.11c) may be integrated out independently to yield an effective free energy and a probability density p_{s} that depends only on the soft internals q^i (see sec. 6.4.1 and chapter 5).

6.2.2 Classical rigid model

If the relations in eq. (5.8) are considered to hold *exactly* and are treated as holonomic constraints, the Hamiltonian function that describes the classical mechanics in the subspace $(\mathcal{E} \times \Sigma) \subset (\mathcal{E} \times \mathcal{I})$, spanned by the coordinates q^u, may be written as follows:

$$H_{\mathrm{r}}(q^u, \eta_u) := \frac{1}{2} \eta_v g^{vw}(q^u) \eta_w + V_\Sigma(q^i) \,, \tag{6.13}$$

where the *reduced mass-metric tensor* $g_{vw}(q^u)$ in $\mathcal{E} \times \Sigma$, that appears in the kinetic energy, is

[110] In this chapter, contrarily to the convention followed in the previous ones and due to the presence of clearly identifiable entropic terms, we find more suggestive to denote the effective potential energy appearing in the exponents of the probability density by F, instead of W.

$$
\begin{aligned}
g_{vw}(q^u) \;=\; & G_{vw}(q^u, f^I(q^i)) + \frac{\partial f^J(q^i)}{\partial q^v} G_{JK}(q^u, f^I(q^i)) \frac{\partial f^K(q^i)}{\partial q^w} \\
& + \; G_{vK}(q^u, f^I(q^i)) \frac{\partial f^K(q^i)}{\partial q^w} + \frac{\partial f^J(q^i)}{\partial q^v} G_{Jw}(q^u, f^I(q^i)) := \\
& \frac{\partial \tilde{f}^\mu}{\partial q^v} G_{\mu\nu}(q^u, f^I(q^i)) \frac{\partial \tilde{f}^\nu}{\partial q^w} \,,
\end{aligned}
\tag{6.14}
$$

and $g^{vw}(q^u)$ is defined to be its inverse in the sense of eq. (6.5). Also, the notation

$$
\tilde{f}^\mu := \begin{cases} q^u & \text{if} \quad u := \mu = 1, \ldots, M+6 \\ f^I(q^i) & \text{if} \quad I := \mu = M+7, \ldots, N \end{cases}
\tag{6.15}
$$

has been introduced for convenience.

Note that eq. (6.14) may derived from the unconstrained Hamiltonian in $(\mathcal{E} \times \mathcal{I})$,

$$
H(q^\mu, \pi_\mu) := \frac{1}{2} \pi_\nu G^{\nu\rho}(q^\mu) \pi_\rho + V(q^a) \,,
\tag{6.16}
$$

using the constraints in eq. (5.8), together with its time derivatives (denoted by an overdot: like in $\dot{A}$)

$$
\dot{q}^I := \frac{\partial f^I(q^i)}{\partial q^j} \dot{q}^j
\tag{6.17}
$$

and defining the momenta η_v as

$$
\eta_v := g_{vw}(q^u)\, \dot{q}^w = g_{vw}(q^u)\, G^{w\mu}(q^u, f^I(q^i))\, \pi_\mu \,.
\tag{6.18}
$$

Hence, the *rigid partition function* is

$$
Z_{\mathrm{r}} = \frac{\alpha_{QM}}{h^{M+6}} \int \exp\left[-\beta \left(\frac{1}{2} \eta_v g^{vw}(q^u) \eta_v + V_\Sigma(q^i) \right) \right] \mathrm{d}q^u \mathrm{d}\eta_u \,.
\tag{6.19}
$$

Integrating over the momenta, we obtain the marginal probability density in the coordinate space analogous to eq. (6.9):

$$
Z_{\mathrm{r}} = \chi_{\mathrm{r}}(T) \int \exp\Big[-\beta \Big(V_\Sigma(q^i) - T \frac{R}{2} \ln \Big[\det g(q^u) \Big] \Big) \Big] \mathrm{d}q^u \,,
\tag{6.20}
$$

where

$$
\chi_{\mathrm{r}}(T) := \left(\frac{2\pi}{\beta} \right)^{\frac{M+6}{2}} \frac{\alpha_{QM}}{h^{\frac{M+6}{2}}} \,.
\tag{6.21}
$$

Repeating the analogy with free energies and entropies in the last paragraphs of the previous section, we define

$$
F_{\mathrm{r}}(q^u) := V_\Sigma(q^i) - T S^{\mathrm{k}}_{\mathrm{r}}(q^u) \,,
\tag{6.22a}
$$

$$
S^{\mathrm{k}}_{\mathrm{r}}(q^u) := \frac{R}{2} \ln \Big[\det g(q^u) \Big] \,,
\tag{6.22b}
$$

being the *rigid equilibrium probability* in the soft subspace $\mathcal{E} \times \Sigma$

$$p_{\mathrm{r}}(q^u) = \frac{\exp\left[-\beta F_{\mathrm{r}}(q^u)\right]}{Z'_{\mathrm{r}}}, \quad \text{with} \quad Z'_{\mathrm{r}} := \int \exp\left[-\beta F_{\mathrm{r}}(q^u)\right] \mathrm{d}q^u . \tag{6.23}$$

Now, it is worth remarking that, although the kinetic entropy Like in the case of G, using the results in chapter 5, we can write the determinant of the reduced mass-metric tensor g as a product of two functions; one depending only on the external coordinates, and the other only on the internal ones q^i. Hence the externals-dependent factor in $\det g(q^u)$ may be integrated out independently to yield a free energy and a probability density p_{r} that depend only on the soft internals q^i (see sec 6.4.1 and chapter 5).

To end this section, we remark that it is frequent in the literature [132–134, 321, 338, 339, 352, 358, 359, 367] to define the so-called *Fixman's compensating potential* [340] as the difference between $F_{\mathrm{s}}(q^u)$, in eq. (6.11), and $F_{\mathrm{r}}(q^u)$, defined above, i.e.,

$$\begin{aligned} V_{\mathrm{F}}(q^u) \;&:=\; TS^{\mathrm{k}}_{\mathrm{r}}(q^u) - TS^{c}_{\mathrm{s}}(q^i) - TS^{\mathrm{k}}_{\mathrm{s}}(q^u) = \\ &\frac{RT}{2} \ln\left[\frac{\det G(q^u)}{\det \mathcal{H}(q^i) \det g(q^u)}\right], \end{aligned} \tag{6.24}$$

so that, performing rigid molecular dynamics simulations, which would yield an equilibrium distribution proportional to $\exp[-\beta F_{\mathrm{r}}(q^u)]$, and adding $V_{\mathrm{F}}(q^u)$ to the potential energy $V_\Sigma(q^i)$ one can reproduce instead the stiff probability density $p_{\mathrm{s}} \propto \exp[-\beta F_{\mathrm{s}}(q^u)]$ [131–133, 338, 339, 352, 357–359, 367]. This allows to obtain at a lower computational cost (due to the timescale problems discussed in the introduction) equilibrium averages that otherwise must be extracted from expensive fully flexible whole-space simulations. In fact, it seems that this particular application of the theoretical tools herein described, and not the search for the correct probability density to sample in Monte Carlo simulations, was what prompted the interest in the study of mass-metric tensors effects.

Finally, to close the theory section, in table 6.1 we summarize the equilibrium probability densities and the different correcting terms herein derived.

Classical Stiff Model	Classical Rigid Model
$p_{\mathrm{s}}(q^u) = \dfrac{\exp\left[-\beta F_{\mathrm{s}}(q^u)\right]}{Z'_{\mathrm{s}}}$	$p_{\mathrm{r}}(q^u) = \dfrac{\exp\left[-\beta F_{\mathrm{r}}(q^u)\right]}{Z'_{\mathrm{r}}}$
$F_{\mathrm{s}}(q^u) := V_\Sigma(q^i) - T(S^{c}_{\mathrm{s}}(q^i) + S^{\mathrm{k}}_{\mathrm{s}}(q^u))$	$F_{\mathrm{r}}(q^u) := V_\Sigma(q^i) - TS^{\mathrm{k}}_{\mathrm{r}}(q^u)$
$S^{\mathrm{k}}_{\mathrm{s}}(q^u) := \dfrac{R}{2} \ln\left[\det G(q^u, f^I(q^i))\right]$	$S^{\mathrm{k}}_{\mathrm{r}}(q^u) := \dfrac{R}{2} \ln\left[\det g(q^u)\right]$
$S^{c}_{\mathrm{s}}(q^i) := -\dfrac{R}{2} \ln\left[\det \mathcal{H}(q^i)\right]$	

Table 6.1: Equilibrium probability densities and correcting terms to the potential energy $V_\Sigma(q^i)$ in the classical stiff and rigid models of constraints.

6.3 Approximations in the literature

Many approximations may be done to simplify the calculation of the different correcting terms introduced in the previous sections. The most frequently found in the literature are the following three:

(i) To neglect the conformational dependence of $\det G$.

(ii) To neglect the conformational dependence of $\det \mathcal{H}$.

(iii) To assume that the hard coordinates are constant, i.e, that the $f^I(q^i)$ in eq. (5.8) do not depend on the soft coordinates q^i.

The conformational dependence of $\det g$ is not included in the points above because it is customarily regarded as important in the literature since it was shown to be non-negligible even for simple systems some decades ago [132, 133, 351, 352]. In these studies, the interest in $\det g$ normally arises in an indirect way, while studying the influence of the Fixman's compensating potential in eq. (6.24) (see discussion below), and, with this same aim, Patriciu et al. [342] have very recently measured the conformational dependence of $\det g$ for a serial polymer with fixed bond lengths and bond angles (i.e., in the approximation (iii)), showing that it is non-negligible and suggesting that it may be so also for more general systems. However, note that, only if approximations (i) and (ii) are assumed, the Fixman's potential depends on $\det g$ alone; if this is not the case, V_{F} cannot be, in general, simplified beyond the expression in eq. (6.24).

An additional relation among the determinants involved and V_{F} comes from a common simplification that is very frequently used: If one assumes approximation (iii), then the reduced mass-metric tensor g turns out to be the subblock of G with soft indices, and in such a situation, the quotient $\det G / \det g$ has been shown to be equal to $1/\det h$ by Fixman [340], where h denotes the subblock of G^{-1} with hard indices, i.e.,

$$h^{IJ}(q^\mu) := \sum_{\sigma=1}^{N} \frac{\partial q^I}{\partial x^\sigma} \frac{1}{m_\sigma} \frac{\partial q^J}{\partial x^\sigma} \,. \tag{6.25}$$

This result has been extensively used in the literature [132, 133, 321, 352, 358, 359], since each of the internal coordinates q^a typically used in macromolecular simulations only involves a small number of atoms, thus rendering the matrix h above sparse and allowing for efficient algorithms to be used in order to find its determinant.

Now, although $\det g$ is customarily regarded as important, the conformational variations of $\det G$ are almost unanimously neglected (approximation (i)) in the literature [313, 371] and may only be said to be indirectly included in h by the authors that use the expression above [132, 133, 321, 342, 352, 359]. This is mainly due to the fact, reported by Gō and Scheraga [313] and, before, by Volkenstein [377], that $\det G$ in a serial polymer may be expressed as in eq. (6.28), being independent of the dihedral angles (which are customarily taken as the soft coordinates). If one also assumes approximation (iii), which, as it will be discussed later, is very common, then $\det G$ is a constant for every conformation of the molecule.

Probably due to computational considerations, but also sometimes to the use of a formulation of the stiff case based on delta functions [367], the conformational dependence of $\det \mathcal{H}$ is almost unanimously neglected (approximation (ii)) in the literature

[313, 338, 340, 342, 355, 371, 373, 374]. Only a few authors include this term in different stages of the reasoning [131, 313, 338, 339, 351, 354, 357], most of them only to argue later that it is negligible.

Although for some simple ad hoc designed potentials that lack long-range terms [132, 133, 321], the aforementioned simplifying assumptions and the ones that will be discussed in the following paragraphs may be exactly fulfilled, in the case of the potential energies used in force fields for macromolecular simulation [104, 105, 117–126], they are not. The typical energy function in this case, has the form

$$\begin{aligned} V_{\mathrm{ff}}(q^a) \;:=\; & \frac{1}{2}\sum_{\alpha=1}^{N_r} K_{r_\alpha}(r_\alpha - r_\alpha^0)^2 + \frac{1}{2}\sum_{\alpha=1}^{N_\theta} K_{\theta_\alpha}(\theta_\alpha - \theta_\alpha^0)^2 \\ & + V_{\mathrm{ff}}^{\mathrm{tors}}(\phi_\alpha) + V_{\mathrm{ff}}^{\mathrm{long-range}}(q^a)\,, \end{aligned} \tag{6.26}$$

where r_α are bond lengths, θ_α are bond angles, ϕ_α are dihedral angles and, for the sake of simplicity, no harmonic terms have been assumed for out-of-plane angles or for hard dihedrals (such as the peptide bond ω). N_r is the number of bond lengths, N_θ the number of bond angles and the quantities K_{r_α}, K_{θ_α}, r_α^0 and θ_α^0 are constants. The term denoted by $V_{\mathrm{ff}}^{\mathrm{tors}}(\phi_\alpha)$ is a commonly included torsional potential that depends only on the dihedral angles ϕ_α and $V_{\mathrm{ff}}^{\mathrm{long-range}}(q^a)$ normally comprises long-range interactions such as Coulomb or van der Waals; hence, it depends on the atomic positions $\vec{x}'_\alpha$ which, in turn, depend on all the internal coordinates q^a.

One of the reasons given for neglecting $\det\mathcal{H}$, when classical force fields are used with potential energy functions such like the one in eq. (6.26), is that the harmonic constraining terms dominate over the rest of interactions and, since the constants appearing on these terms (the K_{r_α}, K_{θ_α} in eq. (6.26)) are independent of the conformation by construction, so is $\det\mathcal{H}$ [313, 338, 354]. In this chapter, we analyze a more realistic quantum-mechanical potential and these considerations are not applicable, however, they also should be checked in the case of classical force fields, since, for a potential energy such as the one in eq. (6.26), the quantities K_{r_α} and K_{θ_α} are finite and the long-range terms will also affect the Hessian at each point of the constrained hypersurface Σ, rendering its determinant *conformation-dependent*.

For the same reason, even in classical force fields, the equilibrium values of the hard coordinates are not the constant quantities r_α^0 and θ_α^0 in eq. (6.26) but some functions $f^I(q^i)$ of the soft coordinates (see eq. (5.8)). This fact, recognized by some authors [313, 360, 361, 378], provokes that, if one chooses to assume approximation (iii) and the constants r_α^0 and θ_α^0 appearing in eq. (6.26) are designated as the equilibrium values, the potential energy in Σ may be heavily distorted, the cause being simply that the long-range interactions between atoms separated by three covalent bonds are not fully relaxed [378]. This effect is probably larger if bond angles, and not only bond lengths, are also constrained, which may partially explain the different dynamical behaviour found in ref. 126 when comparing these types of constraints in molecular dynamics simulations. In quantum mechanical calculations of small dipeptides, on the other hand, the fact that the bond lengths and bond angles depend on the Ramachandran angles (ϕ, ψ) has been pointed out by Schäffer et al. [379]. Therefore, approximation (iii), which is very common in the literature [126, 313, 338–340, 342, 351, 354–359, 367, 368, 371, 373, 374, 380, 381], should be critically analyzed in each particular case.

Much related with the discussion above, one should note that, apart from the typical internal coordinates q^a used until now, in terms of which the constrained hypersurface Σ is described by the relations $q^I = f^I(q^i)$ in eq. (5.8), one may define a different set Q^a such that, on Σ, the corresponding hard coordinates are arbitrary constants $Q^I = C^I$ (the external coordinates q^A and Q^A are irrelevant for this part of the discussion). To do this, for example, let

$$\begin{array}{lll} Q^i := q^i & i = 7, \ldots, M+6 & \text{and} \\ Q^I := q^I - f^I(q^i) + C^I & I = M+7, \ldots, N\,. & \end{array} \tag{6.27}$$

Well then, while the relation between bond lengths, bond angles and dihedral angles (the typical q^a, such as the SASMIC ones) and the Euclidean coordinates is straightforward and simple, the expression of the transformation functions $Q^a(x^\mu)$ needs the knowledge of the f^I, which must be calculated numerically in most real cases. This drastically reduce the practical use of the Q^a, however, it is also true that they are conceptually appealing, since they have a property that closely match our intuition about what the soft and hard coordinates should be (namely, that the hard coordinates Q^I are constant on the relevant hypersurface Σ); and this is why we term them *exactly separable hard and soft coordinates*. Now, we must also point out that, although the real internal coordinates q^a do not have this property, they are usually close to it. The customary labeling of soft and hard coordinates in the literature is based on this circumstance. Somehow, the dihedral angles are the 'softest' of the internal coordinates, i.e., the ones that 'vary the most' when the system visits different regions of the hypersurface Σ; and this is why we term the real q^a *approximately separable hard and soft coordinates*, considering approximation (iii) as a useful reference case.

To sum up, the three simplifying assumptions (i), (ii) and (iii) in the beginning of this section should be regarded as approximations in the case of classical force fields, as well as in the case of the more realistic quantum-mechanical potential investigated in this chapter, and they should be critically assessed in the systems of interest. Here, while studying the model dipeptide HCO-L-Ala-NH_2, no simplifying assumptions of this type have been made.

6.4 Methods

6.4.1 Factorization reminder

In chapter 5, we have shown that the determinant of the mass-metric tensor G in eq. (6.11c) can be written as follows if the SASMIC coordinates introduced in chapter 4 are used:

$$\det G = \left(\prod_{\alpha=1}^{n} m_\alpha^3\right) \sin^2\theta \left(\prod_{\alpha=2}^{n} r_\alpha^4\right)\left(\prod_{\alpha=3}^{n} \sin^2\theta_\alpha\right)\,, \tag{6.28}$$

where the r_α are bond lengths and the θ_α bond angles.

Note that this expression does not explicitly depend on the dihedral angles. However, it may depend on them via the hard coordinates if constraints of the form presented in eq. (5.8) are used.

Also, the term depending on the masses of the atoms in the expression above may be dropped from eq. (6.11c), because it does not depend on the conformation, and the only part of $\det G$ that depend on the external coordinates, $\sin^2\theta$, may be integrated out in eq. (6.9) (θ is one of the externals q^A that describe the overall orientation of the molecule; see sec. 5.2). Hence, the kinetic entropy due to the mass-metric tensor G in the stiff case, may be written, up to additive constants, as

$$S_{\mathrm{s}}^{\mathrm{k}}(q^i) = \frac{R}{2}\left[\sum_{\alpha=2}^{n} \ln\left(r_\alpha^4\right) + \sum_{\alpha=3}^{n} \ln\left(\sin^2\theta_\alpha\right)\right], \tag{6.29}$$

where the individual contributions of each degree of freedom have been factorized.

Also in chapter 5, we have shown that the determinant of the reduced mass-metric tensor g in eq. (6.22b) can be written as follows:

$$\det g = \sin^2\theta \, \det g_2(q^i) \,, \tag{6.30}$$

being the matrix g_2

$$g_2 = \left(\begin{array}{cc|c} m_{tot} I^{(3)} & m_{tot}\, v(\vec{R}) & \cdots \quad m_{tot}\dfrac{\partial \vec{R}}{\partial q^j} \quad \cdots \\ m_{tot}\, v^T(\vec{R}) & \mathcal{J} & \cdots \sum\limits_\alpha m_\alpha \dfrac{\partial \vec{x}'_\alpha}{\partial q^j} \times \vec{x}'_\alpha \cdots \\ \hline \vdots & \vdots & \vdots \\ m_{tot}\dfrac{\partial \vec{R}}{\partial q^i} & \sum\limits_\alpha m_\alpha \left(\dfrac{\partial \vec{x}'_\alpha}{\partial q^i} \times \vec{x}'_\alpha\right)^T & \cdots \sum\limits_\alpha m_\alpha \dfrac{\partial \vec{x}'^{\,T}_\alpha}{\partial q^i}\dfrac{\partial \vec{x}'_\alpha}{\partial q^j} \cdots \\ \vdots & \vdots & \vdots \end{array}\right) \tag{6.31}$$

where $I^{(3)}$ denotes the 3×3 identity matrix and $\vec{x}'_\alpha$ is the position of atom α in the reference frame fixed in the system (the 'primed' reference frame). Additionally, we denote the *total mass* of the system by $m_{tot} := \sum_\alpha m_\alpha$, the position of the *center of mass* of the system in the primed reference frame by $\vec{R} := m_{tot}^{-1} \sum_\alpha m_\alpha \vec{x}'_\alpha$ and the *inertia tensor* of the system, also in the primed reference frame, by $\mathcal{J}$ (see eq. (5.30)).

Then, since $\sin^2\theta$ may be integrated out in eq. (6.20), we can write, omitting additive constants, the kinetic entropy associated to the reduced mass-metric tensor g depending only on the soft internals q^i:

$$S_{\mathrm{r}}^{\mathrm{k}}(q^i) = \frac{R}{2} \ln\left[\det g_2(q^i)\right]. \tag{6.32}$$

Finally, one may note that, since $\sin^2\theta$ divides out in the second line of eq. (6.24) or, otherwise stated, eqs. (6.29) and (6.32) may be introduced in the first line, then the Fixman's potential is independent from the external coordinates as well.

6.4.2 Computational methods

In the particular molecule treated in this chapter (the model dipeptide HCO-L-Ala-NH_2 in fig. 6.1; see also appendix E), the formulae in the preceding sections must be used with $M = 2$, being the internal soft coordinates $q^i \equiv (\phi, \psi)$ the typical Ramachandran angles

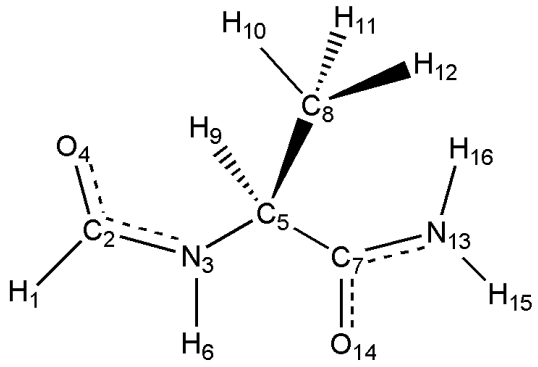


Figure 6.1: Atom numeration of the protected dipeptide HCO-L-Ala-NH_2 (see also appendix E).

(see table 6.2 and sec. 1.2), the total number of coordinates $N = 48$ and the number of hard internals $L = 40$.

Regarding the side chain angle χ, it has been argued in chapter 4 that it is soft with the same right as the angles ϕ and ψ, i.e., the barriers that hinder the rotation on this dihedral are comparable to the ones existing in the Ramachandran surface. However, the height of these barriers is sufficient ($\sim$ 6–12 RT) for the condition (ii) in sec. 6.2.1 to hold and, therefore, its inclusion in the set of hard coordinates is convenient due to its *unimportant* character (see discussion in sec. 6.2.1). Moreover, to describe the behaviour associated to χ with a probability density different from a Gaussian distribution (i.e., its potential energy different from an harmonic oscillator), for example with the tools used in

Atom name	Bond length	Bond angle	Dihedral angle
H_1			
C_2	(2,1)		
N_3	(3,2)	(3,2,1)	
O_4	(4,2)	(4,2,1)	(4,2,1,3)
C_5	(5,3)	(5,3,2)	ω_0 :=**(5,3,2,1)**
H_6	(6,3)	(6,3,2)	(6,3,2,5)
C_7	(7,5)	(7,5,3)	ϕ :=**(7,5,3,2)**
C_8	(8,5)	(8,5,3)	(8,5,3,7)
H_9	(9,5)	(9,5,3)	(9,5,3,7)
H_{10}	(10,8)	(10,8,5)	χ :=**(10,8,5,3)**
H_{11}	(11,8)	(11,8,5)	(11,8,5,10)
H_{12}	(12,8)	(12,8,5)	(12,8,5,10)
N_{13}	(13,7)	(13,7,5)	ψ :=**(13,7,5,3)**
O_{14}	(14,7)	(14,7,5)	(14,7,5,13)
H_{15}	(15,13)	(15,13,7)	ω_1 :=**(15,13,7,5)**
H_{16}	(16,13)	(16,13,7)	(16,13,7,15)

Table 6.2: SASMIC internal coordinates in Z-matrix form of the protected dipeptide HCO-L-Ala-NH_2 (see chapter 4). Principal dihedrals are indicated in bold face and their typical biochemical name is given.

the field of circular statistics [382–384], would severely complicate the derivation of the classical stiff model without adding any conceptual insight to the problem. In addition, although χ is a periodic coordinate with threefold symmetry, the considerable height of the barriers between consecutive minima allows to make the quadratic assumption in eq. (6.2) at each equivalent valley and permits the approximation of the integral on χ by three times a Gaussian integral. The multiplicative factor 3 simply adds a temperature- and conformation-independent reference to the configurational entropy S_s^c in eq. (6.11b).

The same considerations are applied to the dihedral angles, ω_0 and ω_1 (see table 6.2), that describe the rotation around the peptide bonds, and the quadratic approximation described above can also be used, since the heights of the rotation barriers around these degrees of freedom are even larger than the ones in the case of χ.

The ab initio quantum mechanical calculations have been done with the package GAMESS [305] under Linux and in 3.20 GHz PIV machines. The coordinates used for the HCO-L-Ala-NH_2 dipeptide in the GAMESS input files and the ones used to generate them with automatic Perl scripts are the SASMIC coordinates in chapter 4. They are presented in table 6.2 indicating the name of the conventional dihedral angles (see also fig. 6.1 for reference). To perform the energy optimizations, however, they have been converted to *Delocalized Coordinates* [301] in order to accelerate convergence.

First, we have calculated the PES in a regular 12×12 grid of the bidimensional space spanned by the Ramachandran angles ϕ and ψ, with both angles ranging from -165^o to 165^o in steps of 30^o. This has been done by running constrained energy optimizations at the MP2/6-31++G(d,p) level of the theory, freezing the two Ramachandran angles at each value of the grid, starting from geometries previously optimized at a lower level of the theory and setting the gradient convergence criterium to `OPTTOL`=10^{-5} and the self-consistent Hartree-Fock convergence criterium to `CONV`=10^{-6}.

The results of these calculations (which took $\sim$ 100 days of CPU time) are 144 conformations that define Σ and the values of $V_\Sigma(\phi,\psi)$ at these points (the PES itself).

Then, at each optimized point of Σ, we have calculated the Hessian matrix in the coordinates of table 6.2 removing the rows and columns corresponding to the soft angles ϕ and ψ, the result being the matrix $\mathcal{H}(\phi,\psi)$ in eq. (6.11b). This has been done, again, at the MP2/6-31++G(d,p) level of the theory, taking $\sim$ 140 days of CPU time.

Eqs. (6.29) and (6.32) in sec. 6.4.1 have been used to calculate the kinetic entropy terms associated to the determinants of the mass-metric tensors G and g, respectively. The quantities in eq. (6.29), being simply internal coordinates, have been directly extracted from the GAMESS output files via automated Perl scripts. On the other hand, in order to calculate the matrix g_2 in eq. (6.31) that appears in the kinetic entropy of the classical rigid model, the Euclidean coordinates $\vec{x}'_\alpha$ of the 16 atoms in the reference frame fixed in the system, as well as their derivatives with respect to $q^i \equiv (\phi,\psi)$, must be computed. For this, two additional 12×12 grids as the one described above have been computed; one of them displaced 2^o in the positive ϕ-direction and the other one displaced 2^o in the positive ψ-direction. This has been done, again, at the MP2/6-31++G(d,p) level of the theory, starting from the optimized structures found in the computation of the PES described above and taking $\sim$ 75 days of CPU time each grid. Using the values of the positions $\vec{x}'_\alpha$ in these two new grids and also in the original one, the derivatives of these quantities with respect to the angles ϕ and ψ, appearing in g_2, have been numerically obtained as finite differences.

The three calculations (the PES, the Hessian and the displaced PESs) have been repeated for six special points in the Ramachandran space that correspond to important elements of secondary structure (see sec. 6.5), the total CPU time needed for computing all correcting terms at these points has been $\sim$ 16 days. A total of $\sim$ 406 days of CPU time has been needed to perform the whole study at the MP2/6-31++G(d,p) level of the theory.

Finally, we have repeated all the calculations at the HF/6-31++G(d,p) level of the theory in order to investigate if this less demanding method ($\sim$ 10 days for the PES, $\sim$ 8 days for the Hessians, $\sim$ 10 days for each displaced grid, $\sim$ 2 days for the special secondary structure points, making a total of $\sim$ 40 days of CPU time) may be used instead of MP2 in further studies.

6.5 Results

In table 6.3, the maximum variation, the average and the standard deviation in the 12×12 grid defined in the Ramachandran space of the protected dipeptide HCO-L-Ala-NH_2 are shown for the three energy surfaces, V_Σ, F_s and F_r (see eqs. (6.11) and (6.22)), for the three correcting terms, $-TS_s^k$, $-TS_s^c$, and $-TS_r^k$ and for the Fixman's compensating potential V_F (see eq. (6.24)). All the functions have been referenced to zero in the grid.

In fig. 6.2, the Potential Energy Surface V_Σ, at the MP2/6-31++G(d,p) level of the theory, is depicted with the reference set to zero for visual convenience[111]. Neither the surfaces defined by F_s and F_r at the MP2/6-31++G(d,p) level of the theory nor the three energy surfaces V_Σ, F_s and F_r at HF/6-31++G(d,p) are shown graphically since they are visually very similar to the surface in fig. 6.2.

In fig. 6.3, the three correcting terms, $-TS_s^k$, $-TS_s^c$ and $-TS_r^k$ and the Fixman's compensating potential V_F, at the MP2/6-31++G(d,p) level of the theory, are depicted with the reference set to zero. The analogous surfaces at the HF/6-31++G(d,p) level of the theory are visually very similar to the ones in fig. 6.3 and have been therefore omitted.

From the results presented, one may conclude that, although the conformational dependence of the correcting terms $-TS_s^k$, $-TS_s^c$ and $-TS_r^k$ is more than an order of magnitude smaller than the conformational dependence of the Potential Energy Surface V_Σ in the worst case, if *chemical accuracy* (typically defined in the field of ab initio quantum chemistry as 1 kcal/mol [287]) is sought, then they may be relevant. In fact, they are of the order of magnitude of the differences between the energy surfaces V_Σ, F_s and F_r calculated at MP2/6-31++G(d,p) and the ones calculated at HF/6-31++G(d,p).

For the same reasons, we may conclude that, if ab initio derived potentials are used to carry out molecular dynamics simulations of peptides, the Fixman's compensating potential V_F should be included.

Finally, regarding the relative importance of the different correcting terms $-TS_s^k$, $-TS_s^c$ and $-TS_r^k$, the results in table 6.3 suggest that the less important one is the kinetic entropy $-TS_s^k$ of the stiff case (related to the determinant of the mass-metric tensor G) and that the most important one is the one related to the determinant of the Hessian matrix $\mathcal{H}$ of the constraining part of the potential, i.e., the conformational entropy $-TS_s^k$. The first

[111] At the level of the theory used in the calculations, the minimum of $V_\Sigma(\phi,\psi)$ in the grid is −416.0733418995 hartree.

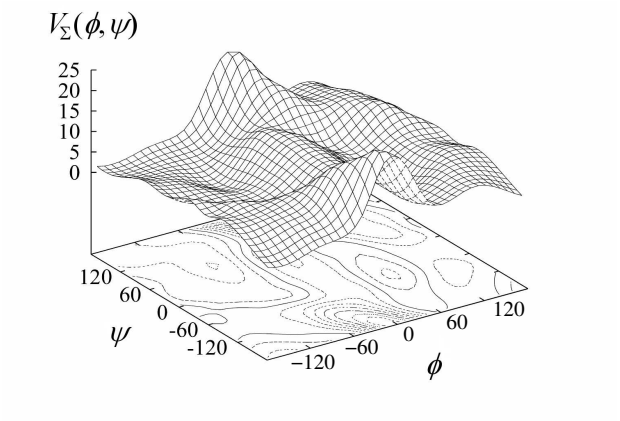


Figure 6.2: Potential Energy Surface (PES) of the model protected dipeptide HCO-L-Ala-NH_2, computed at the MP2/6-31++G(d,p) level of the theory. The surface has been referenced to zero and smoothed with bicubic splines for visual convenience. The units in the z-axis are kcal/mol.

conclusion is in agreement with the approximations typically made in the literature, the second one, however, is not (see sec. 6.3).

Now, although the relative sizes of the conformational dependence of the different terms may be indicative of their importance, the degree of correlation among the surfaces is also relevant (see table 6.5). Hence, in order to arrive to more precise conclusions, we reexamine here the results using the physically meaningful distance to compare potential energy functions that has been introduced in chapter 3, being the working set of conformations the 144 points of the grid.

In table 6.4, which contains the central results of this chapter, the distances between some of the energy surfaces that play a role in the problem are shown. We present the result in units of RT (at 300^o K, where $RT \simeq 0.6$ kcal/mol) because, as it has already been argued, if the distance between two different approximations of the energy of the

	MP2/6-31++G(d,p)			HF/6-31++G(d,p)		
	Max.[a]	Ave.[b]	Std.[c]	Max.[a]	Ave.[b]	Std.[c]
V_Σ	21.64	6.76	3.88	23.62	6.92	4.35
F_s	21.43	6.47	3.93	23.78	7.17	4.38
F_r	21.09	6.46	3.82	23.09	6.76	4.31
$-TS_s^k$	0.24	0.09	0.05	0.23	0.09	0.04
$-TS_s^c$	1.67	0.98	0.32	1.34	0.63	0.30
$-TS_r^k$	0.81	0.37	0.12	0.75	0.38	0.12
V_F	1.68	0.89	0.30	1.35	0.55	0.27

Table 6.3: [a]Maximum variation, [b]average and [c]standard deviation in the 12×12 grid defined in the Ramachandran space of the protected dipeptide HCO-L-Ala-NH_2 for the three energy surfaces, V_Σ, F_s and F_r, the three correcting terms, $-TS_s^k$, $-TS_s^c$, and $-TS_r^k$ and the Fixman's compensating potential V_F. The results at both MP2/6-31++G(d,p) and HF/6-31++G(d,p) levels of the theory are presented and all the functions have been referenced to zero in the grid. The units used are kcal/mol.

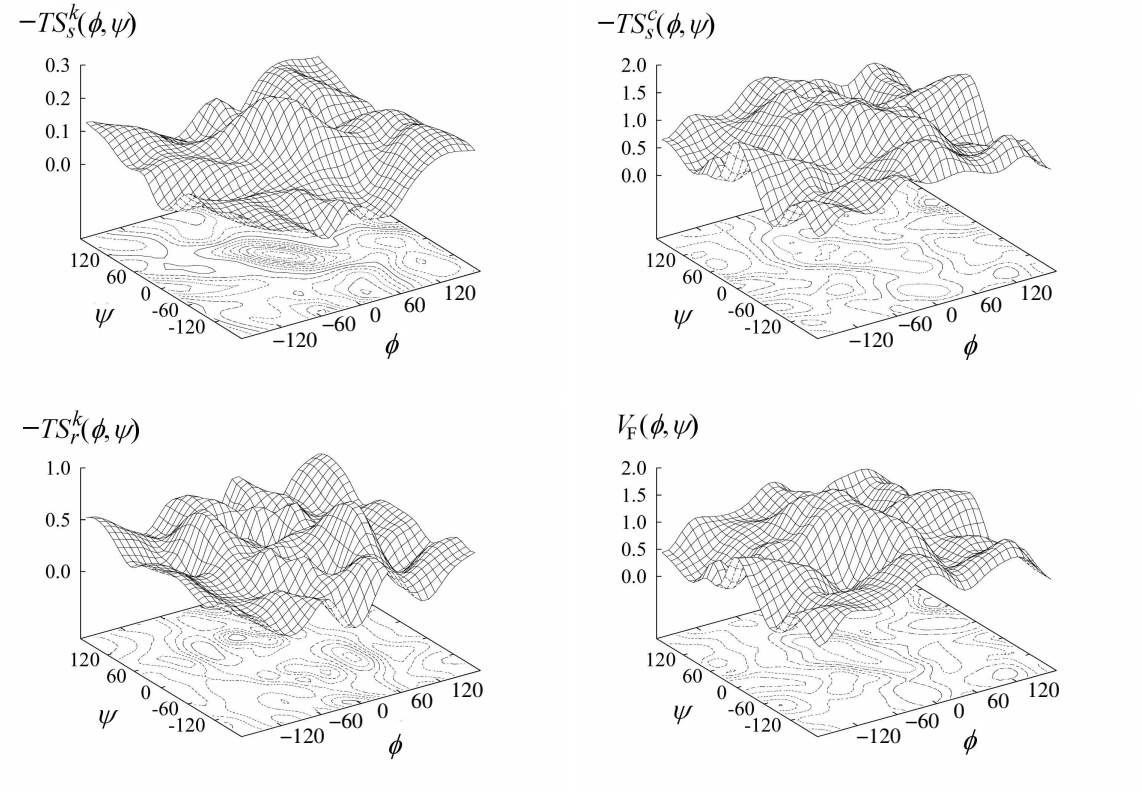


Figure 6.3: Ramachandran plots of the correcting terms appearing in eqs. (6.11) and (6.22), together with the Fixman's compensating potential defined in eq. (6.24), computed at the MP2/6-31++G(d,p) level of the theory in the model dipeptide HCO-L-Ala-NH_2. The surfaces have been referenced to zero and smoothed with bicubic splines for visual convenience. The units in the z-axes are kcal/mol.

same system is less than RT, one may safely substitute one by the other without altering the relevant physical properties. Moreover, if one assumes that the effective energies compared will be used to construct a polypeptide potential and that it will be designed as simply the sum of mono-residue ones, then, the number N_{res} of residues up to which one may go keeping the distance between the two approximations of the the N-residue potential below RT is (see chapter 3):

$$N_{\text{res}} = \left(\frac{RT}{d_{12}}\right)^2 . \tag{6.33}$$

This number is also shown in table 6.4, together with the slope b_{12} of the linear rescaling between V_1 and V_2 and the Pearson's correlation coefficient [296], denoted by r_{12}.

