UNIVERSITAT DE BARCELONA

Development of crystallographic methods for phasing highly modulated macromolecular structures

Iracema Caballero Muñoz

UNIVERSITAT DE BARCELONA

FACULTAT DE FARMÀCIA I CIÈNCIES DE L'ALIMENTACIÓ

(Prof. Dr. Isabel Usón Finkenzeller & Dr. Airlie J. McCoy)

Development of crystallographic methods for phasing highly modulated macromolecular structures

Iracema Caballero Muñoz

2020

UNIVERSITAT DE BARCELONA

FACULTAT DE FARMÀCIA I CIÈNCIES DE L'ALIMENTACIÓ

PROGRAMA DE DOCTORAT EN BIOTECNOLOGIA

Development of crystallographic methods for phasing highly modulated macromolecular structures

Memòria presentada per Iracema Caballero Muñoz

per optar al títol de doctor per la Universitat de Barcelona

Directora: Prof. Dr. Isabel Usón Finkenzeller

Directora: Dr. Airlie J. McCoy

Tutora: Prof. Dr. Josefa Badía Palacín

Iracema Caballero Muñoz

2020

To my father

ACKNOWLEDGEMENTS

I first would like to acknowledge all the support and mentorship of my thesis supervisors. To Isabel Usón, for giving me the opportunity to do the PhD with you and for supporting me during my entire thesis. Also, for valuing my strengths and pushing me to improve my weaknesses. To her lovely family Agustín, Flavia, Miranda and Candela. To Airlie McCoy for supervising me mainly in the distance, and during my short stays in Cambridge. For your patience in answering all my questions, solving all the bugs and even teaching me Australian slang. Doing research with you has been a fantastic experience.

I am also grateful to all my colleagues from the Arcimboldo Team. To Massimo and Claudia, who helped me when I arrived at the lab. They have continued helping me not only in the scientific part but also in the personal one. Both of you have been an inspiration to me, I hope one day I become as good programmer as you. To Ana, we have spent many hours together and learned a lot one from the other, I am really thankful because you always have been there when I have needed it. To Giovanna, Nicolas, Luca, Rafa, Eli, Albert, Alfonso and Pep for all your support and help. I have understood what real teamwork means with all of you. Also, I want to mention the members of the other groups at the Structural Biology Unit. They constitute a fantastic community to work in, and some have become really good friends.

I want to thank all the members of Randy Read's group for hosting me during my visits to Cambridge, I enjoyed a lot the discussions during the coffee breaks and the conferences we attended together.

I also want to thank my parents, who have supported me through all my life. To my mother for her unconditional love. To my father who, when I was a child used to make me sit on his legs to look through the microscope. You taught me what love for science is, and thanks to you, I am here today. You left just before I started the PhD, but I know you would be very proud. I love you, and I wish you were here. Also, to the rest of my family, my brothers, aunts, uncles and cousins.

To all my friends, particularly to my best friend Leti, who I consider a sister, we have always been together since we started university.

Last but not least, to my boyfriend Iñigo, who has supported me throughout all the thesis, and I hope he keeps doing it all my life.

This work was supported by grants BIO2015-64216-P and BES-2016-076329 from the Spanish Ministry of Economy and Competitiveness.

TABLE OF CONTENTS

SUMMARY

Pathologies that result in highly modulated intensities in macromolecular crystal structures pose a challenge for structure solution. To address this issue two studies have been performed: a theoretical study of one of these pathologies, translational non-crystallographic symmetry (tNCS), and a practical study of paradigms of highly modulated macromolecular structures, coiled-coils.

tNCS is a structural situation in which multiple, independent copies of a molecular assembly are found in similar orientations in the crystallographic asymmetric unit. Structure solution is problematic because the intensity modulations caused by tNCS cause the intensity distribution to differ from a Wilson distribution. If the tNCS is properly detected and characterized, expected intensity factors for each reflection that model the modulations observed in the data can be refined against a likelihood function to account for the statistical effects of tNCS.

In this study, a curated database of 80482 protein structures from the PDB was analysed to investigate how tNCS manifests in the Patterson function. These studies informed the algorithm for detection of tNCS, which includes a method for detecting the tNCS order in any commensurate modulation. In the context of automated structure solution pipelines, the algorithm generates a ranked list of possible tNCS associations in the asymmetric unit, which can be explored to efficiently maximize the probability of structure solution.

Coiled-coils are ubiquitous protein folding motifs present in a wide range of proteins that consist of two or more α-helices wrapped around each other to form a supercoil. Despite the apparent simplicity of their architecture, solution by molecular replacement is challenging due to the helical irregularities found in these domains, tendency to form fibers, large dimensions in their typically anisometric asymmetric units, low-resolution and anisotropic diffraction. In addition, the internal symmetry of the helices and their alignment in preferential directions gives rise to systematic overlap of Patterson vectors, a Patterson map that indicates tNCS is present, and intensity modulations similar to those in true tNCS.

In this study, we have explored fragment phasing on a pool of 150 coiled-coils with *ARCIMBOLDO_LITE*, an *ab initio* phasing approach that combines fragment location with *Phaser* and density modification and autotracing with *SHELXE*. The results have been used to identify limits and bottlenecks in coiled-coil phasing that have been addressed in a specific mode for solving coiled-coils, allowing the solution of 95% of the test set and four previously unknown structures, and extending the resolution limit from 2.5 Å to 3.0 Å.

LIST OF ABBREVIATIONS AND ACRONYMS

ACC	**Acc**uracy
ADPs	**A**nisotropic **D**isplacement **P**arameter**s**
AMF	**A**tomic **M**odulation **F**unction
CC	**C**orrelation **C**oefficient
CCP4	**C**ollaborative **C**omputational **P**roject Number **4**
CCTBX	**C**omputational **C**rystallography **T**ool**b**o**x**
CV	**C**haracteristic **V**ector
FN	**F**alse **N**egative
FP	**F**alse **P**ositive
FPR	**F**alse **P**ositive **R**ate
LLG	**L**og **L**ikelihood **G**ain
LTD	**L**attice-**T**ranslocation **D**efect
MAD	**M**ultiple-wavelength **A**nomalous **D**ispersion
MIR	**M**ultiple **I**somorphous **R**eplacement
MPD	**M**ean **P**hase **D**ifference
MR	**M**olecular **R**eplacement
OD	**O**rder-**D**isorder
PDB	**P**rotein **D**ata **B**ank
PREC	**Prec**ision
PyPI	**Py**thon **P**ackage **I**ndex
rmsd	**r**oot-**m**ean-**s**quare **d**eviation
SAD	**S**ingle-wavelength **A**nomalous **D**ispersion
SN	**S**e**n**sitivity
TLS	**T**ranslation-**L**ibration-**S**crew
tNCS	**t**ranslational **N**on-**C**rystallographic **S**ymmetry
TN	**T**rue **N**egative
TP	**T**rue **P**ositive
vrms	**r**oot-**m**ean-**s**quare derived from the likelihood **v**ariance
wMPE	**w**eighted **M**ean **P**hase **E**rror

INTRODUCTION

1. X-ray crystallography

Crystallography allows us to conclusively establish a three-dimensional model of the atomic structure of the molecules down to the atomic level, so knowing the precise stereochemistry of these proteins, we can understand their function and interaction with other molecules. This result has fundamental relevance to structural studies underlying biotechnology and biomedicine, such as the design of catalysts and new drugs.

The word "crystallography" has its origin in the Greek word *krystallos*, meaning "clear ice or ice cold" and was originally used to refer to materials that looked similar to ice, such as quartz. However, the formation of crystals is not a unique property of minerals. There are also crystals of organic compounds, nucleic acids, proteins, and viruses. Crystals are formed by atoms, ions, and/or molecules that pack together in ordered and periodic arrangements in the three dimensions of space, thanks to interatomic and/or intermolecular interactions (Janin & Rodier, 1995).

X-ray diffraction has been the main method used during the past 100 years to determine the three-dimensional structure of crystals. This enables the description of the geometrical arrangement of atoms in the crystals by diffraction patterns obtained through exposure to X-rays (Friedrich *et al.*, 1913; Laue, 1913). Diffraction patterns are arrays of spots called reflections, each reflection is the sum of different waves originated by the diffraction of all atoms in the crystal (Bragg, 1913). As the diffraction pattern depends on the scattering contributions of all the atoms, it will be affected by structural heterogeneity of the protein molecule, the nature of packing interactions (weaker or stronger in different directions), and the large volume fraction of water in the macromolecular crystals, ~50% on average. Thus, the structure gives rise to the diffraction pattern, and from the diffraction pattern, the structure can be determined.

During the diffraction experiment, the measurable parameters are the position and the intensity of each reflection, whereas the phases are lost. This gives rise to the phase problem, a bottleneck in the determination of macromolecular crystal structures (Bragg & Bragg, 1913; Hendrickson, 2013), that may be solved through three kinds of methods. Where the data are high-resolution and the number of atoms in the asymmetric unit of the order of a few hundred atoms or less, the phases for the structure factors that are missing from the results of diffraction experiment can be obtained by *ab initio* phasing (Karle & Hauptman, 1956; Sheldrick *et al.*, 2012; Woolfson, 1987), exploiting probabilistic relations and the possibility of evaluating many starting phase sets through reliable figures of merit. Phases can also be obtained by experimental phasing, using heavy-atom derivatives (Green *et al.*, 1954) or anomalous scattering at particular wavelengths (Hendrickson, 1991). Experimental phasing is a two-step procedure, first, the heavy-atom substructure is determined by *ab initio*, Patterson or dual-space methods, and then the phases of the entire structure are derived using this substructure. Alternatively, phases can be obtained by molecular replacement (MR) (Navaza, 1994; Read, 2001; Rossmann, 1972), which uses previous structural knowledge from a similar structure. Phases calculated from this similar structure placed in the unit cell with a low root-mean-square deviation (rmsd) to the atoms of the target structure are used to estimate the unknown phases and bootstrap structure refinement.

2. Statistics of an ideal crystal, the Wilson distribution

The probability distribution of the intensities in the X-ray diffraction pattern was first considered by Wilson (Wilson, 1949). The eponymous Wilson distribution assumes that the atoms are independent and randomly distributed within the crystal. The Wilson plot shows the falloff in intensity as a function in resolution, this trend is due to the falloff of atomic scattering factors (Debye, 1913). It is the plot of $\ln\left(\overline{I_{hkl}}\,/\sum_i \left(f_i^0\right)^2\right)$ against $(sin^2\theta)/\lambda^2$, where $\overline{I_{hkl}}$ represents the average intensity (on a relative scale) collected for a given interval of θ (the Bragg angle), f are the atomic scattering factors in that angular range, and $(sin^2\theta)/\lambda^2$ is the inverse resolution squared. The plot should give a straight line with a slope -2B, where B is the isotropic overall B-factor.

A linear fit to the high-resolution region of a Wilson plot allows to estimate the Debye-Waller factor, temperature factor or overall B-factor, which is a general measure of the scattering attenuation and describes the overall thermal motion (Debye, 1913). It represents the decrease of intensity in diffraction due to crystal disorder, it includes both the static disorder in the crystal and the dynamic disorder caused by thermal vibrations. It is an approximation of the average atomic B-factors which may be later obtained in refinement. The Wilson distribution can be used to bring the data onto an absolute scale and correct the data through the French and Wilson procedure (French & Wilson, 1978).

Theoretically, the Wilson plot gives a straight line, but if the atoms are not randomly distributed this will cause characteristic departures from linearity (Morris & Bricogne, 2003). At low-resolution this is caused by the presence of solvent regions, where the electron density is much more constant than in the protein regions, at intermediate resolutions (round 5-3 Å) this is caused by the regular peptide structure in secondary structure elements such as the pitch of helices and the translation in beta sheets. Further deviations of the Wilson plot from the expected plot may reveal any anomalies or pathologies in the data. The expected plot is based upon an analysis of high-resolution datasets in the Protein Data Bank (PDB) (Berman *et al.*, 2000; Burley *et al.*, 2018), which takes into account the non-random distribution of atoms within the crystal. Some deviation from this plot is to be expected, however, significant deviation may indicate problems (figure 1).

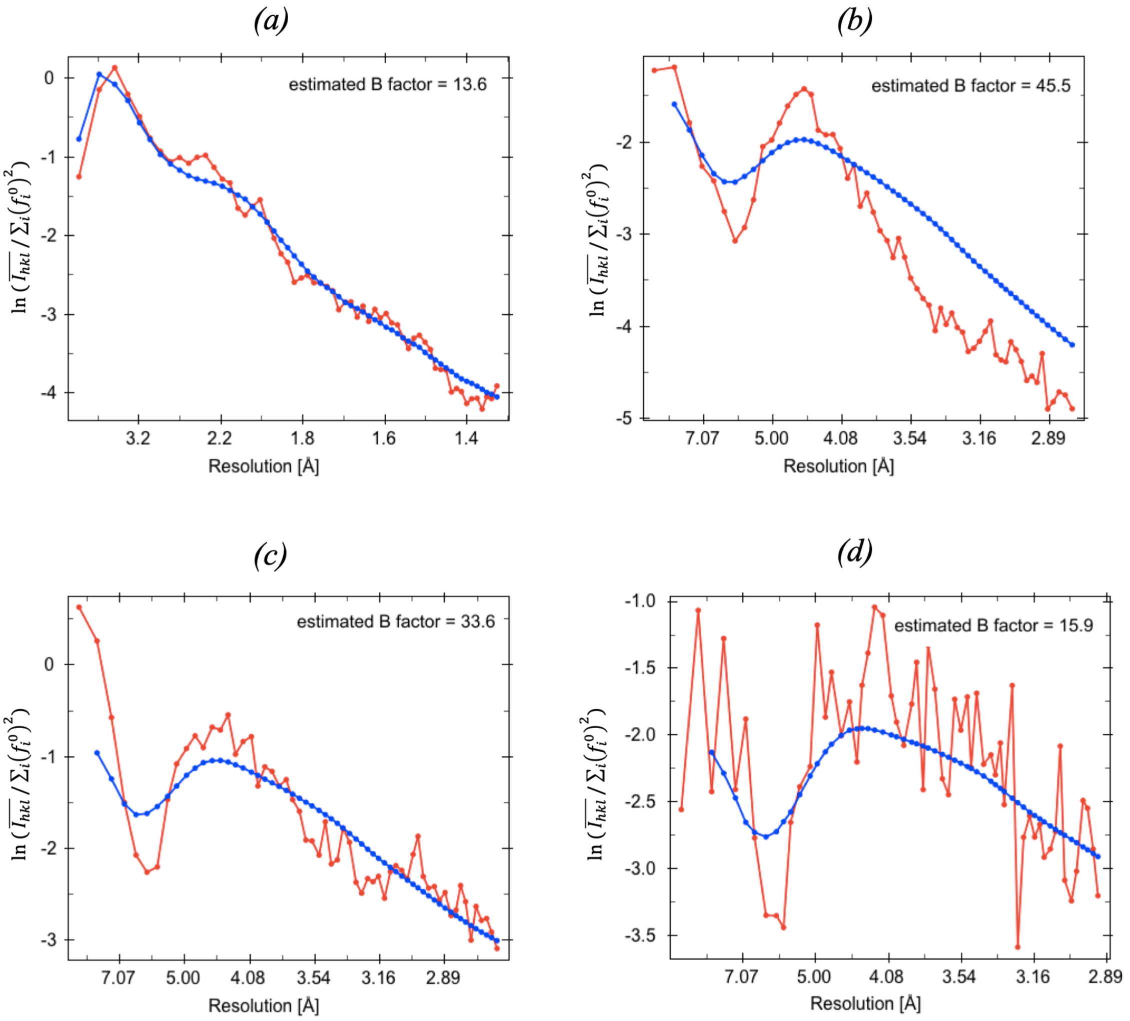


Figure 1. Wilson plots for a) a regular structure (PDB entry 3twe), b) a structure with severe anisotropy (PDB entry 2jee), c) a structure with tNCS (PDB entry 2o1j), d) a structure with merohedral twinning (PDB entry 3v86). The plot from the problem structure is coloured in red, whereas the expected plot is coloured in blue. These plots were generated using *TRUNCATE* (French & Wilson, 1978) from *CCP4* (Winn *et al.*, 2011).

3. Crystal pathologies in macromolecular crystallography

In the context of this work, I use the term “pathology” to denote the systematic deviation of identifiable subsets of reflections from the expected intensity distribution with a deleterious effect on structure solution and refinement. Pathologies are generally considered to be relatively rare in macromolecular crystallography, but this is biased by the population of the PDB with solved structures, which are more likely to not display pathologies. For some classes of proteins, it is more common to encounter crystals with pathologies than high-quality well-diffracting crystal. Therefore, to better understand the deviations from an ideal crystal, the following sections provide an introduction to five common pathologies that occur in macromolecular crystallography, highlighting how to identify and overcome them.

3.1. Anisotropy

Anisotropy is the property of being directionally dependent, implying a variation of certain physical properties as a function of direction, as opposed to isotropy. In crystallography, a physical property of the measured X-ray diffraction data that is the diffraction limits can be anisotropic, also can be anisotropic, the displacement of the atoms that results in the attenuation of the diffracted intensity.

Diffraction anisotropy is a common phenomenon for macromolecular crystals, and it is a directional dependence in diffraction quality when reflections have measurable intensities at higher resolution in one direction, not in another. So, resolution limits vary significantly in different directions in reciprocal space (Sheriff & Hendrickson, 1987). This is related with the attenuation of diffraction, caused by the displacement of the atoms from its mean position due to thermal vibrations, static or dynamic disorder that causes a phase change of the diffracted wave depending on the diffraction angle (Matthews & Czerwinski, 1975; Shakked, 1983; Sheriff & Hendrickson, 1987; Trueblood *et al.*, 1996). Thermal vibrations or disorder in a protein crystal is frequently anisotropic given by the irregular and flexible shape of a protein molecule, as crystal packing interactions can be more uniform in one direction than another, that is, adequate intermolecular interactions may exist in two dimensions or in layers, while contacts in the third dimension may be weak (Janin & Rodier, 1995). If the displacement disorder in a crystal structure varies in different directions of reciprocal space, the diffraction of resolution limits will be anisotropic as well.

Anisotropic properties can be described with the anisotropic displacement tensor U, a three-by-three symmetric matrix, which can be graphically represented as a symmetric ellipsoid (figure 2b) with its principal axes given by the diagonal elements from U (Trueblood *et al.*, 1996). Because of the symmetry of the ellipsoid, six components of the symmetric anisotropic displacement tensor - the diagonal and off-diagonal terms - suffice to describe the orientation and principal axes of an ellipsoid. Anisotropy is subject to the symmetry constraints of the crystal system, which may lead to a reduction of the parameters. In the extreme case of cubic system, anisotropic diffraction is not possible. Another prominent application, which is subject to the same formalism is the description of anisotropic motion of individual atoms, described below and exemplified in figure 2a.

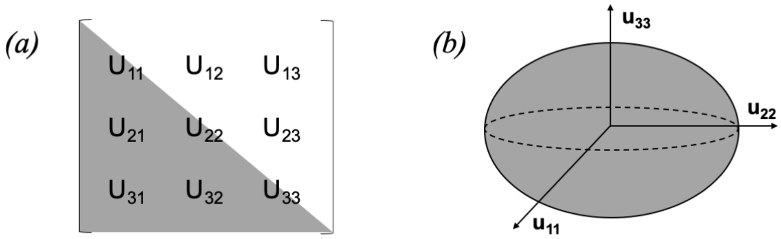

(a)
U_{11} U_{12} U_{13}
U_{21} U_{22} U_{23}
U_{31} U_{32} U_{33}
(b)
u_{33}
u_{22}
u_{11}

Figure 2. a) Anisotropic displacement tensor U describing the anisotropic motion of an atom. The tensor elements U_{ij} are the squares of the displacements u_{ij} from the coordinates in the crystal structure and at angles to these directions. The diagonal (isotropic) elements U_{11}, U_{22} and U_{33} consider the displacements along cell edges and the other (anisotropic) elements the displacements at angle to the cell edges. b) The anisotropic tensor can be visualized as an ellipsoid with principal axes u_{ij}, where the displacement of atoms is of different magnitude along the three principal axes.

3.1.1. Degree of anisotropy

The observed reflections follow a typical intensity falloff as a function of resolution according to different directions in reciprocal space. This directional dependence of the intensity falloff with resolution can be measured with the anisotropic delta-B factor, which also indicates the degree of anisotropy.

The magnitude of the anisotropy is parameterized by three principal components (β_{11}, β_{22}, β_{33}) one for each direction of the crystal, that can be interconverted to anisotropic B factors by a mathematical relationship (Grosse-Kunstleve & Adams, 2002). They are the exponential scale factors used to correct for anisotropy. If the falloff is nearly the same in all three directions, the three principal components will be approximately equal and close to zero. If the falloff has a strong directional dependence, then the component describing the weakest diffracting direction will be large and positive and the component describing the strongest diffracting direction will be large and negative. Regardless of the degree of anisotropy, the sum of the three components is constrained to add up to zero. The anisotropic delta-B is defined as the difference between the two components with the most extreme values (Trueblood *et al.*, 1996). An anisotropic delta-B over 10 Å^2 indicates mild anisotropy, an anisotropic delta-B over 25 Å^2 indicates strong anisotropy and an anisotropic delta-B over 50 Å^2 indicates severe anisotropy (Sawaya, 2014).

3.1.2. Anisotropy correction

In the absence of a model for the contents of the crystal, a first statistical correction of the anisotropy can be done for structure solution. Anisotropic correction is done using the method of Popov & Bourenkov (Popov & Bourenkov, 2003). The six anisotropy parameters of the anisotropy tensor and a scale factor are determined by refinement to maximize the likelihood function given by the Wilson distribution of the data. This essentially weights up the reflections in the poor direction and attenuates the strong direction, removing the anisotropy so creating an isotropic data set.

The first problem for using anisotropic data in phasing and refinement is the inclusion of numerous poorly measured reflections in the data when there is a single (isotropic) resolution cut-off. Ideally, anisotropic correction terms should be applied to both structure-factor amplitudes (*F*) and their errors (σ_F), to obtain corrected normalized structure-factors amplitudes (*E* values) and their errors (σ_E values), and target functions for phasing and refinement should take into account the sigmas in weighting the contribution from each reflection. If the target functions for phasing and refinement that do not account for the sigmas, this procedure is only acceptable as long as the anisotropy is not too severe. If the anisotropy is severe, target functions that do not make effective use of the sigmas should be used with data that have been subjected to ellipsoidal truncation (Sawaya, 2014; Strong *et al.*, 2006; Tickle *et al.*, 2018). In this, an ellipsoidal resolution boundary is imposed on the data so that the weak reflections falling outside this boundary will be discarded from the data set. Alternatively, data with low information content may be removed explicitly.

When calculating electron density maps, a blurring or sharpening factor is normally applied to improve map connectivity and interpretation. If the data is isotropic, then the sharpening factor is normally the Wilson B value. For anisotropic data, it is common to either use the average value of the anisotropic tensor, or the most negative value of the tensor (Sawaya, 2014; Strong *et al.*, 2006).

3.1.3. Anisotropic refinement

In the presence of an atomic model for the contents of the crystal, the anisotropy can be refined more accurately. There is the need for an anisotropic scale factor for comparing the observed (F_{obs}) and the calculated structure factors (F_{calc}) (Usón *et al.*, 1999). The agreement between F_{obs} and F_{calc} is very poor if F_{obs} has a directional dependence and F_{calc} does not.

At resolution of ~ 1.4 Å or higher is it possible to refine individual atomic B-factors anisotropically. The anisotropic displacement parameters (ADPs) provide a description of the anisotropic displacement of an atom from its mean position. The description of an anisotropic B-factor requires six parameters instead of one parameter for the isotropic B-factor (Merritt, 1999; Sheldrick & Schneider, 1997). A difficult task in the parameterization of macromolecular structure models is accounting for correlated dynamic or static displacement, that is the anisotropic motions of the atoms. For example, long side chains have more freedom to swing perpendicularly to bond directions than along the much more tightly restrained bond length. This can be accomplished adding restraints to model a physically sensible behaviour (Sheldrick & Schneider, 1997; Thorn *et al.*, 2012).

An economic description of correlated atomic motion intermediate between individual ADPs and isotropic B-factors is the TLS (Translation-Libration-Screw) parameterization, which describes the movement of whole molecules as rigid bodies within the crystal lattice. This parameterization is appropriate at resolutions lower than 1.4 Å as the TLS description allows to parameterize the model in a physically sensible form with significantly (orders of magnitude) fewer parameters than a description with individual ADPs would require: the TLS parameterization contributes with only 20 parameters per molecule (Murshudov *et al.*, 1998).

3.2. Twinning

A formal definition of twinning is the following: "Twins are regular aggregates consisting of crystals of the same species joined together in some definite mutual orientation" (Giacovazzo, 2002). Twinning arises when the crystal is composed of separate domains of differing orientations related by a symmetry operation (Yeates, 1997), so that reciprocal lattices of twin domains overlap in at least in one dimension (Parsons, 2003). This is produced by a defect of crystal growth, which may be caused when the crystal grows too quickly or is subjected to abrupt variations in the crystallization conditions, such as changes of temperature or pressure, although the cause is not always known.

Twinning is a relatively common phenomenon in protein crystallography (Lebedev *et al.*, 2006) and can be a potentially dangerous crystal pathology, because it can easily be overlooked, hampering structure determination and refinement. It can be a problem as a twinned crystal does not produce a simple diffraction pattern, the recorded data is the sum of the crystal domains in different orientations. On the other hand, if twinning is identified, it is generally a manageable situation and structure determination is often possible (Thompson, 2017).

Twinning can be characterized by the twin law and the twin fractions. The twin law is a set of symmetry operators that relate the different orientations of the twin domains, in protein crystallography the only possible twin laws are rotation axes. The twin law can be expressed as a matrix that transforms the hkl indices of one species into the other. The twin fraction quantifies the fractional volume of the crystal occupied by the twin domains, the sum of the twin fractions for all twin domains must be one (Campeotto *et al.*, 2018; Parsons, 2003). In addition to describing volumes in real space, the twin fraction has an important manifestation in reciprocal space, as the contribution to the observed diffraction intensity from each reciprocal lattice is weighted by its twin fraction (Thompson, 2017).

3.2.1. Types of twinning

Crystal twinning can be classified under four types, which are defined according to the specific way in which the twin domains are oriented relative to one another: non-merohedral, merohedral, pseudo-merohedral and by reticular merohedry (Yeates, 1997).

Non-merohedral twinning

In non-merohedral twinning, the twin domains are oriented in a way that produces an overlapping of the crystal lattices in two-dimensions. Therefore, the reciprocal lattices do not overlap exactly, so this situation produces diffraction patterns where only a subset of reflections from the twin domains overlap (Herbst-Irmer & Sheldrick, 1998; Sevvana *et al.*, 2019), as shown in figure 3. The twin law does not belong to the crystal class of the structure or to the metric symmetry of the cell, it is an arbitrary operator (Thompson, 2017).

A large number of systematic absences and the appearance of one or more unusually long axes can be indicators of non-merohedral twinning, this may cause problems with cell determination and indexing, since more than one orientation matrix is needed to index all reflections (Herbst-Irmer & Sheldrick, 1998). It can usually be identified at the data collection, where software for integrating these data sets, such as *TWINABS*, can be employed (Sheldrick, 2002a). Structures arising from crystals where the pathology is addressed are rare in the PDB.

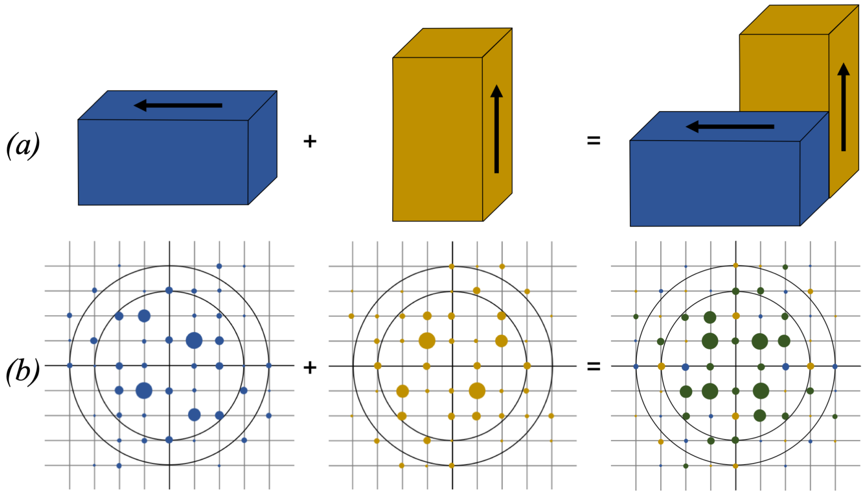


Figure 3. Non-merohedral twinning with two domains that have a unit cell axis which is exactly twice as long as a second axis. a) The relationship of the unit cells in different domains is a 90° rotation. b) Diffraction patterns from the two different domains in the crystal (blue and yellow points), and the diffraction pattern produced when the reciprocal lattices are rotated and superimposed, where only some reflections are overlapped (green points). Image adapted from (Thompson, 2017).

Merohedral and pseudo-merohedral twinning

In merohedral twinning, the twin domains are oriented in such a way that produces an overlapping of the crystal lattices in three-dimensions (Dauter, 2003). Therefore, the reciprocal lattices of the different twin domains are perfectly superimposable, and the twinning is not directly detectable from the reflection pattern (figure 4). The twin domains are related by rotations that are symmetry operations of the crystal system but not of the point group itself. Thus, the rotational symmetry of the crystal lattice must be higher than that of the space group. This is possible for certain space groups in tetragonal, trigonal/hexagonal and cubic lattices (Thompson, 2017).

Additionally, pseudo-merohedral twinning can occur for lattices with lower symmetry if their unit-cell parameters are close to fulfilling the higher symmetry crystal system requirements. For example, an orthorhombic crystal can become pseudo-merohedrally twinned if a ≈ b, making the lattice approximately tetragonal, or a monoclinic if β ≈ 90°, making the lattice approximately orthorhombic. The coincidence of the reflection profiles is close but not required by the crystal symmetry (Thompson, 2017). Otherwise, pseudo-merohedrally twinned crystals have characteristics identical to the classic merohedral twins.

The vast majority of reported cases of merohedral twinning are hemihedral twins, meaning that there are only two different orientations of twin domains are present (most often related by a 180° rotation) (Yeates & Fam, 1999). Higher forms of merohedral twinning then exist as for example: tetartohedral (four twin domains) and ogdohedral (eight twin domains) (Parsons, 2003; Roversi *et al.*, 2012). Hemihedral twins with a twin fraction is 0.5 are said to be "perfectly twinned" and the space group may be mistaken for one of higher symmetry. If the twin fraction is lower than half, the twinning is partial. As the twin fraction deviates from 0.5, the true space group become easier to identify.

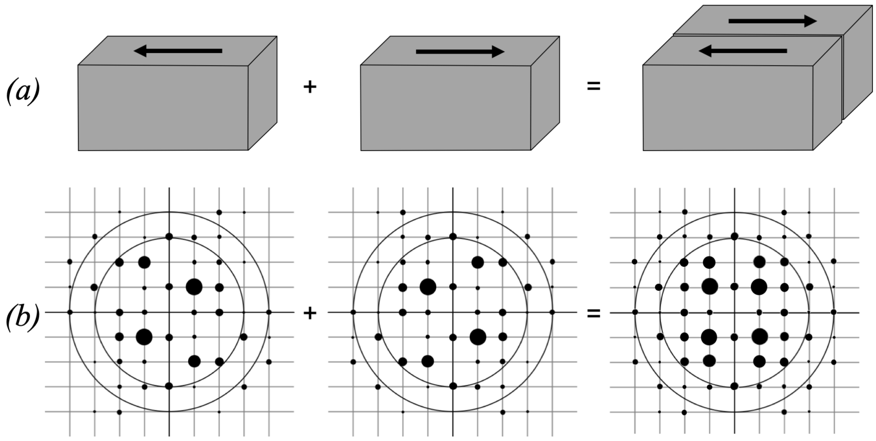


Figure 4. Merohedral twinning. a) The relationship of the unit cells in different domains is a 180° rotation. b) Diffraction patterns from the two different domains in the crystal, and the diffraction pattern with averaged twin reflections, which cause the diffraction pattern to mimic higher symmetry. The twin fraction is 0.5, that is, both domains are present in equal amounts. Image adapted from (Thompson, 2017).

Twinning by reticular merohedry

In twinning by reticular merohedry some of the reflections overlap exactly, while others are non-overlapped. The most frequent example is an obverse/reverse twin in case of a rhombohedral crystal (Herbst-Irmer & Sheldrick, 2002).

3.2.2. The warning signs for merohedral twinning

There are several characteristic warning signs of twinning. The data can appear to have erroneously high symmetry, the merging statistics for a higher symmetry Laue group may be marginally, but statistically significantly, worse than for a lower symmetry one. Occasionally, the asymmetric unit cell volume assigned to the lattice is too small for the size of the molecule or there is an unreasonably high packing density. In some cases, the systematic absences apparent are not consistent with any known space group (Herbst-Irmer, 2016). Additional signs include an inability to solve a structure even though the data have high signal to noise ratio and high-resolution, and, if a model could be built, high R-factors or very noisy difference electron-density maps (Murshudov, 2011).

3.2.3. Effect of twinning on diffraction intensities

The separate twin domains scatter X-rays independently, and therefore, each measured intensity is the sum of the intensities of the individual twin domains. In contrast, within the regular domains, the diffracted X-rays would interfere and the total scattering will correspond to the vector sum of amplitudes (Dauter, 2003). As a consequence, the intensity distribution corresponding to a twinned crystal deviates from the standard Wilson statistics (Wilson, 1949).

This can be explained as a single crystal is characterized by a certain low fraction of very weak and very strong reflections. In a twinned crystal, reflection intensities are added up in sets related by the twin laws, and there is a low probability that the twin operations connect only very weak or very strong reflections, although this will occasionally occur. The diffraction data of a twinned crystal therefore has fewer very weak or very strong intensities than predicted by the Wilson distribution (Stanley, 1972).

3.2.4. Detection and statistical tests for merohedral twinning

Since diffraction patterns originated from all twin domains overlap perfectly, twin detection can be done analysing intensity statistics following two approaches (Dauter, 2003). One is based on the overall statistical properties of diffraction data, such inspection of the Wilson ratio $<F^2>/<I>$ (Wilson, 1949), higher moments of intensity distribution or average intensity ratio $<I^2>/<I>^2$ (Stanley, 1972) and the cumulative intensity distribution N(z)-test (Rees, 1980). Furthermore, the mean value for $|E^2-1|$ is much lower than the expected values of 0.736 for the non-centrosymmetric case, being for a perfect twin 0.541 (Herbst-Irmer & Sheldrick, 1998). For these tests, it does not matter whether the data have been merged in the crystal or the lattice symmetry. The second approach is based on the comparison of twin-related reflection intensities, such as the S(H)-test (Yeates, 1988; Yeates, 1997), negative intensity Britton test (Britton, 1972; Fisher & Sweet, 1980), and Murray-Rust test (Murray-Rust, 1973). These tests require that the diffraction data are merged in the proper low-symmetry point group and can be used also to estimate the twin fraction.

Twinning tests are complicated by the fact that some other pathologies, such as translational pseudosymmetry or anisotropy if present, also perturb the intensity distribution. The L-test or Yeates-Padilla test (Yeates, 1988; Yeates, 1997) is especially robust in such cases. Furthermore, it can be performed successfully without knowing the twin operator, and it is also insensitive to data reduction in the wrong space group. The L-test is based on analysing the cumulative distribution of a ratio, |L|, which is calculated by selecting two intensities proximally located in reciprocal space but not related by any twinning operation, and dividing their difference by their sum. In particular, the cumulative function N(|L|) is linear for untwinned crystals and has a curved shape for twinned ones. The multivariate Z-score for the L-test allows the construction of empirical decision rules to detect twinning.

3.2.5. Structure solution and refinement of twinned data

Once the pathology has been identified, structure determination in the correct space group can often proceed under favorable circumstances. If the hemihedral twin fraction does not approach 0.5, data can be detwinned (Yeates, 1997). This procedure may be useful for the purposes of structure solution, although, the structures can often be solved from the original, not the detwinned data (Yang *et al.*, 2000). Nevertheless, for refinement it is preferable to refine against the original twinned data set as detwinning has the potential to introduce substantial additional error and twinning can be accounted for during refinement (Fisher & Sweet, 1980).

Molecular replacement typically works well with twinned data, although better models are generally required than for the equivalent untwinned data. Experimental phasing with single-wavelength anomalous diffraction (SAD) or multiple-wavelength anomalous diffraction (MAD) is also possible as diffraction data are usually collected from a single specimen with a constant twin fraction, but can be more difficult to both find an initial substructure and interpreting the twinned electron density (McCoy & Read, 2010). Whereas with multiple isomorphous replacement (MIR), the phasing procedure can be severely impaired, as the twin fractions of derivative and native crystals may differ significantly (Dauter, 2003). Thus, more than two derivatives may be necessary for phasing (Yeates & Rees, 1987).

In MR, if calculations are performed on the twinned data, the contrast of the rotation function decreases as it produces multiple solutions from the different orientations of the twin domains with corresponding weights corresponding with the twin fraction. In the translation function, the smaller twin domain contributes with additional noise, but the principal solution should correspond to the main domain (Dauter, 2003).

Refinement of a model against merohedrally or pseudo-merohedrally twinned data can be done using a program with an appropriate twin refinement protocol. The diffraction patterns from the twin domains overlap, so each observed intensity is a weighted sum of the twin-related domains independent crystallographic intensities according to the fractional contribution of the twin fraction. It is necessary to include the twin law relating the twin domains as an additional parameter (usually it is input in the form of a 3x3 matrix), which can then be applied to the calculated structure factors and used to refine the twin fraction. Programs that allow twinning refinement are *SHELXL* (Bernhardt & Herbst-Irmer, 2020; Herbst-Irmer & Sheldrick, 1998; Sheldrick, 2015), *CNS* (Brunger *et al.*, 1998), *Refmac* (Murshudov *et al.*, 1997) or *phenix.refine* (Afonine *et al.*, 2012).

