From isomorphism to polymorphism: connecting interzeolite transformations to structural and graph similarity

Daniel Schwalbe-Koda and Rafael Gomez-Bombarelli

Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139

(Dated: December 7, 2018)

Zeolites are nanoporous crystalline materials with abundant industrial applications. Despite sustained research, only 235 different zeolite frameworks have been realized out of millions of hypothetical ones predicted by computational enumeration. Structure-property relationships in zeolite synthesis are very complex and only marginally understood. Here, we apply structure and graph-based unsupervised machine learning to gain insight on zeolite frameworks and how they relate to experimentally observed polymorphism and phase transformations. We begin by describing zeolite structures using the Smooth Overlap of Atomic Positions method, which clusters crystals with similar cages and density in a way consistent with traditional hand-selected composite building units. To also account for topological differences, zeolite crystals are represented as multigraphs and compared by isomorphism tests. We find that fourteen different pairs and one trio of known frameworks are graph isomorphic. Based on experimental interzeolite conversions and occurrence of competing phases, we propose that the availability of kinetic-controlled transformations between metastable zeolite frameworks is related to their similarity in the graph space. When this description is applied to enumerated structures, over 3,400 hypothetical structures are found to be isomorphic to known frameworks, and thus might be realized from their experimental counterparts. Using a continuous similarity metric, the space of known zeolites shows additional overlaps with experimentally observed phase transformations. Hence, graph-based similarity approaches suggest a venue for realizing novel zeolites from existing ones by providing a relationship between pairwise structure similarity and experimental transformations.

I. INTRODUCTION

Zeolites are inorganic nanoporous materials with an extensive range of industrial applications. Their crystalline structure exhibits regular channels and pores, enabling their usage as selective catalysts, ion exchangers, and separators. Although all zeolite frameworks are metastable with respect to the denser quartz-like structure, they are very robust in harsh operation conditions of temperature and pressure. Because of this operational stability and affordable synthesis, zeolites are commonplace in the chemical, oil and refining industries. They also show much promise in novel green chemistry applications. Traditionally, zeolites have been synthesized by hydrothermal crystallization of aluminosilicate gels. Oftentimes the reactants include organic structure-directing agents (OSDAs) that drive the formation of a given framework, whose nanoporous lattice shape and charge distribution matches the organic molecule in a lock-and-key fashion. Recently, it has been shown that by carefully designing custom OSDAs for a specific chemical reaction, zeolitic catalysts can be tailor-made for a desired reaction. Yet, describing and controlling these synthetic mechanisms is still an open challenge. The multiple phases involved in the nucleation and crystallization of zeolites are very complex and hinder a priori design and descriptions. Alternative synthetic methods, such as the assembly-disassembly-reorganization (ADOR) process or topotactic transformations, favor the conversion of one zeolite structure into another targeted structure while avoiding OSDAs. However, it is unclear how to synthesize arbitrary frameworks using these transformations. In general, understanding and controlling synthetic mechanisms of zeolites are open challenges, and the access of novel frameworks and zeolite discovery remains largely tied to trial-and-error innovation.

From a theoretical point of view, there is no lack of candidates to be synthesized. By connecting tetrahedrally-coordinated TO₄ units (T = Si, Al, P, etc.), one can form millions of different topologies in a three-dimensional space. Databases with millions of zeolite frameworks enumerated by computational methods have been around for more than one decade. In contrast, up to date, a little over 230 zeolite frameworks have been recognized by the International Zeolite Association (IZA). Fewer than 10% of those are commercially used. The practical exploration of these computer-generated databases is mostly focused on classifying structures as feasible or not based on their building units, interatomic distances, tiling, kinetic or thermodynamic properties, framework density-energy relationships and similar metrics. Regardless of the chosen criterion, experimental exceptions are often observed, and most metrics cannot provide predictive insights on the structure-property relationship of zeolites. Other recent computational efforts have addressed the identification of new applications for known zeolites, screening hypothetical structures based on enumeration or performance factors, but hardly any computational
approaches have been proven to suggest candidates with high synthetic potential from hypothetical databases.

