Graph similarity drives zeolite diffusionless transformations and intergrowth

Graph similarity drives zeolite diffusionless transformations and intergrowth

Daniel Schwalbe-Koda,[1, a)] Zach Jensen,[1, b)] Elsa Olivetti,[1, c)] and Rafael Gómez-Bombarelli[1, d)]
Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139

(Dated: 13 August 2019)

Predicting and directing polymorphic transformations is a critical challenge in zeolite synthesis[1–3]. Interzeolite transformations enable selective crystallization[4–7], but are often too complex to be designed by comparing crystal structures. Here, computational and theoretical tools are combined to both exhaustively data-mine polymorphic transformations reported in the literature, and analyze and explain interzeolite relations. It is found that crystallographic building units are weak predictors of topology interconversion and insufficient to explain intergrowth. By introducing a supercell-invariant metric that compares crystal structures using graph theory, we show that diffusionless (topotactic and reconstructive) transformations occur only between graph-similar pairs. Furthermore, all known instances of intergrowth occur between either structurally-similar or graph-similar frameworks. We identify promising pairs for realizing diffusionless transformations and intergrowth, with hundreds of low-distance pairs identified among known zeolites, and thousands of hypothetical frameworks connected to known zeolites counterparts. The theory may enable the understanding and control of zeolite polymorphism.

Traditionally, zeolites are compared according to their framework density[8] or set of constituent units[9]. It is typically understood that the crystallization of certain species is faster when the seed and the product zeolite share the same framework[10,11]. During interzeolite conversion, the crystallization rate can be increased by the presence of similar composite building units (CBUs)[4,7,12], layered precursors[6,13,14] or inorganic agents[15]. The increased stability of denser frameworks is also usually described in terms of Ostwald's rule[8]. This has led to a formulation of heuristic rules such as the common-CBU hypothesis to design organic structure-directing agents (OSDAs)-free routes through hydrothermal treatments[4,12].

Most interzeolite transformations reported in the literature can be described as recrystallization under hydrothermal treatment. Other less common types of zeolite conversion have not been as thoroughly compiled and investigated. We searched through more than 70,000 articles related to zeolites with a combination of natural-language processing and human supervision and identified 391 experimental reports of polymorphic pairwise relations. We classified these pairs into four major groups of pairwise structural relationships: recrystallization, diffusionless transformations, formation of competing phases, and intergrown phases (see Fig. 1a-d and the Supplementary Information for full definitions).

By statistical analysis of the collected pairs, it was found that neither the framework density nor the common-CBU hypothesis by themselves explain the reported phase relations. Fig. 1e shows the tally of transformations reported between zeolite pairs that share a given number of common CBUs. At least 35% of recrystallization, competing and diffusionless relations have initial and final zeolites without any common CBUs. In contrast, 95% of the unique intergrown zeolites pairs have at least one common building unit. Still, nearly 65% of these pairs do not share the same set of CBUs (see Supp. Fig. 1). This data-driven view suggests that the common-CBU rule is not a predictor of interzeolite transitions, despite their roles as topological descriptors and influence on crystallization rates. Common CBUs only partially drive intergrowth.

Absolute changes in framework densities (ΔFD) do not provide compelling trends either. Similarity in density is very common, as evidenced by the skew in the distribution of all pairwise density differences in Fig. 1f. Recrystallization processes and competing phases show broader distribution for ΔFD than other transformations, as low-density frameworks such as FAU or LTA are usually crystallized first in hydrothermal synthesis. Still, 25% of the recrystallizations convert higher to lower density frameworks, contrary to expectations based on Ostwald's rule. On the other hand, more direct diffusionless transformations and intergrown zeolites tend to have smaller differences of framework density. With the exception of the CDO-FER transformation, which is mediated by low density CDO precursors, all diffusionless transformations happen towards denser polymorphs.

The presence of one or more common CBUs or low ΔFD, thus, can help rationalize results, but is not predictive of interzeolite relations. We propose that descriptors based on topology and structure can achieve improved explanatory and predictive power for some classes of experimental relations. Graphs are well-known representations to encode the topology of zeolites[16–19]. Frameworks are represented as multigraphs, or crystal graphs, that label T-O-T (T = Si, Al etc.) covalent bonds as edges and

[a)] ORCID: https://orcid.org/0000-0001-9176-0854
[b)] ORCID: https://orcid.org/0000-0001-7635-5711
[c)] ORCID: https://orcid.org/0000-0002-8043-2385
[d)] Electronic mail: rafagb@mit.edu; ORCID: https://orcid.org/0000-0002-9495-8599