R-factors may not be directly comparable, as they are lower than in single crystals (Murshudov, 2011). In particular, the gap between R_{factor} and R_{free} values as well as their individual values need to be monitored during refinement. Also, difference density maps might have fewer features, as the twinned reflections add noise.

3.3. Order-disorder twinning

Order–disorder twinning is a type of crystal-growth irregularity (Dornberger-Schiff, 1956, 1966; Dornberger-Schiff & Dunitz, 1965; Dornberger-Schiff & Grell-Niemann, 1961), that has been observed for protein structures in several cases deposited in the PDB (Rye *et al.*, 2007; Trame & McKay, 2001; Wang *et al.*, 2005). Order–disorder structures or OD structures arise when a single molecular configuration is maintained in a crystallographically ordered fashion in one layer, but successive layers contain the molecule in an alternative crystallographically ordered fashion. The perfect order within each layer is intercalated with a zone of transition or disorder in the mutual relation of pairs of successive layers, hence the name of the phenomenon (Pletnev *et al*., 2009).

In these structures, the differently oriented molecules are related by specific operations (translational or rotational) that break the crystallographic symmetry (Thompson, 2017). Although cases are known where the distinction between alternate molecular configurations is a difference in orientation also known as rotational OD structures (Pletnev *et al.*, 2009), most cases occur as a difference in relative position between molecules in different layers (Hare *et al.*, 2009; Tsai *et al.*, 2009; Zhu *et al.*, 2008). These cases are also termed as crystals with lattice-translocation defects (LTD) (Wang *et al.*, 2005), in which successive layers of molecules are shifted with respect to each other in a particular direction by a positive or negative fractional displacement in a more or less random fashion while preserving equivalent mutual contacts.

In translational OD structures, the layers of molecules are stacked in such a manner that two or more different stacking vectors can relate neighbouring layers to form geometrically identical interfaces between them. Depending on the sequence of stacking vectors, three types of OD structures can be classified (Lebedev, 2009). Large domains with the same internal organization are individual crystals of OD-twin (figure 5a), and the domains have the same symmetry. An allotwin (figure 5b) contains domains with different sequences of stacking vectors and different crystallographic symmetries. A structure with an irregular sequence of stacking vectors is called disordered OD-structure (figure 5c).

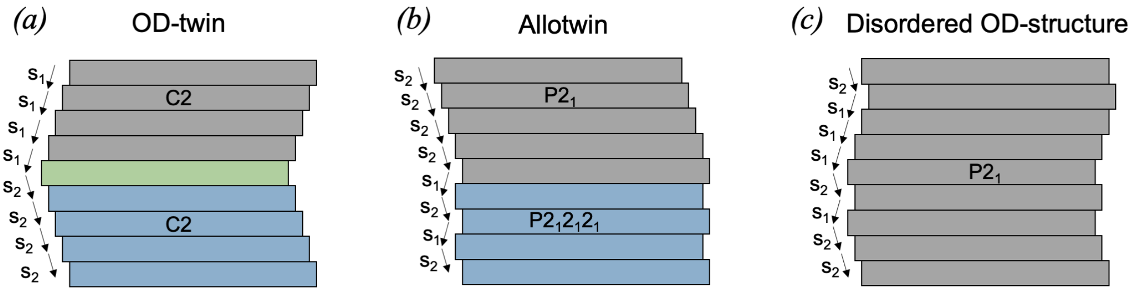


Figure 5. OD structures in which layers are related by two kinds of staking vectors, s_1 and s_2. a) OD-twin with two crystal domains (grey and blue) with local *C*2 symmetry, the intermediate layer (green) can be assigned to any of the two connected crystal domains. b) An allotwin with two different domains (grey and blue) with different crystallographic symmetries *P*2_1 and *P*$2_12_12_1$. c) A disordered OD-structure with an irregular sequence of stacking vectors. Image reproduced from (Lebedev, 2009).

Such irregularity of stacking between the layers introduces a modulation of the intensities of specific reflections. In the diffraction pattern, depending on the degree of randomness and the amount of the shifts in the consecutive layers, OD defects are manifested by the coexistence of streaked or diffused and weak reflections, and strong and sharp reflections (Bragg & Howells, 1954; Cochran & Howells, 1954; Dornberger-Schiff, 1956). Another manifestation is the presence of high non-origin Patterson peaks corresponding to the offsets of the consecutive layers from their positions. If interpreted as vectors between molecules, as in tNCS, they are too short to be a physically possible packing. Figure 6 shows an order–disorder twin crystal of L-2-haloacid dehalogenase (Rye *et al.*, 2007) and illustrates this situation. The presence of a strong-weak/sharp-streaked diffraction pattern is an important feature of this LTD, and helps to distinguish it from tNCS (Dauter & Jaskolski, 2016). Nevertheless, in some cases, structure determination and analysis of crystal packing may be required to distinguish between LTD and tNCS (Hare *et al.*, 2009). Despite the presence of high peaks in the Patterson map and modulation in the diffraction pattern, OD-twinning may be unnoticed. Although structure solution can succeed in some cases (Hare *et al.*, 2009), in others, a correction for this effect can be vital for structure solution (Trame & McKay, 2001).

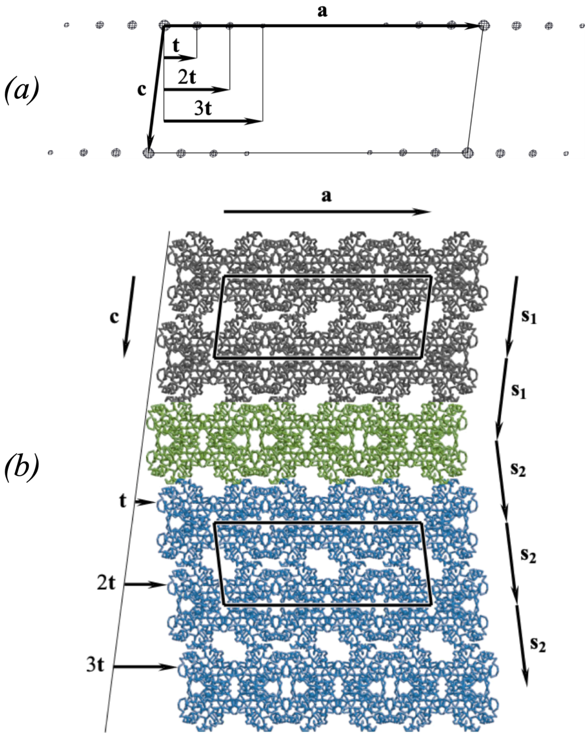


Figure 6. a) A section of the native Patterson map contoured at 4.5σ. Vectors t, 2t and 3t define the positions of non-origin peaks. b) Organization of the crystal with *C*2 space-group symmetry that has two consecutive crystal domains in which layers are related by stacking vectors s_1 (grey) and s_2 (blue). The intermediate layer (green) can be assigned to any of the two connected crystal domains. Vectors t, 2t and 3t define the offsets of three consecutive layers from their positions, these translations are in agreement with the observed non-origin peaks in the Patterson map. Image reproduced from (Lebedev, 2009; Rye *et al.*, 2007).

Furthermore, during the refinement, the map could have parts with uninterpretable density. Thus, applying a procedure to demodulate the data is advisable for the refining step. Several approaches then exist as the ones that rely on the translocation vector being an integral fraction of a unit-cell dimension (Wang *et al.*, 2005) or on the possession of a partially refined model (Tanaka *et al.*, 2008). A more straightforward procedure minimizes the intensity modulation and the non-origin peaks with a demodulation function that is an inverse of the original modulation function that directly subtracts the contribution from the additional lattice to the observed diffraction intensities (Hare *et al.*, 2009).

3.4. Translational non-crystallographic symmetry

Translational non-crystallographic symmetry (tNCS) arises when the asymmetric unit contains two or more copies of a component that are oriented in (nearly) the same way and can be superimposed by a translation that does not correspond to any symmetry operation in the space group (Rossmann & Blow, 1964).

It is particularly insidious when the tNCS operators are very close but not exactly equal to true crystallographic symmetry operators; this situation is referred to as translational pseudosymmetry and is often seen in protein crystallography (Dauter *et al.*, 2005; Zwart *et* *al.*, 2008). In the simplest form, tNCS relates a pair of components, but it can relate any number (*n*) to give *n*-fold pseudo-translation. Although the vectors between the related components in *n*-fold tNCS can differ, the components are commonly related by the same vector. If the *n* copies in the asymmetric unit are related to each other by translating one-*n*th of the unit cell in one or more directions, this type of translation is called *n*-fold commensurate modulation, and its diffraction will show a pattern of one strong reflection at every *n* spots (Chook *et al.*, 1998; Wang & Janin, 1993). An example of a two-fold pseudo-translation with commensurate modulation is shown in figure 7. In this case the tNCS operator is very close to a lattice-centring operator, the effect can be denoted as pseudo-centring (Zwart et al., 2008). This situation produces a diffraction pattern with systematically strong and weak reflections that closely approximates a space group with half the unit cell volume (Chook et al., 1998).

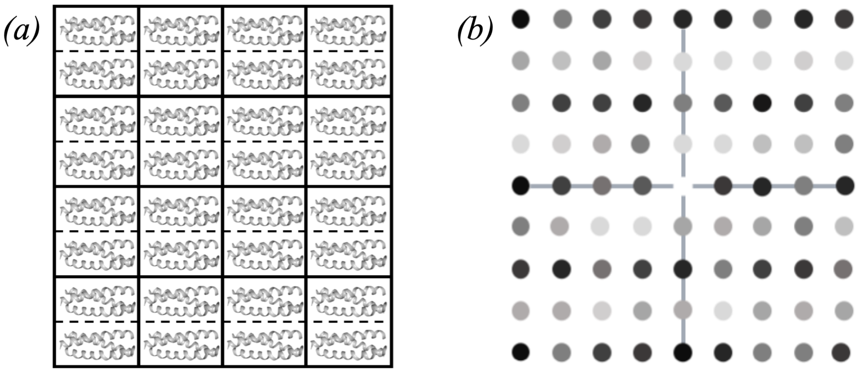


Figure 7. a) A crystal with two parallel molecules shifted by a translation vector of almost ½ of the vertical lattice (0.48, 0, 0). b) The diffraction pattern of that crystal where the reflections of the odd rows are systematically weak while those in even rows that are systematically strong. If the translation in the vertical lattice were to be exactly ½, the odd reflections would be systematically absent, and the unit cell would halve in volume.

The overall modulation of the intensities (systematically strong and systematically weak intensities) arise because the contribution to a structure-factor of molecules related by tNCS have the same (or similar) amplitudes but have relative phases determined by the projection of the translation vector on the diffraction vector, indeed the planes affected by intensity modulation are perpendicular to the translation vectors between copies related by tNCS. As a result, they interfere constructively for some reflections and destructively for others, so that there is a systematic modulation of the sum of their contributions. Whereas in the absence of tNCS, contributions of atoms from symmetry-related molecules are independent.

This effect can also be seen in the intensity distribution, for a structure without tNCS the intensity distribution follows the Wilson distribution, but when tNCS is present, this leads to a systematic broadening of intensity distribution (Read *et al.*, 2013).

The presence of tNCS can cause difficulties in all stages of crystal structure determination, from indexing the diffraction pattern to refining the structure (Read *et al.*, 2013). During the data processing, programs may misindex the reflections, reduce the diffraction images in a unit cell that is too small and/or assign an incorrect space group (Zwart *et al.*, 2008). Structure determination and refinement is problematic if the systematic modulation is not accounted for, because the intensity modulation caused by tNCS breaks the implicit assumption used in likelihood-based methods that the intensities, and the errors in predicting the intensities from the model, follow an isotropic Wilson distribution (Wilson, 1949). Without accounting for the intensity modulations, any placement of components in the same orientation and separated by the appropriate translation vector will reproduce the intensity modulation, improving the fit to the data without necessarily being a correct solution (Sliwiak *et al.*, 2015).

3.4.1. Accounting for the statistical effect of tNCS

The Patterson map can be used to determine the translation vectors between copies related by tNCS (tNCS vectors). The degree of modulation is less significant if there are rotational and/or conformational differences between the copies, and decreases with increasing resolution. For that reason, in addition to the tNCS vector it is also necessary to estimate any small rotational differences in their orientations (tNCS rotations) and the size of random coordinate differences (tNCS rmsd) caused by conformational differences in order to correctly account for tNCS modulation.

The parameters characterizing tNCS (tNCS vector, tNCS rotation and tNCS rmsd) are used to generate expected intensity factors (epsilon factors) for each reflection that model the modulations observed in the data (Read *et al.*, 2013). Note that the total expected intensity factor for a reflection includes the usual integer factor for the number of times the miller index of a reflection is identical under all the distinct pure rotational symmetry operations of the space group (Stewart & Karle, 1976). The tNCS component of the expected intensity factor that models the modulations observed in the data is non-integer (Read *et al.*, 2013), being below 1 for the systematically weak reflections and above 1 for the systematically strong reflections. After initial estimation, the parameters of the tNCS model are refined, via the expected intensity factors for each reflection derived from the tNCS model, using a likelihood function given by the Wilson distribution of the data (McCoy *et al.*, 2007).

A simulation of the probability distributions describing the statistical effects of tNCS illustrates the effects of random coordinate differences and differences in orientation on the strength of modulation for structure factors obtained from a crystal containing two spherical molecules related by tNCS (figure 8). Rotational and conformational differences between the copies can have a similar effect on the strength of the intensity modulation, but there is a direction-dependence of the effect of the rotation difference: a rotation around the diffraction vector has no effect on the modulation along the reciprocal lattice vector corresponding to the tNCS vector (as it does not change the positions of the atoms relative to the Bragg planes), whereas a rotation around an axis perpendicular to the diffraction vector has a large effect (Read *et al.*, 2013).

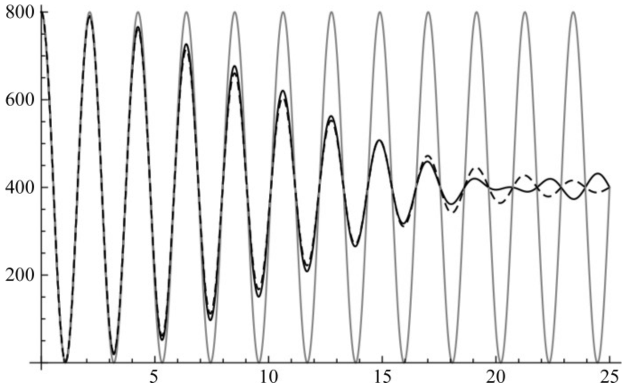


Figure 8. Predicted average intensity in the direction parallel to c* for a crystal in space group *P*1 containing two copies [separated by a fractional translation of (0.47, 0.47, 0.47), i.e., approximately body-centered] of a spherical molecule (r = 20 Å). The solid lines show when the two copies have the same conformation but differ by a 5° rotation around the x axis (black line) or around the z axis (grey line). The dashed line shows when the two copies are in the same orientation but have rmsd of 1.5 Å. Image from (Read *et al.*, 2013).

Furthermore, it is well known that the presence of tNCS can mask the effects of twinning on the intensity statistics (Zwart *et al.*, 2008). Twinning usually decreases the number of very weak intensities, this effect is offset by tNCS, which gives rise to systematically weak and strong intensities (Lebedev *et al.*, 2006). Nevertheless, it was showed that by accounting for the statistical effects of tNCS, it is possible to unmask the competing statistical effects of twinning and tNCS and to more robustly assess the crystal for the presence of twinning (Read *et al.*, 2013).

3.4.2. Types of tNCS

tNCS does not necessarily associate two components in the asymmetric unit but may relate three or more (*n*) components associated by a series of vectors that are multiples of 1, 2, 3 ... (*n*-1) times a translation vector. The order of the tNCS is called *n* and indicate it as tNCS$_n$. Where *n* times the basic translation vector equates to (or is very close to) a unit cell translation vector, the tNCS represents a pseudo-cell, and this case is known as commensurate modulation. Thus, tNCS can range from the very simple case of there being only a single tNCS vector, representing the translation between two molecular assemblies, to cases where multiples of a vector describe the translation between multiple molecular assemblies. Also, different subsets of molecular assemblies in the asymmetric unit may be related by tNCS vectors in different directions. Hence, depending on the type of tNCS that the crystal has, the tNCS correction factors that are going to be applied will differ.

The effect of tNCS on structure-factor intensity statistics (Read *et al.*, 2013) has been better characterised and novel maximum-likelihood algorithms that account for the structure-factor modulations induced by tNCS have been developed during the last years (McCoy *et al.*, 2007). These algorithms account for three different scenarios. The first case is where pairs of molecular components are related by a tNCS vector. The relationship is modelled not as a perfect translation but is rather a translation combined with a small rotation (typically less than 10º), and the tNCS related copies can have conformational differences. Hence, the parameters that are going to be refined are the tNCS vectors, tNCS rotations and tNCS rmsd. The second case is higher-order tNCS, where more than two molecules are related by multiples of the same vector. In these cases, only the tNCS vectors and tNCS rmsd are going to be refined. This necessary simplification is justified, because with more molecules, the rotational differences between the copies and the resulting intensity modulations become less significant. The third case is complex tNCS, where the tNCS copies are related by vectors in different directions. The modulations of the intensities will be much less significant, and probably structure solution will be achieved without any tNCS correction factors being applied, correcting one modulation at a time.

3.4.3. tNCS detection

Detection of translational non-crystallographic symmetry (tNCS) can be critical for success in crystallographic phasing, particularly when MR models are poor or anomalous phasing information is weak. The presence of tNCS is evidenced by the presence of a strong off-origin peak in the Patterson function, caused by the overlap of multiple parallel and equal-length inter-atomic vectors. The Patterson function can be calculated from the measured intensities alone and reveals the interatomic vectors in a structure (Patterson, 1935). The Patterson function represents a convolution of electron density with itself and corresponds to a map of vectors between each pair of atoms in the structure. Thus, when there are copies related by tNCS, each atom is related to itself in the other copy by the same translation vector, so these vectors will fall on top of each other, giving a peak of a significant percentage of the origin peak.

In *phenix.xtriage*, tNCS has been flagged as present if a Patterson function calculated with data from 5-10 Å has a peak at least 20% of the origin peak height and at more than 15 Å from the origin (Zwart *et al.*, 2005). The rationale for the resolution limits is to enhance the signal for the low-resolution molecular transform. Data are truncated since the effect of the disordered (bulk) solvent dominates at low-resolution, and the high-resolution atomic details are not necessary for this purpose. The rationale for the distance threshold is to exclude the Patterson origin peak and internal pseudo-translational symmetry such as in helices; the origin of the cell is the highest peak since every atom is at a distance zero from itself, and at distances shorter than 15 Å, artefacts can be found caused by the periodicity of peptides in secondary structure elements and these also represent predominantly intra-molecular vectors. In addition, peaks heights in the Patterson function are normalized with respect to the origin peak, the peak height in a Patterson map is expressed as a fraction or percentage of the height of the origin peak.

3.5. Modulation

A normal periodic crystal is built by repeating the unit cell by translation along the three directions of space. This translational symmetry generates a periodicity in three-dimensional space as illustrated in figure 10a. In modulated crystals, the short-range translational order from one unit cell to the next is lost, so the atomic structure can no longer be defined by the contents of a single unit cell. However, such modulations possess long-range order that can be used to restore periodicity with a periodic Atomic Modulation Function (AMF) (Lovelace *et al.*, 2013; Petricek *et al.*, 1995; Porta *et al.*, 2011). This AMF is a mathematical description of the modulation that describes the systematic or smoothly varying disorder (van Smaalen, 2007), distinguishing modulated structures from randomly disordered structures.

These modulations can be caused by several phenomena, such as displacement of atomic positions and/or occupational modulations (Schönleber, 2011), due to the weak nature of the crystal packing interactions (Janin & Rodier, 1995).

Modulated structures can be detected from the diffraction pattern from the presence of weak satellite reflections surrounding the Bragg reflections (Porta *et al.*, 2011). The strong main reflections will correspond to the underlying basic unit cell and the much weaker (and closely spaced) satellite reflections will correspond to the periodic AMF wave (Dauter & Jaskolski, 2016). For a periodic crystal, all reflections can be indexed using the three integer indices (hkl) such as:

$$\mathbf{H} = h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^* \quad (1)$$

where a*, b* and c* are the reciprocal lattice vectors of the main reflections of the basic unit cell.

The satellite reflections indicate a violation of the three-dimensional periodicity, but in a specific, regular way. The satellite reflections are at positions different from that of the reciprocal lattice position, but as they are regularly distributed, they can be indexed adding by one or more additional vectors to the reciprocal basis. So, the position of a satellite reflection from a modulated crystal is given by (van Smaalen, 2004):

$$\mathbf{H} = h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^* + m_1\mathbf{q_1} + m_2\mathbf{q_2} + \cdots + m_d\mathbf{q}_d \quad (2)$$

where the q vectors (q_1, q_2, ... q_d), called modulation vectors describe the satellite locations (the direction and distance relative to the main reflections), and *m* (m_1, m_2, ... m_d) are the satellite index that describe the diffraction order of the satellite reflection. Reflections with *m* = 0 are the main reflections, the satellites that are closest to the main reflection are the first order (m = ±1) and the next closest are second order (m = ±2) etc., as shown in figure 9.

With satellite reflections, the diffraction pattern becomes (3 + d) dimensional, where d is the number of satellite directions. The simplest and more frequent case is a four-dimensional crystal with modulation in only one direction (one AMF wave) as shown in figure 9a. Satellite reflections can be indexed by the introduction of a single q vector such that:

$$\mathbf{H} = h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^* + m\mathbf{q} \quad (3)$$

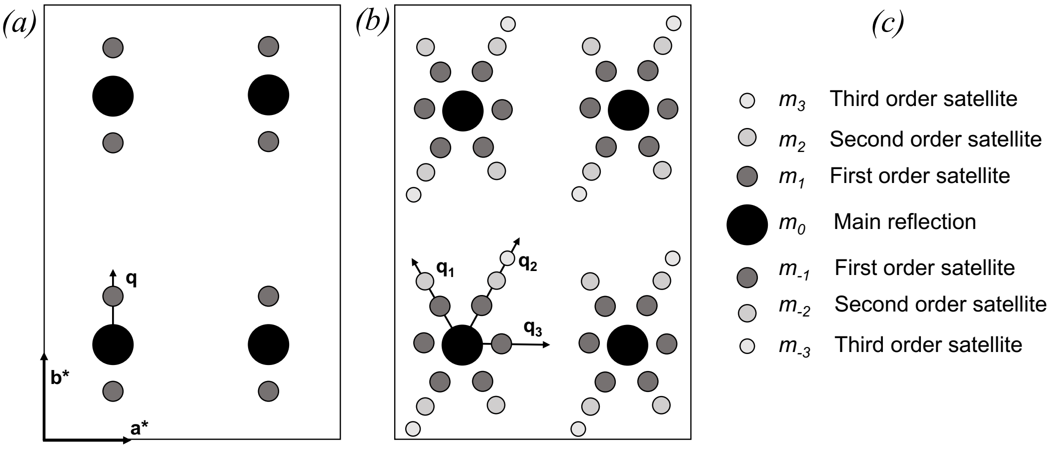

(a)
q
b*
a*
(b)
q1
q2
q3
(c)
m3 Third order satellite
m2 Second order satellite
m1 First order satellite
m0 Main reflection
m-1 First order satellite
m-2 Second order satellite
m-3 Third order satellite

Figure 9. a) A four-dimensional modulated diffraction pattern with a single q vector with first-order satellites (m = ±1) along the b* direction. b) A six-dimensional modulated diffraction pattern with three q vectors. The first q vector (**q₁**) has four satellites described by m = ±2, the second q vector (**q₂**) has six described by m = ±3, and the third q vector (**q₃**) has two satellites described by m = ±1. c) Schematic representation of the main reflections (large black circles) and the satellite reflections (smaller grey circles). Image adapted from (Lovelace *et al.*, 2008; Lovelace *et al.*, 2013).

Modulation can be commensurate or incommensurate with the main lattice and can be distinguished by the spacing of the satellite reflection from the main reflection. For commensurate crystals, all components of the q vector are rational, and for incommensurate crystals, at least one component is irrational and cannot be calculated with a simple fraction (Lovelace *et al.*, 2008).

Commensurate modulation can be described with an integer-multiple relationship to the main lattice (Lovelace *et al.*, 2010), that is, that the periodicity is restored after an integer number of unit cells, as in figure 10b. In this case, the modulation can be interpreted as a case of tNCS (Dauter & Jaskolski, 2016), and the diffraction pattern can be indexed normally by three integer indices and then solved and refined using a supercell consisting of several adjacent unit cells (Wagner & Schönleber, 2009).

Incommensurate modulation cannot be described using integers along the main lattice (Lovelace *et al.*, 2010), as shown in figure 10c. Its proper description is possible within the higher-dimensional superspace approach (Wolff *et al.*, 1981), which allows the recovery of the periodicity of the crystal. This approach consists of a three-dimensional unit cell with modulation of one, two, or three additional dimensions. The additional dimension(s) are described with the q vectors. As a workaround, incommensurately modulated structures can be described not only within the higher-dimensional superspace approach but also by applying the supercell approach, while approximating the irrational component of the modulation wave vector q by a rational number (Schönleber, 2011). Finally, such structures are challenging as they are very difficult to solve and refine with currently existing software (Lovelace *et al.*, 2008; Porta *et al.*, 2011).

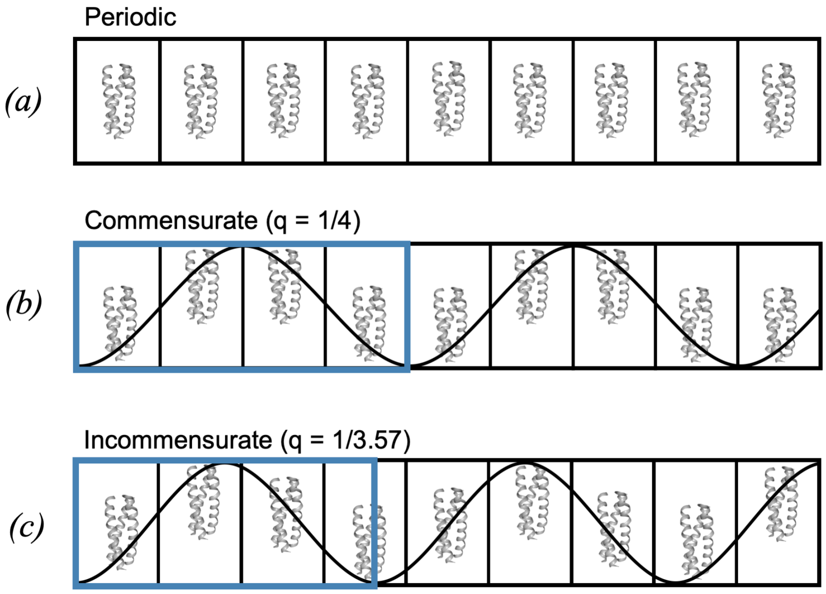


Figure 10. Three types of crystal periodicity. a) Schematic illustration of a perfectly periodic structure with identical unit cells. b) A commensurately modulated crystal where the periodicity

is restored after four unit cells, with a q vector = 1/4. c) An incommensurately modulated crystal with a q vector = 1/3.57, an irrational number, here the modulation period is 3.57 unit cells, so in this case, the structure will never repeat on a unit-cell boundary. In b) and c) the sinusoidal curves represent the AMF and the blue rectangles the modulation period. Protein structure from PDB id 3v86. Image adapted from (Porta *et al.*, 2011).

While modulation is a well-studied phenomenon in small-molecule crystallography in macromolecular protein crystallography it is rarely reported, and as a consequence, structural modulations in this context are poorly understood (Porta *et al.*, 2011). This if probably because samples with modulated diffraction patterns are discarded for being problematic and unsolvable. Nevertheless, in the last years incommensurate crystals of protein structures have been reported (Lovelace *et al.*, 2008), indexing software that can process such data has been developed (Porta *et al.*, 2011; Schreurs *et al.*, 2010), and *in silico* simulations of modulated structures have been performed (Lovelace *et al.*, 2013).

OBJECTIVES

The overall objective of the thesis is to better understand pathologies that results in highly modulated intensities in macromolecular crystal structures, which pose a challenge on structure solution. This goal has been pursued through two main studies.

The first objective was to carry out large-scale studies on crystallographic pathologies that are present in these highly modulated structures, specifically on translational non-crystallographic symmetry. In particular, it has focused on the following aspects:

- Develop an optimal way to characterize tNCS and determine the initial parameters for the model of tNCS so that the refinement of tNCS intensity correction factors can proceed and the statistical effect caused by the modulation in the data can be corrected in *Phaser*.
- Test the traditional parameters to detect tNCS. tNCS has been flagged as present if a Patterson function calculated with data from 5-10 Å has a peak over the 20% of the origin peak height and at more than 15 Å from the origin.
- Determine the degree of modulation in the data.

The second objective is to analyse a set of highly modulated macromolecular structures deeply. In proteins, paradigms of these structures are coiled-coils. Thanks to the solution of the structure of these proteins, we can better understand their biological performance that has a great biomedical and biotechnological relevance. It has focused on the following aspects:

- Identify the difficulties and bottlenecks in coiled-coil phasing, trying to solve each of the structures in a diverse test set of coiled-coils and analysing the performance of the software by comparing its results against the known structures.
- Propose the improvements required to achieve success and implement a specific way to solve general cases of coiled-coils structures.
- As these highly helical structures frequently render resolution worse than 2.5 Å, which is the resolution limit currently set for *ARCIMBOLDO*, the objective is to extend this limit to at least 3 Å resolution. To this end, another aim is to develop a strategy for discriminate solutions at low-resolution and thus validate these solutions.
- Distribute the developed software making it available to the crystallographic community.
- Use the new methods and implementation to phase unknown structures.

MATERIALS AND METHODS

1. Computing setup

The hardware used in this study is described for reproducibility and to provide a framework for performance quantification.

The tNCS calculations were performed in multiprocessing on a workstation with two quad-core Intel Xeon processors X5560 at 2.80 GHz and 24 GB RAM, and on an eighteen-core workstation with Intel(R) Core(TM) i9-9980XE at 3.00 GHz and 64 GB RAM, both with the operating system Debian GNU/Linux 9.

The coiled-coil tests were run on the eight identical eight-core machines of an HP ProLiant BL460c blade system, using them as single, independent workstations with dual quad-core Xeon E5440 processors at 2.83 GHz and 16 GB RAM, and with Debian GNU/LINUX 8.4 operating system. *ARCIMBOLDO_LITE* adjusts the calculations to the available hardware, so that a problem, which failed to be solved on a given setup, might have been solved on a more powerful workstation or on a grid. For example, in multiprocessing, the number of *SHELXE* expansion jobs is equal to the number of physical cores available in the workstation minus one. In contrast, in the case of distributed grid computing, the default is to expand 60 solutions. Hence, additional tests were run on a machine with two 12-core Xeon processors (E5-2680; 2.5 GHz and 128 GB RAM).

The unknown coiled-coil structures described in the results section 2.6.1. were solved in a supercomputing frame, distributing calculations over a grid with HTCondor v.8.4.5 (Tannenbaum *et al.*, 2001) integrated by a maximum of 160 nodes adding up to 250 GFlops. The structures described in the results section 2.6.2. and 2.6.3 were solved on a single machine with six-core Intel processors (i7-3930K, 3.20GHz, and 16 GB RAM) with operation system Debian GNU/Linux 9.

2. Software versions

2.1. tNCS software

The database was generated and queried with SQLite3 (Hipp *et al.*, 2015). The data stored in the database were loaded with Python scripts (version 2.8) written for this purpose, calculating in multiprocessing.

The atomic coordinates of structures deposited with the PDB were analysed and tNCS, if any, was identified using the ncs package from the mmtbx module of the Computational Crystallography Toolbox (cctbx) (Grosse-Kunstleve *et al.*, 2002). In this algorithm, chains with high sequence identity were identified. Then, these were structurally superimposed, testing each crystal symmetry operation, including the identity, and if they superimposed with a translation, the pair was added to a growing list of tNCS-related chains in the asymmetric unit. The translation can include a rotational tolerance defined by an angular threshold. After all combinations of sequence-matched chains and symmetry operations had been considered, the list was analysed to find the largest tNCS order. Importantly, the analysis forced the tNCS related molecules to form a closed group; so, for example, if the rotational tolerance was 3°, and A superimposes on B with a 2° rotation, B superimposes on C with a 2° rotation and A superimposes on C with a 4° rotation, then A, B and C form a tNCS group order three even though A and C do not superimpose within the tolerance of 3°. In the limit of high angular tolerances, high order rotational symmetry may be misidentified as high-order translational symmetry (e.g., PDB id 2gtt (Albertini *et al.*, 2006)). The package reports the chain identifier of the tNCS related chains, the tNCS vector in fractional and orthogonal coordinates, the rotational difference, and the percentage of total scattering for the pairs of molecules related by tNCS.

The Patterson function was calculated from the deposited data with *Phasertng* (McCoy, 2020). Where mean intensities were available, reflections recorded as net positive were used for the calculation. If only anomalous intensities were available, a mean intensity was calculated as a simple average of the Friedel mates, or using the singleton intensity if only one Friedel mate was present. If only structure factor amplitudes were available and these had been generated by the French and Wilson (French & Wilson, 1978) procedure, then the transformation was reversed to obtain intensities (Read & McCoy, 2016); the information loss meant that reflections with negative experimental intensity were set to zero intensity. If only structure factor amplitudes were available and these had not been subjected to the French and Wilson algorithm, the intensity was taken as the square of the structure factor amplitude. All data were used without applying an $I/\sigma(I)$ selection criterion.

The tNCS correction terms were calculated with the *Phasertng* software package using algorithms like those implemented in *Phaser* (Jamshidiha *et al.*, 2019; McCoy *et al.*, 2007; Read *et al.*, 2013; Read & McCoy, 2016; Sliwiak *et al.*, 2014). When the tNCS order was greater than 2, the relative orientations between the components related by tNCS were not included in the model for tNCS but their effect was absorbed approximately by the tNCS rmsd parameter. Correction terms were applied to the observed and calculated structure factors during all likelihood calculations involved in MR.

The decision tree was generated using the scikit-learn python library (Pedregosa *et al.*, 2011).

2.2. *ARCIMBOLDO_LITE*

ARCIMBOLDO_LITE is deployed for Linux and Macintosh and can be downloaded through the Python Package Index (PyPI) (https://pypi.org/project/arcimboldo/) or as part of the *CCP4* program suite starting from release 7.0 (Winn *et al.*, 2011). Requires Python versions between 2.6 or newer, *Phaser* version 2.8 or higher from the *PHENIX* (Liebschner *et al.*, 2019) or *CCP4* (Winn *et al.*, 2011) distributions for fragment-placement and *SHELXE* (Usón & Sheldrick, 2018) version 2018 or higher from *SHELX* distribution server for density modification and autotracing.

The figures of merit used in decision making were:

- *Phaser*'s intensity-based log-likelihood gain (LLG) (Read & McCoy, 2016) explains how the model fits the data by calculating the difference between the likelihood of the model and the likelihood calculated from a Wilson distribution, so it measures how much better the data can be predicted with the model than with a random distribution of the same atoms. The more positive the LLG, the higher will be the signal in a MR search.
- Correlation coefficient between observed and calculated normalized intensities (CC) (Fujinaga & Read, 1987) was calculated by *SHELXE* (Sheldrick, 2002b). It measures the linear correlation between the native structure factors and those calculated from the partial structure.
- Structure-amplitude-weighted mean phase errors (wMPE) (Lunin & Woolfson, 1993) were calculated with *SHELXE* against the refined models from the PDB to assess performance.

2.3. Data analysis, model building and refinement

For data analysis and subsequent refinement, crystallographic software has been used notably from the program suites: *PHENIX* (Liebschner *et al.*, 2019), *CCP4* (Winn *et al.*, 2011), and *SHELX* (Sheldrick, 2008b).

XPREP v.2015/1 (Sheldrick, 2008a) and *phenix.xtriage* (Zwart *et al.*, 2005) were used for data analysis. Additionally, *Phaser* (McCoy *et al.*, 2007) was used to calculate the anisotropic delta-B factor in the coiled-coil study.

Model and maps were examined with *Coot* v.0.8.7 (Emsley *et al.*, 2010). It was also employed for manual model building in the solution of the unknown coiled-coil structures.

Phenix.refine (Afonine *et al.*, 2012) was used to refine the unknown coiled-coil structure described in section 1.6.1., *BUSTER* (Bricogne *et al.*, 2017) was used to refine the structure described in section 1.6.2., and the twinning refinement in *SHELXL* (Sheldrick, 2015) was used to refine the structure described in the results section 2.6.3.

For the structures described in the results sections 2.6.2. and 2.6.3. further completion was done modelling side chains with *SEQUENCE SLIDER* (Borges *et al.*, 2020), another program developed in our laboratory. *SEQUENCE SLIDER* models side chains on partial polypeptide traces in a brute force approach. At resolution of 2 Å or worse, the electron density maps generated from partial *ARCIMBOLDO* solutions may not display density to distinguish side chains. All possible sequence assignments allowed by the known sequence may be assembled and individually tried. Whereas the distributed *SEQUENCE SLIDER* uses an ancillary program, *SPROUT*, to model side chains, the alpha version used in this work relied on *Scwrl4* (Krivov *et al.*, 2009). The sequence may be matched to the trace based on the secondary structure prediction to reduce the number of possibilities. Possible models are sent to refinement, and crystallographic indicators are used for discrimination. The model extension and improvement of phases for correct models allows model completion. Iteration reveals better discrimination among the possibilities evaluated.

2.4. Graphics

Figures were prepared with the *PyMOL* Molecular Graphics System v.1.2r2 (Schrodinger, 2015) and *Matplotlib* v.1.5.3 (Hunter, 2007).

3. Test data

3.1. tNCS database

The database was derived from an initial subset of 90083 crystal structures from the PDB (Berman *et al.*, 2000; Burley *et al.*, 2018) deposited between 1976 and 2018 and for which there were also deposited X-ray intensities or amplitudes.

3.2. Coiled-coils test set

The test set is composed of two pools of coiled-coil crystal structures from the PDB.