In this article, we present a method to explore the space of zeolites with unsupervised machine learning. Based on structural and graph descriptors, we derive similarities between zeolite topologies. We then postulate an explanation for observed transitions between zeolite phases through graph theory and kinetic arguments. Based on a database of hundreds of thousands of hypothetical zeolites, we present more than 3,400 crystals which could be accessed by such transformations. Finally, we employ a graph similarity tool as a way to explore the space of known and hypothetical zeolites.

II. RESULTS

A. Structural similarity

We start with the assumption that the structures of zeolites ultimately embed most important information about their feasibility [26]. Fundamentally unfeasible zeolites should have structural patterns inaccessible through standard kinetic and nucleation processes because of thermodynamic instability, kinetic traps, etc. On the other hand, fundamentally feasible zeolites are necessarily a superset of existing materials and should also have structural patterns that determine their accessibility. Since all known zeolites are feasible, we could use this fact to derive structures maximally similar to known ones in terms of framework topology. The crystallization processes in zeolites involve a series of transformations between metastable phases, according to Ostwald’s rule [27], often related by the framework density of the product [28-29]. The stability of these phases tends to increase with conversion towards denser structures [29], with the limiting case being quartz. It is known that the presence of certain substructures in the crystallization pathways can favor such interconversion mechanisms [29-31].

Typical ways to describe the construction of zeolites, therefore, use the concept of secondary building units (SBUs) [32]. These are arrangement motifs of tetrahedra observed throughout the crystalline structure of many zeolites. The increasing number of substructures in which zeolites can be divided, however, makes the use of SBUs cumbersome for analyzing the patterns within the structure. Other sets of building units have been adopted to contain that problem [33], but some zeolites cannot be described even when these sets of structures are adopted [19]. Nevertheless, composite building units (CBUs) are still a standard to communicate the structural features of a zeolite [35]. Their complexity and hierarchical analyses require topological analyses with human supervision or classification [16-20], rendering them inappropriate to purely autonomous data-driven approach [36]. Particularly, we are interested in an algorithmic way to assign a similarity between zeolite structures with different CBUs, space groups, unit cell sizes, framework densities, ener-

FIG. 1. t-SNE plot of known zeolites based on their SOAP fingerprint calculated with (a) Euclidean metric; and (b) kernel metric. The color indicates the framework density of the topology. Zeolites sharing CBUs tend to be clustered together in certain regions of the space.
in zeolite structures, however, proximity in this space does not imply that zeolites are necessarily made by the same building units. Overall, the distribution of known zeolites in the t-SNE plot correlates with the framework density of the materials; the Pearson correlation coefficient between the Euclidean distance of the SOAP fingerprints and the difference of framework densities between is 0.950. Most probably, this is a consequence of the averaging of SOAP fingerprints rather than a failure of the descriptor. By smearing out local influences of chemical environments, a mean global descriptor is obtained for each one of the crystal structures.

This descriptor is able to encode useful information about the chemical intuition embedded in zeolites, nevertheless. Vector descriptors can be compared pairwise with the Euclidean distance for a more quantitative approach. The closest pair in the SOAP space is AFR-SFO. Despite the fact that both zeolitic frameworks have different space groups, with orthorhombic and monoclinic unit cells, both crystals have similar framework densities and share the CBU \textit{sti} \[35\]. Topologically speaking, the building process of both structures are related solely by an inversion and mirror symmetries \[42\]. Furthermore, the crystallization of both zeolites in their hydrothermal synthesis depends on hydroxyl or fluoride ions. In the calcination stage, however, instead of being removed from the frameworks along with OSDAs, these ions become trapped in their structures. This process, analogous in both zeolites, creates a discrepancy in the sizes of the cages inside the framework. \[43\]-\[45\].

Other pairs among the closest ones include the SBS and SBT zeolites, with large cages and multidimensional channels \[46\]; SFE and SSY, which are formed by the expansion of MTT and TON \[47\]-\[48\] (also close in the SOAP space); AFS and BPH, sharing their characteristic CBUs; and so on.

