Machine learning applied to zeolite synthesis: the missing link for realizing high-throughput discovery

Manuel Moliner,[1] Yuriy Román-Leshkov,[2] Avelino Corma[1*]

[1] Instituto de Tecnología Química, Universitat Politècnica de València-Consejo Superior de Investigaciones Científicas, Avenida de los Naranjos s/n, 46022 València, Spain

[2] Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States

* Corresponding author: E-mail address: acorma@itq.upv.es

Conspectus

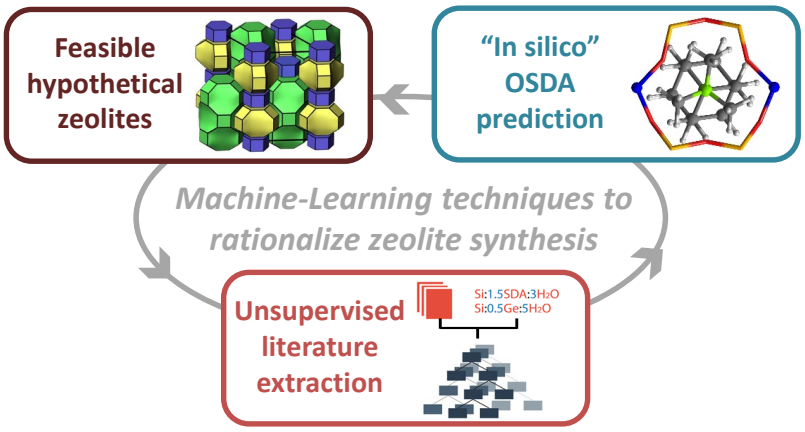
Zeolites are microporous crystalline materials with well-defined cavities and pores, which can be prepared under different pore topologies and chemical compositions. Their preparation is typically defined by multiple interconnected variables (e.g. reagent sources, molar ratios, ageing treatments, reaction time and temperature, among others), but unfortunately their distinctive influence, particularly on the nucleation and crystallization processes, is still far from being understood. Thus, the discovery and/or optimization of specific zeolites is closely related to the exploration of the parametric space through trial-and-error methods, generally by studying the influence of each parameter individually.

In the last decade, machine learning (ML) methods have rapidly evolved to address complex problems involving highly non-linear or massively combinatorial processes that conventional approaches cannot solve. Considering the vast and interconnected multiparametric space in zeolite synthesis, coupled with our poor understanding of the mechanisms involved in their nucleation and crystallization, the use of ML is especially timely for improving zeolite synthesis. Indeed, the complex space of zeolite synthesis requires drawing inferences from incomplete and imperfect information, for which ML methods are very well-suited to replace the intuition-based approaches traditionally used to guide experimentation.

In this Account, we contend that both existing and new ML approaches can provide the "missing link" needed to complete the traditional zeolite synthesis workflow used in our quest to rationalize zeolite synthesis. Within this context, we have made important efforts on developing ML tools in different critical areas, such as 1) data-mining tools to process the large amount of data generated using high-throughput platforms; 2) novel complex algorithms to predict the formation of energetically-stable hypothetical zeolites and guide the synthesis of new zeolite structures; 3) new "ab-initio" OSDA predictions to direct the synthesis of hypothetical or known zeolites; 4) an automated tool for non-supervised data extraction and classification from published research articles.

ML has already revolutionized many areas in materials science by enhancing our ability to map intricate behavior to process variables, especially in the absence of well-understood mechanisms. Undoubtedly, ML is a burgeoning field with many future opportunities for further breakthroughs to advance the design of molecular sieves. For this reason, this Account includes an outlook of future research directions based on current challenges and opportunities. We envision this Account will become a hallmark reference for both well-established and new researchers in the field of zeolite synthesis.

Graphical conspectus



1.- Introduction

Designing crystalline materials with tailored physicochemical properties is critical to industries spanning chemicals and petroleum to pharmaceuticals and electronics. Zeolites are crystalline, microporous aluminosilicates with well-defined cavities and pore topologies of molecular dimensions that directly impact the global economy with their ubiquitous use in many large-scale catalytic and absorption processes. Few crystalline materials exhibit the level of synthetic complexity encountered in the preparation of zeolites, where multiple parameters (i.e. reagent sources, molar ratios, ageing treatments, reaction time and temperature, among many others) can be used to alter the outcome.[1] The objective of selecting the appropriate synthesis conditions is to arrive at a crystal structure that has the desired physicochemical properties, but this point is very difficult because zeolite crystallization is not well understood. Zeolite crystallization is an interfacial phenomenon where the nucleation and crystallization of solute molecules is mediated by interactions with structure directing agents (SDAs), mainly organic and inorganic cations, that vary both in size and composition. In general, zeolites require the use of organic SDAs (OSDAs), which typically are amines and ammonium cations, featuring sizes and shapes commensurate with the geometry of porous channels/cages, to direct pore formation.[2-3] Preferentially, these OSDAs are amines, ammonium However, today, most zeolite discovery efforts continue to be based on trial-and-error approaches with minimal control over the resulting structures.[1-2]

Recent advances in artificial intelligence (AI) coupled with increased accessibility to large data sets has allowed the development of new algorithms and statistical methods capable of extracting relationships between variables in multidimensional systems.[4] In particular, the use of machine learning (ML)—a subfield of AI that relies on complex mathematical models that can effectively "learn" from past data to find complex patterns embedded within large data sets—in materials science has revolutionized our ability to map intricate behavior to process variables, especially in the absence of well-understood mechanisms. Considering the vast and interconnected multiparametric space in zeolite synthesis, our poor understanding of the control of mechanisms involved in their nucleation/crystallization, and the large amount of empirical data existing in the field, the use of ML is especially timely for improving zeolite synthesis. We expect these advances will have a dramatic impact on predicting hypothetical and known zeolites, as well as their synthesis conditions, undoubtedly accelerating the discovery of target microporous materials and ultimately improving our fundamental understanding.

To accomplish this goal, we must incorporate both existing and new ML approaches within the "traditional" zeolite synthesis workflow featuring well-established high-throughput synthesis/characterization devices and data-mining software (see Figure 1). We contend that ML will have a pivotal role in extracting, classifying, and interconnecting information across four critical research areas, namely i) high-throughput synthesis efforts, ii) design of feasible hypothetical zeolites, iii) "In-silico" prediction of OSDAs for target zeolites, and iv) automated, non-supervised data extraction from published literature. We note that most ML algorithms require "learning" from existing data sets to improve their accuracy, thus requiring effective methods to generate, organize, extract, and utilize existing and new information from computational and experimental outputs.

In this account, we present, within the context of concurrent efforts by many other research groups, the main tools developed by our group at the ITQ that have helped advance each of the above-mentioned research areas. First, we describe data-mining tools developed to process the large amount of data generated at the advent of high-throughput infrastructure. Next, we describe the development of complex algorithms to predict the formation of energetically-stable hypothetical zeolites, the "ab-initio" OSDA predictions to direct the synthesis of hypothetical or known zeolites, and, finally, the non-supervised data extraction and classification from literature. We conclude with an outlook of future research directions based on current challenges and opportunities.

2.- "High-Throughput" platforms for zeolite synthesis

In the late 1990's, the empirical data acquisition process for zeolite synthesis was greatly accelerated with the implementation of "high-throughput" (HT) synthesis methods.[5-7] These HT systems featured robotic multireactor systems operating under the tenets of automation, parallelization, and miniaturization that could explore many synthetic parameters automatically with drastic reductions in cost and time.[8] Unlike those used in the pharmaceutical industry, HT reactors for zeolite synthesis needed to be re-engineered to i) handle harsher temperature (~150-200°C), pressure (~1.5 MPa) and alkaline conditions, ii) dispense both liquid and solids of varying physicochemical properties, and iii) be compatible with automated and parallelized characterization techniques using microgram scale solids. The use of these HT devices fast-tracked the discovery of some novel zeolitic materials.[9-10] The fast rates and large amounts of data generated in these systems required the development of "data-mining" methods for rapid and unsupervised analyses (see Figure 1).[11]

2.1.- High-Throughput synthesis

The first multiautoclave designs consisted of a metal block containing Teflon-lined cylindrical chambers presenting diverse volumes, mostly between ~0.5-1 ml (see Figure 2A).[6-7, 12-13] We developed a multiautoclave for an in-house HT robotic system (see Figure 2B), based on 15 individual portable Teflon vessels with intermediate volumes of ~1-2 ml.[14] This feature was an important breakthrough in the design of multiautoclaves for zeolites because it allowed precise weight control during the entire gel preparation process. Modern multiautoclave designs also permit the *in-situ* filtration of the gels by connecting the vessels to a sealed, vacuum-pumped chamber.[15-16]

In general, the manufacture of fully automated devices for HT zeolite synthesis is restricted to large HT technology producers (e.g., Unchained Labs, Avantium, HTE, Bosch, Chemspeed and Zinsser), who can integrate robotics, engineering, and data-management into customer-tailored commercial instruments. Unlike simpler multiautoclave reactors, these highly-modular systems integrate liquid/powder dosing, stirring/mixing, milling/grinding, pH control, heating/cooling, among other requirements that can be included depending on the customer necessities. Several years ago, we developed an in-house automated system for zeolite synthesis at the ITQ (see Figure 2B), composed of a robotic arm that handled the vials, a liquid/solid dosing station, and a stirring/evaporation station.[14] Seven calibrated syringe pumps allowed precise liquid dosing, while the accurate control of the liquid/solid additions and liquid evaporations were accomplished through analytical balance measurements.