The first pool was adopted from a previous study (Thomas *et al.*, 2015) and comprises 94 cases with resolutions ranging between 0.9 and 2.9 Å, sizes between 15 to 618 residues distributed in the asymmetric unit in one to four chains that belong to 32 different space groups in which $C2$ predominates, followed by $P2_12_12_1$ and $P2_1$. They were deposited with the PDB between 1997 and 2012. Eight structures, PDB entries 1s9z, 2pnv, 3h00, 3h7z, 3ra3, 3s0r, 3v86, and 4dzk, are annotated as merohedrally twinned.

The PDB entries for these 94 structures are: 1byz, 1d7m, 1deb, 1env, 1ezj, 1g1j, 1gmj, 1jcd, 1k33, 1kql, 1kyc, 1m3w, 1m5i, 1mi7, 1n7s, 1nkd, 1p9i, 1s35, 1s9z, 1t6f, 1uii, 1uix, 1usd, 1wt6, 1x8y, 1y66, 1ybk, 1yod, 1zv7, 1zvb, 2akf, 2b22, 2bez, 2efr, 2fxm, 2ic6, 2ic9, 2no2, 2ovc, 2pnv, 2q5u, 2q6q, 2qih, 2v71, 2w6a, 2w6b, 2wpq, 2xu6, 2xus, 2xv5, 2ykt, 2zzo, 3a2a, 3ajw, 3azd, 3bas, 3cve, 3cvf, 3etw, 3h00, 3h7z, 3hfe, 3hrn, 3k29, 3k9a, 3ljm, 3m91, 3mqc, 3ni0, 3okq, 3p7k, 3pp5, 3q8t, 3qh9, 3ra3, 3s0r, 3s4r, 3s9g, 3swf, 3swk, 3swy, 3t97, 3trt, 3twe, 3tyy, 3u1a, 3u1c, 3v86, 3vgy, 3vir, 3vp9, 4dzk, 4dzn and 4e61.

This test set was expanded with a second pool of 56 structures selected from the PDB in the range of 2-3 Å resolution and sizes spanning 45-635 amino acids arranged in one to 12 chains. These structures, which were deposited in the years 2001–2016, belong to 26 different space groups, with $P2_1$, $C2$ and $P2_12_12_1$ predominating. Three of them, PDB entries 3miw, 4bl6, and 5ajs, are merohedrally twinned.

The PDB entries for these 56 structures are as follows: 1kdd, 1pl5, 1t3j, 1u4q, 1unx, 1urq, 1w5h, 2ahp, 2b9c, 2jee, 2nps, 2o1j, 2oqq, 2wz7, 3a7o, 3cyo, 3efg, 3g9r, 3iv1, 3m9h, 3miw, 3nwh, 3onx, 3r3k, 3r47, 3r4h, 3thf, 3tul, 3v2r, 4bl6, 4bry, 4cgc, 4gif, 4hu6, 4l2w, 4ltb, 4m3l, 4n6j, 4nad, 4oh8, 4pn8, 4pn9, 4pna, 4pxj, 4pxu, 4qkv, 4w7y, 4xa3, 4yv3, 5ajs, 5c9n, 5cx2, 5d3a, 5djn, 5eoj and 5jxc.

The joint set thus covered 0.9-3 Å resolution, asymmetric unit contents ranging from 15 to 635 amino acids, and 38 different space groups. No isomorphous structures were included, although PDB entries 3mqc and 3nwh are closely related. Table 1 characterizes both test sets.

Table 1. Characteristics of the test sets used.

	Test Set 1	**Test Set 2**
Number of structures	94	56
Range of resolution limit (Å)	0.9 - 2.9	2.0 - 3.0
Residues in the asymmetric unit	15 - 618	45 - 635
Polypeptide chains in the asymmetric unit	1 - 4	1 - 12
Nº of different space groups	32	26
Most frequent space groups (and presence in the test set)	*C*2 (13.8%) $P2_12_12_1$ (12.8%) $P2_1$ (10.6%)	$P2_1$ (17.9%) *C*2 (14.3%) $P2_12_12_1$ (7.1%)
Used previously in	(Thomas *et al.*, 2015)	

3.3. Unknown coiled-coil structures

Four novel coiled-coil crystal structures were solved in our group thanks to the *coiled_coil* mode implemented in *ARCIMBOLDO_LITE* and their characteristics are described below.

Mus musculus **SYCP3 homotetramer in two crystal forms ($P2_1$ and *P*1)**

The structures of the synaptonemal complex protein 3 (SYCP3) from *Mus musculus* in two crystal forms ($P2_1$ and *P*1) (West *et al.*, 2018; West *et al.*, 2019) have been deposited in the PDB with the codes 6dd8 and 6dd9 respectively.

The $P2_1$ data set was generated by merging three individual data sets from different crystals collected at the Advanced Photon Source, beamline 24ID-E obtaining an R(int) of 0.2245, and cut to a final resolution of 2.5 Å. Data were used as intensities. The unit-cell parameters are a=45.89 Å, b=49.49 Å, c=150.56 Å, α=90º, β=90.79º, and γ=90º. The asymmetric unit contains a single tetramer, totalling 576 residues, along with 50% of solvent content.

The *P*1 data set was generated by merging five independent data sets collected at the Advanced Photon Source beamline 24ID-E and the Stanford Synchrotron Radiation Lightsource beamline 14-1, obtaining an R(int) of 0.1935, and cut to a final resolution of 2.2 Å. Data were used as intensities. The unit-cell parameters are a=45.84 Å, b=52.40 Å, c=75.33 Å, α=94.73º, β=103.99º and γ=110.47º. The asymmetric unit contains a single tetramer, totalling 576 residues, along with 47% of solvent content.

A peptide-based nanotube with tNCS

The data set was generated by merging all equivalents (including Friedel pairs) to give 7142 reflections and an R(int) of 0.2245 and an R(sigma) of 0.058. The resolution of the data set was 2 Å, and data were used as intensities. The crystal belonged to space group *P*1, with unit cell parameters a=23.48 Å, b=27.92 Å, c=45.43 Å, α=93.60º, β=90.81º and γ=113.13º. The asymmetric unit contains four helices, totalling 128 residues, corresponding to a solvent content of 36%.

A peptide-based nanotube with twinning

The resolution of the data, available as intensities, was 1.3 Å. The crystal presented space group *R*3 with the hexagonal unit cell of a=b=40.45 Å, c=59.34 Å, α=β=90º and γ=120º. The asymmetric unit contained one helix of 36 residues, corresponding to a solvent content of 44%.

RESULTS AND DISCUSSION

1. DETECTION OF TRANSLATIONAL NON-CRYSTALLOGRAPHIC SYMMETRY IN PATTERSON FUNCTIONS

1.1. Introduction

Translational non-crystallographic symmetry (tNCS) arises when the asymmetric unit contains two or more copies of a component that are oriented in (nearly) the same way and can be superimposed by a translation that does not correspond to any symmetry operation in the space group (Rossmann & Blow, 1964).

This causes an overall modulation with systematically strong and weak intensities (Chook *et al.*, 1998), affecting structure determination and refinement. The maximum-likelihood methods used for MR depend on an accurate statistical model and they are highly sensitive to the difficulties to account for the statistical effects of tNCS. For that reason, it is fundamentally important to address this pathology for structure solution.

To characterize the statistical effects of tNCS accurately, it is necessary to determine the translation relating the copies (tNCS vectors), and the size of random coordinate differences caused by conformational differences from exact translation. In the case of a pair of molecules related by tNCS, the algorithm implemented in *Phaser* models a small rotation (tNCS rotation) and a rms deviation (tNCS rmsd). For higher-order tNCS, rotation differences are not modelled explicitly. These parameters are used to generate expected intensity factors for each reflection that model the modulations observed in the data (Read *et al.*, 2013), which are refined against a likelihood function (McCoy *et al.*, 2007) given by the Wilson distribution of the data.

The Patterson map can be used to determine the translation vector(s) relating the copies. To this end, the Patterson map is traditionally calculated with data truncated from 5-10 Å and inspected to find a Patterson peak over the 20% of the origin peak height and at more than 15 Å from the origin (Zwart *et al.*, 2005). However, there has been no systematic study of the parameters underlying this approach, nor an assessment of how accurate it is in the detection of tNCS. Also, this approach does not automatically give the order of the tNCS, which is critical for correcting the modulations. We are also interested in ranking alternative hypotheses for tNCS, in the context of developing automated structure solution strategies.

1.2. Database curation

The database was derived from an initial subset of 90083 crystal structures from the PDB. Curation of this database was essential to carry out this study, to ensure control on tests and correct rating of results. Curation included the following checks on data quality:

Retracted entries were deleted, and obsolete structures were replaced by their respective valid entries as of October 2018. Also, a small subset of structures for which our scripts failed was substituted with data or coordinates from PDB_REDO (Joosten *et al.*, 2009) whenever possible, if that solved the issue, or else deleted without further examination of the causes.

When PDB entries contained MTRIX cards that represent NCS operators, to reconstruct the crystallographic asymmetric unit, the *phenix.pdb.mtrix_*reconstruction script was used (Liebschner *et al.*, 2019). The first matrix is a unit matrix and it is trivial since it corresponds to the deposited coordinates in the PDB file. Other MTRIX records will appear only when other transformations are required to generate the entire asymmetric unit. These transformations operate on the coordinates to apply non-crystallographic symmetry operations.

Similarly, some structures required taking into account the operations present on their SCALE remarks that contain the transformation required to place the model in the asymmetric unit, these remarks have the transformation from the orthogonal coordinates to fractional crystallographic coordinates.

Furthermore, data in the form of unmerged intensities were converted to merged intensities with *phenix.reflection_file_converter* using the *--non-anomalous* option (Liebschner *et al.*, 2019).

Additionally, various structural classes with characteristic high intensity modulation even in the absence of tNCS as collagen, structures containing nucleic acids, or highly α-helical proteins (75% or more helical content), such as coiled-coils were excluded. The helical content was calculated following the distribution of characteristic vectors (CVs) (Medina *et al.*, 2020) defined by the centroids of α-carbons and carbonyl oxygens from consecutive and overlapping heptapeptides. The intensity modulations generated by the helical repeats in these structures cannot be corrected by modelling them as tNCS-generated modulations, and so are beyond the scope of this study.

Also excluded from the database were viruses, small non-proteins (antibiotics and peptides), structures that have been refined as ensembles, disordered structures with a mean occupancy less than 0.75, and structures where only the C-α atoms are deposited.

Finally, since the tNCS modulations of intensities becomes less pronounced at high resolution, where data extended to high-resolution they were truncated to 3 Å resolution to save run time in the calculations. Also, in the course of this study, initially unforeseen criteria were found to play a role. This is the case for data completeness, as will be reported. To account for our findings, cases with completeness below 80% were segregated in the database. So, the primary database was further curated to remove cases where the data were less than 80% complete, and a separate database maintained to further study the effects of incompleteness.

The final curated database contains 80482 structures. Its characteristics and genesis are summarized in table 2. The small database of structures with data completeness less than 80% consisted of 1294 cases.

Table 2. Summary of database curation. The removed structures are shown in red and the replaced ones in blue.

Initial database	**90083**
Obsolete pdb files	296
Substituted by data from PDB_REDO	357
Failure of our scripts and not in PDB_REDO or still error	331
MTRIX | The script could not apply the MTRIX cards	15 | 2
SCALE	16
Structures refined as ensembles	79
Disordered structures, mean occupancy < 0.75	92
C-α-only structures	21
Contains nucleic acids	5445
Highly helical structures (coiled-coils, transmembrane proteins...)	1712
Collagen	32
Virus	202
Antibiotics	36
Peptides	59
Overall completeness below 80%	1294
Final database	**80482**

It is worth mentioning that although this is a vast database, it is biased to the structures that have been solved. Prior to the correction of maximum likelihood target functions for tNCS modulations, structure solution was severely hindered by this pathology, and MR solution was usually possible only in the cases where good models and good data were available, or where the tNCS intensity modulation less severe.

1.3. tNCS in real space

The first question to arise when studying tNCS is “What constitutes tNCS?” This is not a simple question to answer. The effects of tNCS form a continuum between exact tNCS and molecules in the asymmetric unit oriented with large rotation angles with respect to one another (general NCS).

Our initial approach was to use the coordinates for decision making. Whether or not coordinates have tNCS depends on the choice of a rotational tolerance. In our experience in tNCS parameter refinement, tNCS rotations can refine to values up to 10º (Read *et al.*, 2013). Coordinate analysis was therefore carried out exploring a wide range of rotational tolerances, from 0º to 20º. The results are shown in table 3.

Table 3. Results of the coordinate analysis depending on different rotational tolerance ranges (cumulative). The results show the number of structures with tNCS and the percentage of the total database (including monomeric structures), the number of structures with 2 molecules related by tNCS and the structures with more than 2 molecules related by tNCS.

Rotational tolerance	tNCS	tNCS order = 2	tNCS order > 2
0-2º	2523 (3.13%)	2375	148
0-5º	4818 (6%)	4332	486
0-10º	7503 (9.3%)	6660	843
0-15º	9549 (11.86%)	8396	1153
0-20º	11230 (13.95%)	9822	1408

At small angular tolerances, less than 5**º**, one in 20 structures in the database were flagged as having tNCS; at 10**º** tolerance, this had increased to nearly one in ten; and by 20**º** it was one in seven. Furthermore, in some cases the order of the tNCS also increased with tolerance; 6% of the tNCS was higher-order tNCS (*n*>2) at 2**º** tolerance and 14% at 20º tolerance. Most of the increase in the order of the tNCS was observed when increasing the tolerance from 2º to 5º because higher-order tNCS often has subsets of components more closely related than others, and what, at small tolerances, appears to be complex low order tNCS reduces to a simple high order tNCS at larger tolerances, for definition of these categories see the introduction section 3.4.2. We refer to the coordinates-based test for tNCS as the pdb-tNCS(*r***º**), where the angle *r* is the angular tolerance, and the value is true/false.

1.4. Patterson vector length threshold

Patterson function intra-molecular vectors cluster around the Patterson function origin peak. These peaks, which constitute noise in the context of searching for tNCS vectors, can be excluded by setting a minimum vector length threshold. The shortest tNCS vector that is possible in any given case will depend on the shortest extent of the molecules, and this distance could be used a constraint on the tNCS vector. However, the shortest extent is not known before structure determination; only by assuming a spherical molecule could a reasonable estimate of the average molecular extent be made from the molecular weight for a completely unknown structure. Independently, there is a need to exclude short vectors because of pseudo-symmetry in secondary structure elements, such as alpha-helices and beta-sheets. The distances arising from these pseudo-symmetries are less than 15 Å, which has been used as the threshold distance for exclusion (Zwart et al., 2005).

To determine whether this distance was larger than any tNCS vectors in the PDB, the vector relating the copies was used to calculate the vector length that corresponds to the distance from the origin of the Patterson peak. The shortest tNCS vector in the database was 22.4 Å, for PDB entry 3i57 (MacKenzie *et al.*, 2009) with a fractional translation vector of (0.5, 0, 0) and a rotational tolerance of 6.7º. The structure of 3i57 is shown in figure 11a and its Patterson function in figure 11b. We conclude that the 15 Å distance from the origin of the Patterson peak is suitable for excluding self-vectors while not excluding any true tNCS vectors.

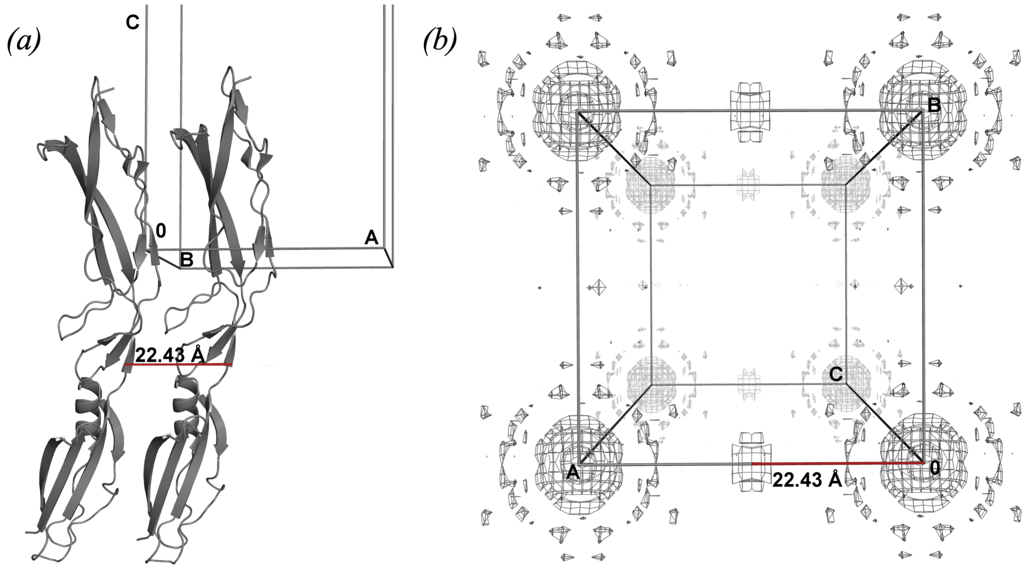


Figure 11. a) tNCS related molecules of PDB id 3i57. b) Patterson function of PDB id 3i57, drawn in 3D perspective projection, showing the origin peaks and the peak 22.43 Å from the origin, which corresponds to the tNCS translation (0.5, 0.0, 0).

1.5. Patterson peak threshold

The next step was to investigate the correlation of the pdb-tNCS with the peak heights in the Patterson function. Several resolution ranges to calculate the Patterson were explored: 3-10 Å, 4-10 Å, 5 -10 Å, 3-15 Å, 4-15 Å and 5-15 Å. As stated above, only peaks further than 15 Å from the origin peak were considered. Figure 12 shows the histograms for the distribution of top non-origin Patterson peak heights, results are shown for Patterson functions calculated with data between 5-10 Å and with different pdb-tNCS(*r*°) angular tolerances.

The top non-origin peak was expressed as a percentage of the height of the Patterson origin peak and as a Z-score value (number of standard deviations above the mean value of all the peaks). The mean height of the Patterson is disproportionately affected by the large origin peak. However, since all Pattersons have a large origin peak, the effect of this on the mean was relatively consistent across all Pattersons, and therefore the Z-score was a valid discriminator even though it was not a good absolute measure of significance.

For pdb-tNCS(2°), figure 12 showed that the traditional Patterson-20% origin peak threshold was broadly correct; this gave an accuracy (defined below) of 96%. However, for pdb-tNCS(15°) the accuracy began to break down (94%), and by pdb-tNCS(20°) was 92%.

Correspondingly, the appendix figure A2 provides graphs for other resolution ranges, a) and b) illustrate the decrease in modulation for structures with tNCS as rotational tolerance increases for all the resolution ranges used for calculating the Patterson map. Appendix figure A2 c) and d) show that the majority of the structures without tNCS have a low Patterson peak.

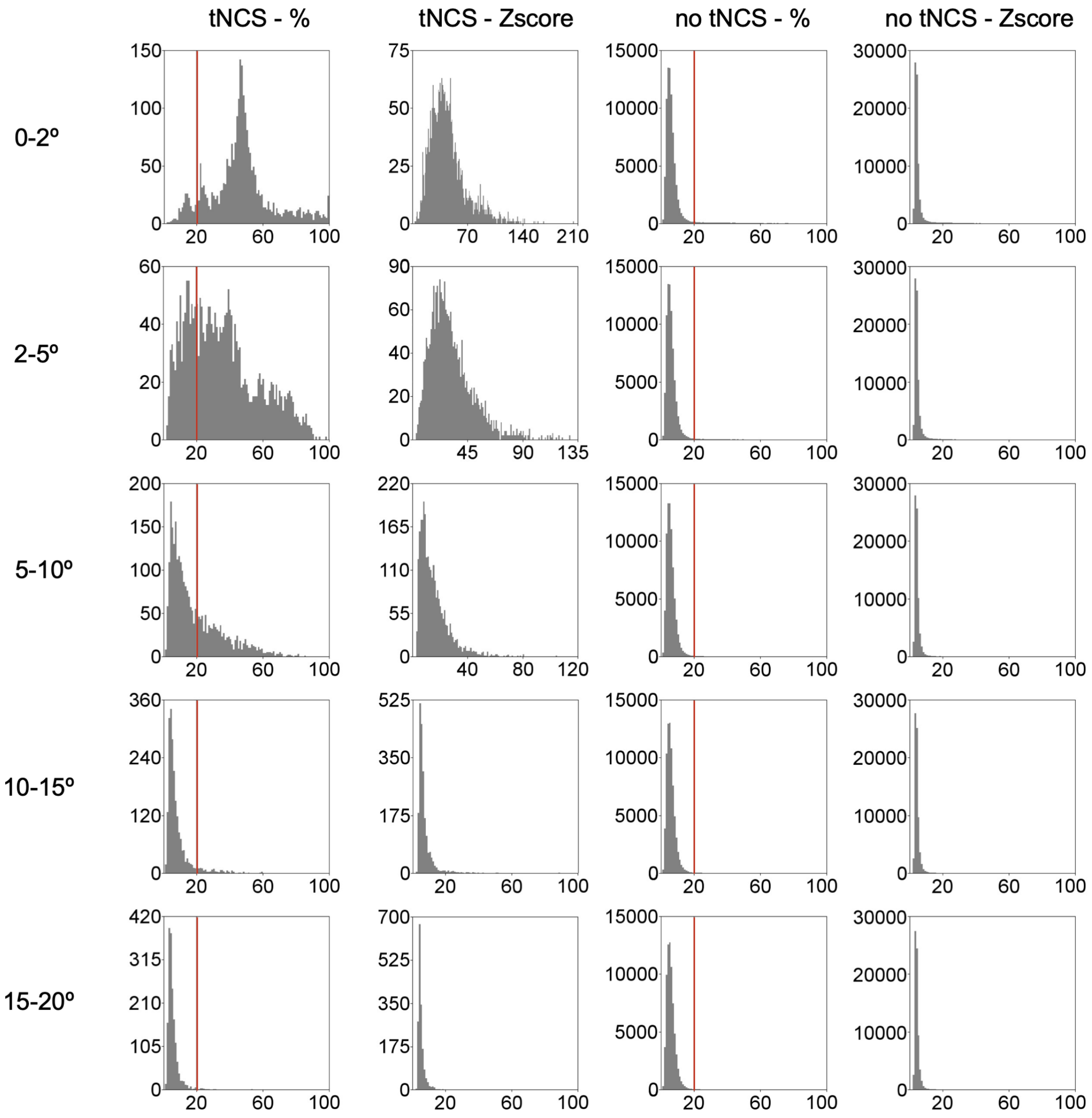


Figure 12. Non-cumulative histograms of the number of structures leading to different values for the highest non-origin peak, depending on rotational tolerances. The Patterson function was calculated with data from 5-10 Å. The first and second columns display cases with tNCS, whereas the third and fourth columns show cases without tNCS; the first and third columns express the maximal non-origin peak height as a percentage of the origin peak height, while the second and fourth columns express it as a Z-score. A red line is drawn at Patterson-20%, which is the previous threshold for determining the presence of tNCS.

1.6. Decision tree

To develop criteria for distinguishing between the presence and absence of tNCS depending on the height of the Patterson highest non-origin peak a decision tree was employed (Breiman *et al.*, 1984), which is a predictive modelling approach used in statistics, data mining, and machine learning.

The database was divided randomly into a training set (75%) and a test set (25%). The Gini index, that is a measure of statistical dispersion, was used as a criterion for calculating discrimination. A value of zero indicates no discrimination, and a value of one indicates maximal discrimination.

The training set was used to train the algorithm and included information on pdb-tNCS, and the highest non-origin Patterson peaks. The algorithm resulting from the decision tree was then applied to the test set which only had the information for the highest non-origin Patterson peak. Since there was only one parameter to fit for each decision tree (the height of the Patterson peak) cross-validation to avoid overfitting was not performed. A confusion matrix was generated in order to compute the Accuracy (ACC), Sensitivity (SN), Precision (PREC) and False Positive Rate (FPR) of the algorithm, where, given TP are true positives, TN are true negatives, FP are false positives, and FN are false negatives.

$$ACC = \frac{TP+TN}{P+N} = \frac{TP+TN}{TP+TN+FP+FN} \qquad (4)$$

$$SN = \frac{TP}{TP+\mathrm{FN}} \qquad (5)$$

$$PREC = \frac{TP}{TP+FP} \qquad (6)$$

$$FPR = \frac{FP}{TN+FP} \qquad (7)$$

The Patterson function resolution ranges explored were: 3-10 Å, 4-10 Å, 5-10 Å, 3-15 Å, 4-15 Å, and 5-15 Å. Following the study of the length of tNCS vectors, only peaks further than 15 Å from the origin peak were accepted.

Tables 4 and 5 shows that whatever the Patterson resolution or pdb-tNCS(*r*°) rotational tolerance, suitable Patterson thresholds based on either percentages of the origin peak or Z-scores could be found for high accuracy decision making; we call the associated threshold *t* values the Patterson-*t*% and Patterson-Z*t*, respectively. Smaller rotational tolerances favoured the use of higher resolution data.

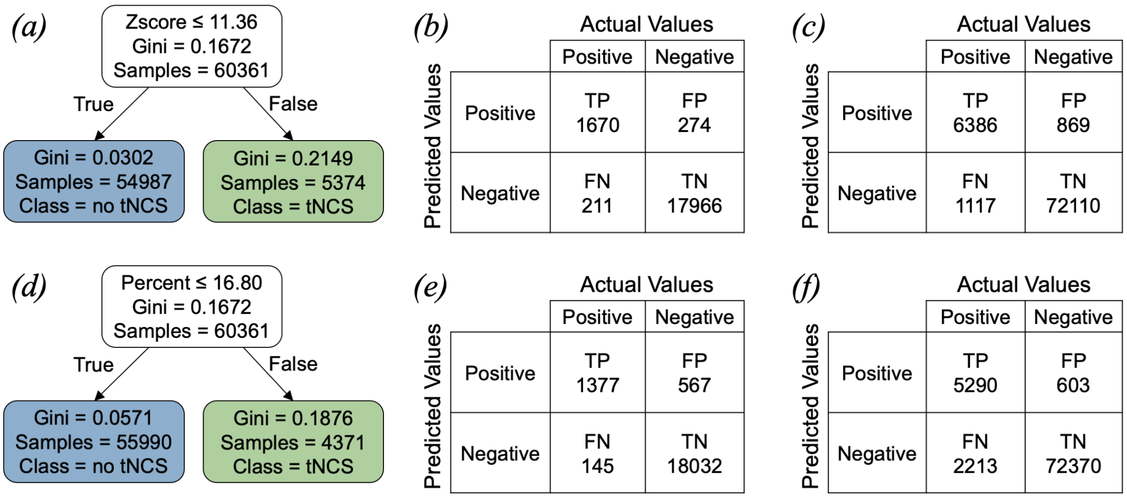
Table 4. Accuracy (ACC) of the decision trees and best value of Patterson-Zt, depending on the rotational tolerance and resolution ranges used for calculating the Patterson function. The cell highlighted in grey has the highest accuracy for pdb-tNCS(10º) and is discussed in the text (figure 13).

	0-2º		0-5º		**0-10º**		0-15º		0-20º	
	ACC	Z-score	ACC	Z-score	ACC	Z-score	ACC	Z-score	ACC	Z-score
3-10 Å	98.10	46.81	98.23	28.90	96.97	12.81	94.96	9.80	93.10	9.80
4-10 Å	97.68	33.70	98.19	20.33	97.17	11.49	95.14	10.35	93.20	9.60
5-10 Å	97.22	24.97	97.94	16.51	97.36	10.82	95.29	9.35	93.36	8.65
3-15 Å	98.03	46.91	98.23	28.82	97.07	12.86	95.31	10.09	93.28	9.57
4-15 Å	97.67	36.00	98.09	21.04	97.26	10.84	95.45	9.63	93.47	9.60
5-15 Å	97.02	26.39	97.74	17.90	97.59	11.36	95.63	9.66	93.83	9.06

Table 5. Accuracy of the decision trees and best value of Patterson-*t*%, depending on the rotational tolerance and resolution ranges used for calculating the Patterson function. The cell highlighted in grey is the prediction for the traditional resolution range of 5-10 Å for pdb-tNCS(10°) and is discussed in the text.

	0-2º		0-5º		**0-10º**		0-15º		0-20º	
	ACC	Percent	ACC	Percent	ACC	Percent	ACC	Percent	ACC	Percent
3-10 Å	97.95	28.13	97.66	15.83	95.99	8.31	94.05	7.63	91.97	8.31
4-10 Å	97.75	32.38	97.59	18.17	96.21	11.86	94.17	11.70	92.05	11.59
5-10 Å	97.34	34.39	97.37	19.85	96.46	16.80	94.39	15.40	92.31	15.53
3-15 Å	98.04	30.48	97.65	15.52	96.16	8.31	94.36	7.523	92.21	7.55
4-15 Å	97.79	34.20	97.56	18.67	96.39	11.62	94.43	10.71	92.30	10.73
5-15 Å	97.22	36.25	97.24	19.24	96.61	16.41	94.70	15.56	92.64	15.52

Taking pdb-tNCS(10**º**) as a useful measure of tNCS, the best predictions, with 97.6% accuracy (figure 13b, 13c), used Patterson functions calculated between 5-15 Å and a Patterson-Z*t* where *t*=11.36 threshold (figure 13a). Only slightly poorer accuracy, at 96.5% (figure 13e, 13f), could be obtained using the traditional 5-10 Å resolution range and a Patterson-*t*% threshold, but this required *t*=16.8% (figure 13d) rather than the traditionally used *t*=20%, implying that the previous Patterson-*t*% threshold for tNCS is too conservative. Since altering the resolution range and using a Patterson-Z*t* threshold had a marginal effect on accuracy, we decided to use the traditional 5-10 Å resolution range and Patterson-*t*% threshold for our algorithm, although with lowered threshold value from 20% to 16.8%. Using the narrower resolution range also guards against any technical problems with collecting data at low resolution.



(b)

		Actual Values	
		Positive	Negative
Predicted Values	Positive	TP 1670	FP 274
	Negative	FN 211	TN 17966

(c)

		Actual Values	
		Positive	Negative
Predicted Values	Positive	TP 6386	FP 869
	Negative	FN 1117	TN 72110

(e)

		Actual Values	
		Positive	Negative
Predicted Values	Positive	TP 1377	FP 567
	Negative	FN 145	TN 18032

(f)

		Actual Values	
		Positive	Negative
Predicted Values	Positive	TP 5290	FP 603
	Negative	FN 2213	TN 72370

Figure 13. a), b) and c) represent the best prediction using 5-15 Å resolution and a Patterson-Zt where t=11.36. d), e) and f) represent the prediction using the traditional resolution cut-off of 5-10 Å and a Patterson-t% where t=16.8%. a) and d) show the decision trees for pdb-tNCS(10°). b) and e) are the confusion matrix for the test set (25%), the accuracies calculated from the confusion matrices, using equation 4, are 97.6% and 96.5% respectively. c) and f) display the confusion matrix for the entire database (100%), where the overall accuracies given by the confusion matrices from the entire database yield 97.5% and 96.5% respectively.

The false negatives and false positives were further investigated. The sensitivity (equation 5) of the algorithm was 85% and the precision (equation 6) was 88%, while the false positive rate (equation 7) was 1%, indicating that the algorithm identifies cases of no tNCS exceptionally well, but fails to identify some cases with tNCS. With only one parameter to fit, there is a simple trade-off between identifying false negatives and false positives. The bias in the classifier towards no tNCS comes about because the database contains a higher proportion of structures without tNCS. If we assume that novel datasets will be no more biased towards having tNCS than deposited structures, then the bias is appropriate for accuracy. It is possible that the proportion of crystals that grow with tNCS is higher than that represented by the database, because these structures are less likely to be solved, however we cannot quantify this.

Both false negatives and false positives will impact structure solution by MR or experimental phasing. False negatives occurred where the Patterson peak was below the threshold proposed by the decision tree but where pdb-tNCS(*r*°) was true. False negatives will mean that intensity modulations are not corrected, and structure solution by MR will then require high-quality models to succeed, or, for SAD phasing, the anomalous signal will need to be strong. False positives occurred where the top peak in the Patterson was above the threshold but pdb-tNCS(*r*°) was false. False positives are particularly severe in the context of structure solution with Phaser because tNCS will be forced to apply to the components in the asymmetric unit (whether MR models or heavy atoms) when there is none. Therefore, the false positive rate (equation 7) of 1% was significant for practical applications even though low.

1.7. tNCS in reciprocal space, epsilon factor distribution and completeness (addressing the false negatives)

Some of the false negatives in the pdb-tNCS(10°) confusion matrix could be rescued by considering a larger angular tolerance. Indeed 353 out of 869 of the false negatives from the best decision tree (figure 13c) have tNCS according to pdb-tNCS(20°). Note that this is not equivalent to using the decision tree generated with pdb-tNCS(20°), which includes additional false negatives. This phenomenon was true for every pdb-tNCS(*r*°) analysed; false negatives could be rescued by considering larger perturbations in the rotation angles.

The studies in real space showed that using a Patterson function peak threshold gave high accuracy for detecting tNCS when using pdb-tNCS(*r*°) as the definition of tNCS. However, the optimal Patterson function peak threshold depended critically on the rotation *r* used for the classification, with the Patterson function peak threshold getting lower as *r* increased. Furthermore, an increasing number of structures that did not have pdb-tNCS(*r*°) were detected as having tNCS as the Patterson function peak threshold was lowered. The studies using the real space classifier clearly demonstrated the problem of tNCS being a continuum between exact tNCS and NCS. The problem of false negatives lay not in the threshold, but in the real space classifier of pdb-tNCS(*r*°).

There are several reasons why pdb-tNCS(*r*°) may not correspond with significant modulations in the data. If the tNCS-related components are large, the radius of the molecular G-function (Rossmann & Blow, 1962) is small so that the modulations fall off faster with orientational differences (Read *et al.*, 2013). If the tNCS-related copies differ substantially in conformation, the modulations fall off faster with resolution. Finally, if the symmetry-related tNCS vectors are very different, modulations arising from the symmetry-related copies will tend to cancel.

The scope of this study is to determine initial parameters for the model of tNCS so that the refinement of tNCS intensity correction factors can proceed. Therefore, if the resulting modulations are not significant, then tNCS is effectively not present for our purposes, and there is nothing to correct: if the (insignificant) tNCS epsilon factors are omitted there will be no impact on structure solution.

Thus, we examined the distribution of epsilon factors after refinement as a classifier for the presence or absence of tNCS. Refined epsilon factors that cluster around one define unmodulated data, while those that refine to the extremes of the distribution define high modulation. Epsilon factors measure the statistical weight of a reflection and constitute the scale factors applied for having a Wilson distribution, so if the intensity of a reflection is multiplied by one no correction is taking place. The variance about one (σ_1^2) is used as the statistical metric for measuring the degree of modulation.

$$\sigma_1^2 = \frac{1}{n}\sum\nolimits_n (x-1)^2 \qquad (8)$$

This indicator is called eps-tNCS, and takes a range of values between 0 and $(^n/_2)^2+(^n/_2-1)^2$, although in practice it is less than one in all but extraordinary circumstances. Histograms showing examples of the distribution of epsilon factors and their associated eps-tNCS are presented in figure 14.

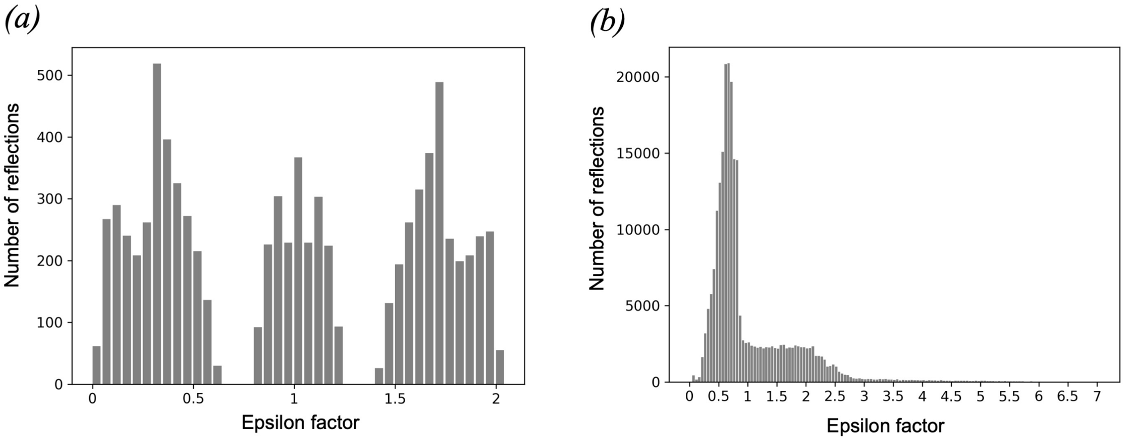


Figure 14. Histograms showing the distribution of refined tNCS epsilon factors for a) 2cc0 with σ_1^2=0.63 for tNCS$_2$ (Taylor *et al.*, 2006) and b) 4n3e with σ_1^2=0.61 for tNCS$_7$ (Sliwiak *et al.*, 2014).

The distribution of eps-tNCS values versus Patterson-*t*% is shown in figure 15. There is a clear linear relationship between the two: Patterson peak height is directly related to modulation in the data. The Patterson-Z*t* had a lower correlation coefficient (0.82) with the eps-tNCS than Patterson-*t*%. The correlation coefficient between eps-tNCS and Patterson-*t*% was 0.934, and was calculated with eps-tNCS refined against 5-10 Å data and Patterson functions calculated with 5-10 Å data.

This analysis demonstrated that the false negatives in the algorithm, as determined by pdb-tNCS (a binary measure) were cases where the eps-tNCS (a real number) was low, and therefore their mis-classification should not strongly impact structure solution. It also demonstrates that the Patterson function peak height is a good measure for the ranking of a tNCS hypothesis.