The collapse of the SOAP fingerprint to the framework density is mainly a result of the differences between the norms of the vectors. As frameworks with higher densities tend to have more overlap between atomic positions within each chemical environment, the norm of the power spectrum tends to scale accordingly. To address this problem, the Euclidean distance was replaced by a cosine similarity, whose normalization mitigates the scaling issues induced by the $L^2$ distance. Under this metric, the similarity between crystal structures becomes proportional to the inner product of their vector representations. By evaluating pairwise combinations of known zeolites, we derive an alternative similarity matrix, giving rise to the t-SNE plot shown in Fig. 1b.

Density plays a smaller role on zeolites similarity as computed by the kernel over SOAP fingerprints. The Pearson correlation coefficient between the kernel distance and the difference of framework densities is 0.686. Second-order similarity factors can now be seen in the more scattered t-SNE plot. Neighboring zeolites are still surrounded by counterparts with same CBUs and approximately equal framework density, as depicted in Fig. 1b. The two SOAP-based metrics are still related, the Pearson correlation between them being 0.641.

A distance metric is an essential tool to explore zeolite space. By providing a quantitative measure of how (dis)similar two crystal structures are, it is possible to go beyond qualitative or classification analyses. The SOAP descriptors prove to be a useful tool not only to visualize the space of zeolites, but also to provide semiquantitative explanations on the synthesis of zeolites. The strong correlation of the framework density with the SOAP fingerprints, however, may limit the understanding of zeolites, even if a kernel metric is employed. Such emphasis on the role of density may overlook frameworks that are structurally similar, but with different densities. Alternative descriptions with complementary predictive power are necessary to provide further insights.

### B. Graph isomorphism

By averaging the power spectrum of chemical descriptors under the SOAP method, it is possible to assess the geometrical configuration of crystal structures in a way that is connected to building units and density. Nevertheless, important features of zeolites lie on how these units are connected to each other, giving rise to a unique topology. To account for this connectivity between atoms in a zeolite framework, we propose the use of graph theory to compare similarity between the structures. This approach has been used several times in the literature for enumerating the space of zeolites \[12\]-\[21\], determining unique topological descriptors \[49\] or the prediction of zeolitic surfaces using minimal cuts \[50\]. Recently, a general manner to generate crystal graphs which include periodic boundary conditions using multigraphs was proposed \[51\]. Here, we represent each T atom in a zeolite as a node in the graph. Two nodes are connected in the graph if there is an oxygen between their respective atoms in the crystal structure. Boundary conditions are included following the approach of Xie and Grossman \[51\].

Using graph theory, we can define similarities between crystals based on their crystal graphs. The simplest way to compare graphs is to identify graph isomorphism between the crystal graphs, i.e. if they are equivalent up to a relabeling of their nodes \[52\]. To avoid confusion with the crystallographic concept of isomorphism, we herein refer to crystal graph isomorphism (CGI) if two crystals have isomorphic graph representations.

Starting from the set of known zeolites from the IZA database, we check for the existence of CGI for all pairwise combinations of known zeolites. If the graph representations were distinct, no isomorphism would be found. The isomorphism test returned 14 pairs of different zeolite topologies with the same multigraph connectivity when periodic boundary conditions are considered. A trio of zeolite topologies which have the same graph representation was also found. Table \[1\] shows the isomor-
Interpolation between crystal structures of GME-AFI (above) based on the equivalence of their atoms within the CGI (below). One extra period is added to the $c$ direction to better illustrate the reconstructive phase transformation. Atoms, bonds and reconstructed (or broken) bonds are depicted with red dots, solid and dashed lines, respectively. The unit cell, depicted in blue, shrinks in the $c$ direction and expands in the $a$ and $b$ directions during the transformation.