The results at both MP2/6-31++G(d,p) and HF/6-31++G(d,p) levels of the theory are presented. The first three rows in each of the first two blocks are related to the classical stiff model, the next row to the classical rigid model and the last one in each block to the comparison between the two models. The third block in the table is associated to the comparison between the two different levels of the theory used.

The F_s vs. V_Σ row (in the first two blocks) assess the importance of the two correcting

Corr.[a]	V_1[b]	V_2[c]	d_{12}[d]	N_{res}[e]	b_{12}[f]	r_{12}[g]
		MP2/6-31++G(d,p)				
$-TS_{\text{s}}^{\text{k}} - TS_{\text{s}}^{\text{c}}$	F_{s}	V_Σ	0.74 RT	1.82	0.98	0.9967
$-TS_{\text{s}}^{\text{c}}$	F_{s}	$V_\Sigma - TS_{\text{s}}^{\text{k}}$	0.74 RT	1.83	0.98	0.9967
$-TS_{\text{s}}^{\text{k}}$	F_{s}	$V_\Sigma - TS_{\text{s}}^{\text{c}}$	0.11 RT	80.45	1.00	0.9999
$-TS_{\text{r}}^{\text{k}}$	F_{r}	V_Σ	0.29 RT	11.62	1.01	0.9995
V_{F}	F_{s}	F_{r}	0.67 RT	2.24	0.97	0.9972
		HF/6-31++G(d,p)				
$-TS_{\text{s}}^{\text{k}} - TS_{\text{s}}^{\text{c}}$	F_{s}	V_Σ	0.73 RT	1.90	0.99	0.9975
$-TS_{\text{s}}^{\text{c}}$	F_{s}	$V_\Sigma - TS_{\text{s}}^{\text{k}}$	0.71 RT	2.00	0.99	0.9976
$-TS_{\text{s}}^{\text{k}}$	F_{s}	$V_\Sigma - TS_{\text{s}}^{\text{c}}$	0.10 RT	90.99	1.00	0.9999
$-TS_{\text{r}}^{\text{k}}$	F_{r}	V_Σ	0.26 RT	14.83	1.01	0.9997
V_{F}	F_{s}	F_{r}	0.61 RT	2.69	0.98	0.9982
		MP2/6-31++G(d,p) vs. HF/6-31++G(d,p)				
	V_Σ	V_Σ	1.25 RT	0.64	1.12	0.9925
	F_{s}	F_{s}	1.18 RT	0.72	1.11	0.9934
	F_{r}	F_{r}	1.18 RT	0.72	1.12	0.9932

Table 6.4: Comparison of different energy surfaces involved in the study of the constrained equilibrium of the protected dipeptide HCO-L-Ala-NH_2. [a]Correcting term whose importance is measured in the corresponding row, [b]reference potential energy V_1 (the 'correct' one, the one containing the correcting term), [c]approximated potential energy V_2 (i.e, V_1 minus the correcting term in column *a*), [d]statistical distance between V_1 and V_2, [e]maximum number of residues in a polypeptide potential up to which the correcting term in column *a* may be omitted, [f]slope of the linear rescaling between V_1 and V_2 and [g]Pearson's correlation coefficient. All quantities are dimensionless, except for d_{12} which is given in units of the thermal energy RT at 300° K.

terms, $-TS_{\text{s}}^{\text{k}}$ and $-TS_{\text{s}}^{\text{c}}$, in the stiff case. The result $d_{12} = 0.74\,RT$ indicates that, for the alanine dipeptide, V_Σ may be used as an approximation of F_{s} with caution if accurate results are sought. In fact, the low value of $N_{\text{res}} = 1.82 < 2$ shows that, if we wanted to describe a 2-residue peptide omitting the stiff correcting terms, we would typically make an error greater than the thermal noise in the energy differences. The next two rows investigate the effect of each one of the individual correcting terms. The conclusion that can be extracted from them (as the relative sizes in table 6.3 already suggested) is that the conformational entropy associated to the determinant of the Hessian matrix $\mathcal{H}$ is much more relevant than the correcting term $-TS_{\text{s}}^{\text{k}}$, related to the mass-metric tensor G, allowing to drop the latter up to ~ 80 residues (according to MP2/6-31++G(d,p) calculations). As it has been already remarked, this second conclusion is in agreement with the approximations frequently done in the literature; however, it turns out that the importance of the

V_1 [a]		V_2 [b]	r_{12} [c]
		MP2/6-31++G(d,p)	
V_Σ	vs.	$-TS_s^c$	0.1572
V_Σ	vs.	$-TS_s^k$	-0.0008
V_Σ	vs.	$-TS_r^k$	-0.3831
V_Σ	vs.	V_F	0.3334
		HF/6-31++G(d,p)	
V_Σ	vs.	$-TS_s^c$	0.0682
V_Σ	vs.	$-TS_s^k$	0.0897
V_Σ	vs.	$-TS_r^k$	-0.3544
V_Σ	vs.	V_F	0.2404
		MP2/6-31++G(d,p) vs. HF/6-31++G(d,p)	
$-TS_s^c$	vs.	$-TS_s^c$	0.9136
$-TS_s^k$	vs.	$-TS_s^k$	0.9808
$-TS_r^k$	vs.	$-TS_r^k$	0.9316
V_F	vs.	V_F	0.9217

Table 6.5: Correlation between the different correcting terms involved in the study of the constrained equilibrium of the protected dipeptide HCO-L-Ala-NH_2. [a]Reference potential energy, [b]approximated potential energy, [c]Pearson's correlation coefficient.

Hessian-related term has been persistently underestimated (see sec. 6.3).

The F_r vs. V_Σ row, in turn, shows the data associated to the kinetic entropy term $-TS_r^k$, which is related to the determinant of the reduced mass-metric tensor g in the classical rigid model. From the results there ($d_{12} = 0.29\,RT$ and $N_{res} = 11.62$ at the MP2/6-31++G(d,p) level), we can conclude that the only correction term in the rigid case is less important than the ones in the stiff case and that V_Σ may be used as an approximation of F_r for oligopeptides of up to ~ 12 residues.

The last row in each of the first two blocks in table 6.4 is related to the interesting question in molecular dynamics of whether or not one should include the Fixman's compensating potential V_F (see eq. (6.24)) in rigid simulations in order to obtain the stiff equilibrium distribution, $\exp(-\beta F_s)$, instead of the rigid one, $\exp(-\beta F_r)$. This question is equivalent to asking whether or not F_r is a good approximation of F_s. From the results in the table, we can conclude that the Fixman's potential is relevant for peptides of more than 2 residues and its omission may cause an error greater than the thermal noise in the energy differences.

The appreciable sizes of the different correcting terms, shown in table 6.3, together with their low correlation with the Potential Energy Surface V_Σ, presented in the first two blocks of table 6.5, explain their considerable relevance discussed in the preceding paragraphs.

Moreover, from the comparison of the MP2/6-31++G(d,p) and the HF/6-31++G(d,p)

blocks, one can tell that the study herein performed may well have been done at the lower level of the theory (if we had known) with a tenth of the computational effort (see sec. 6.4). This fact, explained by the high correlation, presented in the third block of table 6.5, between the correcting terms calculated at the two levels, is very relevant for further studies on more complicated dipeptides or longer chains and it indicates that the differences in size between the different correcting terms at MP2/6-31++G(d,p) and HF/6-31++G(d,p), which are presented in table 6.3, are mostly due to a harmless linear scaling effect similar to the well-known empirical scale factor frequently used in ab initio vibrational analysis [240, 316, 385]. This view is supported by the data in the third block of table 6.4, related to the comparison between the energy surfaces calculated at MP2/6-31++G(d,p) and HF/6-31++G(d,p), where the slopes b_{12} are consistently larger than unity.

A last conclusion that may be extracted from the block labeled 'MP2/6-31++G(d,p) vs. HF/6-31++G(d,p)' in table 6.4 is that the typical error in the energy differences (given by the distances d_{12}) produced when one reduces the level of the theory from MP2/6-31++G(d,p) to HF/6-31++G(d,p) is comparable (less than twice) to the error made if the most important correcting terms of the classical constrained models studied in this chapter are dropped. This is a useful hint for researchers interested in the conformational analysis of peptides with quantum chemistry methods [163, 207, 210–214] and also to those whose aim is the design and parametrization of classical force fields from ab initio quantum mechanical calculations [159, 163, 164].

Finally, in order to enrich and qualify the analysis, a new *working set* of conformations, different from the 144 points of the grid in the Ramachandran space, have been selected and the whole study has been repeated on them. These new conformations are six important secondary structure elements which form repetitive patterns stabilized by hydrogen bonds in polypeptides. Their conventional names and the corresponding values of the ϕ and ψ angles have been taken from ref. 10, have already been presented in table 1.1 and are recalled in table 6.6 for convenience.

In fig. 6.4, the relative energies of these conformations are shown for the three relevant potentials, V_Σ, F_s and F_r, at both MP2/6-31++G(d,p) and HF/6-31++G(d,p) levels of the theory. Since the antiparallel β-sheet is the structure with the minimum energy in all the cases, it has been set as the reference and the rest of energies in the figure should be regarded as relative to it.

	ϕ	ψ
α-helix	−57	−47
3_{10}-helix	−49	−26
π-helix	−57	−70
polyproline II	−79	149
parallel β-sheet	−119	113
antiparallel β-sheet	−139	135

Table 6.6: Ramachandran angles (in degrees) of some important secondary structure elements in polypeptides. Data taken from ref. 10.

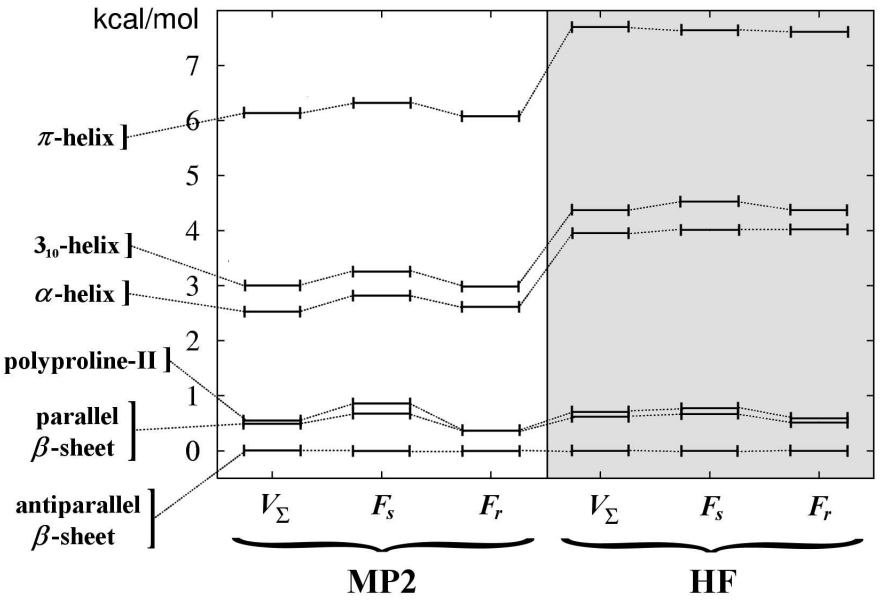


Figure 6.4: Relative energies of some important elements of secondary structure for the three potentials V_Σ, F_s and F_r, in the model dipeptide HCO-L-Ala-NH_2 and at both MP2/6-31++G(d,p) and HF/6-31++G(d,p) levels of the theory. The energy of the antiparallel β-sheet has been taken as reference. The units are kcal/mol.

The meaningful assessment, using the statistical distance described above, of the typical error made in the energy differences has been also performed on this new working set of conformations. The results are presented in table 6.7.

The distances between the free energies, F_s and F_r, and their corresponding approximations obtained dropping the correcting entropies, $-TS_s^k$, $-TS_s^c$ and $-TS_r^k$, or the Fixman's compensating potential V_F, in the first two blocks of the table, are consistently smaller than the ones found in the study of the grid defined in the whole Ramachandran space (cf. table 6.4). And so are the distances between the three relevant potentials, V_Σ, F_s and F_r, calculated at the MP2/6-31++G(d,p) and HF/6-31++G(d,p) levels of the theory.

Although the distance d_{12} used is a statistical quantity and, therefore, one must be cautious when working with such a small set of conformations (of size six, in this case), the conclusion drawn from this second part of the study is that, if one is interested only in the 'lower region' of the Ramachandran surface, where the typical secondary structure elements lie, then, one may safely neglect the conformational dependence of the different correcting terms appearing in the study of the constrained equilibrium of peptides, at least up to oligopeptides (poly-alanines) of ~ 10 residues in the worst case (the neglect of the Fixman's compensating potential V_F in the F_s vs. F_r comparison at MP2/6-31++G(d,p)).

This difference between the two working set of conformations may be explained looking at one of the ways of expressing the distance d_{12} (eq. (3.12a) in chapter 3, which we repeat here):

$$d_{12} = \sqrt{2}\,\sigma_2(1 - r_{12}^2)^{1/2}\,, \tag{6.34}$$

where r_{12} is the Pearson's correlation coefficient between the potential energies de-

Corr.[a]	V_1[b]	V_2[c]	d_{12}[d]	N_{res}[e]	b_{12}[f]	r_{12}[g]
		MP2/6-31++G(d,p)				
$-TS_s^k - TS_s^c$	F_s	V_Σ	0.22 RT	19.72	0.99	0.9990
$-TS_s^c$	F_s	$V_\Sigma - TS_s^k$	0.26 RT	14.07	0.98	0.9985
$-TS_s^k$	F_s	$V_\Sigma - TS_s^c$	0.06 RT	298.13	1.01	0.9999
$-TS_r^k$	F_r	V_Σ	0.20 RT	25.64	0.99	0.9992
V_F	F_s	F_r	0.34 RT	8.73	0.99	0.9977
		HF/6-31++G(d,p)				
$-TS_s^k - TS_s^c$	F_s	V_Σ	0.14 RT	47.94	1.00	0.9997
$-TS_s^c$	F_s	$V_\Sigma - TS_s^k$	0.15 RT	46.12	1.00	0.9997
$-TS_s^k$	F_s	$V_\Sigma - TS_s^c$	0.05 RT	380.30	1.00	0.9999
$-TS_r^k$	F_r	V_Σ	0.15 RT	41.85	0.99	0.9997
V_F	F_s	F_r	0.18 RT	30.12	1.01	0.9996
		MP2/6-31++G(d,p) vs. HF/6-31++G(d,p)				
	V_Σ	V_Σ	0.77 RT	1.68	1.28	0.9929
	F_s	F_s	0.77 RT	1.69	1.26	0.9928
	F_r	F_r	0.71 RT	1.96	1.28	0.9939

Table 6.7: Comparison of different approximations to the energies of some important elements of secondary structure (see table 6.6) in the study of the constrained equilibrium of the protected dipeptide HCO-L-Ala-NH_2. See the caption of table 6.4 for an explanation of the keys in the different columns.

noted by V_1 and V_2 and σ_2 is the standard deviation in the values of V_2 on the relevant working set of conformations.

This last quantity, σ_2, is the responsible of the differences between tables 6.4 and 6.7, since the set of conformations comprised by the six secondary structure elements in table 6.6 spans a smaller energy range than the whole PES in fig. 6.2 (or F_s, or F_r, which have very similar variations). Accordingly, the dispersion in the energy values is smaller: $\sigma_2 \simeq 2$ kcal/mol in the case of the secondary structure elements and $\sigma_2 \simeq 4$ kcal/mol for the grid in the whole Ramachandran space (see table 6.3). Since the correlation coefficient in both cases are of similar magnitude, the differences in σ_2 produce a smaller distance d_{12} for the second set of conformations studied, i.e., a smaller typical error made in the energy differences when omitting the correcting terms derived from the consideration of constraints.

To end this section, an important remark: Although this 'lower region' of the Ramachandran space contains the most relevant secondary structure elements (which are also the most commonly found in experimentally resolved native structures of proteins [41, 52, 386, 387]) and may be the only region explored in the dynamical or thermody-

namical study of small peptides, if the aim is the design of effective potentials for computer simulation of polypeptides [159, 163, 164], then, some caution is recommended, since long-range interactions in the sequence may temporarily compensate local energy penalizations and the higher regions of the energy surfaces studied could be important in transition states or in some relevant dynamical paths of the system.

In the following section, the many results discussed in the preceding paragraphs are summarized.

6.6 Conclusions

In this chapter, the theory of classical constrained equilibrium has been collected for the stiff and rigid models. The pertinent correcting terms, which may be regarded as effective entropies, as well as the Fixman's compensating potential, have been derived and theoretically discussed (see eqs. (6.11), (6.22) and (6.24), together with the formulae in sec. 6.4.1). In addition, the common approximation of considering that, for typical internals, the equilibrium values of the hard coordinates do not depend on the soft ones, has also been discussed and related to the rest of simplifications. The treatment of the assumptions in the literature is thoroughly reviewed and discussed in sec. 6.3.

In the central part of the work presented in this chapter (sec. 6.5), the relevance of the different correcting terms has been assessed in the case of the model dipeptide HCO-L-Ala-NH_2, with quantum mechanical calculations including electron correlation. Also, the possibility of performing analogous studies at the less demanding Hartree-Fock level of the theory has been investigated. The results found are summarized in the following points:

- In Monte Carlo simulations of the classical stiff model at room temperature, the effective entropy $-TS_{\mathrm{s}}^{\mathrm{k}}$, associated to the determinant of the mass-metric tensor G, may be neglected for peptides of up to ~ 80 residues. Its maximum variation in the Ramachandran space is 0.24 kcal/mol.

- In Monte Carlo simulations of the classical stiff model at room temperature, the effective entropy $-TS_{\mathrm{s}}^{\mathrm{c}}$, associated to the determinant of the Hessian $\mathcal{H}$ of the constraining part of the potential, should be included for peptides of more than 2 residues. Its maximum variation in the Ramachandran space is 1.67 kcal/mol.

- In Monte Carlo simulations of the classical rigid model at room temperature, the effective entropy $-TS_{\mathrm{r}}^{\mathrm{k}}$, associated to the determinant of the reduced mass-metric tensor g, may be neglected for peptides of up to ~ 12 residues. Its maximum variation in the Ramachandran space is 0.81 kcal/mol.

- In rigid molecular dynamics simulations intended to yield the stiff equilibrium distribution at room temperature, the Fixman's compensating potential V_{F} should be included for peptides of more than 2 residues. Its maximum variation in the Ramachandran space is 1.68 kcal/mol.

- If the assumption that only the more stable region of the Ramachandran space, where the principal elements of secondary structure lie, is relevant, then, the importance of the correcting terms decreases and the limiting number of residues in a

polypeptide potential up to which they may be omitted is approximately four times larger in each of the previous points.

- In both cases (i.e., either if the whole Ramachandran space is considered relevant, or only the lower region), the errors made if the most important correcting terms are neglected are of the same order of magnitude as the errors due to a decrease in the level of theory from MP2/6-31++G(d,p) to HF/6-31++G(d,p).

- The whole study of the relevance of the different correcting terms (or future analogous investigations) may be performed at the HF/6-31++G(d,p) level of the theory, yielding very similar results to the ones obtained at MP2/6-31++G(d,p) and using a tenth of the computational effort.

To end this discussion, some qualifications should be made. On the one hand, the conclusions above refer to the case in which a classical potential directly extracted from the quantum mechanical (Born-Oppenheimer) one is used; for the considerably simpler force fields typically used for macromolecular simulations, the study should be repeated and different results may be obtained. On the other hand, the investigation performed in this chapter has been done in one of the simplest dipeptides; both its isolated character and the relatively small size of its side chain play a role in the results obtained. Hence, for bulkier residues included in polypeptides, these conclusions should be approached with caution and much interesting work remains to be done.

Chapter 7

Efficient model chemistries for peptides. Split-valence Gaussian basis sets and the heterolevel approximation.

> As soon as an Analytical Engine exists, it will necessarily guide the future course of the science. Whenever any result is sought by its aid, the question will then arise — what course of calculation can these results be arrived at by the machine in the shortest time? [388]
>
> — Charles Babbage, 1864

7.1 Introduction

The study and characterization of the conformational behaviour of oligopeptides [163, 245, 389, 390] and, specially, dipeptides [163, 207–215] is an unavoidable first step in any bottom-up approach to the protein folding problem (see refs. 74, 87, 107, 111, 112 and chapter 1 for an introduction to it). Although classical force fields [104, 105, 117–126] are the only computationally feasible choice for simulating large molecules at present, they have been shown to yield inaccurate potential energy surfaces (PESs) for dipeptides [208] and it is widely recognized that they are unable to capture the fine details needed to correctly describe the intricacies of the whole protein folding process [33, 84, 98, 100, 158–161]. On the other hand, albeit prohibitively demanding in terms of computational resources, ab initio quantum mechanical calculations (see chapter 2) are not only regarded as the correct physical description that in the long run will be the preferred choice to directly tackle proteins (given the exponential growth of computer power; see fig. 1.2d), but they are also used in small peptides as the reference against which less accurate methods must be compared [163, 208] in order to parameterize additive, classical force fields for polypeptides.

Now, despite the sound theoretical basis, in practical quantum chemistry calculations a plethora of approximations must be typically made if one wants to obtain the final results in a reasonable human time. The exact ‘recipe’ that includes all the assumptions and steps

needed to calculate the relevant observables for any molecular system has been termed *model chemistry* (MC) by John Pople. In his own words, a MC is an "approximate but well-defined general and continuous mathematical procedure of simulation" [220].

The two starting approximations to the exact Schrödinger equation that a MC must contain have been described in chapter 2 and they are, first, the truncation of the N-electron space (in wavefunction-based methods) or the choice of the functional (in DFT) and, second, the truncation of the one-electron space, via the LCAO scheme (in both cases). The extent up to which the first truncation is carried (or the functional chosen in the case of DFT) is commonly called the *method* and it is denoted by acronyms such as RHF, MP2, B3LYP, CCSD(T), FCI, etc., whereas the second truncation is embodied in the definition of a finite set of atom-centered Gaussian functions termed *basis set* (see sec. 2.9), which is also designated by conventional short names, such as 6-31+G(d), TZP or cc-pVTZ(–f). If we denote the method by a capital M and the basis set by a B, the specification of both shall be denoted by M/B and called a *level of the theory*. Typical examples of this are RHF/3-21G or MP2/cc-VDZ.

Such levels of the theory are, by themselves, valid model chemistries, however, it is very common [163, 207, 246, 255] to use different levels to perform, first, a possibly constrained geometry optimization and, then, a single-point energy calculation on the resulting structures. If we denote by $L_i := M_i/B_i$ a given level of the theory, this 'mixed' calculation is indicated by $L_E//L_G$, where the level L_E at which the single-point energy calculation is performed is written first [220]. Herein, if $L_E \neq L_G$, we shall call $L_E//L_G$ an *heterolevel* model chemistry; whereas, if $L_E = L_G$ it will be termed a *homolevel* one, and it will be typically abbreviated omitting the 'double slash' notation.

Apart from the approximations described above, which are the most commonly used and the only ones that are considered in this chapter, the model chemistry concept may include a lot of additional features: protocols for extrapolating to the infinite-basis set limit [219, 391–394], additivity assumptions [395–398], extrapolations of the Møller-Plesset series to infinite order [399], removal of the so-called *basis set superposition error* (BSSE) [400–406], etc. The reason behind most of these techniques is the urging need to reduce the computational cost (and hence the price) of the calculations. For example, in the case of the heterolevel approximation, this economy principle forces the level L_E at which the single-point energy calculation is performed to be more accurate and more numerically demanding than L_G; the reason being simply that, while we must compute the energy only once at L_E, we need to calculate several times the energy and its gradient with respect to the unconstrained internal nuclear coordinates at level L_G (the actual number depending on the starting structure, the algorithms used and the size of the system). Therefore, it would be pointless to use an heterolevel model chemistry $L_E//L_G$ in which L_G is more expensive than L_E, since, at the end of the geometry optimization, the energy at level L_E is available.

Now, although general applicability is a requirement that all model chemistries must satisfy, general accuracy is not mandatory. The truth is that the different procedures that conform a given MC are typically parameterized and tested in very particular systems, which are often small molecules. Therefore, the validity of the approximations outside that native range of problems must be always questioned and checked, and, while the computational cost of a given model chemistry is easy to evaluate, its expected accuracy on a particular problem could be difficult to predict a priori, specially if we are dealing

with large molecules in which interactions in very different energy scales are playing a role. The description of the conformational behaviour of peptides (or, more generally, flexible organic species), via their potential energy surfaces in terms of the soft internal coordinates, is one of such problems and the one that is treated in this chapter.

Our aim is here is to provide an exhaustive study of the *Restricted Hartree Fock* (RHF) and *Møller-Plesset 2* (MP2) methods (described in chapter 2), using the split-valence family of basis sets devised by Pople and collaborators [231–238]. To this end, we compare the PES of the model dipeptide HCO-L-Ala-NH_2 (see appendix E and fig. 7.1) calculated with a large number of homo- and heterolevels, and assess the efficiency of the different model chemistries by comparison with a reference PESs.

Special interest is devoted to the evaluation of the influence of polarization and diffuse functions in the basis sets, distinguishing between those placed at heavy atoms and those placed at hydrogens, as well as the effect of different contraction and valence-splitting schemes.

The second objective of this study, and probably the main one, is the investigation of the *heterolevel assumption*, which is defined here to be that which states that *heterolevel model chemistries are more efficient than homolevel ones*. The heterolevel assumption is very commonly assumed in the literature [163, 246, 255], but it is seldom checked. As far as we know, the only tests for peptides or related systems, have been performed using a small number of conformers [242, 243], and this is the first time that this potentially very economical approximation is tested in full PESs.

In sec. 7.2.1, the methodological details regarding the quantum mechanical calculations performed in this work are provided. In sec. 7.2.2, a brief summary of the meaning and the properties of the distance introduced in chapter 3 is given for reference. Next, in sec. 7.2.3, we discuss the rules and criteria that have been used in order to reasonably sample the enormous space of all Pople's basis sets. In sec. 7.3, the main results of the investigation are presented. For convenience, they are organized into four different subsections: in sec. 7.3.1, a RHF//RHF-intramethod study is performed, whereas the MP2 analogous is presented in sec. 7.3.2. In sec. 7.3.3, a small interlude is dedicated to reflect on the general ideas and concepts underlying an investigation such as this one, and also to compare the RHF and MP2 results obtained in the previous two sections. In sec. 7.3.4, heterolevel model chemistries in which the geometry is calculated at RHF and the energy at MP2 are evaluated. Finally, sec. 7.4 is devoted to give a brief summary of the most important conclusions of the work.

7.2 Methods

7.2.1 Quantum mechanical calculations and internal coordinates

All ab initio quantum mechanical calculations have been performed using the Gaussian03 program [49] under Linux and in 3.20 GHz PIV machines with 2 GB RAM memory. The internal coordinates used for the Z-matrix of the HCO-L-Ala-NH_2 dipeptide in the Gaussian03 input files (automatically generated with Perl scripts) are the SASMIC ones introduced in chapter 4, and which are again presented in table 7.1 (see also fig. 7.1 for reference). For the geometry optimizations, the SASMIC scheme has been used too (`Opt=Z-matrix` option) instead of the default redundant internal coordinates provided by

Atom name	Bond length	Bond angle	Dihedral angle
H_1			
C_2	(2,1)		
N_3	(3,2)	(3,2,1)	
O_4	(4,2)	(4,2,1)	(4,2,1,3)
C_5	(5,3)	(5,3,2)	(5,3,2,1)
H_6	(6,3)	(6,3,2)	(6,3,2,5)
C_7	(7,5)	(7,5,3)	ϕ :=**(7,5,3,2)**
C_8	(8,5)	(8,5,3)	(8,5,3,7)
H_9	(9,5)	(9,5,3)	(9,5,3,7)
H_{10}	(10,8)	(10,8,5)	(10,8,5,3)
H_{11}	(11,8)	(11,8,5)	(11,8,5,10)
H_{12}	(12,8)	(12,8,5)	(12,8,5,10)
N_{13}	(13,7)	(13,7,5)	ψ :=**(13,7,5,3)**
O_{14}	(14,7)	(14,7,5)	(14,7,5,13)
H_{15}	(15,13)	(15,13,7)	(15,13,7,5)
H_{16}	(16,13)	(16,13,7)	(16,13,7,15)

Table 7.1: Internal coordinates in Z-matrix form of the protected dipeptide HCO-L-Ala-NH_2 according to the SASMIC scheme introduced in chapter 4. The numbering of the atoms is that in fig. 7.1, and the soft Ramachandran angles ϕ and ψ are indicated.

Gaussian03, since we have seen that, when soft coordinates, such as the Ramachandran angles, are held fixed and mostly hard coordinates are let vary, the use of the SASMIC scheme slightly reduces the time to converge with respect to the redundant internals (for unconstrained optimizations, on the other hand, the redundant coordinates seem to slightly outperform the SASMIC ones).

All PESs in this study have been discretized into a regular 12×12 grid in the bidimensional space spanned by the Ramachandran angles ϕ and ψ, with both of them ranging from -165^o to 165^o in steps of 30^o.

To calculate the geometry at a particular level of the theory, we have run constrained energy optimizations at each point of the grid, freezing the two Ramachandran angles

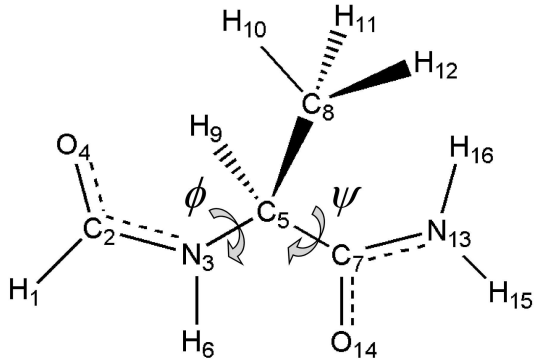


Figure 7.1: Atom numeration of the protected dipeptide HCO-L-Ala-NH_2 according to the SASMIC scheme introduced in chapter 4 (see also appendix E). The soft Ramachandran angles ϕ and ψ are also indicated.

ϕ and ψ at the corresponding values, and, in order to save computational resources, the starting structures were taken, when possible, from PESs previously optimized at a lower level of the theory. The convergence criterium for RHF optimizations has been set to `Opt=Tight`, while, in the case of MP2, an intermediate option of `IOp(1/7=100)` has been used (note that `Opt=Tight` corresponds to `IOp(1/7=10)`, whereas the default criterium is `IOp(1/7=300)`). The resulting geometries have been automatically extracted by Perl scripts and used to construct the input files for the heterolevel calculations.

The self-consistent Hartree-Fock convergence criterium have been set in all cases to `SCF=(Conver=10)` (tighter than `SCF=Tight`) and the MP2 calculations have been performed in the (default) frozen-core approximation (see sec. 2.10).

At the Hartree-Fock level, 142 potential energy surfaces have been calculated, taking a total of $\sim$ 3.7 years of CPU time, whereas, at MP2, 35 PESs have been computed and the time invested amounts to $\sim$ 4.5 years, from which, the highest level PES computed in this study, the MP2/6-311++G(2df,2pd) one depicted in fig 7.7, has taken $\sim$ 3 years of computer time. Finally, 88 PESs have been calculated with MP2//RHF-intermethod model chemistries, taking $\sim$ 6 months. In total, 265 potential energy surfaces of the model dipeptide formyl-L-alanine-amide have been computed for this study, taking $\sim$ 8.7 years of CPU time.

Finally, let us note that the correcting terms coming from mass-metric tensors determinants have not been included in this study for obvious computational reasons. Due to the conclusions arrived in chapter 6, one of the future research directions that must be followed is precisely the exploration of their influence in the relative efficiency of the model chemistries studied here.

7.2.2 Physically meaningful distance

In order to compare the PESs produced by the different (homo- and heterolevel) model chemistries, the distance introduced in chapter 3 has been used. Let us recall here that this *distance*, denoted by d_{12}, profits from the complex nature of the system studied to compare any two different potential energy functions, V_1 and V_2, and, from a working set of conformations (in this case, the 144 points of each PES), it statistically measures the typical error that one makes in the *energy differences* if V_2 is used instead of the more accurate V_1, admitting a linear rescaling and a shift in the energy reference.

This distance, which has energy units, presents better properties than other quantities customarily used to perform these comparisons, such as the energy RMSD, the average energy error, etc., and it may be related to the Pearson's correlation coefficient r_{12} by

$$d_{12} = \sqrt{2}\,\sigma_2(1 - r_{12}^2)^{1/2}\,, \tag{7.1}$$

where σ_2 is the standard deviation of V_2.

Also, due to its physical meaning, it has been argued in chapter 3 that, if the distance between two different approximations of the energy of the same system is less than RT, one may safely substitute one by the other without altering the relevant dynamical or thermodynamical behaviour. Consequently, we shall present the results in units of RT (at 300^o K, so that $RT \simeq 0.6$ kcal/mol).

Finally, if one assumes that the effective energies compared will be used to construct a polypeptide potential and that it will be designed as simply the sum of mono-residue

ones (as a first exploratory approximation), then, the number N_{res} of residues up to which one may go keeping the distance d_{12} between the two approximations of the the N-residue potential below RT is

$$N_{\text{res}} = \left(\frac{RT}{d_{12}}\right)^2 . \tag{7.2}$$

Now, according to the value taken by N_{res} for a comparison between a fixed reference PES, denoted by V_1, and a candidate approximation, denoted by V_2, we divide all the efficiency plots in sec 7.3 in three regions depending on the accuracy: the *protein region*, corresponding to $0 < d_{12} \leq 0.1RT$, or, equivalently, to $100 \leq N_{\text{res}} < \infty$; the *peptide region*, corresponding to $0.1RT < d_{12} \leq RT$, or $1 \leq N_{\text{res}} < 100$; and, finally, the *inaccurate region*, where $d_{12} > RT$, and even for a dipeptide it is not advisable to use V_2 as an approximation to V_1.

7.2.3 Basis set selection

In the whole study presented in this chapter, only Pople's split-valence basis sets [231–238] have been investigated. Among the many reasons behind this choice, we would like to mention the following ones:

- They are very popular and they are implemented in almost every quantum chemistry package, in such a way that they are readily available for most researchers and the results here may be easily checked or extended.
- There exist a lot of data calculated using these basis sets in the literature, so that the knowledge about their behaviour in different problems is constantly growing and may also be enriched by the study presented here.
- Pople's split-valence basis sets incorporate, and hence allow to investigate, most of the features and improvements that are commonly used in the literature, such as contraction, valence-splitting, diffuse functions and polarizations.
- The number of different basis sets available is very large (see, for example the EMSL database at `http://www.emsl.pnl.gov/forms/basisform.html`), so that, for obvious computational reasons, one cannot explore them all, and some choice must be made.

Now, even restricting oneself to this particular family of basis sets, the number of variants that can be formed by independently adding each type of diffuse or polarization function to each one of the basic 6-31G and 6-311G sets is huge (to get to the sought point, there is no need to consider the addition of functions to 3-21G, 4-31G, etc.). Using that the largest set of diffuse and polarization shells that we may add is the '++G(3df,3pd)' one [234], we can express the different basis sets that may be constructed as a product of all the independent possibilities:

$$\begin{aligned}&\{\text{6-31G, 6-311G}\} \times \{\cdot, +\}_{\text{heavy}} \times \{\cdot, +\}_{\text{hydrogen}} \\ &\times \{\cdot, \text{d}, \text{2d}, \text{3d}\}_{\text{heavy}} \times \{\cdot, \text{p}, \text{2p}, \text{3p}\}_{\text{hydrogen}} \times \{\cdot, \text{f}\}_{\text{heavy}} \times \{\cdot, \text{d}\}_{\text{hydrogen}} ,\end{aligned} \tag{7.3}$$

First-stage, rules-complying basis sets (24)		
3-21G	6-31G	6-311G
3-21G(d,p)	6-31G(d,p)	6-311G(d,p)
3-21++G	6-31G(2d,2p)	6-311G(2d,2p)
3-21++G(d,p)	6-31G(2df,2pd)	6-311G(2df,2pd)
4-31G	6-31++G	6-311++G
4-31G(d,p)	6-31++G(d,p)	6-311++G(d,p)
4-31++G	6-31++G(2d,2p)	6-311++G(2d,2p)
4-31++G(d,p)	6-31++G(2df,2pd)	**6-311++G(2df,2pd)**
First violation of the rules (5)		
6-31+G(d,p)	6-31++G(d)	6-31G(f,d)
6-31 · +G(d,p)	6-31++G(· ,p)	
Second violation of the rules (10)		
4-31G(d)	6-311G(d)	6-31G(d)
4-31+G(d)	6-311+G(d)	6-31+G(d)
4-31+G(d,p)	6-311+G(d,p)	
4-31++G(d)	6-311++G(d)	

Table 7.2: Basis sets investigated in this chapter. They are organized in three groups: in the first one, we have the basis sets that comply with some heuristic restrictions commonly found in the literature; in the second group, these restrictions are broken in an exploratory manner; finally, in the third group, 10 new basis sets are selected according to what has been learned by violating the rules. The number of basis sets in each group is shown in brackets, the dot · is used to indicate that no shell of a particular type is added to the heavy atoms, and the largest basis set is written in bold face. See also fig. 7.2.

where the dot · indicates, here, that no function is added from a particular group.