2.2.- High-Throughput characterization

HT zeolite synthesis requires concomitant topological, textural, and/or chemical analysis to be performed at commensurate timescales to avoiding workflow bottlenecks. Powder X-ray diffraction (PXRD) is the most common technique to identify crystalline microporous structures. Commercial vendors tackled automated collection of multiple PXRD patterns from large sample libraries by incorporating flat stages that could be moved in all directions (XYZ-stages, see Figure 2C). Gas sorption to probe zeolite porosity also required unique adaptation for HT analysis because these measurements may last several hours. In order to circumvent this bottleneck, a system for screening the porosity of large number of microporous materials was developed using the heat generated during gas adsorption for quantification.[17] Other characterization techniques, such as X-ray fluorescence (XRF),[18] IR spectroscopy combined with the adsorption/desorption of probe molecules,[19] and temperature programmed desorption (TPD),[20] have also been adapted for HT systems.

2.3.- Data-Mining Techniques

The development of the first automated HT zeolite synthesis systems drastically increased the number of experiments that could be performed in parallel and, consequently, increased the number of variables that could be explored simultaneously (see Figure 3).[6, 13-14] Accordingly, data-mining techniques begun their development almost simultaneously with HT synthesis as a means to aid in the exploration of broad synthesis spaces through careful design of experiments (DoE) and to analyze the large quantities of data generated by these experiments (see Figure 1).

2.4.1.- Design of Experiments (DoE)

In HT zeolite synthesis, selecting which variables to investigate during the initial DoE is a very challenging task because the effects that individual variables have on nucleation/crystallization mechanisms are highly intercorrelated. The first DoE for early HT zeolite synthesis campaigns were based almost exclusively on classic exploratory factorial designs.[6, 13-14] These designs involve generating possible combinations of two or more variables, which presenting different levels or values. Full factorial designs allowed to exploring the effect of each synthesis variable on zeolite crystallization, as well as the influence of interconnected variables (see Figure 3A).[6] We systematically studied the system TEA:SiO_2:Na_2O:Al_2O_3:H_2O by HT methods, where TEA is tetraethylammonium, in order to synthesize the Beta zeolite with high yields while using a low OSDA content in the synthesis gel.[14] Following a full-factorial design (see Figure 3B), a high-silica Beta with low OSDA contents (TEA/Si~0.27) and excellent crystallinity was obtained when using concentrated synthesis gels (e.g., H_2O/Si~5). The factorial design approach has also yielded new zeolite structures. For instance, researchers at UOP systematically explored simple mixtures of tetramethyl and tetraethylammonium OSDAs, ultimately discovering conditions that crystallized UZM-4 (12x8-rings) and UZM-5 (8x8-rings).[9]

In this respect, our group initiated some of the first efforts to incorporate simple statistics to properly evaluate the impact of the different synthesis variables during HT synthesis processes. Our objective was to direct the synthesis conditions more effectively in the second generation of experiments.

The discovery of ITQ-33 started by performing a very large initial factorial design (3×4^3) to explore unusual synthesis conditions using flexible OSDA molecules (e.g. hexamethonium).[10, 21] The initial proposed conditions totaled 192 experiments, spanning the following precursor ranges: Si/Ge~2-30, B/(Si+Ge)~0-0.05, OH/(Si+Ge)~0.1-0.5, H_2O/(Si+Ge)~5-30. This campaign

resulted in some phase-pure zeolites and several multi-phase mixtures, one of which contained an unknown phase that was named ITQ-33. The results were subjected to Pareto analysis in order to plan a second generation of experiments aimed at isolating ITQ-33.[21] The Pareto analysis showed that ITQ-33 was most influenced by Si/Ge and OH/(Si+Ge) (see Figure 3C). In light of these findings, a second set of 18 experiments was proposed that yielded phase-pure ITQ-33—a unique extra-large pore zeolite interconnected bidirectionally with 10-ring channels exhibiting remarkable selectivity to diesel and propylene in the cracking of vacuum gasoil.[10, 21]

In a similar way, the discovery of ITQ-30—a zeolite with excellent catalytic performance for the alkylation of benzene with propylene to produce cumene—involved studying the directing role of the rigid and bulky *N*-methyl-sparteinium OSDA across 144 experiments spanning a broad range of synthesis conditions.[22] These experiments generated ITQ-21 (a large-pore zeolite) and an unknown phase named ITQ-30. Further statistical analysis revealed that the crystallization of ITQ-30 was negatively influenced by increasing water and Al contents, regardless the Si/Ge ratio (see Figure 3D). Accordingly, the next generation of experiments afforded the crystallization phase-pure ITQ-30 zeolite under Ge-free conditions.[22]

Advanced methods, including artificial neural networks (ANNs) and genetic algorithms (GAs), are necessary for cases in which simple statistics cannot effectively guide experimental design.[14, 23] ANNs are non-linear systems that can model complex multidimensional studies through nodes with connections reminiscent of those found in a biological brain (see Figure 4A). GAs operate with similar mechanisms to those behind Darwinian evolution, where the best variables dominate the next generation population by selecting the proper operators (e.g., selection, crossover, and mutation). As a standout example, we combined ANNs and GAs to improve the catalytic behavior of Ti-silicates for the selective epoxidation of olefins.[23] Specifically, we used ANNs to predict the internal relationships between different synthesis variables after being properly trained with previous data, and then used GAs to optimize the next generation of material synthesis experiments considering the knowledge extracted by the ANN. Different synthesis variables were considered (e.g. surfactant, organic modifier, OH or titanium contents). Three generations of 38 samples were synthesized using the NN-GA optimization process, achieving an outstanding improvement of both catalytic activity and epoxide selectivity each generation (see Figure 4B). The improved catalytic behavior was found when decreasing the amount of the organic modifier while keeping the OH/Si molar ratio at ~0.2.[23]

2.4.2.- Data extraction and classification

Data extraction and classification from HT experiments can become a major bottleneck if not managed properly. For this reason, significant efforts have been dedicated to develop non-supervised data analysis tools with the aim of generating new Data-Mining mapping/exploration methods that allow facile and rapid data extraction and visualization.

Clustering using k-means and Principal Component Analysis (PCA)

Clustering analyses of raw PXRD data permits non-supervised classification of crystalline materials into diverse groups based on similarities in the diffraction patterns. This technique is particularly useful for discriminating and identifying pure-phases in mixed systems. By considering the PXRD patterns as structural vectors, the classification or clustering of the achieved solids can be carried out by applying statistical tools, such as k-means clustering and PCA.

The k-means clustering algorithm assigns *n* samples into *k* clusters with the nearest mean (a value which is updated every time a new component is added to the cluster) and this process is repeated until all components are classified into different clusters. We utilized the k-means clustering algorithm for classifying HT synthesis raw data, by selecting all the PXRD patterns obtained from the 144 syntheses carried out during the ITQ-21/ITQ-30 synthesis campaign described in the previous section (see Figure 5A).[11] The k-clustering analysis binned the raw PXRD data results into three well-defined clusters: amorphous (cluster 1), ITQ-21 (cluster 2), and ITQ-30 (cluster 3) (see Figure 5B). Interestingly, the overall match between real phases and the proposed clusters following the k-means analysis was ~90%, demonstrating the high potential of this tool.

The PCA uses statistical methods to reduce the information contained within a long descriptor (e.g., an individual PXRD pattern containing all the diffraction intensities), into three structural principal components (SPCs) while conserving all the information of the original data. We applied the PCA analyses to the data from the same ITQ-21/ITQ-30 synthesis campaign,[11] achieving a dimensional reduction to just three components for each PXRD pattern, ultimately providing a very simple 3D cluster visualization (see Figure 5C) that allowed us to correlate the SPC projections with the crystallinity and chemical composition of each synthesized material.