TABLE I. Pairs of known zeolites which are multigraph isomorphic. Bold pairs share the same secondary building units.

| CDO-FER | SFO-AFR | RSN-VSV | AWO-UEI |
|---------|---------|---------|---------|
| AFI-GME | AHT-ATV | CGF-HEU | JBW-NPO |
| ABW-BCT | AWW-RTE | APC-APD | BOF-LAU |
| MER-PHI | SBN-THO | ACO-GIS-LTJ |

phic pairs found among the known zeolites. A striking feature among these pairs is that only four of these pairs share the exact same set of CBUs, namely CDO-FER, SFO-AFR, RSN-VSV, and AWO-UEI. All the remaining pairs share the same connectivity between atoms inside and across the unit cell with their counterparts, but have different building schemes in terms of CBUs.

Many of the pairs in Tab. I are connected through known phase transformations or appear together as polymorphs in synthetic processes. Among the pairs identified, the most reported transformation is from CDO to FER. Interconversion between both frameworks include naturally or induced synthetic processes \[53-55\], as well as reassembly \[56-58\]. One can be obtained by the other just by a relative dislocation of the precursor layers \[59, 60\]. Both APC-APD \[61\] and CDO-FER are related by a topotactic transformation, recognized by the relationship between crystallographic symmetries of the precursor layers and the final crystal structure \[62\]. The stacking pattern for these specific layers preserves the graph isomorphism even with the rearrangement. A favorable kinetic transition benefits from this fact, so for some of these layered silicates, no bonds have to be broken nor rearranged in the condensation process \[6\].

Not all observed transformations between graph isomorphic pairs in Tab. I occur in the context of condensation of layered precursors, however. The conversion between GME and AFI is an example of an isochemical phase transformation with breaking and reforming of T-O bonds. Alberti et al. studied this reconstructive phase transformation under the increase of temperature and identified the existence of an intermediate, “transient” phase with three-connected T atoms \[63\]. Dusseiller et al. later explained this mechanism using powder X-ray diffraction (XRD) pattern measurements, verifying a number of 18 T-O bonds breaking under compression of the gme cage \[64\]. Using the graph isomorphism criterion and its equivalence between nodes, we explain this same phase transformation by proposing a kinetic criterion derived from the node and edge equivalences. Even though the kinetic process involves bond breaks, the net number of bonds formed per unit cell is zero. This symmetry allows for smaller barriers for these zeolite interconversions to happen. We can visualize such phase transition between both structures by interpolating the equivalent atomic positions from each crystal. Fig. 2 depicts the evolution of the GME-AFI transformation, compatible with both the three-connected intermediate and cage compression mechanisms \[63, 64\]. Similar mechanism of reconstructive phase transition have been proposed for the pair ABW-BCT \[65\]. The ABW zeolite has an abw cage, whose bonds could be broken to form the more compact lau cage in the BCT framework while keeping the graph isomorphism, also increasing the framework density in the process. Even though both zeolites have been synthesized with different compositions, this analysis suggests the topology of the tectosilicate BCT is accessible for the synthetic lithium zeolite ABW, just like the AFI zeolite can be formed from the GME framework.

Finally, some pairs related by the graph isomorphism criterion appear as competing phases in the synthesis of zeolites. Seed-assisted synthesis of zeolite MER from PHI seeds \[66-68\] or experimental similarities in the synthesis conditions \[69, 72\] were already reported, even though they do not share the same CBUs \[73\]. Likewise, the discovery of the RSN topology was also associated to the
To the best of our knowledge, some of the pairs in Tab. I have not been jointly reported in experimental results. However, some share synthetic and topological commonalities, not only such as the aforementioned pair AFR-SFO [42, 44, 45, 75], but also in the case of AWO-UEI [76, 77], SBN-THO [78, 79], and ACO-GIS [80]. Other graph isomorphic pairs have not been analyzed in the literature before, partly because the realized materials have quite different compositions. The reduced practical interest in some of these zeolite topologies may also have occluded the simultaneous apparition of both phases in experimental investigations, or the possibility of the direct interconversion.