Therefore, there are $2 \times 2 \times 2 \times 4 \times 4 \times 2 \times 2 = 512$ different Pople's split-valence basis sets just considering the 6-31G and 6-311G families. This number is prohibitively large to carry out a full study even at the RHF level, so that, here, the following strategy has been devised to render the investigation feasible:

To begin with, we impose several constraints on the basis sets that will be considered in a first stage:

(i) The maximum set of diffuse and polarization shells added is '++G(2df,2pd)', instead of '++G(3df,3pd)'. This is consistent with the thumb-rule concept of *balance* [177], according to which, the most efficient (*balanced*) basis sets are typically those that contain, for each angular momentum l, one shell more than the ones included for $l + 1$; so that 6-311++G(3df,3pd), for example, should be regarded as *unbalanced*.

(ii) There must be the same number and type of shells in hydrogens as in heavy atoms.

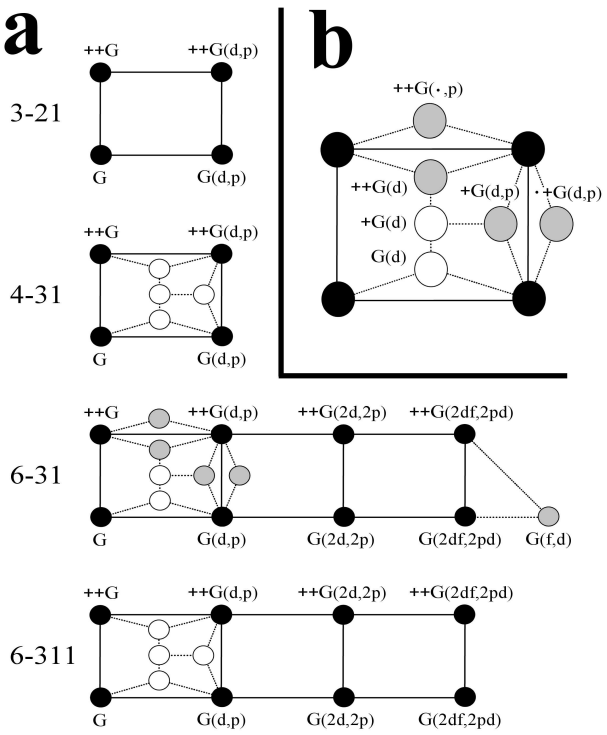


Figure 7.2: Basis sets investigated in this chapter. They are organized in three groups: in the first one, depicted as *black circles*, we have the basis sets that comply with some heuristic restrictions commonly found in the literature; in the second group, represented as *grey-filled circles*, these restrictions are broken in an exploratory manner; finally, in the third group, shown as *white-filled circles*, 10 new basis sets are selected according to what has been learned by violating the rules. In **(a)**, a general view of all the 39 basis sets is presented, while in **(b)**, the left-most region of the 6-31 family has been enlarged so that the basis sets belonging to the second and third groups could be more easily appreciated. The dot · is used to indicate that no shell of a particular type is added to the heavy atoms, and the geometric arrangement of the basis sets has no deep meaning whatsoever, it is only intended to provide visual comfort. See also table 7.2.

This has to be interpreted in the proper way: for example, if we add two d-type polarization shells to the heavy atoms, we must add two p-type ones to hydrogens. They are of the same type in the sense that they are one momentum angular unit larger than the largest one in the respective valence shell.

(iii) The higher angular momentum f-type (for heavy atoms) and d-type shells (for hydrogens), are not included unless the lower polarizations are doubly split, i.e., unless we have already included the (2d,2p)-shells. This is again consistent with the balance rule mentioned in point (i).

(iv) The investigation of smaller basis sets is restricted to the 3-21G and 4-31G families and the largest set of extra shells that is added to them is '++G(d,p)'. For consistence, the diffuse and polarization functions used for 3-21G and 4-31G are the same as the ones for 6-31G and 6-311G [233, 234, 237, 238].

These *rules*, most of which are typically obeyed (often tacitly[207, 242, 243]) in the literature [396], produce the list of 24 basis sets labeled as 'First-stage, rules-complying basis sets' in table 7.2 and depicted as black circles in fig. 7.2.

Even if their exhaustive study is already a demanding computational task and the space of all Pople's split-valence basis sets may be thought to be reasonably sampled by this 'first-stage' group, we wanted to check the validity of some of the rules, since, in

the same spirit of the arguments given in the introduction, what is good for a particular system or a particular purpose is not necessarily good for a different one, which may be far away from the native playground where the methods have been traditionally tested and parameterized. Therefore, to this end, we have chosen the medium-sized and reasonably RHF-efficient 6-31++G(d,p) basis set (see sec. 7.3.1), and we have modified it in order to break restrictions (ii) and (iii). On the one hand, as representants of breaking rule (ii), we have selected the basis sets 6-31+G(d,p), 6-31++G(d), 6-31$\cdot$+G(d,p) and 6-31++G($\cdot$,p), where, in the first two cases, a diffuse and a polarization shell has been respectively removed from the hydrogens, while, in the last two ones, the removal has been carried out on the heavy atoms. This second modification is so unusual (in fact, we have not found any work where it is performed) that there is no notation for it in the literature; herein, a dot $\cdot$ is used in the place where the unexisting heavy-atom shell would appear. On the other hand, as a representant of breaking rule (iii), we have selected 6-31G(f,d). This new group of 5 basis sets is labeled as 'First violation of the rules' in table 7.2 and depicted as grey-filled circles in fig. 7.2. We have decided to violate neither rule (i), mainly for computational reasons, nor rule (iv), due to the fact that the study of the smaller basis sets was intended to be only exploratory (and, in any case, the 3-21G and 4-31G families have proved to be rather inefficient for this problem; see sec. 7.3).

The conclusions extracted from the study of the 'first violation of the rules' group within RHF are discussed later, however, it suffices to say for the moment that we learn from them that breaking rule (iii) is not advantageous, and that one may benefit from breaking rule (ii) only if the functions are removed from the hydrogens. Therefore, in the final step of the selection of the basis sets that shall be investigated, we include a new group of 10 basis sets which come from removing diffuse and/or polarization shells from some of the most efficient ones in the other two groups. This new block is labeled as 'Second violation of the rules' in table 7.2 and depicted as white-filled circles in fig. 7.2.

The basis sets used in the RHF part of the study are those in table 7.2, whereas, in the MP2 part, we have considered the smaller subgroup that may be found in table 7.4 (see also fig. 7.2). All of them have been taken from the Gaussian03 internally stored library except for 6-31$\cdot$+G(d,p), 6-31++G($\cdot$,p) and all the basis sets extracted from the 3-21G and 4-31G ones by adding extra functions. The first two have no accepted notation and cannot be specified in the program, while the ones derived from 3-21G and 4-31G have been constructed, for consistence, using the diffuse and/or polarization shells of the 6-31G and 6-311G families. For these exceptions, the data has been taken from the EMSL repository at `http://www.emsl.pnl.gov/forms/basisform.html` (see also footnote 68 in chapter 2) and the basis sets have been read using the `Gen` keyword. In all cases, spherical GTOs have been preferred, thus having 5 d-type and 7 f-type functions per shell (see sec. 2.9).

7.3 Results

7.3.1 RHF//RHF-intramethod model chemistries

The study in this chapter begins by performing an exhaustive comparison of all the homolevel MCs and most of the heterolevel ones that can be constructed using the 39 different basis sets described above and within the RHF method. The original aim was to

identify the most efficient basis sets for doing geometry optimizations and those that perform best for single-point energy calculations, in order to extract the information needed to carry out, in successive stages, a (necessarily) more restrictive study of MP2//MP2 and MP2//RHF model chemistries. However, due to the considerations made in sec. 7.3.3, all mentions to the accuracy of any given MC in this section must be understood as relative to the RHF//RHF reference, and not to the (surely better) MP2//MP2 one or to the exact result. In this spirit, this part of the study should be regarded as an evaluation of the most efficient model chemistries for approximating *the infinite basis set Hartree-Fock limit* (for which the best RHF//RHF homolevel here is probably a good reference), and also as a way of introducing the relevant concepts and the systematic approach that shall be used in the rest of the computationally more useful sections.

Having this in mind, the *efficiency* of a particular MC is laxly defined as a balance between accuracy (in terms of the distance introduced in sec. 7.2.2) and computational cost (in terms of time). It is graphically extracted from the *efficiency plots*, where the distance d_{12} between any given model chemistry and a reference one is shown in units of RT in the x-axis, while, in the y-axis, one can find in logarithmic scale the average computational time taken for each model chemistry, per point of the 12×12 grid defined in the Ramachandran space of the model dipeptide HCO-L-Ala-NH_2 (see the following pages for several examples). As a general thumb-rule, *we shall consider a MC to be more efficient for approximating the reference when it is placed closer to the origin of coordinates in the efficiency plot*. This approach is intentionally non-rigorous due to the fact that many factors exist, such as the algorithms used, the actual details of the computers (frequency of the processor, size of the RAM and cache memories, system bus velocity, disk access velocity, operating system, libraries, etc.), the starting guesses for the SCF orbitals, the starting structures in geometry optimizations, etc., which influence the computational time but may vary from one practical calculation to another. Taking this into account, the only conclusions that shall be drawn in this chapter about the relative efficiency of the model chemistries studied are those deduced from strong signals in the plots and, therefore, those that can be extrapolated to future calculations; *the small details shall be typically neglected*.

The efficiency plots that we will discuss in this section are the ones used to compare *RHF//RHF-intramethod* homo- and heterolevel model chemistries with the RHF *reference*, defined as the homolevel MC with the largest basis set, i.e., RHF/6-311++G(2df,2pd) (since, in this section, there is no possible ambiguity, the levels shall be denoted in what follows omitting the 'RHF' keyword and specifying only the basis set). The plots corresponding to this first intramethod part comprise the figures from 7.3 to 7.6.

In fig. 7.3, the *homolevel* MCs corresponding to all the basis sets in table 7.2 are compared to the reference one. In fig. 7.3a, a general picture is presented, whereas, in fig. 7.3b, a detailed zoom of the most efficient region of the plot is shown. It takes an average of ∼ 30 hours per grid point[112] to calculate the PES of the model dipeptide HCO-L-Ala-NH_2 at the reference homolevel 6-311++G(2df,2pd); this time is denoted by t_{best} and the most efficient region is defined as that in which $d_{12} < RT$ and $t < 10\%$ of t_{best}. Additionally, we indicate in the plots the *peptide region* ($0.1RT < d_{12} \leq RT$), containing the

[112] The time per point for homolevels is calculated assuming that all geometry optimizations take 20 iterations to converge. This is done in order to avoid the ambiguity due to the choice of the starting structures and it allows to place all MCs on a more equivalent footing.

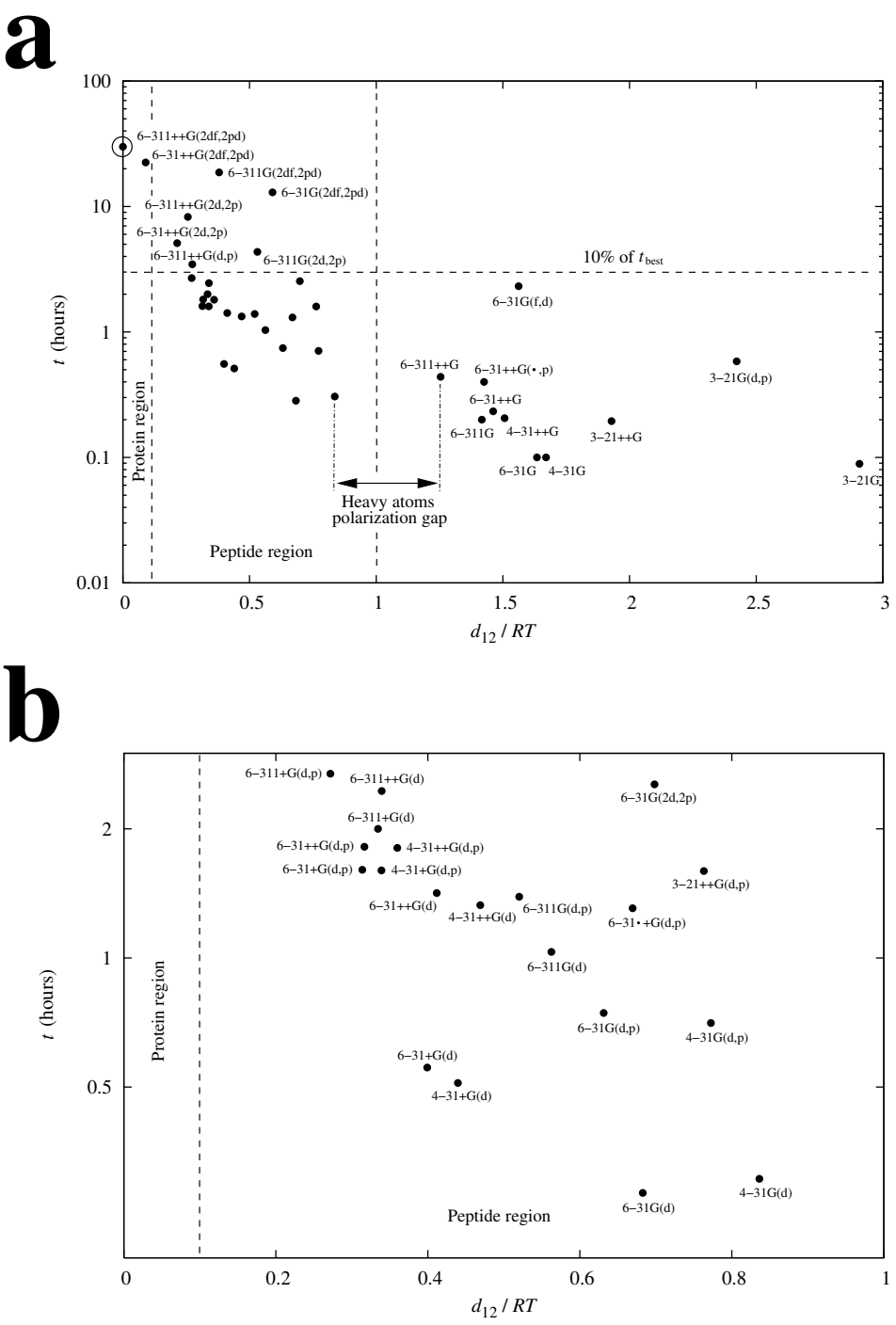


Figure 7.3: Efficiency plots of the *RHF-homolevel* model chemistries corresponding to the basis sets in table 7.2. In the x-axis, we show the distance d_{12}, in units of RT at 300º K, between any given model chemistry and the reference one (indicated by an encircled point), while, in the y-axis, we present in logarithmic scale the average computational time taken for each model chemistry, per point of the 12×12 grid defined in the Ramachandran space of the model dipeptide HCO-L-Ala-NH_2. **(a)** General view containing all basis sets. **(b)** Detailed zoom of the most efficient region of the plot ($d_{12} < RT$ and $t < 10\%$ of t_{best}).

MCs that may correctly approximate the reference one for chains of 1–100 residues, and the *protein region* ($0 < d_{12} \leq 0.1RT$), including the MCs that are accurate for polypeptides over 100 residues (see sec. 7.2).

From these two plots, several conclusions may be drawn:

- Regarding the check of rules (ii) and (iii) via the basis sets in the second group in table 7.2, we see that 6-31+G(d,p) is more efficient than the 6-31++G(d,p) one (it is cheaper and, despite being smaller, more accurate![113]), that 6-31++G(d) is as efficient as the most efficient basis sets of the rules-complying group (being outperformed only by some of the ones in the third group in table 7.2), that 6-31 $\cdot$ +G(d,p) has drifted a little towards the inefficiency region and that that the 6-31++G($\cdot$,p) is well deep in it. This suggests that *it may be profitable to break rule* (ii) *but only in the direction of removing shells from the hydrogens, and not from the heavy atoms*, which is in agreement with the common practice in the literature based on the intuition that 'hydrogens are typically more passive atoms sitting at the end of bonds' [177]. On the other hand, 6-31G(f,d) turns out to be very inefficient, being about as accurate as the simple 6-31G basis set but far more expensive. This confirms that *it is a good idea to follow restriction* (iii), which is consistent with the already mentioned thumb-rule of basis set 'balance' [177].

- *The whole* 3-21G *family of basis sets is very inefficient*. Only the 3-21++G(d,p) one lies in the accurate region and, anyway, it is less efficient than most of the other basis sets there.

- Contrarily, *the* 4-31G *family results are quite parallel to and only slightly worse than those of the* 6-31G *family*, suggesting that, to account for conformational energy differences within the RHF method, the contraction of valence orbitals is more important than the contraction of the core ones if homolevel model chemistries are used.

- In fig. 7.3a, we can notice the existence of a *gap* in the values of the distance d_{12} that lies around $d_{12} = RT$ and that separates the model chemistries in two groups. Notably, all the basis sets in the most accurate group share a common characteristic: *they contain heavy atoms polarization functions*, whereas those in the inaccurate group do not, with the only exceptions of 3-21G(d,p) and 6-31G(f,d), whose bad quality has been explained in the previous points for other reasons.

- *All the basis sets with extra polarizations*, (2d,2p) *or* (2df,2pd), *and no diffuse functions are less efficient than their diffuse functions-containing counterparts.*

- Out of some of the specially inefficient cases discussed in the preceding points, *the addition of diffuse functions to singly-polarized* ((d) or (d,p)) *basis sets always increases the accuracy*.

- *The only basis set whose homolevel MC lies in the protein region is the expensive* 6-31++G(2df,2pd).

[113] Note that the Hartree-Fock method has a variational origin (see sec. 2.7), in such a way that, if we were interested in the absolute value of the energy, and not in the energy differences, an enlargement of the basis set would always improve the results.

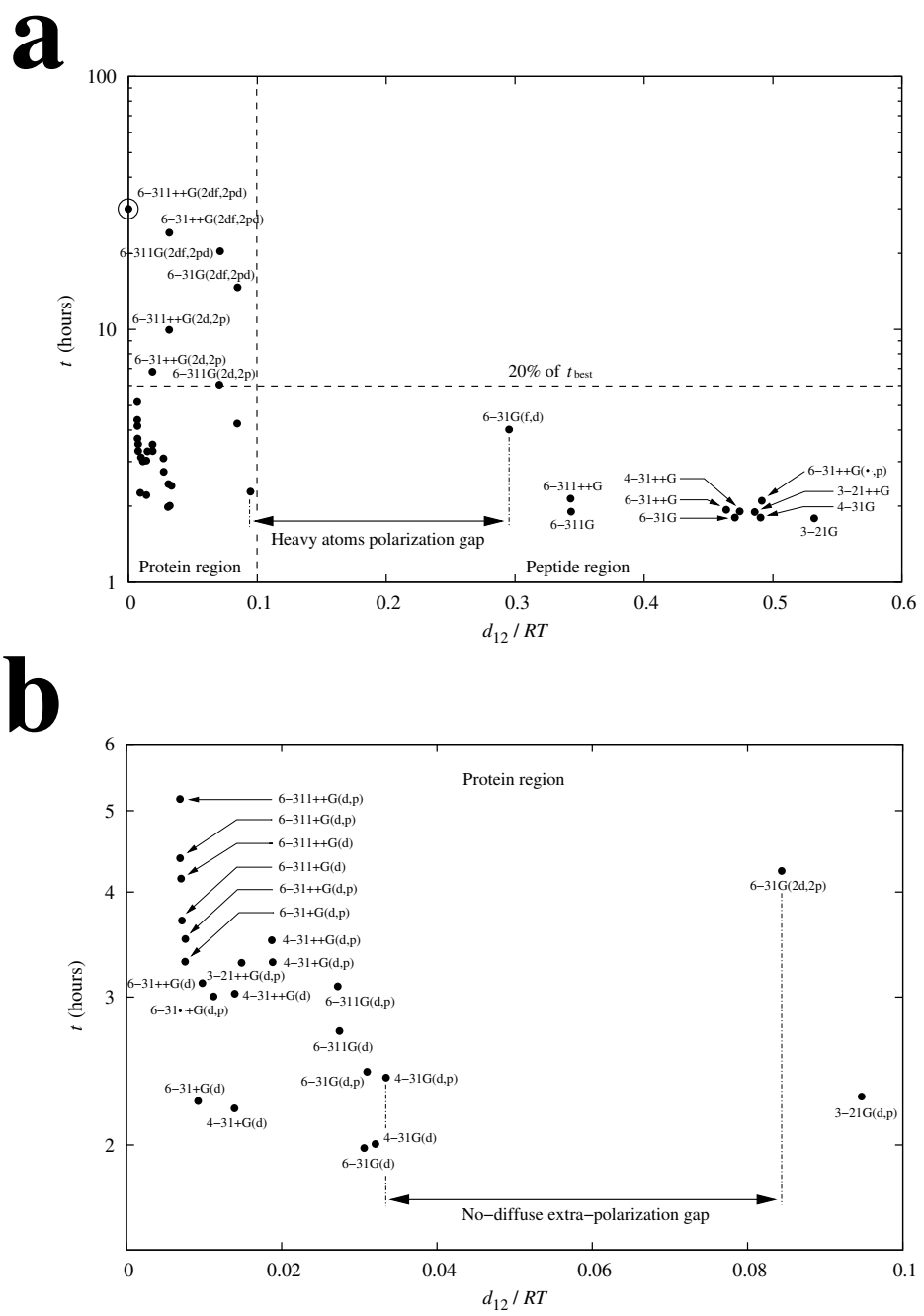


Figure 7.4: Efficiency plots of the *RHF-heterolevel* model chemistries $L_E^{\text{best}}//L_G^i$ obtained computing the geometries with all the basis sets in table 7.2 but the largest one and then performing a single-point energy calculation at the best level of the theory, L^{best} :=6-311++G(2df,2pd), on each one of them. In the x-axis, we show the distance d_{12}, in units of RT at 300$^{\text{o}}$ K, between any given model chemistry and the reference one (the *homolevel* 6-311++G(2df,2pd), indicated by an encircled point), while, in the y-axis, we present in logarithmic scale the average computational time taken for each model chemistry, per point of the 12×12 grid defined in the Ramachandran space of the model dipeptide HCO-L-Ala-NH_2. **(a)** General view containing all basis sets. **(b)** Detailed zoom of the most efficient region of the plot ($d_{12} < RT$ and $t < 20\%$ of t_{best}).

- If we look at the most efficient basis sets (those that lie at the lower-left envelope of the 'cloud' of points), we can see that *no accumulation point is reached*, i.e., that, although the distance between 6-311++G(2df,2pd) and 6-31++G(2df,2pd) is small enough to consider that we are close to the Hartree-Fock limit for this particular problem (see chapter 2), if the basis set is intelligently enlarged, we obtain increasingly better model chemistries.

- For less than 10% the cost of the reference calculation, some particularly efficient basis sets for RHF-homolevel model chemistries that can be used without altering the relevant conformational behaviour of short peptides (i.e., whose distance d_{12} with 6-311++G(2df,2pd) is less than RT) are 6-31+G(d,p), 6-31+G(d), 4-31+G(d) and 6-31G(d).

Next, in fig. 7.4, the reference homolevel 6-311++G(2df,2pd) is compared to the *RHF//RHF-intramethod-heterolevel* model chemistries $L_E^{\text{best}}//L_G^i$ obtained computing the geometries with the 38 remaining basis sets in table 7.2 and then performing a single-point energy calculation at the best level of the theory, L^{best} :=6-311++G(2df,2pd), on each one of them. The aim of this comparison is twofold: on the one hand, we want to measure the relative efficiency of the different basis sets for calculating the *geometry* (not the energy), on the other hand, we want to find out whether or not the *heterolevel assumption* described in the introduction is a good approximation within RHF.

Like in the previous case, in fig. 7.4a, a general picture is presented, whereas, in fig. 7.4b, a detailed zoom of the most efficient region of the plot is shown. The average time per point t of the heterolevel MCs has been calculated adding the average cost of performing a single-point at L^{best} :=6-311++G(2df,2pd) ($\sim$ 1.7 hours) to the average time per point needed to calculate the geometry at each one of the levels L_G^i (see footnote 112). This $\sim$ 1.7 hours 'offset' in all the times, has rendered advisable to set the limit used to define the efficient region in this case to the 20% of t_{best} (instead of the former 10%), so that most of the relevant basis sets are included in the second plot in fig. 7.4b.

In this second part of the study, several interesting conclusions may be extracted from the plots:

- About the test of rules (ii) and (iii), more or less the same remarks as before can be made, the only difference being that, in this case, for computing the geometry, the basis set 6-31$\cdot$+G(d,p) is not as bad as for the homolevel calculation. The signal, however, is rather weak and *the main conclusions stated in the first point above should not be modified*.

- Regarding the 3-21G family of basis sets, we see here that, differently from what happened for the homolevels, *they are not so bad to account for the geometry*, and, in the case of 3-21G, 3-21++G and 3-21++G(d,p), their efficiency is close to that of the corresponding 4-31G and 6-31G counterparts.

- *Moreover, the* 4-31G *basis sets performance is still quite close to that of the* 6-31G *family*. This point, together with the previous one, and differently from what happened in the case of homolevel model chemistries, suggests that, to account for the equilibrium geometry within RHF, the contraction scheme is only mildly important both for valence and core orbitals.

- In fig. 7.4a, we see again, the *gap* in the values of the distance d_{12} separating the MCs with the geometry calculated using basis sets that contain heavy atoms polarization functions from those that do not. The only differences are that, this time, the gap is even more evident, it lies around $d_{12} = 0.2RT$, and the basis set 3-21G(d,p) is placed below it.

- The signal noticed in the homolevel case regarding the relative inefficiency of the the basis sets with *extra polarizations*, (2d,2p) *or* (2df,2pd), *and no diffuse functions* has become stronger here and a second *gap* can be seen separating them from their diffuse functions-containing counterparts and also from the basis sets with only one polarization shell. This gap separates, for example, the model chemistries whose geometries have been calculated with the basis sets 6-31G(2df,2pd) and 6-31G(d,p), in such a way that the smaller one is not only more efficient, *but also more accurate*. This clearly illustrates one of the points raised in this chapter, namely, that model chemistries parameterized and tested in concrete systems may behave in an unexpected way when used in a different problem, and that the investigation of the quality of the most popular model chemistries, as well as the design of new ones, for the study of the conformational preferences of peptides, is a necessary, albeit enormous, task.

- Also for geometry optimizations, *the addition of diffuse functions to singly-polarized* ((d) or (d,p)) *basis sets increases the accuracy*.

- Contrarily to the situation for homolevels, where the only basis set that lied in the protein region was the 6-31++G(2df,2pd) one and some MCs presented distances of near $3RT$ with the reference one, here, all model chemistries lie well below $d_{12} = RT$, and those for which the geometry has been computed with a basis set that contains heavy atoms polarization functions (except for 6-31G(f,d)) are *all in the protein region*, so that they can correctly approximate the reference MC for chains of more than 100 residues. Remarkably, some of this heterolevel MCs, such as 6-311++G(2df,2pd)//6-31+G(d) for example, are physically equivalent to the reference homolevel up to peptides of *ten thousand residues* at less of 10% the computational cost. Indeed, all these results *confirm the heterolevel assumption*, discussed in the introduction and so commonly used in the literature [163, 246, 255], for RHF//RHF-intramethod model chemistries.

- Differently from the homolevel case, *an accumulation point is reached* here in the basis sets, since, in fig. 7.4b, we can see that there is no noticeable increase in accuracy, say, beyond 6-311+G(d).

- Finally, let us mention 6-311+G(d), 6-31+G(d) and 6-31G(d) as some examples of particularly efficient basis sets for calculating the geometry in RHF-heterolevel model chemistries. They can be used without altering the relevant conformational behaviour of polypeptides of more than a hundred residues (i.e., their distance d_{12} with the homolevel 6-311++G(2df,2pd) is less than $0.1RT$), and their computational cost is less than 20% that of the reference calculation.

Now, after the geometry, we shall investigate the efficiency for performing energy calculations within RHF of the all the basis sets in table 7.2 but the largest one. To

render the study meaningful, the geometry on top of which the single-points are computed must be the same, and we have chosen it to be the one calculated at the level L^{best} :=6-311++G(2df,2pd). Of course, since the reference to which the $L_E^i//L_G^{\text{best}}$ heterolevel MCs must be compared is the L^{best} homolevel, and they take more computational time than this MC (the time t_{best} plus the one required to perform the single-point at L_E^i), *all of them are computationally inefficient a priori*. Therefore, in the efficiency plots in fig. 7.5, the time shown in the y-axis is not the one needed to calculate the actual PES with the $L_E^i//L_G^{\text{best}}$ model chemistry, but just the one required for the single-point computation. In principle therefore, the study and the conclusions drawn should be regarded only as providing *hints* about how efficient a given basis set will be if it is used to calculate the energy on top of some less demanding geometry than the L^{best} one (in order to have a model chemistry that could have some possibility of being efficient). However, in the fourth part of the RHF//RHF-intramethod investigation (see below), we show that the performance of the different basis sets for single-point calculations depends weakly on the underlying geometry, so that the range of validity of the present part of study must be thought to be wider.

In fig. 7.5a, a general picture of the comparison is presented, whereas, in fig. 7.5b, a detailed zoom of the most efficient region of the plot is shown. As we have already mentioned, the time t shown is the average one per point required to perform the single-point energy calculation on the best geometry, and, consequently, the time t_{best} used for defining the efficient region has been redefined as the one needed for a single-point at L^{best} (i.e., $t_{\text{best}} \simeq 1.7$ hours) .

We extract the following conclusions from the plots:

- Regarding the check of rules (ii) and (iii), *the situation is the same as in the two former cases*, with the only difference that we can see that, for single-point calculations, the basis set 6-31G(f,d) is much more inefficient than for geometry optimizations, being of an accuracy close to that of the smallest basis set studied, the 3-21G, and taking considerably more time.

- *The* 3-21G *family of basis sets is very inefficient for energy calculations*.

- On the other hand, like it happened in the homolevels case, *the* 4-31G *basis sets performance is quite close to that of the* 6-31G *family*. This suggests that, for energy calculations in RHF//RHF model chemistries, to use a considerable number of primitive Gaussian shells to form the contracted ones, is more important in the valence orbitals than in the core ones.

- *The heavy atoms polarization gap in the distance* d_{12} *also occurs for single-point calculations* (see fig. 7.5a). This time, the basis set 3-21G(d,p) is placed above it.

- *The relative inefficiency of the the basis sets with extra polarizations*, (2d,2p) *or* (2df,2pd), *and no diffuse functions is also observed here for energy calculations*. It is mild, like in the homolevels case, and no gap appears.

- Like in the two studies above, *the addition of diffuse functions to singly-polarized* ((d) or (d,p)) *basis sets increases the accuracy* for single-point energy calculations as well.

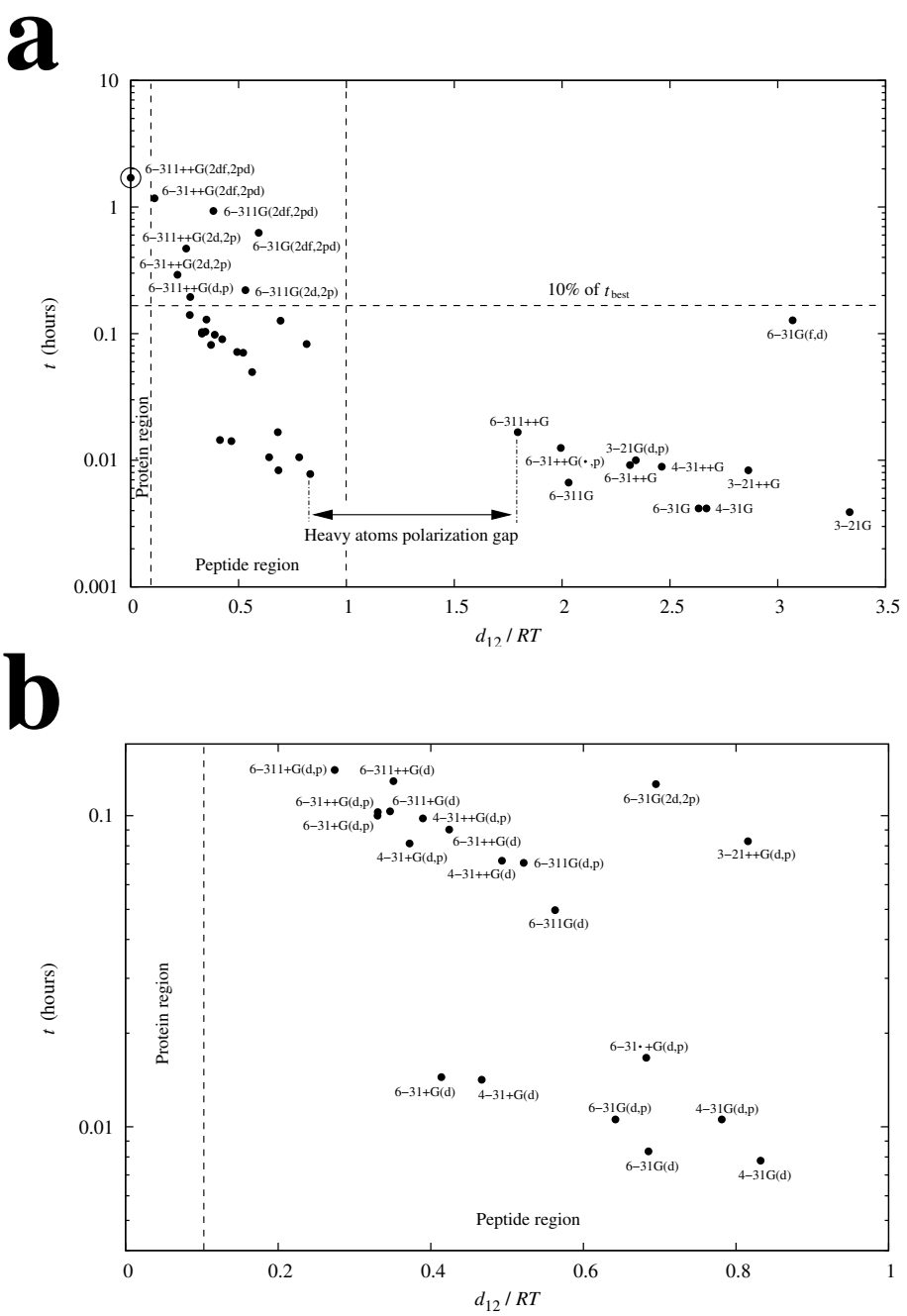


Figure 7.5: Efficiency plots of the *RHF-heterolevel* model chemistries $L_E^i // L_G^{\text{best}}$ obtained computing the geometry at the best level of the theory, L^{best} :=6-311++G(2df,2pd), and then performing a single-energy calculation with all the basis sets in table 7.2 but the largest one. In the x-axis, we show the distance d_{12}, in units of RT at 300$^{\text{o}}$ K, between any given model chemistry and the reference one (the *homolevel* 6-311++G(2df,2pd), indicated by an encircled point), while, in the y-axis, we present in logarithmic scale the average computational time taken for the corresponding single-point, per point of the 12×12 grid defined in the Ramachandran space of the model dipeptide HCO-L-Ala-NH_2. **(a)** General view containing all basis sets. **(b)** Detailed zoom of the most efficient region of the plot ($d_{12} < RT$ and $t < 10\%$ of t_{best}).

- Regarding the accuracy of the investigated MCs, the situation seen in the homolevel case is even worse here, since not even the 6-31++G(2df,2pd) single-point MC lies in the protein region and the worst basis sets (see 3-21G, for example) present distances over $3RT$. This enriches and supports the ideas that underlie the *heterolevel assumption*, showing that, *whereas the level of the theory may be lowered in the calculation of the (constrained) equilibrium geometries, it is necessary to perform high-level energy single-points if a good accuracy is sought*.

- Regarding the basis set convergence issue, the situation here is analogous to the one seen in the case of homolevel MCs: *No accumulation point is reached*, and the accuracy can always be increased by intelligently enlarging the basis set.

- Finally, let us mention 6-311+G(d,p), 6-31+G(d) and 6-31G(d) as some examples of particularly efficient basis sets also for calculating the energy in RHF-heterolevel model chemistries. They can be used without altering the relevant conformational behaviour of short peptides, and their computational cost is less than 10% that of the reference single-point calculation.

Next, in order to close the RHF//RHF-intramethod section, we evaluate a group of heterolevel model chemistries which are constructed by simultaneously decreasing the level of the theory used for the geometry and the one used for the energy single-point, relatively to the reference 6-311++G(2df,2pd).

Using the basis sets in table 7.2, there exist $38 \times (38 - 1) = 1406$ different model chemistries of the form $L_E^i//L_G^i$, with $L_E^i \neq L_G^i$ and excluding 6-311++G(2df,2pd). This number is too large to perform an exhaustive study and, therefore, any investigation of the MCs in this particular group must be necessarily exploratory. Here, we are specially interested in the most efficient MCs, so that, using the lessons learned in the preceding paragraphs, we have considered only heterolevels with L_G^i being 6-31G(d), 6-31+G(d) or 6-311+G(d), which we have proved to perform well at least when the single-point is calculated at 6-311++G(2df,2pd). For L_E^i, different criteria have been followed. On the one hand, since the energy at level L_G^i is readily available as an output of the geometry optimization step, it is clear that to perform a single-point calculation with a level of similar accuracy to L_G^i will not pay. On the other hand, some hints may be extracted from the study in fig. 7.5 about which could be the most efficient basis sets for calculating the energy. Taking these two points into consideration, and also including, for checking purposes, some levels that are expected to be inefficient, the basis sets that are investigated for performing single-points within RHF are those shown in fig. 7.6a, where t_{best} is again the time taken by the reference homolevel 6-311++G(2df,2pd).