Adaptable Time Warping (ATW) models

Our group has developed protocols to extract and predict structural parameters of synthesized materials in order to classify and relate structural features to synthesis variables.[24-25] For

example, a common issue in the analysis of diffraction data is that the PXRD pattern of a specific crystalline structure can present large differences, both in peak intensity and 2θ shifts, depending on its crystal size or chemical composition. This complicates non-supervised structural recognition, particularly when mixed phases are present. Our ATM algorithm allows searching distances to detect 2θ shifts between the input and the reference pattern instead of comparing the value of the input pattern at a specific 2θ angle with a reference at the same 2θ angle (see Figure 6A).[24-25] The method was successfully validated when the diffraction patterns for eight different crystalline zeolites were correctly identified with a classification error of <3% from the complex diffractogram of the mixed solids (see Figure 6B).[24-25]

3.- Computational methods and machine learning techniques for zeolite synthesis

Modern zeolite synthesis requires close integration between experiments and computation to gain insight into the fundamental underpinnings linking structure and property to the synthesis recipe. Computational methods, including molecular dynamics simulations, electronic structure calculations using first principles, Monte Carlo techniques, and continuum macroscopic approaches, have been developed hand-in-hand with experimental methodologies to understand the assembly of microporous materials.[26] Theoretical simulations can in principle require less time compared to experimental measurements, thereby accelerating the discovery of new materials, reducing both time and cost expenditures. However, unlike well-established HT experimental protocols, most computational techniques used to date cannot be implemented in a HT manner without jeopardizing accuracy given the need to use expensive high-level quantum chemistry methods to correctly calculate the complex energy landscape of hydrothermal crystallization processes. In this respect, the use of ML algorithms offers an attractive avenue to accelerate the discovery and optimization of molecular sieve synthesis by bypassing the need for resource-intensive simulations and instead use learned patterns from training examples to estimate properties or predict outcomes under unexplored conditions. In the next section we present some of our efforts towards the development of computational tools for enabling the use of ML for zeolite synthesis.

3.1.- Hypothetical structures and phase identification

To date, approximately 240 distinct zeolite structures have been successfully synthesized, which is in stark contrast to the millions of hypothetical structures generated using mathematical constructs. Hypothetical zeolites have been generated using symmetry-constrained geometric linkage of subunits, tiling theory, and genetic algorithms coupled with bonding rules and lattice energy minimization programs to down select chemically-feasible structures.[27-28] Deem and co-

workers who used Monte Carlo simulations coupled with interatomic potential refinement to investigate the arrangement of Si atom positions, unit cells, space groups, and framework densities in porous materials, generating over 2.6 M predicted zeolite-like materials.[29] Since then, the chemical feasibility of these hypothetically structures has been evaluated by various groups using more complex methods, including local interatomic distances (LIDs), TTT angles, minimum 5-th neighbor distance, average tetrahedral order parameter, and pore dimensionality to further refine the subgroup of synthetically accessible materials.[30-31]

Evidently, these hypothetical structures can be used to identify new zeolites synthesized in the laboratory. For instance, when we obtained the structure of ITQ-51 (see Figure 7A), an extra-large pore zeolite with 16-rings synthesized using bulky proton sponges as OSDAs, we realized that the structure was included in Deem's database as a pure-silica analog.[32] This encouraged us to explore analogous structures, an exercise that revealed extensive similarities between ITQ-51 and AlPO-31. Using this information, we surmised that if the 4-rings forming the 12-ring channel in AlPO-31 were substituted by six helical 4-ring chains, a hypothetical 18-ring zeolite (denoted as T18MR, see Figure 7B) would be formed.[32] This approach was extended to generate three additional hypothetical large-pore structures that are currently important synthetic targets in our laboratory.

Hypothetical zeolite construction is also a powerful tool when used to narrow down the structural space for unknown, highly-complex crystals for which limited available characterization data is available. ITQ-43 is a hierarchical zeolite featuring a very open structure (11.4 T-atoms/1000 $Å^3$) and cloverleaf-like channels formed by 28-rings (see Figure 7C).[33] When we first synthesized ITQ-43, we knew from PXRD data that the crystal structure was built either by *C222*, *Cmm2*, *Cm2m*, *C2mm*, or *Cmmm* space groups. However, due to its large cell parameters and low stability in its calcined form, we could not extract reliable structural information from either diffraction or HR-TEM data. With limited characterization data, we relied on an in-house simulation program to generate a feasible set of potential structural candidates.[33-34] More specifically, we developed an evolutionary algorithm deployed using GPU hardware that independently manipulated fixed arrays of variables corresponding to atoms coordinates belonging to the asymmetric unit cell. By using suitable fitness evaluation and optimization criteria we generated the 50 most-viable structures. The resolved crystal structure of ITQ-43 was shown to be one of these predicted structures, thereby demonstrating the usefulness of hypothetical structures generated through experimentally-imposed constrains.

In a very elegant attempt to predict the synthesis of hypothetical zeolites, Yu et al. have described a multidatabase, Zeobank, containing synthesis conditions, known structures, and hypothetical structures to perform computational-guided studies between synthetic parameters and zeolite structures.[35-37] Different data-mining techniques, as support vector machines (SVM) and neural networks (NN), were investigated for correlating experimental conditions and crystalline products.

3.2.- OSDA-zeolite prediction

For zeolites, we now understand that coupled thermodynamic and kinetic factors, mainly in the form weak van der Waals interactions between OSDAs and inorganic moieties that influence nucleation events, are responsible for determining the synthesis product. Accurately capturing the fine interplay between organic and inorganic species at the molecular level necessitates the development of the appropriate computational tools.

Inspired by the work of Catlow et. al.,[38] we used Monte Carlo and energy minimization molecular dynamics simulations to rationalize the effect of OSDA stabilization on the zeolite structure.[39-40] Specifically, by explicitly including the OSDA-OSDA and OSDA-zeolite interactions in the potential energy function, we established a simple, yet powerful framework to approximate the energy change of the system upon OSDA incorporation. This approach allowed us to isolate one product out of two closely related zeolite structures, namely ITQ-7 and ITQ-17, by identifying an optimal OSDA out of several structurally-similar azocompounds.[39] We used similar molecular simulations to predicted an optimal OSDA to synthesize phase pure Ti-containing BEC.[40] Notably, the generality of the approach was demonstrated when a commercial tert-butyl-iminotris(dimethylamino)phosphorane OSDA was identified amongst multiple phosphazenes to stabilize the structure of the elusive boggsite zeolite, enabling, for the first time, the synthesis of a molecular sieve that had only been obtained as a naturally-occurring mineral.[41]

These results suggest that these computational methods allow the *a priori* prediction of an OSDA molecule to synthesize a desired framework. However, a major shortcoming of this approach is that molecules used in the calculation could be difficult to synthesize. A computational method to predict chemically synthesizable OSDAs for crystalline molecular sieves was reported by Deem, wherein transformations from organic chemistry were applied to a library of available reagents to generate molecules that were scored based on rigidity, volume, stability under synthesis conditions, and energy of interaction with the zeolite.[42] Davis et. al. validated the

method experimentally by successfully synthesizing the SFW zeolite,[43] and the enantioenriched polycrystalline STW zeolite (see Figure 8).[44-45]

The time requirement to perform accurate molecular dynamics simulations is inextricably correlated with computing power and availability. For example, when predicting the suitability of a molecule to serve as an OSDA for a target zeolite, calculating the stabilization energy within the framework is one of the most intensive computational steps, requiring several hours of CPU time. Comparatively, trained ML algorithms are inherently more efficient and less computationally-intensive, making them ideally suited to replace computationally expensive molecular dynamics evaluations of the stabilization energy of the OSDA inside zeolites. Deem et. al. used a data set of 4781 OSDA stabilization energies previously computed for BEA zeolite to train a NN on the molecular structure descriptors of OSDAs to forecast their stabilization energies.[46] Notably, the trained network was able to predict stabilization energies for new putative molecules with comparable accuracy to that obtained with molecular dynamics simulations, generating a list of new, chemically-synthesizable molecules that could be used to crystallize the elusive polymorph A of BEA.

This approach could be applied to an even more ambitious goal: the synthesis of custom-designed zeolites that are tailored for specific applications. We recently demonstrated a new concept in which a zeolite is prepared using OSDAs that mimic the transition state (TS) of preestablished reactions, resulting in drastically enhanced reaction rates and selectivities.[47-49] The idea of imprinting a TS within a rigid crystalline structure represents a disruptive departure from traditional catalysis and provides exciting opportunities for designing more selective, active, and responsive solids that can be further extended with the help of ML.