Based on these experimental reports, we conjecture that the isomorphism test suggests the existence of kinetic phase transitions between certain zeolite frameworks. Such transformations between zeolite structures related by CGI imply that both the initial and final structures are related by a simple geometrical rearrangement of the atoms. The perfect equivalence between the graphs ensures both structures have the same number of atoms and bonds inside the unit cell. In addition, a bijection between nodes is guaranteed by the graph isomorphism: for each atom in one crystal, there is one equivalent atom in the final crystal with the exact same neighborhood. This suggests the transformation between metastable structures is kinetically favorable and can be satisfied with small displacements of atoms within the boundary conditions, even though new sets of building units have to be formed.

If the isomorphism between these structures is an indicator that they may be synthetically related, then new frameworks could potentially be accessed from known ones as parent structures, regardless of the final and initial CBUs. To investigate this possibility, we look for isomorphic pairs in hypothetical zeolite databases [13, 14] which have energy above quartz (see Methods). When all pairwise combinations of 269,515 enumerated structures and the known zeolites are assessed with the isomorphism criterion, 3,446 different hypothetical frameworks are found to have a known isomorphic counterpart. Based on previous evidence of accessibility, we then suggest these structures could be the first candidates to explore the space of new zeolites. Fig. 3 shows how these hypothetical zeolites relate to known ones in terms of frequency of isomorphism. The high number of polymorphs isomorphic to the APC/APD, DON, SAF, AWO/UEI, ATV/AHT and SIV may be related to the richness of graph-equivalent topologies that can be constructed using an orthorhombic unit cell within the search parameters used by Deem et al. [13] to design the zeolite cells. Furthermore, all of these crystals are better characterized by chains instead of just building blocks, whose regularity may play a role in the definition of the graph topology. In Fig. 3b, we present the hypothetical zeolite
PCOD8128559, which is graph isomorphic to the ATV zeolite. Despite their structural similarity, the distribution of rings are not exactly corresponded. Furthermore, PCOD8128559 has a slightly higher framework density (19.8 T atoms/1000 Å³) when compared to ATV (18.9 T atoms/1000 Å³).

The spirit behind combinatorial approaches is to exhaust a given space, and thus theoretical databases would aspire to recover all known zeolite topologies. Deem et al. reported finding 119 known zeolites within their SLC-optimized dataset with energy above quartz [13]. Their detection method is based on comparing XRD patterns of hypothetical zeolites with known ones, giving a reliable method to assess crystal equivalence. We used the isomorphism test for the same task, and were able to recover 89 of the 119 hits reported by those authors. The discrepancy arises because the size of the graph also influences the result of the isomorphism test. Therefore, the choice of the unit cell for a crystal graph is relevant to compare graphs according to the isomorphism criterion. Since some of the optimized structures assigned to known lattices are assigned to a supercell and not to their primitive unit cell, their multigraphs are a scaled up version of the structure provided by IZA. Fig. 2 illustrates this fact with the example of hypothetical zeolite PCOD8156534. Although its counterpart is ATV, this mapping was not detected due to the choice of the unit cell. To address this, we expanded the original primitive cell for every zeolite topology by up to a factor of two in all directions, augmenting the database of unit cells by 9 times its original size. This expanded database allowed us to find 115 out of the 119 zeolites reported by Deem et al. [13]. Furthermore, when we look for zeolite topologies beyond the correspondence provided by Deem et al., we detect 144 unique frameworks equivalent to those found in known zeolite unit cells. From these 144 graphs, 40 are graphs which have not been detected by them. At the same time, the isomorphism test was unable to find 15 topologies retrieved by those authors. The use of supercells rather than primitive cells, as well as reassignments of periodic boundary conditions in the enumerated database, lead to this discrepancy. The absence of nearly 100 IZA-recognized structures from the hypothetical database suggests that rule-based pruning approaches may have inadvertently removed valid structures from the enumerated library. This is likely because their predicted formation energy was too high. This could be due to inaccuracies in the theoretical methods, and could be improved by more accurate potentials. If the calculations are indeed accurate, the implication is that high-energy zeolite phases can still be metastable and kinetically accessible. These absences further strengthen arguments about avoiding hard, hand-made rules in favor of data-driven approaches.