There are no essentially new conclusions to extract from this part of the study, since it mainly confirms those drawn from the previous parts and shows that they can be combined rather independently. For example, the approximate verticality of the dotted lines joining the MCs with equal L_E^i indicates, as we have already mentioned, that, in the RHF//RHF case, *the accuracy of a given model chemistry depends much more strongly on the level used for calculating the energy than on the one used for the geometry*. Also, the fact that the MCs with L_G^i =6-31G(d) lie in the lower-left envelope of the plot shows that the 6-31G(d) keeps its character of efficient basis set for computing the geometry even if the single-point is calculated with levels that are different from the reference one. Finally,

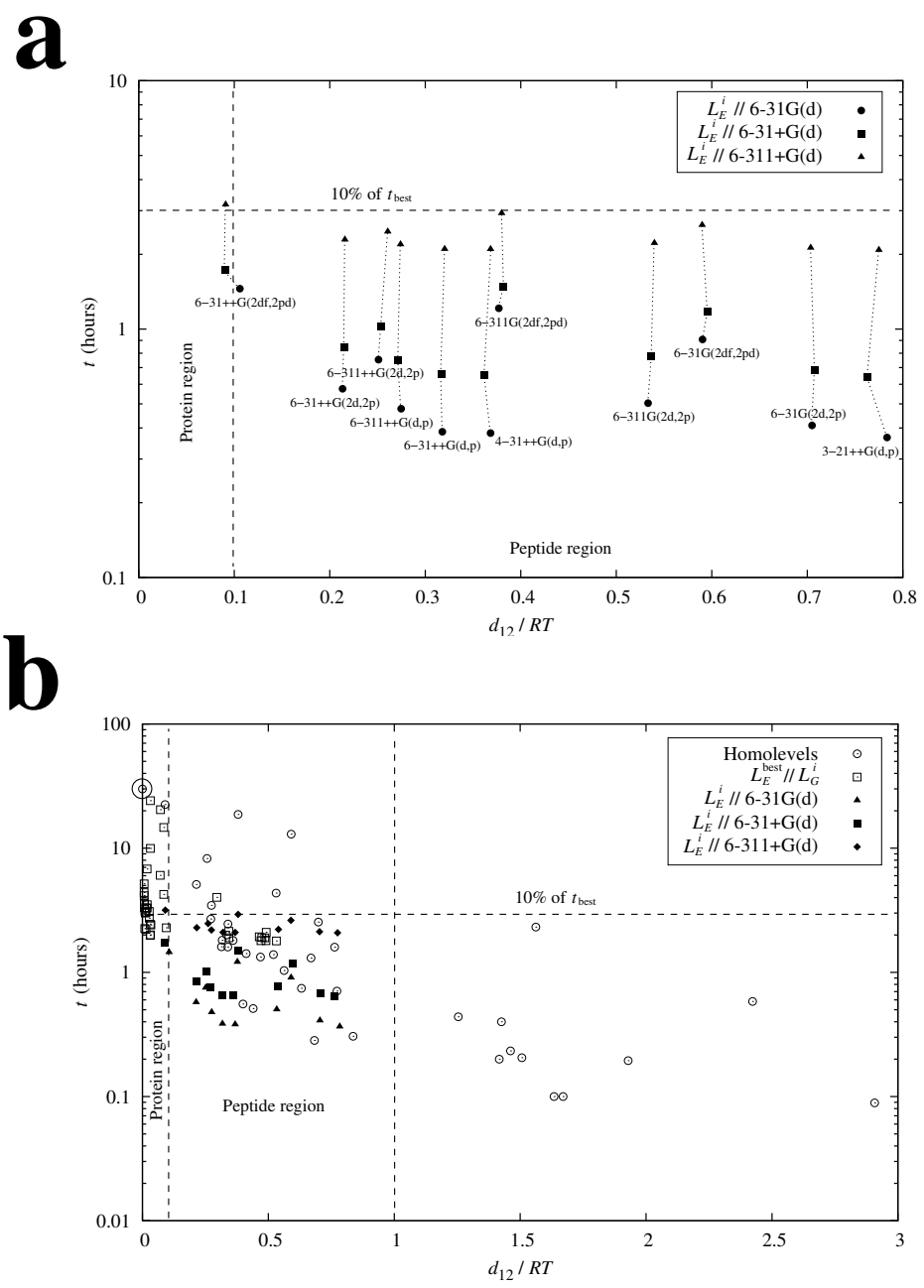


Figure 7.6: **(a)** Efficiency plot of some selected *RHF-heterolevel* model chemistries $L_E^i//L_G^i$ with $L_E^i \neq L_G^i$ and both of them different from the best level 6-311++G(2df,2pd). The MCs calculated on top of the same geometry are joined by broken lines. **(b)** Efficiency plot of all the model chemistries in figs. 7.3, 7.4 and in the (a)-part of this figure. In both figures, in the x-axis, we show the distance d_{12}, in units of RT at 300º K, between any given model chemistry and the reference one (the *homolevel* 6-311++G(2df,2pd), indicated by an encircled point in (b)), while, in the y-axis, we present in logarithmic scale the average computational time per point of the 12×12 grid defined in the Ramachandran space of the model dipeptide HCO-L-Ala-NH_2. The different accuracy regions depending on d_{12}, are labeled, and the 10% of the time t_{best} taken by the reference homolevel 6-311++G(2df,2pd) is also indicated.

Efficient RHF//RHF model chemistries	d_{12}/RT [a]	N_{res} [b]	t (% of t_{best}) [c]
6-311++G(2df,2pd)//6-31++G(d,p)	0.008	17382.5	11.74%
6-311++G(2df,2pd)//6-31+G(d)	0.009	11752.4	7.53%
6-311++G(2df,2pd)//4-31+G(d)	0.014	5163.4	7.38%
6-311++G(2df,2pd)//6-31G(d)	0.031	1066.0	6.62%
6-31++G(2df,2pd)//6-31G(d)	0.106	89.3	4.86%
6-31++G(2d,2p)//6-31G(d)	0.213	22.0	1.92%
6-311++G(d,p)//6-31G(d)	0.275	13.2	1.60%
6-31++G(d,p)//6-31G(d)	0.318	9.9	1.29%
4-31++G(d,p)//6-31G(d)	0.368	7.4	1.27%
6-31G(d)//6-31G(d)	0.683	2.1	0.95%
6-31G//6-31G	1.634	0.4	0.33%
3-21G//3-21G	2.908	0.1	0.30%

Table 7.3: List of the most efficient RHF//RHF-intramethod model chemistries located at the lower-left envelope of the cloud of points in fig. 7.6b. The first block contains MCs of the form $L_E^{best}//L_G^i$ (see fig. 7.4), the second one those of the form $L_E^i//L_G^i$ (see fig. 7.6a), and the third one the homolevels in fig. 7.3. [a]Distance with the reference MC (the homolevel 6-311++G(2df,2pd)), in units of RT at 300º K. [b]Maximum number of residues in a polypeptide potential up to which the corresponding model chemistry may correctly approximate the reference. [c]Required computational time, expressed as a fraction of t_{best}.

note that, in this particular problem, and within RHF, if one wants to correctly approximate the reference MC beyond 100-residue peptides, the energy must be calculated at 6-31++G(2df,2pd).

In fig. 7.6b, all the 110 model chemistries studied up to now are depicted as a summary (the 38 inefficient $L_E^i//L_G^{best}$ ones are not shown). Now, if we look at the lower-left envelope of the plot, we can see that, depending on the target accuracy sought, the most efficient model chemistries may belong to different groups among the ones investigated above. From $\sim 0RT$ to $\sim 0.1RT$, for example, the most efficient MCs are the $L_E^{best}//L_G^i$ ones; from $\sim 0.1RT$ to $\sim 0.5RT$, on the other hand, the model chemistries of the form $L_E^i//L_G^i$, where the single-point level has also been lowered with respect to the reference one, clearly outperform those in the rest of groups; finally, for distances $d_{12} > 0.5RT$, it is recommendable to use homolevel model chemistries. In table 7.3, these efficient model chemistries are shown together with their distance d_{12} to the reference homolevel 6-311++G(2df,2pd), the number of residues N_{res} up to which they can be used as a good approximation of it, and the required computational time t, expressed as a fraction of t_{best}.

7.3.2 MP2//MP2-intramethod model chemistries

Now, using all the information gathered in the previous RHF//RHF-intramethod section (see however sec. 7.3.3 and the first paragraph of sec. 7.3.1), we open the second part of

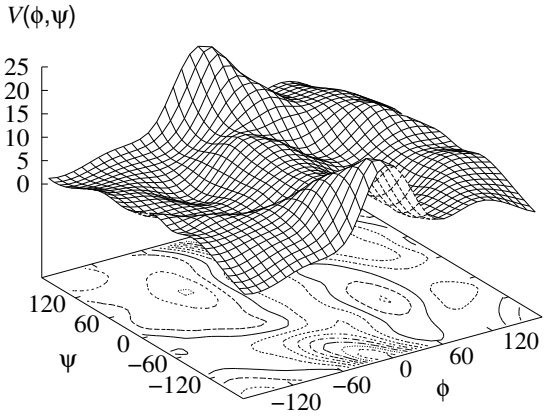


Figure 7.7: Potential energy surface of the model dipeptide HCO-L-Ala-NH_2 computed at the MP2/6-311++G(2df,2pd) level of the theory. The PES has been originally calculated in a 12×12 discrete grid in the space spanned by the Ramachandran angles ϕ and ψ and later smoothed with bicubic splines for visual convenience.

the study, in which we shall perform an *MP2//MP2-intramethod* investigation with some selected basis sets among those in table 7.2. The choice of MP2 [250] as the method immediately 'above' RHF is justified by several reasons. In the first place, it is typically regarded as accurate and as the reasonable starting point to include correlation in the literature [166, 241, 244, 246, 251, 252], where it is also commonly used as a reference calculation to evaluate or parameterize less demanding methods [163, 164, 253–255]. Secondly, and contrarily to DFT, MP2 is a wavefunction-based method that allows to more or less systematically improve the calculations by going to higher orders of the Møller-Plesset perturbation expansion (see sec. 2.10). The majority of the rest of methods devised to add correlation to the RHF wavefunction-based results, such as coupled cluster, configuration interaction, or MCSCF, are more computationally demanding than MP2 [177, 248, 249]. Finally, although, for some particular problems, DFT may rival MP2 [244, 407, 408], the latter is known to account better for weak dispersion forces, which are present and may be important in peptides [246, 247].

The basis sets investigated in this MP2 part are the 11 ones in table 7.4 and they have been originally chosen in order to adequately sample the larger set studied at RHF and check if the same effects are observed at MP2. Some kind of selection must be done due to the higher computational cost of MP2 calculations (see sec. 2.10), so that, with the hope that the RHF results were relatively transferable to MP2, the basis sets that have proved to be relatively more efficient at RHF were included in table 7.4, together with the largest one, the smallest one and a small number of other basis sets (such as 6-31G(d,p) or 6-31G(2d,2p), for example) intended to analyze the tendencies observed in the pre-

3-21G	6-31G(d)	6-31+G(d)	6-311+G(d)
6-31G	6-31G(d,p)	6-31++G(d,p)	**6-311++G(2df,2pd)**
6-31++G	6-31G(2d,2p)	6-31++G(2d,2p)	

Table 7.4: Basis sets investigated in the *MP2//MP2-intramethod* part of the study. The largest one is indicated in bold face.

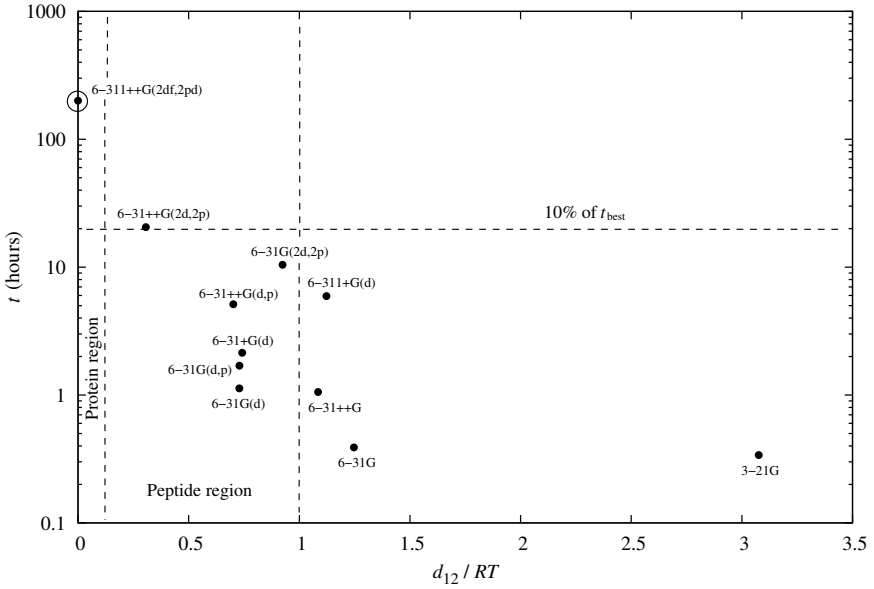


Figure 7.8: Efficiency plot of the *MP2-homolevel* model chemistries corresponding to all the basis sets in table 7.4. In the x-axis, we show the distance d_{12} in units of RT, at 300^{o} K, between any given model chemistry and the reference one (the *homolevel* 6-311++G(2df,2pd), indicated by an encircled point), while, in the y-axis, we present in logarithmic scale the average computational time taken for each model chemistry, per point of the 12×12 grid defined in the Ramachandran space of the model dipeptide HCO-L-Ala-NH_2.

vious section. In the following discussion and in sec. 7.3.3, however, the RHF → MP2 transferability of the results is shown to be imperfect, so that, despite the valuable lessons learned in this chapter, in further studies, one of the research directions that will have to be followed is the addition of more basis sets to table 7.4.

We would also like to stress that the MP2-reference PES of formyl-L-alanine-amide, with the 6-311++G(2df,2pd) basis set, that has been calculated to carry out the investigation presented here is, as far as we are aware, *the one computed at the highest level of the theory at present*. Although coupled cluster methods have been used to perform single-points on top of the geometries optimized at lower levels for some selected conformers [207], the highest levels used to calculate full PESs in the literature after the one used in this study seem to be MP2/6-311G(d,p) in ref. 210 and B3LYP/6-311++G(d,p) in ref. 207 (assuming that the accuracy of the B3LYP method lies somewhere between RHF and MP2). The MP2/6-311++G(2df,2pd) PES computed for this work is shown in fig. 7.7, where the energy reference has been set to zero[114] and the surface has been smoothed using bicubic splines for visual convenience.

Now, the structure of the MP2//MP2-intramethod study is the same as in the RHF//RHF case, so that we begin by evaluating the *MP2 homolevels*, and, just as we did before, the 'MP2' keyword is omitted from the MCs specification, since, in this section, no possible

[114] At this level of the theory, the absolute energy of the minimum point in the 12×12 grid (located at $(-75^o, 75^o)$) is −416.4705201527 hartree

ambiguity may appear.

In fig. 7.8, the *homolevel* MCs corresponding to all the basis sets in table 7.4 are compared to the reference one. It takes an average of $\sim$ 200 hours $\simeq$ 8 days of CPU time per grid point (see footnote 112) to calculate the PES of the model dipeptide HCO-L-Ala-NH_2 at the reference homolevel 6-311++G(2df,2pd); this time is denoted by t_{best}.

Regarding the conclusions that can be extracted from this plot, let us focus, stating the differences, on the issues parallel to the ones studied in the RHF case, although, since the number of basis sets in table 7.4 is smaller than that in table 7.2, some details will have to be left out:

- *The basis set* 3-21G *is again the worst one for homolevel calculations*, with a distance close to 3 *RT*.

- The heavy atoms polarization gap that we saw in fig. 7.3a, *is absent here*, and, for example, the 6-31++G basis set is more accurate than the larger and polarized 6-311+G(d).

- *The only basis set with extra polarizations and no diffuse functions that we have studied in the MP2 case, the* 6-31G(2d,2p) *one, is less efficient than its diffuse functions-containing counterpart, the* 6-31++G(2d,2p) *one*.

- Whereas, in the RHF case, the addition of diffuse functions to singly-polarized ((d) or (d,p)) basis sets always increased the accuracy, here, it is sometimes slightly advantageous (in the 6-31G(d,p) $\rightarrow$ 6-31++G(d,p) case) and sometimes slightly disadvantageous (in the 6-31G(d) $\rightarrow$ 6-31+G(d) case). So that no clear conclusion may be drawn to this respect.

- *There is no basis set whose homolevel MC lies in the protein region*, although we remark that the second largest basis set studied with RHF, the 6-31++G(2df,2pd) one, which lied in the protein region then, has not been included in this MP2 part of the work.

- If we look at the most efficient basis sets (those that lie at the lower-left envelope of the 'cloud' of points), we can see that, like in RHF, *no accumulation point is reached*, i.e., that, although the distance between 6-311++G(2df,2pd) and 6-31++G(2d,2p) is small enough to consider that we are close to the MP2 limit for this particular problem (see chapter 2), if the basis set is intelligently enlarged, we obtain increasingly better model chemistries. Also note that, if we compare fig. 7.8 here to fig. 7.3 in the previous section, we do not observe a strong signal indicating the slower basis set convergence of the MP2 method that is commonly mentioned in the literature [177]. Therefore, from these limited data, we must conclude that, *for conformational energy differences in peptides, the homolevel model chemistries converge approximately at the same pace towards the infinite basis set limit for RHF and MP2*.

- For less than 10% the cost of the reference calculation, some particularly efficient basis sets for MP2-homolevel model chemistries that can be used without altering the relevant conformational behaviour of short peptides (i.e., whose distance

d_{12} with 6-311++G(2df,2pd) is less than RT) are 6-31++G(d,p), 6-31G(d,p) and 6-31G(d).

Next, in fig. 7.9, the reference homolevel 6-311++G(2df,2pd) is compared to the *MP2//MP2-intramethod-heterolevel* model chemistries $L_E^{\text{best}}//L_G^i$ obtained computing the geometries with the 10 remaining basis sets in table 7.4 and then performing a single-point energy calculation at the best level of the theory, L^{best} :=6-311++G(2df,2pd), on each one of them. Like in the RHF case, the aim of this comparison is twofold: on the one hand, we want to measure the relative efficiency of the different basis sets for calculating the *geometry* (not the energy), on the other hand, we want to find out whether or not the *heterolevel assumption* described in the introduction is a good approximation within MP2.

The average time per point t of the heterolevel MCs has been calculated adding the average cost of performing a single-point at L^{best} :=6-311++G(2df,2pd) ($\sim$ 2.7 hours) to the average time per point needed to calculate the geometry at each one of the levels L_G^i (see footnote 112).

The following remarks may be made about fig. 7.9:

- Although the only representant of the 3-21G family of basis sets in this MP2//MP2-intramethod study is one of the most inaccurate levels for calculating the geometry, the signal observed in the RHF case, indicating that the 3-21G basis sets are not so bad to account for the geometry, *also occurs here*, where we can see that 3-21G is more accurate (and hence more efficient) than the larger 6-31G and 6-31++G basis sets.

- Contrarily to the homolevel case, here we can appreciate, like we did in RHF, a rather wide *gap* in the values of the distance d_{12} separating the MCs with the geometry calculated using basis sets that contain heavy atoms polarization functions from those that do not.

- The signal noticed in the homolevel case regarding the relative inefficiency of the the basis sets with extra polarizations and no diffuse functions *has been inverted here*, since the 6-31++G(2d,2p) is less accurate than the smaller 6-31G(2d,2p) one.

- Again, and contrarily to the RHF case, the addition of diffuse functions to singly-polarized ((d) or (d,p)) basis sets it is sometimes slightly advantageous (in the 6-31G(d,p) $\rightarrow$ 6-31++G(d,p) case) and sometimes slightly disadvantageous (in the 6-31G(d) $\rightarrow$ 6-31+G(d) case). So that no clear conclusion may be drawn to this respect.

- Like in the RHF case, and contrarily to the situation for MP2 homolevels, where no basis sets lied in the protein region and some MCs presented distances of near $3RT$ with the reference one, here, most model chemistries lie well below $d_{12} = RT$, and those for which the geometry has been computed with a basis set that contains heavy atoms polarization functions are *all in the protein region*, so that they can correctly approximate the reference MC for chains of more than 100 residues. Remarkably, some of this heterolevel MCs, such as 6-311++G(2df,2pd)//6-31G(d) for example, are physically equivalent to the reference homolevel up to peptides of 400 residues at less of 10% the computational cost. Indeed, all these results *confirm the*

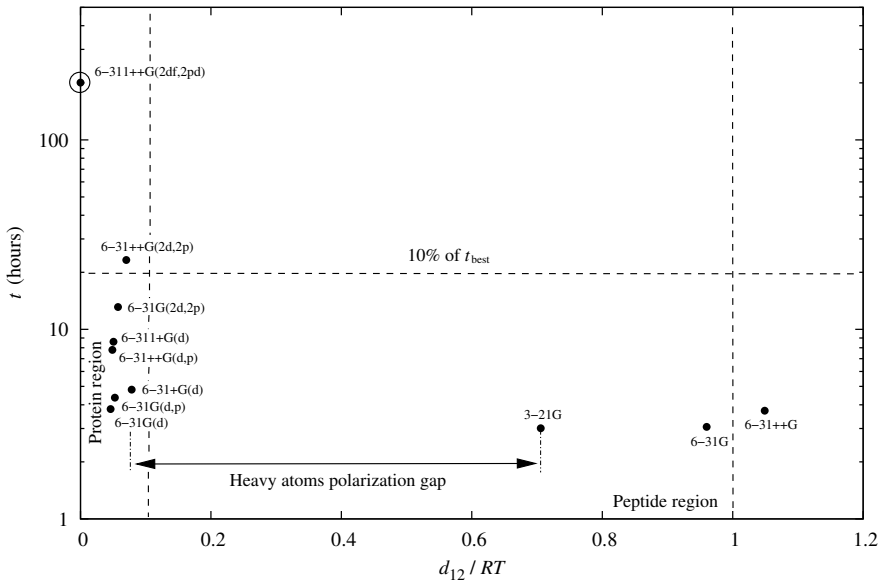


Figure 7.9: Efficiency plot of the *MP2-heterolevel* model chemistries $L_E^{\text{best}}//L_G^i$ obtained computing the geometries with all the basis sets in table 7.4 but the largest one and then performing a single-point energy calculation at the best level of the theory, L^{best} :=6-311++G(2df,2pd), on each one of them. In the x-axis, we show the distance d_{12}, in units of RT at 300$^{\text{o}}$ K, between any given model chemistry and the reference one (the *homolevel* 6-311++G(2df,2pd), indicated by an encircled point), while, in the y-axis, we present in logarithmic scale the average computational time taken for each model chemistry, per point of the 12×12 grid defined in the Ramachandran space of the model dipeptide HCO-L-Ala-NH_2.

heterolevel assumption, discussed in the introduction and so commonly used in the literature [163, 246, 255], for MP2//MP2-intramethod model chemistries.

- Differently from the homolevel case, *an accumulation point is reached* here in the basis sets, since we can see that there is no noticeable increase in accuracy beyond 6-31G(d). Regarding the convergence towards the infinite basis set limit, we observe again that, whereas it is slightly slower here than in fig. 7.4, the signal is too weak to conclude anything and we repeat what we said in the homolevel case: that, *for conformational energy differences in peptides, the ability of accounting for the geometry in heterolevel model chemistries of the form $L_E^{\text{best}}//L_G^i$ converge approximately at the same pace towards the infinite basis set limit for RHF and MP2*.

- Finally, let us mention 6-31G(d) as the only clear example of a particularly efficient basis set for calculating the geometry in MP2-heterolevel model chemistries. It can be used without altering the relevant conformational behaviour of polypeptides of around 400 residues (i.e., its distance d_{12} with the homolevel 6-311++G(2df,2pd) is $\sim 0.05RT$), and its computational cost is $\sim$ 2% that of the reference calculation. The rest of the basis sets in fig. 7.9 are either less accurate and not significantly cheaper, or more expensive and not more accurate.

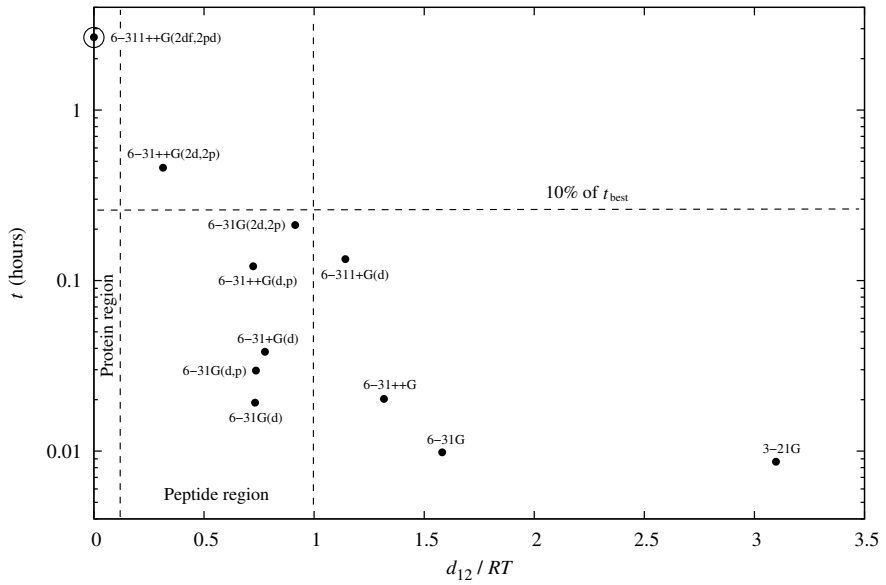


Figure 7.10: Efficiency plot of the *MP2-heterolevel* model chemistries $L_E^i//L_G^{\text{best}}$ obtained computing the geometry at the best level of the theory, L^{best} :=6-311++G(2df,2pd), and then performing a single-energy calculation with all the basis sets in table 7.2 but the largest one. In the x-axis, we show the distance d_{12}, in units of RT at 300$^{\text{o}}$ K, between any given model chemistry and the reference one (the *homolevel* 6-311++G(2df,2pd), indicated by an encircled point), while, in the y-axis, we present in logarithmic scale the average computational time taken for the corresponding single-point, per point of the 12×12 grid defined in the Ramachandran space of the model dipeptide HCO-L-Ala-NH_2.

Now, after the geometry, we shall investigate the efficiency for performing energy calculations of all the basis sets in table 7.4 but the largest one. To render the study meaningful, the geometry on top of which the single-points are computed must be the same, and we have chosen it to be the one calculated at the level L^{best} :=6-311++G(2df,2pd), like in the RHF case. Of course, since the reference to which the $L_E^i//L_G^{\text{best}}$ heterolevel MCs must be compared is the L^{best} homolevel, and they take more computational time than this MC (the time t_{best} plus the one required to perform the single-point at L_E^i), *all of them are computationally inefficient a priori*. Therefore, in the efficiency plot in fig. 7.10, the time shown in the y-axis is not the one needed to calculate the actual PES with the $L_E^i//L_G^{\text{best}}$ model chemistry, but just the one required for the single-point computation. In principle therefore, the study and the conclusions drawn should be regarded only as providing *hints* about how efficient a given basis set will be if it is used to calculate the energy on top of some less demanding geometry than the L^{best} one (in order to have a model chemistry that could have some possibility of being efficient). However, in the fourth part of the MP2//MP2-intramethod investigation (see below), we show, like we did in the RHF case, that the performance of the different basis sets for single-point calculations depends weakly on the underlying geometry, so that the range of validity of the present part of study must be thought to be wider. Again, the time t_{best} used for defining the efficient

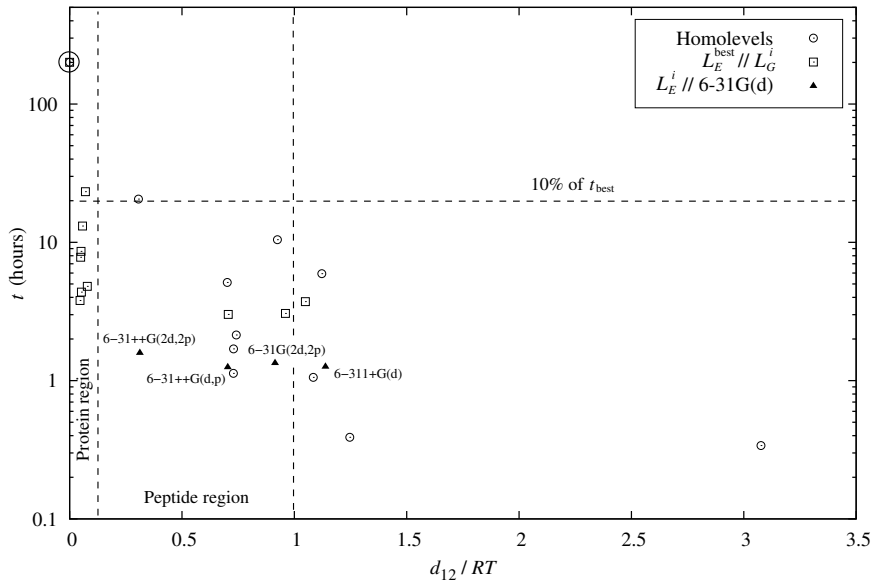


Figure 7.11: Efficiency plot of all the model chemistries in figs. 7.8, 7.9 and also of four additional ones of the form L_E^i//6-31G(d). Only the latter are labeled. In the x-axis, we show the distance d_{12}, in units of RT at 300º K, between any given model chemistry and the reference one (the *homolevel* 6-311++G(2df,2pd), indicated by an encircled point), while, in the y-axis, we present in logarithmic scale the average computational time per point of the 12×12 grid defined in the Ramachandran space of the model dipeptide HCO-L-Ala-NH_2. The different accuracy regions depending on d_{12}, are labeled, and the 10% of the time t_{best} taken by the reference homolevel 6-311++G(2df,2pd) is also indicated.

region in fig. 7.10 has been redefined as the one needed for a single-point at L^{best}.

The conclusions of this part of the study are:

- Like in RHF, *the* 3-21G *basis set is very inefficient for energy calculations*.
- Although all basis sets containing heavy atoms polarization functions are more accurate than the ones that do not, differently from the geometry case, *we do not observe a clear gap in the distance* d_{12} *separating the two groups* for MP2 single-point energy calculations.
- Like in the MP2-homolevel case and like in RHF, the respective positions in the plot of 6-31G(2d,2p) and 6-31++G(2d,2p) constitute a signal that indicates that *the basis sets with extra polarizations and no diffuse functions are less efficient than their diffuse functions-containing counterparts for energy calculations*.
- Like in the rest of the MP2 study, nothing conclusive can be said about the addition of diffuse functions to the singly polarized 6-31G(d) and 6-31G(d,p) basis sets.
- Regarding the accuracy of the investigated MCs, the situation here is analogous to the one found for RHF, and, again this supports the ideas that underlie the *heterolevel assumption*, showing that, *also at MP2, whereas the level of the theory*

Efficient MP2//MP2 model chemistries	d_{12}/RT [a]	N_{res} [b]	t (% of t_{best}) [c]
6-311++G(2df,2pd)//6-31G(d)	0.046	468.3	1.90%
6-31++G(2d,2p)//6-31G(d)	0.312	10.2	0.79%
6-31++G(d,p)//6-31G(d)	0.703	2.0	0.62%
6-31G(d)//6-31G(d)	0.729	1.9	0.56%
6-31G//6-31G	1.247	0.6	0.19%
3-21G//3-21G	3.076	0.1	0.17%

Table 7.5: List of the most efficient MP2//MP2-intramethod model chemistries located at the lower-left envelope of the cloud of points in fig. 7.11. The first block contains MCs of the form $L_E^{\text{best}}//L_G^i$ (see fig. 7.9), the second one those of the form L_E^i//6-31G(d) (see fig. 7.11), and the third one the homolevels in fig. 7.8. [a]Distance with the reference MC (the homolevel 6-311++G(2df,2pd)), in units of RT at 300° K. [b]Maximum number of residues in a polypeptide potential up to which the corresponding model chemistry may correctly approximate the reference. [c]Required computational time, expressed as a fraction of t_{best}.

may be lowered in the calculation of the (constrained) equilibrium geometries, it is necessary to perform high-level energy single-points if a good accuracy is sought.

- Regarding the basis set convergence issue, the situation here is analogous to the one seen in the case of homolevel MCs: *No accumulation point is reached*, and the accuracy can always be increased by intelligently enlarging the basis set. The convergence velocity towards the MP2 limit is again very similar to the one in RHF.

- Finally, let us mention 6-31G(d,p) and 6-31G(d) as some examples of particularly efficient basis sets for calculating the energy in MP2-heterolevel model chemistries. They can be used without altering the relevant conformational behaviour of short peptides, and their computational cost is less than 10% that of the reference single-point calculation.

To close the MP2//MP2-intramethod section, we have calculated four PESs with model chemistries of the form L_E^i//6-31G(d), since the geometry computed at the 6-31G(d) has proved to be very accurate when a single-point at the highest level was performed on top of it. Due to the same computational arguments presented in the previous section, only those basis sets significantly larger than 6-31G(d) have been used to calculate the energy. The results are presented in fig. 7.11 together with a summary of the rest of the MP2//MP2 model chemistries studied in this section (except for the inefficient $L_E^i//L_G^{\text{best}}$ ones).

We have already advanced a conclusion that may be extracted from this last plot, namely, that if we compare the distance d_{12} of the L_E^i//6-31G(d) model chemistries in fig. 7.11 to the distance of the $L_E^i//L_G^{\text{best}}$ ones in fig. 7.10 for the same L_E^i, we see that they are very close. Therefore, like in the RHF case, we conclude that *the accuracy of a given model chemistry depends much more strongly on the level used for calculating the energy than on the one used for the geometry.*

Finally, in table 7.5, we present the most efficient MCs that lie at the lower-left envelope of the plot in fig. 7.11. Like in RHF, we can see that, depending on the target accuracy sought, these most efficient model chemistries may belong to different groups among the ones investigated above. From $\sim 0RT$ to $\sim 0.1RT$, for example, the most efficient MCs are the $L_E^{\text{best}}//L_G^i$ ones; from $\sim 0.1RT$ to $\sim 0.75RT$, on the other hand, the model chemistries of the form L_E^i//6-31G(d) outperform those in the rest of groups; finally, for distances $d_{12} > 0.75RT$, it is recommendable to use homolevel model chemistries.

7.3.3 Interlude

The general abstract framework behind the investigation presented in this chapter (and also behind most of the works found in the literature), may be described as follows:

The objects of study are the *model chemistries* defined by Pople [220] and discussed in the introduction. The space containing all possible MCs is a rather complex and multidimensional one and it is denoted by $\mathcal{M}$ in fig. 7.12. The model chemistries under scrutiny are applied to a particular *problem* of interest, which may be thought to be formed by three ingredients: the *physical system*, the *relevant observables* and the *target accuracy*. The model chemistries are then selected according to their ability to yield numerical values of the relevant observables for the physical system studied within the target accuracy. The concrete numerical values that one wants to approach are those given by the *exact model chemistry* MC_ε, which could be thought to be either the experimental data or the exact solution of the electronic Schrödinger equation. However, the computational effort needed to perform the calculations required by MC_ε is literally infinite, so that, in practice, one is forced to work with a *reference model chemistry* MC^{ref}, which, albeit different from MC_ε, is thought to be close to it. Finally, the set of model chemistries that one wants to investigate are compared to MC^{ref} and the nearness to it is seen as approximating the nearness to MC_ε.

These comparisons are commonly performed using a numerical quantity d that is a function of the relevant observables. In order for the intuitive ideas about relative proxim-

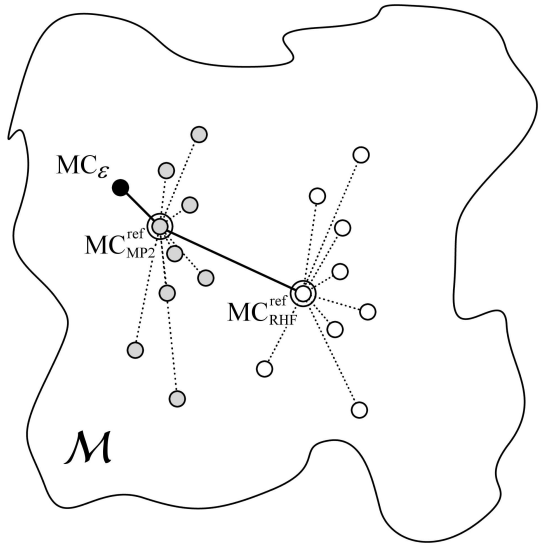


Figure 7.12: Space $\mathcal{M}$ of all model chemistries. The exact model chemistry MC_ε is shown as a black circle, MP2 model chemistries are shown as grey-filled circles and RHF model chemistries as white-filled ones. The homolevel reference PESs are indicated with an additional circle around the points. The situation depicted is (schematically) the one found in this study.

ity in the space $\mathcal{M}$ to be captured and the above reasoning to be meaningful, this numerical quantity d must have some of the properties of a mathematical distance (see sec. 3.9). In particular, it is advisable that the *triangle inequality* is obeyed, so that, for any model chemistry MC, one has that

$$d(\mathrm{MC}_\varepsilon, \mathrm{MC}) \leq d(\mathrm{MC}_\varepsilon, \mathrm{MC}^{\mathrm{ref}}) + d(\mathrm{MC}^{\mathrm{ref}}, \mathrm{MC}) \,, \tag{7.4a}$$

$$d(\mathrm{MC}_\varepsilon, \mathrm{MC}) \geq \left| d(\mathrm{MC}_\varepsilon, \mathrm{MC}^{\mathrm{ref}}) - d(\mathrm{MC}^{\mathrm{ref}}, \mathrm{MC}) \right| \,, \tag{7.4b}$$

and, assuming that $d(\mathrm{MC}_\varepsilon, \mathrm{MC}^{\mathrm{ref}})$ is small (and d is a positive function), we obtain

$$d(\mathrm{MC}_\varepsilon, \mathrm{MC}) \simeq d(\mathrm{MC}^{\mathrm{ref}}, \mathrm{MC}) \,, \tag{7.5}$$

which is the sought result in agreement with the ideas stated at the beginning of this section.