3.3.- Literature Data extraction

The full-scale implementation of ML techniques for zeolite synthesis is hindered by the challenges associated with data sparsity and scarcity. Indeed, open-access datasets and synthesis protocols for zeolites are smaller and more diverse compared to other efforts like the Materials Genome Initiative. Low availability of materials data causes underfitting and large prediction bias in ML models,[50] and these shortcomings can be further exacerbated if negative examples are not included. In this respect, the prolific peer-reviewed manuscript and patent literature for zeolite synthesis offers a vast amount of data collected over a span of six decades. However, collecting the relevant data from tables, figures, and experimental sections of thousands of documents is an impossible task without automation. Leveraging recent advances

in natural language processing and text markup parsing tools, we recently developed a tool in collaboration with Olivetti and co-workers to automatically extract synthesis information and trends from zeolite journal articles (see Figure 9).[51] Specifically, our pipeline automatically located, extracted, and organized zeolite synthesis data from both the tables and main text of thousands of articles. We validated the accuracy of the extracted data using a subset of articles related to the preparation of germanium-containing zeolites for which the pipeline accurately identified the complex relationships between the synthesis parameters and resulting topology. We envision that with future improvements and small changes in data engineering, this tool can be used to solve several other research questions in zeolite synthesis chemistry.

4.- Frontiers of ML for zeolite synthesis

The main challenge in zeolite synthesis is the incomplete understanding of the molecular-level interactions and the kinetic and thermodynamic driving forces that govern the adsorption and binding specificity of OSDAs to precursors leading to specific nucleation/crystallization events. A fundamental understanding of these processes and the ability to *a priori* control crystallization requires synergistic research efforts to probe atomic to macroscopic length scales. The complex and multidimensional space of zeolite synthesis requires drawing inferences from incomplete and imperfect information, for which ML methods are very well-suited to replace the intuition-based, trial-and-error approaches traditionally used to guide experimentation.

We showed how databases of hypothetical zeolites can play an important role in zeolite discovery. However, the standard representation of crystal structures has been optimized for human learning, which might not necessarily be optimal for ML. We need to develop improved descriptors that can capture the properties we intend to model in more effective ways. Thus, developing "ML-friendly" representations of crystal structures that are easily transferable across methods is essential for reaching the level of predictive sophistication we have acquired in molecular systems. In organic synthesis, NN have been used to create fingerprints or molecular fragments for molecules in reactions, leading to improved prediction capabilities.[52] The field of zeolite crystallization could benefit tremendously from developing new approaches to define more efficient nomenclature and structural representations.

Undoubtedly, new ML approaches will play an important role for enabling the computer-assisted synthesis of organic molecules that could replace expensive OSDAs for known zeolites or predict the structure of molecules leading to new topologies. Recent work showing the use of ML to predict the stabilization energies of chemically-synthesizable OSDAs with an accuracy

commensurate to that of high-level molecular dynamic simulations is a true testament of the versatility of these algorithms. Future directions should leverage current efforts in ML-based automated organic molecule retrosynthesis,[53] including the unsupervised selection of reaction conditions,[54] catalysts, and reagents, to deploy AI-driven experimental platforms including full automation for highly parallelized OSDA synthesis. OSDA design could benefit from cutting-edge algorithms, such as generative adversarial networks and reinforcement learning already used in the design of biological compounds, in which new molecules with specific physicochemical features are produced using a punishment/reward system analogous to that in psychological conditioning.[55] Further, given the limited amount of data compared to other fields, ML tools applied to zeolite synthesis will benefit from cutting-edge approaches in meta-learning, including neural Turing machines,[56] and imitation learning.[57]

Naturally, the efficacy of ML schemes hinges on both the quality and amount of data used to train the algorithms. We showed the tremendous advantages of natural language parsing tools capable of accessing the vast amount of experimental data published in patents and journal articles in an automated fashion. Future efforts should focus on developing improved ML and visual recognition AI algorithms used in face-recognition and self-driving vehicles to extract and classify data from figures (including diffractograms and spectrograms) in addition to written records, journal articles, patents, laboratory notebooks, and internal databases. In many cases, however, data is collected and stored in many disjoint formats without having validation or standardized metadata. It is essential, that we, as a community, create and adopt robust standardization protocols to make data/metadata accessible in a computer-readable form akin to those implemented in parallel fields, while also allowing for easily implementing changes as the data are updated or corrected.[58]

The growth of ML tools used for zeolite design is exciting, but they cannot be used yet as a "silver-bullet" for solving all open questions in the field. It is imperative we understand the limitations of ML tools so that we can help them learn properly. We should keep in mind that predictive models developed by ML tools might not be interpretable to humans, given that the way ML models represent knowledge rarely mirrors that used by scientists. Therefore, as we embrace the ML-based design, we have to continue working on representing data in a manner that maximizes the amount that humans and machines learn from each other.

Biographical sketches

Manuel Moliner obtained his Ph.D. at the Polytechnic University of Valencia in 2008 under the guidance of Prof. Corma. After a two-year Postdoc at Caltech with Prof. Davis, he joined the ITQ as a Tenured Scientist. His research interests are at the interface between heterogeneous catalysis and materials design.

Yuriy Román-Leshkov is a Professor of Chemical Engineering at MIT. He received his Ph.D. from the University of Wisconsin-Madison in 2008 under the guidance of Prof. Dumesic and completed his postdoctorate with Prof. Davis at Caltech. At MIT, the Román Group specializes in elucidating the structure-activity relationships of heterogeneous catalysts and the design of novel catalytic materials.

Avelino Corma is a Professor at the ITQ (UPV-CSIC). He received his Ph.D. at the Complutense University of Madrid in 1976 and performed his Postdoc at Queen's University. His current research field is catalysis, covering aspects of synthesis, characterization, and reactivity in acid-base and redox catalysis.

Acknowledgements

This work has been supported by the EU through ERC-AdG-2014-671093, by the Spanish Government through SEV-2016-0683 and RTI2018-101033-B-I00 (MCIU/AEI/FEDER, UE), and by La Caixa-Foundation through MIT-SPAIN MISTI program (LCF/PR/MIT17/11820002). Y.R.-L. thanks the DoE for funding through the Office of Basic Energy Sciences (DE-SC0016214).

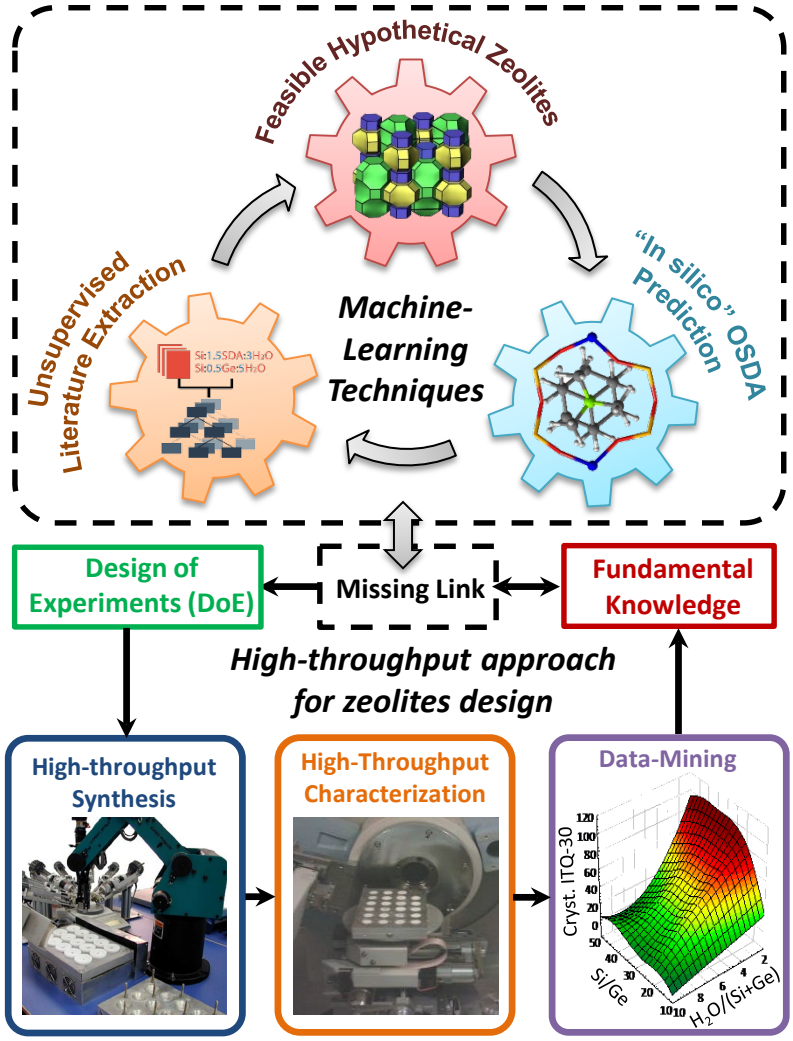


Figure 1. High Throughput discovery workflow for the synthesis of microporous materials