C. Graph similarity

If comparing graphs based on isomorphism limits the analysis to a binary measure, a continuous similarity measure between graphs can relax this constraint. Possibilities may include many different graph kernels. In this article, we adapted the D-measure [81] for usage in multigraphs. This similarity compares the connectivity between graphs based on the node distances, generating a metric space that recovers the results from isomorphism tests with minimum number of false positives [81].

When the description power of D-measure is compared to SOAP, we observe that both representations are complementary. While the latter tends to correlate with framework density, the D-measure also assigns dissimilarity between different graphs based on their size, as seen in Fig. 4. The Pearson correlation coefficient between the D-measure and the ratio between the sizes of the graphs is equal to -0.803, suggesting that ratios closer to 1 lead to smaller dissimilarities. However, the connectivity of both graphs also play a role in defining the dissimilarity between two different crystals, as crystals with different sizes can have higher dissimilarities. Based on all possible pairs of known zeolites, we can use this fact to draw relationships between crystals from the IZA database. Fig. 4 plots the joint distribution of zeolite pairs according to both criteria. There is no significant correlation between dissimilar crystals according to the metrics under comparison. When we select crystals with normalized distance lower than 0.1 in both criteria, 174 pairs of zeolites are found. 11 out of the 14 pairs reported on Tab. 1 are included in this range. The other three are ABW-BCT, AFI-GME and SBN-THO. Of the trio, the pairs ACO-LTJ and GIS-LTJ were not found within this range. The difference of framework density leads to a higher SOAP distance, even though their graph dissimilarity is zero. Using the fact that ABW-BCT and AFI-GME are related by a reconstructive phase transition, this result is compatible with the increased SOAP distance and can indicate that analogous isomorphic pairs may also be related by such a transition. Since both measures together address structural and connectivity issues in phase transformations, we suggest that pairs close in both spaces could be favored in terms of phase transformations.

Finally, D-measure is applied to analyze the similarities between hypothetical and existing materials in a continuous window, allowing a richer interpretation than a binary metric such as CGI. Based on the distribution of distances to hypothetical frameworks, we can detect which known frameworks have denser environments within the hypothetical zeolite database. Fig. 5 shows the distribution of pairwise distances between every known IZA framework and enumerated hypothetical zeolites. The size of the graphs distinguish some frameworks from the theoretical database. Bigger frameworks, such as PAU, TSC and MWF, as well as smaller ones, such as DFT, EDI, JBW and NPO, have less closer neighbors than most of the zeolites. This inhomogeneity suggests the
space of hypothetical zeolites, as selected in [13], has “gaps”, making some frameworks more distant than all others within this space.

**DISCUSSION**

Similarity metrics based on multigraphs, both binary isomorphism test and continuous distance metrics, suggest that zeolite structures with similar connectivity may be related through topotactic, reconstructive phase transformations, and also through the formation of competing phases. Since zeolites are metastable structures, the existence of kinetically-controlled pathways is crucial for their synthesis. The transitions are kinetically favorable if no net chemical bonds have to be broken to rearrange the framework geometry or if they are concertedly broken and reformed. This enables transitions regardless of building units. We thus postulate that novel frameworks can be accessed through transformation of existing isomorphic structures, possibly in combination with novel OSDAs.

The use of a continuous distance metric also allows us to extend the analysis to transformations where a small number of net bonds are formed or broken but where most tetrahedra retain their connectivities. Furthermore, the graph method is atom-agnostic and can be applied not only to these frameworks, but also expanded to metal-organic frameworks and other materials with interesting polymorphism.

For this analysis we have assumed all zeolite to be the ideal pure silicate phase, and that all information derives from their structures. In practice, some isomorphic pairs have been realized with very different composition. The crystallization rate of aluminosilicates in the presence of OSDAs is very sensitive to the Si/Al ratio, but interzeolite transformations may be governed by different rules and allow an access new region of composition space for known frameworks.