The distance introduced in sec. 7.2.2, measured in this case on the conformational energy surfaces (the relevant observable) of the model dipeptide formyl-L-alanine-amide (the physical system), approximately fulfills the triangle inequality and thus captures the nearness concept in the space $\mathcal{M}$ of model chemistries.

Now, as we have advanced and after having completed the intramethod parts of the study with both the RHF and MP2 methods, we shall use the ideas discussed above to tackle the natural question about the transferability of the RHF results to the more demanding and more accurate MP2-based model chemistries.

As a first step to answer this question, we point out that the distance between the reference RHF/6-311++G(2df,2pd) model chemistry and the MP2 one depicted in fig. 7.7 is $\sim 1.42RT$. This prevents us from using the former as an approximation of the latter even for dipeptides if we want that the conformational behaviour at room temperature be unaltered. It also indicates that, whereas basis set convergence has been reasonably achieved, within the family of Pople's Gaussian basis sets, both for homo- and heterolevel model chemistries inside the two methods, the *convergence in method has not been achieved in the* RHF $\rightarrow$ MP2 *step, even with the largest basis set investigated* 6-311++G(2df,2pd).

Complementarily to this, in fig. 7.13, we show the distance of all RHF//RHF model chemistries studied in sec. 7.3.1 (except for the inefficient $L_E^i // L_G^{\mathrm{best}}$ ones), with both the RHF reference (in the y-axis) and the MP2 one (in the x-axis). Some relevant remarks may be made about the situation encountered:

- *The distance of all RHF-intramethod model chemistries to the MP2 reference is larger than RT*, therefore, none of the former may be used to approximate the latter, not even in dipeptides.

- Although a general trend could be perceived and, for example, the RHF homolevels can be clearly divided *in both axes* by the heavy atoms polarization gap found in the previous sections, *the correlation between the distance to the MP2 reference and the distance to the RHF one is as low as* $r \simeq 0.66$, being r Pearson's correlation coefficient. Therefore, almost all details are lost and the accuracy with respect to RHF/6-311++G(2df,2pd) cannot be translated into accuracy with respect to the MP2 reference.

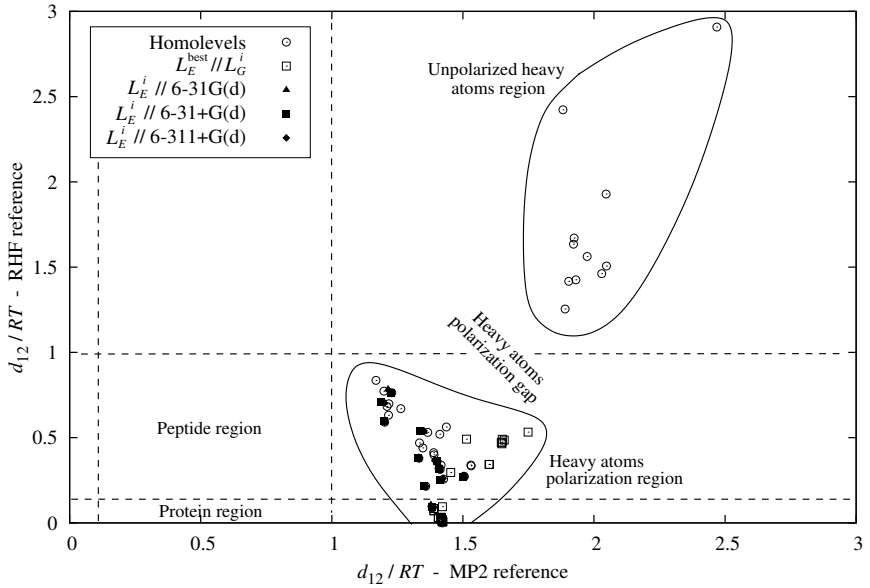


Figure 7.13: All *RHF-intramethod* model chemistries studied in sec. 7.3.1, except for the inefficient $L_E^i//L_G^{\text{best}}$ ones. The distance d_{12}, in units of RT at 300° K, with the homolevel MP2/6-311++G(2df,2pd) reference is shown in the x-axis, while the distance with the RHF reference is shown in the y-axis. The different accuracy regions depending on d_{12}, are labeled, and two groups of *homolevel* MCs are distinguished: those that contain heavy atoms polarization shells and those that do not.

- Related to the previous point, *some strange behaviours are present*. For example, not only are there RHF//RHF model chemistries that are closer to the MP2 reference than RHF/6-311++G(2df,2pd), but the one that is closest is the small RHF/4-31G(d,p) homolevel. *This is probably caused by fortituous cancellations that shall not allow systematization and that may unpredictably vary from one problem to another*. Similar compensations have already been observed in the literature [207].

- If we denote by $\text{MC}^{\text{ref}}_{\text{MP2}}$ the MP2 reference model chemistry and, by $\text{MC}^{\text{ref}}_{\text{RHF}}$, the RHF one, we may use eqs. (7.4),

$$d(\text{MC}^{\text{ref}}_{\text{MP2}}, \text{MC}) \leq d(\text{MC}^{\text{ref}}_{\text{MP2}}, \text{MC}^{\text{ref}}_{\text{RHF}}) + d(\text{MC}^{\text{ref}}_{\text{RHF}}, \text{MC}) \,, \tag{7.6a}$$

$$d(\text{MC}^{\text{ref}}_{\text{MP2}}, \text{MC}) \geq \left| d(\text{MC}^{\text{ref}}_{\text{MP2}}, \text{MC}^{\text{ref}}_{\text{RHF}}) - d(\text{MC}^{\text{ref}}_{\text{RHF}}, \text{MC}) \right| \,, \tag{7.6b}$$

  to notice that, since $d(\text{MC}^{\text{ref}}_{\text{MP2}}, \text{MC}^{\text{ref}}_{\text{RHF}}) \simeq 1.42RT$, *for any model chemistry MC that is close to the RHF-intramethod reference*, i.e., that present a small $d(\text{MC}^{\text{ref}}_{\text{RHF}}, \text{MC})$, we have that

$$d(\text{MC}^{\text{ref}}_{\text{MP2}}, \text{MC}) \simeq d(\text{MC}^{\text{ref}}_{\text{MP2}}, \text{MC}^{\text{ref}}_{\text{RHF}}) \simeq 1.42RT \,. \tag{7.7}$$

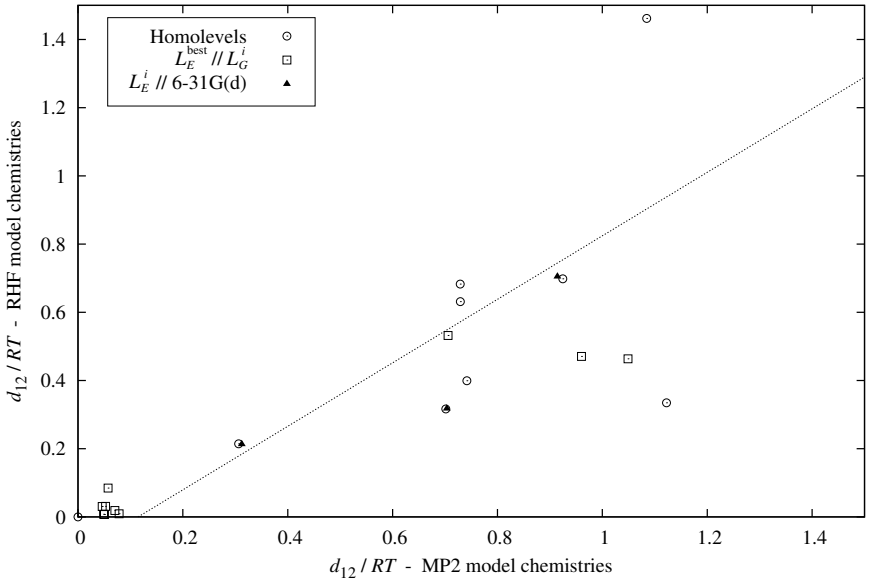


Figure 7.14: Distance to their respective references of the *MP2*- and *RHF-intramethod* model chemistries corresponding to the same combination of basis sets $B_E^i // B_G^i$, expressed in units of RT at 300º K. Only the region with $d_{12} < 1.5RT$ is shown, and the best-fit line is depicted with a dotted line.

This set of RHF-intramethod model chemistries that are close to the RHF reference and that present the approximately constant value of $d(\mathrm{MC}^{\mathrm{ref}}_{\mathrm{MP2}}, \mathrm{MC})$ above is, more or less, the lower group encircled in fig. 7.13.

All the points above illustrate what we have already advanced at the beginning of sec. 7.3.2: that the accuracy (or the efficiency, if computational time is included in the discussion) of any model chemistry with respect to a good RHF reference, such as the RHF/6-311++G(2df,2pd) one, *cannot be transferred to higher levels of the theory* and, therefore, any such comparison must be seen as providing information only about the infinite basis set Hartree-Fock limit.

To close this section, let us approach the question of the RHF → MP2 transferability of the results from a different angle.

We have proved in the preceding paragraphs that the study of RHF-intramethod model chemistries comparing them to a good RHF reference cannot be used for predicting the accuracy of these MCs with respect to a probably better MP2 reference. Now, in sec. 7.3.2, MP2-intramethod model chemistries have been compared to the MP2/6-311++G(2df,2pd) homolevel, which, in turn, has been shown to be close to the infinite basis set MP2 limit. However, this level of the theory is very demanding computationally: the whole 12×12 grid of points in the PES of HCO-L-Ala-NH_2 has taken ∼ 3 years of CPU time in 3.20 GHz PIV machines, while the one calculated at RHF/6-311++G(2df,2pd) has taken 'only' ∼ 6 months (see sec. 7.2.1).

Therefore, we have decided to check whether or not the accuracy of a given RHF-intramethod MC with respect to the RHF reference is indicative of the accuracy of the

MP2-intramethod MC that uses the same basis sets with respect to its own MP2 reference. The answer to this question is in fig. 7.14. There, each point corresponds to a given combination of basis sets $B^i_E//B^i_G$ and, in the x-axis, the distance between the associated MP2 model chemistry and the MP2/6-311++G(2df,2pd) reference is shown. In the y-axis, on the other hand, we present the distance of the analogous RHF model chemistry to the RHF/6-311++G(2df,2pd) homolevel.

Although, since we have had to restrict ourselves to that combinations that were present both in sec. 7.3.1 and in sec. 7.3.2, the set of MCs is smaller in this case, the conclusion extracted is that *the correlation is more significant than before*: $r \simeq 0.92$ if we use all the MCs, and $r \simeq 0.80$ if we remove the 3-21G homolevel, which is very inaccurate in both cases, from the set. This indicates that, although some details might be lost, *the relative efficiency of Gaussian basis sets in RHF-intramethod studies provides hints about their performance at MP2*, and it partially justifies the structure of the investigation presented in this chapter.

Finally, the overall situation described in this section and the relations among all the intramethod model chemistries studied are schematically depicted in fig. 7.12.

7.3.4 MP2//RHF-intermethod model chemistries

In the final part of the study presented here, we investigate the efficiency of *heterolevel* model chemistries in which the geometry is calculated at RHF and, then, a single-point energy calculation is performed on top of it at MP2. They shall be termed *MP2//RHF-intermethod model chemistries*.

To this end, the RHF geometries that are used are those computed with the 8 basis sets in table 7.6. Like in sec. 7.3.2, they have been selected from those in table 7.2 looking for the most efficient ones, but also trying to reasonably sample the whole group of basis sets, in order to check whether or not the behaviours and signals observed in the remaining parts of the study are repeated here. The MP2 single-points, on the other hand, are computed with the whole set of possibilities in table 7.2.

In fig. 7.15, we present an efficiency plot, using the MP2/6-311++G(2df,2pd) homolevel as reference MC, and containing all the MP2//MP2 model chemistries studied in sec. 7.3.2 together with the new 88 possible MP2//RHF-intermethod combinations of the form MP2/B^i_E//RHF/B^i_G.

Some conclusions can be drawn from this plot:

- Due to the larger computational demands of the MP2 method, even the model chemistries whose geometry has been computed at the highest RHF level, the one with the 6-311++G(2df,2dp) basis set, are much cheaper than the MP2 reference.

3-21G	6-31G(d)	6-31+G(d)	6-311+G(d)
6-31G	6-31G(2d,2p)	6-31++G(2d,2p)	6-311++G(2df,2pd)

Table 7.6: Basis sets investigated for calculating the geometry in the *MP2//RHF-intermethod* part of the study.

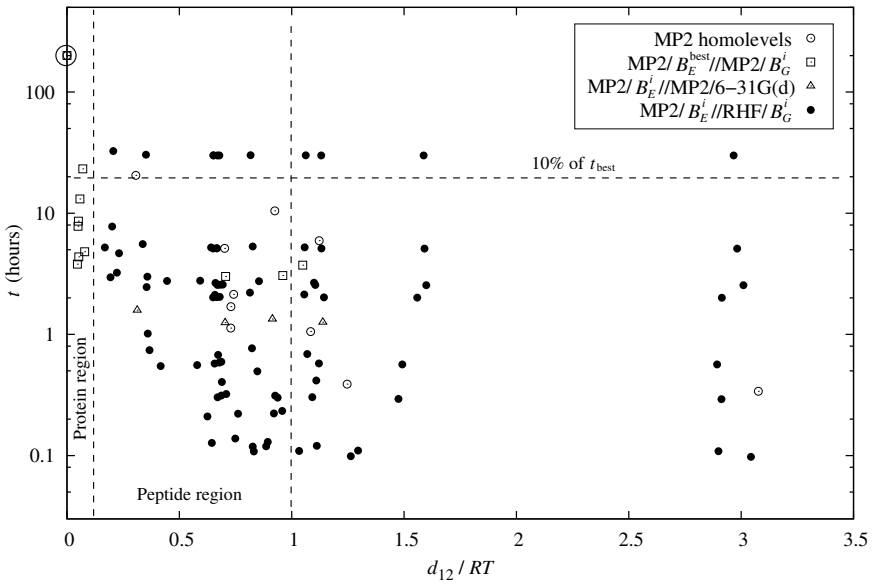


Figure 7.15: Efficiency plot of all the MP2//MP2 model chemistries in fig. 7.11 together with the new 88 possible MP2//RHF-intermethod combinations of the form MP2/B_E^i//RHF/B_G^i introduced in this section. In the x-axis, we show the distance d_{12}, in units of RT at 300º K, between any given model chemistry and the reference one (the *homolevel* MP2/6-311++G(2df,2pd), indicated by an encircled point), while, in the y-axis, we present in logarithmic scale the average computational time per point of the 12×12 grid defined in the Ramachandran space of the model dipeptide HCO-L-Ala-NH_2. The different accuracy regions depending on d_{12}, are labeled, and the 10% of the time t_{best} taken by the reference homolevel MP2/6-311++G(2df,2pd) is also indicated.

Their times are slightly larger than the 10% of t_{best}, whereas all the rest of MP2//RHF model chemistries take less than that bound.

- For all the RHF geometries, the model chemistries whose MP2 single-point has been calculated with 3-21G, 6-31G, 6-31++G and most of the 6-311+G(d) ones lie above $d_{12} = RT$, so that they should not be used even on dipeptides. This is related to the heavy atoms polarization gap observed in previous sections, although the signal is not so strong here.

- The rest of MP2//RHF model chemistries not included in the two previous points lie at the *efficient region*, defined as that for which $d_{12} < RT$ and $t < 10\%$ of t_{best}. This confirms the *heterolevel assumption* also in the intermethod context.

- However, no MP2//RHF-intermethod model chemistry, not even the ones with the single-point calculated at the highest MP2/6-311++G(2df,2pd) level, lie in the protein region. Therefore, *if we want to approximate the reference MP2 results for peptides longer than 100 residues, the single-point energy calculation must be performed at MP2*.

- *There is no accuracy region where the MP2-homolevel model chemistries are more*

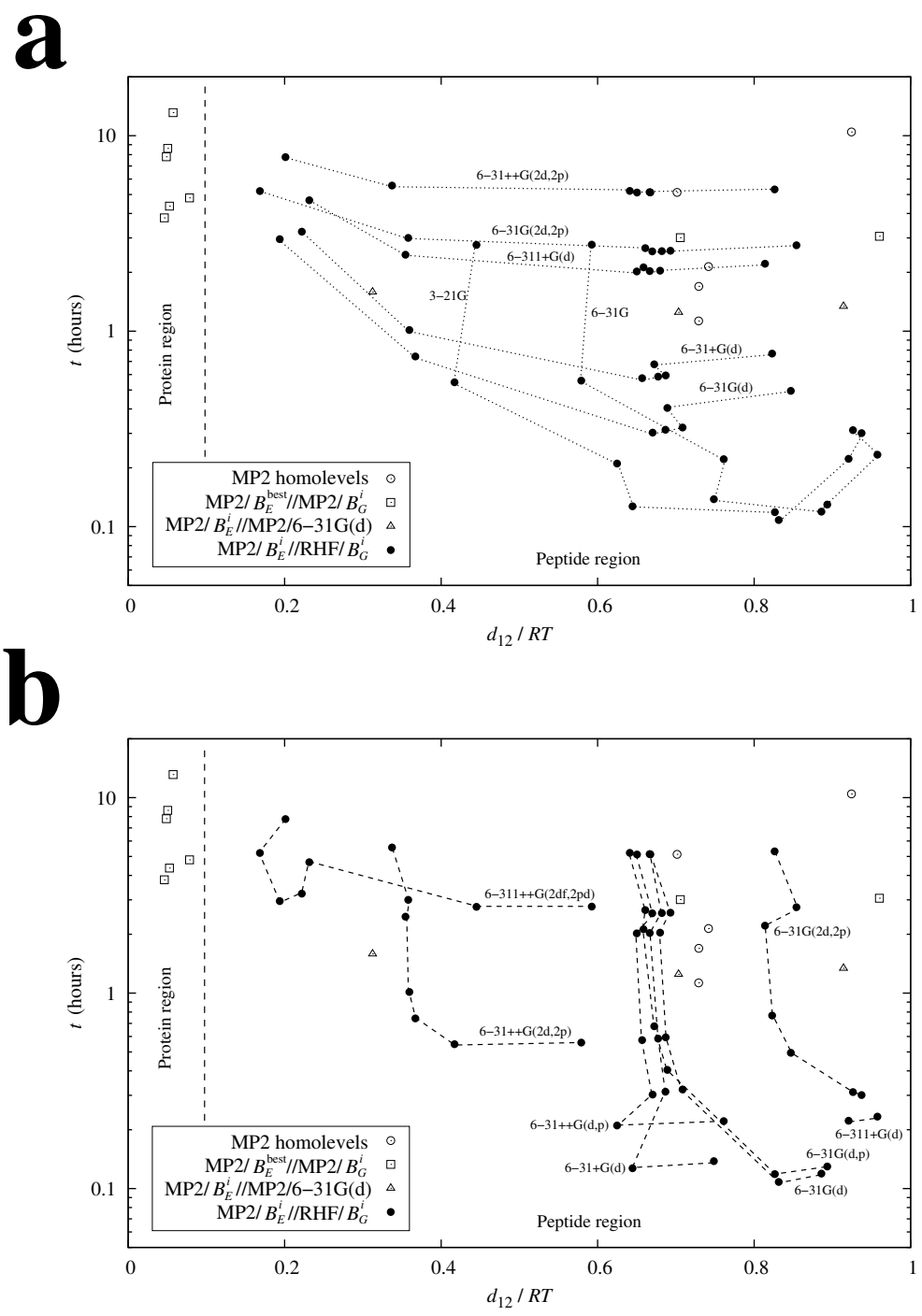


Figure 7.16: Selected region of the efficiency plot in fig. 7.15. In **(a)**, the MCs sharing the same RHF level for the geometry have been joined by *dotted lines* and the basis set used for that part of the calculation is indicated. In **(b)**, the MCs sharing the same MP2 level for the single-point calculations have been joined by *broken lines* and the corresponding basis set labels are also shown. The order in which the points have been joined in both cases has no meaning at all and it is only intended for visual convenience.

Efficient MP2//MP2 and MP2//RHF MCs	d_{12}/RT [a]	N_{res} [b]	t (% of t_{best}) [c]
MP2/6-311++G(2df,2pd)//MP2/6-31G(d)	0.046	468.3	1.90%
MP2/6-311++G(2df,2pd)//RHF/6-31G(d)	0.194	26.7	1.48%
MP2/6-31++G(2d,2p)//RHF/6-31G(d)	0.367	7.4	0.37%
MP2/6-31++G(2d,2p)//RHF/3-21G	0.417	5.7	0.27%
MP2/6-31+G(d)//RHF/3-21G	0.645	2.4	0.06%
MP2/6-31G(d)//RHF/3-21G	0.831	1.4	0.05%
MP2/6-31++G//RHF/3-21G	1.033	0.9	0.05%
MP2/6-31G//RHF/3-21G	1.263	0.6	0.05%
MP2/3-21G//RHF/3-21G	3.043	0.1	0.05%

Table 7.7: List of the most efficient MP2//MP2 and MP2//RHF model chemistries located at the lower-left envelope of the cloud of points in fig. 7.15. The first block contains the only MP2//MP2 model chemistry in the list, the second one the MP2//RHF ones with a distance d_{12} below RT, and the third one those that are inaccurate even for dipeptides. [a]Distance with the reference MC (the homolevel MP2/6-311++G(2df,2pd)), in units of RT at 300° K. [b]Maximum number of residues in a polypeptide potential up to which the corresponding model chemistry may correctly approximate the reference. [c]Required computational time, expressed as a fraction of t_{best}.

efficient than the rest.

Now, in fig. 7.16, the efficient region of the previous plot is enlarged and, due to the large number of MP2//RHF model chemistries studied, two subplots are produced: the one in fig. 7.16a, in which the MCs sharing the same RHF level for the geometry have been joined by dotted lines, and the one in fig. 7.16b, in which the MCs sharing the same MP2 level for the single-point calculations have been joined by broken lines.

Let us remark some interesting facts that can be seen in these two more detailed plots:

- The leftmost group of five MP2//RHF model chemistries that show the highest accuracy are those in which the geometry has been obtained with basis sets containing heavy atoms polarization functions and the single-point energy calculation has been performed at MP2/6-311++G(2df,2pd). In particular, the MP2/6-311++G(2df,2pd) //RHF/6-31G(d) potential energy surface can correctly approximate the reference one up to peptides of $\sim$ 25 residues at around 1% its computational cost. This supports the *heterolevel assumption* for MP2//RHF-intermethod model chemistries.

- The RHF geometries calculated with the unpolarized basis sets 3-21G and 6-31G are, in general, less accurate than the rest, however, due to their low computational cost, they turn out to be the most efficient ones from $d_{12} \simeq 0.4RT$ on. Remarkably, 3-21G is more efficient than 6-31G.

- In fig. 7.16b, we can observe that, for the medium-sized basis sets 6-31++G(d,p), 6-31+G(d), 6-31G(d,p) and 6-31G(d), the single-point accuracy is rather insensitive to their differences and they may be used interchangeably. There is, however,

a weak signal, in the region of unpolarized RHF geometries, indicating that the addition of diffuse functions may increase the quality of the energy calculations at MP2.

- The relative accuracy of the MCs whose MP2 single-point has been computed at 6-31++G(2d,2p) and at 6-31G(2d,2p) suggests that, like in previous parts of the study, *it is a good idea to add diffuse functions to basis sets that contain doubly-split polarizations shells*, also for the MP2 energy calculations of MP2//RHF-intermethod model chemistries.

- Like it happened in sec 7.3.1, in fig. 7.16a, we notice that there is no real improvement if we calculate the RHF geometry beyond 6-31G(d). So that, *an accumulation point is reached for RHF geometries in MP2//RHF-intermethod model chemistries*.

Finally, in table 7.7, we present the most efficient MCs that lie at the lower-left envelope of the plot in fig. 7.15. These are *the most efficient model chemistries found in this work*.

7.4 Conclusions

In this study, we have investigated more than 250 potential energy surfaces of the model dipeptide HCO-L-Ala-NH_2 calculated with homo- and heterolevel RHF//RHF, MP2//MP2 and MP2//RHF model chemistries. As far as we are aware, the highest-level PESs in the literature, the MP2/6-311++G(2df,2pd) homolevel in fig. 7.7, has been used as a reference and all the rest of calculations have been compared to it (except for sec. 7.3.1, where the RHF//RHF model chemistries have been compared to RHF/6-311++G(2df,2pd)). The data and the results extracted are so extense that we have decided to give here a brief summary of the most important ones.

The first conclusion that we want to point out is that, for the largest basis set evaluated here, the 6-311++G(2df,2pd) one, for which the RHF and MP2 limits appear to have been reached, *the convergence in method has not been achieved*. I.e., the distance between the MP2 and RHF references is $d_{12} \simeq 1.42RT$, so that the latter cannot be used to approximate the former even for dipeptides. Therefore, *we discourage the use of RHF//RHF model chemistries for peptides*, and, unless otherwise stated, most of the conclusions below should be understood as referring either to MP2//MP2-intramethod or to MP2//RHF-intermethod model chemistries, which have proved to be acceptably accurate with respect to the best MP2 calculation.

Regarding the relative efficiency of the Pople split-valence basis sets investigated:

- In the whole study, the polarization shells in heavy atoms have been shown to be essential to accurately account for both the conformational dependence of the geometry and of the energy of the system. Except for some particular model chemistries with 3-21G geometries, which may be used if we plan to describe short oligopeptides, *our recommendation is that polarization functions in heavy atoms be included*.

- In most cases, we have also observed a strong signal indicating that *no basis sets should be used containing doubly-split polarization shells and no diffuse functions*.

- *The* 6-31G(d) *basis set*, which is frequently used in the literature, [163, 215, 240, 241, 247, 389, 409], *has turned out to be a very efficient one for calculating the geometry both at RHF and MP2*.

- Regarding the basis set convergence issue, we can conclude that, *for the largest basis sets in the Pople split-valence family, both the RHF and MP2 infinite basis set limits are approximately reached*.

- Finally, some *weaker signals* have been observed suggesting that to add higher angular momentum polarization shells (f,d) before adding the lower ones may be inefficient, that it is not recommendable to put polarization or diffuse functions on hydrogens only, and that it may be efficient in some cases to add diffuse functions to singly-polarized basis sets.

Regarding the heterolevel assumption, which, as far as we are aware, has been tested in this work for the first time in full PESs:

- As a general and very clear conclusion, since only some small-basis set homolevels lie in the lower-left envelope of the efficiency plots presented in the previous sections, and, in all cases, it happens for distances d_{12} greater than RT, we can say that *the heterolevel assumption is correct for the description of the conformational behaviour of the system studied here with MP2//MP2 and MP2//RHF model chemistries* (also for RHF//RHF-heterolevels but, as we remarked above, this has little computational interest)

- Due to the much stronger dependence of the accuracy of MCs on the level used for the single-point than on the one used for the geometry optimization, together with the lower computational cost of the former, *the general recommendation is that the greatest computational effort be dedicated to the energy calculation*.

- Despite this general thumb rule, *if one wants to approximate the MP2 reference calculation for peptides of more than 100 residues, the geometry must be calculated using MP2*. Nevertheless, with small and cheap basis sets, such as 6-31G(d), very accurate MP2//MP2 results may be obtained at a low computational cost.

Finally, let us repeat the remark at the end of chapters 4 and 6: The investigation performed here has been done in one of the simplest dipeptides. The fact that we have treated it as an isolated system, the small size of its side chain and also its aliphatic character, all play a role in the results obtained. Hence, for bulkier residues included in polypeptides, and, specially for those that are charged or may participate in hydrogen-bonds, the conclusions drawn about the relative importance of the different type of functions in the basis set, as well as those regarding the comparison between RHF and MP2, should be approached with caution and much interesting work remains to be done.

Appendices

A The meaning of probability density functions

Let us define a *stochastic* or *random variable*[115] as a pair (X, p), with X a subset of $\mathbb{R}^n$ for some n and p a function that takes n-tuples $x \equiv (x_1, \ldots, x_n) \in X$ to positive real numbers,

$$\begin{array}{rccc} p: & X & \longrightarrow & [0,\infty) \\ & x & \longmapsto & p(x) \end{array}$$

Then, X is called *range*, *sample space* or *phase space*, and p is termed *probability distribution* or *probability density function* (PDF). The phase space can be discrete, a case with which we shall not deal here, or continuous, so that $p(x)\,\mathrm{d}x$ (with $\mathrm{d}x := \mathrm{d}x_1 \cdots \mathrm{d}x_n$) represents the probability of occurrence of some n-tuple in the set defined by $(x, x+\mathrm{d}x) := (x_1, x_1 + \mathrm{d}x_1) \times \cdots \times (x_n, x_n + \mathrm{d}x_n)$, and the following normalization condition is satisfied:

$$\int_X p(x)\,\mathrm{d}x = 1\ . \tag{A.1}$$

It is precisely in the continuous case where the interpretation of the function $p(x)$ alone is a bit problematic, and playing intuitively with the concepts derived from it becomes dangerous. On one side, it is obvious that $p(x)$ is not the probability of the value x happening, since the probability of any specific point in a continuous space must be zero (what is the probability of selecting a random number between 3 and 4 and obtaining *exactly* π?). In fact, the correct way of using $p(x)$ to assign probabilities to the n-tuples in X is 'to multiply it by differentials' and say that it is the probability that any point in a differentially small interval occurs (as we have done in the paragraph above eq. (A.1)). The reason for this may be expressed in many ways: one may say that $p(x)$ is an object that only makes sense under an integral sign (like a Dirac delta), or one may realize that only probabilities of finite subsets of X can have any meaning. In fact, it is this last statement the one that focuses the attention on the fact that, if we decide to reparameterize X and perform a change of variables $x'(x)$, what should not change are the integrals over finite subsets of X, and, therefore, $p(x)$ cannot transform as a scalar quantity (i.e., satisfying $p'(x') = p(x(x'))$), but according to a different rule.

If we denote the *Jacobian matrix* of the change of variables by $\partial x/\partial x'$, we must have that

$$p'(x') = \left|\det\left(\frac{\partial x}{\partial x'}\right)\right| p(x(x'))\ , \tag{A.2}$$

[115] See Van Kampen [410] for a more complete introduction to probability theory.

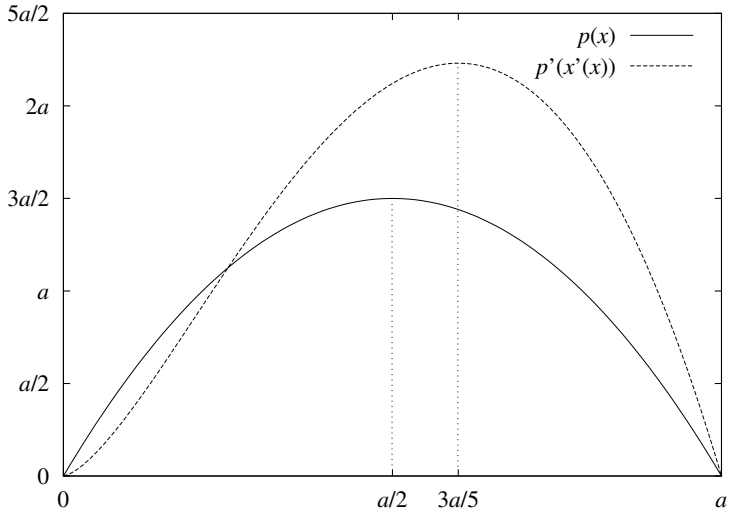


Figure A.1: Probability density functions $p(x)$ and $p'(x'(x))$ in eqs. (A.4) and (A.5) respectively. In the axes, the quantities x and $p(x)$ are shown for convenience. Note that the area enclosed by the two curves is different; this is because $p'(x'(x))$ is normalized with the measure $\mathrm{d}x'$ and not with $\mathrm{d}x$, which is the one implicitly assumed in this representation.

so that, for any finite set $Y \subset X$ (with its image by the transformation denoted by Y'), and indicating the probability of a set with a capital P, we have the necessary property

$$P(Y) := \int_Y p(x)\,\mathrm{d}x = \int_{Y'} p'(x')\,\mathrm{d}x' =: P'(Y')\,. \tag{A.3}$$

All in all, the object that has meaning content is P and not p. If one needs to talk about things such as the *most probable regions*, or *the most probable states*, or *the most probable points*, or if one needs to compare in any other way the relative probabilities of different parts of the phase space X, an *arbitrary* partition of X into finite subsets $(X_1, \ldots, X_i, \ldots)$ must be defined[116]. These X_i should be considered more useful *states* than the individual points $x \in X$ and their probabilities $P(X_i)$, which, contrarily to $p(x)$, do not depend on the coordinates chosen, should be used as the meaningful quantities about which to make well-defined probabilistic statements.

To illustrate this, let us see an example: suppose we have a 1-dimensional PDF

$$p(x) = \frac{6}{a^3} x\,(a - x)\,. \tag{A.4}$$

The maximum of $p(x)$ is at $x = a/2$, however, it would not be very clever to declare that $x = a/2$ *is the most probable value of* x, since one may choose to describe the problem with a different but perfectly legitimate variable x', whose relation to x is, say, $x = x'^2$, and find the PDF in terms of x' using eq. (A.2):

$$p'(x') = \frac{12}{a^3} x'^3\,(a - x'^2)\,. \tag{A.5}$$

[116] Two additional reasonable properties should be asked to such a partition: (i) the sets in it must be exclusive, i.e., $X_i \cap X_j = \emptyset, \forall i \neq j$, and (ii) they must fill the phase space, $\bigcup_i X_i = X$

Now, insisting on the mistake, we may find the maximum of $p'(x')$, which lies at $x' = (3a/5)^{1/2}$ (see fig. A.1), and declare it *the most probable value of* x'. But, according to the change of variables given by $x = x'^2$, the point $x' = (3a/5)^{1/2}$ corresponds to $x = 3a/5$ and, certainly, it is not possible that $x = a/2$ and $x = 3a/5$ are the most probable values of x at the same time!

To sum up, only finite regions of continuous phase spaces can be termed *states* and meaningfully assigned a probability that do not depend on the coordinates chosen. In order to do that, an *arbitrary* partition of the phase space must be defined.

Far for being an academic remark, this is relevant in the study of the equilibrium of proteins, where, very commonly, Anfinsen's *thermodynamic hypothesis* is invoked (see sec. 1.4). Loosely speaking, it says that *the functional native state of proteins lies at the minimum of the effective potential energy* (i.e., the maximum of the associated Boltzmann PDF, proportional to $e^{-\beta W}$, in eq. (1.7)), but, according to the properties of PDFs described in the previous paragraphs, much more qualifying is needed.

First, one must note that all complications arise from the choice of integrating out the momenta (for example, in eqs. (1.6) or (6.9)) to describe the equilibrium distribution of the system with a PDF dependent only on the potential energy. If the momenta were kept and the PDF expressed in terms of the complete Hamiltonian as $p(q^\mu, \pi_\mu) = e^{-\beta H}/Z$, then, it would be invariant under canonical changes of coordinates (which are the physically allowed ones), since the Jacobian determinant that appears in eq. (A.2) equals unity in such a case (see footnote 108 in chapter 6). If we now look, using this complete description in terms of H, for the *most probable point* (q^μ, π_μ) in the whole dynamical phase space, the answer does not depend on the coordinates chosen: It is the point with all momenta π_μ set to zero (since the kinetic energy is a positive defined quadratic form on the π_μ), and the positions q^μ set to those that minimize the potential energy $V(q^\mu)$, denoted by $q^\mu_{\min}$. If we now perform a point transformation, which is a particular case of the larger group of canonical transformations [343],

$$q^\mu \to q'^\mu(q^\mu) \qquad \text{and} \qquad \pi_\mu \to \pi'_\mu = \frac{\partial q^\nu}{\partial q'^\mu}\pi_\nu \,, \tag{A.6}$$

the *most probable point* in the new coordinates turns out to be 'the same one', i.e., the point $(q'^\mu, \pi'_\mu) = (q'^\mu(q^\mu_{\min}), 0)$, and all the insights about the problem are consistent.

However, if one decides to integrate out the momenta, the marginal PDF on the positions that remains has a more complicated meaning than the joint one on the whole phase space and lacks the reasonable properties discussed above. The central issue is that the marginal $p(q^\mu)$ (for example, the one in eq. (6.12)) quantifies the probability that the positions of the system be in the interval $(q^\mu, q^\mu + \mathrm{d}q^\mu)$ *without any knowledge about the momenta*, or, otherwise stated, *for any value of the momenta*.