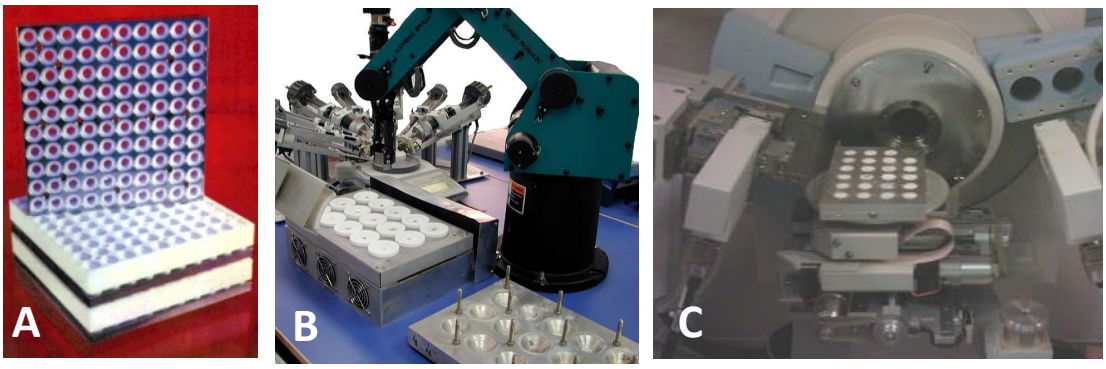


Figure 2. Images of the SINTEF multiautoclave (A), the in-house developed system for hydrothermal synthesis of zeolites built at ITQ (B) and a multisample preparation over an XYZ-stage in a PANalytical diffractometer at ITQ (C). Reproduced with permission from ref. [6, 14]. Copyright (1998 and 2005) Wiley and Elsevier, respectively.

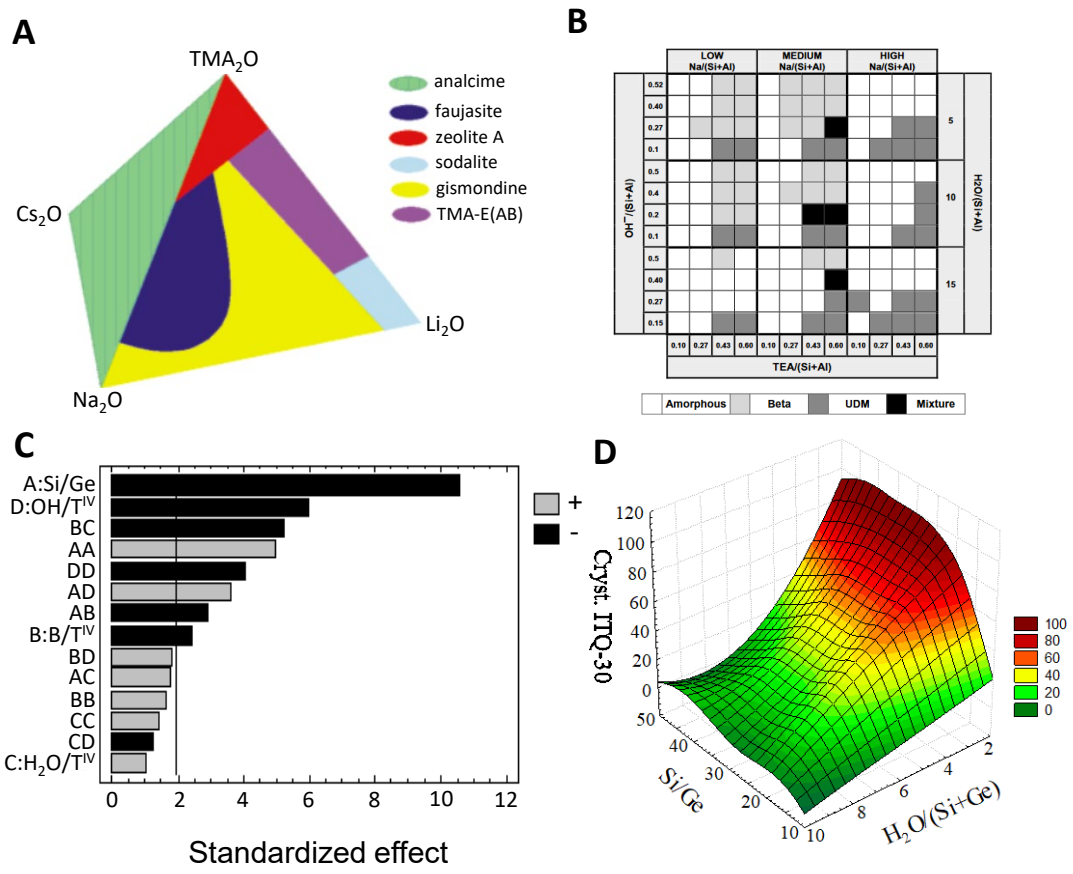


Figure 3. (A,B) Factorial-based designs proposed for HT synthesis of zeolites by Akporiaye et al. and Corma et al., respectively. Reproduced with permission from ref. [6, 14]. Copyright (1998 and 2005) Wiley and Elsevier, respectively. (C) Statistical evaluation of the influence of different variables in HT zeolite synthesis. Reproduced with permission from ref. [21]. Copyright (2008) Elsevier. (D) 3-D representation of the influence of different variables on the crystallinity of ITQ-30. Reproduced with permission from ref. [22]. Copyright (2006) Elsevier.

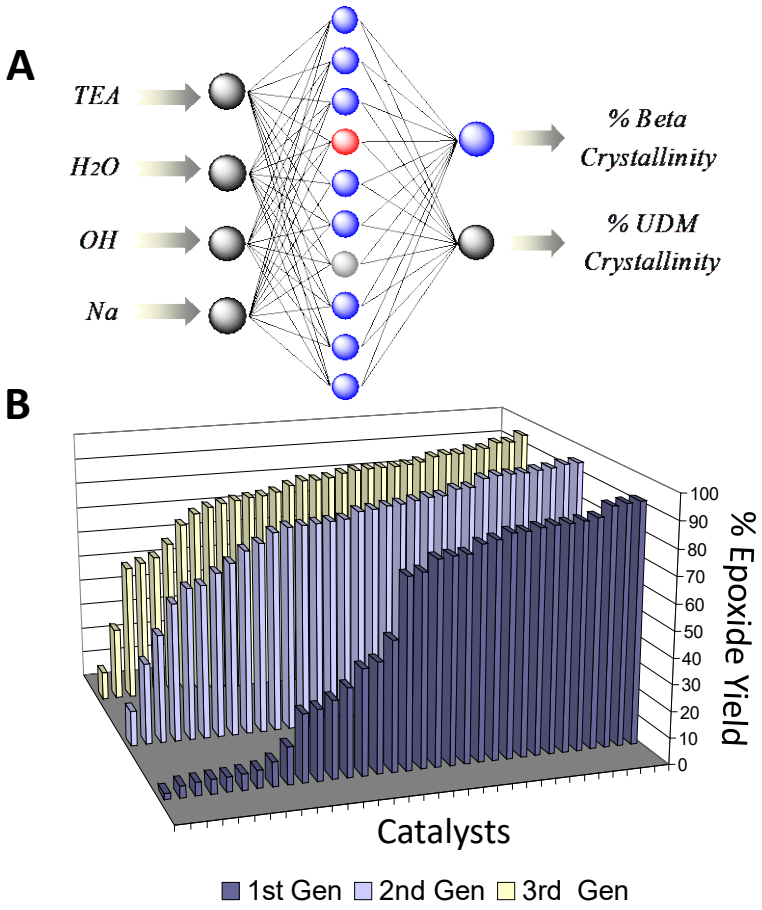


Figure 4. (A) Scheme of a Neural Network employed for modelling the dataset obtained during the HT synthesis study for Beta zeolite. Reproduced with permission from ref. [14]. Copyright (2005) Elsevier. (B) Evolution of the catalytic activity for the epoxidation reaction after three evolved generations. Reproduced with permission from ref. [23]. Copyright (2005) Elsevier.