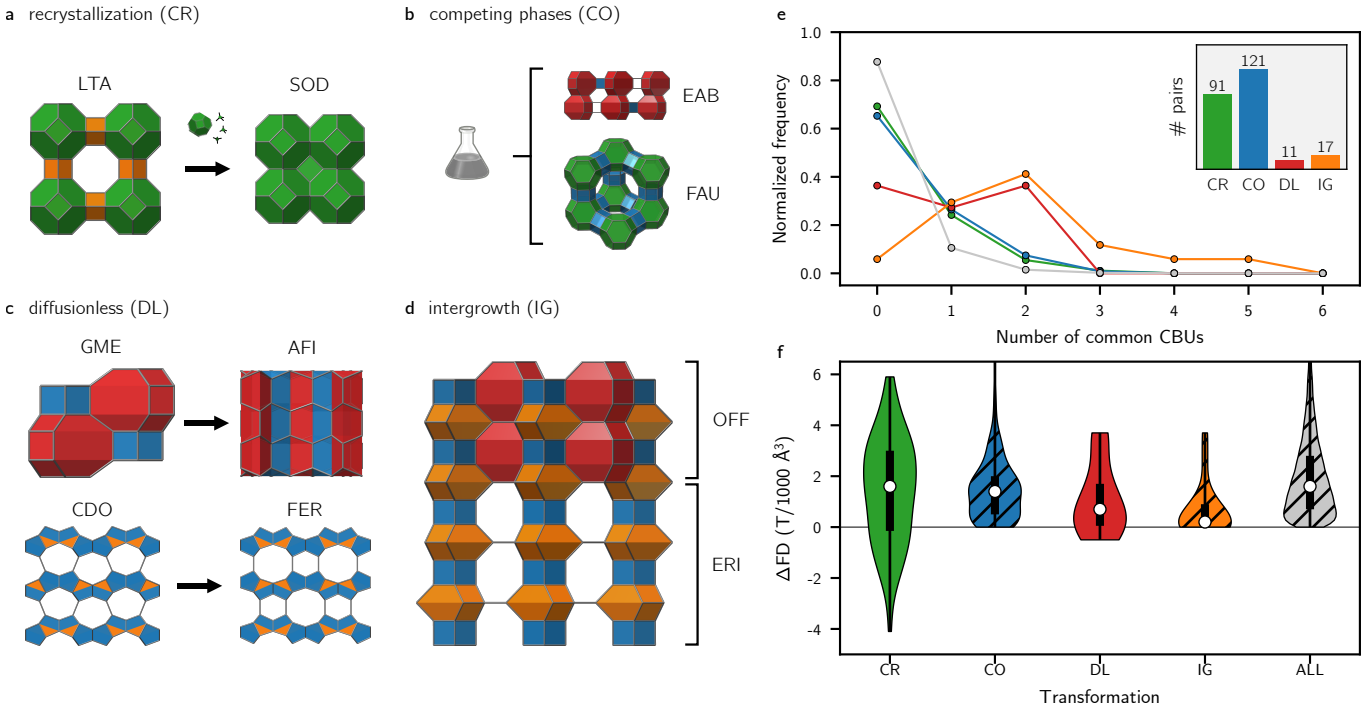

FIG. 1. **Types of zeolite transformations and classical explanations. a-d**, Diagrams of interzeolite transformations found in the literature. Colored substructures depict CBUs. **e**, Fraction of pairs experimentally observed within interzeolite transformations and their number of common CBUs. **e (inset)**, Histogram of the literature extraction. The number above the bar corresponds to the number of unique pairs found in the literature under that category. **f**, Statistical distribution of differences of framework density (ΔFD) between the source and the product zeolite for each transformation. The white dot indicates the median of the distribution. The bottom and top of the thicker solid lines are the first and third quartiles, respectively, while the whiskers indicate the maximum and the minimum of the range. For **e**, **f**, the zeolite conversions are abbreviated as following: recrystallization (CR), competing phases (CO), diffusionless (DL), and intergrowth (IG). For comparison, the distributions for all pairs of known zeolites regardless of experimental relation (ALL) are also shown. As the notion of source and product does not apply for competing phases and intergrown frameworks, only the absolute value of ΔFD is reported in the hatched plots.

capture periodic boundary conditions (Fig. 2a). To compare different topologies, a metric of distance between graphs is necessary. We first applied graph isomorphism, which verifies if two graphs are equivalent up to a relabeling of their nodes. Starting from a set of 245 known zeolites from the International Zeolite Association (IZA) database, we checked for the existence of crystal graph isomorphism for all 29,890 pairwise combinations of zeolite graphs. If all graph representations were distinct, no isomorphism would be detected. However, we found 14 pairs and one trio of different zeolite topologies with the same multigraph connectivity in their unit cell. Table I shows the isomorphic pairs identified among the known zeolites. 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 have identical graph representations, but different CBUs.