In Euclidean coordinates, the volume in momenta space does not depend on the positions, however, in general curvilinear coordinates, the accessible momenta volume is different from point to point, and one can say the same about the *kinetic entropy* (see chapter 6) associated with the removed π_μ, which, apart from the potential energy, also enters the coordinate PDF.

If, despite these inconveniences, the description in terms of only the positions q^μ is chosen to be kept (which is typically recommendable from the computational point of view), two different approaches may be followed to assure the meaningfulness of the

statements made: Either some partition of the conformational space into finite subsets must be defined, as it is described in the beginning of this appendix and as it is done in ref. 116, or the position-dependent kinetic entropies that appear when curvilinear coordinates are used and that are introduced in chapter 6 must be included in the effective potential energy function.

B Functional derivatives

A *functional* $\mathcal{F}[\Psi]$ is a mapping that takes functions to numbers (in this document, only functionals in the real numbers are going to be considered):

$$\begin{array}{rccc} \mathcal{F}: & \mathcal{G} & \longrightarrow & \mathbb{R} \\ & \Psi & \longmapsto & \mathcal{F}[\Psi] \end{array}$$

For example, if the function space $\mathcal{G}$ is the Hilbert space of square-integrable functions L^2 (the space of states of quantum mechanics), the objects in the domain of $\mathcal{F}$ (i.e., the functions in L^2) can be described by infinite-tuples $(c_1, c_2, \ldots)$ of complex numbers and $\mathcal{F}$ may be pictured as a function of infinite variables.

When dealing with function spaces $\mathcal{G}$ that meet certain requirements[117], the limit on the left-hand side of the following equation can be written as the integral on the right-hand side:

$$\lim_{\varepsilon \to 0} \frac{\mathcal{F}[\Psi_0 + \varepsilon\delta\Psi] - \mathcal{F}[\Psi_0]}{\varepsilon} := \int \frac{\delta\mathcal{F}[\Psi_0]}{\delta\Psi}(x)\delta\Psi(x)\mathrm{d}x \,, \tag{B.1}$$

where x denotes a point in the domain of the functions in $\mathcal{G}$, and the the object $(\delta\mathcal{F}[\Psi_0]/\delta\Psi)(x)$ (which is a function of x not necessarily belonging to $\mathcal{G}$) is called the *functional derivative* of $\mathcal{F}[\Psi]$ in the in the *point* Ψ_0.

One common use of this functional derivative is to find stationary points of functionals. A function Ψ_0 is said to be an *stationary point* of $\mathcal{F}[\Psi]$ if:

$$\frac{\delta\mathcal{F}[\Psi_0]}{\delta\Psi}(x) = 0 \,. \tag{B.2}$$

In order to render this definition operative, one must have a method for computing $(\delta\mathcal{F}[\Psi_0]/\delta\Psi)(x)$. Interestingly, it is possible, in many useful cases (and in all the applications of the formalism in this document), to calculate the sought derivative directly from the left-hand side of eq. (B.1). The procedure, in such a situation, begins by writing out $\mathcal{F}[\Psi_0 + \varepsilon\delta\Psi]$ and clearly separating the different orders in ε. Secondly, one drops the terms of zero order (by virtue of the subtraction of the quantity $\mathcal{F}[\Psi_0]$) and those of second order or higher (because they vanish when divided by ε and the limit $\varepsilon \to 0$ is taken). The remaining terms, all of order one, are divided by ε and, finally, $(\delta\mathcal{F}[\Psi_0]/\delta\Psi)(x)$ is identified out of the resulting expression (which must written in the form of the right-hand side of eq. (B.1)). For a practical example of this process, see secs. 2.6 and 2.7.

[117] We will not discuss the issue further but let it suffice to say that L^2 does satisfy these requirements.

C Lagrange multipliers and constrained stationary points

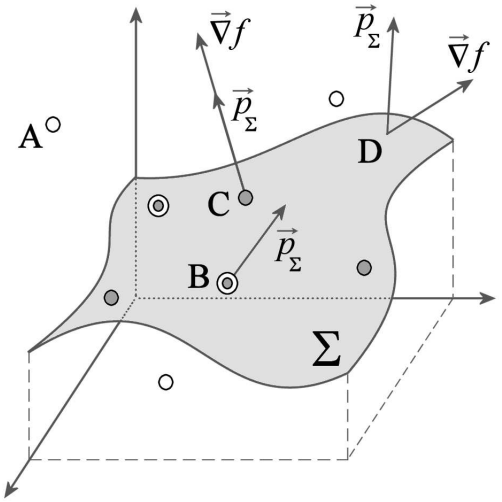


Figure C.1: Schematic depiction of a constrained stationary points problem. Σ is the 2-dimensional search space, which is embedded in $\mathbb{R}^3$. The *white-filled circles* are solutions of the unconstrained problem only, the *gray-filled circles* are solutions of only the constrained one and the *gray-filled circles inside white-filled circles* are solutions of both. A, B, C and D are examples of different situations discussed in the text.

Very often, when looking for the stationary points of a function (or a functional), the search space is not the whole one, in which the derivatives are taken, but a certain subset of it defined by a number of constraints. An elegant and useful method for solving the constrained problem is that of the *Lagrange multipliers*.

Although it can be formally generalized to infinite dimensions (i.e., to functionals, see appendix B), here we will introduce the method in $\mathbb{R}^N$ in order to gain some geometrical insight and intuition.

The general framework may be described as follows: we have a differentiable function $f(\vec{x})$ that takes points in $\mathbb{R}^N$ to real numbers and we want to find the stationary points of f restricted to a certain subspace Σ of $\mathbb{R}^N$, which is defined by K *constraints*[118]:

$$L_i(\vec{x}) = 0 \quad i = 1, \ldots, K \,. \tag{C.1}$$

The points that are the solution of the constrained problem are those $\vec{x}$ belonging to Σ where the first order variation of f would be zero if the derivatives were taken 'along' Σ. In other words, the points $\vec{x}$ where the gradient $\vec{\nabla} f$ has only components (if any) in directions that 'leave' Σ (see below for a rigorous formalization of these intuitive ideas). Thus, when comparing the solutions of the unconstrained problem to the ones of the constrained problem, three distinct situations arise (see fig. C.1):

[118] If the constraints are functionally independent, one must also ask that $K < N$. If not, Σ will be either a point (if $K = N$) or empty (if $K > N$).

i) A point $\vec{x}$ is a solution of the unconstrained problem (i.e. it satisfies $\vec{\nabla} f(\vec{x}) = 0$) but it does not belong to Σ. Hence, it is not a solution of the constrained problem. This type of point is depicted as a white-filled circle in fig. C.1.

ii) A point $\vec{x}$ is a solution of the unconstrained problem (i.e. it satisfies $\vec{\nabla} f(\vec{x}) = 0$) and it belongs to Σ. Hence, it is also a solution of the constrained problem, since, in particular, the components of the gradient in directions that do not leave Σ are zero. This type of point is depicted as a gray-filled circle inside a white-filled circle in fig. C.1.

iii) A point $\vec{x}$ is not a solution of the unconstrained problem (i.e., one has $\vec{\nabla} f(\vec{x}) \neq 0$) but it belongs to Σ and the only non-zero components of the gradient are in directions that leave Σ. Hence, it is a solution of the constrained problem. This type of point is depicted as a gray-filled circle in fig. C.1.

From this discussion, it can be seen that, in principle, no conclusions about the number (or existence) of solutions of the constrained problem may be drawn only from the number of solutions of the unconstrained one. This must be investigated for each particular situation.

In fig. C.1, an schematic example in $\mathbb{R}^3$ is depicted. The constrained search space Σ is a 2-dimensional surface and the direction[119] in which one leaves Σ is shown at several points as a perpendicular vector $\vec{p}_\Sigma$. In such a case, the criterium that $\vec{\nabla} f$ has only components in the direction of leaving Σ may be rephrased by asking $\vec{\nabla} f$ to be parallel to $\vec{p}_\Sigma$, i.e., by requiring that there exists a number λ such that $\vec{\nabla} f = -\lambda \vec{p}_\Sigma$. The case $\lambda = 0$ is also admitted and the explanation of the minus sign will be given in the following.

In this case, $K = 1$, and one may note that the perpendicular vector $\vec{p}_\Sigma$ is precisely $\vec{p}_\Sigma = \vec{\nabla} L_1$. Let us define $\widetilde{f}$ as

$$\widetilde{f}(\vec{x}) := f(\vec{x}) + \lambda L_1(\vec{x}) \,. \tag{C.2}$$

It is clear that, requiring the gradient of $\widetilde{f}$ to be zero, one recovers the condition $\vec{\nabla} f = -\lambda \vec{p}_\Sigma$, which is satisfied by the points solution of the constrained stationary points problem. If one also asks that the derivative of $\widetilde{f}$ with respect to λ be zero, the constraint $L_1(\vec{x}) = 0$ that defines Σ is obtained as well.

This process illustrates the *Lagrange multipliers* method in this particular example. In the general case, described by eq. (C.1) and the paragraph above it, it can be proved that the points $\vec{x}$ which are stationary subject to the constraints imposed satisfy

$$\vec{\nabla} \widetilde{f}(\vec{x}) = 0 \quad \text{and} \quad \frac{\partial \widetilde{f}(\vec{x})}{\partial \lambda_i} = 0 \quad i = 1, \ldots, K \,, \tag{C.3}$$

where

$$\widetilde{f}(\vec{x}) := f(\vec{x}) + \sum_{i=1}^{K} \lambda_i L_i(\vec{x}) \,. \tag{C.4}$$

[119] Note that, only if $K = 1$, i.e., if the dimension of Σ is $N - 1$, there will be a vector perpendicular to the constrained space. For $K > 1$, the dimensionality of the vector space of the directions in which one 'leaves' Σ will be also larger than 1.

Of course, if one follows this method, the parameters λ_i (which are, in fact, the *Lagrange multipliers*) must also be determined and may be considered as part of the solution.

Also, it is worth remarking here that any two pair of functions, f_1 and f_2, of $\mathbb{R}^N$ whose restrictions to Σ are equal (i.e., that satisfy $f_1|_\Sigma = f_2|_\Sigma$) obviously represent the same constrained problem and they may be used indistinctly to construct the auxiliary function $\widetilde{f}$. This fact allows us, after having constructed $\widetilde{f}$ from a particular f, to use the equations of the constraints to change f by another simpler function which is equal to f when restricted to Σ. This freedom is used to derive the Hartree and Hartree-Fock equations, in secs. 2.6 and 2.7, respectively.

The formal generalization of these ideas to functionals (see appendix B) is straightforward if the space $\mathbb{R}^N$ is substituted by a functions space $\mathcal{F}$, the points $\vec{x}$ by functions, the functions f, L_i and $\widetilde{f}$ by functionals and the requirement that the gradient of a function be zero by the requirement that the functional derivative of the analogous functional be zero.

Finally, let us stress something that is rarely mentioned in the literature: *There is another (older) method, apart from the Lagrange multipliers one, for solving a constrained optimization problem: simple substitution.* I.e., if we can find a set of $N - K$ independent *adapted coordinates* that parameterize Σ and we can write the score function f in terms of them, we would be automatically satisfying the constraints. Actually, in practical cases, the method chosen is a suitable combination of the two; in such a way that, if substituting the constraints in f is difficult, the necessary Lagrange multipliers are introduced to force them, and vice versa.

As a good example of this, the reader may want to check the derivation of the Hartree equations in sec. 2.6 (or the Hartree-Fock ones in sec. 2.7). There, we start by proposing a particular form for the total wavefunction Φ in terms of the one-electron orbitals ϕ_i (see eq. (2.21)) and we write the functional $\mathcal{F}$ (which is the expected value of the energy) in terms of that special Φ (see eq. (2.23)). In a second step, we impose the constraints that the one-particle orbitals be normalized ($\langle\phi_i|\phi_i\rangle = 1, i = 1, \ldots, N$) and force them by means of N Lagrange multipliers λ_i. Despite the different treatments, both conditions are constraints standing on the same footing. The only difference is not conceptual, but operative: for the first condition, it would be difficult to write it as a constraint; while, for the second one, it would be difficult to define adapted coordinates in the subspace of normalized orbitals. So, in both cases, the easiest way for dealing with them is chosen.

D General mathematical argument for the factorization of external coordinates

In this appendix, we present a mathematical argument that shows that the factorization of the determinants of the mass-metric tensors G and g achieved in chapter 5 is a result of more general underlying geometrical properties and could have been expected a priori. Anyway, we would like to stress that we first found the explicit formulae in the two practical cases and then suspected that an argument such as the one herein presented must exist.

Let Ω be a finite dimensional differentiable manifold equipped with a riemannian metric tensor. Take local coordinates q^μ on Ω and denote by $G_{\mu\nu}(q)$ the components of the metric tensor in these coordinates.

The transformation

$$q'^\mu = q^\mu + \epsilon\,\xi^\mu(q) + O(\epsilon^2) \tag{D.1}$$

is said an *isometry* and $\xi^\mu(x)$ is said a *Killing vector field* if

$$G_{\mu\nu}(q'(q)) = J^\rho{}_\mu(q'(q))\, G_{\rho\sigma}(q)\, J^\sigma{}_\nu(q'(q))\,, \tag{D.2}$$

where

$$J^\mu{}_\nu(q'(q)) := \left(\frac{\partial q^\mu}{\partial q'^\nu}\right)(q'(q))\,. \tag{D.3}$$

Now, expanding eq. (D.2) up to first order in ϵ and noticing that $\det(J^\mu{}_\nu) = 1-\epsilon\,\partial_\mu\xi^\mu(q)$, we obtain the following differential equation for $\mathcal{G} := \det(G_{\mu\nu})$:

$$\xi^\mu(q)\partial_\mu\mathcal{G}(q) = -2(\partial_\mu\xi^\mu(q))\,\mathcal{G}(q)\,. \tag{D.4}$$

Let us apply this machinery to the case considered in this work. For concreteness, we shall derive the factorization of the external coordinates in the unconstrained case and shall argue that this still holds in the constrained one.

Simultaneous translations and rotations of all the particles[120] are isometries of the mass-matrix tensor in eq. (5.12). The important point for us is that, in the coordinates q^μ introduced in sec. 5.2, these transformations change the external coordinates $(X, Y, Z, \phi, \theta, \psi)$ and leave the internal coordinates q^a untouched (see eq. (5.7)).

A global translation is given in Euclidean coordinates by $x^p_\alpha \mapsto x^p_\alpha + \epsilon$. In the coordinates q^μ, it takes $(X, Y, Z) \mapsto (X, Y, Z) + \epsilon\,(1, 1, 1)$ and does not affect the remaining coordinates. With the above notation, $\xi^\mu = 1$, $\mu = 1, 2, 3$ and $\xi^\mu = 0$, $\forall\mu > 3$. Hence, eq. (D.4) implies that

$$\partial_X\mathcal{G} = \partial_Y\mathcal{G} = \partial_Z\mathcal{G} = 0\,, \tag{D.5}$$

i.e., the determinant of the mass-metric tensor does not depend on the coordinates X, Y, Z.

[120] Notice that the isometry group of the mass-metric tensor is much bigger, since translations and rotations acting independently on each particle are also isometry transformations.

A global rotation in the coordinates q^μ rotates (X, Y, Z) and changes the Euler angles (in a complicated way which will not be important for our purposes) but does not affect the internal coordinates. Hence, $\xi^\mu = 0, \forall \mu > 6$. In addition, the matrix $J^\mu_{\ \nu}$ does not depend on X, Y, Z because the rotation acts linearly on them. Let us abbreviate $\alpha \equiv \alpha^p \equiv (\phi, \theta, \psi)$, $p = 1, \ldots, 3$. Recalling that $\mathcal{G}$ does not depend on X, Y, Z, the differential equation (D.4) reads

$$\xi^p(\alpha)\partial_p \mathcal{G}(\alpha, q^a) = -2(\partial_p \xi^p(\alpha))\, \mathcal{G}(\alpha, q^a)\ . \tag{D.6}$$

The group of rotations in $\mathbb{R}^3$ has three linearly independent Killing vector fields which are complete in the sense that one can join two arbitrary points (ϕ, θ, ψ) and (ϕ', θ', ψ') by moving along integral curves of the Killing vector fields. This guarantees that the solution of eq. (D.6) is of the form

$$\mathcal{G}(\alpha, q^a) = \mathcal{G}_1(\alpha)\, \mathcal{G}_2(q^a) \tag{D.7}$$

and we have the desired result.

To derive the factorization of the external coordinates in the constrained case, simply notice that the constraints in chapter 5 do not involve the external coordinates. Therefore, global translations and rotations are still isometries of the reduced mass-metric tensor and the result follows.

E Model dipeptides. Notation and definitions

In the last decades, due to the exponential growth of computer power, quantum mechanical calculations in small organic molecules have become feasible. In this context, one of the most studied systems are the *model dipeptides* [163, 207–215, 411].

They are made up of a central amino acid residue with some additional chemical groups attached to the N- and C-termini. The intention is that dipeptides *model* residues in proteins (see sec. 1.2), so that the information provided by their study could be used to design effective polypeptide potentials that can tackle the protein folding problem. In this spirit, two of the substituents of the central α-carbon (apart from the α-hydrogen and the R-group) are peptide planes that end in the chemical groups *P* (in the N-terminus) and *Q* (in the C-terminus), so that, if the R-group is different from hydrogen, the model dipeptide is chiral just as free amino acids or residues in proteins. See fig. E.1a for a schematic representation of an L-model dipeptide with generic P-, Q- and R-groups.

The R-group may be chosen to be any of the 20 side chains belonging to the genetically encoded amino acids (see fig. 1.7), while the P- and Q-groups may independently be either a single hydrogen atom (H) or a methyl group (denoted by CH_3 or Me) [211]. This makes the part of the model dipeptide that simulates the backbone neutral (with a possible charge at the side chain of titratable residues; see sec. 1.2). The only charged amino and acid groups in the backbone of proteins may be those located at the termini (depending on the pH), therefore, much in the spirit of the 'modeling' aim stated above, this neutrality of model dipeptides more closely resembles the environment of a typical residue inserted in the middle of the chain.

Depending on the P- and Q-groups, the model dipeptides are named in many different ways in the literature, including spaces or not, including dashes or not, specifying the chirality or not, etc. [207, 208, 210–212, 215, 409, 412]. In this dissertation, we shall use a designation of the form *P-group-(L,D)-residue-Q-group*, where the P-group shall be *formyl* if P=H, or *acetyl* if P=Me, and the Q-group shall be *amide* if Q=H, or *methylamide*

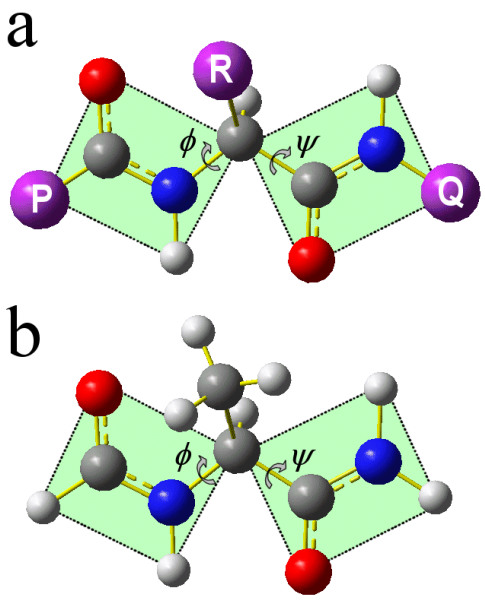


Figure E.1: Schematic depiction of model dipeptides indicating the soft Ramachandran internal coordinates ϕ and ψ. The color code used for the atoms is that in fig. 1.4. **(a)** L-dipeptide with both peptide bonds in the trans-conformation and with generic groups P-, Q- and R-. **(b)** Choosing P,Q=H and R=CH_3, we produce the model dipeptide HCO-L-Ala-NH_2 (formyl-L-alanine-amide), which is extensively investigated in this dissertation.

if Q=Me. More commonly, the notation *PCO-(L,D)-Res-NHQ*, closely resembling the chemical formula, shall be used, where *Res* stands for the three letter code of the residue that can be found in fig. 1.7).

According to these conventions, the model dipeptide that is extensively studied in this dissertation and that is depicted in fig. E.1b, is named *formyl-L-alanine-amide* and denoted by *HCO-L-Ala-NH*$_2$.

Additionally, note that there exists a certain ambiguity in the way of indicating the length oligopeptides. On the one hand, it seems reasonable to use the prefix or the number that corresponds to the quantity of amino acid residues [389]. On the other hand, the fact that the systems discussed in this appendix contain two peptide planes, has generalized the term *dipeptide* for naming them. Although in this dissertation we only deal with one-residue dipeptides, let us remark that the designation $(N+1)$-*peptide* shall denote *N-residue peptides* in future works.

Regarding their conformational behaviour, it is common to describe model dipeptides in terms of their Ramachandran angles ϕ and ψ, shown in fig. E.1, and assume that the rest of internal coordinates are either fixed or located at their constrained equilibrium values for each (ϕ,ψ)-pair (see sec. 1.2, as well as chapters 4, 5, 6 and 7 for further information about this issue). The plot of the so-called *potential energy surface* (PES), in terms of ϕ and ψ, is termed *Ramachandran map* and it is commonly used as a tool for studying the conformational preferences of dipeptides [207, 208, 215]. Figs. 4.9, 6.2 and 7.7 represent some Ramachandran maps calculated and studied in this dissertation.

The Ramachandran space spanned by the angles ϕ and ψ is, of course, periodic, and may be described using two different *cuts* shown in fig. E.2: the *topological* or *traditional* one, with the angles ranging from 0^o to 360^o, and the *standard* or *IUPAC* one, with the angles ranging from -180^o to 180^o [211, 411]. Although the topological cut presents some advantages regarding the position of the minima and it is sometimes used in the works about peptide systems [411], in this dissertation, *we have preferred the standard cut* for consistence with the general biochemical literature.

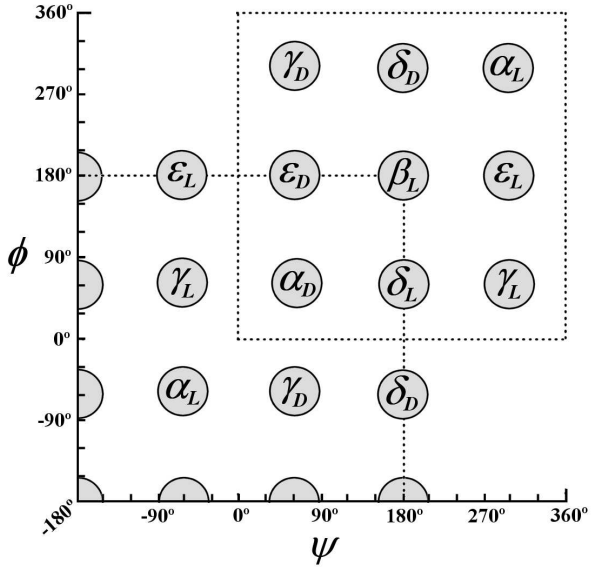


Figure E.2: Topological (or traditional) and standard (or IUPAC) cuts used for specifying the conformations of dipeptides in terms of the Ramachandran angles ϕ and ψ in the literature. The first goes from 0^o to 360^o, the second from -180^o to 180^o. Additionally, the nine ideal minima that are predicted using MDCA are shown using the notation in fig. E.3a. Note that they lie in the border of the standard cut, while they are located in the interior in the topological case.

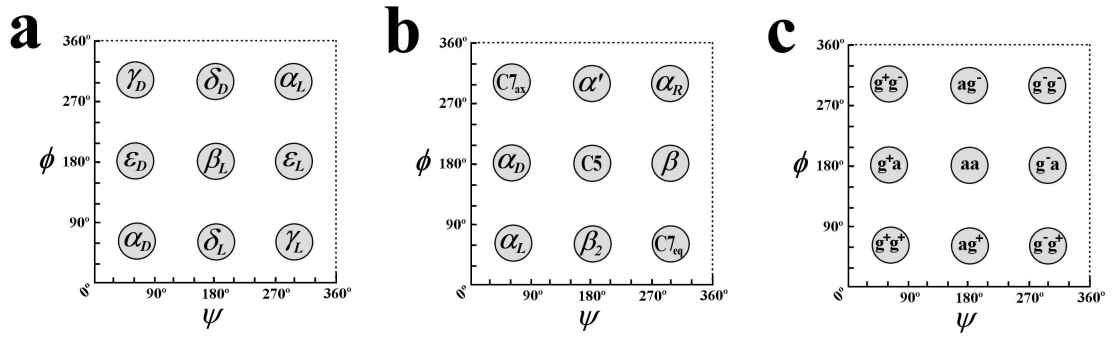


Figure E.3: Location and different notations of the nine ideal minima predicted by MDCA in the bidimensional Ramachandran space. **(a)** Notation inspired in topological issues. **(b)** Notation inspired in the secondary structure elements found in polypeptides. **(c)** Standard IUPAC notation for general molecules associated to the relative position of the substituents located at opposite sides of the same rotating bond.

Any angle φ_{std} in the standard cut may be transformed to one in the topological cut using that

$$\begin{aligned} \varphi_{\text{top}} &= 360^o + \varphi_{\text{std}} \qquad \text{if } \varphi_{\text{std}} < 0^o \,, \\ \varphi_{\text{top}} &= \varphi_{\text{std}} \qquad \text{if } \varphi_{\text{std}} \geq 0^o \,, \end{aligned} \tag{E.1}$$

and the inverse transformation is given by

$$\begin{aligned} \varphi_{\text{std}} &= \varphi_{\text{top}} \qquad \text{if } \varphi_{\text{top}} \leq 180^o \,, \\ \varphi_{\text{std}} &= \varphi_{\text{top}} - 360^o \qquad \text{if } \varphi_{\text{top}} > 180^o \,, \end{aligned} \tag{E.2}$$

Finally, if we approximate the bonds on which the Ramachandran angles are defined as ideal rotors with three equivalent minima (similar to the central bond of ethane (CH_3–CH_3), for example), we may hazard a guess and predict nine possible minima in the bidimensional Ramachandran space. This approach is called *multidimensional conformational analysis* (MDCA) [411] and the nine ideal minima are labeled according to different notations in the literature [207, 211]. In this dissertation, we shall mostly stick to the designation using subscripted Greek letters that appears in fig. E.2 and in fig. E.3a.

Bibliography

[1] J. L. Alonso and P. Echenique, *Relevant distance between two different instances of the same potential energy in protein folding*, Biophys. Chem. **115** (2004) 159–168. 86, 89

[2] J. L. Alonso, G. A. Chass, I. G. Csizmadia, P. Echenique, and A. Tarancón, *Do theoretical physicists care about the protein folding problem?*, in *Meeting on Fundamental Physics 'Alberto Galindo'*, edited by R. F. Álvarez-Estrada et al., Aula Documental, Madrid, 2004, (arXiv:q-bio.BM/0407024). 86, 93, 94, 115, 132, 151

[3] J. L. Alonso and P. Echenique, *A physically meaningful method for the comparison of potential energy functions*, J. Comp. Chem. **27** (2006) 238–252.

[4] P. Echenique and J. L. Alonso, *Definition of Systematic, Approximately Separable and Modular Internal Coordinates (SASMIC) for macromolecular simulation*, J. Comp. Chem. **27** (2006) 1076–1087.

[5] P. Echenique, I. Calvo, and J. L. Alonso, *Quantum mechanical calculation of the effects of stiff and rigid constraints in the conformational equilibrium of the Alanine dipeptide*, J. Comp. Chem. **27** (2006) 1748–1755.

[6] P. Echenique and I. Calvo, *Explicit factorization of external coordinates in constrained Statistical Mechanics models*, J. Comp. Chem. **27** (2006) 1733–1747.

[7] P. Echenique, J. L. Alonso, and I. Calvo, *Effects of constraints in general branched molecules: A quantitative ab initio study in HCO-L-Ala-NH_2*, in *From Physics to Biology. The Interface between Experiment and Computation. BIFI 2006 II International Congress*, edited by J. Clemente-Gallardo, Y. Moreno, J. F. Sáenz Lorenzo, and A. Velázquez-Campoy, volume 851, pp. 108–116, AIP Conference Proceedings, Melville, New York, 2006.

[8] P. Echenique and J. L. Alonso, *Efficient model chemistries for peptides. I. Split-valence Gaussian basis sets and the heterolevel approximation*, In progress, 2006.

[9] I. Ramsen, *Unsolved problems in chemistry*, in *Modern Inventions and Discoveries*, J. A. Hill and Co., New York, 1904. 1

[10] A. M. Lesk, *Introduction to Protein Architecture*, Oxford University Press, Oxford, 2001. 1, 4, 17, 18, 171

[11] Y. Cho, S. Gorina, P. D. Jeffrey, and N. P. Pavletich, *Crystal structure of a p53 tumor suppressor-DNA complex: Understanding tumorigenic mutations*, Science **265** (1994) 346–355. 2

[12] C. M. Dobson, *Protein-misfolding diseases: Getting out of shape*, Nature **729** (2002) 729. 2, 26

[13] J. W. Kelly, *Towards and understanding of amyloidgenesis*, Nat. Struct. Biol. **9** (2002) 323. 2

[14] E. H. Koo, P. T. Lansbury Jr., and J. W. Kelly, *Amyloid diseases: Abnormal protein aggregation in neurodegeneration*, Proc. Natl. Acad. Sci. USA **96** (1999) 9989–9990. 2

[15] M. H. Nielsen, F. S. Pedersen, and J. Kjems, *Molecular strategies to inhibit HIV-1 replication*, Retrovirology **2** (2005) 10. 2

[16] H. Ohtaka and E. Freire, *Adaptive inhibitors of the HIV-1 protease*, Prog. Biophys. Mol. Biol. **88** (2005) 193–208. 2

[17] U. Bacha, J. Barrila, A. Velázquez-Campoy, S. Leavitt, and E. Freire, *Identification of novel inhibitors of the SARS associated coronavirus main proteinase 3CLpro*, Biochemistry **43** (2004) 4906–4912. 2

[18] S. Venkatraman, F. G. Njoroge, V. M. Girijavallabhan, V. S. Madison, N. H. Yao, A. J. Prongay, N. Butkiewicz, and J. Pichardo, *Design and synthesis of depeptidized macrocyclic inhibitors of Hepatitis C NS3-4A protease using structure-based drug design*, J. Med. Chem. **48** (2005) 5088–5091. 2

[19] J. Poehlsgaard and S. Douthwaite, *The bacterial ribosome as a target for antibiotics*, Nat. Rev. Microbiol. **3** (2005) 870–881. 2

[20] C. Smith, *Drug target validation: Hitting the target*, Nature **422** (2003) 341–347. 2

[21] J. Scheller, K.-H. Gührs, F. Grosse, and U. Conrad, *Production of spider silk proteins in tobacco and potato*, Nat. Biotech. **19** (2001) 573–577. 2

[22] F. Vollrath and D. Porter, *Spider silk as archetypal protein elastomer*, Soft Matter **2** (2006) 377-385. 2

[23] C. M. Bellingham and F. W. Keeley, *Self-ordered polymerization of elastin-based biomaterials*, Curr. Opin. Sol. State Mat. Sci. **8** (2004) 135–139. 2

[24] A. Y. Wang, X. Mo, C. S. Chen, and S. M. Yu, *Facile modification of collagen directed by collagen mimetic peptides*, J. Am. Chem. Soc. **127** (2005) 4130–4131. 2

[25] S. A. Maskarinec and D. A. Tirrell, *Protein engineering approaches to biomaterials design*, Curr. Opin. Biotech. **16** (2005) 422–426. 2

[26] M. Strong, *Protein nanomachines*, PLoS Biology **2** (2004) 0305. 2

[27] Many authors, *What don't we know?*, Science **309** (2005) 78–102. 4

[28] L. D. Stein, *Human genome: End of the beginning*, Nature **431** (2004) 915–916. 4

[29] W. W. Gibbs, *The unseen genome: Gems among the junk*, Scientific American **289** (2003) 46–53. 5

[30] M. A. Nobrega, I. Ovcharenko, V. Afzal, and E. M. Rubin, *Scanning human gene deserts for long-range enhancers*, Science **302** (2003) 413. 5

[31] A. Woolfe, M. Goodson, D. K. Goode, P. Snell, G. K. McEwen, T. Vavouri, S. F. Smith, P. North, H. Callaway, K. Kelly, K. Walter, I. Abnizova, W. Gilks, Y. J. K. Edwards, J. E. Cooke, and G. Elgar, *Highly conserved non-coding sequences are associated with vertebrate development*, PLoS Biology **3** (2005) 0116. 5

[32] G. Gibson and S. V. Muse, *A Primer of Genome Science*, Sinauer, Sunderland, 2nd edition, 2004. 5

[33] C. Gómez-Moreno Calera and J. Sancho Sanz, editors, *Estructura de Proteínas*, Ariel ciencia, Barcelona, 2003. 5, 12, 22, 24, 31, 35, 36, 86, 177

[34] B. Boeckmann, A. Bairoch, R. Apweiler, M.-C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. 0'Donovan, I. Phan, S. Pilbout, and M. Schneider, *The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003*, Nucleic Acids Research **31** (2003) 365–370. 5

[35] S. Orchard, H. Hermjakob, and R. Apweiler, *Annotating the human proteome*, Mol. Cell. Proteomics **4** (2005) 435–440. 5

[36] D. A. Case, H. J. Dyson, and P. E. Wright, *Use of chemical shifts and coupling constants in nuclear magnetic resonance structural studies on peptides and proteins*, Methods Enzymol. **239** (1994) 392–416. 6, 135

[37] J. T. Pelton and L. R. McLean, *Spectroscopic methods for analysis of protein secondary structure*, Anal. Biochem. **277** (2000) 167–176. 6, 135

[38] J. Drenth, *Principles of Protein X-Ray Crystallography*, Springer-Verlag, New York, 1999. 6

[39] J. P. Glusker, *X-ray crystallography of proteins*, Methods Biochem. Anal. **37** (1994) 1–72. 6

[40] G. J. Kleywegt, *Validation of protein crystal structures*, Acta Crystallogr. D **56** (2000) 249–265. 6

[41] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, *The Protein Data Bank*, Nucleic Acids Research **28** (2000) 235–242. 6, 24, 104, 173

[42] Millennium Ecosystem Assessment. *Ecosystems and Human Well-being: Biodiversity Synthesis*, Washington, DC, 2005, http://www.millenniumassessment.org/en/Products.Synthesis.aspx. 6

[43] T. Castrigianò, P. D'Onorio De Meo, D. Cozzetto, I. G. Talamo, and A. Tramontano, *The PMDB Protein Model Database*, Nucleic Acids Research **34** (2006) 306–309. 6

[44] A. Tramontano, *Of men and machines*, Nat. Struct. Biol. **10** (2003) 87–90. 6, 25, 26

[45] M. D. S. Kumar, K. A. Bava, M. M. Gromiha, P. Prabakaran, K. Kitajima, H. Uedaira, and A. Sarai, *ProTherm and ProNIT: Thermodynamic databases for proteins and protein-nucleic acid interactions*, Nucleic Acids Research **34** (2006) 204–206. 6

[46] T.-Y. Lee, H.-D. Huang, J.-H. Hung, H.-Y. Huang, Y.-S. Yang, and T.-H. Wang, *dbPTM: An information repository of protein post-translational modification*, Nucleic Acids Research **34** (2006) 622–627. 7

[47] M. D. S. Kumar and M. M. Gromiha, *PINT: Protein-protein Interactions Thermodynamic Database*, Nucleic Acids Research **34** (2006) 195-198. 7

[48] R. Dawkins, *River Out of Eden: A Darwinian View of Life*, BasicBooks, New York, 1995. 7

[49] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery, Jr., T. Vreven, K. N. Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, P. Y. Ayala, K. Morokuma, G. A. Voth, P. Salvador, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich, A. D. Daniels, M. C.