We have also re-encountered previous results suggesting that many known materials are missing from enumerated databases, and thus that their coverage is not complete. These “gaps” in the space of enumerated zeolites are possibly due to the combination of hand-made rules and semiclassical simulation, and may ruling out important candidates in the selection process.

**Limitations and further steps**

Our proposed approaches bypass complex CBU or tiling enumerations to capture relevant structure-property relationships. Yet, they are not perfect metrics: graph isomorphism is purely binary and cannot establish intermediate similarities; SOAP overestimates the influence of framework density and local structure and D-measure is overly dependant on graph size, to the extent that it fails to detect supercell and primitives as related graphs. Novel ways to convert graphs to continuous variables, using unsupervised machine learning, or inspired by quantum computers could provide richer and more predictive metrics.
Furthermore, the proposed CGI model does not explain all known phases among zeolites, much less all crystalline transformations. We hypothesize that it is particularly applicable to metastable systems, such as zeolites, whose synthesis and interphase transformations are driven majorly by kinetic considerations. In these, the existence of concerted low-barrier transformations in the unit cell can be explained by graph similarity of reactant and products. Other processes, such thermodynamically-driven transformations, or direct assembly of building blocks from solution may be better explained through other features, such as traditional CBU reassembly.

Further analysis on how crystal graph similarity is related to phase transformations/competitions is ongoing. Further research is underway to validate these hypothesis and to identify how new OSDAs can be combined with interzeolite conversion between graph isomorphic structures.

**METHODS**

**Database of zeolites**  A database of 230 known zeolite frameworks was downloaded from the Database of Zeolite Structures, kept by the Structure Commission of the International Zeolite Association (IZA) [15]. No partially disordered frameworks were included in the database. The hypothetical zeolite database was generated and optimized using the Sanders-Leslie-Catlow (SLC) force field [82] by Deem et al. [13][14]. From the complete database with about 314k entries, we remove those whose energies are lower than quartz, and end up with 269,515 zeolite structures.

**Construction of zeolite multigraphs** We start from the multigraph representation of crystals satisfying periodic boundary conditions [31]. To maximize the amount of information embedded in the graph and minimize its size, we remove the oxygen atoms from the graph. Each
crystal graph contains as many nodes as T atoms in its input unit cell. We then consider only silicon atoms as nodes for the graph which are connected if a硅氧原子 exists. This avoids the usage of nearest-neighbors search using Voronoi diagrams for porous materials.

**Comparing crystal graphs and structures** Multi-graph isomorphism is performed using the VF2 algorithm [32] as implemented in NetworkX [63]. The graph similarity D-measure is implemented as reported by Schieber et al. [52].

To describe crystal structures within the SOAP formalism, for each atomic environment \( \chi_i \) in the crystal structure, a power spectrum \( p(\chi_i^*) \) is calculated [37, 84] using \( r_{\text{cut}} = 10 \), radial basis size of 8 with \( L_{\text{max}} = 5 \) as hyperparameters. Instead of adopting the regularized entropy match kernel [38] to compare crystal structures, however, we opt for the average SOAP fingerprint given by

\[
p(Z) = \frac{1}{N} \sum_i p(\chi_i^Z) \quad (1)
\]

for each crystal structure \( Z \). This allows us to generalize the analysis to hundreds of thousands of zeolites without incurring into excessive computational cost.

The Euclidean distance between two crystal structures using the SOAP fingerprint is simply the \( L^2 \) norm of the difference between their power spectra,

\[
d_{\text{Eucl}}(A, B) = \| p(A) - p(B) \|_2. \quad (2)
\]

On the other hand, the unnormalized average structure kernel \( K \) can be defined as [38]

\[
K(A, B) = p(A) \cdot p(B), \quad (3)
\]

leading to the normalized kernel

\[
\tilde{K}(A, B) = \frac{K(A, B)}{\sqrt{K(A, A)K(B, B)}}. \quad (4)
\]

This kernel induces a metric \( d_{\text{kernel}} \) given by [38]

\[
d_{\text{kernel}}(A, B) = \sqrt{2 - 2\tilde{K}(A, B)}. \quad (5)
\]

**Acknowledgements**

DSK acknowledges the MIT Nicole and Ingo Wender Fellowship and the MIT Presidential Graduate Fellowship for financial support. RGB thanks MIT DMSE and Toyota Faculty Chair for support.