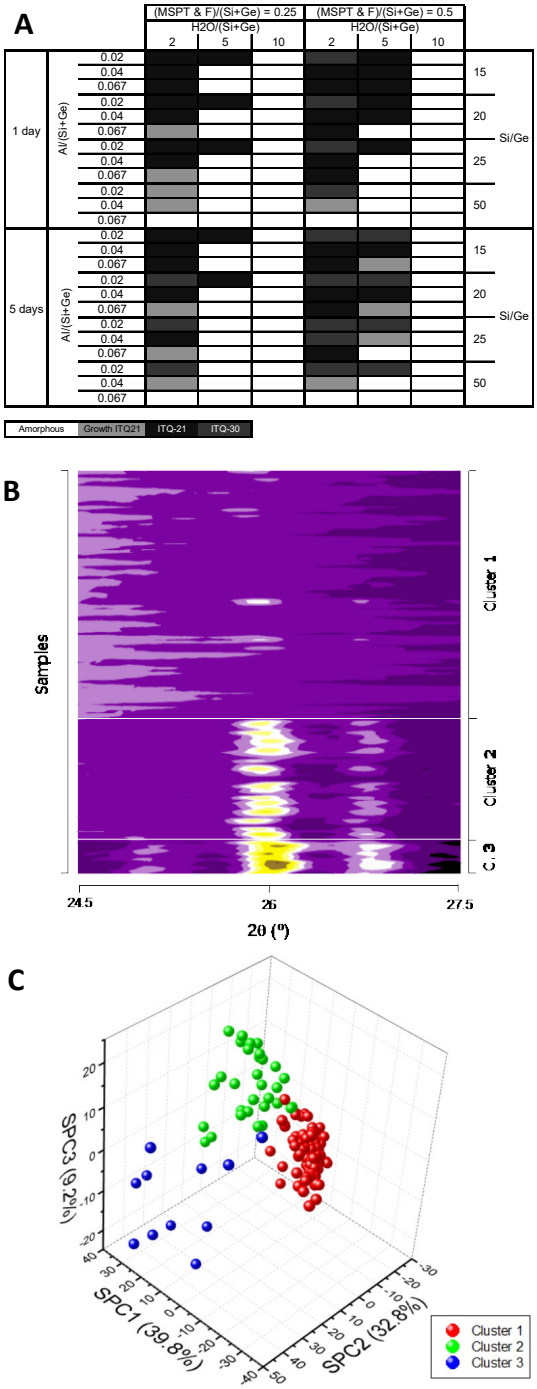


Figure 5. (A) Phase diagram achieved when varying multiple variables using methyl-sparteine as OSDA. (B) Clusters achieved when applying the k-clustering analysis to the raw PXRD data results of the ITQ-21/ITQ-30 study (note that the 2θ angle section comprised between 24.5-27.5° is presented). (C) Simple 3-D zeolite cluster representation achieved by the statistical dimensional reduction of the entire PXRD patterns to just three interrelated variables using the Principal Component Analysis (PCA). Reproduced with permission from ref. [11]. Copyright (2006) American Chemical Society.

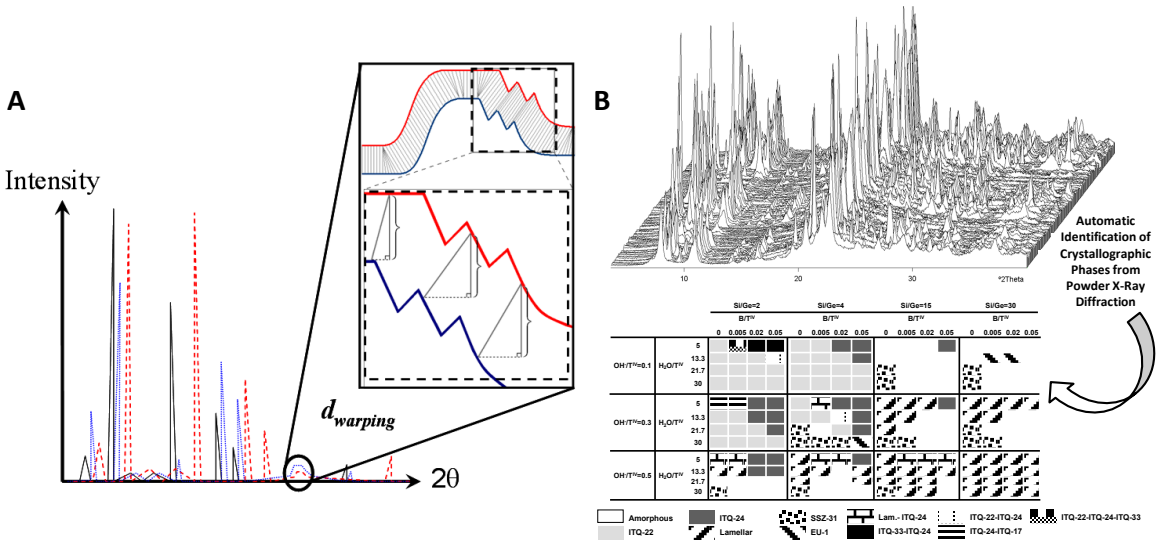


Figure 6. (A) Adaptable time warping (ATW) approach applied to powder X-ray diffractograms, which allows excellent identification accuracies even when peak intensity and 2θ shifts are present (see inset). (B) Automatic analysis of a PXRD dataset achieved using hexamethonium as OSDA to identify the different crystallographic phases. Reproduced with permission from ref. [24-25]. Copyright (2009 and 2008) Wiley and RSC, respectively.

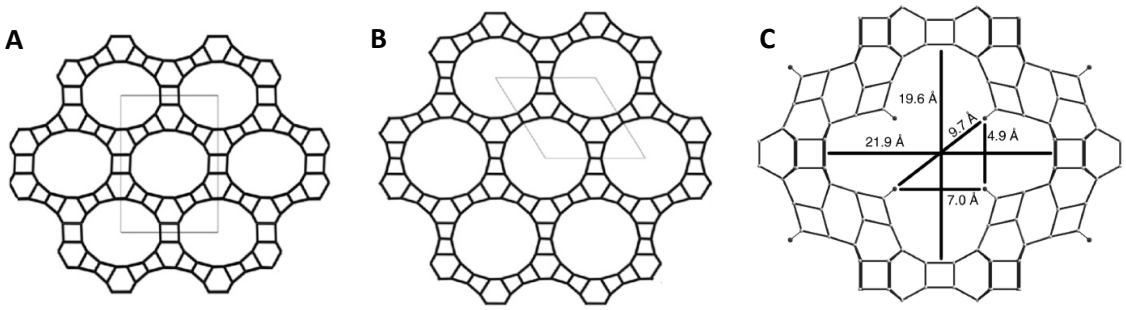


Figure 7. Zeolite structures of the ITQ-51 (A), the hypothetical T18MR (B) and ITQ-43 (C). Reproduced with permission from ref. [32-33]. Copyright (2013 and 2011) National Academy of Sciences and AAAS, respectively.

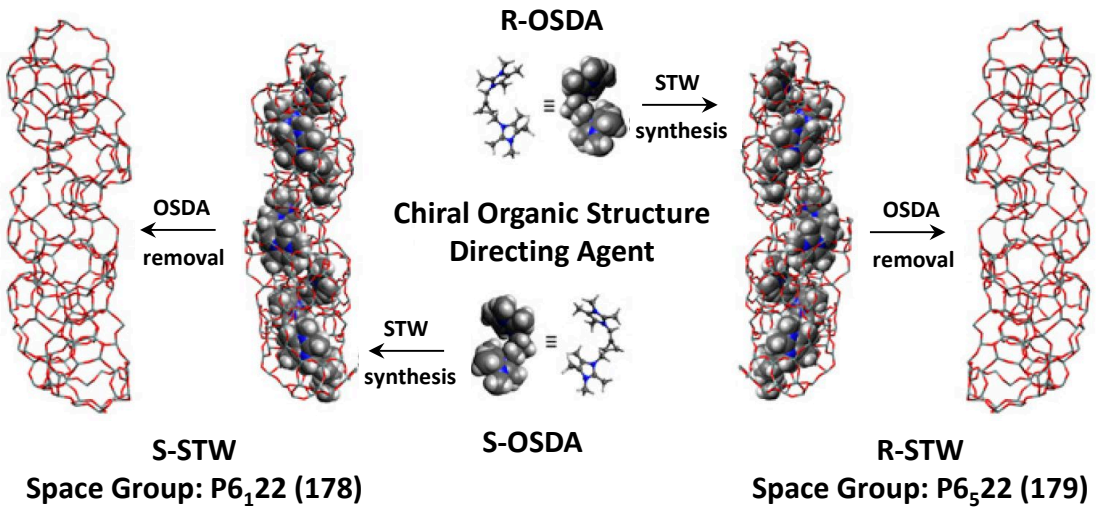


Figure 8. Representation of the OSDA-zeolite interaction for the synthesis of a chiral zeolite. Reproduced with permission from ref. [45]. Copyright (2017) National Academy of Sciences.