Strain, O. Farkas, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul, S. Clifford, J. Cioslowski, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, C. Gonzalez, and J. A. Pople, *Gaussian 03, Revision C.02*, Gaussian, Inc., Wallingford, CT, 2004. 8, 111, 113, 114, 179

[50] R. S. Cahn, S. C. Ingold, and V. Prelog, *Specification of molecular chirality*, Angew. Chem. Int. Ed. **5** (1966) 385–415. 8, 12

[51] M. Klussmann, H. Iwamura, S. P. Mathew, D. H. Wells Jr., U. Pandya, A. Armstrong, and D. G. Blackmond, *Thermodynamic control of asymmetric amplification in amino acid catalysis*, Nature **441** (2006) 621–623. 8

[52] T. E. Creighton, *Proteins: Structures and Molecular Properties*, Freeman, W. H., New York, 2nd edition, 1992. 13, 14, 24, 31, 173

[53] M. Di Giulio, *The origin of the genetic code: Theories and their relationships, a review*, Biosystems **80** (2005) 175–184. 14

[54] R. D. Knight, S. J. Freeland, and L. F. Landweber, *Selection, history and chemistry: The three faces of the genetic code*, Trends Biochem. Sci. **24** (1999) 241–247. 14

[55] M. S. Weiss, A. Jabs, and R. Hilgenfeld, *Peptide bonds revisited*, Nat. Struct. Biol. **5** (1998) 676. 16

[56] J. F. Swain and L. M. Gierash, *A new twist for an Hsp70 chaperone*, Nat. Struct. Biol. **9** (2002) 406–408. 16

[57] V. I. Lim and A. S. Spirin, *Stereochemical analysis of ribosomal transpeptidation conformation of nascent peptide*, J. Mol. Biol. **188** (1986) 565–574. 16

[58] G. N. Ramachandran and C. Ramakrishnan, *Stereochemistry of polypeptide chain configurations*, J. Mol. Biol. **7** (1963) 95–99. 17

[59] L. Bragg, J. C. Kendrew, and M. F. Perutz, *Polypeptide chain configurations in crystalline proteins*, Proc. Roy. Soc. London Ser. A **203** (1950) 321–357. 18, 20

[60] L. Pauling, R. B. Corey, and H. R. Branson, *The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain*, Proc. Natl. Acad. Sci. USA **37** (1951) 205–211. 18, 20

[61] D. Eisenberg, *The discovery of the α-helix and β-sheet, the principal structural features of proteins*, Proc. Natl. Acad. Sci. USA **100** (2003) 11207–11210. 18

[62] J. C. Kendrew, G. Bodo, H. M. Dintzis, R. G. Parrish, H. Wyckoff, and D. C. Phillips, *A three-dimensional model of the myoglobin molecule obtained by x-ray analysis*, Nature **181** (1958) 662–666. 20

[63] M. F. Perutz, M. G. Rossman, A. F. Cullis, H. Muirhead, G. Will, and A. C. T. North, *Structure of haemoglobin: A three-dimensional Fourier synthesis at 5.5 Å resolution, obtained by x-ray analysis*, Nature **185** (1960) 416–422. 20

[64] E. N. Baker and R. E. Hubbard, *Hydrogen bonding in globular proteins*, Prog. Biophys. Mol. Biol. **44** (1984) 97–179. 20

[65] M. N. Fodje and S. Al-Karadaghi, *Occurrence, conformational features and amino acid propensities for the π-helix*, Protein Eng. **15** (2002) 353–358. 20

[66] M. E. Karpen, P. L. De-Haseth, and K. E. Neet, *Differences in the amino acid distributions of 3_{10}-helices and α-helices*, Prot. Sci. **1** (1992) 1333–1342. 20

[67] M. L. Higgins, *The structure of fibrous proteins*, Chem. Rev. **32** (1943) 195–218. 20

[68] B. W. Low and R. B. Baybutt, *The π helix – A hydrogen bonded configuration of the polypeptide chain*, J. Am. Chem. Soc. **74** (1952) 5806–5807. 20

[69] J. Donohue, *Hydrogen bonded helical configurations of the polypeptide chain*, Proc. Natl. Acad. Sci. USA **39** (1953) 470–478. 20

[70] R. V. Pappu and G. D. Rose, *A simple model for polyproline II structure in unfolded states of alanine-based peptides*, Prot. Sci. **11** (2002) 2437–2455. 20

[71] B. J. Stapley and T. P. Creamer, *A survey of left-handed polyproline II helices*, Prot. Sci. **8** (1999) 587–595. 20

[72] B. Zagrovic, J. Lipfert, E. J. Sorin, I. S. Millett, W. F. van Gunsteren, S. Doniach, and V. S. Pande, *Unusual compactness of a polyproline type II structure*, Proc. Natl. Acad. Sci. USA **102** (2005) 11698–11703. 20

[73] J. C. Kendrew, *Myoglobin and the structure of proteins*, Nobel Lecture, 1962, http://nobelprize.org/nobel_prizes/chemistry/laureates/1962/. 22

[74] C. B. Anfinsen, *Principles that govern the folding of protein chains*, Science **181** (1973) 223–230. 22, 27, 34, 177

[75] C. M. Dobson, *The nature and significance of protein folding*, in *Mechanism of Protein Folding*, edited by R. H. Pain, pp. 1–33, Oxford University Press, New York, 2000. 22, 23, 35

[76] C. M. Dobson, *Protein folding and misfolding*, Nature **426** (2003) 884–890. 22, 27, 35, 151

[77] J. Ellis, *Proteins as molecular chaperones*, Nature **328** (1987) 378–379. 22

[78] F. U. Hartl, *Molecular chaperones in cellular protein folding*, Nature **381** (2002) 571–580. 22

[79] F. U. Hartl and M. Hayer-Hartl, *Molecular chaperones in the cytosol: from nascent chain to folded protein*, Science **295** (2002) 1852–1858. 22

[80] A. L. Horwich, E. U. Weber-Ban, and D. Finley, *Chaperone rings in protein folding and degradation*, Proc. Natl. Acad. Sci. USA **96** (1999) 11033–11040. 22

[81] Y. Duan and P. A. Kollman, *Computational protein folding: From lattice to all-atom*, IBM Systems Journal **40** (2001) 297–309. 23

[82] C. Hardin, T. V. Pogorelov, and Z. Luthey-Schulten, *Ab initio protein structure prediction*, Curr. Opin. Struct. Biol. **12** (2002) 176–181. 23

[83] E. Krieger, S. B. Nabuurs, and G. Vriend, *Homology modeling*, in *Structural Bioinformatics*, edited by P. E. Bourne and H. Weissig, pp. 507–521, Wiley-Liss, 2003. 24

[84] K. Ginalski, N. V. Grishin, A. Godzik, and L. Rychlewski, *Practical lessons from protein structure prediction*, Nucleic Acids Research **33** (2005) 1874–1891. 24, 26, 36, 37, 177

[85] M. Jacobson and A. Sali, *Comparative protein structure modeling and its applications to drug discovery*, Ann. Rep. Med. Chem. **39** (2004) 259–276. 24, 26

[86] V. Daggett and A. Fersht, *The present view of the mechanism of protein folding*, Nat. Rev. Mol. Cell Biol. **4** (2003) 497. 24, 26

[87] B. Honig, *Protein folding: From the Levinthal paradox to structure prediction*, J. Mol. Biol. **293** (1999) 283–293. 24, 27, 32, 177

[88] D. Baker and A. Sali, *Protein structure prediction and structural genomics*, Science **294** (2001) 93–96. 24

[89] M. A. Marti-Renom, A. C. Stuart, A. Fiser, R. Sánchez, F. Melo, and A. Sali, *Comparative protein structure modeling of genes and genomes*, Annu. Rev. Biophys. Biomol. Struct. **29** (2000) 291–325. 24

[90] C. Chothia and A. M. Lesk, *The relation between the divergence of sequence and structure in proteins*, EMBO J. **5** (1986) 823–826. 24

[91] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and L. D. J., *Gapped BLAST and PSI-BLAST: A new generation of protein database search programs*, Nucleic Acids Research **25** (1997) 3389–33402. 24

[92] S. Henikoff, *Scores for sequence searches and alignments*, Curr. Opin. Struct. Biol. **6** (1996) 352–360. 24

[93] U. Pieper, N. Eswar, F. P. Davis, H. Braberg, M. S. Madhusudhan, A. Rossi, M. Marti-Renom, R. Karchin, B. M. Webb, D. Eramian, M.-Y. Shen, L. Kelly, F. Melo, and A. Sali, *MODBASE: A database of annotated comparative protein structure models and associated resources*, Nucleic Acids Research **34** (2006) 291–295. 24

[94] J. U. Bowie, R. Luthy, and D. Eisenberg, *A method to identify protein sequences that fold into a known three-dimensional structure*, Science **253** (1991) 164–170. 24

[95] J. Moult, K. Fidelis, B. Rost, T. Hubbard, and A. Tramontano, *Critical Assessment of Methods of Protein Structure Prediction (CASP)—Round 6*, PROTEINS: Struct. Funct. Bioinf. **7** (2005) 3–7. 24, 25

[96] C. A. Orengo and J. M. Thornton, *Protein families and their evolution—A structural perspective*, Annu. Rev. Biochem. **74** (2005) 867–900. 24

[97] P. Bradley, K. M. S. Misura, and D. Baker, *Toward high-resolution de novo structure prediction for small proteins*, Science **309** (2005) 1868–1871. 24

[98] O. Schueler-Furman, C. Wang, P. Bradley, K. Misura, and D. Baker, *Progress in modeling of protein structures and interactions*, Science **310** (2005) 638–642. 24, 36, 37, 177

[99] J. Moult, *A decade of CASP: Progress, bottlenecks and prognosis in protein structure prediction*, Curr. Opin. Struct. Biol. **15** (2005) 285–289. 25

[100] R. Bonneau and D. Baker, *Ab initio protein structure prediction: Progress and prospects*, Annu. Rev. Biophys. Biomol. Struct. **30** (2001) 173–189. 26, 36, 37, 177

[101] C. A. Rohl, C. E. Strauss, D. Chivian, and D. Baker, *Modeling structurally variable regions in homologous proteins with Rossetta*, PROTEINS: Struct. Funct. Bioinf. **55** (2004) 656–677. 26

[102] C. A. Rohl, C. E. Strauss, K. M. Misura, and D. Baker, *Protein structure prediction using Rossetta*, Methods Enzymol. **383** (2004) 66–93. 26

[103] M. Levitt and A. Warshel, *Computer simulation of protein folding*, Nature **253** (1975) 694–698. 26

[104] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, *CHARMM: A program for macromolecular energy, minimization, and dynamics calculations*, J. Comp. Chem. **4** (1983) 187–217. 26, 29, 36, 43, 86, 95, 104, 125, 152, 161, 177

[105] A. D. MacKerell Jr., B. Brooks, C. L. Brooks III, L. Nilsson, B. Roux, Y. Won, and M. Karplus, *CHARMM: The energy function and its parameterization with an overview of the program*, in *The Encyclopedia of Computational Chemistry*, edited by P. v. R. Schleyer et al., pp. 217–277, John Wiley & Sons, Chichester, 1998. 26, 29, 36, 43, 86, 95, 104, 125, 152, 161, 177

[106] D. Baker, *Prediction and design of macromolecular structures and interactions*, Phil. Trans. R. Soc. London B Biol. Sci. **361** (2006) 459–463. 26

[107] J. Skolnick, *Putting the pathway back into protein folding*, Proc. Natl. Acad. Sci. USA **102** (2005) 2265–2266. 26, 27, 37, 177

[108] M. A. Basharov, *Protein folding*, J. Cell. Mol. Med. **7** (2003) 223–237. 27

[109] B. Hardesty and G. Kramer, *Folding of a nascent peptide on the ribosome*, Prog. Nucleic Acid Res. Mol. Biol. **66** (2001) 41–66. 27

[110] A. R. Fersht and V. Daggett, *Protein folding and unfolding at atomic resolution*, Cell **108** (2002) 573–582. 27

[111] V. Daggett and A. R. Fersht, *Is there a unifying mechanism for protein folding?*, Trends Biochem. Sci. **28** (2003) 18–25. 27, 177

[112] C. Levinthal, *Are there pathways for protein folding?*, J. Chim. Phys. **65** (1968) 44–45. 27, 33, 177

[113] T. Lazaridis and M. Karplus, *Thermodynamics of protein folding: a microscopic view*, Biophys. Chem. **100** (2003) 367–395. 28, 29, 30, 31, 34, 35, 36, 127

[114] W. Greiner, H. Stocker, and L. Neise, *Thermodynamics and Statistical Mechanics*, Classical Theoretical Physics, Springer, New York, 2004. 28, 30

[115] B. A. Dubrovin, A. T. Fomenko, and S. P. Novikov, *Modern Geometry — Methods and Applications*, volume I. The Geometry of Surfaces, Transformation Groups and Fields, Springer, New York, 1984. 28

[116] T. Lazaridis and M. Karplus, *Effective energy function for proteins in solution*, PROTEINS: Struct. Funct. Gen. **35** (1999) 133–152. 29, 218

[117] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, J. Merz, K. M., D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman, *A second generation force field for the simulation of proteins, nucleic acids, and organic molecules*, J. Am. Chem. Soc. **117** (1995) 5179–5197. 29, 36, 125, 152, 161, 177

[118] T. A. Halgren, *Merck Molecular Force Field. I. Basis, form, scope, parametrization, and performance of MMFF94*, J. Comp. Chem. **17** (1996) 490–519. 29, 36, 125, 152, 161, 177

[119] T. A. Halgren, *Merck Molecular Force Field. II. MMFF94 van der Waals and electrostatica parameters for intermolecular interactions*, J. Comp. Chem. **17** (1996) 520–552. 29, 36, 152, 161, 177

[120] T. A. Halgren, *Merck Molecular Force Field. III. Molecular geometrics and vibrational frequencies for MMFF94*, J. Comp. Chem. **17** (1996) 553–586. 29, 36, 152, 161, 177

[121] T. A. Halgren, *Merck Molecular Force Field. IV. Conformational Energies and Geometries for MMFF94*, J. Comp. Chem. **17** (1996) 587–615. 29, 36, 152, 161, 177

[122] T. A. Halgren, *Merck Molecular Force Field. V. Extension of MMFF94 using experimental data, additional computational data, and empirical rules*, J. Comp. Chem. **17** (1996) 616–641. 29, 36, 152, 161, 177

[123] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives, *Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids*, J. Am. Chem. Soc. **118** (1996) 11225–11236. 29, 36, 125, 152, 161, 177

[124] W. L. Jorgensen and J. Tirado-Rives, *The OPLS potential functions for proteins. Energy minimization for crystals of cyclic peptides and Crambin*, J. Am. Chem. Soc. **110** (1988) 1657–1666. 29, 36, 43, 152, 161, 177

[125] D. A. Pearlman, D. A. Case, J. W. Caldwell, W. R. Ross, T. E. Cheatham III, S. DeBolt, D. Ferguson, G. Seibel, and P. Kollman, *AMBER, a computer program for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to elucidate the structures and energies of molecules*, Comp. Phys. Commun. **91** (1995) 1–41. 29, 36, 43, 152, 161, 177

[126] W. F. Van Gunsteren and M. Karplus, *Effects of constraints on the dynamics of macromolecules*, Macromolecules **15** (1982) 1528–1544. 29, 36, 151, 152, 157, 161, 177

[127] V. I. Arnold, *Mathematical Methods of Classical Mechanics*, Graduate Texts in Mathematics, Springer, New York, 2nd edition, 1989. 30, 155

[128] K. A. Dill and H. S. Chan, *From Levinthal to pathways to funnels: The "new view" of protein folding kinetics*, Nat. Struct. Biol. **4** (1997) 10–19. 30, 33

[129] A. Fersht, *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding*, Freeman, W. H., New York, 1998. 31

[130] C. M. Dobson, A. Šali, and M. Karplus, *Protein folding: A perspective from theory and experiment*, Angew. Chem. Int. Ed. **37** (1998) 868–893. 32, 34, 94

[131] E. Helfand, *Flexible vs. rigid constraints in Statistical Mechanics*, J. Chem. Phys. **71** (1979) 5000. 32, 151, 152, 153, 159, 161

[132] M. R. Pear and J. H. Weiner, *Brownian dynamics study of a polymer chain of linked rigid bodies*, J. Chem. Phys. **71** (1979) 212. 32, 152, 153, 154, 159, 160, 161

[133] D. Perchak, J. Skolnick, and R. Yaris, *Dynamics of rigid and flexible constraints for polymers. Effect of the Fixman potential*, Macromolecules **18** (1985) 519–525. 32, 137, 152, 153, 154, 159, 160, 161

[134] S. Reich, *Smoothed Langevin dynamics of highly oscillatory systems*, Physica D **118** (2000) 210–224. 32, 151, 152, 153, 159

[135] T. Lazaridis and M. Karplus, *"New view" of protein folding reconciled with the old through multiple unfolding simulations*, Science **278** (1997) 1928–1931. 32, 34

[136] T. Lazaridis and M. Karplus, *Discrimination of the native from misfolded protein models with an energy function including implicit solvation*, J. Mol. Biol. **288** (1999) 477–487. 32

[137] K. A. Dill, *Polymer principles and protein folding*, Prot. Sci. **8** (1999) 1166–1180. 32, 34, 86, 93, 94, 151

[138] C. Levinthal, *How to fold graciously*, in *Mossbauer Spectroscopy in Biological Systems*, edited by J. T. P. DeBrunner and E. Munck, pp. 22–24, Allerton House, Monticello, Illinois, 1969, University of Illinois Press. 32, 35, 151

[139] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, *Funnels, pathways, and the energy landscape of protein folding: A synthesis*, Proteins **21** (1995) 167–195. 33, 34

[140] H. S. Chan and K. A. Dill, *Protein folding in the landscape perspective: Chevron plots and non-Arrhenius kinetics*, PROTEINS: Struct. Funct. Gen. **30** (1998) 2–33. 33

[141] R. L. Baldwin, *The nature of protein folding pathways: The classical versus the new view*, J. Biomol. NMR **5** (1995) 103–109. 34

[142] J. D. Bryngelson and P. G. Wolynes, *Spin-glasses and the statistical-mechanics of protein folding*, Proc. Natl. Acad. Sci. USA **84** (1987) 7524–7528. 34, 36

[143] J. N. Onuchic and P. G. Wolynes, *Theory of protein folding*, Curr. Opin. Struct. Biol. **14** (2004) 70–75. 34, 36

[144] S. S. Plotkin and J. Onuchic, *Understanding protein folding with energy landscape theory. Part I: Basic concepts*, Quart. Rev. Biophys. **35** (2002) 111–167. 34

[145] R. Day and V. Daggett, *Ensemble versus single-molecule protein unfolding*, Proc. Natl. Acad. Sci. USA **102** (2005) 13445–1450. 34

[146] S. E. Radford and C. M. Dobson, *Insights into protein folding using physical techniques: Studies of lysozyme and alpha-lactalbumin*, Phil. Trans. R. Soc. London B Biol. Sci. **348** (1995) 17–25. 34

[147] C. D. Snow, H. Nguyen, V. S. Pande, and M. Gruebele, *Absolute comparison of simulated and experimental protein-folding dynamics*, Nature **420** (2002) 102–106. 34

[148] A. A. Deniz, T. A. Laurence, G. S. Beligere, M. Dahan, A. B. Martin, D. S. Chemla, P. E. Dawson, P. G. Schultz, and S. Weiss, *Single-molecule protein folding: Diffusion fluorescence resonance energy transfer studies of the denaturation of chymotrypsin inhibitor 2*, Proc. Natl. Acad. Sci. USA **97** (2000) 5179–5184. 34

[149] D. Baker, *Metastable states and folding free energy barriers*, Nat. Struct. Biol. **5** (1998) 1021–1034. 35

[150] J. L. Sohl, S. S. Jaswal, and D. A. Agard, *Unfolded conformations of α-lytic protease are more stable than its native state*, Nature **395** (1998) 817–819. 35

[151] V. Cerny, *A thermodynamical approach to the travelling salesman problem: An efficient simulation algorithm*, J. Optimiz. Theory App. **45** (1985) 41–51. 35, 46

[152] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, *Optimization by simulated annealing*, Science **220** (1983) 671–680. 35, 46

[153] R. Dawkins, *The Blind Watchmaker: Why the Evidence of Evolution Reveals a Universe without Design*, W. W. Norton & Company, New York, 1987. 35

[154] N. Gō and H. Taketomi, *Respective roles of short- and long-range interactions in protein folding*, Proc. Natl. Acad. Sci. USA **75** (1978) 559–563. 36

[155] G. I. Makhatadze and P. L. Privalov, *Energetics of protein structure*, Adv. Prot. Chem. **47** (1995) 307–425. 36

[156] R. A. Abagyan, *Protein structure prediction by global energy optimization*, in *Computer Simulations of Biomolecular Systems*, edited by W. F. Van Gunsteren, volume 3, Kluwer academic publishing, Dordrecht, 1997. 36

[157] P. Derreumaux, *Ab initio polypeptide structure prediction*, Theo. Chem. Acc. **104** (2000) 1–6. 36

[158] M. Karplus and J. A. McCammon, *Molecular dynamics simulations of biomolecules*, Nat. Struct. Biol. **9** (2002) 646–652. 36, 177

[159] A. R. MacKerell Jr., M. Feig, and C. L. Brooks III, *Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations*, J. Comp. Chem. **25** (2004) 1400–1415. 36, 37, 125, 154, 171, 174, 177

[160] A. V. Morozov, T. Kortemme, K. Tsemekhman, and D. Baker, *Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum mechanical calculations*, Proc. Natl. Acad. Sci. USA **101** (2004) 6946–6951. 36, 37, 177

[161] C. D. Snow, E. J. Sorin, Y. M. Rhee, and V. S. Pande, *How well can simulation predict protein folding kinetics and thermodynamics?*, Annu. Rev. Biophys. Biomol. Struct. **34** (2005) 43–69. 36, 37, 177

[162] M.-H. Hao and H. A. Scheraga, *Designing potential energy functions for protein folding*, Curr. Opin. Struct. Biol. **9** (1999) 184–188. 37

[163] M. Beachy, D. Chasman, R. Murphy, T. Halgren, and R. Friesner, *Accurate ab initio quantum chemical determination of the relative energetics of peptide conformations and assessment of empirical force fields*, J. Am. Chem. Soc. **119** (1997) 5908–5920. 37, 66, 78, 154, 171, 174, 177, 178, 179, 191, 197, 201, 214, 227

[164] A. J. Bordner, C. N. Cavasotto, and R. A. Abagyan, *Direct derivation of van der Waals force fields parameters from quantum mechanical interaction energies*, J. Phys. Chem. B **107** (2003) 9601–9609. 37, 78, 154, 171, 174, 197

[165] R. A. Friesner and M. D. Beachy, *Quantum mechanical calculations on biological systems*, Curr. Opin. Struct. Biol. **8** (1998) 257–262. 37

[166] F. Jensen, *An introduction to the state of the art in quantum chemistry*, Ann. Rep. Comp. Chem. **1** (2005) 1–17. 37, 78, 197

[167] A. V. Morozov, K. Tsemekhman, and D. Baker, *Electron density redistribution accounts for half the cooperativity of α-helix formation*, J. Phys. Chem. B **110** (2006) 4503–4505. 37

[168] C. J. Barden and H. F. Schaffer III, *Quantum chemistry in the* 21^{st} *century*, Pure Appl. Chem. **72** (2000) 1405–1423. 37

[169] J. Simons, *An experimental chemists's guide to ab initio quantum chemistry*, J. Chem. Phys. **95** (1991) 1017–1029. 37

[170] B. N. Taylor, *Guide for the Use of the International System of Units (SI)*, NIST Special Publication 811, National Institute of Standards and Technology, 1995. 38

[171] D. R. Hartree, *The wave mechanics of an atom with a non-Coulomb central field. Part I. Theory and methods*, Proc. Camb. Philos. Soc. **24** (1927) 89. 38, 47

[172] H. Shull and G. G. Hall, *Atomic units*, Nature **184** (1959) 1559. 38

[173] M. Born and K. Huang, *Dynamical Theory of Crystal Lattices*, chapter Appendices VII and VIII, Oxford University Press, London, 1954. 40

[174] M. Born and J. R. Oppenheimer, *Zur Quantentheorie der Molekeln*, Ann. Phys. Leipzig **84** (1927) 457–484. 40

[175] M. P. Marder, *Condensed Matter Physics*, Wiley-Interscience, New York, 2000. 41

[176] C. J. Cramer, *Essentials of Computational Chemistry: Theories and Models*, John Wiley & Sons, Chichester, 2nd edition, 2002. 41, 111, 136

[177] F. Jensen, *Introduction to Computational Chemistry*, John Wiley & Sons, Chichester, 1998. 41, 67, 69, 71, 72, 78, 79, 81, 82, 83, 84, 183, 188, 197, 199

[178] R. G. Parr and W. Yang, *Density-Functional Theory of Atoms and Molecules*, volume 16 of *International series of monographs on chemistry*, Oxford University Press, New York, 1989. 41

[179] T. Shida, *The Chemical Bond: A Fundamental Quantum-Mechanical Picture*, Springer Series in Chemical Physics, Springer-Verlag, Berlin, 2006. 41

[180] A. Szabo and N. S. Ostlund, *Modern Quantum Chemistry: Introduced to Advanced Electronic Structure Theory*, Dover Publications, New York, 1996. 41, 65, 69, 78, 79, 81, 82, 83

[181] B. T. Sutcliffe, *The coupling of nuclear and electronic motions in molecules*, J. Chem. Soc. Faraday Trans. **89** (1993) 2321–2335. 41

[182] B. T. Sutcliffe, *The nuclear motion problem in molecular physics*, Adv. Quantum Chem. **28** (1997) 65–80. 41

[183] B. T. Sutcliffe and R. G. Woolley, *Molecular structure calculations without clamping the nuclei*, Phys. Chem. Chem. Phys. **7** (2005) 3664–3676. 41

[184] G. Hunter, *Conditional probability amplitudes in wave mechanics*, Intl. J. Quant. Chem. **9** (1975) 237–242. 42

[185] W. Hunziger and I. M. Sigal, *The quantum N-body problem*, J. Math. Phys. **41** (2000) 3448–3510. 42

[186] W. Hunziker, *On the spectra of Schrödinger multiparticle Hamiltonians*, Helv. Phys. Acta **39** (1966) 451–462. 42

[187] M. B. Ruskai and J. P. Solovej, *Asymptotic neutrality of polyatomic molecules*, Lect. Notes Phys. **403** (1992) 153–174. 42

[188] B. Simon, *Schrödinger operators in the twentieth century*, J. Math. Phys. **41** (2000) 3523–3555. 42

[189] C. Van Winter, *Theory of finite systems of particles I. The Green function*, Mat. Fys. Skr. Dan. Vid. Selsk **2** (1964) 1–60. 42

[190] H. Yserentant, *On the electronic Schrödinger equation*, Technical report, Universität Tübingen, 2003, http://www.math.tu-berlin.de/~yserenta/. 42

[191] G. Friesecke, *The multiconfiguration equations for atoms and molecules: Charge quantization and existence of solutions*, Arch. Rational Mech. Anal. **169** (2003) 35–71. 42

[192] G. M. Zhislin, *Discussion of the spectrum of Schrödinger operators for systems of many particles*, Trudy Moskovskogo matematiceskogo obscestva **9** (1960) 81–120, (In Russian). 42

[193] H. P. Hratchian and H. B. Schlegel, *Finding minima, transition states, and following reaction pathways on ab initio potential energy surfaces*, in *Theory and Applications of Computational Chemistry: The First Forty Years*, edited by C. D. et al., chapter 10, Elsevier, 2005. 43, 154

[194] T. E. Cheatham III and M. A. Young, *Molecular dynamics simulation of nucleic acids: Successes, limitations and promise*, Biopolymers **56** (2001) 232–256. 43

[195] J. W. Ponder and D. A. Case, *Force fields for protein simulations*, Adv. Prot. Chem. **66** (2003) 27–85. 43, 125

[196] P. A. M. Dirac, *Quantum Mechanics of Many-Electron Systems*, Proc. Roy. Soc. London **123** (1929) 714. 45, 47

[197] R. Seeger and J. A. Pople, *Self-consistent molecular orbital methods. XVIII. Constraints and stability in Hartree-Fock theory*, J. Chem. Phys. **66** (1977) 3045–3050. 46, 51, 61

[198] E. Marinari and G. Parisi, *Simulated tempering: A new Monte Carlo scheme*, Europhys. Lett. **19** (1992) 451–458. 46

[199] J. C. Slater, *Note on Hartree's method*, Phys. Rev. **35** (1930) 210–211. 47, 52

[200] E. Cancès, M. DeFranceschi, W. Kutzelnigg, C. Le Bris, and Y. Maday, *Computational quantum chemistry: A primer*, in *Handbook of numerical analysis. Volume X: Special volume: Computational chemistry*, edited by P. Ciarlet and C. Le Bris, pp. 3–270, Elsevier, 2003. 50, 51, 61, 62, 63, 68, 71

[201] P. L. Lions, *Solutions of Hartree-Fock equations for Coulomb systems*, Commun. Math. Phys. **109** (1987) 33–97. 51

[202] E. H. Lieb and B. Simon, *On solutions to the Hartree-Fock problem for atoms and molecules*, J. Chem. Phys. **61** (1974) 735–736. 51, 61

[203] E. H. Lieb and B. Simon, *The Hartree-Fock theory for Coulomb systems*, Commun. Math. Phys. **53** (1977) 185–194. 51, 61

[204] V. Fock, *Näherungsmethode zur Lösung des quantenmechanischen Mehrkörperproblems*, Z. Phys. **61** (1930) 126. 52

[205] T. Koopmans, *Über die Zuordnung von Wellenfunktionen und Eigenwerten zu den einzelnen Electronen eines Atoms*, Physica **1** (1934) 104. 60

[206] H. B. Schlegel and J. J. W. McDouall, *Do you have SCF stability and convergence problems?*, in *Computational Advances in Organic Chemistry: Molecular Structure and Reactivity*, edited by C. Ögretir and I. G. Csizmadia, pp. 167–185, Kluwer Academic, The Netherlands, 1991. 62, 63, 64, 65

[207] A. Perczel, Ö. Farkas, I. Jákli, I. A. Topol, and I. G. Csizmadia, *Peptide models. XXXIII. Extrapolation of low-level Hartree-Fock data of peptide conformation to large basis set SCF, MP2, DFT and CCSD(T) results. The Ramachandran surface of alanine dipeptide computed at various levels of theory*, J. Comp. Chem. **24** (2003) 1026–1042. 66, 78, 85, 87, 96, 107, 125, 129, 137, 154, 171, 177, 178, 184, 198, 207, 227, 228, 229

[208] A. M. Rodríguez, H. A. Baldoni, F. Suvire, R. Nieto Vázquez, G. Zamarbide, R. D. Enriz, Ö. Farkas, A. Perczel, M. A. McAllister, L. L. Torday, J. G. Papp, and I. G. Csizmadia, *Characteristics of Ramachandran maps of L-alanine diamides as computed by various molecular mechanics, semiempirical and ab initio MO methods. A search for primary standard of peptide conformational stability*, J. Mol. Struct. **455** (1998) 275–301. 66, 78, 125, 126, 177, 227, 228

[209] R. F. Frey, J. Coffin, S. Q. Newton, M. Ramek, V. K. W. Cheng, F. A. Momany, and L. Schäfer, *Importance of correlation-gradient geometry optimization for molecular conformational analyses*, J. Am. Chem. Soc. **114** (1992) 5369–5377. 66, 78, 112, 113, 177, 227

[210] C.-H. Yu, M. A. Norman, L. Schäfer, M. Ramek, A. Peeters, and C. van Alsenoy, *Ab initio conformational analysis of N-formyl L-alanine amide including electron correlation*, J. Mol. Struct. **567–568** (2001) 361–374. 66, 113, 125, 126, 137, 154, 171, 177, 198, 227

[211] A. G. Császár and A. Perczel, *Ab initio characterization of building units in peptides and proteins*, Prog. Biophys. Mol. Biol. **71** (1999) 243–309. 66, 137, 154, 171, 177, 227, 228, 229

[212] A. Láng, I. G. Csizmadia, and A. Perczel, *Peptide models. XLV: Conformational properties of N-formyl-L-methioninamide ant its relevance to methionine in proteins*, PROTEINS: Struct. Funct. Bioinf. **58** (2005) 571–588. 66, 137, 154, 171, 177, 227

[213] R. Vargas, J. Garza, B. P. Hay, and D. A. Dixon, *Conformational study of the alanine dipeptide at the MP2 and DFT levels*, J. Phys. Chem. A **106** (2002) 3213–3218. 66, 137, 154, 171, 177, 227

[214] M. Elstner, K. J. Jalkanen, M. Knapp-Mohammady, and S. Suhai, *Energetics and structure of glycine and alanine based model peptides: Approximate SCC-DFTB, AM1 and PM3 methods in comparison with DFT, HF and MP2 calculations*, Chem. Phys. **263** (2001) 203–219. 66, 171, 177, 227

[215] H. A. Baldoni, G. Zamarbide, R. D. Enriz, E. A. Jauregui, Ö. Farkas, A. Perczel, S. J. Salpietro, and I. G. Csizmadia, *Peptide models XXIX. Cis-trans iosmerism of peptide bonds: Ab initio study on small peptide model compound; the 3D-Ramachandran map of formyl-glycinamide*, J. Mol. Struct. **500** (2000) 97–111. 66, 177, 214, 227, 228

[216] J. Kobus, *Diatomic molecules: Exact solutions of HF equations*, Adv. Quantum Chem. **28** (1997) 1–14. 66

[217] C. C. J. Roothaan, *New developments in molecular orbital theory*, Rev. Mod. Phys. **23** (1951) 69–89. 66, 67, 68

[218] G. G. Hall, *The molecular orbital theory of chemical valency VIII. A method of calculating ionization potentials*, Proc. Roy. Soc. London Ser. A **205** (1951) 541–552. 66, 68

[219] F. Jensen, *Estimating the Hartree-Fock limit from finite basis set calculations*, Theo. Chem. Acc. **113** (2005) 267–273. 67, 178

[220] J. A. Pople, *Nobel lecture: Quantum chemical models*, Rev. Mod. Phys. **71** (1999) 1267–1274. 67, 70, 71, 178, 205

[221] J. M. García de la Vega and B. Miguel, *Basis sets for computational chemistry*, in *Introduction to Advanced Topics of Computational Chemistry*, edited by L. A. Montero, L. A. Díaz, and R. Bader, chapter 3, pp. 41–80, Editorial de la Universidad de la Habana, 2003. 69

[222] T. Helgaker and P. R. Taylor, *Gaussian basis sets and molecular integrals*, in *Modern Electronic Structure Theory. Part II*, edited by D. R. Yarkony, pp. 725–856, World Scientific, Singapore, 1995. 69

[223] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, Dover, New York, 9 edition, 1964. 70, 71

[224] J. C. Slater, *Atomic shielding constants*, Phys. Rev. **36** (1930) 57–54. 70

[225] C. Zener, *Analytic atomic wave functions*, Phys. Rev. **36** (1930) 51–56. 70

[226] R. J. Mathar, *Mutual conversion of three flavors of Gaussian Type Orbitals*, Intl. J. Quant. Chem. **90** (2002) 227–243. 71, 72, 73

[227] T. Kato, *On the eigenfunctions of many-particle systems in quantum mechanics*, Commun. Pure Appl. Math. **10** (1957) 151–177. 71

[228] S. F. Boys, *Electronic wavefunctions. I. A general method of calculation for stationary states of any molecular system*, Proc. Roy. Soc. London Ser. A **200** (1950) 541–554. 72

[229] H. B. Schlegel and M. J. Frisch, *Transformation between Cartesian and pure spherical harmonic Gaussians*, Intl. J. Quant. Chem. **54** (1995) 83–87. 72, 73

[230] W. J. Hehre, R. F. Stewart, and J. A. Pople, *Self-consistent molecular-orbital methods. I. Use of Gaussian expansions of Slater-type atomic orbitals*, J. Chem. Phys. **51** (1969) 2657–2664. 73, 74, 75

[231] R. Ditchfield, W. J. Hehre, and J. A. Pople, *Self-consistent molecular-orbital methods. IX. An extended Gaussian-type basis for molecular-orbital studies of organic molecules*, J. Chem. Phys. **54** (1971) 724–728. 73, 77, 179, 182

[232] W. J. Hehre, R. Ditchfield, and J. A. Pople, *Self-consistent molecular-orbital methods. XII. Further extensions of Gaussian-type basis sets for use in molecular-orbital studies of organic molecules*, J. Chem. Phys. **56** (1972) 2257–2261. 73, 76, 77, 179, 182

[233] P. C. Hariharan and J. A. Pople, *The influence of polarization functions on molecular orbital hydrogenation energies*, Theor. Chim. Acta **28** (1973) 213–222. 73, 77, 179, 182, 184

[234] M. J. Frisch, J. A. Pople, and J. S. Binkley, *Self-consistent molecular-orbital methods. 25. Supplementary functions for Gaussian basis sets*, J. Chem. Phys. **80** (1984) 3265–3269. 73, 77, 179, 182, 184

[235] R. Krishnan, J. S. Binkley, R. Seeger, and J. A. Pople, *Self-consistent molecular-orbital methods. XX. A basis set for correlated wave functions*, J. Chem. Phys. **72** (1980) 650–654. 73, 77, 179, 182

[236] J. S. Binkley, J. A. Pople, and W. J. Hehre, *Self-consistent molecular-orbital methods. 21. Small split-valence basis sets for first-row elements*, J. Am. Chem. Soc. **102** (1980) 939–947. 73, 77, 179, 182

[237] G. W. Spitznagel, T. Clark, J. Chandrasekhar, and P. v. R. Schleyer, *Stabilization of methyl anions by first row substituents. The superiority of diffuse function-augmented basis sets for anion calculations*, J. Comp. Chem. **3** (1982) 363–371. 73, 77, 179, 182, 184

[238] T. Clark, J. Chandrasekhar, G. W. Spitznagel, and P. v. R. Schleyer, *Efficient diffuse function-augmented basis sets for anion calculations. III. The 3-21+G basis set for first-row elements Li–F*, J. Comp. Chem. **4** (1983) 294–301. 73, 77, 179, 182, 184

[239] G. K. Woodgate, *Elementary Atomic Structure*, Oxford University Press, USA, 2 edition, 1983. 75

[240] A. P. Scott and L. Radom, *Harmonic vibrational frequencies: An evaluation of Hartree-Fock, Møller-Plesset, Quadratic Configuration Interaction, Density Functional Theory, and semiempirical scale factors*, J. Phys. Chem. **100** (1996) 16502–16513. 78, 171, 214

[241] M. D. Halls and H. B. Schlegel, *Comparison study of the prediction of Raman intensities using electronic structure methods*, J. Chem. Phys. **111** (1999) 8819–8824. 78, 197, 214