---

References:

[1] M. E. Davis, *Nature* **417**, 813 (2002).
[2] S. F. Abdo and S. T. Wilson, *Zeolites in Catalysis: Properties and Applications*, 310 (2017).
[3] Y. Li, L. Li, and J. Yu, “Applications of Zeolites in Sustainable Chemistry,” (2017).
[4] C. S. Cundy and P. A. Cox, *Chemical Reviews* **103**, 663 (2003).
[5] C. S. Cundy and P. A. Cox, “The hydrothermal synthesis of zeolites: Precursors, intermediates and reaction mechanisms,” (2005).
[6] E. M. Gallego, M. T. Portilla, C. Paris, A. León-Escamilla, M. Boronat, M. Moliner, and A. Corma, *Science* **355**, 1051 (2017).
[7] C. R. A. Catlow, *Interdisciplinary Science Reviews* **40**, 294 (2015).
[8] W. J. Roth, P. Nachtigall, R. E. Morris, P. S. Wheatley, V. R. Seymour, S. E. Ashbrook, P. Chlbna, L. Grajciar, M. Polozij, A. Zukal, O. Shvets, and J. Čejka, *Nature Chemistry* **5**, 628 (2013).
[9] B. Marler and H. Gies, *European Journal of Mineralogy* **24**, 405 (2012).
[10] P. Eliášová, M. Opanasenko, P. S. Wheatley, M. Shamzhy, M. Mazur, P. Nachtigall, W. J. Roth, R. E. Morris, and J. Čejka, *Chemical Society Reviews* **44**, 7177 (2015).
[11] C. Li, M. Moliner, and A. Corma, *Angewandte Chemie International Edition* (2018), 10.1002/anie.201711422.
[12] M. Treacy, I. Rivin, E. Balkovsky, K. Randall, and M. Foster, *Microporous and Mesoporous Materials* **74**, 121 (2004).
[13] M. W. Deem, R. Pophale, P. A. Cheeseman, and D. J. Earl, *The Journal of Physical Chemistry C* **113**, 21353 (2009).
[14] R. Pophale, P. A. Cheeseman, and M. W. Deem, *Physical Chemistry Chemical Physics* **13**, 12407 (2011).
[15] Ch. Baerlocher and L.B. McCusker, http://www.iza-structure.org/databases/.
[16] P. Guo, N. Yan, L. Wang, and X. Zou, *Crystal Growth & Design* **17**, 6821 (2017).
[17] Y. Li, J. Yu, and R. Xu, *Angewandte Chemie International Edition* **52**, 1673 (2013).
[18] J. Lu, L. Li, H. Cao, Y. Li, and J. Yu, *Physical Chemistry Chemical Physics* **19**, 1276 (2017).
[19] V. A. Blatov, G. D. Ilyushin, and D. M. Proserpio, *Chemistry of Materials* **25**, 412 (2013).
[20] E. D. Kuznetsova, O. A. Blatova, and V. A. Blatov, *Chemistry of Materials* **30**, 2829 (2018).
[21] M. D. Foster, A. Simperler, R. G. Bell, O. D. Friedrichs, F. A. A. Paz, and J. Klinowski, *Nature Materials* **3**, 234 (2004).
[22] M. Mazur, P. S. Wheatley, M. Navarro, W. J. Roth, M. Polozij, A. Mayoral, P. Eliášová, P. Nachtigall, J. Čejka, and R. E. Morris, *Nature Chemistry* **8**, 58 (2016).
APPENDIX

FIG. A1. Distributions of SOAP fingerprint Euclidean distances between each pairwise combination of a known zeolite with a hypothetical one. All distances and frequencies are normalized. The color indicates the average distance between all neighbors.