The isomorphism test hints at the presence of kinetic transformation channels between crystal phases[20]. Indeed, three of the graph isomorphic pairs are related experimentally through diffusionless transformations: CDO-FER, GME-AFI, and APC-APD. The perfect equivalence between graphs ensures that both structures have the same number of atoms and bonds inside the unit cell, so there is no net bond breaking or formation. A bijection between nodes is also guaranteed: for each atom in the starting crystal, there is an equivalent atom in the final crystal with the same neighborhood. The transition between the frameworks is either purely displacive (no covalent bonds are broken) or concerted (military), resembling martensitic transitions in metallurgy[21]. In contrast, hydrothermal treatments lead to the formation of an amorphous phase, essentially breaking down the topology and rebooting the crystallization process. This implies a less selective transformation if not assisted by OSDAs. Other experimental parameters may have a larger influence on the outcomes

TABLE I. Pairs of known zeolites whose crystal graphs are multigraph isomorphic. Bold pairs share the same set of CBUs.

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	

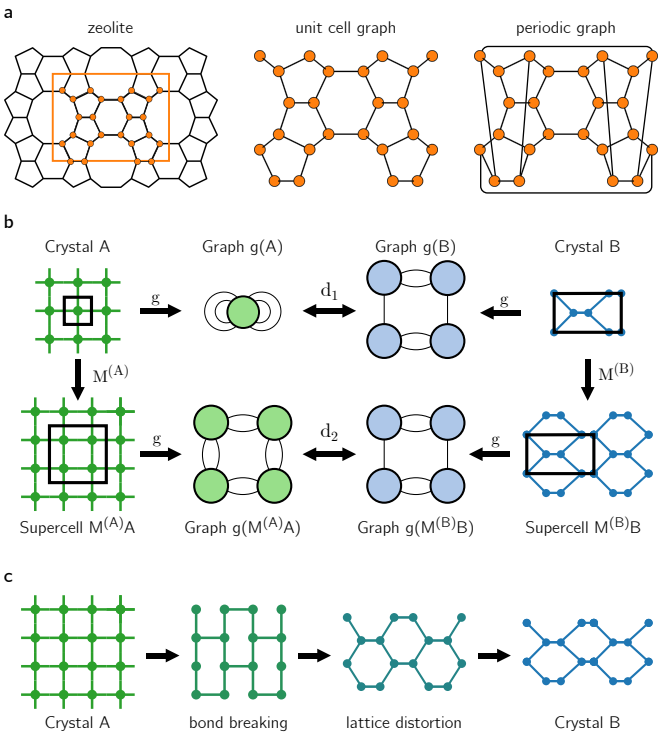

FIG. 2. **Graph and supercell matching.** **a**, Representation of a zeolite using a graph. The unit cell graph is modified to satisfy periodic boundary conditions by looping bonds back into the unit cell. **b**, Graph distance between different hypothetical crystal structures. The distance d between crystal graphs varies with the choice of the crystallographic unit cell. In the given example, the transformation matrices are $M^{(A)} = 2\mathbb{I}$ and $M^{(B)} = \mathbb{I}$, with $\mathbb{I}$ the identity matrix. With the choice of an appropriate metric, $d_2 \leq d_1$. **c**, Hypothetical A-B transformation. The bond breaking step removes two extra edges from the $M^{(A)}A$ crystal graph to match the $M^{(B)}B$ crystal graph, and is followed by a diffusionless transformation at constant graph.

than the precrystallized reactants themselves, as in the case of competing phases.

The relationship between graph isomorphism and interzeolite conversions is illustrated with the isochemical phase transition between GME and AFI. Alberti *et al.* and Dusselier *et al.* identified an intermediate, "transient" phase with three-connected T atoms resulting from the compression of the *gme* cage upon heating[22,23]. Using the graph isomorphism criterion, this transition is explained through node and edge equivalences. Although the kinetic process involves bond breaking, the net number of bonds formed per unit cell is zero. Such a mechanism between the structures can be visualized by interpolating the equivalent atomic positions of each crystal. Supp. Fig. 2 depicts the evolution of the GME-AFI transformation, compatible with both the three-connected intermediate and cage compression mechanisms[22,23].

Crystal graph isomorphism has three obvious limitations as a similarity metric for crystals: (i) it is a binary metric; (ii) it is a computationally expensive test for large graphs; and (iii) it is not invariant to the choice of the unit cell. To address (i) and (ii), we adapted the D-measure[24] for multigraphs. This similarity compares graph connectivities based on distributions of node distances, generating a continuous metric space which recovers graph isomorphism with zero distance[24] and has lower computational cost. Then, a variational approach is proposed to solve (iii). When comparing a zeolite pair, we search for two supercells with equal number of T atoms and minimum discrepancy in terms of lattice geometry (see Methods). This allows the comparison between crystals with different point groups, as is the case of most transformations in the literature (Supp. Tab. I). The topological distance between two crystals is then taken as the D-measure between the best-matched supercell graphs (Figs. 2b,c). Furthermore, to investigate the role of 3D atomic arrangement, we combined structural similarities under a single descriptor by using the Smooth Overlap of Atomic Positions (SOAP) approach[25]. We observed that kernel distance is well correlated with framework density and with CBUs (see Suppl. Figs. 3-6).