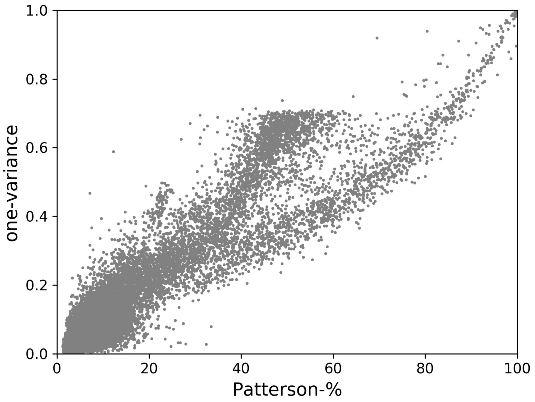


Figure 15. Scatter plot showing the distribution of refined tNCS epsilon factor variances about one (σ_1^2; one-variance; equation 8) for all cases with pdb-tNCS(20°). Data range 5-10 Å.

It has long been known that complete, good quality data are key for successful MR using Patterson methods (Navaza, 1994). In the course of the study, became apparent that the completeness of the data has a significant effect on the accuracy of the Patterson-based decision tree. Figure 16 shows that low completeness data resulted in several outliers in the Patterson-*t*% versus σ_1^2 scatter plot. Figure 17 shows how the accuracy of the decision tree deteriorated with decreasing completeness.

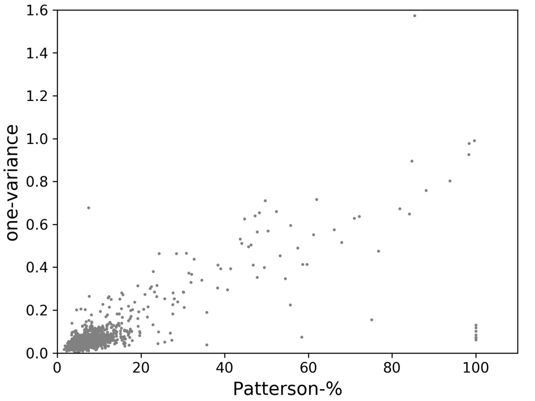


Figure 16. Scatter plot showing the distribution of refined tNCS epsilon factor variances for all cases with pdb-tNCS(20°) and data completeness less than 80%, which were excluded from the database. Many outliers are present in the distribution, with eight cases in the bottom right part of the plot (3c6o (Hayashi *et al.*, 2008), 1jpn (Padmanabhan & Freymann, 2001), 1sxh (Schumacher *et al.*, 2004), 1n8o, 1eam (Hu *et al.*, 1999), 1wwr (Kuratani *et al.*, 2005), 3it5 (Spencer *et al.*, 2010) and 1lbs (Uppenberg *et al.*, 1995)) having high Patterson peaks but no significant epsilon factor dispersion. There was one outlier in the top right part of the plot (3he1 (Osipiuk *et al.*, 2011)) with one-variance of nearly 1.6 for tNCS_6, the only case that was observed for which the σ_1^2 was greater than one.

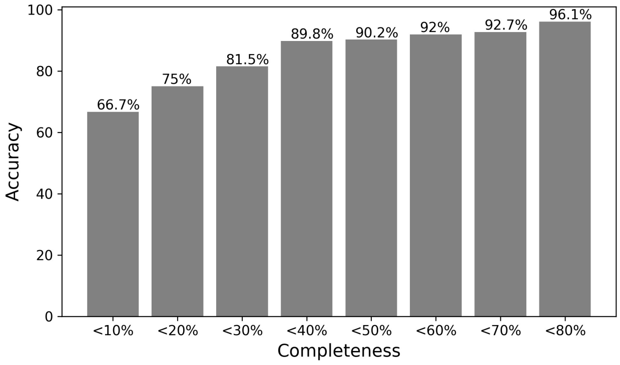


Figure 17. Histogram showing how the completeness of the data affects the accuracy of the decision tree. Low data completeness causes the algorithm to become much less reliable.

Distribution of missing data in these data sets was not investigated; however, when large percentages of data are missing, it is normally because the user has failed to collect a wedge of data, either through initial miss-identification of the true space group, radiation damage, ice rings, or severe overlapping of a section of the data (e.g., due to one long unit cell dimension). Lacking a wedge of data will impact the eps-tNCS refinement because systematic omission of data for a direction in reciprocal space leaves parameters in real space perpendicular to that direction undefined. In addition, missing wedges of data complicate data processing, and if due to overlaps, some reflections may be integrated including partial intensity from a neighbouring reflection causing a strong modulation, but also measuring intensity from neighbours can affect the weak reflections leading to missing modulation, affecting the Patterson and so the study of the modulation.

1.8. Calculated data and lattice translocation disorder (addressing the false positives)

Several entries in the database had significant Patterson peaks despite not having tNCS. For these cases, Patterson functions were calculated from the coordinates and compared with the observed ones (figure 18).

Although it is demonstrated before that the epsilon factor distribution is a better decision algorithm than the decision tree, for illustrating this example, the traditional 20% threshold, resolution cut-off of 5-10 Å, and a rotational tolerance of 20° were chosen. There were 213 false positives, where pdb-tNCS(20°) was false, and the highest non-origin Patterson peak from the observed data was above the 20% threshold. For 158 cases (75%) the highest non-origin Patterson peak from the calculated data was below the 20% threshold. In these cases, modulation of the data could not be explained by the calculated intensities.

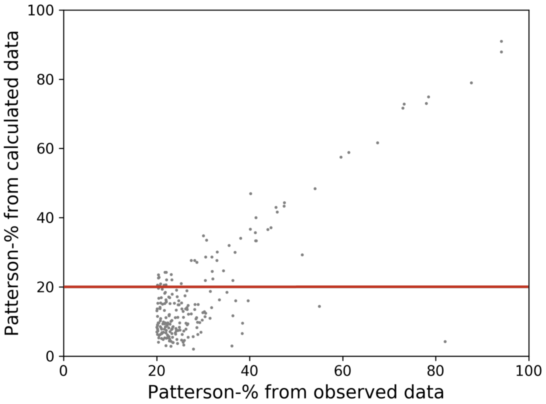


Figure 18. Scatter plot showing the distribution of highest non-origin Patterson peaks in the calculated and observed data, as a percentage of the origin peak, for cases with pdb-tNCS(20°) false and observed Patterson-20% true. Above the red line cases with calculated Patterson-20% true, and below the line cases with calculated Patterson-20% false.

It is possible that these structures show a degree of lattice translocation disorder, with stacking heterogeneity between mosaic blocks (Dauter *et al.*, 2005; Rye *et al.*, 2007). Interestingly, the distribution of space groups in these structures differed significantly from the distribution across all deposited structures, with space group *P*2$_1$ present at three times the number expected (see table 6). The 2$_1$ screw axis has been implicated as an important component of polytropism for crystals (Aquilano *et al.*, 2002). Furthermore, the low symmetry space groups such as *P*2$_1$, *C*2 and *P*1 might be affected by poor scaling, given the lack of symmetry equivalents. For instance, if the beam is unstable, showing a periodic flux oscillation, frames of strong and weak reflections will be measured.

Table 6. Space group propensity for 158 cases where there was no tNCS in coordinates and a high peak in the Patterson from the observed data was absent when using calculated data. PDB average following (Wukovitz & Yeates, 1995).

Space group	Number	Percent	PDB average
*P*2$_1$	60	38%	11.1%
*C*2	30	19%	6.1%
*P*1	23	15%	2.6%
*P*2$_1$2$_1$2$_1$	8	5%	36.1%
*P*2$_1$2$_1$2	5	3%	3.7%
*C*222$_1$	5	3%	3.7%
*R*32	5	3%	—
*R*3	5	3%	—
Other space groups with only 4 or less structures in each one	17	11%	—

One of the cases that had significant Patterson peaks despite not having tNCS was the proteolytic domain of *Archaeoglobus fulgidus* Lon protease (1z0v (Dauter *et al.*, 2005), a structure known to be an allotwin (Lebedev, 2009). Individual crystals belonged to the space group $P2_1$ and $P2_12_12_1$, with the transition layers in plane space group $P2_12_1(2)$ giving a sequence of stacking vectors.

Another case was Lipase B from *Candida antarctica*, also known to be an OD-twin. In this case, the two space groups involved were *C*2 and $P2_12_12_1$, with the transition layers again in plane space group $P2_12_1(2)$. The deposited data for 1lbs (Uppenberg *et al.*, 1995) were processed in the larger, orthorhombic lattice, which resulted in apparent data completeness of 27.5% although the completeness in the actual *C*2 space group was 82.4%. In terms of the study, this structure was included in the small database of structures with less than 80% complete data, however, had it been included in the main database, it would have been the most extreme false positive outlier.

In another case, Ftsk motor domain from *Escherichia coli* (2ius) (Massey *et al.*, 2006) the indexing and space group determination for the crystal was problematic (Jan Löwe, *pers. comm.*). We thus hypothesize that these outliers are as a result of a structure with a lattice-translocation defect, rather than tNCS. In the context of automated structure determination, it is therefore important to consider the absence of tNCS even in the context of large Patterson peaks being present.

1.9. High order tNCS

In the course of the study, a few cases in which sub-groups of components were related by different tNCS vectors were noted. These cases tended toward pseudo-centring in multiple directions. For example, a small ligand-bound complex of von Hippel-Lindau (VHL) E3 Ubiquitin Ligase and the Hypoxia Inducible Factor (HIF) Alpha Subunit (PDB id 4w9d, $P4_122$) (Galdeano *et al.*, 2014), showed a pseudo-centring in the *a* (0.5, 0.04, 0.0) and *a-b* diagonal (0.54, 0.5, 0.0)) directions, and similarly, the crystal structure of SOAR domain (PDB id 3teq, $P4_12_12$) (Yang *et al.*, 2012) showed pseudo-centring in the *a* (0.49, 0.01, 0.0) and *a-b* diagonal (0.49, 0.51, 0.0) directions. If there are sub-groups of components related by different tNCS vectors or if only some components of the asymmetric unit are related by a tNCS vector, then the modulations of the expected intensities due to the tNCS will be much less significant, and structure solution may be achieved without any tNCS correction being applied, as indeed was the case in these examples. However, if structure solution fails, detecting and correcting the dominant order of tNCS within the asymmetric unit may be enough.

1.10. tNCS detection

An algorithm for characterizing and ranking tNCS hypotheses by analysis of the intensities prior to structure solution was developed. Correct identification of tNCS can have a profound impact on the ability to place components in the asymmetric unit, whether they be components by MR or heavy atoms by experimental phasing. In the context of a pipeline for structure solution with *Phaser*, the fastest route to structure solution on average should be by exploring the tNCS hypotheses in order of ranking by our criteria.

Our algorithm for tNCS detection not only determines the tNCS vector and the tNCS order but also involves tests that aim to exclude pathological cases. First, a Patterson function is calculated from the data, by default using **5-10 Å** resolution data. Peaks are picked in the Patterson function and filtered by two criteria; the peak height must be over a given percentage of the origin peak height and the peak distance must be above more than a given distance from the origin. As guided by this study, the default distance threshold is **15 Å** and the default Patterson function threshold is **16.8%.** Cases where at least one of the unit cell dimensions is less than the origin distance threshold, are considered pathological (most likely peptides) and are excluded from further analysis. If there are no surviving non-origin distinct peaks over the Patterson-% threshold, the algorithm terminates with status “tNCS not indicated”, otherwise the algorithm proceeds to the analysis of the tNCS order. The simplest interpretation of surviving peaks is that each (if there are more than one) presents an independent tNCS_2 vector and with Patterson-% indicating the strength of the associated modulation, which provides a ranking for the hypotheses.

Then, further analysis was performed to determine if the Patterson peaks are due to a higher-order tNCS commensurate modulation and, if so, the order of that commensurate modulation. Noise in the Patterson function is removed by setting all values below 8% of the Patterson origin peak to zero, and the noise-reduced Patterson function is transformed to reciprocal space, where commensurate modulation is detected as strong low order Fourier terms. The hypothesis for a given commensurate modulation will predict a set of equal-height peaks in the Patterson function. In practice, because the components are not related by a perfect translation (as previously discussed), these predicted peaks will have different heights, and some may be below the Patterson-*t*% threshold of the analysis.

Following the studies on eps-tNCS and the high correlation with the height of the highest Patterson peak, we rank commensurate modulations that predict the highest-ranked peak higher than those that do not. The result of the algorithm is a ranked list of tNCS modulations representing high-order commensurate tNCS_n and commensurate and non-commensurate tNCS_2. Following the observation that high Patterson peaks in the data may be due to order-disorder effects, the case of no tNCS is also always included in the list of hypotheses. Note that the ranking is not necessary for structure solution. In the context of an automated pipeline, as long as the correct hypothesis is in the list, it will be explored. The ranking only affects the order in which the hypotheses are explored, and hence the efficiency of structure solution.

An unoptimized part of the algorithm attempts to prevent the misclassification of coiled-coils and amyloid peptide repeats as having tNCS. As previously discussed, pseudo-symmetry in secondary structure elements generates large peaks in the Patterson function close to the origin. Although coiled-coils were excluded from our curated database, by looking at a small number of cases it was observed that the 15 Å minimum vector exclusion around the origin was not sufficient to exclude peaks generated by the coiled-coil pseudo-symmetry (Kondo *et al*., 2008). Taking a heuristic approach, we exclude peaks from the tNCS analysis if they cluster together within the short distance separation characteristic of coiled-coils. Future work will perform a systematic study of coiled-coils and amyloid peptide repeats to optimize the tNCS detection algorithm in these cases.

Finally, in this work, to model either the tNCS-rotation or the tNCS-rmsd from the Patterson function was not attempted. This decision is in line with having seen the limited sensitivity of the Patterson as compared with the epsilon refinement, since some information about these parameters is contained in the Patterson peak height relative to the origin peak, with lower peak heights indicating more deviation from perfect translation, there may also be information about rotational deviations in the 3-dimensional Patterson peak shape. However, in practice, refinement of these parameters to correct the modulation starting from several different tNCS-rotation perturbations works extremely well, and in most cases, all perturbations converge on refinement to the same final tNCS-rotation and tNCS-rmsd.

To conclude, an analysis of a curated database of protein structures from the PDB to investigate how tNCS manifests in the Patterson was performed. These studies informed our algorithm for detection of tNCS, which includes a method for detecting the number of vectors involved in any commensurate modulation (the tNCS order). Our algorithm generates a ranked list of possible tNCS associations in the asymmetric unit, for exploration during structure solution.

2. COILED-COILS

2.1. Introduction

Coiled-coils are protein structure domains that consist of two or more α-helices wrapped around each other to form a supercoil (Mason & Arndt, 2004). The sequences underlying this fold contain characteristic repeats of seven residues leading to left-handed coiling or 11 residues in the case of right-handed coiling (Lupas & Gruber, 2005). The helices are packed together in a specific knobs-into-holes manner (Crick, 1953), with interhelical interactions playing a dominant role in folding (Burkhard *et al.*, 2001). The architecture of a coiled-coil domain determines its oligomerization state, so slight variations in the knobs-into-holes packing and different degrees of supercoiling cause these structures to adopt a wide range of geometries and topologies (Burkhard *et al.*, 2001; Lupas & Gruber, 2005; Rackham *et al.*, 2010).

These ubiquitous protein folding motifs represent a substantial portion of structural studies due to their versatility, as they are found in a variety of proteins involved in a wide range of biological functions (Lupas, 1996) of which notable examples are transcription, ATP synthesis, catalytic activity, molecular spacing, intracellular transport, transmembrane signalling, membrane fusion and re-modelling, proteostasis, the formation of the extracellular matrix and several cytoskeletal and nuclear structures of the eukaryotic cell or mediation of protein-protein interaction (Baxevanis & Vinson, 1993; Kuhn *et al.*, 2014; Mier *et al.*, 2016; West *et al.*, 2019). These diverse functions agree with the fact that approximately 10% of the eukaryotic proteins are coiled-coils (Liu & Rost, 2001).

Furthermore, they have a widespread biomedical significance, since different oligomerization states have been associated with disease-causing mutations (Kalman *et al.*, 2020). Also, thanks to their self-assembling nature, they have been exploited in the biomaterials field to construct self-assembled materials such as peptide-based nanotubes that are highly attractive for many biomedical applications such as drug delivery, scaffolds for tissue engineering, and many others (Burgess *et al.*, 2015). Their ubiquity and versatility underlines the importance of developing computational methods to solve these structures.

In general, mainly helical structures constitute favourable cases for phasing with *ARCIMBOLDO_LITE*, where polyalanine helices provide ideal search fragments as they are constant, rigid, and nearly ubiquitous. Despite the apparent simplicity of the coiled-coil architecture, the modulation dominating the data causes problems in the case of experimental phasing and MR (Blocquel *et al.*, 2014; Dauter, 2015; Franke *et al.*, 2014; Franke *et al.*, 2011; Thomas *et al.*, 2020).

Several factors act in combination to make coiled-coils notoriously difficult for phasing. They are highly sensitive to experimental conditions of crystallization; long proteins formed with these domains tend to aggregate into fibers instead of rendering single crystals (Lupas & Gruber, 2005), compromising the obtaining of high-resolution structures. Besides this, their filamentous nature can lead to crystalline lattices with lateral association of the molecules, which also entails highly anisotropic dimensions of their asymmetric units and anisotropic diffraction. Moreover, deviations from the canonical repeat cause helical and superhelical fold irregularities (Lupas & Gruber, 2005) making the solution of the structures by MR difficult. In addition, SAD and/or MAD phasing by seleno-methionine incorporation is significantly hampered by the repetitive nature of their sequences that are often deficient in methionine (Franke *et al.*, 2011).

Phasing of coiled-coil crystal structures with fragments has been implemented in the *AMPLE* (Sanchez Rodriguez *et al.*, 2020; Thomas *et al.*, 2015; Thomas *et al.*, 2020) and *CCsolve* (Rämisch *et al.*, 2015) pipelines, which combine de novo structure prediction (Das *et al.*, 2009), MR search and autotracing (Sheldrick, 2010) or automated model building (Terwilliger *et al.*, 2008). In addition, recent notable improvement in methods for phasing coiled-coils structures have been added. Such as the *ab initio* modelling of elongated helices and oligomeric coiled-coils from two to four helices (Thomas *et al.*, 2020), and the use of libraries of helical ensembles (Sanchez Rodriguez *et al.*, 2020).

In the present work, fragment phasing on a pool of 150 coiled-coils have been explored, and the results have been used to identify hurdles, develop ways to overcome them and equip *ARCIMBOLDO_LITE* (Millán *et al.*, 2015) with a specialized protocol with optimized strategies and parameters values for coiled-coil structures.

2.2. Overall performance of *ARCIMBOLDO_LITE*

The name of the phasing method *ARCIMBOLDO* came from the analogy with the paintings from Giuseppe Arcimboldo, who used to compose portraits out of fruits, vegetables and flowers, as the program assembles secondary structure elements to compose the structure of a protein. Since the fragments represent a low fraction of the total scattering mass, most attempts remain a “still-life” but if these fragments are properly placed, density modification and main-chain tracing reveals the true portrait of the protein.

There are three implementations of the *ARCIMBOLDO* method depending on the most suitable search model for the problem structure. *ARCIMBOLDO_LITE* (Millán *et al.*, 2015; Sammito *et al.*, 2015) employs small and very accurate model fragments such as polyalanine α-helices, *ARCIMBOLDO_BORGES* (Sammito *et al.*, 2013) libraries of local folds and *ARCIMBOLDO_SHREDDER* (Sammito *et al.*, 2014) libraries of fragments from an homologous template.

The *ARCIMBOLDO* (Rodríguez *et al.*, 2012; Rodríguez *et al.*, 2009) workflow schematized in figure 19 starts when *Phaser* (McCoy *et al.*, 2007; Read & McCoy, 2016) sequentially locates the search fragments that will depend on the implementation of *ARCIMBOLDO* used. In the first stage performs a rotation search, then the rotation peaks are clustered and translation search is performed for each of these rotation groups. Finally, if there are any clashes between atoms because the solution is not compatible with the crystallographic symmetry restrictions imposed by the space group, the solution will not pass the packing filter. After all fragment-location operations have been performed, *SHELXE* (Sheldrick, 2010) performs iterative density modification and main-chain autotracing starting from all selected substructures, to expand them into a nearly complete structure.

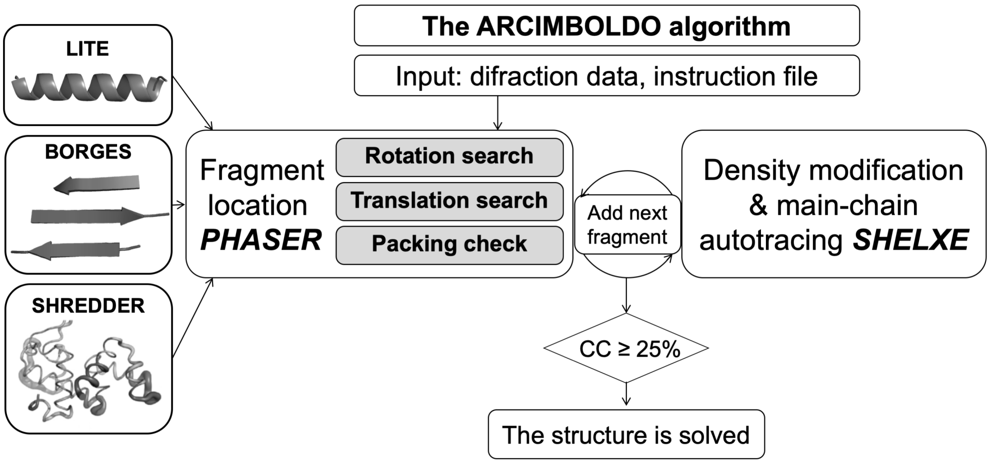


Figure 19. The *ARCIMBOLDO* workflow.

This study focused on *ARCIMBOLDO_LITE* with ideal polyalanine α-helices as search fragments. It was employed to attempt phasing a set of 150 coiled-coil test structures with resolutions ranging from 0.9 to 3 Å, sizes from 15 to 635 residues, and including 38 different space groups. 94 structures in this test set were adopted from an earlier study on coiled-coil phasing with *AMPLE* (Thomas *et al.*, 2015). The lower resolution range was supplemented with a further 28 structures at 2.0–2.5 Å resolution and 28 at 2.5–3.0 Å resolution as the negative impact of poor resolution had been previously identified in their study. Those cases were selected by resolution, aiming at a distribution in size and space group. A few low-resolution cases discussed in the *CCsolve* (Rämisch *et al.*, 2015) paper were also adopted, if not already in the previous test set.

In general cases, correct solutions can be distinguished from wrong ones by the correlation coefficient (CC) of the partial structure against the experimental data, a CC of 25% or higher indicates successful *SHELXE* expansion results (Thorn & Sheldrick, 2013).

This indication is correlated with small values of weighted mean phase error (wMPE), that confirms the indication provided by the CC. This value is calculated against the refined models available from the PDB (Sheldrick, 2002b), and their values span from 0 to 90º, if smaller than 80º indicate not random solutions and correlation between the phases under study with respect to the true phases. The CC is reliable at atomic resolution with individual atoms (Sheldrick *et al.*, 2012), at lower resolution a polypeptide trace that takes into account the context is needed. But even with a polypeptide trace, in the case of coiled-coils, a random partial structure can show high CC values when the modulation of the data is explained. For that reason, to ensure the reliability of the results and for the purpose of this study, a structure is considered solved when the wMPE versus the reference deposited with the PDB was below 65º.

An initial baseline to identify cases that could be solved straightforwardly with *ARCIMBOLDO_LITE* was set running the program with general default parameters on the pool of 150 structures. The fragment search was configured to find four polyalanine helices of 18 residues and using the standard resolution-dependent *SHELXE* parameterization (Sammito *et al.*, 2015). This straightforward approach was successful in 78 out of the 150 cases (52%) and led to the identification of the most interesting cases. It is worth mentioning that the high success rate achieved with minimum intervention underscores the generality of the *ARCIMBOLDO_LITE* approach and robustness of its default parameterization in solving the most disparate cases of coiled-coils.

The following sections describe the particular problems that prevented some of the remaining 72 structures from being immediately solved. Solutions to these problems are then proposed and tested.

2.2.1. Number of fragments to search and helix length

In general, choice of search fragments is based on the secondary-structure prediction for the contents of the asymmetric unit and the signal that can be expected from a fragment of a given size for the particular data (McCoy *et al.*, 2017; Oeffner *et al.*, 2018). Furthermore, some trial and error may be necessary, as seen in a case where the effect of helix length was systematically tested (Schoch *et al.*, 2015).

A first search configured to find four helices of 18 residues demonstrated that this could be a good starting strategy, but the structures that were not solved in the first blind run were run again selecting a more appropriate helix length and the number of helices to be placed. The helix sizes spanned from six to 50 residues, and the number of fragments placed was from one to 12.

In the case of high-resolution structures, the best results were obtained employing shorter helices as the model is very accurate and does not account for the deviations in the helices. In the case of structures with long and curved helices, trying to search for the entire helix with the full length, straight model helix built into *ARCIMBOLDO_LITE* may do not lead to any solution, while dividing the search in two or three shorter helices that can adopt different orientations along the true helix can lead to a successful expansion in *SHELXE*.

Finally, it is worth mentioning that the minimum percentage of the total structure that is required to phase depends on the structure. As demonstrated in previous studies (Millán *et al.*, 2015), a bare 15% of the main-chain atoms is enough to solve a structure at 2 Å, but as the high helical content of coiled-coils is known, there is no reason to artificially limit the use of model helices.

2.2.2. rmsd and vrms

MR requires the location of a model of a known structure close enough to the target structure, which can subsequently be used to derive starting phases. The degree of similarity between the target and model is quantified in terms of rmsd. This value can be inferred by the sequence identity (Chothia & Lesk, 1986), as the target structure is yet undetermined. It is worth mentioning that a recent study improved prior estimates of the coordinate error (Hatti *et al.*, 2020; Oeffner *et al.*, 2013). In most cases of successful MR, the protein of interest shares at least 35% sequence identity with its structural homologous, corresponding to an rmsd around 1.5 Å (Abergel, 2013). This rms coordinate error expected for the model along with the scattering power in the asymmetric unit that this model contributes is used to compute the Sigma(A) curve, which estimates how the accuracy of the model falls off as a function of resolution and is fundamental to how *Phaser* uses the models (Read, 1986).

When phasing with ideal helices, the rmsd cannot be inferred by sequence identity. In addition, the ideal 14-residue polyalanine helix typically used in *ARCIMBOLDO_LITE* as search fragment usually represents a small percentage of the scattering. For this reason, only parts of the target structure presenting a low value of rmsd are susceptible of being located, so our choice is setting a default value of 0.2 Å for the fragment search in *Phaser*. Furthermore, at low rmsd values, the sampling of conformational space is finer, so that one of our models is expected to have an rmsd of 0.2, which increases the probability of finding a more accurate solution.

Longer helices were used in most of the test cases, and the accumulated curvature in coiled-coils was expected to lead to higher deviations, but in practice, all structures but one were solved by setting the rmsd to 0.2 Å. PDB id 3thf in space group $P2_12_12$ with 349 independent residues at 2.7 Å was only solved by increasing the rmsd to 0.5 Å.

In *Phaser*'s rigid-group refinement stage the input rmsd parameter can be refined to maximize the LLG (Oeffner *et al.*, 2013) through the root-mean-square derived from the likelihood variance (vrms) and consequently increase the accuracy in the determination, improve the scoring of alternative solutions and thus the probability of MR success. In the context of *ARCIMBOLDO,* the LLG serves two purposes, to evaluate if the fragments are properly placed or not, and to prioritize solutions that will be expanded.

In solved structures, the vrms refined to values around 0.1 Å, ranging from 0.05 to 0.53 Å. This roughly corresponds to the default rmsd parameterization, and therefore refining the rmsd as a parameter does not have a large effect. The only exception was noted for PDB id 3v86 at 2.91 Å, structure which is merohedrally twinned, and where the correct substructure was only discriminated by refining the rmsd. All other cases were insensitive to this parameter. As it has not been observed to have adverse effects in any case, this calculation is activated by default in the *coiled_coil* mode.

2.2.3. Translational non-crystallographic symmetry

Translational non-crystallographic symmetry (tNCS) arises when the asymmetric unit contains components that are oriented in nearly the same way and can be superimposed by a translation that does not correspond to any symmetry operation in the space group. See the introduction section 3.4 for more details.

The presence of translational non-crystallographic symmetry (tNCS) is deduced by the currently distributed *Phaser* (version 2.8) from the presence of peaks separated from the origin by more than 15 Å and above 20% of the origin peak in the Patterson function calculated using data from 5 to 10 Å resolution. If tNCS is identified, *Phaser* will correct the effect of the modulation in the input data and search for pairs of molecules (groups in a more general case) related by the tNCS vector (Sliwiak *et al.*, 2014). Parameters describing the translation, small rotation, and conformational differences between copies are determined and used to compute correction factors to the target function (Read *et al.*, 2013). By default, *ARCIMBOLDO_LITE* exploits this feature in *Phaser*, simultaneously placing tNCS-related copies associated with a given rotation.

In the case of coiled-coils, the internal symmetry of a single helix along with the accidental overlap of vectors derived from the systematic alignment of helices along predominant directions gives rise to strong peaks in the Patterson function (Urzhumtsev *et* *al.*, 2016). In these cases, the Patterson function looks like rosary beads, with peaks near the origin indicating the direction of the helices, and other peaks that are related to the molecular position and the rotation angle between the helices (Urzhumtsev et al., 2016). This has been frequently observed in RNA and DNA, which are molecules with pseudo-helical symmetry that are usually packed more or less parallel to each other (Kondo et al., 2008). Additionally, in coiled-coils it has been observed that preferential orientation along a common axis produces linear arrays of Patterson peaks separated by 5.1 Å, which arise from the pitch along the coiled-coil axis of 5.1 Å per α-helical turn, that are orientated in the same direction and indicate the direction of the coiled-coil axis (Thomas et al., 2020).

Therefore, the correction for tNCS should be disabled when high peaks in the Patterson are due to internal repeats in the coiled-coiled rather than true tNCS. The *coiled_coil* keyword entails its deactivation.

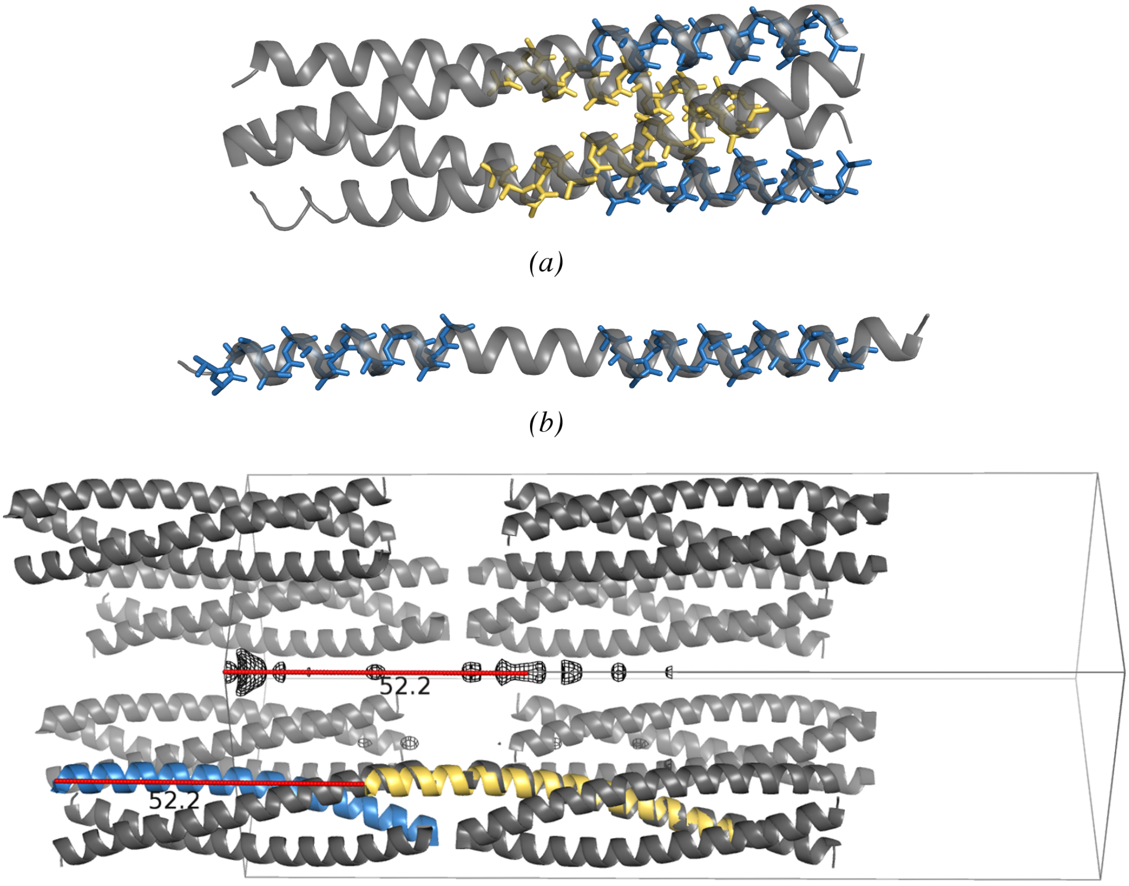
Within the first pool of 94 structures, 19 cases show peaks in the Patterson function, which would trigger tNCS pairwise location. Of these, PDB entries 1byz, 1g1j, 1kyc, 1nkd, 1p9i, 1x8y, 1yod, 2b22, 2bez, 2ic6, 2wpq, 3bas, 3hfe, 3k9a, 3m91, 3p7k, 3v86 and 3vgy have been solved, while 3mqc remains unsolved. Within the second, lower resolution pool of 56 structures, tNCS was identified from the Patterson function in ten solved cases: PDB entries 2ahp, 3efg, 3r3k, 5c9n, 1unx, 2wz7, 1w5h, 2o1j, 3v2r and 3nwh. A further three cases, PDB entries 3iv1, 3tul, and 4pna, remain unsolved. These structures are summarized in appendix table A1.

Notwithstanding, structures containing several helices in the asymmetric unit may display true tNCS, as illustrated in figure 20a. This was the case for the PDB entries 1g1j, 2o1j and 3nwh, where phasing was only successful accounting for this feature and placing tNCS-related pairs.

In the cases with apparent tNCS, as a fragment of a helix can be translated and superimposed onto another part of the helix, aiming to place tNCS-related copies might cause fragment overlap and structure solution to fail. All of these structures were tried with pairwise placement turned off (keyword tNCS: False); that is, placing single helices sequentially as well as placing pairs of tNCS-related helices. In eight cases, either setting led to a correct solution as no overlap between fragments occurs. Figure 20b shows the case of PDB id 3efg, with a single copy in the asymmetric unit, which could not display real tNCS, where the fragments related by tNCS did not overlap. In 17 cases, a solution was only found by placing single-fragment copies sequentially, whereas pairs of fragments placed as related by the translation vector derived from the Patterson map were either misplaced despite their high scores or discarded at the packing check because of partial overlap with symmetry equivalents. For example, PDB id 3p7k in space group $P6_322$ at 2.3

Å resolution, whose packing is shown in figure 20c, contains a single, curved helix of 45 amino acids in the asymmetric unit. Displacing it by 52.2 Å in the direction of the c axis partially superimposes it on two symmetry equivalents, one of them in the reversed direction. The corresponding Patterson peak displayed in the figure is the maximum identified by *Phaser*, but generating pairs of helices related by such a translation would, in this case, prevent the location of a correct solution. Thus, to solve this structure, the pairwise placement feature needs to be turned off.

As differentiating genuine tNCS from Patterson artefacts is difficult in the presence of fragment-derived modulation, the default behaviour in *ARCIMBOLDO_LITE* for coiled-coils will be to avoid the tNCS-related search, but if no solution is achieved this alternative should be tried.



(a)

(b)

(c)

Figure 20. a) PDB id 2o1j with true tNCS, the helices coloured in blue and yellow are pairs of molecules related by tNCS. b) The structure 3efg could be solved placing tNCS-related copies because there was no fragment overlap. c) Apparent translational non-crystallographic symmetry in the case of PDB id 3p7k. The structure is shown as a blue cartoon, with symmetry equivalents as a grey cartoon and the Patterson map contoured at 2 σ as black mesh. The yellow helix corresponds to PDB id 3p7k translated 52.2 Å by the vector corresponding to the Patterson function peak. It coincides with different portions of symmetry-related helices.

2.2.4. Packing filter at translation search

Partially overlapping solutions are usually discarded after the translation search, during the packing filter: if there are any clashes between atoms the ensuing structure renders a physically impossible model or it is not compatible with the crystallographic symmetry restrictions imposed by the space group.

The selection of translation function solutions is done in relationship with the maximum value the function adopts. The number of peaks that will pass to the packing check is defined by a cut-off of 75% between the top solution and the mean value, where the value of the top peak is defined as 100% and the value of the mean is defined as 0%, shown in figure 21a. If the maximum corresponds to a very high value for a physically impossible solution, this and all other peaks selected may be discarded in the packing check. In space groups where proper rotational symmetry operations (i.e., not combined with translations) are present, a recurrent problem is that helices placed on pure rotation axes may be characterized by extremely high LLG scores, while correct solutions may be well below 75% of these values. In all space groups, a second helix placed on top of a previous helix may also lead to disproportionately high LLG scores. In this case, no solution with feasible packing will be output in the list of translation-function solutions, and the process halts as the packing filter discard everything.

To address this recurring problem in helical fragment searches a new packing filter within the analysis of the translation function was proposed (Caballero *et al.*, 2018) and implemented by our collaborators in *Phaser*. This ensures that the top solution used as a reference for selection will not be rejected later in the packing check, shown in figure 21b. *ARCIMBOLDO_LITE* uses a very stringent default for either check, allowing no overlap at all.