[242] A. G. Császár, *On the structures of free glycine and α-alanine*, J. Mol. Struct. **346** (1995) 141–152. 78, 179, 184

[243] P. Hudáky, I. Jákli, A. G. Császár, and A. Perczel, *Peptide models. XXXI. Conformational properties of hydrophobic residues shaping the core of proteins. An ab initio study of N-formyl-L-valinamide and N-formyl-L-phenylalaninamide*, J. Comp. Chem. **22** (2001) 732–751. 78, 179, 184

[244] A. St.-Amant, W. D. Cornell, P. A. Kollman, and T. A. Halgren, *Calculation of molecular geometries relative conformational energies, dipole moments, and molecular electrostatic potential fitted charges of small organic molecules of biochemical interest by Density Functional Theory*, J. Comp. Chem. **12** (1995) 1483–1506. 78, 197

[245] C. P. Chun, A. A. Connor, and G. A. Chass, *Ab initio conformational analysis of N- and C-terminally-protected valyl-alanine dipeptide model*, J. Mol. Struct. **729** (2005) 177–184. 78, 177

[246] P. Hobza and J. Šponer, *Toward true DNA base-stacking energies: MP2, CCSD(T) and Complete Basis Set calculations*, J. Am. Chem. Soc. **124** (2002) 11802–11808. 78, 178, 179, 191, 197, 201

[247] T. Van Mourik, P. G. Karamertzanis, and S. L. Price, *Molecular conformations and relative stabilities can be as demanding of the electronic structure method as intermolecular calculations*, J. Phys. Chem. **110** (2006) 8–12. 78, 197, 214

[248] P. Knowles, M. Schütz, and H.-J. Werner, *Ab initio nethods for electron correlation in molecules*, in *Modern Methods and Algorithms of Quantum Chemistry*, edited by J. Grotendorst, volume 3, pp. 97–179, Jülich, 2000, John von Neumann Institute for Computing. 78, 81, 82, 197

[249] W. Kutzelnigg and P. Von Herigonte, *Electron correlation at the dawn the* 21^{st} *century*, Adv. Quantum Chem. **36** (1999) 185–229. 78, 81, 82, 197

[250] C. Møller and M. S. Plesset, *Note on an approximation treatment for many-electron systems*, Phys. Rev. **46** (1934) 618–622. 78, 197

[251] P. Hobza and J. Šponer, *Structure, energetics, and dynamics of the nucleic acid base pairs: Nonempirical ab initio calculations*, Chem. Rev. **99** (1999) 3247–3276. 78, 197

[252] R. A. DiStasio Jr., Y. Jung, and M. Head-Gordon, *A Resolution-of-The-Identity implementation of the local Triatomics-In-Molecules model for second-order Møller-Plesset perturbation theory with application to alanine tetrapeptide conformational energies*, J. Chem. Theory Comput. **1** (2005) 862–876. 78, 197

[253] W. Wang, *Method-dependent relative stability of hydrogen bonded and π–π stacked structures of the formic acid tetramer*, Chem. Phys. Lett. **402** (2005) 54–56. 78, 197

[254] Y. Zhao and D. G. Truhlar, *Infinite-basis calculations of binding energies for the hydrogen bonded and stacked tetramers of formic acid and formamide and their use for validation of hybrid DFT and ab initio methods*, J. Phys. Chem. A **109** (2005) 6624–6627. 78, 197

[255] J. C. Sancho-García and A. Karpfen, *The torsional potential in 2,2′ revisited: High-level ab initio and DFT results*, Chem. Phys. Lett. **411** (2005) 321–326. 78, 178, 179, 191, 197, 201

[256] A. Galindo and P. Pascual, *Quantum Mechanics*, Springer-Verlag, Berlin, 1990. 79

[257] C. Cohen-Tannoudji, B. Diu, and F. Laloë, *Quantum Mechanics*, Hermann and John Wiley & Sons, Paris, 2nd edition, 1977. 79

[258] L. K. William Thompson, *Popular Lectures and Addresses*, 1891–1894. 85

[259] D. Borgis, N. Lévy, and M. Marchi, *Computing the electrostatic free-energy of complex molecules: The variational Coulomb field approximation*, J. Chem. Phys. **119** (2003) 3516. 85, 96, 97, 99

[260] B. N. Dominy and C. L. Brooks III, *Development of a generalized Born model parametrization for proteins and nucleic acids*, J. Phys. Chem. B **103** (1999) 3765–3773. 85, 94, 96, 104

[261] O. Donini and D. F. Weaver, *Development of modified force field for cation-amino acid interactions: Ab initio-derived empirical correction terms with comments on cation-π interactions*, J. Comp. Chem. **19** (1998) 1515–1525. 85, 96

[262] E. Gallicchio and R. M. Levy, *AGBNP: An analytic implicit solvent model suitable for molecular dynamics simulations and high-resolution modeling*, J. Comp. Chem. **25** (2004) 479–499. 85, 96

[263] A. Ghosh, C. S. Rapp, and R. A. Friesner, *Generalized Born model based on a surface integral formulation*, Journal of Physical Chemistry B **102** (1998) 10983–10990. 85, 94, 96, 97

[264] W. Im, M. S. Lee, and C. L. Brooks III, *Generalized Born model with a simple smoothing function*, J. Comp. Chem. **24** (2003) 1661–1702. 85, 94, 96

[265] A. Onufriev, D. Bashford, and D. A. Case, *Modification of the generalized Born model suitable for macromolecules*, J. Phys. Chem. B **104** (2000) 3712–3720. 85, 94, 96

[266] N. Pokala and T. M. Handel, *Energy functions for protein design I: Efficient and accurate continuum electrostatics and solvation*, Prot. Sci. **13** (2004) 925–936. 85, 94, 96

[267] D. Qiu, P. S. Shenkin, F. P. Hollinger, and W. C. Still, *The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate Born radii*, J. Phys. Chem. A **101** (1997) 3005–3014. 85

[268] M. Scarsi, J. Apostolakis, and A. Caflisch, *Continuum electrostatic energies of macromolecules in aqueous solutions*, J. Phys. Chem. A **101** (1997) 8098–8106. 85, 94, 96

[269] M. Scarsi, J. Apostolakis, and A. Caflisch, *Comparison of a GB solvation model with explicit solvent simulations: Potential of mean force and conformational preferences of alanine dipeptide and 1,2-dichloroetane*, J. Phys. Chem. B **102** (1998) 3637–3641. 85, 96

[270] F. Wagner and T. Simonson, *Implicit solvent models: Combining an analytical formulation of continuum electrostatics with simple models of the hydrophobic effect*, J. Comp. Chem. **20** (1999) 322–335. 85, 96

[271] L. David, R. Luo, and G. M. K., *Comparison of generalized Born and Poisson models: Energetics and dynamics of HIV protease*, J. Comp. Chem. **21** (2000) 295–309. 85, 96, 97

[272] S. R. Edinger, C. Cortis, P. S. Shenkin, and R. A. Friesner, *Solvation free energies of peptides: Comparison of approximate continuum solvation models with accurate solution of the Poisson-Boltzmann equation*, J. Phys. Chem. B **101** (1997) 1190–1197. 85, 96

[273] M. Feig, A. Onufriev, M. S. Lee, W. Im, D. A. Case, and C. L. Brooks III, *Performance comparison of generalized Born and Poisson methods in the calculation of electrostatic solvation energies for protein structures*, J. Comp. Chem. **25** (2004) 265–284. 85, 96, 97

[274] R. M. Levy, L. Y. Zhang, E. Gallicchio, and A. K. Felts, *On the nonpolar hydration free energy of proteins: Surface area and continuum solvent models for the solute-solvent interaction energy*, J. Am. Chem. Soc. **125** (2003) 9523–9530. 85, 96

[275] H. Nymeyer and A. E. García, *Simulation of the folding equilibrium of α-helical peptides: A comparison of the generalized Born approximation with explicit solvent*, Proc. Natl. Acad. Sci. USA **100** (2003) 13934–13939. 85, 96

[276] A. Onufriev, D. A. Case, and D. Bashford, *Effective Born radii in the generalized Born approximation: The importance of being perfect*, J. Comp. Chem. **23** (2002) 1297–1304. 85

[277] P. R. Bevington and D. K. Robinson, *Data reduction and error analysis for the physical sciences*, Mc. Graw–Hill, New York, 3rd edition, 2003. 89, 130

[278] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical recipes in C. The art of scientific computing*, Cambridge University Press, New York, 2nd edition, 2002. 89, 130

[279] P. B. Liebelt, *An introduction to optimal estimation*, Addison-Wesley, 1967. 89

[280] J. N. Onuchic, H. Nymeyer, A. E. García, J. Chahine, and N. D. Socci, *The energy landscape theory of protein folding: Insights into folding mechanisms and scenarios*, Adv. Prot. Chem. **53** (2000) 87–130. 94

[281] V. S. Pande, A. Grosberg, and T. Tanaka, *How accurate must potentials be for successful modeling of protein folding?*, J. Chem. Phys. **103** (1995) 9482–9491. 94

[282] A. F. Pereira de Araújo and T. C. Pochapsky, *Monte Carlo simulations of protein folding using inexact potentials: How accurate must parameters be in order to preserve essential features of the energy landscape?*, Folding and Design **1** (1996) 299–314. 94

[283] A. F. Pereira de Araújo and T. C. Pochapsky, *Estimates for the potential accuracy required in realistic folding simulations and structure recognition experiments*, Folding and Design **1** (1997) 135–139. 94

[284] A. Leach, *Molecular modelling: Principles and applications*, Prentice Hall, Harlow, 2nd edition, 2001. 94

[285] N. Metropolis, A. N. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, *Equation of state calculation by fast computing machines*, J. Chem. Phys. **21** (1953) 1087–1092. 94

[286] H. J. Rothe, *Latice gauge theories: An introduction*, volume 43 of *World Scientific Lecture Notes in Physics*, World Scientific, Singapore, 2nd edition, 1997. 94

[287] I. P. Daykov, T. A. Arias, and T. D. Engeness, *Robust ab initio calculation of condensed matter: Transparent convergence through semicardinal multiresolution analysis*, Phys. Rev. Lett. **90** (2003) 216402. 94, 166

[288] B. Honig and A. Nicholls, *Classical electrostatics in biology and chemistry*, Science **213** (1995) 1144–1149. 94

[289] M. Orozco and F. J. Luque, *Theoretical methods for the representation of solvent in biomolecular systems*, Chem. Rev. **100** (2000) 4187–4225. 94

[290] B. Roux, *Implicit solvent models*, in *Computational Biophysics*, edited by O. Becker, A. D. MacKerell, B. Roux, and M. Watanabe, Marcel Dekker Inc., New York, 2001. 94

[291] C. Zhang, S. R. Kimura, Z. Weng, S. Vadja, R. C. Brower, and C. Delisi, *The waters of life*, J. Frank. Inst. **335** (1997) 231–240. 94

[292] D. Bashford and D. A. Case, *Generalized Born models of macromolecular solvation effects*, Ann. Rev. Phys. Chem. **51** (2000) 129–152. 94

[293] W. Still, A. Tempczyk, R. Hawley, and T. Hendrickson, *Semianalytical treatment of solvation for molecular mechanics and dynamics*, J. Am. Chem. Soc. **112** (1990) 6127–6129. 94

[294] L. A. Campos, M. M. García-Mira, R. Godoy-Ruiz, J. M. Sánchez-Ruiz, and J. Sancho, *Do proteins always benefit from a stability increase? Relevant and residual stabilisation in a three-state protein by charge optimisation*, J. Mol. Biol. **344** (2004) 223–237. 96

[295] K. Lindorff-Larsen, E. Paci, L. Serrano, C. M. Dobson, and M. Vendruscolo, *Calculation of mutational free energy changes in transition states for protein folding*, Biophys. Chem. **85** (2003) 1207–1214. 96

[296] J. D. Dobson, *Applied multivariate data analysis*, volume I, Springer-Verlag, New York, 1991. 98, 104, 168

[297] J. W. Neidigh, R. M. Fesenmeyer, and N. H. Andersen, *Designing a 20-residue protein*, Nat. Struct. Biol. **9** (2002) 425. 104

[298] E. A. Coutsias, C. Seok, and K. A. Dill, *Using quaternions to calculate RMSD*, J. Comp. Chem. **25** (2004) 1849–1857. 105

[299] G. A. Chass, M. A. Sahai, J. M. S. Law, S. Lovas, Ö. Farkas, A. Perczel, J.-L. Rivail, and I. G. Csizmadia, *Toward a Computed Peptide Structure Database: The Role of a Universal Atomic Numbering System of Amino Acids in Peptides and Internal Hierarchy of Database*, Intl. J. Quant. Chem. **90** (2002) 933–968. 111, 112, 113, 115, 132, 136

[300] M. A. Sahai, S. Lovas, G. A. Chass, P. Botond, and I. G. Csizmadia, *A modular numbering system of selected oligopeptides for molecular computations: using pre-computed amino acid building blocks*, J. Mol. Struct. **666-667** (2003) 169–218. 111, 112, 113, 132

[301] J. Baker, A. Kessi, and B. Delley, *The generation and use of delocalized internal coordinates in geometry optimization*, J. Chem. Phys. **105** (1996) 192–212. 111, 127, 165

[302] K. Németh and M. Challacombe, *The quasi-independent curvilinear coordinate approximation for geometry optimization*, J. Chem. Phys. **121** (2004) 2877. 111, 112

[303] B. Paizs, J. Baker, S. Suhai, and P. Pulay, *Geometry optimization of large biomolecules in redundant internal coordinates*, J. Chem. Phys. **113** (2000) 6566. 111, 112

[304] M. von Arnim and R. Ahlrichs, *Geometry optimization in generalized natural internal coordinates*, J. Chem. Phys. **111** (1999) 9183. 111

[305] M. W. Schmidt, K. K. Baldridge, J. A. Boatz, S. T. Elbert, M. S. Gordon, H. J. Jensen, S. Koseki, N. Matsunaga, K. A. Nguyen, S. Su, T. L. Windus, M. Dupuis, and J. A. Montgomery, *General Atomic and Molecular Electronic Structure System*, J. Comp. Chem. **14** (1993) 1347–1363. 111, 113, 114, 127, 165

[306] G. Fogarasi, X. Zhou, P. W. Taylor, and P. Pulay, *The Calculation of Ab initio Molecular Geometries: Efficient Natural Internal Coordinates and Empirical Correction by offset Forces*, J. Am. Chem. Soc. **114** (1992) 8191–8201. 111, 112

[307] P. Pulay and G. Fogarasi, *Geometry optimization in redundant internal coordinates*, J. Chem. Phys. **96** (1992) 2856. 111, 112

[308] P. Pulay, G. Fogarasi, F. Pang, and J. E. Boggs, *Systematic ab initio gradient calculation of molecular geometries, force constants, and dipole moment derivatives*, J. Am. Chem. Soc. **101** (1979) 2550–2560. 111, 112

[309] C. Peng, P. Y. Ayala, H. B. Schlegel, and M. J. Frisch, *Using redundant internal coordinates to optimize equilibrium geometries and transition states*, J. Comp. Chem. **17** (1996) 49–56. 112

[310] R. A. Abagyan and A. K. Mazur, *New Methodology for Computer-Aided Modelling of Biomolecular Structure and Dynamics. 2. Local Deformations and Cycles*, J. Biomol. Struct. Dyn. **6** (1989) 833–845. 112, 113, 115, 132, 136

[311] R. A. Abagyan, M. M. Totrov, and D. A. Kuznetsov, *ICM: A New Method For Protein Modeling and Design: Applications To Docking and Structure Prediction From The Distorted Native Conformation*, J. Comp. Chem. **15** (1994) 488–506. 112, 113, 115, 132, 136, 151, 152

[312] A. K. Mazur and R. A. Abagyan, *New Methodology For Computer-Aided Modelling of Biomolecular Structure and Dynamics. 1. Non-Cyclic Structures*, J. Biomol. Struct. Dyn. **6** (1989) 815–832. 112, 113, 115, 132, 136

[313] N. Gō and H. A. Scheraga, *On the use of classical statistical mechanics in the treatment of polymer chain conformation*, Macromolecules **9** (1976) 535. 112, 115, 136, 137, 138, 150, 151, 152, 154, 160, 161

[314] M. Karplus and J. N. Kushick, *Method for estimating the configurational entropy of macromolecules*, Macromolecules **14** (1981) 325–332. 112, 154, 157

[315] W. J. Hehre, W. A. Lathan, R. Ditchfield, M. D. Newton, and J. A. Pople, *Gaussian 70*, Quantum Chemistry Program Exchange, 1970, Program No. 237. 112, 147

[316] I. N. Levine, *Quantum Chemistry*, Prentice Hall, Upper Saddle River, 5th edition, 1999. 112, 147, 171

[317] K. J. Jalkanen, R. M. Nieminen, M. Knapp-Mohammady, and S. Suhai, *Vibrational analysis of various isotopomers of L-alanyl-L-alanine in aqueous solution: Vibrational Circular Dichroism, Raman, and Raman Optical Activity spectra*, Intl. J. Quant. Chem. **92** (2002) 239–259. 112, 113

[318] K. J. Jalkanen and S. Suhai, *N-acetyl-L-alanine N'-methylamide: A density functional analysis of the vibrational absorption and vibrational circular dichroism spectra*, Chem. Phys. **208** (1996) 81–116. 112, 113

[319] H. B. Schlegel, *Some practical suggestions for optimizing geometries and locating transition states*, in *New Theoretical Concepts for Understanding Organic Reactions*, edited by J. Bertrán and I. G. Csizmadia, pp. 33–53, Kluwer Academic, The Netherlands, 1989. 113

[320] P. Davis and R. Hersh, *The Mathematical Experience*, Birkäuser, Boston, 1981. 135

[321] N. G. Almarza, E. Enciso, J. Alonso, F. J. Bermejo, and M. Álvarez, *Monte Carlo simulations of liquid n-butane*, Mol. Phys. **70** (1990) 485–504. 135, 136, 137, 152, 159, 160, 161

[322] S. B. Chen, *Monte Carlo simulations of conformations of chain molecules in a cylindrical pore*, J. Chem. Phys. **123** (2005) 074702. 135

[323] U. H. E. Hansmann and Y. Okamoto, *New Monte Carlo algorithms for protein folding*, Curr. Opin. Struct. Biol. **9** (1999) 177–183. 135

[324] J. Klos and T. Pakula, *Lattice Monte Carlo simulations of three-dimensional charged polymer chains*, J. Chem. Phys. **120** (2005) 2496–2501. 135

[325] L. Nivón and E. I. Shakhnovich, *All-atom Monte Carlo simulation of GCAA RNA folding*, J. Mol. Biol. **344** (2004) 29–45. 135

[326] H. Senderowitz and W. C. Still, *Sampling potential energy surface of glycyl glycine peptide: Comparison of Metropolis Monte Carlo and stochastic dynamics*, J. Comp. Chem. **19** (1998) 1294–1299. 135

[327] D. Shental-Benchor, S. Kirca, N. Ben-Tal, and T. Haliloglu, *Monte Carlo studies of folding, dynamics and stability in α-helices*, Biophys. J. **88** (2005) 2391–2402. 135

[328] J. Shimada and E. I. Shakhnovich, *The ensemble folding kinetics of protein G from an all-atom Monte Carlo simulation*, Proc. Natl. Acad. Sci. USA **99** (2002) 11175–11180. 135

[329] R. D. Taylor, P. J. Jewsbury, and J. W. Essex, *A review of protein-small molecule docking methods*, J. Comput. Aid. Mol. Des. **16** (2002) 151–166. 135

[330] G. R. Smith and M. J. E. Sternberg, *Prediction of protein-protein interactions by docking methods*, Curr. Opin. Struct. Biol. **12** (2002) 28–35. 135

[331] L. D. Barron and H. L., *Vibrational Raman optical activity: from fundamentals to biochemical applications*, in *Circular Dichroism Principles and Applications*, edited by K. Nakanishi, N. Berova, and R. W. Woody, pp. 667–701, Wiley-VCH, New York, 2nd edition, 2000. 135

[332] M. Levantino, Q. Huang, A. Cupane, M. Laberge, A. Hagarman, and R. Schweitzer-Stenner, *The importance of vibronic perturbations in ferrocytochrome c spectra: A reevaluation of spectral properties based on low-temperature optical absortion, resonance Raman, and molecular-dynamics simulations*, J. Chem. Phys. **123** (2000) 054508. 135

[333] H. H. Mantsch and C. D., *Infrared Spectroscopy of Biomolecules*, Wiley-Liss, Chichester, UK, 1996. 135

[334] S. Yang and M. Cho, *IR spectra of N-methylacetamide in water predicted by combined quantum mechanical/molecular mechanical molecular dynamics simulations*, J. Chem. Phys. **123** (2005) 134503. 135

[335] N. Sreerama and R. W. Woody, *Circular Dichroism of peptides and proteins*, in *Circular Dichroism Principles and Applications*, edited by K. Nakanishi, N. Berova, and R. W. Woody, pp. 601–620, Wiley-VCH, New York, 2nd edition, 2000. 135

[336] J.-H. Choi, J.-S. Kim, and M. Cho, *IR spectra of N-methylacetamide in water predicted by combined quantum mechanical/molecular mechanical molecular dynamics simulations*, J. Chem. Phys. **122** (2005) 174903. 135

[337] T. A. Keiderling, *Peptide and protein conformational studies with vibrational circular dichroism and related spectroscopies*, in *Circular Dichroism Principles and Applications*, edited by K. Nakanishi, N. Berova, and R. W. Woody, pp. 621–666, Wiley-VCH, New York, 2nd edition, 2000. 135

[338] W. K. Den Otter and W. J. Briels, *Free energy from molecular dynamics with multiple constraints*, Mol. Phys. **98** (2000) 773–781. 136, 137, 150, 152, 153, 159, 161

[339] D. C. Morse, *Theory of constrained Brownian motion*, Adv. Chem. Phys. **128** (2004) 65–189. 136, 137, 150, 152, 154, 159, 161

[340] M. Fixman, *Classical Statistical Mechanics of constraints: A theorem and application to polymers*, Proc. Natl. Acad. Sci. USA **71** (1974) 3050–3053. 137, 150, 151, 152, 153, 159, 160, 161

[341] E. B. Wilson Jr., J. C. Decius, and P. C. Cross, *Molecular Vibrations: The Theory of Infrared and Raman Vibrational Spectra*, Dover Publications, New York, 1980. 137

[342] A. Patriciu, G. S. Chirikjian, and R. V. Pappu, *Analysis of the conformational dependence of mass-metric tensor determinants in serial polymers with constraints*, J. Chem. Phys. **121** (2004) 12708–12720. 138, 150, 152, 160, 161

[343] H. Goldstein, C. Poole, and J. Safko, *Classical Mechanics*, Addison-Wesley, 3rd edition, 2002. 139, 217

[344] D. Adams and M. Cawardine, *Last Chance to See*, Ballantine Books, New York, 1990. 151

[345] S. He and H. A. Scheraga, *Brownian dynamics simulations of protein folding*, J. Chem. Phys. **108** (1998) 287. 151, 152

[346] H. M. Chun, C. E. Padilla, D. N. Chin, M. Watanabe, V. I. Karlov, H. E. Alper, K. Soosaar, K. B. Blair, O. M. Becker, L. S. D. Caves, R. Nagle, D. N. Haney, and B. L. Farmer, *MBO(N)D: A multibody method for long-time Molecular Dynamics simulations*, J. Comp. Chem. **21** (2000) 159–184. 151, 152

[347] S. Reich, *Multiple time scales in classical and quantum-classical molecular dynamics*, J. Comput. Phys. **151** (1999) 49–73. 151

[348] T. Schlick, E. Barth, and M. Mandziuk, *Biomolecular dynamics at long timesteps: Bridging the timescale gap between simulation and experimentation*, Annu. Rev. Biophys. Biomol. Struct. **26** (1997) 181-222. 151

[349] N. G. Van Kampen and J. J. Lodder, *Constraints*, Am. J. Phys. **52** (1984) 419–424. 151

[350] N. Gō and H. A. Scheraga, *Analysis of the contributions of internal vibrations to the statistical weights of equilibrium conformations of macromolecules*, J. Chem. Phys. **51** (1969) 4751. 151, 152, 157

[351] J. M. Rallison, *The role of rigidity constraints in the rheology of dilute polymer solutions*, J. Fluid Mech. **93** (1979) 251–279. 151, 152, 160, 161

[352] D. Chandler and B. J. Berne, *Comment on the role of constraints on the conformational structure of n-butane in liquid solvent*, J. Chem. Phys. **71** (1979) 5386–5387. 152, 153, 159, 160

[353] M. Gottlieb and R. B. Bird, *A Molecular Dynamics calculation to confirm the incorrectness of the random-walk distribution for describing the Kramers freely jointed bead-rod chain*, J. Chem. Phys. **65** (1976) 2467. 152

[354] H. J. C. Berendsen and W. F. Van Gunsteren, *Molecular Dynamics with constraints*, in *The Physics of Superionic Conductors and Electrode Materials*, edited by J. W. Perram, volume NATO ASI Series B92, pp. 221–240, Plenum Press, 1983. 152, 161

[355] H. J. C. Berendsen and W. F. Van Gunsteren, *Molecular Dynamics simulations: Techniques and approaches*, in *Molecular Liquids-Dynamics and Interactions*, edited by A. J. e. a. Barnes, pp. 475–500, Reidel Publishing Company, 1984. 152, 161

[356] G. Ciccotti and J. P. Ryckaert, *Molecular dynamics simulation of rigid molecules*, Comput. Phys. Rep. **4** (1986) 345–392. 152, 161

[357] W. K. Den Otter and W. J. Briels, *The calculation of free-energy differences by constrained molecular-dynamics simulations*, J. Chem. Phys. **109** (1998) 4139. 152, 153, 159, 161

[358] M. Fixman, *Simulation of polymer dynamics. I. General theory*, J. Chem. Phys. **69** (1978) 1527. 152, 153, 159, 160, 161

[359] M. Pasquali and D. C. Morse, *An efficient algorithm for metric correction forces in simulations of linear polymers with constrained bond lengths*, J. Chem. Phys. **116** (2002) 1834. 152, 153, 159, 160, 161

[360] J. Zhou, S. Reich, and B. R. Brooks, *Elastic molecular dynamics with self-consistent flexible constraints*, J. Chem. Phys. **111** (2000) 7919. 152, 154, 161

[361] B. Hess, H. Saint-Martin, and H. J. C. Berendsen, *Flexible constraints: An adiabatic treatment of quantum degrees of freedom, with application to the flexible and polarizable mobile charge densities in harmonic oscillators model for water*, J. Chem. Phys. **116** (2002) 9602. 152, 161

[362] R. F. Álvarez-Estrada and G. F. Calvo, *Models for biopolymers based on quantum mechanics*, Mol. Phys. **100** (2002) 2957–2970. 152

[363] R. F. Álvarez-Estrada, *Models of macromolecular chains based on Classical and Quantum Mechanics: comparison with Gaussian models*, Macromol. Theory Simul. **9** (2000) 83–114. 152

[364] H. C. Andersen, *Rattle: A "velocity" version of the Shake algorithm for molecular dynamics calculations*, J. Comput. Phys. **52** (1983) 24–34. 152

[365] E. Barth, K. Kuczera, B. Leimkuhler, and R. D. Skeel, *Algorithms for constrained Molecular Dynamics*, J. Comp. Chem. **16** (1995) 1192–1209. 152

[366] J. P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen, *Numerical integration of the Cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes*, J. Comput. Phys. **23** (1977) 327–341. 152

[367] J. Schlitter and M. Klän, *The free energy of a reaction coordinate at multiple constraints: a concise formulation*, Mol. Phys. **101** (2003) 3439–3443. 152, 153, 154, 159, 160, 161

[368] W. F. Van Gunsteren, *Methods for calculation of free energies and binding constants: Successes and problems*, in *Computer Simulations of Biomolecular Systems*, edited by W. F. Van Gunsteren and P. K. Weiner, pp. 27–59, Escom science publishers, Netherlands, 1989. 152, 154, 161

[369] A. R. Dinner, *Local deformations of polymers with nonplanar rigid main-chain coordinates*, J. Comp. Chem. **21** (2000) 1132–1144. 152

[370] E. W. Knapp and A. Irgens-Defregger, *Off-lattice Monte Carlo method with constraints: Long-time dynamics of a protein model without nonbonded interactions*, J. Fluid Mech. **14** (1993) 19–29. 152

[371] A. J. Pertsin, J. Hahn, and H. P. Grossmann, *Incorporation of bond-lengths constraints in Monte Carlo simulations of cyclic and linear molecules: conformational sampling for cyclic alkanes as test systems*, J. Comp. Chem. **15** (1994) 1121–1126. 152, 160, 161

[372] J. Schofield and M. A. Ratner, *Monte Carlo methods for short polypeptides*, J. Chem. Phys. **109** (1998) 9177. 152

[373] M. P. Allen and D. J. Tildesley, *Computer simulation of liquids*, Clarendon Press, Oxford, 2005. 154, 161

[374] D. Frenkel and S. B., *Understanding molecular simulations: From algorithms to applications*, Academic Press, Orlando FL, 2nd edition, 2002. 154, 161

[375] B. Viskolcz, S. N. Fejer, and I. G. Csizmadia, *Thermodynamic Functions of Conformational Changes. 2. Conformational Entropy as a Measure of Information Accumulation*, *To be published in J. Phys. Chem. A.*, 2006. 156

[376] I. Andricioaei and M. Karplus, *On the calculation of entropy from covariance matrices of the atomic fluctuations*, J. Chem. Phys. **115** (2001) 6289–6292. 157

[377] M. V. Volkenstein, *Configurational Statistical of Polymeric Chains*, Interscience, New York, 1959. 160

[378] J. Chen, W. Im, and C. L. Brooks III, *Application of torsion angle molecular dynamics for efficient sampling of protein conformations*, J. Comp. Chem. **26** (2005) 1565–1578. 161

[379] L. Schäfer and C. Ming, *Predictions of protein backbone bond distances and angles from first principles*, J. Mol. Struct. **333** (1995) 201–208. 161

[380] M. Mazars, *Statistical physics of the freely jointed chain*, Phys. Rev. E **53** (1996) 6297. 161

[381] M. Mazars, *Canonical partition function of freely jointed chains*, J. Phys. A: Math. Gen. **31** (1998) 1949–1964. 161

[382] E. Demchuk and H. Singh, *Statistical thermodynamics of hindered rotation from computer simulations*, Mol. Phys. **99** (2001) 627–636. 165

[383] V. Hnizdo, A. Fedorowicz, H. Singh, and E. Demchuk, *Statistical thermodynamics of internal rotation in a hindering potential of mean force obtained from computer simulations*, J. Comp. Chem. **24** (2003) 1172–1183. 165

[384] K. V. Mardia and P. E. Jupp, *Directional Statistics*, John Wiley & Sons, Chichester, 2000. 165

[385] M. D. Halls, J. Velkovski, and H. B. Schlegel, *Harmonic Frequency Scaling Factors for Hartree-Fock, S-VWN, B-LYP, B3-LYP, B3-PW91 and MP2 and the Sadlej pVTZ Electric Property Basis Set*, Theo. Chem. Acc. **105** (2001) 413. 171

[386] P. Chakrabarti and D. Pal, *The interrelationships of side-chain and main-chain conformations in proteins*, Prog. Biophys. Mol. Biol. **76** (2001) 1–102. 173

[387] K. Gunasekaran, C. Ramakrishnan, and P. Balaram, *Disallowed Ramachandran conformations of amino acid residues in protein structures*, J. Mol. Biol. **264** (1996) 191–198. 173

[388] C. Babbage, *Passages from the Life of a Philosopher*, Rutgers University Press, 1864, http://www.fourmilab.ch/babbage/lpae.html. 177

[389] I. A. Topol, S. K. Burt, E. Deretey, T.-H. Tang, A. Perczel, A. Rashin, and I. G. Csizmadia, *α- and 3_{10}-helix interconversion: A quantum-chemical sttudy on polyalanine systems in the gas phase and in aqueous solvent*, J. Am. Chem. Soc. **123** (2001) 6054–6060. 177, 214, 228

[390] M. Elstner, K. J. Jalkanen, M. Knapp-Mohammady, T. Frauenheim, and S. Suhai, *DFT studies on helix formation in N-acetyl-(L-alanyl)$_n$-N′-methylamide for n=1–20*, Chem. Phys. **256** (2001) 15–27. 177

[391] P. Jurečka, J. Šponer, J. Černý, and P. Hobza, *Benchmark database of accurate (MP2 and CCSD(T) complete basis set limit) interaction energies of small model complexes, DNA base pairs, and amino acid pairs*, Phys. Chem. Chem. Phys. **8** (2006) 1985–1993. 178

[392] G. A. Petersson, D. K. Malick, M. J. Frisch, and M. Braunstein, *The convergence of complete active space self-consistent-field energies to the complete basis set limit*, J. Chem. Phys. **123** (2005) 074111. 178

[393] Z.-H. Li and M. W. Wong, *Scaling of correlation basis set extension energies*, Chem. Phys. Lett. **337** (2001) 209–216. 178

[394] M. R. Nyden and G. A. Petersson, *Complete basis set correlation energies. I. The assymptotic convergence of pair natural orbital expansions*, J. Chem. Phys. **75** (1981) 1843–1862. 178

[395] P. Jurečka and P. Hobza, *On the convergence of the ($\Delta E^{\mathrm{CCSD(T)}} - \Delta E^{\mathrm{MP2}}$) term for complexes with multiple H-bonds*, Chem. Phys. Lett. **365** (2002) 89–94. 178

[396] E. W. Ignacio and H. B. Schlegel, *On the additivity of basis set effects in some simple fluorine containing systems*, J. Comp. Chem. **12** (1991) 751–760. 178, 184

[397] J. S. Dewar and A. J. Holder, *On the validity of polarization and correlation additivity in ab initio molecular orbital calculations*, J. Comp. Chem. **3** (1989) 311–313. 178

[398] R. H. Nobes, W. J. Bouma, and L. Radom, *The additivity of polarization function and electron correlation effects in ab initio molecular-orbital calculations*, Chem. Phys. Lett. **89** (1982) 497–500. 178

[399] J. A. Pople, M. J. Frisch, B. T. Luke, and J. S. Binkley, *A Moller-Plesset study of the energies of AH_n molecules (A = Li to F)*, Intl. J. Quant. Chem. **17** (1983) 307–320. 178

[400] R. Crespo-Otero, L. A. Montero, W.-D. Stohrer, and J. M. García de la Vega, *Basis set superposition error in MP2 and density-functional theory: A case of methane-nitric oxide association*, J. Chem. Phys. **123** (2005) 134107. 178

[401] M. L. Senent and S. Wilson, *Intramolecular basis set superposition errors*, Intl. J. Quant. Chem. **82** (2001) 282–292. 178

[402] I. Mayer and P. Valiron, *Second order Møller-Plesset perturbation theory without basis set superposition error*, J. Chem. Phys. **109** (1998) 3360–3373. 178

[403] F. Jensen, *The magnitude of intramolecular basis set superposition error*, Chem. Phys. Lett. **261** (1996) 633–636. 178

[404] I. Mayer, *On the non-additivity of the basis set superposition error and how to prevent its appearance*, Theo. Chem. Acc. **72** (1987) 207–210. 178

[405] S. F. Boys and F. Bernardi, *The calculation of small molecular interactions by the differences of separate total energies. Some procedures with reduced errors*, Mol. Phys. **19** (1970) 553–566. 178

[406] H. B. Jansen and P. Ros, *Non-empirical molecular orbital calculations on the protonation of carbon monoxide*, Chem. Phys. Lett. **3** (1969) 140–143. 178

[407] C. Tuma, A. D. Boese, and N. C. Handy, *Predicting the binding energies of H-bonded complexes: A comparative DFT study*, Phys. Chem. Chem. Phys. **1** (1999) 3939–3947. 197

[408] L. González, O. Mó, and M. Yáñez, *High-level ab initio versus DFT calculations on $(H_2O_2)_2$ and H_2O_2-H_2O complexes as prototypes of multiple hydrogen bonds*, J. Comp. Chem. **18** (1998) 1124–1135. 197

[409] T. Beke, I. Csizmadia, and A. Perczel, *Theoretical study on tertiary structural elements of β-peptides: Nanotubes formed from parallel-sheet-derived assemblies of β-peptides*, J. Am. Chem. Soc. **128** (2006) 5158–5167. 214, 227

[410] N. G. van Kampen, *Stochastic processes in Physics and Chemistry*, North-Holland, Amsterdam, 1981. 215

[411] A. Perczel, J. G. Angyán, M. Kajtar, W. Viviani, J.-L. Rivail, J.-F. Marcoccia, and I. G. Csizmadia, *Peptide models. 1. Topology of selected peptide conformational potential energy surfaces (glycine and alanine derivatives)*, J. Am. Chem. Soc. **113** (1991) 6256-6265. 227, 228, 229

[412] I. Bágyi, B. Balogh, A. Czaklik, O. Éliás, Z. Gáspári, V. Gergely, I. Hudáky, P. Hudáky, A. Kalászi, L. Károlyházy, K. Keserû, R. Kiss, G. Krajsovszky, B. Láng, T. Nagy, Á. Rácz, A. Szentesi, T. Tábi, P. Tapolcsányi, J. Vaik, J. C. P. Koo, G. A. Chass, Ö. Farkas, A. Perczel, and P. Mátyus, *Generation and analysis of the conformational potential energy surfaces of N-acetyl-N-methyl-L-alanine-N′-methylamide. An exploratory ab initio study*, J. Mol. Struct. **625** (2003) 121–136. 227

Index