Fig. 3a illustrates the distribution of the pairwise distances between zeolites according to the two descriptors. When plotting all known interzeolite synthesis relations from our database, all diffusionless transformations fall in the lower graph distance region (Figs. 3b-d). This advances the argument that graph similarity is a powerful predictor of kinetic transformations in zeolites. The only exception to this rule is the reconstructive LTA-IFY transformation[26], which reportedly requires an external stress of 3 GPa to be induced. This very high stress suggests that diffusionless transitions between low graph-similar zeolites require extreme conditions. A large number of intergrown frameworks also occur for low graph-distance pairs. Interestingly, the SOAP descriptor complements the predictions of zeolite intergrowth, by capturing intergrown pairs with high structural similarity. This indicates the intergrowth may be possible if: the two polymorphs have similar atomic connectivities, resulting in low strain energy when similar free surfaces are put in contact (low graph distance); or if they have similar building units, through which most stacking faults can be formed (low SOAP distance). For recrystallization transformations and competing phases, the majority of the pairs have higher dissimilarities across both metrics.

Since the topological similarity between two structures is a strong indicator that they may be related by intergrowth or diffusionless transformations, new frameworks could be accessed from known ones as parent structures. We looked for isomorphic pairs in hypothetical zeolite databases[27] which have energy above quartz and assessed pairwise combinations of 269,515 enumerated structures and 245 known zeolites with the isomorphism criterion. 3,879 different hypothetical frameworks are found to be isomorphic to at least one of 150 known structures. Of those theoretical structures, 2,096 have density higher than their isomorphic known zeolite. Fig. 3e and Supp.