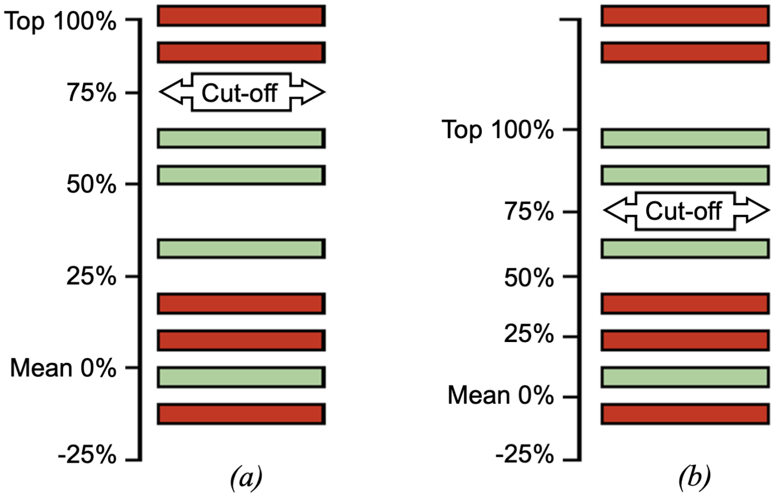


Figure 21. a) List of translation solutions without packing filter, selected solutions are above the 75% cut-off. b) List of translation solutions with packing filter, the top solution used as a reference will pass later the packing symmetry check.

The cases of PDB entries 2v71 in space group *C*2, 1d7m in *C*222$_1$, 4bl6 in *P*6$_1$, 3miw in *P*4$_2$, 5jxc in *P*2$_1$, 3r47 in *P*4$_2$, 4bry in *I*4$_1$22 and 3thf in *P*2$_1$2$_1$2 could only be solved when *ARCIMBOLDO_LITE* was run activating *Phaser*'s packing filter at translation. The only drawback is an increase in running time, but for coiled-coils activating this option is the default, as this issue frequently hinders solution, especially at resolutions worse than 2 Å.

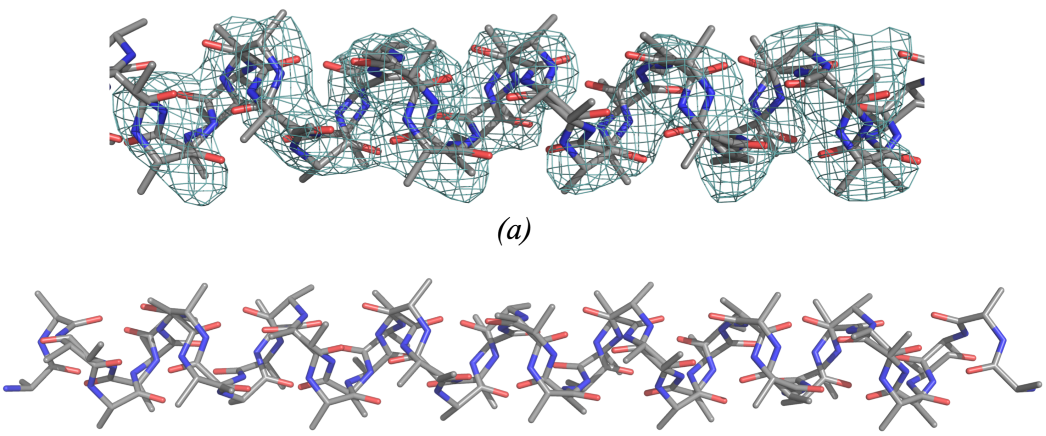
2.3. Performance of *ARCIMBOLDO_LITE* at resolutions between 2 and 3 Å

From the outset, it became evident that lower resolution posed particular difficulties. This prompted us to extend the original test set with 56 structures at resolution worse than 2.0 Å to a total of 106. Among them, 43 corresponded to resolutions between 2.5 and 3.0 Å (15 structures in the first set and 28 in the second). Five of the eight structures that remain unsolved correspond to the lower resolution span.

2.3.1. Reversed helices

At resolutions worse than 2 Å it was observed that placement of the first helices occasionally took place in the correct position but in reversed direction since properties of the helix account for main low-resolution diffraction features in either direction. This is illustrated in figure 22a.

This problem can be solved by phasing with substructures with reversed helices but first was required to generate a template in reversed direction, shown in figure 22b. *ARCIMBOLDO_LITE* has a template of an ideal α-helix of 70 residues which is cut depending on the length of the search fragment. To generate its reverse model first this ideal helix was manually reverted with *Coot*, after that, to achieve the best position, the helix was shifted along its axis 0.5 Å obtaining several models. Then, as this phenomenon was observed for the first time in the case of PDB entry 3miw, these models were submitted to rigid body refinement (McCoy *et al.*, 2017) against the experimental data. The one rendering better results, which was the one where the Cα and the carbonyl Carbon atom coincide, was selected for using it as a reversed template.



(a)

(b)

Figure 22. a) Helices in both directions with the electron density map contoured at 1.5 σ obtained from *ARCIMBOLDO_LITE* from the correct solution of PDB entry 3miw. b) The ideal helix from *ARCIMBOLDO_LITE* and its reversed model, this is an example of 30 residues that was cut from the template of 70 residues.

Some examples with substructures with reversed helices are PDB entries 3p7k at 2.3 Å resolution, 3h7z at 2.5 Å resolution, and 2nps at 2.5 Å resolution, where coexisting correct substructures led to a full solution, even though some of the substructures with reversed helices were sent to expansion as well. In the case of 3p7k, 37.8% of the substructures present at least one helix in reversed direction, for 3h7z this value is even higher 65.6%, and for 2nps only the 12.7% of the substructures render reversed helices. As mentioned before, the final solution was accomplished by a substructure with all the helices in the correct position and in no case by these partially incorrect substructures.

In the cases of PDB id 3onx at 2.9 Å resolution, PDB id 2jee at 2.8 Å resolution, PDB id 2fxm at 2.7 Å resolution and PDB id 3miw at 2.5 Å resolution all the substructures have reversed helices and this issue prevented solution of the structure. Such non-random but partially incorrect solutions are often not corrected by *SHELXE*'s density modification and autotracing, as the initial fragments cause phase bias in the map to be traced. Therefore, the incorrect helices are found and built again every cycle and the process is locked on these errors, despite showing deceptively promising figures of merit and trace extension.

Even though the presence of reversed helices in the substructure tends to persist throughout tracing, the problem can be solved by also phasing with substructures with reversed helices, after the placement of several fragments. After rigid-body refinement and rescoring, discrimination of the correct, more complete partial substructures improves, allowing solutions where some of the first fragments had been reversed to be rescued. If combinatorial perturbation of helix direction produces less than 1000 solutions, all of them will be explored. Otherwise, a sparse selection of them will be trialled to make the number of solutions tractable.

One example is the case of PDB id 3onx, where the best solution displays three helices correctly located and one helix in reversed direction, illustrated in figure 23a. After three cycles of iterating density modification and autotracing, *SHELXE* could not revert this helix and the final solution was characterized by a wMPE of 61.1º for 146 residues traced and a misleading high CC of 34.4%, figure 23b shows the lack of progress during the tracing. In a run probing the helix direction, a solution with all the helices in the correct direction was reached by reversing the first helix, which was the one incorrect. After density modification and autotracing, this solution renders a CC of 41.3% a wMPE of 51.3º and 181 residues traced out of 250.

In the case of PDB id 2jee the best solution, in terms of best CC, presented five of six helices in reversed direction and one helix mistranslated, the CC of this solution was 35.8%, with a wMPE of 83.3º and 111 residues traced. Reversing the helices during *SHELXE* tracing gave a correct solution with all the helices in the proper direction; during the course of the run, two of the six helices were reversed. After the expansion, a CC of 37.9% and a wMPE of 64.3º was reached and 158 residues were traced out of 312.

In the case of PDB id 2fxm, all helices placed by *ARCIMBOLDO_LITE* were in the correct direction, but the six were mistranslated, and the solution has a CC of 35.1% with a wMPE 89.5º and 144 residues traced. Performing the run with helix reversal leads to a solution with all the helices in the correct positions and directions and characterized with a CC of 53.9% with a wMPE of 44.3º and 190 residues traced out of 238.

Another example is provided by PDB id 3miw, which contained ten chains of 53 residues in the asymmetric unit and where the deposited data present severe anisotropy and twinning. After a search configured to find ten helices of 30 residues followed by two cycles of density modification and autotracing, a solution was identified that was characterized by 298 traced residues and a CC of up to 35.4%. Its wMPE was 62.9° and it contained 7.9% incorrect trace. Examination of the original solution revealed that of the ten placed helices, two were reversed. A fresh run with the version of *ARCIMBOLDO_LITE* that probes the helix direction rendered a substructure with all fragments correctly placed. This solution was reached by reversing three of the ten helices during the course of the run. The final solution was characterized by a wMPE of 59.7° for 301 residues traced out of 530. Remarkably, the errors in the trace decreased to 3.7% while the CC increased to 37.8%. Furthermore, solution of this structure was accomplished on a more powerful hardware, with three times more cores, that leads to the generation and extension of a larger number of partial solutions. Figure 23 displays the electron-density map for the partially incorrect (figure 23c) and the correct (figure 23d) solutions.

As can be seen from the CC values quoted above, the discrimination between correct and partially incorrect solutions can be narrow; therefore, the *coiled_coil* mode triggers systematic probing of both helix directions.

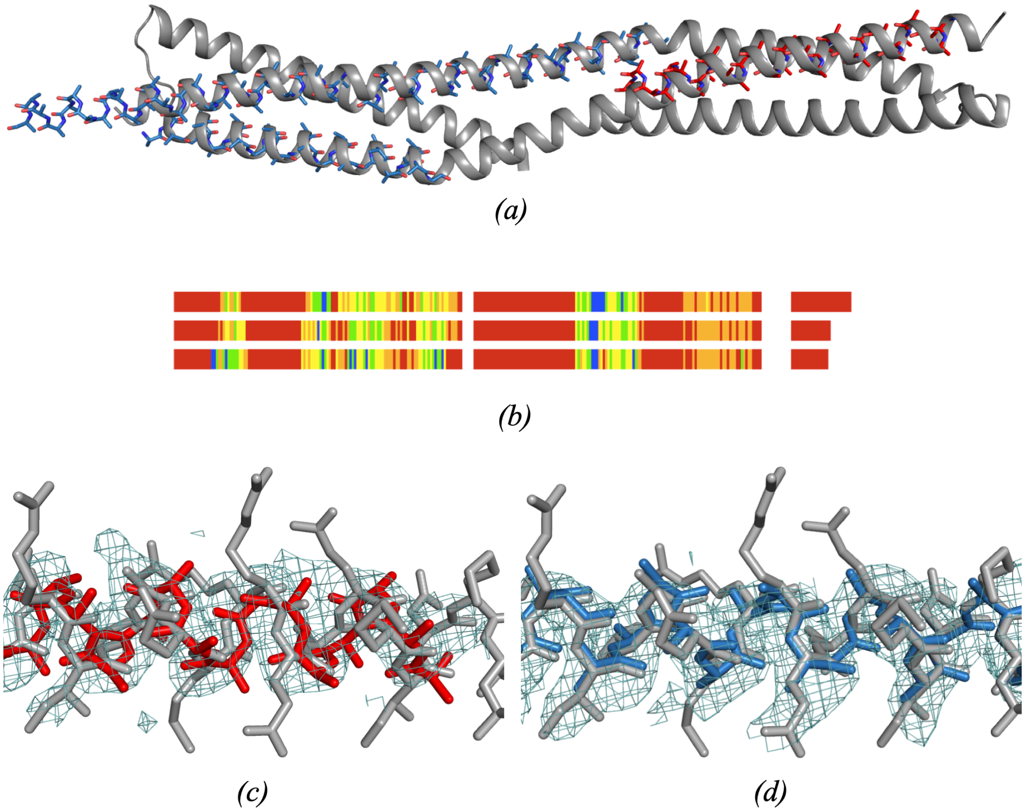


Figure 23. a) PDB id 3onx at 2.9 Å resolution with three helices correctly placed in blue and one helix placed reversed in red; fragments placed, shown as sticks, are superimposed on the origin-shifted PDB structure shown as grey cartoon. b) Lack of progress in three cycles *SHELXE* autotracing for PDB id 3onx. The first two blocks represent the length of the polypeptide chains traced by *SHELXE*, with the rmsd of the traces colour-coded from blue (<0.3 Å rmsd), green (<0.6 Å rmsd) and yellow (<1.0 Å rmsd) to red, where no trace can be matched within 2.0 Å rmsd. The third block represents the length of traced residues that cannot be assigned to any part of the correct structure. The consistent orange-coloured sections, indicating up to 2.0 Å rmsd, correspond to persistent reversed traces. c) Electron density map contoured at 1 σ after density modification and autotracing of an inverted helix in the solution for PDB id 3miw with errors. d) Electron density for the same region in the correct structure.

2.3.2. *SHELXE* autotracing with helical restraints

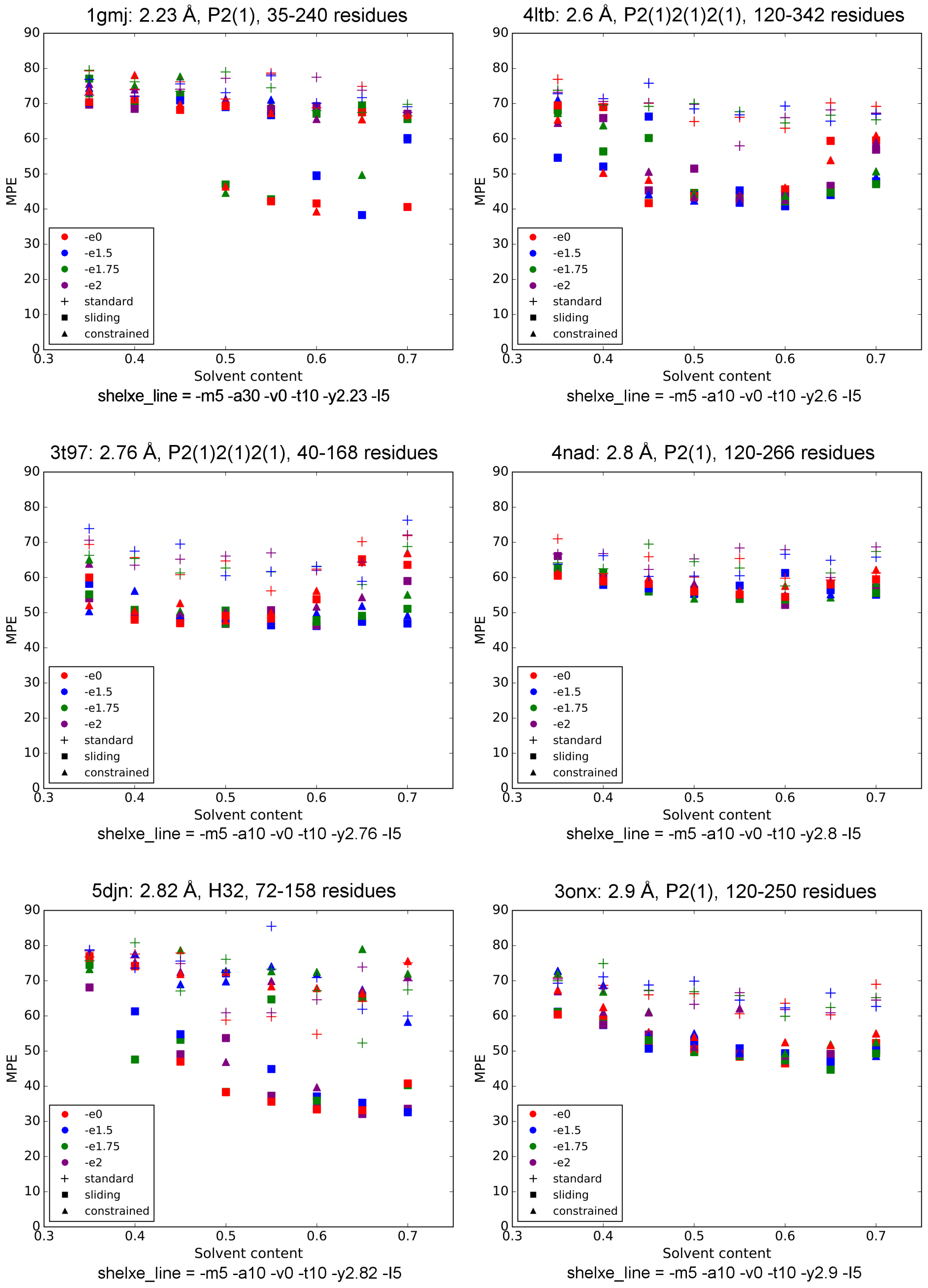
Whereas for coiled-coils with diffraction data to resolutions of 2.0 Å or better are generally solved using the standard resolution-dependent *SHELXE* parameterization (Sammito *et al.*, 2015), as the resolution becomes more limited the coverage of the traced model generated by *SHELXE* tracing decreases. Electron density in areas where the helices are bent degrades, leading to extended rather than helical polypeptide traces. As automatic map interpretation stalls, the discrimination of solutions becomes more uncertain. At resolutions worse than 2.5 Å this often leads to incorrect traces that are nevertheless characterized by a CC above 30%. High numbers of false positives are why *ARCIMBOLDO_LITE* generally fails to find a solution in coiled-coils structures if the experimental data does not reach better than 2.0 Å resolution.

A helically constrained main-chain tracing has been incorporated into *SHELXE* (Usón & Sheldrick, 2018). This choice is automatically triggered within the *coiled_coil* mode and leads to all autotracing cycles apart from the last being seeded from longer helices and extension of the main chain with helical restraints for Ramachandran angles or helical sliding. The last cycle reverts to *SHELXE* defaults, allowing the tracing of missing non-helical areas such as loops. The model characterized by the best CC will be kept.

All test structures with resolutions between 2.0 and 3.0 Å were subjected to different parameterizations of *SHELXE* in its standard and constrained autotracing modes to derive default parameters for *ARCIMBOLDO_LITE* in its *coiled_coil* mode. Figure 24 displays the results of a range of parameterizations on six challenging cases with low-resolution and/or a small fraction of the complete structure to start the extension. These graphs show how helically constrained autotracing is decisive in extending the trace and in lowering the weighted mean phase error, allowing a solution to be reached in cases where the standard autotracing would not lead to a solution. While the constrained autotracing (-q8 to -q14) uses larger helical seeds of eight to 14 residues and constraints on the extension of each amino acid to Ramachandran angles in the helical region, the sliding autotracing (-Q) additionally extends the sliding helical fragments of the polypeptide chain and is used by default for coiled-coils.

Additionally, the lack of completeness can introduce systematic aberrations and errors that greatly affect the quality of the map. The “free lunch algorithm” implemented in *SHELXE* was used to extrapolate reflections at different the resolution limits. The algorithm basically invents the unmeasured data and to use density-modification techniques to extrapolate the phases of these data improving the interpretability of the electron-density maps (Usón *et al.*, 2007). Table 8 summarizes the default coiled-coil resolution-dependent *SHELXE* parameterization, including the resolution limit of the extrapolated reflections.

Leaving the *SHELXE* line unset in the input .bor file will activate *SHELXE* defaults in the *coiled_coil* mode that differ from the standard defaults. Finally, *ARCIMBOLDO* will stop by default once a solution characterized by a CC above 30% has been reached, but in *coiled_coil* mode, it will continue to complete the predetermined number of *SHELXE* expansion cycles.


1gmj: 2.23 Å, P2(1), 35-240 residues

4ltb: 2.6 Å, P2(1)2(1)2(1), 120-342 residues

3t97: 2.76 Å, P2(1)2(1)2(1), 40-168 residues

4nad: 2.8 Å, P2(1), 120-266 residues

5djn: 2.82 Å, H32, 72-158 residues

3onx: 2.9 Å, P2(1), 120-250 residues

Figure 24. Scatter plots summarizing the results of different parameterizations of three alternative autotracing algorithms in *SHELXE* on six different structures. The colour represents the resolution limit of extrapolated reflections (-e) and the shape represents the autotracing algorithms. In the shelxe_line, -m sets the number of density-modification cycles, -a the main-chain autotracing cycles, -v the density-sharpening factor, -t the time factor for peptide searches and -y the highest resolution for the starting phases from the model; -l leads to the use of extrapolated reflections in all density modification cycles.

Table 8. *SHELXE* resolution-dependent parameterization for the case in coiled-coils.

Resolution	***shelxe_line***
<= 1.0	-m200 -a8 -s0.25 -v0.5 -t10 -Q -l200 -y(resolution) -e1.0 -m200 -a1 -s0.2 -v0.5 -t10 -q -l200 -y(resolution) -e1.0
]1.0, 1.3]	-m100 -a8 -s0.35 -v0.5 -t10 -Q -l100 -y(resolution) -e1.0 -m100 -a1 -s0.3 -v0.25 -t10 -q -l100 -y(resolution) -e1.0
]1.3, 1.5]	-m50 -a8 -s0.45 -v0.1 -t10 -Q -l50 -y(resolution) -e(resolution-0.3) -m50 -a1 -s0.4 -v0.1 -t10 -q -l50 -y(resolution) -e(resolution-0.5)
]1.5, 2.0]	-m15 -a8 -s0.5 -v0 -t10 -Q -l15 -y(resolution) -e(resolution-0.3) -m15 -a1 -s0.45 -v0 -t10 -q -l15 -y(resolution) -e(resolution-0.5)
]2.0, 2.5]	-m10 -a8 -s0.6 -v0 -t10 -Q -l10 -y(resolution) -e(resolution-0.3) -m10 -a1 -s0.55 -v0 -t10 -q -l10 -y(resolution) -e(resolution-0.5)
]2.5, 3.0]	-m5 -a8 -s0.6 -v0 -t10 -Q -l5 -y(resolution) -e(resolution-0.3) -m5 -a1 -s0.55 -v0 -t10 -q -l5 -y(resolution) -e(resolution-0.5)

2.3.3. True solutions, non-random solutions and false solutions, and how to distinguish them

ARCIMBOLDO, along with other fragment-based phasing methods, uses the extension of the main-chain trace output by *SHELXE* and the CC characterizing it to identify correct solutions. Cases where the resolution extends to 2 Å or better usually afford a good correlation between the CC of the trace and the wMPE of the structure, and hence a clear-cut discrimination of correct solutions. In such cases, a CC value above 30% typically corresponds to a trace covering over two-thirds of the true structure and a map in which side chains can be recognized unequivocally. Exceptions have been observed for false, mistranslated solutions (i.e., solutions containing incorrectly positioned helices but in correct orientations). Side-chain assignment in coiled-coils tends to be obscured compared with the main chain. Partially correct solutions containing mistranslated or reversed helices may be characterized by high figures of merit more frequently than in other kinds of structures, with the exception of DNA (Urzhumtsev *et al.*, 2016). We were interested in investigating the discrimination of best-scoring incorrect solutions from true solutions within the pool of coiled-coil test structures in order to avoid misleading program users with an incorrectly identified outcome of the phasing process.

Figure 25 shows bars representing the CC and coverage of the traces for correct and best-scoring incorrect solutions for 18 difficult coiled-coil test cases, ordered by resolution. In this graph, correct solutions tend to exceed CC values of 40% and in all cases, the correct solution was characterized by a CC at least 4.5% above that of the incorrect solution. At resolutions of 2.5 Å or better, both the CC and the percentage of traced residues show a clear-cut difference between correct and incorrect solutions. The situation becomes more complicated as the resolution decreases, especially since the graph compares the correct solution with partially incorrect solutions in which one or more of the helices in the starting substructure were reversed. Such cases include PDB entries 3p7k at 2.3 Å resolution (one reversed fragment), 3h7z at 2.5 Å resolution (two reversed fragments), and 2nps at 2.5 Å resolution (three reversed fragments). Thus, incorrect solutions are not random. Although the trace coverage tends to be significantly higher for the correct solution, this is not true in the case of two of these structures (3p7k and 3h7z), in which the reversed helix is also extended. It is not possible to give an absolute number differentiating both situations, as CC values above 40% have been observed for incorrect solutions, such as PDB id 2o1j at 2.7 Å resolution. This structure displays true tNCS of order two and could only be solved by accounting for it in *Phaser* as well as placing fragments pairwise. Such pathologies tend to arise in coiled-coils, and, as seen in figure 23, even in manual building, error identification may not be trivial. Thus, unfortunately, partially correct solutions cannot be distinguished from the correct solution without the latter's higher CC value for comparison, that is without the observation of a bimodal distribution in the figures of merit rendered. Therefore, an additional step has been proposed and implemented to verify the final solution.

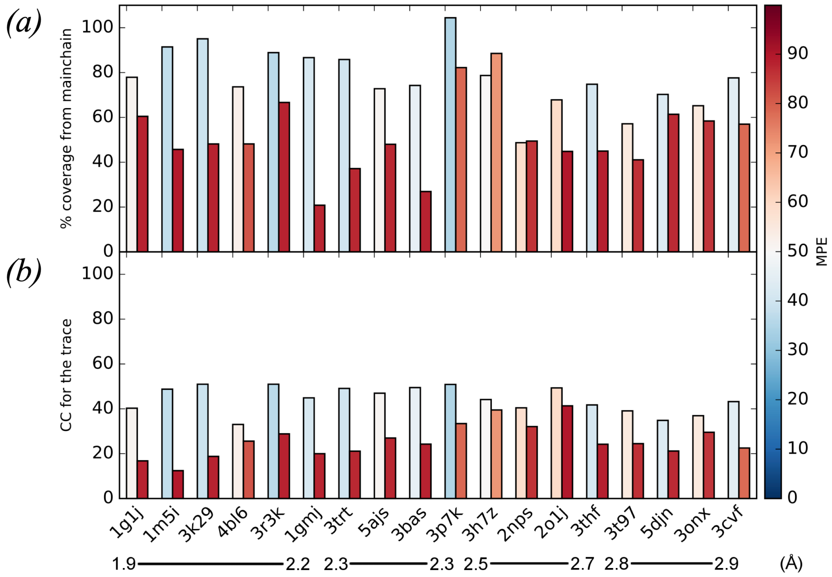


Figure 25. Bar plots representing correct (left) and best-scoring incorrect (right) solutions of the 18 most challenging test cases ordered from high to low-resolution. The wMPE versus the deposited structure is colour-coded from red (random) to blue (solved). a) Structure coverage in the trace. b) CC of the trace.

2.3.4. Final verification of the best-ranking solution

The general mode of ARCIMBOLDO has a resolution limit currently established at 2.5 Å, the rationale is not an absolute impossibility that the program would produce correct solutions at lower resolutions, but rather barring the risk of yielding false positives, which might go unidentified. Thus, the addition of a verification step that could rule out the false positives allows us to extend the resolution limit of our methods in the case of coiled-coils and should be extended to other lower resolution scenarios in *ab initio* phasing.

So, given the concern raised about partially correct solutions bearing good figures of merit, the incorporation in the *coiled_coil* mode in *ARCIMBOLDO_LITE* of an additional step that generates perturbations of the substructure leading to the best solution and compares their scores before and after extension is derived from the present work. In view of the obtained results, this step is compulsory and activated if the resolution is lower than 2 Å, allowing us to extend the resolution limit from 2.5 Å for the general mode of *ARCIMBOLDO_LITE* to 3.0 Å for the *coiled_coil* mode.

First, two types of perturbations were generated, a random solution and a group of substructures by reversing the direction of the helices. The random solution was generated by applying a fractional translation vector of (0.1, 0.1, 0.1), except for the space group *P*1, where half of the helices were translated by a vector of (0.5, 0.5, 0.5) and the other half remained in the same position. This is necessary because in *P*1 there is no symmetry within the unit cell, and translating the substructure would not modify the solution. In the case where this strategy did not give a clear random solution (where the mean phase difference (MPD) of the random solution with respect to the best solution was similar and less than 70º), correct from incorrect solutions cannot be discriminated. Regarding the substructures with reversed helices, taking the *Phaser* substructure that led to the best final CC, a systematic reversal of the helices in the substructure was performed, generating a maximum of 999 additional substructures, so the sparsity or completeness depended on the number of fragments.

Subsequently, rigid-body refinement and rescoring in *Phaser* is performed and then a maximum of 58 best-scoring combinations in terms of LLG and CC are subjected to extension in *SHELXE* together with the randomly translated solution and the best solution, totalling 60 substructures. An illustration of the substructures that are involved in the verification step is shown in figure 26.

The results of the extensions are then compared for evidence of discrimination between the randomly translated solution or groups of consistent solutions. The rationale is that if the discrimination from the best solution and the random solution persists or the final solutions are equivalent, confidence in this solution will be justified. Thus, the best solution is validated if it can be clearly discriminated from the random solution or if different perturbations develop into a group of equivalent solutions. Conversely, it is not validated if the best solution cannot be discriminated from the random solution or the extension of inconsistent solutions leads to inconclusive results with structurally different structures characterized by comparable figures of merit.

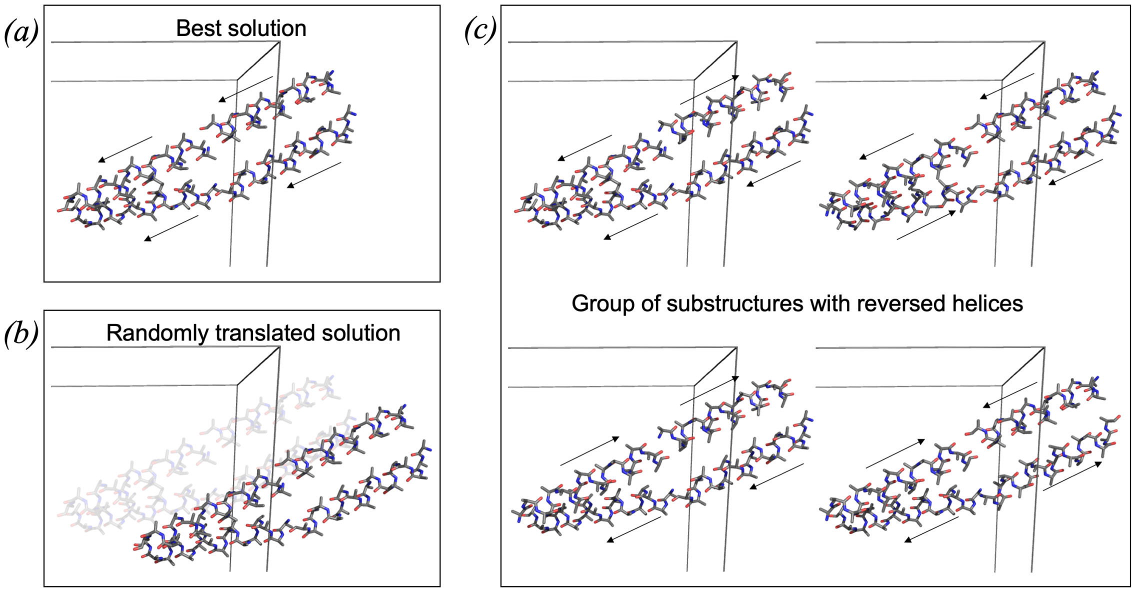

(a)
Best solution
(c)
(b)
Randomly translated solution
Group of substructures with reversed helices

Figure 26. Substructures involved in the verification step for the case of PDB id 1deb. a) Substructure leading to the best solution of *ARCIMBOLDO_LITE*. b) Randomly translated solution generated by applying a fractional translation vector of (0.1, 0.1, 0.1), behind there is the picture of the best solution for comparison. c) Group of substructures with reversed helices, the arrows indicate the direction of the helices. This is an illustration of the type of substructures that will be generated, as in the real case, with four helices the number of possible combinations is 16, as each helix could be in two directions.

This procedure was tested in all the cases in the test set presenting a resolution between 2 and 3 Å. Three scenarios in the solved structures were found, as reversed helices can sometimes be corrected by iterative tracing in *SHELXE*. In the first scenario, *SHELXE* has successfully reversed all the helices that were in reverse direction (figure 27a), in the second scenario, *SHELXE* has successfully reversed all the helices or none (figure 27b), in the third scenario, *SHELXE* has successfully reversed all the helices, some helices or none (figure 27c).

In all these scenarios, the best solution can be clearly discriminated from the random solution, since the difference between their correlation coefficients is greater than 15%. In the unsolved structures, all the traces are random and the figures of merit similar to the wrong solution (figure 27d). In these cases, the best solution cannot be discriminated against from the random solution and the difference between their correlation coefficients is less than 9%.

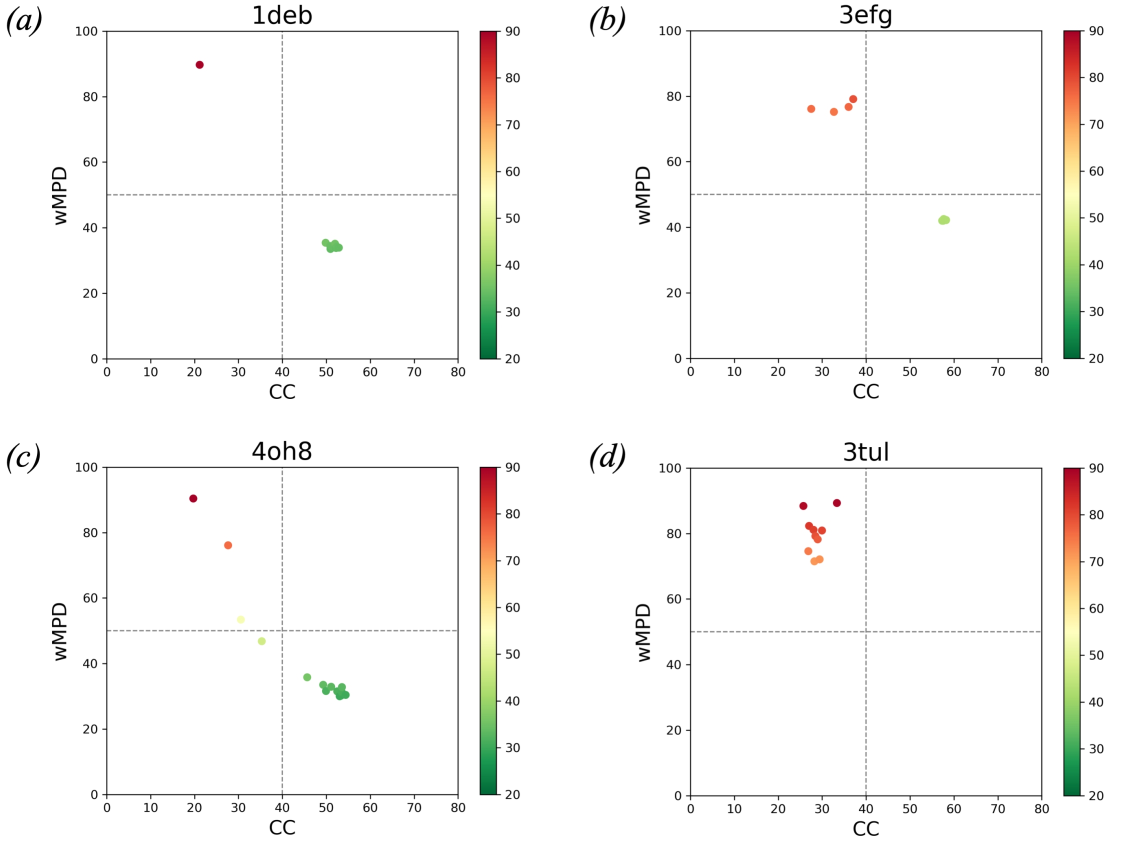


Figure 27. Verification results in the case of a) PDB id 1deb, where the final traces are equivalent and correct, two differentiated groups can be observed, the random solution and the group of substructures with all the helices corrected by *SHELXE* within the best solution. b) PDB id 3efg with two differentiated groups, the random solution with some substructures with all the helices in reversed direction and the best solution with some substructures with corrected helices. c) PDB id 4oh8 with a disperse group where *SHELXE* has corrected all the helices, some of them or none. d) PDB id 3tul that remains unsolved and there is one group where the substructures are close to the random solution. The x axis is the CC of the trace (%) and the y axis is the MPD (°) with respect to the best substructure, which is also coloured from green (more structurally similar) to red (more structurally different).

There is an interval where the difference between the CC from the best and the wrong solution, between 15% and 9%, is not enough to discriminate if a structure is solved or not. In all these cases, the group of substructures with reversed helices have a similar CC, either because there are equivalent correct solutions or because there are structurally different structures characterized by comparable CC. We observed that if all are correct solutions, they have small structural differences, as all are the same solution, whereas if the structure is not solved, the inconsistent solutions are structurally different. Furthermore, in the case of solved structures, the structural difference between the random solution and a correct solution is larger than in the case of unsolved structures. This is illustrated by the case of PDB entries 3vir (solved) and 4pna (not solved) in figure 28.

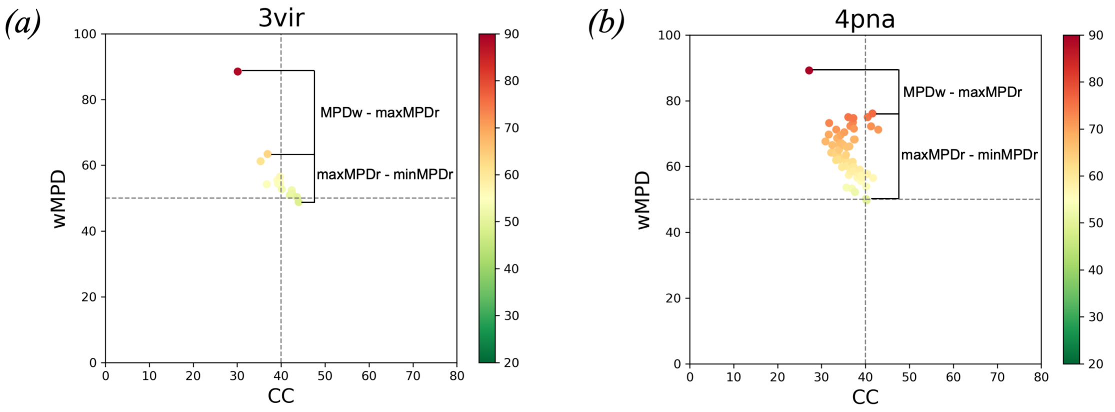


Figure 28. Verification results in the case of a) PDB id 3vir, the substructure expansion led to equivalent correct solutions where tracing had reversed the incorrect portions. The minor differences in CC or MPD are irrelevant and are derived from slight differences in the extension of the trace and its deviation from the ideal geometry. b) PDB id 4pna, the extension of inconsistent solutions leads to inconclusive results with structurally different structures characterized by comparable CC. The MPD is calculated with respect to the best solution, the difference between the minimum and maximum value of MPD of the reverse group is smaller in solved structures and larger in unsolved structures. In contrast, the difference between the MPD of the random solution and the maximum value of the of MPD of the reverse group is larger in solved structures and smaller in unsolved structures.