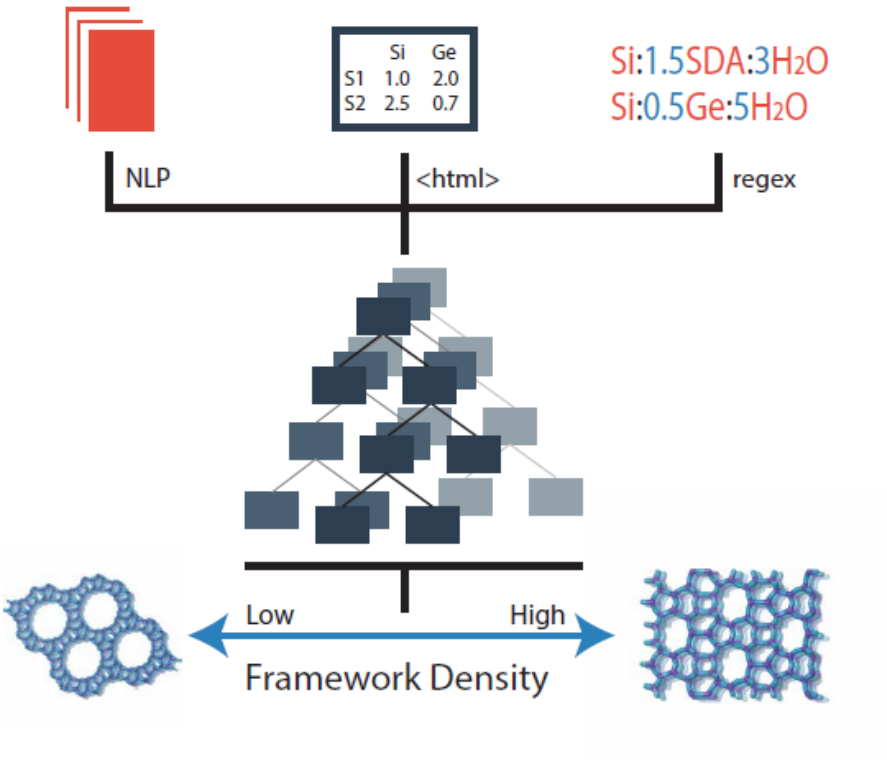


Figure 9. Scheme of the methodology employed for zeolite literature extraction from multiple aspects of a journal article (i.e. text and table data), modeling, and structure prediction (i.e. zeolite framework densities). Reproduced with permission from ref. [51]. Copyright (2019) American Chemical Society.

REFERENCES:

1. Cundy, C. S.; Cox, P. A., The Hydrothermal Synthesis of Zeolites: Precursors, Intermediates and Reaction Mechanism. *Micropor. Mesopor. Mater.* **2005,** *82*, 1-78.
2. Burton, A., Recent trends in the synthesis of high-silica zeolites. *Catal. Rev.* **2018,** *60*, 132-175.
3. Moliner, M.; Rey, F.; Corma, A., Towards the Rational Design of Efficient Organic Structure-Directing Agents for Zeolite Synthesis. *Angew. Chem. Int. Ed.* **2014,** *52*, 13880-13889.
4. Jordan, M. I.; Mitchell, T. M., Machine learning: Trends, perspectives, and prospects. *Science* **2015,** *349*, 255-260.
5. Schultz, P. G.; Xiang, X.; Goldwasser, I., *US 5,985,356 (1994)*.
6. Akporiaye, D. E.; Dahl, I. M.; Karlsson, A.; Wendelbo, R., Combinatorial Approach to the Hydrothermal Synthesis of Zeolites. *Angew. Chem., Int. Ed.* **1998,** *37*, 609-611.
7. Klein, J.; Lehmann, C. W.; Schmidt, H. W.; Maier, W. E., Combinatorial Material Libraries on the Microgram Scale with an example of Hydrothermal synthesis. *Angew. Chem., Int. Ed.* **1998,** *37*, 3369-3372.
8. Potyrailo, R. A.; Rajan, K.; Stowe, K.; Takeuchi, I.; Chisholm, B.; Lam, H., Combinatorial and High-Throughput Screening of Materials Libraries: Review of State of the Art. *ACS Comb. Sci.* **2011,** *13*, 579-633.
9. Blackwell, C. S.; Broach, R. W.; Gatter, M. G.; Holmgren, J. S.; Jan, D. Y.; Lewis, G. J.; Mezza, B. I.; Mezza, T. M.; Miller, M. A.; Moscoso, J.; Patton, R. L.; Rohde, L. M.; Schoonover, M. W.; Sinkler, W.; Wilson, B. A.; Wilson, S. T., Open-Framework Materials Synthesized in the TMA+/TEA+ Mixed-Template System: The New Low Si/Al Ratio Zeolites UZM-4 and UZM-5. *Angew. Chem., Int. Ed.* **2003,** *42*, 1737-1740.
10. Corma, A.; Díaz-Cabañas, M. J.; Jordá, J. L.; Martínez, C.; Moliner, M., High-throughput synthesis and catalytic properties of a molecular sieve with 18- and 10-member rings. *Nature* **2006,** *443*, 842-845.
11. Corma, A.; Moliner, M.; Serra, J. M.; Serna, P.; Díaz-Cabañas, M. J.; Baumes, L. A., A New Mapping/Exploration Approach for HT Synthesis of Zeolites. *Chem. Mater.* **2006,** *18*, 3287-3296.
12. Lai, R.; Kang, B. S.; Gavalas, G. R., Parallel Synthesis of ZSM-5 Zeolite Films from Clear Organic-Free Solutions. *Angew. Chem., Int. Ed.* **2001,** *113*, 422-425.
13. Choi, K.; Gardner, D.; Hilbrandt, N.; Bein, T., Combinatorial Methods for the Synthesis of Aluminophosphate Molecular Sieves. *Angew. Chem., Int. Ed.* **1999,** *38*, 2891-2894.
14. Moliner, M.; Serra, J. M.; Corma, A.; Argente, E.; Valero, S.; Botti, V., Application of artificial neural networks to high-throughput synthesis of zeolites. *Micropor. Mesopor. Mater.* **2005,** *78*, 73-81.
15. Caremans, T. P.; Kirschhock, C. E. A.; Verlooy, P.; Paul, J. S.; Jacobs, P. A.; Martens, J. A., Prototype high-throughput system for hydrothermal synthesis and X-ray diffraction of microporous and mesoporous materials. *Micropor. Mesopor. Mater.* **2006,** *90*, 62-68.
16. Anderson, B. J.; Gillespie, R. D.; Bricker, M. L. Multiautoclave with set of vessels for combinatorial synthesis of zeolites and other materials 20080124, 2008.
17. Wollman, P.; Leistner, M.; Stoweck, U.; Grunker, R.; Gedrich, K.; Klein, N.; Throl, O.; Grahlert, W.; Senkovska, I.; Dreisbach, F.; Kaskel, S., High-throughput screening: speeding up porous materials discovery. *Chem. Commun.* **2011,** *47*, 5151-5153.
18. Havrilla, G. J.; Miller, T. C., High-Throughput screening with micro-x-ray fluorescence. *Rev. Sci. Instrum.* **2005,** *76*, 062201.
19. Kubanek, P.; Schmidt, H. W.; Spliethoff, B.; Schüth, F., Parallel IR spectroscopic characterization of CO chemisorption on Pt loaded zeolites. *Micropor. Mesopor. Mater.* **2005,** *77*, 89-96.
20. Wang, H.; Liu, Z.; Shen, J., Quantified MS Analysis Applied to Combinatorial Heterogeneous Catalyst Libraries. *J. Comb. Chem.* **2003,** *5*, 802-808.