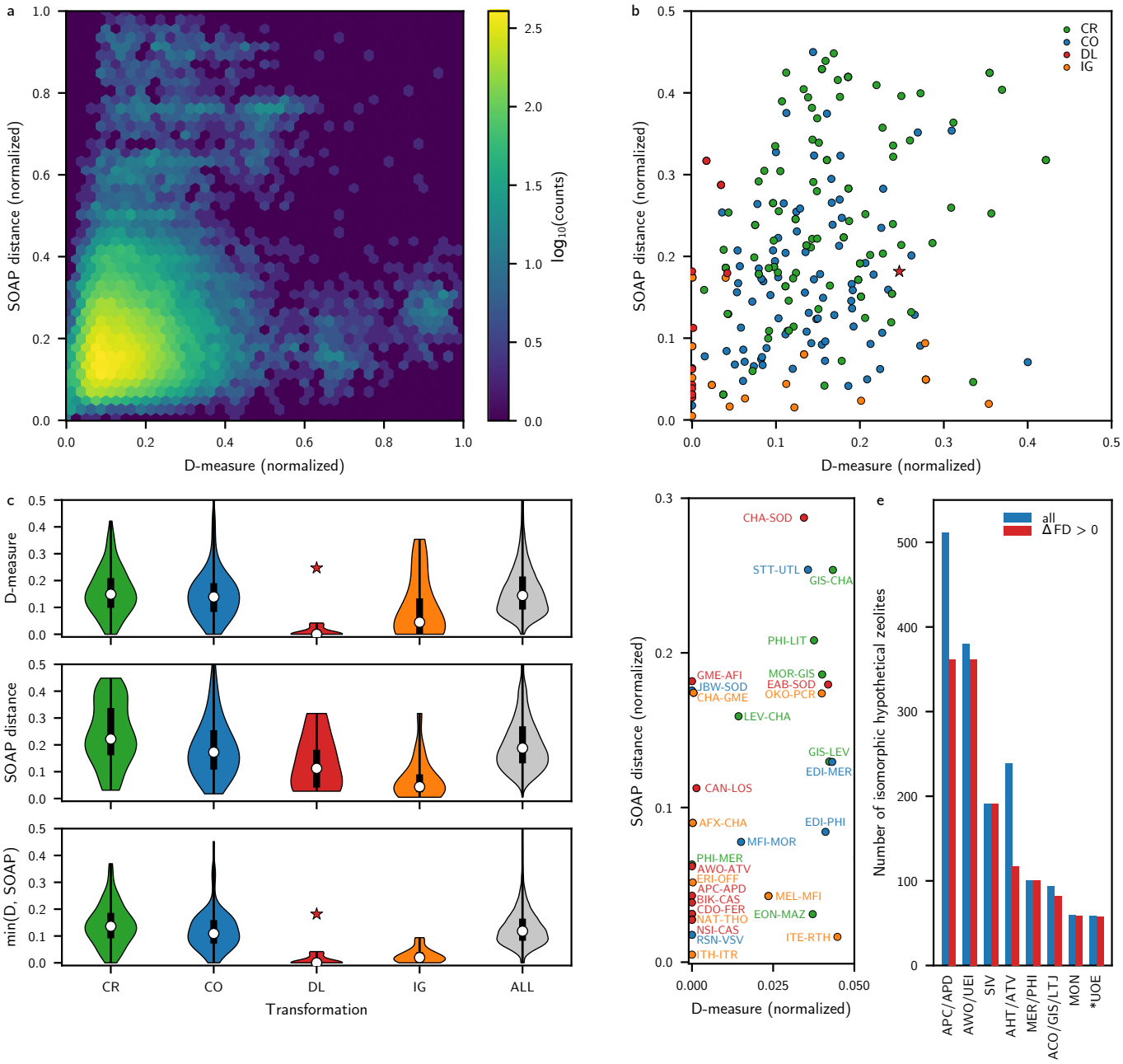

FIG. 3. **Structural and graph similarities in zeolites.** **a**, Distribution of zeolite pairs according to their (normalized) SOAP and graph distances. **b**, Distribution of experimentally known zeolite transformations in the joint SOAP-graph space and **c**, separately. Diffusionless transformations (DL) and intergrowth (IG) have smaller graph distances when compared to recrystallization (CR) and competing (CO) processes, as also evidenced by **d** and Supp. Fig. 11. The only exception is the LTA-IFY transformation[26], indicated by a star. In **c**, the white dot indicates the median of the distribution. The bottom and top of the thicker solid lines are the first and third quartiles, respectively. The whiskers indicate the maximum and the minimum of the range. As opposed to similarity in density or in CBUs, high topological similarity is rare, as evidenced by the narrow base of the ALL violin plots. **e**, Histogram of the number of hypothetical zeolites which are graph isomorphic to known frameworks. Blue bars indicate the count of all hypothetical frameworks isomorphic to the known dataset, whereas red bars indicate the number of hypothetical frameworks which are both isomorphic to and have higher density than their known counterparts. Only the eight most frequent codes are shown in the figure, sorted in descending order of frequency of denser hypothetical isomorphs. Supp. Fig. 7 reports the full list.

Fig. 7 show the number of hypothetical zeolites related to known ones. Based on previous evidence of accessibility and transitions towards denser crystals in diffusionless transformations (see Fig. 1f), we suggest that known zeolites with high number of isomorphic hypothetical frameworks would be strong starting points for the synthesis of novel zeolites. As a matter of fact, further transformations leading to novel structures have already been observed for the pair with most isomorphic frameworks, APC/APD, upon heating APD to temperatures above 600 °C[28]. However, the resulting phase seems to not have been characterized yet. A high number of polymorphs are graph isomorphic to (and denser than) APC/APD, AWO/UEI, SIV, ATV/AHT, and MER/PHI. They may be related to the richness of graph-equivalent topologies that can be constructed using chains in an orthorhombic symmetry.

Analogously, we propose new diffusionless transformations and intergrowths for known zeolites from their graph and structural similarities. 237 pairs of known zeolites have normalized graph distance under 0.01 (Supp. Tab. II), and 136 have SOAP distance smaller or equal to 0.04 (Supp. Tab. III). This tight threshold captures 60% of all pairs related by diffusionless transformations and intergrowth. These pairs would be the best starting points for further experimental verification of graph-based relationships, as well as to pursue new synthesis routes or intergrown frameworks with desired chemical compositions. Based on our literature corpus, we suggest that diffusionless transformations between graph-similar pairs could be induced by calcinating a low-density zeolite with temperatures in the range of 200-800 °C in ambient pressure. The solid-state transition can be further enhanced by hydrating the initial zeolite, adding cations such as Li^+ or Na^+ to it[29], and applying anisotropic stresses.[26] In the case of targeted intergrowth, we suggest that dual-templating or seed-assisted growth can enable the realization of two frameworks in a single grain. From a theoretical perspective, comparing crystallographic data from our candidates allows the identification of fault planes expected for such disordered crystals (as depicted in Fig. 1d). Additionally, combining topological predictions with growth models[30] can provide insights on the composition and supersaturation of the initial gel, stacking fault rates, and microstructural shape of the final crystal.

Finally, based on the distribution of distances between zeolites, we can determine which known frameworks are unique in terms of structure and topology. Supp. Figs. 9 and 10 show the distribution of pairwise distances for every known IZA framework. Whereas SOAP distances identify zeolites with lower framework densities (see Supp. Fig. 6, 9), the graph distance also detects similar and exotic topologies through pairwise comparisons (Supp. Figs. 8 and 10).