Figure 29 shows a workflow to decide whether a structure is solved, unsolved or no conclusive discrimination can be reached. The verification step will only be performed if the best scoring solution from *ARCIMBOLDO_LITE* has reached a CC above 25% and thus it is susceptible of being correct.

As mentioned above, in the case that the generation of a random solution failed, the MPD between the best solution and the random one (MPDw) will be small (less than 70º), and in this case, no conclusive distinction between the best solution and the random one could be reached. In those cases, the user could attempt model building and refinement to see whether the structure refines to acceptable R-values. A good approach is the use of *SEQUENCE SLIDER* (Borges *et al.*, 2020), since the correct assignment of the sequence in coiled-coils is challenging and this strategy generates all possible sequence hypotheses which then are refined.

A structure is solved if the difference between the CC between the best (CCbest) and the wrong or random solution (CCw) is above 15%. In contrast, it is not solved if this difference is less than 9%.

If this value is between 15% and 9% first it is necessary to check that the group of substructures with reversed helices have a similar CC. Thus, the difference between the highest CC and the lowest CC (diff_CC) must be lower than 13% and their coefficient of variation (CV_CC), calculated dividing the standard deviation by the mean of the MPD of the reversed group and multiplied by 100, must be lower than the 10%. Then structural differences are examined. The verification step will output that the structure is solved if the difference between the minimum and maximum value of MPD of the reverse group (maxMPDr - minMPDr) is lower than 16º and the difference between the MPD of the random solution and the maximum value of the of MPD of the reverse group (MPDw – maxMPDr) is above 21º, on the contrary, the verification step will output that the structure is not solved.

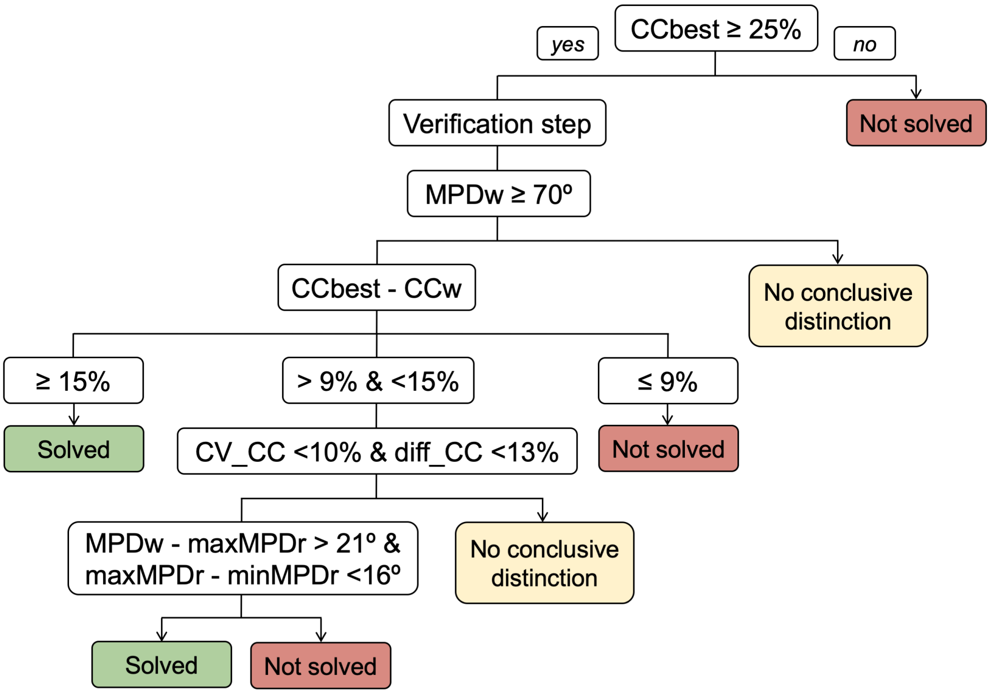

CCbest ≥ 25%
yes
no
Verification step
Not solved
MPDw ≥ 70º
CCbest - CCw
No conclusive distinction
≥ 15%
> 9% & <15%
≤ 9%
Solved
CV_CC <10% & diff_CC <13%
Not solved
MPDw - maxMPDr > 21º & maxMPDr - minMPDr <16º
No conclusive distinction
Solved
Not solved

Figure 29. Decision workflow representing the verification step. Colours represent if a structure is solved (green), not solved (red), or not discriminable (yellow).

2.4. Unsolved Structures

Only eight of 150 structures have remained unsolved. The percentage of unsolved structures for the first pool was 3.2% (three in 94), and for the second pool was 8.9% (five in 56). The unsolved structures can be classed as having very poor data or a large content in their asymmetric unit.

Three structures (PDB entries 3azd, 4pna and 3s4r) diffract at resolution better than 2.5

Å. PDB id 3azd shows an alarming validation report, with an unusually high clash score and poor side-chain geometry. Furthermore, its deposited data are extremely incomplete, the resolution of the deposited data spans from 0.9 to 2.7 Å, but it totally lacks resolution data below 2.7 Å. Also, over one half (57.57%) of the amplitudes have their associated sigmas set to zero. Improving the diffraction experiment and/or correcting an error during deposition would likely increase the chances of solving this structure, but this is outside the scope of this work. PDB id 4pna is affected by anisotropy and ice rings and could not be solved. It is noteworthy that the related PDB id 5f2y (not in the test set), which is a point mutant of the same protein in space group *I*2 that diffracted to the same nominal resolution and also affected by anisotropy, was trivially solvable. PDB id 3s4r also has completeness issues (data are only 85% complete) and severe anisotropy. Indeed, this structure could be solved with ideal data calculated from the model.

The remaining five structures (PDB entries 3iv1, 1u4q, 4xa3, 3tul, and 3mqc) all diffracted to 2.5 Å resolution or worse. Data from PDB id 4xa3 were highly anisotropic so the authors (Taylor *et al.*, 2015) performed ellipsoidal truncation and anisotropic scaling applying a reciprocal space resolutions cut-offs of 3.3 Å, 2.5 Å, and 2.6 Å for a*, b*, and c* respectively. Thus, the deposited data then have completeness issues (data are only 77% complete) due to anisotropic truncation. PDB id 3mqc has a large number of outliers in the data and anisotropy, and is isostructural with the solved PDB id 3nwh, although the former contains a somewhat longer construct. PDB entries 4xa3 and 3mqc could be solved with ideal data calculated from the model. PDB entries 1u4q (635 residues), 3iv1 (624 residues) and 3tul (521 residues), with more than 500 residues in the asymmetric unit are characterized by an expected LLG (Oeffner *et al.*, 2018) of 11.2 or less for placement of a helix of 30 residues due to the limits imposed by the number of reflections; thus, it is not surprising that they cannot be solved on a workstation even with ideal data calculated from the model to the experimental resolution.

These structures have not been solved, and they provide good examples of the reliability of the figures of merit as a metric for identifying correct solutions, as illustrated in table 9. Only 3azd has a resolution better than 2 Å, and at atomic resolution, the CC indicates unequivocally that the structure is not solved. For the other structures at resolution worse than 2 Å, PDB entries 3iv1 and 1u4q have a CC lower than 25%. For the other five structures, this figure of merit cannot be used to discriminate if a structure is solved or not, but thanks to the verification step, introduced in the present work, the final solution can be discriminated as incorrect. So, in a real case, despite having wrong solutions with good figures of merit, with the verification step confidence in the result will be justified.

Table 9. Results for the unsolved structures and decision of the verification step. The verification step is not performed if the CC is below 25%.

PDB	Resolution (Å)	# of residues	CC (%)	Residues traced	wMPE (º)	Verification
3azd	0.98	60	5.2	26	88.9	Not performed
4pna	2.1	208	42.4	173	86.4	Not solved
3s4r	2.45	179	31.0	159	89.7	Not solved
3iv1	2.50	624	24.8	257	89.6	Not performed
1u4q	2.50	635	16.6	143	89.1	Not performed
4xa3	2.55	284	28.2	144	89.0	Not solved
3tul	2.79	521	33.6	275	89.6	Not solved
3mqc	2.80	400	40.2	239	87.3	Not solved

2.5. Performance of the *coiled_coil* mode

Figure 30 summarizes the single-workstation performance of *ARCIMBOLDO_LITE* on the set of 150 test structures. The previous sections describe the particular problems that prevented some of the remaining 72 structures from being immediately solved and solutions for these problems, which led to phasing solutions in a further, previously unsuccessful, 64 cases. In total, of the 150 structures, 142 (95%) were solved. Appendix table A2 condenses characteristics and results for each of the structures probed.

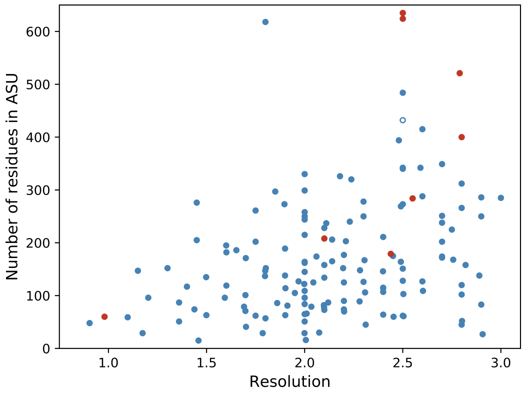


Figure 30. Performance of *ARCIMBOLDO_LITE* on a pool of 150 coiled-coil test structures. A total of 142 structures, corresponding to 95% of the cases were solved. Structures phased on eight-core machines are represented by blue dots. Open dots mark cases where more powerful hardware (a 24-core workstation) was required. The red dots mark the eight unsolved cases. The abscissa represents the resolution and ordinate represents the asymmetric unit content characterizing the test cases.

Regarding the characteristics of successfully solved cases, they covered the full range of resolution data in the set, from the highest resolution structure with 0.9 Å (PDB id 1byz) to the lowest resolution structure with 3 Å (PDB id 4qkv). In terms of length and complexity, a wide range is covered as well, from the smallest structure with just a single chain in the asymmetric unit comprising 15 residues (PDB id 1kyc) to the largest structure with four chains in the asymmetric unit totalling 618 residues (PDB id 2efr).

The total run time for *ARCIMBOLDO_LITE* jobs in *coiled_coil* mode typically took a few hours but ranged from five minutes for PDB id 1s9z searching for one helix of 18 amino acids to 19 hours for PDB id 5jxc searching for 12 helices of 18 residues, when the data extended resolutions better than 2 Å and when executed on the eight-core machines described in materials and methods. Lower resolution cases required more intensive computations owing to helix-orientation reversion and verification of potential solutions, which proved to be critical for ruling out false positives.

This *coiled_coil* mode in *ARCIMBOLDO_LITE* incorporates a new search algorithm to probe and verify alternative helix directions. It relies on advances in the MR search (Oeffner *et al.*, 2018) and autotracing (Usón & Sheldrick, 2018). Finally, the results of our tests show that the new mode substantially extends the range of data suitable for fragment phasing of coiled-coil structures, and thus the resolution limit has been extended from 2.5 Å for the general mode of *ARCIMBOLDO_LITE* to 3.0 Å for the *coiled_coil* mode.

2.6. Practical application: unknown coiled-coil structures

This section describes four novel structures that were solved with the *coiled_coil* mode; two structures of the synaptonemal complex protein 3 (SYCP3) in two crystal forms ($P2_1$ and $P1$), and two structures of peptide-based nanotubes. In addition to the complications of a typical coiled-coil, the solution of the peptide-based nanotubes was complicated by one structure having tNCS and the other having twinning.

2.6.1. *Mus musculus* SYCP3 homotetramer in two crystal forms ($P2_1$ and $P1$)

The synaptonemal complex protein 3 (SYCP3) is an axis core protein that plays a key role in meiotic chromosome organization and recombination. It consists of a coiled-coil assembly that further oligomerizes into micron-length filaments.

Crystals of this protein were obtained and described in (West *et al.*, 2018; West *et al.*, 2019) and belonged to two different space groups ($P2_1$ and $P1$), at 2.5 Å and 2.2 Å resolution respectively. Both crystal forms contain a single tetramer in the asymmetric unit, totalling 576 residues. Crystals were expected to contain a heterotetramer of *M. musculus* SYCP2 and SYCP3. However, upon solution and model building, it was apparent that these crystals contained SYCP3 homotetrameric complexes.

The *coiled_coil* mode implemented in *ARCIMBOLDO_LITE* allowed the solution of the structure in both space groups with a search configured to find four helices of 30 polyalanine residues. The best solution of the structure in the $P2_1$ crystal form coming from the *ARCIMBOLDO_LITE* run, which had an initial CC of 22.4%, and after seven cycles of iterative density modification and main chain autotracing with *SHELXE* reached a solution with a final CC of 38.9% and 393/576 amino acids placed. The structure in the *P*1 crystal form was successfully determined after eight cycles of expansion with *SHELXE*, which improved the starting CC from 17.2% to 35.6% and traced 371 of 576 residues. Finally, the verification step confirmed the confidence in these solutions.

The partial structures rendered by *ARCIMBOLDO_LITE* after *SHELXE* autotracing were completed through iterative manual building in *Coot* (Emsley *et al.*, 2010) and refinement using *phenix.refine* (Afonine *et al.*, 2012). Coordinate files and structure factors for structures in both monoclinic and triclinic crystal forms have been deposited in the Protein Data Bank with PDB codes 6dd8 and 6dd9 respectively. Data collection and refinement statistics are shown in table 10 and the structures are represented in figure 31.

Table 10. Data Collection and Refinement Statistics of the *Mus musculus* SYCP3 homotetramer in two crystal forms ($P2_1$ and *P*1).

	***Mm* SYCP3^{CC} $P2_1$ (6dd8)**	***Mm* SYCP3^{CC} *P*1 (6dd9)**
Data collection		
Space Group	$P2_1$	*P*1
Unit Cell Dimensions (a, b, c) Å	45.89, 49.49, 150.56	45.84, 52.40, 75.33
Unit cell Angles (α, β, γ) °	90, 90.79, 90	94.73, 103.99, 110.47
Resolution (Å)	47 – 2.5	48.26 - 2.2
Completeness %	99.86	99.67
Overall I/σ	15.0	16.1
Refinement		
R_{work} %	25.4	28.7
R_{free} %	32.3	33.1

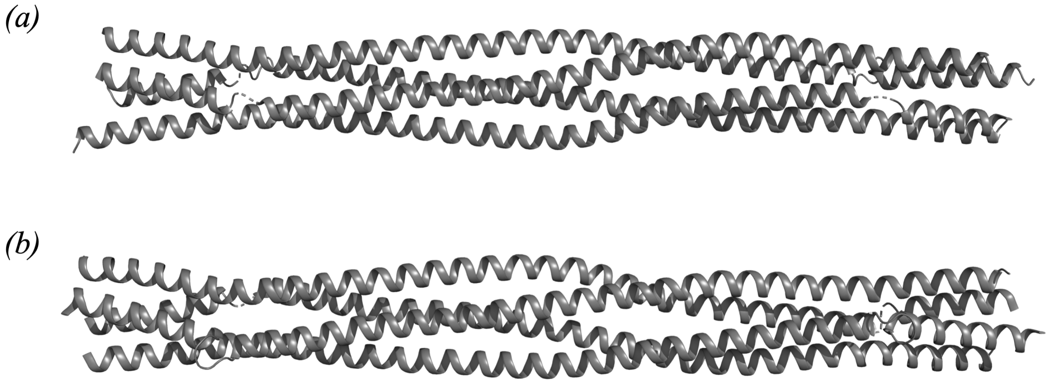


Figure 31. Cartoon representation of the homotetramer of SYP3 in a) monoclinic ($P2_1$) and b) triclinic (*P*1) crystal forms.

2.6.2. A peptide-based nanotube with tNCS

Peptide-based nanotubes are self-assembling peptides that undergo spontaneous assembling into ordered nanostructures. These artificial tubular constructs constitute a new class of biomaterials that are highly attractive for many biomedical applications such as drug delivery, scaffolds for tissue engineering, molecular electronics and many others (Burgess *et al.*, 2015). The motifs are usually based on repeat motifs that form helical assemblies that can potentially close into peptide-based nanotubes. Naturally occurring self-assembly motifs are coiled-coils.

One of these structures belongs to space group *P*1 at 2 Å resolution and contains four copies of a helix, totalling 128 residues in the asymmetric unit. In the analysis of tNCS carried out by *Phaser*, a peak in the Patterson function was found separated from the origin peak by 11.7 Å, a height of 57.7% of the origin and a translation vector of one half in the direction (0.5, 0, 0), corresponding to a tNCS order of two with pseudo-centring. The epsilon factors follow a bimodal distribution (figure 32), demonstrating the presence of tNCS.

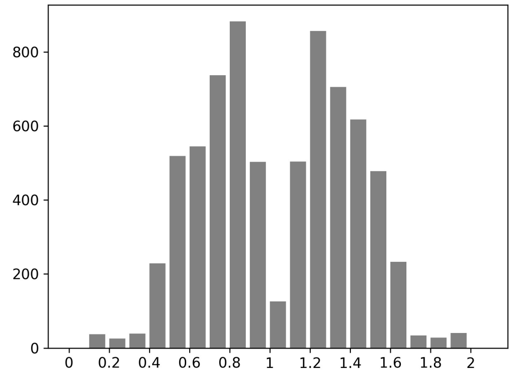


Figure 32. Histogram showing the distribution of refined tNCS epsilon factors. The structure presents clear modulation in the data indicated by a variance about one (σ_1^2) of 0.36. In the abscissa, there are the values of the epsilon factors.

The calculation of the epsilon factor distribution is essential to detect if a coiled-coil structure presents true intramolecular tNCS, as the internal symmetry of the helix itself can generate significant Patterson peaks in the absence of tNCS, and trying to place tNCS-related copies might cause fragment overlap and structure solution to fail.

So, the *ARCIMBOLDO_LITE* run with the *coiled_coil* mode was performed activating the search for pairs of helices related by tNCS. The structure was solved with a search for four helices of 20 polyalanine residues, followed by eight cycles expansion with *SHELXE*. A final solution with CC of 42.2% and 98 residues traced was reached and confirmed by the verification step.

Beyond the solution obtained with *ARCIMBOLDO_LITE*, further completion was done modelling side chain with *SEQUENCE SLIDER* (Borges *et al.*, 2020), another program developed in our laboratory that supports solution of coiled-coil and partial model structures extending polyalanine helices with all possible sequence assignments compatible with the structure.

The model coming from the *ARCIMBOLDO_LITE* run with a CC of 42.2% is composed of four chains; chain A with 36 residues, chain B with 31, chain C with 21 residues and chain D with 14 residues. Terminal residues were eliminated due to lack of electron density. Long refinement of this polyalanine model with *BUSTER* (Bricogne *et al.*, 2017) gives an R/R_{free} of 36.5%/44.3%. Table 11 shows the results of the first run of *SEQUENCE SLIDER.*

Table 11. Results for the first run of *SEQUENCE SLIDER.*

Chain	# of residues	# of models generated	# of distinguished solutions
A	27	11	2
B	28	10	1
C	19	19	-
D	12	26	1

Results for chains A, B and D had clear discrimination between correct and incorrect placements. The best map from *SEQUENCE SLIDER* was used to complete the original polyalanine model by extending helix C to 32 and D to 29 residues respectively. This polyalanine model improved R/R_{free} factors to 35%/43.3% and the side-chain electron densities of chains A and D were in agreement with *SEQUENCE SLIDER* solutions.

Subsequently, established sequences were assigned and modelled onto chains A and D, while chains B and C were left as polyalanine. Side-chains correctness was further supported by improvement in R/R_{free} upon refinement with *BUSTER*, 31%/38.2%. This new model was used for a fresh *SEQUENCE SLIDER* run evaluating only chains B and C (table 12).

Table 12. Results for the second run of *SEQUENCE SLIDER.*

Chain	# of residues	# of models generated	# of distinguished solutions
A	27	fixed	fixed
B	29	9	1
C	32	6	1
D	29	fixed	fixed

Upon phase improvement, a clear distinction among the models generated for chains B and C was obtained. Therefore, the correct sequence was set for all chains and further manual building with *Coot* build all residues for which electron density could be seen, that results in the total number of residues that the structure has, 128. A final refinement with *BUSTER* gave R/R_{free} of 24.6%/33.2%. All the refinement steps and results are summarized in table 13, the data collection and refinement statistics are shown in table 14, and the structure is represented in figure 33.

Table 13. Summary of the refinement steps and results.

	R (%)	**R_{free} (%)**
Solution from *ARCIMBOLDO_LITE* without some terminal residues (polyAla) containing four chains and 86 residues.	36.5	44.3
Previous model with helices enlargement (polyAla) containing four chains and 117 residues.	35.0	43.3
Model from the first run of *SEQUENCE SLIDER* with chains A and D containing side chains and B and C composed of polyalanine. Four chains and 117 residues (56 residues containing side chains).	31.0	38.2
Model from the second run of *SEQUENCE SLIDER* with all chains containing side chains and enlarged to 32 residues. Four chains and 128 residues.	24.6	33.2

Table 14. Data Collection and Refinement Statistics of the peptide-based nanotube with tNCS.

Data collection	
Space Group	*P*1
Unit Cell Dimensions (a, b, c) Å	23.48, 27.92, 45.43
Unit cell Angles (α, β, γ) º	93.60º, 90.81º, 113.13º
Resolution (Å)	17.63 - 2.0
Completeness %	99.82
Overall I/σ	12.2
Refinement	
R_{work} %	24.6
R_{free} %	33.2

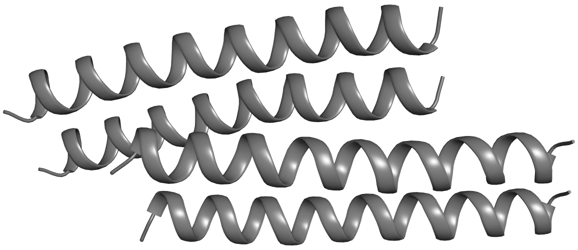


Figure 33. Cartoon representation of the peptide-based nanotube.

2.6.3. A peptide-based nanotube with twinning

Another peptide-based nanotube at 1.3 Å resolution was solved with the *coiled_coil* mode in the presence of twinning.

During the space group determination, *XPREP* indicated two equally probable space groups, *R*3 and *R*32. After a cell content analysis with *Phaser*, one copy of the peptide could only fit in space group *R*3 and not in *R*32, making it clear that the correct space group was *R*3. As data appear to have erroneously high symmetry, this led us to think that the crystal could be twinned, and indeed, twinning in *R*3 may give apparent *R*32 symmetry.

For that reason, data analysis was made with *phenix.xtriage*. The results showed that the intensity statistics are significantly different from those expected from untwinned data, indicating merohedral twinning. The values are summarized in table 15. Furthermore, the twin fraction was estimated by the H test of Yeates and the negative intensity Britton test giving a twin fraction of 0.46 and 0.44 respectively. Eventually confirmed by refinement, with a twin fraction of 0.48.

Table 15. Intensity statistics from *phenix.xtriage.*

Type of statistic	Value	Reference values
$<I^2>/<I>^2$	1.764	untwinned: 2.0, perfect twin: 1.5
$<F>^2/<F^2>$	0.841	untwinned: 0.785, perfect twin: 0.885
$<\|E^2-1\|>$	0.626	untwinned: 0.736, perfect twin: 0.541
$<\|L\|>$	0.434	untwinned: 0.5, perfect twin: 0.375
$<L^2>$	0.248	untwinned: 0.333, perfect twin: 0.2
Multivariate Z-score L-test	4.256	values > 3.5 indicates twinning

The analysis also indicates tNCS as there is a significant peak of 25% height of the origin peak. With one copy in the asymmetric unit, genuine tNCS is not possible. But high peaks in the Patterson are frequently present in coiled-coils due to the internal periodicity of the helix itself.

The solution of the structure was accomplished with the *coiled_coil* mode. A run configured to find one helix of 18 residues results in 32 residues traced out of 36, and is characterized by a CC of 34.1%.

Correct assignment of the sequence in coiled-coils is challenging. Our *SEQUENCE SLIDER* strategy generates all possible sequence hypotheses, which are refined with *SHELXL* (Sheldrick, 2015). To account for twinning during the refinement, the twin law must be specified. The twin law is: h,-h-k,-l, and this is input in the form of the 3 × 3 matrix components (1 0 0 -1 -1 0 0 0 -1).

Nine different assignments of the sequence were tested, the hypothesis with the best R-factors gives an R/R_{free} of 22.2%/32.3% and the worse an R/R_{free} of 25.8%/37.7%. After ten cycles of iterative refinement and model building of the best hypothesis with *SHELXL* and *Coot*, an R/R_{free} of 17.3%/23% was accomplished, the structure is represented in figure 34a and the data collection and refinement statistics are shown in table 16. Incorporating hydrogen atoms in riding positions to the model leads to an increase in the gap between R and R_{free} (16.9%/23.2%) and therefore, hydrogens were left out.

R-factors have higher values if twinning is not accounted for in the refinement step. If the R-factors with (figure 34b) and without twin refinement (figure 34c) are compared, the last model without twinning refinement gives an R/R_{free} of 26.4%/35.4% in comparison with 17.3%/23% with twin refinement.

Table 16. Data Collection and Refinement Statistics of the peptide-based nanotube with twinning.

Data collection	
Space Group	*R*3
Unit Cell Dimensions (a, b, c) Å	40.45, 40.45, 59.34
Unit cell Angles (α, β, γ) °	90°, 90°, 120°
Resolution (Å)	16.8 - 1.3
Completeness %	92.69
Overall I/σ	19.8
Refinement	
R_{work} %	17.3
R_{free} %	23

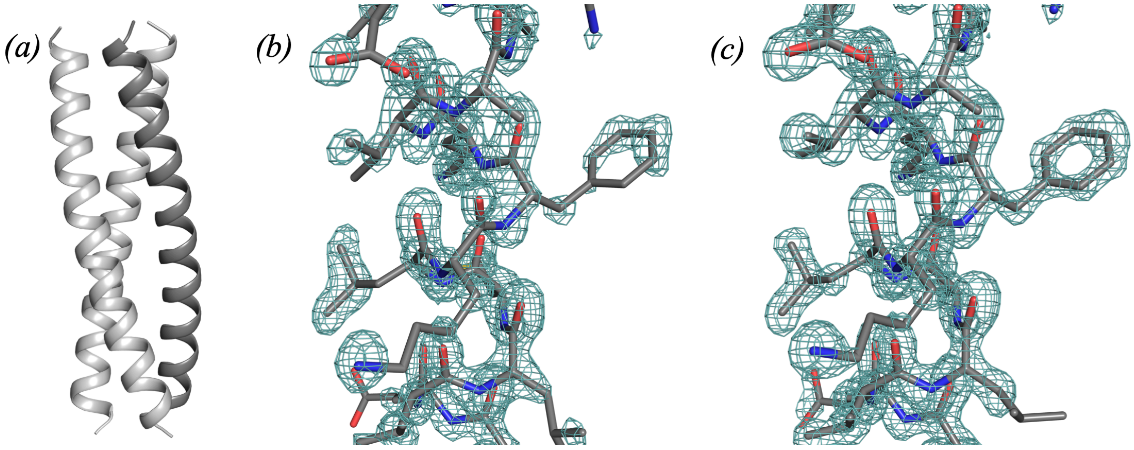


Figure 34. a) Cartoon representation of the peptide-based nanotube in dark grey with its symmetry equivalents in light grey. Comparison between equivalent portions of the electron density map contoured at 2 σ without b) twinning refinement and c) with twinning refinement.

CONCLUSIONS

The results of these studies led to the following conclusions.

In the first study, tNCS was analysed using a curated database of 80482 protein structures, to inform an algorithm for the detection of tNCS, which includes a method for detecting the number of vectors involved in any commensurate modulation (the tNCS order) (Caballero *et al.*, 2020). Our algorithm generates a ranked list of possible tNCS associations in the asymmetric unit, for exploration during structure solution.

- The parameters used in the Patterson function to detect tNCS were determined in the course of the following analysis:
    - A coordinate analysis was performed to detect tNCS depending on the choice of a rotational tolerance. Thanks to this, we can be sure that the 15 Å distance from the origin exclusion is safe.
    - Then the correlation of the tNCS in the coordinates with the Patterson highest non-origin peak was investigated and showed that up to 10º of rotational tolerance there is a clear modulation in the data.
    - A decision tree to predict the presence and absence of tNCS depending on the height of the Patterson highest non-origin peak was performed. The traditional resolution cut-off is maintained to 5-10 Å and the Patterson % threshold needs to be lowered from 20% to 16.8%.
- We also determined if the modulation in the data is significant and needs to be corrected:
    - Data modulation can be quantified through their epsilon factor distribution. Refined epsilon factors that cluster around one define unmodulated data, while those that refine to the extremes of the distribution indicate high modulation.
    - Finally, high Patterson peaks in the data may be due to order-disorder effects, for that reason, the case of no tNCS is also always included in the list of hypotheses.

The second study has led to the identification of limits and bottlenecks in coiled-coil phasing that have been addressed in a specific mode for solving coiled-coils (Caballero *et al.*, 2018) that was implemented in *ARCIMBOLDO_LITE*. It allows the solution of 142 of 150 test structures (95%), showing a higher success rate than the initial baseline, where only 52% of the test set was solved. This study significantly advances the phasing of coiled-coils structures.

- The problems identified for general cases of coiled-coils structures and their solution trigger the following defaults:
    - The ideal polyalanine helix used as a search fragment usually represents a small percentage of the scattering. For this reason, only parts of the target structure presenting a low value of rmsd are susceptible of being located, so our choice is setting a default value of 0.2 Å. Also, as rmsd refinement was required for solution identification in at least in one case, and it has not been observed to have adverse effects in any case, this calculation will be performed by default.
    - The internal symmetry of the helix itself can originate the same Patterson peaks derived from identical molecules related by tNCS. It makes it difficult to differentiate genuine intermolecular tNCS from Patterson artefacts, and overlapping solutions can occur. For this reason, the placement of pairs of tNCS related helices is deactivated.
    - Overlapping solutions may be characterized by extremely high LLG scores and no solutions pass the packing filter. The incorporation of a new packing filter during the translation search in *Phaser* was prompted by this study and its use will ensure that the top solution has acceptable packing. The peak height to accept further translation solutions will be relative to this first well-packed solution.
- In addition to the previous bottlenecks, at low-resolution, more difficulties were found, but thanks to the following improvements, the resolution limit currently established for *ARCIMBOLDO* from 2.5 to 3 Å has been extended:
    - At low-resolution, the placement of helices occasionally took place in the correct position but in reversed direction. To avoid this, helices reversed in the same positions are generated and tested.
    - Also, at low-resolution, the geometry of helical trace can degrade. This problem is solved thanks to a new autotracing algorithm in *SHELXE* with restrictions for helical tracing, the use of larger helical seeds and the extension of the polypeptide chain by sliding helical fragments.
    - Finally, wrong solutions can have deceptively high figures of merit. Thus, to verify the most promising solution, its original substructure will be perturbed by helix reversal and a random translation. The results of the various extensions are compared for evidence of discrimination between groups of consistent solutions.

    - Extension of the resolution limit imposed in *ARCIMBOLDO* from 2.5 to 3.0 Å in the case of coiled-coils has been enabled by the introduction of the verification step.

- All these new features are available in *ARCIMBOLDO_LITE*, which is distributed through PyPI and *CCP4*. The *coiled_coil* mode can be activated by setting a keyword named *coiled_coil* to true in the instruction file, or via a checkbox in the *CCP4* interface.
- Finally, this implementation has allowed solving four previously unknown structures.

REFERENCES

Abergel, C. (2013). *Molecular replacement: tricks and treats.* Acta Crystallogr D Biol Crystallogr 69, 2167-2173.

Afonine, P. V., Grosse-Kunstleve, R. W., Echols, N., Headd, J. J., Moriarty, N. W., Mustyakimov, M., Terwilliger, T. C., Urzhumtsev, A., Zwart, P. H. & Adams, P. D. (2012). *Towards automated crystallographic structure refinement with phenix.refine.* Acta Crystallographica Section D 68, 352-367.

Albertini, A. A., Wernimont, A. K., Muziol, T., Ravelli, R. B., Clapier, C. R., Schoehn, G., Weissenhorn, W. & Ruigrok, R. W. (2006). *Crystal structure of the rabies virus nucleoprotein-RNA complex.* Science 313, 360-363.

Aquilano, D., Pastero, L., Veeesler, S. & Astier, J. P. (2002). Space groups of crystals and polytypism. The interplay among symmetry glide elements, face characters and screw dislocations., Vol. ISBN: 88-218-0903-X, Crystal Growth: from basic to applied,, Rome (Italy).

Baxevanis, A. D. & Vinson, C. R. (1993). *Interactions of coiled coils in transcription factors: where is the specificity?* Curr Opin Genet Dev 3, 278-285.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *The Protein Data Bank.* Nucleic Acids Research 28, 235-242.

Bernhardt, E. & Herbst-Irmer, R. (2020). *Phase transition and structures of the twinned low-temperature phases of (Et4N)[ReS4].* Acta Crystallogr C Struct Chem 76, 231-235.

Blocquel, D., Habchi, J., Durand, E., Sevajol, M., Ferron, F., Erales, J., Papageorgiou, N. & Longhi, S. (2014). *Coiled-coil deformations in crystal structures: the measles virus phosphoprotein multimerization domain as an illustrative example.* Acta Crystallographica Section D 70, 1589-1603.

Borges, R. J., Meindl, K., Trivino, J., Sammito, M., Medina, A., Millan, C., Alcorlo, M., Hermoso, J. A., Fontes, M. R. d. M. & Usón, I. (2020). *SEQUENCE SLIDER: expanding polyalanine fragments for phasing with multiple side-chain hypotheses.* Acta Crystallographica Section D 76, 221-237.

Bragg, W. H. & Bragg, W. L. (1913). *The Reflection of X-rays by Crystals.* Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences 88, 428-438.

Bragg, W. L. (1913). *The Structure of Some Crystals as Indicated by Their Diffraction of X-rays.* Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences 89, 248-277.

Bragg, W. L. & Howells, E. R. (1954). *X-ray diffraction by imidazole methaemoglobin.* Acta Crystallographica 7, 409-411.

Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984). *Classification and Regression Trees*. Wadsworth, New York: Chapman and Hall.

Bricogne, G., Blanc, E., Brandl, M., Flensburg, C., Keller, P., Paciorek, W., Roversi, P., Sharff, A., Smart, O. S., Vonrhein, C. & Womack, T. O. (2017). *BUSTER.* Global Phasing Ltd., Cambridge, UK.

Britton, D. (1972). *Estimation of twinning parameter for twins with exactly superimposed reciprocal lattices.* Acta Crystallographica Section A 28, 296-297.

Brunger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Crystallography & NMR system: A new software suite for macromolecular structure determination.* Acta Crystallogr D Biol Crystallogr 54, 905-921.

Burgess, N. C., Sharp, T. H., Thomas, F., Wood, C. W., Thomson, A. R., Zaccai, N. R., Brady, R. L., Serpell, L. C. & Woolfson, D. N. (2015). *Modular Design of Self-Assembling Peptide-Based Nanotubes.* J Am Chem Soc 137, 10554-10562.

Burkhard, P., Stetefeld, J. & Strelkov, S. V. (2001). *Coiled coils: a highly versatile protein folding motif.* Trends in Cell Biology 11, 82-88.

Burley, S. K., Berman, H. M., Bhikadiya, C., Bi, C., Chen, L., Di Costanzo, L., Christie, C., Dalenberg, K., Duarte, J. M., Dutta, S., Feng, Z., Ghosh, S., Goodsell, D. S., Green, R. K., Guranović, V., Guzenko, D., Hudson, B. P., Kalro, T., Liang, Y., Lowe, R., Namkoong, H., Peisach, E., Periskova, I., Prlić, A., Randle, C., Rose, A., Rose, P., Sala, R., Sekharan, M., Shao, C., Tan, L., Tao, Y.-P., Valasatava, Y., Voigt, M., Westbrook, J., Woo, J., Yang, H., Young, J., Zhuravleva, M. & Zardecki, C. (2018). *RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy.* Nucleic Acids Research 47, D464-D474.

Caballero, I., Sammito, M., Afonine, P. V., Usón, I., Read, J. R. & McCoy, A. J. (2020). *Detection of translational non-crystallographic symmetry in Pattersons.* Acta Crystallogr D Struct Biol In revision.

Caballero, I., Sammito, M., Millán, C., Lebedev, A., Soler, N. & Usón, I. (2018). *ARCIMBOLDO on coiled coils.* Acta Crystallogr D Struct Biol 74, 194-204.

Campeotto, I., Lebedev, A., Schreurs, A. M. M., Kroon-Batenburg, L. M. J., Lowe, E., Phillips, S. E. V., Murshudov, G. N. & Pearson, A. R. (2018). *Pathological macromolecular crystallographic data affected by twinning, partial-disorder and exhibiting multiple lattices for testing of data processing and refinement tools.* Sci Rep 8, 14876.

Chook, Y. M., Lipscomb, W. N. & Ke, H. (1998). *Detection and Use of Pseudo-Translation in Determination of Protein Structures.* Acta Crystallographica Section D 54, 822-827.

Chothia, C. & Lesk, A. M. (1986). *The relation between the divergence of sequence and structure in proteins.* The EMBO Journal 5, 823-826.

Cochran, W. & Howells, E. R. (1954). *X-ray diffraction by a layer structure containing random displacements.* Acta Crystallographica 7, 412-415.

Crick, F. (1953). *The packing of [alpha]-helices: simple coiled-coils.* Acta Crystallographica 6, 689-697.

Das, R., Andre, I., Shen, Y., Wu, Y., Lemak, A., Bansal, S., Arrowsmith, C. H., Szyperski, T. & Baker, D. (2009). *Simultaneous prediction of protein folding and docking at high resolution.* Proc Natl Acad Sci U S A 106, 18978-18983.