21. Moliner, M.; Díaz-Cabañas, M. J.; Fornés, V.; Martínez, C.; Corma, A., Synthesis methodology, stability, acidity, and catalytic behavior of the 18×10 member ring pores ITQ-33 zeolite. *J. Catal.* **2008,** *254*, 101-109.
22. Corma, A.; Díaz-Cabañas, M. J.; Moliner, M.; Martínez, C., Discovery of a new catalytically active and selective zeolite (ITQ-30) by high-throughput synthesis techniques. *J. Catal.* **2006,** *241*, 312-318.
23. Corma, A.; Serra, J. M.; Serna, P.; Valero, S.; Argente, E.; Botti, V., Optimisation of olefin epoxidation catalysts with the application of high-throughput and genetic algorithms assisted by artificial neural networks (softcomputing techniques). *J. Catal.* **2005,** *229*, 513-524.
24. Baumes, L. A.; Moliner, M.; Corma, A., Design of a Full-Profile-Matching Solution for High-Throughput Analysis of Multiphase Samples Through Powder X-ray Diffraction. *Chem. Eur. J.* **2009,** *15*, 4258-4269.
25. Baumes, L. A.; Moliner, M.; Nicoloyannis, N.; Corma, A., A reliable methodology for high throughput identification of a mixture of crystallographic phases from powder X-ray diffraction data *CrystEngComm* **2008,** *10*, 1321-1324
26. Chen, L.; Deem, M. W., Strategies for high throughput, templated zeolite synthesis. *Molecular Physics* **2002,** *100*, 2175-2181.
27. Treacy, M.; Rivin, I.; Balkovsky, E.; Randall, K.; Foster, M., Enumeration of periodic tetrahedral frameworks. II. Polynodal graphs. *Micropor. Mesopor. Mater.* **2004,** *74*, 121-132.
28. Woodley, S. M.; Catlow, R., Crystal structure prediction from first principles *Nat. Mater.* **2008,** *7*, 937-946.
29. Pophale, R.; Cheeseman, P. A.; Deem, M. W., A database of new zeolite-like materials. *Phys. Chem. Chem. Phys.* **2011,** *13*, 12407-12412.
30. Li, Y.; Yu, J.; Xu, R., Criteria for zeolite frameworks realizable for target synthesis. *Angew. Chem. Int. Ed.* **2013,** *52*, 1673-1677.
31. Zimmermann, N.; Salcedo Perez, J. L.; Haranczyk, High-Throughput Assessment of Hypothetical Zeolite Materials for Their Synthesizability and Industrial Deployability. **2019**.
32. Martínez-Franco, R.; Moliner, M.; Yun, Y.; Sun, J.; Wan, W.; Zou, X.; Corma, A., Synthesis of an extra-large molecular sieve using proton sponges as organic structure-directing agents. *Proc. Natl. Acad. Sci. U.S.A.* **2013,** *110*, 3749-3754.
33. Jiang, J.; Jordá, J. L.; Yu, J.; Baumes, L. A.; Mugnaioli, E.; Diaz-Cabañas, M. J.; Kolb, U.; Corma, A., Synthesis and Structure Determination of the Hierarchical Meso-Microporous Zeolite ITQ-43. *Science* **2011,** *333*, 1131.
34. Baumes, L. A.; Kruger, F.; Jimenez, S.; Collet, P.; Corma, A., Boosting theoretical zeolitic framework generation for the determination of new materials structures using GPU programming. *Phys. Chem. Chem. Phys.* **2011,** *13*, 4674-4678.
35. Li, J.; Qi, M.; Kong, J.; Wang, J.; Yan, Y.; Huo, W.; Yu, J.; Xu, R.; Xu, Y., Computational prediction of the formation of microporous aluminophosphates with desired structural features. *Micropor. Mesopor. Mater.* **2010,** *129*, 251-255.
36. Gao, N.; Li, J.; Li, J.; Kong, J.; Yu, J.; Xu, R., Syntheses and characterizations of aluminophosphate molecular sieves AFI guided by missing value estimation on database of aluminophosphate syntheses. *Micropor. Mesopor. Mater.* **2013,** *174*, 14-19.
37. Yu, J.; Xu, R., Rational Approaches toward the Design and Synthesis of Zeolitic Inorganic Open-Framework Materials. *Acc. Chem. Res.* **2010,** *43*, 1195-1204.
38. Jackson, R.; Catlow, C., Computer simulation studies of zeolite structure. *Mol. Simul.* **1988,** *1*, 207-224.
39. Sastre, G.; Cantin, A.; Diaz-Cabañas, M. J.; Corma, A., Searching organic structure directing agents for the synthesis of specific zeolitic structures: An experimentally tested computational study. *Chem. Mater.* **2005,** *17*, 545-552.
40. Moliner, M.; Serna, P.; Cantín, A.; Sastre, G.; Díaz-Cabañas, M. J.; Corma, A., Synthesis of the Ti-Silicate Form of BEC Polymorph of β-Zeolite Assisted by Molecular Modeling. *J. Phys. Chem. C* **2008,** *112*, 19547-19554.

41. Simancas, R.; Dari, D.; Velamazán, N.; Navarro, M. T.; Cantín, A.; Jordá, J. L.; Sastre, G.; Corma, A.; Rey, F., Modular Organic Structure-Directing Agents for the Synthesis of Zeolites. *Science* **2010,** *330*, 1219-1222.
42. Pophale, R.; Daeyaert, F.; Deem, M. W., Computational prediction of chemically synthesizable organic structure directing agents for zeolites. *J. Mater. Chem. A* **2013,** *1*, 6750-6760.
43. Davis, T. M.; Liu, A. T.; Lew, C. M.; Xie, D.; Benin, A. I.; Elomari, S.; Zones, S. I.; Deem, M. W., Computationally guided synthesis of SSZ-52: A zeolite for engine exhaust clean-up. *Chem. Mater.* **2016,** *28*, 708-711.
44. Schmidt, J. E.; Deem, M. W.; Davis, M. E., Synthesis of a Specified, Silica Molecular Sieve by Using Computationally Predicted Organic Structure-Directing Agents. *Angew. Chem. Int. Ed.* **2014,** *53*, 8372-8374.
45. Brand, S. K.; Schmidt, J. E.; Deem, M. W.; Daeyaert, F.; Ma, Y.; Terasaki, O.; Orazov, M.; Davis, M. E., Enantiomerically enriched, polycrystalline molecular sieves. *Proc. Natl. Acad. Sci. U. S. A.* **2017,** *114*, 5101-5106.
46. Daeyaert, F.; Ye, F.; Deem, M. W., Machine-learning approach to the design of OSDAs for zeolite beta. *Proc. Natl. Acad. Sci. U. S. A.* **2019,** *116*, 3413-3418.
47. Gallego, E. M.; Portilla, M. T.; Paris, C.; León-Escamilla, A.; Boronat, M.; Moliner, M.; Corma, A., "Ab initio" synthesis of zeolites for preestablished catalytic reactions. *Science* **2017,** *355*, 1051-1054.
48. Li, C.; Paris, C.; Martínez-Triguero, J.; Boronat, M.; Moliner, M.; Corma, A., Synthesis of reaction-adapted zeolites as methanol-to-olefins catalysts with mimics of reaction intermediates as organic structure-directing agents. *Nat. Catal.* **2018,** *1*, 547-554.
49. Gallego, E. M.; Paris, C.; Cantín, A.; Moliner, M.; Corma, A., Conceptual similarities between zeolites and artificial enzymes. *Chem. Sci.* **2019,** *DOI: 10.1039/C9SC02477H*.
50. Zhang, Y.; Ling, C., A strategy to apply machine learning to small datasets in materials science. *Npj Comp. Mater.* **2018,** *4*, 25.
51. Jensen, Z.; Kim, E.; Kwon, S.; Gani, T. Z.; Román-Leshkov, Y.; Moliner, M.; Corma, A.; Olivetti, E., A Machine Learning Approach to Zeolite Synthesis Enabled by Automatic Literature Data Extraction. *ACS Central Sci.* **2019,** *5*, 892-899.
52. Duvenaud, D.; Maclaurin, D.; Aguilera-Iparraguirre, J.; Gómez-Bombarelli, R.; Hirzel, T.; Aspuru-Guzik, A.; P., A. R., Convolutional Networks on Graphs for Learning Molecular Fingerprints. *In the Proceedings of Advances in Neural Information Processing Systems 28 (NIPS 2015)* **2015**, 2215-2223.
53. Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A., Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Sci.* **2018,** *4*, 268-276.
54. Gao, H.; Struble, T. J.; Coley, C. W.; Wang, Y.; Green, W. H.; Jensen, K. F., Using Machine Learning To Predict Suitable Conditions for Organic Reactions. *ACS Central Sci.* **2018,** *4*, 1465-1476.
55. Sanchez-Lengeling, B.; Aspuru-Guzik, A., Inverse molecular design using machine learning: Generative models for matter engineering. *Science* **2018,** *361*, 360-365.
56. Graves, A.; Wayne, G.; Danihelka, I., Neural turing machines. *arXiv preprint arXiv:1410.5401* **2014**.
57. Duan, Y.; Andrychowicz, M.; Stadie, B.; Ho, O. J.; Schneider, J.; Sutskever, I.; Abbeel, P.; Zaremba, W. I., One-shot imitation learning. *In Advances in neural information processing systems* **2017**, 1087-1098.
58. Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A., Machine learning for molecular and materials science. *Nature* **2018,** *559*, 547-555.