In summary, we have developed a computational approach based on graph theory to explain and predict interzeolite transformations. A literature mining was performed to clarify prevalent mechanisms of crystal phase relations. The extracted data was analyzed with structural and graph similarity metrics. We show how typical descriptors cannot fully explain pairwise interactions between zeolite crystals. Instead, topological similarity provides an alternative metric to understand and quantify zeolite transformations. We employ the graph isomorphism test to explain diffusionless interzeolite conversions regardless of their building units. Then, by generalizing this comparison to a topological similarity metric applied over supercells, we obtain remarkable agreement with the experimental results retrieved from the literature. Furthermore, we identify hundreds of high-similarity pairs among known zeolites, and thousands of hypothetical frameworks that are related to known topologies. These structural pairwise relationships can enable: (i) the organotemplate-free synthesis of known zeolites from suitable starting frameworks; (ii) the discovery of new frameworks by interconverting known ones; and (iii) the design of intergrown systems. Finally, the method for determining zeolite similarities is atom-agnostic and can be extended to metal-organic frameworks, covalent-organic frameworks, and other systems potentially related by diffusionless transformations.

[1] Davis, M. E. Ordered porous materials for emerging applications. *Nature* **417**, 813–821 (2002).

[2] Maldonado, M., Oleksiak, M. D., Chinta, S. & Rimer, J. D. Controlling crystal polymorphism in organic-free synthesis of Na-zeolites. *J. Am. Chem. Soc.* **135**, 2641–2652 (2013).

[3] Gallego, E. M. *et al.* "Ab initio" synthesis of zeolites for preestablished catalytic reactions. *Science* **355**, 1051–1054 (2017).

[4] Honda, K. *et al.* Role of structural similarity between starting zeolite and product zeolite in the interzeolite conversion process. *J. Nanosci. Nanotechnol.* **13**, 3020–6 (2013).

[5] Marler, B. & Gies, H. Hydrous layer silicates as precursors for zeolites obtained through topotactic condensation: a review. *Eur. J. Mineral.* **24**, 405–428 (2012).

[6] Eliášová, P. *et al.* The ADOR mechanism for the synthesis of new zeolites. *Chem. Soc. Rev.* **44**, 7177–7206 (2015).

[7] Li, C., Moliner, M. & Corma, A. Building Zeolites from Precrystallized Units: Nanoscale Architecture. *Angew. Chem. Int. Ed.* **57**, 15330–15353 (2018).

[8] Goel, S., Zones, S. I. & Iglesia, E. Synthesis of Zeolites via Interzeolite Transformations without Organic Structure-Directing Agents. *Chem. Mater.* **27**, 2056–2066 (2015).

[9] Baerlocher, C., McCusker, L. B. & Olson, D. H. *Atlas of Zeolite Framework Types* (Elsevier Science, Amsterdam, 2007), 6 edn.

[10] Xie, B. *et al.* Organotemplate-free and fast route for synthesizing beta zeolite. *Chem. Mater.* **20**, 4533–4535 (2008).

[11] Iyoki, K., Itabashi, K. & Okubo, T. Progress in seed-assisted synthesis of zeolites without using organic structure-directing agents. *Microporous Mesoporous Mater.* **189**, 22–30 (2014).

[12] Itabashi, K., Kamimura, Y., Iyoki, K., Shimojima, A. & Okubo, T. A Working Hypothesis for Broadening Framework Types of Zeolites in Seed-Assisted Synthesis without Organic Structure-Directing Agent. *J. Am. Chem. Soc.* **134**, 11542–11549 (2012).

[13] Verheyen, E. *et al.* Design of zeolite by inverse sigma transformation. *Nat. Mater.* **11**, 1059–1064 (2012).

[14] Zhao, Z. *et al.* Insights into the topotactic conversion process from layered silicate RUB-36 to FER-type zeolite by layer reassembly. *Chem. Mater.* **25**, 840–847 (2013).

[15] Van Tendeloo, L., Gobechiya, E., Breynaert, E., Martens, J. A. & Kirschhock, C. E. A. Alkaline cations directing the transformation of fau zeolites into five different framework types. *Chem. Commun.* **49**, 11737–11739 (2013).