Dauter, Z. (2003). *Twinned crystals and anomalous phasing.* Acta Crystallogr D Biol Crystallogr 59, 2004-2016.

Dauter, Z. (2015). *Solving coiled-coil protein structures.* IUCrJ 2, 164-165.

Dauter, Z., Botos, I., LaRonde-LeBlanc, N. & Wlodawer, A. (2005). *Pathological crystallography: case studies of several unusual macromolecular crystals.* Acta Crystallographica Section D 61, 967-975.

Dauter, Z. & Jaskolski, M. (2016). *Crystal pathologies in macromolecular crystallography.* Postepy Biochem 62, 401-407.

Debye, P. (1913). *Interference of X rays and heat movement.* Annalen der Physik 348, 49-92.

Dornberger-Schiff, K. (1956). *On order-disorder structures (OD-structures).* Acta Crystallographica 9, 593-601.

Dornberger-Schiff, K. (1966). *Reinterpretation of pseudo-orthorhombic diffraction patterns.* Acta Crystallographica 21, 311-322.

Dornberger-Schiff, K. & Dunitz, J. D. (1965). *Pseudo-orthorhombic diffraction patterns and OD structures.* Acta Crystallographica 19, 471-472.

Dornberger-Schiff, K. & Grell-Niemann, H. (1961). *On the theory of order-disorder (OD) structures.* Acta Crystallographica 14, 167-177.

Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. (2010). *Features and development of Coot.* Acta Crystallogr D Biol Crystallogr 66, 486-501.

Fisher, R. G. & Sweet, R. M. (1980). *Treatment of diffraction data from crystals twinned by merohedry.* Acta Crystallographica Section A 36, 755-760.

Franke, B., Gasch, A., Rodriguez, D., Chami, M., Khan, M. M., Rudolf, R., Bibby, J., Hanashima, A., Bogomolovas, J., von Castelmur, E., Rigden, D. J., Uson, I., Labeit, S. & Mayans, O. (2014). *Molecular basis for the fold organization and sarcomeric targeting of the muscle atrogin MuRF1.* Open Biol 4, 130172-130172.

Franke, B., Rodriguez, D., Usón, I. & Mayans, O. (2011). *Phasing of an unpredicted palindromic coiled-coil motif.* Acta Crystallographica Section A 67, C65.

French, S. & Wilson, K. (1978). *On the treatment of negative intensity observations.* Acta Crystallographica Section A 34, 517-525.

Friedrich, W., Knipping, P. & Laue, M. (1913). *Interferenzerscheinungen bei Röntgenstrahlen.* Annalen der Physik 346, 971-988.

Fujinaga, M. & Read, R. J. (1987). *Experiences with a new translation-function program.* Journal of Applied Crystallography 20, 517-521.

Galdeano, C., Gadd, M. S., Soares, P., Scaffidi, S., Van Molle, I., Birced, I., Hewitt, S., Dias, D. M. & Ciulli, A. (2014). *Structure-guided design and optimization of small molecules targeting the protein-protein interaction between the von Hippel-Lindau (VHL) E3 ubiquitin ligase and the hypoxia inducible factor (HIF) alpha subunit with in vitro nanomolar affinities.* J Med Chem 57, 8657-8663.

Giacovazzo, C. (2002). *Fundamentals of Crystallography*. Oxford: Oxford University Press.

Green, D. W., Ingram, V. M. & Perutz, M. F. (1954). *The Structure of Haemoglobin. IV. Sign Determination by the Isomorphous Replacement Method.* Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences 225, 287-307.

Grosse-Kunstleve, R. W. & Adams, P. D. (2002). *On the handling of atomic anisotropic displacement parameters.* Journal of Applied Crystallography 35, 477-480.

Grosse-Kunstleve, R. W., Sauter, N. K., Moriarty, N. W. & Adams, P. D. (2002). *The Computational Crystallography Toolbox: crystallographic algorithms in a reusable software framework.* Journal of Applied Crystallography 35, 126-136.

Hare, S., Cherepanov, P. & Wang, J. (2009). *Application of general formulas for the correction of a lattice-translocation defect in crystals of a lentiviral integrase in complex with LEDGF.* Acta Crystallogr D Biol Crystallogr 65, 966-973.

Hatti, K. S., McCoy, A. J., Oeffner, R. D., Sammito, M. D. & Read, R. J. (2020). *Factors influencing estimates of coordinate error for molecular replacement.* Acta Crystallogr D Struct Biol 76, 19-27.

Hayashi, K., Tan, X., Zheng, N., Hatate, T., Kimura, Y., Kepinski, S. & Nozaki, H. (2008). *Small-molecule agonists and antagonists of F-box protein-substrate interactions in auxin perception and signaling.* Proc Natl Acad Sci U S A 105, 5632-5637.

Hendrickson, W. A. (1991). *Determination of macromolecular structures from anomalous diffraction of synchrotron radiation.* Science 254, 51-58.

Hendrickson, W. A. (2013). *Evolution of diffraction methods for solving crystal structures.* Acta Crystallogr A 69, 51-59.

Herbst-Irmer, R. (2016). *Twinning in chemical crystallography – a practical guide.* Zeitschrift für Kristallographie - Crystalline Materials 231.

Herbst-Irmer, R. & Sheldrick, G. M. (1998). *Refinement of Twinned Structures with SHELXL97.* Acta Crystallographica Section B 54, 443-449.

Herbst-Irmer, R. & Sheldrick, G. M. (2002). *Refinement of obverse/reverse twins.* Acta Crystallogr B 58, 477-481.

Hipp, D. R., Kennedy, D. & Mistachkin, J. (2015). *SQLite.* Version 3.8.10.2.

Hu, G., Gershon, P. D., Hodel, A. E. & Quiocho, F. A. (1999). *mRNA cap recognition: Dominant role of enhanced stacking interactions between methylated bases and protein aromatic side chains.* Proceedings of the National Academy of Sciences 96, 7149.

Hunter, J. D. (2007). *Matplotlib: A 2D graphics environment.* Computing in Science & Engineering 9, 90-95.

Jamshidiha, M., Perez-Dorado, I., Murray, J. W., Tate, E. W., Cota, E. & Read, R. J. (2019). *Coping with strong translational noncrystallographic symmetry and extreme anisotropy in molecular replacement with Phaser: human Rab27a.* Acta Crystallogr D Struct Biol 75, 342-353.

Janin, J. & Rodier, F. (1995). *Protein–protein interaction at crystal contacts.* Proteins: Structure, Function, and Bioinformatics 23, 580-587.

Joosten, R. P., Salzemann, J., Bloch, V., Stockinger, H., Berglund, A. C., Blanchet, C., Bongcam-Rudloff, E., Combet, C., Da Costa, A. L., Deleage, G., Diarena, M., Fabbretti, R., Fettahi, G., Flegel, V., Gisel, A., Kasam, V., Kervinen, T., Korpelainen, E., Mattila, K., Pagni, M., Reichstadt, M., Breton, V., Tickle, I. J. & Vriend, G. (2009). *PDB_REDO: automated re-refinement of X-ray structure models in the PDB.* J Appl Crystallogr 42, 376-384.

Kalman, Z. E., Mészáros, B., Gáspári, Z. & Dobson, L. (2020). *Distribution of disease-causing germline mutations in coiled-coils suggests essential role of their N-terminal region.* bioRxiv, 2020.2004.2007.029165.

Karle, J. & Hauptman, H. (1956). *A theory of phase determination for the four types of non-centrosymmetric space groups 1P222, 2P22, 3P12, 3P22.* Acta Crystallographica 9, 635-651.

Kondo, J., Urzhumtseva, L. & Urzhumtsev, A. (2008). *Patterson-guided ab initio analysis of structures with helical symmetry.* Acta Crystallogr D Biol Crystallogr 64, 1078-1091.

Krivov, G. G., Shapovalov, M. V. & Dunbrack, R. L., Jr. (2009). *Improved prediction of protein side-chain conformations with SCWRL4.* Proteins 77, 778-795.

Kuhn, M., Hyman, A. A. & Beyer, A. (2014). *Coiled-coil proteins facilitated the functional expansion of the centrosome.* PLoS Comput Biol 10, e1003657.

Kuratani, M., Ishii, R., Bessho, Y., Fukunaga, R., Sengoku, T., Shirouzu, M., Sekine, S. & Yokoyama, S. (2005). *Crystal structure of tRNA adenosine deaminase (TadA) from Aquifex aeolicus.* J Biol Chem 280, 16002-16008.

Laue, M. (1913). *Eine quantitative Prüfung der Theorie für die Interferenzerscheinungen bei Röntgenstrahlen.* Annalen der Physik 346, 989-1002.

Lebedev, A. A. (2009). *On some implications of non-crystallographic symmetry.* thesis*,* University of York.

Lebedev, A. A., Vagin, A. A. & Murshudov, G. N. (2006). *Intensity statistics in twinned crystals with examples from the PDB.* Acta Crystallogr D Biol Crystallogr 62, 83-95.

Liebschner, D., Afonine, P. V., Baker, M. L., Bunkoczi, G., Chen, V. B., Croll, T. I., Hintze, B., Hung, L.-W., Jain, S., McCoy, A. J., Moriarty, N. W., Oeffner, R. D., Poon, B. K., Prisant, M. G., Read, R. J., Richardson, J. S., Richardson, D. C., Sammito, M. D., Sobolev, O. V., Stockwell, D. H., Terwilliger, T. C., Urzhumtsev, A. G., Videau, L. L., Williams, C. J. & Adams, P. D. (2019). *Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix.* Acta Crystallographica Section D 75, 861-877.

Liu, J. & Rost, B. (2001). *Comparing function and structure between entire proteomes.* Protein Science 10, 1970-1979.

Lovelace, J., Murphy, C., Daniels, L., Narayan, K., Schutt, C., Lindberg, U., Svensson, C. & Borgstahl, G. (2008). *Protein crystals can be incommensurately modulated.* Journal of Applied Crystallography - J APPL CRYST 41, 600-605.

Lovelace, J., Winn, M. & Borgstahl, G. (2010). *Simulation of modulated reflections.* Journal of Applied Crystallography 43.

Lovelace, J. J., Simone, P. D., Petricek, V. & Borgstahl, G. E. (2013). *Simulation of modulated protein crystal structure and diffraction data in a supercell and in superspace.* Acta Crystallogr D Biol Crystallogr 69, 1062-1072.

Lunin, V. Y. & Woolfson, M. M. (1993). *Mean phase error and the map-correlation coefficient.* Acta Crystallogr D Biol Crystallogr 49, 530-533.

Lupas, A. (1996). *Coiled coils: new structures and new functions.* Trends in Biochemical Sciences 21, 375-382.

Lupas, A. N. & Gruber, M. (2005). *The structure of alpha-helical coiled coils.* Adv Protein Chem 70, 37-78.

MacKenzie, D. A., Tailford, L. E., Hemmings, A. M. & Juge, N. (2009). *Crystal structure of a mucus-binding protein repeat reveals an unexpected functional immunoglobulin binding activity.* J Biol Chem 284, 32444-32453.

Mason, J. M. & Arndt, K. M. (2004). *Coiled coil domains: stability, specificity, and biological implications.* Chembiochem 5, 170-176.

Massey, T. H., Mercogliano, C. P., Yates, J., Sherratt, D. J. & Lowe, J. (2006). *Double-stranded DNA translocation: structure and mechanism of hexameric FtsK.* Mol Cell 23, 457-469.

Matthews, B. W. & Czerwinski, E. W. (1975). *Local scaling: a method to reduce systematic errors in isomorphous replacement and anomalous scattering measurements.* Acta Crystallographica Section A 31, 480-487.

McCoy, A. J. (2020). In revision.

McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. & Read, R. J. (2007). *Phaser crystallographic software.* J Appl Crystallogr 40, 658-674.

McCoy, A. J., Oeffner, R. D., Wrobel, A. G., Ojala, J. R., Tryggvason, K., Lohkamp, B. & Read, R. J. (2017). *Ab initio solution of macromolecular crystal structures without direct methods.* Proc Natl Acad Sci U S A 114, 3637-3641.

McCoy, A. J. & Read, R. J. (2010). *Experimental phasing: best practice and pitfalls.* Acta Crystallogr D Biol Crystallogr 66, 458-469.

Medina, A., Trivino, J., Borges, R. J., Millan, C., Usón, I. & Sammito, M. D. (2020). *ALEPH: a network-oriented approach for the generation of fragment-based libraries and for structure interpretation.* Acta Crystallographica Section D 76, 193-208.

Merritt, E. A. (1999). *Expanding the model: anisotropic displacement parameters in protein structure refinement.* Acta Crystallographica Section D 55, 1109-1117.

Mier, P., Alanis-Lobato, G. & Andrade, M. (2016). *Protein-protein interactions can be predicted using coiled coil co-evolution patterns.* Journal of Theoretical Biology 412.

Millán, C., Sammito, M. & Usón, I. (2015). *Macromolecular ab initio phasing enforcing secondary and tertiary structure.* IUCrJ 2, 95-105.

Morris, R. J. & Bricogne, G. (2003). *Sheldrick's 1.2 A rule and beyond.* Acta Crystallographica Section D 59, 615-617.

Murray-Rust, P. (1973). *The crystal structure of [Co(NH3)6]4Cu5Cl17: a twinned cubic crystal.* Acta Crystallographica Section B 29, 2559-2566.

Murshudov, G. (2011). *Some properties of crystallographic reliability index – R factor : effect of twinning.* Applied and Computational Mathematics 10.

Murshudov, G. N., Davies, G. J., Isupov, M., Krzywda, S. & Dodson, E. J. (1998). *The Effect of Overall Anisotropic Scaling in Macromolecular Refinement.* CCP4 Newsletter on Protein Crystallography 35, contribution 12.

Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Refinement of Macromolecular Structures by the Maximum-Likelihood Method.* Acta Crystallographica Section D 53, 240-255.

Navaza, J. (1994). *AMoRe: an automated package for molecular replacement.* Acta Crystallographica Section A Foundations of Crystallography 50, 157-163.

Oeffner, R. D., Afonine, P. V., Millán, C., Sammito, M., Usón, I., Read, R. J. & McCoy, A. J. (2018). *On the application of the expected log-likelihood gain to decision making in molecular replacement.* Acta Crystallogr D Struct Biol 74, 245-255.

Oeffner, R. D., Bunkoczi, G., McCoy, A. J. & Read, R. J. (2013). *Improved estimates of coordinate error for molecular replacement.* Acta Crystallogr D Biol Crystallogr 69, 2209-2215.

Osipiuk, J., Xu, X., Cui, H., Savchenko, A., Edwards, A. & Joachimiak, A. (2011). *Crystal structure of secretory protein Hcp3 from Pseudomonas aeruginosa.* J Struct Funct Genomics 12, 21-26.

Padmanabhan, S. & Freymann, D. M. (2001). *The conformation of bound GMPPNP suggests a mechanism for gating the active site of the SRP GTPase.* Structure 9, 859-867.

Parsons, S. (2003). *Introduction to twinning.* Acta Crystallogr D Biol Crystallogr 59, 1995-2003.

Patterson, A. L. (1935). *A direct method for the determination of the components of interatomic distances in crystals.* Z. Kristallogr 90, 517–542.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011). *Scikit-learn: Machine Learning in Python.* Journal of Machine Learning Research 12, 2825-2830.

Petricek, V., van der Lee, A. & Evain, M. (1995). *On the use of crenel functions for occupationally modulated structures.* Acta Crystallographica Section A 51, 529-535.

Pletnev, S., Morozova, K. S., Verkhusha, V. V. & Dauter, Z. (2009). *Rotational order-disorder structure of fluorescent protein FP480.* Acta crystallographica. Section D, Biological crystallography 65, 906-912.

Popov, A. N. & Bourenkov, G. P. (2003). *Choice of data-collection parameters based on statistic modelling.* Acta Crystallogr D Biol Crystallogr 59, 1145-1153.

Porta, J., Lovelace, J. J., Schreurs, A. M., Kroon-Batenburg, L. M. & Borgstahl, G. E. (2011). *Processing incommensurately modulated protein diffraction data with Eval15.* Acta Crystallogr D Biol Crystallogr 67, 628-638.

Rackham, O. J., Madera, M., Armstrong, C. T., Vincent, T. L., Woolfson, D. N. & Gough, J. (2010). *The evolution and structure prediction of coiled coils across all genomes.* J Mol Biol 403, 480-493.

Rämisch, S., Lizatovic, R. & Andre, I. (2015). *Automated de novo phasing and model building of coiled-coil proteins.* Acta Crystallographica Section D 71, 606-614.

Read, R. J. (1986). *Improved Fourier coefficients for maps using phases from partial structures with errors.* Acta Crystallographica Section A 42, 140-149.

Read, R. J. (2001). *Pushing the boundaries of molecular replacement with maximum likelihood.* Acta Crystallogr D Biol Crystallogr 57, 1373-1382.

Read, R. J., Adams, P. D. & McCoy, A. J. (2013). *Intensity statistics in the presence of translational noncrystallographic symmetry.* Acta Crystallogr D Biol Crystallogr 69, 176-183.

Read, R. J. & McCoy, A. J. (2016). *A log-likelihood-gain intensity target for crystallographic phasing that accounts for experimental error.* Acta Crystallogr D Struct Biol 72, 375-387.

Rodríguez, D., Sammito, M., Meindl, K., de Ilarduya, I. M., Potratz, M., Sheldrick, G. M. & Usón, I. (2012). *Practical structure solution with ARCIMBOLDO.* Acta Crystallogr D Biol Crystallogr 68, 336-343.

Rodríguez, D. D., Grosse, C., Himmel, S., Gonzalez, C., de Ilarduya, I. M., Becker, S., Sheldrick, G. M. & Usón, I. (2009). *Crystallographic ab initio protein structure solution below atomic resolution.* Nat Methods 6, 651-653.

Rossmann, M. (1972). *The Molecular Replacement Method*. Gordon & Breach.

Rossmann, M. G. & Blow, D. M. (1962). *The detection of sub-units within the crystallographic asymmetric unit.* Acta Crystallographica 15, 24-31.

Rossmann, M. G. & Blow, D. M. (1964). *Solution of the phase equations representing non-crystallographic symmetry.* Acta Crystallographica 17, 1474-1475.

Roversi, P., Blanc, E., Johnson, S. & Lea, S. M. (2012). *Tetartohedral twinning could happen to you too.* Acta Crystallographica Section D 68, 418-424.

Rye, C. A., Isupov, M. N., Lebedev, A. A. & Littlechild, J. A. (2007). *An order-disorder twin crystal of l-2-haloacid dehalogenase from Sulfolobus tokodaii.* Acta Crystallographica Section D 63, 926-930.

Sammito, M., Meindl, K., de Ilarduya, I. M., Millán, C., Artola-Recolons, C., Hermoso, J. A. & Usón, I. (2014). *Structure solution with ARCIMBOLDO using fragments derived from distant homology models.* FEBS J 281, 4029-4045.

Sammito, M., Millán, C., Frieske, D., Rodriguez-Freire, E., Borges, R. J. & Usón, I. (2015). *ARCIMBOLDO_LITE: single-workstation implementation and use.* Acta Crystallogr D Biol Crystallogr 71, 1921-1930.

Sammito, M., Millán, C., Rodríguez, D. D., de Ilarduya, I. M., Meindl, K., De Marino, I., Petrillo, G., Buey, R. M., de Pereda, J. M., Zeth, K., Sheldrick, G. M. & Usón, I. (2013). *Exploiting tertiary structure through local folds for crystallographic phasing.* Nat Meth 10, 1099-1101.

Sanchez Rodriguez, F., Simpkin, A. J., Davies, O. R., Keegan, R. M. & Rigden, D. J. (2020). *Helical ensembles outperform ideal helices in molecular replacement.* Acta Crystallographica Section D 76, 962-970.

Sawaya, M. R. (2014). *Methods to refine macromolecular structures in cases of severe diffraction anisotropy.* Methods Mol Biol 1091, 205-214.

Schoch, G. A., Sammito, M., Millán, C., Usón, I. & Rudolph, M. G. (2015). *Structure of a 13-fold superhelix (almost) determined from first principles.* IUCrJ 2, 177-187.

Schönleber, A. (2011). *Organic molecular compounds with modulated crystal structures.* Zeitschrift für Kristallographie 226, 499-517.

Schreurs, A., Xian, X. & Kroon-Batenburg, L. M. J. (2010). *EVAL15: A diffraction data integration method based on ab initio predicted profiles.* Journal of Applied Crystallography 43.

Schrodinger, LLC (2015). The PyMOL Molecular Graphics System, Version 1.8.

Schumacher, M. A., Allen, G. S., Diel, M., Seidel, G., Hillen, W. & Brennan, R. G. (2004). *Structural basis for allosteric control of the transcription regulator CcpA by the phosphoprotein HPr-Ser46-P.* Cell 118, 731-741.

Sevvana, M., Ruf, M., Usón, I., Sheldrick, G. M. & Herbst-Irmer, R. (2019). *Non-merohedral twinning: from minerals to proteins.* Acta Crystallogr D Struct Biol 75, 1040-1050.

Shakked, Z. (1983). *Anisotropic scaling of three-dimensional intensity data.* Acta Crystallographica Section A 39, 278-279.

Sheldrick, G. M. (2002a). *TWINABS.* University of Göttingen.

Sheldrick, G. M. (2002b). *Macromolecular phasing with SHELXE.* Z. Kristallogr. 217, 644-650.

Sheldrick, G. M. (2008a). *XPREP.* Version 2008/2.

Sheldrick, G. M. (2008b). *A short history of SHELX.* Acta Crystallogr A 64, 112-122.

Sheldrick, G. M. (2010). *Experimental phasing with SHELXC/D/E: combining chain tracing with density modification.* Acta Crystallogr D Biol Crystallogr 66, 479-485.

Sheldrick, G. M. (2015). *Crystal structure refinement with SHELXL.* Acta Crystallogr C Struct Chem 71, 3-8.

Sheldrick, G. M., Gilmore, C. J., Hauptman, H. A., Weeks, C. M., Miller, R. & Usón, I. (2012). *International Tables for Crystallography, Vol. F*, 2nd online ed. Chester: International Union of Crystallography.

Sheldrick, G. M. & Schneider, T. R. (1997). *SHELXL: high-resolution refinement.* Methods Enzymol 277, 319-343.

Sheriff, S. & Hendrickson, W. A. (1987). *Description of overall anisotropy in diffraction from macromolecular crystals.* Acta Crystallographica Section A 43, 118-121.

Sliwiak, J., Dauter, Z., Kowiel, M., McCoy, A. J., Read, R. J. & Jaskolski, M. (2015). *ANS complex of St John's wort PR-10 protein with 28 copies in the asymmetric unit: a fiendish combination of pseudosymmetry with tetartohedral twinning.* Acta Crystallogr D Biol Crystallogr 71, 829-843.

Sliwiak, J., Jaskolski, M., Dauter, Z., McCoy, A. J. & Read, R. J. (2014). *Likelihood-based molecular-replacement solution for a highly pathological crystal with tetartohedral twinning and sevenfold translational noncrystallographic symmetry.* Acta Crystallogr D Biol Crystallogr 70, 471-480.

Spencer, J., Murphy, L. M., Conners, R., Sessions, R. B. & Gamblin, S. J. (2010). *Crystal structure of the LasA virulence factor from Pseudomonas aeruginosa: substrate specificity and mechanism of M23 metallopeptidases.* J Mol Biol 396, 908-923.

Stanley, E. (1972). *The identification of twins from intensity statistics.* Journal of Applied Crystallography 5, 191-194.

Stewart, J. M. & Karle, J. (1976). *The calculation of [epsilon] associated with normalized structure factors, E.* Acta Crystallographica Section A 32, 1005-1007.

Strong, M., Sawaya, M. R., Wang, S., Phillips, M., Cascio, D. & Eisenberg, D. (2006). *Toward the structural genomics of complexes: crystal structure of a PE/PPE protein complex from Mycobacterium tuberculosis.* Proc Natl Acad Sci U S A 103, 8060-8065.

Tanaka, S., Kerfeld, C. A., Sawaya, M. R., Cai, F., Heinhorst, S., Cannon, G. C. & Yeates, T. O. (2008). *Atomic-level models of the bacterial carboxysome shell.* Science 319, 1083-1086.

Tannenbaum, T., Wright, D., Miller, K. & Livny, M. (2001). *Beowulf Cluster Computing with Linux*, edited by T. Sterling, pp. 307-350: MIT Press.

Taylor, E. J., Gloster, T. M., Turkenburg, J. P., Vincent, F., Brzozowski, A. M., Dupont, C., Shareck, F., Centeno, M. S., Prates, J. A., Puchart, V., Ferreira, L. M., Fontes, C. M., Biely, P. & Davies, G. J. (2006). *Structure and activity of two metal ion-dependent acetylxylan esterases involved in plant cell wall degradation reveals a close similarity to peptidoglycan deacetylases.* J Biol Chem 281, 10968-10975.

Taylor, K. C., Buvoli, M., Korkmaz, E. N., Buvoli, A., Zheng, Y., Heinze, N. T., Cui, Q., Leinwand, L. A. & Rayment, I. (2015). *Skip residues modulate the structural properties of the myosin rod and guide thick filament assembly.* Proc Natl Acad Sci U S A 112, E3806-3815.

Terwilliger, T. C., Grosse-Kunstleve, R. W., Afonine, P. V., Moriarty, N. W., Zwart, P. H., Hung, L. W., Read, R. J. & Adams, P. D. (2008). *Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard.* Acta Crystallogr D Biol Crystallogr 64, 61-69.

Thomas, J. M., Keegan, R. M., Bibby, J., Winn, M. D., Mayans, O. & Rigden, D. J. (2015). *Routine phasing of coiled-coil protein crystal structures with AMPLE.* IUCrJ 2, 198-206.

Thomas, J. M. H., Keegan, R. M., Rigden, D. J. & Davies, O. R. (2020). *Extending the scope of coiled-coil crystal structure solution by AMPLE through improved ab initio modelling.* Acta Crystallographica Section D 76, 272-284.

Thompson, M. C. (2017). *Identifying and Overcoming Crystal Pathologies: Disorder and Twinning.* Methods Mol Biol 1607, 185-217.

Thorn, A., Dittrich, B. & Sheldrick, G. M. (2012). *Enhanced rigid-bond restraints.* Acta Crystallographica Section A 68, 448-451.

Thorn, A. & Sheldrick, G. M. (2013). *Extending molecular-replacement solutions with SHELXE.* Acta Crystallogr D Biol Crystallogr 69, 2251-2256.

Tickle, I. J., Flensburg, C., Keller, P., Paciorek, W., Sharff, A., Vonrhein, C. & Bricogne, G. (2018). *STARANISO.* Global Phasing Ltd., Cambridge, UK.

Trame, C. B. & McKay, D. B. (2001). *Structure of Haemophilus influenzae HslU protein in crystals with one-dimensional disorder twinning.* Acta Crystallogr D Biol Crystallogr 57, 1079-1090.

Trueblood, K. N., Burgi, H.-B., Burzlaff, H., Dunitz, J. D., Gramaccioli, C. M., Schulz, H. H., Shmueli, U. & Abrahams, S. C. (1996). *Atomic Dispacement Parameter Nomenclature. Report of a Subcommittee on Atomic Displacement Parameter Nomenclature.* Acta Crystallographica Section A 52, 770-781.

Tsai, Y., Sawaya, M. R. & Yeates, T. O. (2009). *Analysis of lattice-translocation disorder in the layered hexagonal structure of carboxysome shell protein CsoS1C.* Acta Crystallogr D Biol Crystallogr 65, 980-988.

Uppenberg, J., Ohrner, N., Norin, M., Hult, K., Kleywegt, G. J., Patkar, S., Waagen, V., Anthonsen, T. & Jones, T. A. (1995). *Crystallographic and molecular-modeling studies of lipase B from Candida antarctica reveal a stereospecificity pocket for secondary alcohols.* Biochemistry 34, 16838-16851.

Urzhumtsev, A., Urzhumtseva, L. & Baumann, U. (2016). *Helical Symmetry of Nucleic Acids: Obstacle or Help in Structure Solution?* New York: Methods in Molecular Biology.

Usón, I., Pohl, E., Schneider, T. R., Dauter, Z., Schmidt, A., Fritz, H.-J. & Sheldrick, G. M. (1999). *1.7 A structure of the stabilized REIv mutant T39K. Application of local NCS restraints.* Acta Crystallographica Section D 55, 1158-1167.

Usón, I. & Sheldrick, G. M. (2018). *An introduction to experimental phasing of macromolecules illustrated by SHELX; new autotracing features.* Acta Crystallogr D Struct Biol 74, 106-116.

Usón, I., Stevenson, C., Lawson, D. & Sheldrick, G. (2007). *Structure determination of the O-methyltransferase NovP using the 'free lunch algorithm' as implemented in SHELXE.* Acta crystallographica. Section D, Biological crystallography 63, 1069-1074.

van Smaalen, S. (2004). *An elementary introduction to superspace crystallography.* Zeitschrift Fur Kristallographie - Z KRISTALLOGR 219, 681-691.

van Smaalen, S. (2007). *Incommensurate Crystallography.* Oxford: Oxford University Press.

Wagner, T. & Schönleber, A. (2009). *A non-mathematical introduction to the superspace description of modulated structures.* Acta crystallographica. Section B, Structural science 65, 249-268.

Wang, J., Kamtekar, S., Berman, A. J. & Steitz, T. A. (2005). *Correction of X-ray intensities from single crystals containing lattice-translocation defects.* Acta Crystallogr D Biol Crystallogr 61, 67-74.

Wang, X. & Janin, J. (1993). *Orientation of non-crystallographic symmetry axes in protein crystals.* Acta Crystallographica Section D 49, 505-512.

West, A. M. V., Rosenberg, S. C., Ur, S. N., Lehmer, M. K., Ye, Q., Hagemann, G., Caballero, I., Usón, I., Herzog, F. & Corbett, K. D. (2018). *A conserved mechanism for meiotic chromosome organization through self-assembly of a filamentous chromosome axis core.* bioRxiv, 375220.

West, A. M. V., Rosenberg, S. C., Ur, S. N., Lehmer, M. K., Ye, Q., Hagemann, G., Caballero, I., Usón, I., MacQueen, A. J., Herzog, F. & Corbett, K. D. (2019). *A conserved filamentous assembly underlies the structure of the meiotic chromosome axis.* eLife 8, e40372.

Wilson, A. J. C. (1949). *The probability distribution of X-ray intensities.* Acta Crystallographica 2, 318-321.

Winn, M. D., Ballard, C. C., Cowtan, K. D., Dodson, E. J., Emsley, P., Evans, P. R., Keegan, R. M., Krissinel, E. B., Leslie, A. G., McCoy, A., McNicholas, S. J., Murshudov, G. N., Pannu, N. S., Potterton, E. A., Powell, H. R., Read, R. J., Vagin, A. & Wilson, K. S. (2011). *Overview of the CCP4 suite and current developments.* Acta Crystallogr D Biol Crystallogr 67, 235-242.

Wolff, P. M., Janssen, T. & Janner, A. (1981). *The superspace groups for incommensurate crystal structures with a one-dimensional modulation.* Acta Crystallographica Section A 37, 625-636.

Woolfson, M. (1987). *Direct methods - from birth to maturity.* Acta Crystallographica Section A 43, 593-612.

Wukovitz, S. W. & Yeates, T. O. (1995). *Why protein crystals favour some space-groups over others.* Nat Struct Biol 2, 1062-1067.

Yang, F., Dauter, Z. & Wlodawer, A. (2000). *Effects of crystal twinning on the ability to solve a macromolecular structure using multiwavelength anomalous diffraction.* Acta Crystallogr D Biol Crystallogr 56, 959-964.

Yang, X., Jin, H., Cai, X., Li, S. & Shen, Y. (2012). *Structural and mechanistic insights into the activation of Stromal interaction molecule 1 (STIM1).* Proceedings of the National Academy of Sciences 109, 5657-5662.

Yeates, T. (1988). *Simple statistics for intensity data from twinned specimens.* Acta Crystallographica Section A 44, 142-144.

Yeates, T. O. (1997). *Detecting and overcoming crystal twinning.* Methods Enzymol 276, 344-358.

Yeates, T. O. & Fam, B. C. (1999). *Protein crystals and their evil twins.* Structure 7, R25-29.

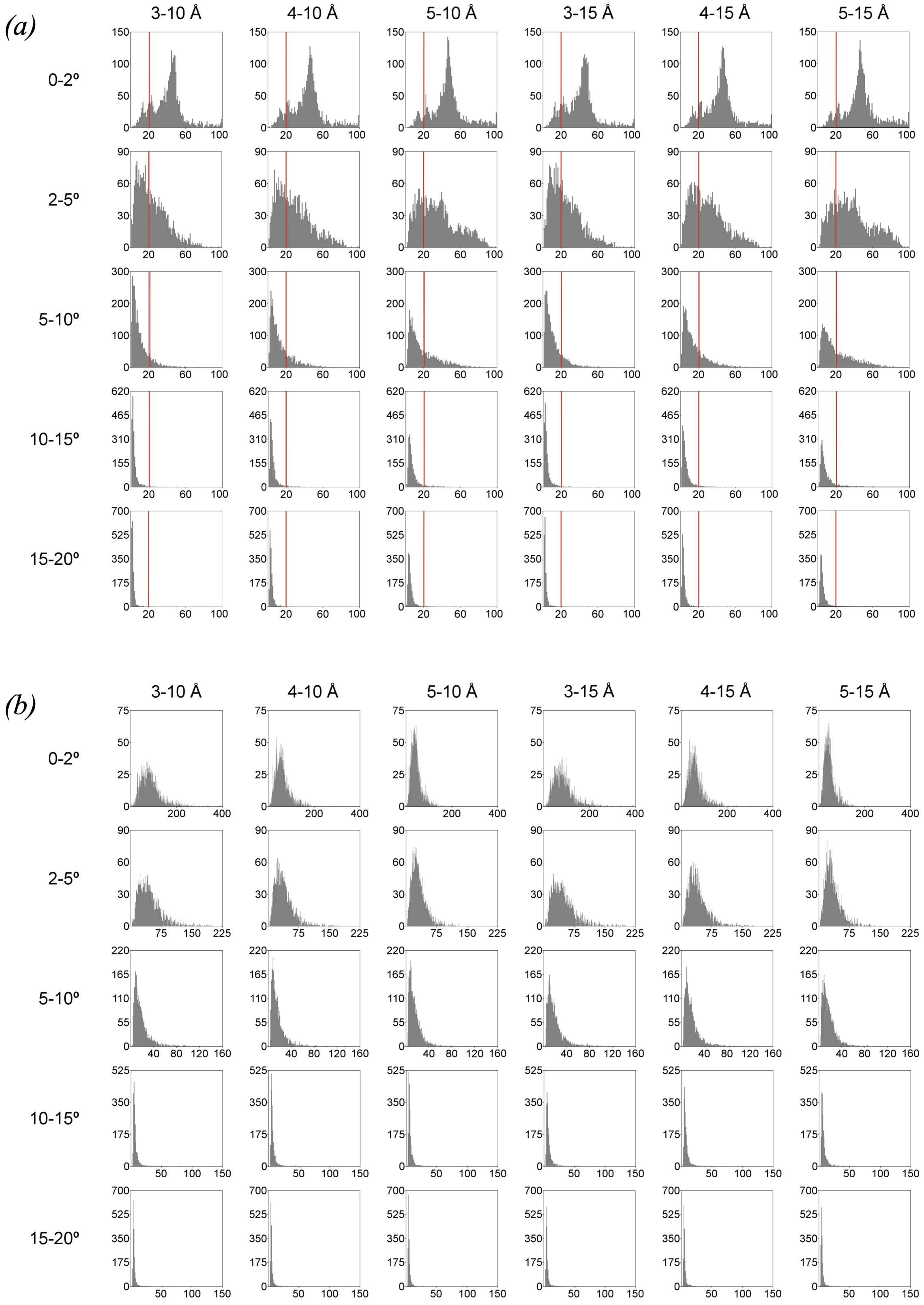
Yeates, T. O. & Rees, D. C. (1987). *An isomorphous replacement method for phasing twinned structures.* Acta Crystallographica Section A 43, 30-36.

Zhu, X., Xu, X. & Wilson, I. A. (2008). *Structure determination of the 1918 H1N1 neuraminidase from a crystal with lattice-translocation defects.* Acta Crystallogr D Biol Crystallogr D64, 843-850.

Zwart, P. H., Grosse-Kunstleve, R. W. & Adams, P. D. (2005). *Xtriage and Fest: automatic assessment of X-ray data and substructure structure factor estimation.* CCP4 Newsletter 42, contribution 10.

Zwart, P. H., Grosse-Kunstleve, R. W., Lebedev, A. A., Murshudov, G. N. & Adams, P. D. (2008). *Surprises and pitfalls arising from (pseudo)symmetry.* Acta Crystallogr D Biol Crystallogr 64, 99-107.