[16]O'Keeffe, M. & Hyde, S. T. The asymptotic behavior of coordination sequences for the 4-connected nets of zeolites and related structures. *Z. Kristallogr.* **211**, 73–78 (1996).
[17]Foster, M. D. *et al.* Chemically feasible hypothetical crystalline networks. *Nat. Mater.* **3**, 234–238 (2004).
[18]Treacy, M., Rivin, I., Balkovsky, E., Randall, K. & Foster, M. Enumeration of periodic tetrahedral frameworks. II. Polynodal graphs. *Microporous Mesoporous Mater.* **74**, 121–132 (2004).
[19]Witman, M. *et al.* Cutting Materials in Half: A Graph Theory Approach for Generating Crystal Surfaces and Its Prediction of 2D Zeolites. *ACS Cent. Sci.* **4**, 235–245 (2018).
[20]Blatov, V. A. Topological relations between three-dimensional periodic nets. I. Uninodal nets. *Acta Crystallogr. A* **63**, 329–343 (2007).
[21]Porter, D. A., Easterling, K. E. & Sherif, M. *Phase Transformations in Metals and Alloys* (CRC Press, Boca Raton, 2009), 3 edn.
[22]Alberti, A., Cruciani, G. & Martucci, A. Reconstructive phase transitions induced by temperature in gmelinite-Na zeolite. *Am. Mineral.* **102**, 1727–1735 (2017).
[23]Dusselier, M., Kang, J. H., Xie, D. & Davis, M. E. CIT-9: A Fault-Free Gmelinite Zeolite. *Angew. Chem. Int. Ed.* **56**, 13475–13478 (2017).
[24]Schieber, T. A. *et al.* Quantification of network structural dissimilarities. *Nat. Commun.* **8**, 13928 (2017).
[25]Bartók, A. P., Kondor, R. & Csányi, G. On representing chemical environments. *Phys. Rev. B* **87**, 184115 (2013).
[26]Jordá, J. L. *et al.* Synthesis of a Novel Zeolite through a Pressure-Induced Reconstructive Phase Transition Process. *Angew. Chem. Int. Ed.* **52**, 10458–10462 (2013).
[27]Deem, M. W., Pophale, R., Cheeseman, P. A. & Earl, D. J. Computational Discovery of New Zeolite-Like Materials. *J. Phys. Chem. C* **113**, 21353–21360 (2009).
[28]Keller, E. B., Meier, W. M. & Kirchner, R. M. Synthesis, structures of AlPO4-C and AlPO4-D, and their topotactic transformation. *Solid State Ion.* **43**, 93–102 (1990).
[29]Alberti, A. & Martucci, A. Reconstructive phase transitions in microporous materials: Rules and factors affecting them. *Microporous Mesoporous Mater.* **141**, 192–198 (2011).
[30]Anderson, M. W. *et al.* Predicting crystal growth via a unified kinetic three-dimensional partition model. *Nature* **544**, 456–459 (2017).

Acknowledgements D.S.-K. acknowledges the MIT Nicole and Ingo Wender Fellowship, the MIT Robert Rose Presidential Fellowship, and the MIT Energy Initiative (MITEI) Storage Seed Fund for financial support. R.G.-B. thanks MIT DMSE, Toyota Faculty Chair and MITEI for support. The work of E.O. and Z.J. was partially funded by the National Science Foundation Award #1534340, DMREF, and the Office of Naval Research (ONR) under Contract No. N00014-16-1-2432. D.S.-K. and R.G.-B. thank A. Corma, M. Moliner and Y. Román-Leshkov for fruitful discussions.

Author Contributions R.G.-B. conceived the project. D.S.-K. and R.G.-B. formulated the hypothesis of graph-similar transformations. D.S.-K. developed the graph and supercell matching methods, wrote computer code, and performed all calculations. Z.J. and E.O. performed the literature mining and database query. Z.J., E.O. and D.S.-K. reviewed the extracted articles. D.S.-K. and R.G.-B. wrote the first version of the manuscript and made the figures. All authors contributed to the final version of the manuscript.

Competing Interests The authors declare no competing interests.

METHODS

Literature extraction Papers describing interzeolite transitions were found from a database of over 2 million material science and chemistry articles using a natural language processing (NLP) pipeline[31]. First, a subsection of approximately 70,000 papers related to zeolite materials was extracted by searching for variants of the word "zeolite" within the text of each paper. Then, papers containing interzeolite transitions were found by searching for pairs of zeolite materials with different IZA structures within the title, abstract, and synthesis paragraphs of the paper, as determined by the NLP pipeline[32]. These sections were also searched for keywords relating to interzeolite transitions such as "inter-*", "intergrowth", "topological", "reconstruction", and "ADOR". More than 540 papers contained at least one pair of structures and at least one keyword match. They were manually checked to determine the type of transition and pairs of zeolites involved in the interzeolite transformation.

Database of zeolites A database of 245 known zeolite frameworks was downloaded from the Database of Zeolite Structures as of March 2019, kept by the Structure Commission of the International Zeolite Association (IZA)[33], including interrupted and partially disordered frameworks. Their framework density and CBUs were also extracted from the website. We selected the hypothetical zeolite database generated and optimized using the Sanders-Leslie-Catlow (SLC) force field[34] by Deem *et al.*[27,35] From the complete database with about 314,000 entries, we removed those whose energies were lower than quartz, ending up with 269,515 zeolite structures. All zeolites are considered in their pure silicate form.