APPENDIX


(a)
3-10 Å
4-10 Å
5-10 Å
3-15 Å
4-15 Å
5-15 Å
0-2º
2-5º
5-10º
10-15º
15-20º
150
100
50
0
90
60
30
300
200
620
465
310
155
700
525
350
175
20
60
100
(b)
3-10 Å
4-10 Å
5-10 Å
3-15 Å
4-15 Å
5-15 Å
0-2º
2-5º
5-10º
10-15º
15-20º
75
50
25
0
200
400
90
60
30
75
150
225
220
165
110
55
40
80
120
160
525
350
175
50
100
150
700

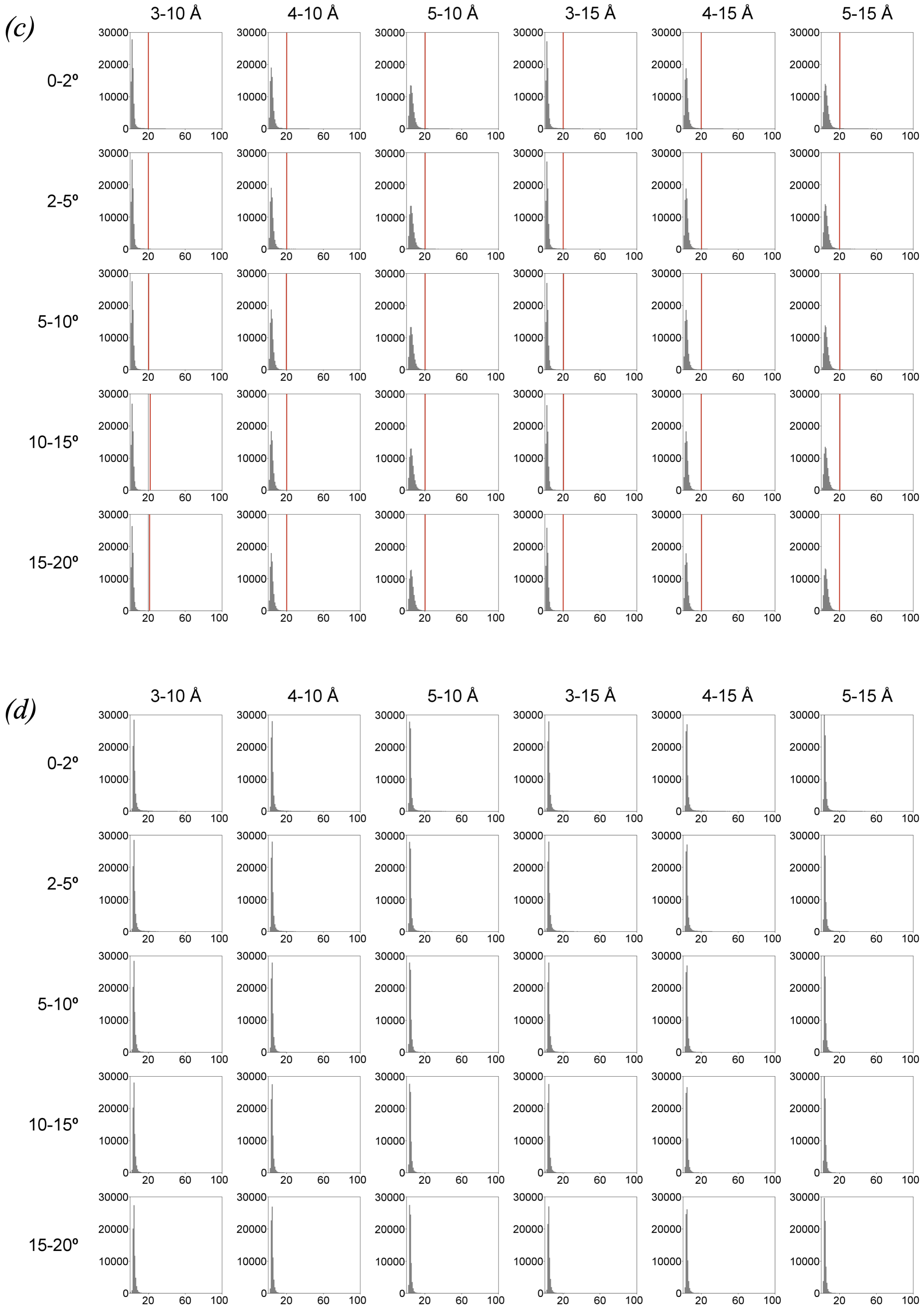


Figure A1. Non-cumulative histograms of the number of structures with the highest non-origin peak value, depending on the rotational tolerances and the resolution used for calculating the Patterson map. a) Patterson peak percentages for structures with tNCS, with a red line drawn at 20%. b) Patterson peak Z-scores for structures with tNCS. c) Patterson peak percentages for structures without tNCS. d) Patterson peak Z-scores for structures without tNCS.

Table A1. Structures where *Phaser* has found a significant Patterson peak. The third column displays whether a structure presents tNCS matching the vector derived in *Phaser*.

PDB	Chains & helices	tNCS	Patterson %	Distance	Vector	tNCS corrected		tNCS not corrected	
						CC	wMPE	CC	wMPE
1byz	4C 4H	yes	30.5	16.3	0.5, 0, 0.5	50.8	15.6	48.9	14.0
1g1j	2C 2H	yes	28.3	33.9	0.5, 0.5, 0.375	33.5	57.1	31.0	87.6
1kyc	1C 1H	no	21.5	20.0	0.33, 0.67, 0.08	15.5	91.8	40.5	52.6
1nkd	1C 2H	no	23.2	17.5	0.1667, 0, 0.55	9.2	88.5	63.4	17.9
1p9i	1C 1H	no	20.1	21.1	0, 0.1528, 0.5	14.2	89.0	65.5	23.7
1x8y	1C 1H	no	25.2	37.8	0, 0, 0.5	Packing excludes all		53.2	32.4
1yod	2C 2H	no	21.1	65.9	0.33, 0.67, 0.5	54.7	24.2	55.5	24.0
2b22	1C 1H	no	25.7	19.9	0, 0, 0.3750	Packing excludes all		51.7	49.8
2bez	2C 2H	no	24.5	37.8	0.33, 0.67, 0.07	Packing excludes all		59.0	24.5
2ic6	2C 4H	yes	29.6	29.3	0, 0.5, 0.1562	61.3	19.4	61.3	19.3
2wpq	3C 3H	no	37.2	89.2	0, 0, 0.5	35.0	56.4	35.4	57.7
3bas	2C 2H	no	50.7	19.9	0, 0, 0.5	26.7	88.2	49.0	45.8
3hfe	3C 3H	no	21.2	21.9	0.58, 0, 0.625	13.4	87.9	52.9	41.5
3k9a	1C 2H	no	30.3	77.5	0, 0, 0.5	Packing excludes all		62.4	29.2
3m91	4C 4H	no	21.8	36.3	0.367, 0.5, 0.73	18.3	88.2	55.0	30.1
3p7k	1C 1H	no	30.7	52.2	0, 0, 0.3438	Packing excludes all		53.0	33.6
3v86	1C 1H	no	24.9	26.0	0.33, 0.67, 0.6	43.5	82.6	49.4	51.7
3vgy	2C 2H	no	26.4	28.1	0.33, 0.67, 0.15	Packing excludes all		58.2	38.4
3mqc	4C 4H	no	21.4	29.5	0, 0.5, 0	Unsolved			
2ahp	2C 2H	no	22.1	17.7	0.067, 0.5, 0.75	Packing excludes all		54.9	44.0
3efg	1C 1H	no	23.6	40.9	0, 0, 0.5	57.3	41.7	58.3	42.3
3r3k	3C 3H	no	44.3	31.5	0.5, 0.5, 0	Packing excludes all		34.7	48.6
5c9n	2C 2H	no	46.5	60.0	0.5, 0.5, 0.5	43.4	31.6	51.2	29.0
1unx	2C 2H	yes	29.9	68.3	0.5, 0.5, 0.5	46.3	39.5	47.1	38.6
2wz7	6C 6H	no	22.0	51.2	0, 0, 0.2153	28.0	89.1	41.8	57.1
1w5h	2C 2H	yes	64.4	52.3	0, 0, 0.5	40.4	49.6	53.1	44.5
2o1j	4C 4H	yes	61.4	34.4	0.5, 0, 0.5	46.2	62.6	36.1	90.7
3v2r	5C 5H	no	22.0	29.0	0.5, 0, 0.5	34.1	84.8	38.0	62.8
3nwh	4C 4H	yes	28.4	80.9	0.25, 0.5, 0.53	48.1	57.9	35.2	88.6
3iv1	8C 8H	no	38.4	28.1	0, 0.5, 0	Unsolved			
3tul	4C 16H	no	36.7	79.7	0, 0.5, 0.0938	Unsolved			
4pna	7C 7H	no	33.8	31.1	0.5, 0.0125, 0.5	Unsolved			

Table A2. Characteristics and results for the coiled-coil test set.

PDB	Test set	Space group	Resolution	Nres	Solvent content	Anisotropic delta B-factor	ASU content (C=chain, H=helix)	Architecture	Twin fraction	tNCS	CC (%)	Nres traced	wMPE (º)	Search fragments NHM (N=number of helices, M=helix length) & parameters
1byz	**1**	P1	0.90	48	25.5	0.96	4C of 1H	Left-handed	-	real	48.93	52	14.0	2H10 - TNCS false
3azd	**1**	$P3_1$	0.98	60	45.1	0.27	2C of 1H	Left-handed	-	no	-	-	-	-
1nkd	**1**	C2	1.09	59	40.6	6.88	1C of 2H	Left-handed	-	apparent	63.35	59	17.9	2H12 - TNCS false
2ic6	**1**	$P2_12_12_1$	1.15	147	41.5	6.86	2C of 2H	Left-handed	-	real	61.27	149	19.3	4H15 - TNCS false
1p9i	**1**	$C222_1$	1.17	29	32.7	3.50	1C of 1H	Left-handed	-	apparent	65.46	30	23.7	2H10 - TNCS false
2akf	**1**	P1	1.20	96	36.3	7.28	3C of 1H	Left-handed	-	no	32.07	85	48.0	4H18
1jcd	**1**	P1	1.30	152	34.1	5.71	3C of 1H	Left-handed	-	no	60.04	130	42.6	4H18
3ljm	**1**	C2	1.36	87	45.8	0.77	3C of 1H	Left-handed	-	no	49.36	91	27.8	4H18
3twe	**1**	$I4_1$	1.36	51	36.7	3.08	2C of 1H	Left-handed	-	no	56.45	49	35.1	4H18
2w6a	**1**	$P2_1$	1.40	117	42.0	9.47	2C of 1H	Left-handed	-	no	60.51	114	31.7	4H18
1kyc	**1**	H32	1.45	15	28.2	0.06	1C of 1H	Left-handed	-	apparent	40.54	15	52.6	1H6 - TNCS false
1n7s	**1**	$P2_12_12_1$	1.45	276	43.6	13.67	4C of 1H	Left-handed	-	no	53.26	259	25.7	4H18
1ybk	**1**	$P3_121$	1.45	205	76.9	1.73	4C of 1H	Right-handed	-	no	45.13	186	40.1	4H18
1t6f	**1**	$P2_12_12_1$	1.47	74	41.2	3.49	2C of 1H	Left-handed	-	no	53.32	75	39.5	4H18
2q5u	**1**	I222	1.50	135	44.3	9.60	3C of 1H	Left-handed	-	no	43.43	105	53.2	4H18
3pp5	**1**	$P6_3$	1.50	63	63.9	6.72	1C of 1H	Left-handed	-	no	50.85	60	22.4	4H18
4dzn	**1**	$P2_12_12_1$	1.59	96	42.2	8.58	3C of 1H	Left-handed	-	no	51.60	94	35.6	4H18
1wt6	**1**	$P2_12_12_1$	1.60	195	56.2	15.47	3C of 1H	Left-handed	-	no	53.54	184	29.0	4H18
2bez	**1**	H32	1.60	119	49.2	5.07	2C of 1H	Left-handed	-	apparent	58.96	92	24.5	2H25 - TNCS false
3ni0	**1**	$I2_12_12_1$	1.60	182	45.6	18.78	2C of 1H	Left-handed	-	no	37.06	122	40.1	4H18
1y66	**1**	$P2_12_12_1$	1.65	186	48.1	0.93	4C of 1H	Left-handed	-	no	48.51	170	28.2	4H30
1usd	**1**	I4	1.70	41	43.9	16.29	1C of 1H	Right-handed	-	no	54.38	41	37.2	4H18
1zv7	**1**	C222	1.70	71	53.0	1.71	2C of 1H	Left-handed	-	no	57.76	66	30.2	2H30
1zvb	**1**	C2	1.70	101	48.0	2.63	3C of 1H	Left-handed	-	no	46.99	85	51.2	4H18
3hfe	**1**	C2	1.70	79	40.7	12.64	3C of 1H	Left-handed	-	apparent	52.87	77	41.5	3H20 - TNCS false
3swk	**1**	$P2_1$	1.70	171	44.2	20.35	2C of 1H	Left-handed	-	no	40.19	129	45.7	4H18
1k33	**1**	H3	1.75	62	43.7	6.04	1C of 2H	Left-handed	-	no	53.99	53	37.8	4H18
3cve	**1**	$P2_1$	1.75	261	44.6	3.04	4C of 1H	Left-handed	-	no	50.11	247	48.6	4H18 - sliding autotracing
4dzk	**1**	P321	1.79	29	57.1	9.52	1C of 1H	Left-handed	0.429	no	42.70	25	49.2	4H18
1uix	**1**	C2	1.80	137	50.3	4.29	2C of 1H	Left-handed	-	no	52.32	126	36.7	4H18
1yod	**1**	$P6_122$	1.80	57	54.4	3.05	2C of 1H	Left-handed	-	apparent	55.46	62	24.0	4H18 - TNCS false
2efr	**1**	P1	1.80	618	50.6	20.38	4C of 1H	Left-handed	-	no	29.99	382	57.1	7H30
3m91	**1**	$P2_1$	1.80	152	64.1	26.39	4C of 1H	Left-handed	-	apparent	55.01	132	30.1	4H20 - TNCS false
3u1c	**1**	$P4_32_12$	1.80	202	51.7	24.63	2C of 1H	Left-handed	-	no	27.85	136	56.1	4H30
3vp9	**1**	C2	1.80	147	63.2	23.92	2C of 1H	Left-handed	-	no	42.08	82	53.1	4H18
2wpq	**1**	$P2_1$	1.85	297	50.2	11.90	3C of 1H	Left-handed	-	apparent	35.37	205	57.7	6H30 - TNCS false

1g1j	**1**	$P2_12_12$	1.86	86	29.8	13.29	2C of 1H	Left-handed	-	real	33.54	55	57.1	2H30 - TNCS true
1ezj	**1**	$P42_12$	1.90	114	46.4	5.34	1C of 4H	Left-handed	-	no	49.55	85	44.4	1H30
2qih	**1**	P3	1.90	273	59.8	4.59	2C of 1H	Left-handed	-	no	41.51	204	50.6	4H18
3hrn	**1**	H32	1.90	63	50.7	10.41	1C of 1H	Left-handed	-	no	51.28	62	43.9	4H18
3q8t	**1**	C2	1.90	189	57.3	5.40	2C of 1H	Left-handed	-	no	55.14	174	36.0	4H18
3swy	**1**	$P2_1$	1.90	138	39.3	12.68	3C of 1H	Left-handed	-	no	53.10	120	44.9	6H15
2xus	**1**	H32	1.91	81	46.8	0.02	2C of 1H	Left-handed	-	no	48.42	61	42.0	4H18
2q6q	**1**	$P2_12_12_1$	1.97	127	46.7	0.70	2C of 1H	Left-handed	-	no	53.36	115	42.7	4H18
1m5i	**1**	$P3_221$	2.00	105	55.1	15.57	1C of 3H	Left-handed	-	no	51.05	103	37.2	4H18
1uii	**1**	$P2_12_12_1$	2.00	122	80.0	14.81	2C of 1H	Left-handed	-	no	53.36	127	32.9	4H25
2b22	**1**	$I4_122$	2.00	29	38.9	9.35	1C of 1H	Left-handed	-	apparent	51.70	23	49.8	2H10 - TNCS false
2ic9	**1**	C2	2.00	145	55.3	12.96	2C of 2H	Left-handed	-	no	40.97	108	51.1	4H18
3a2a	**1**	$P4_3$	2.00	164	48.6	2.31	4C of 1H	Left-handed	-	no	57.30	153	37.5	4H18
3etw	**1**	$P6_1$	2.00	109	76.4	28.78	1C of 2H	Left-handed	-	no	49.48	106	34.9	4H18
3k29	**1**	C2	2.00	162	50.1	12.62	1C of 2H	Left-handed	-	no	51.77	149	41.9	2H30
3u1a	**1**	$P2_12_12_1$	2.00	330	50.8	28.65	4C of 1H	Left-handed	-	no	55.24	306	37.5	4H35
4pn8	**2**	$P2_1$	2.00	299	50.8	1.03	10C of 1H	Left-handed	-	no	44.84	214	38.8	4H18
3m9h	**2**	F222	2.00	258	43.7	2.07	6C of 1H	Left-handed	-	no	52.09	172	38.3	4H18
1urq	**2**	C2	2.00	250	54.8	7.99	4C of 1H	Left-handed	-	no	37.89	187	45.7	4H35
2ahp	**2**	C2	2.00	65	42.0	7.53	2C of 1H	Left-handed	-	apparent	54.87	58	44.0	4H18 - TNCS false
3efg	**2**	$P6_3$	2.00	51	52.5	9.45	1C of 1H	Left-handed	-	apparent	58.25	44	42.3	4H18 - TNCS false
3g9r	**2**	$P2_1$	2.00	244	52.6	21.62	6C of 1H	Partially right-handed	-	no	35.85	168	52.1	4H18
4n6j	**2**	$P3_221$	2.00	96	47.2	7.43	2C of 1H	Left-handed	-	no	47.78	92	45.4	4H18
4yv3	**2**	$P2_1$	2.00	215	44.2	23.83	3C of 1H	Left-handed	-	no	44.86	170	45.5	4H18
2oqq	**2**	C2	2.00	84	46.0	11.42	2C of 1H	Left-handed	-	no	49.83	64	37.0	4H18
1s9z	**1**	$P6_3$	2.01	16	33.2	16.67	1C of 1H	Left-handed	0.12	no	55.36	14	41.9	4H18
3qh9	**1**	$C222_1$	2.01	66	53.0	1.67	1C of 1H	Left-handed	-	no	62.10	58	43.9	4H18
3vgy	**1**	P321	2.03	79	44.9	8.27	2C of 1H	Left-handed	-	apparent	58.22	68	38.4	1H30 + 1H15 - TNCS false
3okq	**1**	$C222_1$	2.04	125	63.8	20.84	1C of 2H	Left-handed	-	no	60.52	125	28.0	4H18
4pxj	**1**	C2	2.06	174	77.2	31.44	3C of 1H	Left-handed	-	no	55.25	175	22.5	4H18
2ovc	**1**	I4	2.07	30	42.2	7.68	1C of 1H	Left-handed	-	no	64.46	25	41.1	4H18
4m3l	**2**	$P2_1$	2.10	228	47.0	12.62	4C of 1H	Left-handed	-	no	37.04	132	45.5	4H18
4pna	**2**	$P22_12_1$	2.10	208	53.0	12.53	7C of 1H	Left-handed	-	apparent	-	-	-	-
3cyo	**2**	$P2_13$	2.10	73	38.5	0.00	1C of 2H	Left-handed	-	no	36.07	49	43.6	4H18
2pnv	**1**	$P6_3$	2.10	78	52.6	13.21	2C of 1H	Left-handed	0.278	no	54.22	72	63.2	4H18
3ajw	**1**	$P6_522$	2.10	134	50.0	12.39	1C of 2H	Left-handed	-	no	56.01	111	32.4	4H18
3k9a	**1**	$P6_322$	2.10	82	59.1	8.02	1C of 2H	Left-handed	-	apparent	62.35	74	29.2	2H10 - TNCS false
3s9g	**1**	I4	2.10	158	58.2	10.38	2C of 2H	Left-handed	-	no	41.64	127	40.8	2H35
2ykt	**1**	$C222_1$	2.11	237	56.2	5.83	2C of 3H	Left-handed	-	no	51.19	205	33.5	5H20

PDB	Test set	Space group	Resolution	Nres	Solvent content	Anisotropic delta B-factor	ASU content (C=chain, H=helix)	Architecture	Twin fraction	tNCS	CC (%)	Nres traced	wMPE (º)	Search fragments NHM (N=number of helices, M=helix length) & parameters
5eoj	**2**	$P4_322$	2.12	87	53.5	17.20	3C of 1H	Left-handed	-	no	49.15	70	35.8	4H18
3swf	**1**	$P2_12_12_1$	2.14	165	64.9	11.91	3C of 1H	Left-handed	-	no	56.22	160	31.2	4H18
1kdd	**2**	$P4_12_12$	2.14	206	59.0	10.94	6C of 1H	Left-handed	-	no	47.25	155	30.6	4H18
4bl6	**2**	$P6_1$	2.18	326	55.6	19.01	4C of 1H	Left-handed	0.477	no	31.86	225	50.7	4H25 - pack_tra true
4pn9	**2**	$C222_1$	2.20	177	54.3	57.59	6C of 1H	Left-handed	-	no	47.55	147	37.8	4H18
3r3k	**2**	I222	2.20	90	58.1	4.20	3C of 1H	Left-handed	-	apparent	34.65	69	48.6	3H25 - TNCS false
5c9n	**2**	$P2_12_12_1$	2.20	125	75.3	20.77	2C of 1H	Left-handed	-	apparent	51.19	119	29.0	4H18 - TNCS false
1gmj	**1**	$P2_1$	2.20	240	73.3	61.81	4C of 1H	Left-handed	-	no	33.21	139	52.9	1H10 + 1H25 - sliding autotracing
1x8y	**1**	$P6_522$	2.20	74	74.8	24.52	1C of 1H	Left-handed	-	apparent	53.16	65	32.4	2H30 - TNCS false
2zzo	**1**	P321	2.20	70	47.3	9.18	2C of 1H	Left-handed	-	no	57.77	73	43.7	4H18
3h00	**1**	C2	2.20	152	73.3	24.86	4C of 1H	Left-handed	0.47	no	50.39	138	44.7	4H18
5cx2	**2**	C2	2.21	203	48.7	8.87	4C of 1H	Left-handed	-	no	32.18	120	46.9	4H18
2v71	**1**	C2	2.24	320	62.7	11.17	2C of 1H	Left-handed	-	no	40.19	239	50.5	4H40 - pack_tra true
4oh8	**2**	$P2_1$	2.28	89	51.8	10.75	2C of 1H	Left-handed	-	no	44.38	74	38.6	4H18
2b9c	**2**	$P6_5$	2.30	278	62.5	1.20	2C of 1H	Left-handed	-	no	41.37	151	46.1	4H18
5ajs	**2**	$P4_3$	2.30	250	53.4	36.93	4C of 1H	Left-handed	0.492	no	43.66	160	55.2	4H25
4hu6	**2**	P1	2.30	126	54.9	25.92	4C of 1H	Left-handed	-	no	48.60	93	41.8	4H18
3bas	**1**	C2	2.30	167	55.3	37.89	2C of 1H	Left-handed	-	apparent	48.98	121	45.8	4H30 - TNCS false
3p7k	**1**	$P6_322$	2.30	45	64.9	5.93	1C of 1H	Left-handed	-	apparent	52.95	45	33.6	2H15 - TNCS false
3trt	**1**	I222	2.30	148	53.0	30.09	2C of 1H	Left-handed	-	no	48.15	117	41.1	1H50
3ra3	**1**	$P3_2$	2.31	106	62.5	11.42	4C of 1H	Left-handed	0.181	no	48.42	105	36.1	4H18
1unx	**2**	$P4_132$	2.40	64	51.8	0.00	2C of 1H	Left-handed	-	real	47.08	52	38.6	4H18 - TNCS false
1deb	**1**	$P6_522$	2.40	107	63.9	13.26	2C of 1H	Left-handed	-	no	58.48	101	35.3	4H18
1s35	**1**	$P4_12_12$	2.40	211	68.5	23.11	1C of 5H	Left-handed	-	no	37.13	177	42.7	1H30 + 2H18
2xv5	**1**	$P2_12_12_1$	2.40	114	57.4	10.21	2C of 1H	Partially right-handed	-	no	57.70	89	37.0	4H18
3tyy	**1**	$P3_221$	2.40	146	65.4	4.69	2C of 1H	Left-handed	-	no	55.94	136	36.9	4H18
3s0r	**1**	$I4_1$	2.45	60	67.3	2.90	2C of 1H	Left-handed	0.499	no	68.26	60	48.1	4H18
3s4r	**1**	H32	2.45	179	65.8	54.67	2C of 1H	Left-handed	-	no	-	-	-	-
4e61	**1**	$P2_1$	2.45	175	59.0	3.63	2C of 2H	Left-handed	-	no	50.05	119	44.7	4H18
2wz7	**2**	$P2_12_12_1$	2.48	394	42.7	11.22	6C of 1H	Left-handed	-	apparent	41.81	218	57.1	4H18 - TNCS false
5d3a	**2**	$P3_121$	2.49	164	66.4	12.53	2C of 1H	Left-handed	-	no	48.69	141	39.1	4H18
4l2w	**2**	C2	2.49	269	72.7	2.47	4C of 1H	Left-handed	-	no	30.55	170	58.5	4H30
3miw	**2**	$P4_2$	2.50	432	45.6	49.63	10C of 1H	Left-handed	0.499	no	37.75	301	59.7	10H30 - pack_tra - helices reverted
5jxc	**2**	$P2_1$	2.50	484	48.3	31.34	6C of 1H	Left-handed	-	no	35.82	277	50.1	12H18 - pack_tra
3r47	**2**	$P4_2$	2.50	340	56.7	12.88	12C of 1H	Left-handed	-	no	46.07	252	46.3	6H25 - pack_tra
1w5h	**2**	$P4_22_12$	2.50	62	43.1	3.57	2C of 1H	Left-handed	-	real	53.11	53	44.5	4H18 - TNCS false

3a7o	**2**	$P4_32_12$	2.50	342	69.1	15.29	6C of 1H	Left-handed	-	no	37.97	272	36.8	6H30
1u4q	**2**	C2	2.50	635	52.0	31.66	2C of 7H	Left-handed	-	no	-	-	-	-
3iv1	**2**	$P3_221$	2.50	624	43.1	10.95	8C of 1H	Left-handed	-	apparent	-	-	-	-
4w7y	**2**	$P6_5$	2.50	128	70.7	4.76	2C of 1H	Left-handed	-	no	49.03	107	32.9	4H18
2nps	**2**	C2	2.50	273	48.4	24.81	4C of 1H	Left-handed	-	no	38.88	127	59.9	4H40
1t3j	**2**	$F4_132$	2.50	62	79.9	0.00	1C of 1H	Left-handed	-	no	49.68	67	28.8	4H18
1pl5	**2**	$P6_5$	2.50	151	83.7	17.63	2C of 1H	Left-handed	-	no	46.75	144	30.6	4H18
1mi7	**1**	$P6_122$	2.50	103	75.4	9.30	1C of 4H	Left-handed	-	no	45.15	91	38.4	4H18
3h7z	**1**	$P6_3$	2.51	61	60.8	20.90	1C of 1H	Partially right-handed	0.457	no	51.53	56	51.4	4H18
4xa3	**2**	$C222_1$	2.55	284	61.2	3.25	2C of 3H	Left-handed	-	no	-	-	-	-
4ltb	**2**	$P2_12_12_1$	2.59	342	55.8	53.42	2C of 3H	Partially right-handed	-	no	32.19	183	57.9	3H40
3nwh	**2**	$P2_1$	2.60	415	55.2	6.31	4C of 1H	Left-handed	-	real	48.07	215	57.9	8H18 - TNCS true
4pxu	**2**	$P6_422$	2.60	288	72.5	1.56	2C of 1H	Left-handed	-	no	31.72	133	53.1	4H30
1env	**1**	H32	2.60	115	68.7	14.74	1C of 2H	Left-handed	-	no	54.45	105	41.9	4H18
2o1j	**2**	$C222_1$	2.70	174	49.3	16.84	4C of 1H	Left-handed	-	real	50.72	121	55.0	4H18 - TNCS true - sliding autotracing
3thf	**2**	$P2_12_12$	2.70	349	66.5	47.62	2C of 3H	Left-handed	-	no	32.81	232	50.5	6H25 - RMSD 0.5 - pack_tra
3r4h	**2**	$P4_32_12$	2.70	172	57.2	23.53	6C of 1H	Left-handed	-	no	41.25	118	45.6	4H18
1d7m	**1**	$C222_1$	2.70	202	72.2	31.53	2C of 1H	Left-handed	-	no	44.21	157	40.9	6H20 - pack_tra
1kql	**1**	$P6_422$	2.70	109	70.9	2.89	2C of 1H	Left-handed	-	no	53.69	91	35.7	4H18
2fxm	**1**	$C222_1$	2.70	238	60.3	57.41	2C of 1H	Left-handed	-	no	53.91	190	44.3	6H20 - helices reverted
2xu6	**1**	$P4_12_12$	2.70	127	62.4	16.79	2C of 1H	Left-handed	-	no	50.66	105	41.1	4H18
3vir	**1**	C2	2.70	251	78.2	20.34	4C of 1H	Left-handed	-	no	37.68	209	57.9	4H50
3v2r	**2**	$P2_1$	2.75	225	33.9	18.00	5C of 1H	Left-handed	-	apparent	44.41	132	58.6	4H30 - TNCS false - sliding autotracing
3tul	**2**	$P2_122_1$	2.79	521	60.2	16.30	4C of 3H	Left-handed	-	apparent	-	-	-	-
4gif	**2**	P321	2.80	45	64.2	36.46	1C of 1H	Left-handed	-	no	42.90	34	47.0	4H18
4nad	**2**	$P2_1$	2.80	266	52.3	56.50	2C of 3H	Left-handed	-	no	32.23	127	57.9	4H30
2jee	**2**	P1	2.80	312	63.5	77.26	4C of 1H	Left-handed	-	no	37.93	158	64.3	6H30 - helices reverted
1m3w	**1**	$P2_1$	2.80	120	37.5	13.87	4C of 1H	Left-handed	-	no	60.79	78	58.9	4H18
2no2	**1**	$P4_22_12$	2.80	102	74.7	62.55	1C of 1H	Left-handed	-	no	45.86	94	41.9	4H18
2w6b	**1**	$P6_3$	2.80	52	64.0	8.28	1C of 1H	Left-handed	-	no	59.89	38	59.4	4H18
3mqc	**1**	$P2_1$	2.80	400	56.9	26.79	4C of 1H	Left-handed	-	apparent	-	-	-	-
3t97	**1**	$P2_12_12_1$	2.80	168	47.5	33.39	3C of 1H	Left-handed	-	no	37.16	104	55.4	1H40
5djn	**2**	H32	2.82	158	77.1	19.45	2C of 1H	Left-handed	-	no	28.76	121	53.7	4H18
4bry	**2**	$I4_122$	2.89	138	77.7	2.29	2C of 1H	Left-handed	-	no	37.08	105	49.7	4H18 - pack_tra
4cgc	**2**	$P4_32_12$	2.90	83	60.8	19.52	3C of 1H	Left-handed	-	no	33.91	45	47.3	4H18
3onx	**2**	$P2_1$	2.90	250	65.7	47.91	2C of 2H	Left-handed	-	no	41.31	181	51.3	4H30 - helices reverted
3cvf	**1**	$P6_5$	2.90	286	85.9	11.73	4C of 1H	Left-handed	-	no	45.99	206	47.8	4H18
3v86	**1**	P321	2.91	27	46.2	5.92	1C of 1H	Left-handed	0.495	apparent	49.38	22	51.7	2H12 - TNCS false - VRMS true
4qkv	**2**	$P2_12_12_1$	3.00	285	63.5	61.48	3C of 1H	Left-handed	-	no	40.27	164	49.1	6H30

CURRICULUM VITAE

SCIENTIFIC PRODUCTION

- Iracema Caballero, Massimo D. Sammito, Pavel V. Afonine, Isabel Usón, Randy J. Read & Airlie J. McCoy. (2020). *Detection of translational non-crystallographic symmetry in Pattersons*. In revision.

- Alan M.V. West, Scott C. Rosenberg, Sarah N. Ur, Madison K. Lehmer, Qiaozhen Ye, Götz Hagemann, Iracema Caballero, Isabel Usón, Amy J. MacQueen, Franz Herzog & Kevin D. Corbett. (2019). *A conserved filamentous assembly underlies the structure of the meiotic chromosome axis*. eLife; 8:e40372

- Iracema Caballero, Massimo D. Sammito, Claudia Millan, Andrey Lebedev, Nicolás Soler & Isabel Usón. (2018). *ARCIMBOLDO on coiled-coils*. Acta Crystallogr D Struct Biol 74, 194-204.

POSTERS PRESENTATIONS AND TALKS

Presenting author Iracema Caballero:

- Iracema Caballero. *Using ARCIMBOLDO_LITE for phasing with helical models*. Oral presentation at the Phasing@Home Online Meeting in July 14th, 2020.

- Iracema Caballero. *Data quality assessment and detection of pathologies: things you should look at once you have a native data set*. Oral presentation at the Methods in Structural Biology Programme Seminar held in Barcelona in November 28th, 2019.

- Iracema Caballero, Massimo D. Sammito, Isabel Usón, Randy J. Read & Airlie J. McCoy. *Dealing with Modulated Macromolecular Structures with Translational Non-Crystallographic Symmetry*. Oral presentation at the 32nd European Crystallographic Meeting held in Vienna in August 22th, 2019.

- Iracema Caballero, Massimo D. Sammito, Isabel Usón, Randy J. Read & Airlie J. McCoy. *Determination of Translational Non-Crystallographic Symmetry*. Oral presentation at the CCP4 study weekend held in Nottingham in January 9th, 2019.

- Iracema Caballero. *Detecting tNCS for crystallographic phasing*. Oral presentation at the Structural and Computational Biology Programme Seminar held in Barcelona in December 12th, 2018.

- Iracema Caballero. *ARCIMBOLDO on coiled-coils*. Software demonstration at the Software Fayre of the ECM31 held in Oviedo in August 25th, 2018.

- Iracema Caballero, Massimo Sammito, Claudia Millán, Nicolas Soler, Andrey Lebedev & Isabel Usón. *Overcoming phasing difficulties in coiled-coils with ARCIMBOLDO_LITE: verifying solutions*. Poster presentation at the 31st European Crystallographic Meeting held in Oviedo in August 22nd-27th, 2018.

- Iracema Caballero. *Phasing coiled-coils with ARCIMBOLDO, low resolution challenges*. Oral presentation at the Structural and Computational Biology Programme Seminar held in Barcelona in March 20th, 2017.

Iracema Caballero as co-author:

- Airlie J. McCoy, Robert D. Oeffner, Tristan I. Croll, Kaushik S. Hatti, Massimo D. Sammito, Duncan H. Stockwell, Iracema Caballero, Claudia Millán, Isabel Usón & Randy J. Read. *Phaser - The Next Generation*. Oral presentation at the CCP4 study weekend held in Nottingham in January 9th, 2019.

- Randy J. Read, Robert D. Oeffner, Iracema Caballero, Isabel Usón & Airlie J. McCoy. *Information content in molecular replacement*. Oral presentation at the CCP4 study weekend held in Nottingham in January 10th, 2019.

- Isabel Usón, Claudia Millán, Massimo Sammito, Rafael Borges, Nicolas Soler, Iracema Caballero & Ana Medina*. All is fair in phasing: the combined artillery in ARCIMBOLDO.* Oral presentation at the 31st European Crystallographic Meeting held in Oviedo in August 24th, 2018.

- Rafael Borges, Massimo Sammito, Claudia Millán, Nicolas Soler, Ana Medina, Iracema Caballero, Marcos R. M. Fontes & Isabel Usón. *Expanding partial structures by assembling most probable side chain composition.* Poster presentation at the 31st European Crystallographic Meeting held in Oviedo in August 22nd-27th, 2018.

- Nicolas Soler, Claudia Millán, Massimo Sammito, Iracema Caballero, Rafael Borges & Isabel Usón. *Recent advances in ARCIMBOLDO towards low resolution*. Poster presentation at the International school of crystallography held in Erice (Italy), in June 2nd-11th, 2017.

- Isabel Usón, Iracema Caballero, Massimo D. Sammito, Claudia Millan, Andrey Lebedev & Nicolás Soler. *ARCIMBOLDO at low resolution*. Oral presentation at the CCP4 study weekend held in Nottingham in January 10th, 2017.

PARTICIPATION IN CONFERENCES, MEETINGS, SYMPOSIUMS, SCHOOLS AND COURSES

- August 18th-23th, 2019 (Vienna, Austria) - The 32nd European Crystallographic Meeting

- June 20th, 2019 (Barcelona) - Annual Scientific Meeting 2019 of the Structural Biology Unit

- January 8th-10th, 2019 (Nottingham, UK) - CCP4 study weekend

- December 19th, 2018 (Barcelona) - IBMB Xmas Meeting

- August 22nd-27th, 2018 (Oviedo, Asturias) - The 31st European Crystallographic Meeting

- August 18th-22nd, 2018 (Mieres, Asturias) - European Crystallographic Computing Forum

- June 18th-19th, 2018 (Barcelona) - Ignacio Fita's 65th Anniversary Symposium

- May 7th-8th, 2018 (l'Espluga de Francolí, Tarragona) - Annual Scientific Meeting 2018 of the Structural Biology Unit

- December 18th, 2017 (Barcelona) - Xmas Meeting IBMB

- September 19th-21st, 2017 (Barcelona) - Conference on methods and applications in the frontier between MX and CryoEM

- May 29th-30th, 2017 (Prades, Tarragona) - Annual Scientific Meeting 2017 of the Structural Biology Unit

- May 5th-10th, 2017 (Madrid) - Macromolecular Crystallography School

- December 20th, 2016 (Barcelona) - Xmas Meeting IBMB

- June 6th-7th, 2016 (The Valley of Nuria, Girona) - Annual Scientific Meeting 2016 of the Structural Biology Unit

CSIC and UB courses:

- 2020, Introduction to applied statistics with R software, online, 30 hours
- 2020, Adobe InDesign CS5, online, 40 hours
- 2019, UB Python Course
- 2019, DRUPAL 8: creation of interactive and efficient websites, online, 30 hours
- 2019, PowerPoint basic 2016, online, 40 hours
- 2018, Advanced Python, online, 60 hours
- 2018, Complete Photoshop C5, online, 50 hours
- 2018, Specific English: English presentations, online, 30 hours
- 2017, Applied Statistics I, Barcelona (Research and Development Center "Pascual Vila"), 25 hours / April 24th-28th, 2017 (Barcelona) - Course in Applied Statistics

Research stays:

I have completed several research stays at the laboratory of Prof. Dr. Randy J. Read, FRS, in the Haematology Department of the University of Cambridge at the Cambridge Institute for Medical Research. The group is distinguished by having introduced Bayesian methods of maximum probability in resolution of crystallographic structures, implemented in its *Phaser* molecular replacement program:

- October 23th-26th, 2018
- November 27th-30th, 2018
- February 4th-8th, 2019
- July 15th to August 17th, 2019