Construction of zeolite multigraphs We adopted a multigraph representation of crystals satisfying periodic boundary conditions[36] for pure zeolite silicate structures. To maximize the amount of information embedded in the graph while minimizing its size, we removed the oxygen atoms from the graph. Each crystal graph contains as many nodes as Si atoms in its input unit cell. Nodes are connected if and only if their correspondent Si atoms share an oxygen atom. This avoids errors from the usage of nearest-neighbors search using Voronoi diagrams for porous materials. No node or edge labels are included in the graph.

Comparing crystal graphs Multigraph isomorphism is performed using the VF2 algorithm[37] as implemented at NetworkX[38]. The graph similarity D-measure is implemented as reported by Schieber *et al.*[24] As the graph complement cannot be computed for multigraphs, the alpha-centrality is taken with respect to the graph without parallel edges. This still preserves notions of connectivities and distances.

The variational approach for the graph similarity starts with two crystals A, B with lattice matrices

$$A = \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \mathbf{a}_3 \end{bmatrix}, \tag{1}$$

$$B = \begin{bmatrix} \mathbf{b}_1 & \mathbf{b}_2 & \mathbf{b}_3 \end{bmatrix}, \tag{2}$$

with $\mathbf{a}_i, \mathbf{b}_j$ lattice vectors associated to each of the crystals. Each unit cell given by A, B contains a number $n_T^{(A,B)}$ of T atoms in it. Given the lattice and the atomic positions, we can construct a graph $G_X = g(X)$ for the crystal X using the crystal graph constructor $g : \mathcal{C} \to \mathcal{G}$. This maps the space of crystal structures (made from lattice and atomic basis) $\mathcal{C}$ to the space of graphs $\mathcal{G}$. As demonstrated by the isomorphism, the constructor g is not necessarily injective.

Ultimately, we want to compare the topology between two graphs with the same number of nodes. This determines a scaling $m^{(A,B)}$ necessary to equalize the mass of both supercells,

$$m^{(A,B)} = \frac{\mathrm{LCM}\left(n_T^{(A)}, n_T^{(B)}\right)}{n_T^{(A,B)}}, \tag{3}$$

where $\mathrm{LCM}\left(n_T^{(A)}, n_T^{(B)}\right)$ is the least common multiple between $n_T^{(A)}$ and $n_T^{(B)}$. We can then look for the transformation matrices $M^{(A,B)}$ that minimize the graph distance $d_{\mathcal{G}}(G_A, G_B)$ between two graphs G_A, G_B created using the function g,

$$d(A,B) = \min_{M^{(A)}, M^{(B)}} d_{\mathcal{G}}\left(g\left(M^{(A)}A\right), g\left(M^{(B)}B\right)\right), \tag{4}$$

subject to the constraints

$$\det M^{(A,B)} = m^{(A,B)}, \tag{5}$$

$$m_{ij}^{(A,B)} \in \mathbb{Z}. \tag{6}$$

The resulting supercells are described by the matrices $M^{(A,B)}$, in a generalization of the coincidence lattice method for surfaces[39] with a custom graph metric. The crystal A, B supercells have lattice matrices given by $M^{(A)}A$, $M^{(B)}B$ plus the resulting atomic basis. A practical implementation is described in the Supplementary Information.

SOAP For each atomic environment $\mathcal{X}_i$ in the crystal structure, a SOAP power spectrum $\mathbf{p}(\mathcal{X}_i)$ is calculated[25,40] using $r_{\mathrm{cut}} = 10$, radial basis size of 8 with $L_{\max} = 5$ as hyperparameters. Instead of adopting the regularized entropy match kernel[41] to compare crystal

structures, we opt for the average SOAP fingerprint given by

$$\mathbf{p}(Z) = \frac{1}{N}\sum_i \mathbf{p}(\mathcal{X}_i^Z) \tag{7}$$

for each crystal structure Z. This allow us to simplify the analysis to hundreds of thousands of zeolites with varying number of atoms.
The unnormalized average structure kernel K can be defined as the inner product of their power spectra,[41]

$$K(A,B) = \mathbf{p}(A)\cdot\mathbf{p}(B), \tag{8}$$

leading to the normalized kernel equivalent to a cosine similarity,

$$\bar{K}(A,B) = \frac{K(A,B)}{\sqrt{K(A,A)K(B,B)}}. \tag{9}$$

This kernel induces a metric d_{kernel} given by[41]

$$d_{\text{kernel}}(A,B) = \sqrt{2 - 2\bar{K}(A,B)}. \tag{10}$$

Data Availability The zeolite datasets analyzed during the current study are available online at the International Zeolite Association Database (`http://www.iza-structure.org/databases/`), and at the Predicted Crystallography Open Database (`http://www.crystallography.net/pcod/`). Complete references for the literature analyzed in this work, pairwise distances between known zeolites, and isomorphism between hypothetical and known zeolites are available in the Supplementary Information.

Code Availability The code used to download journal articles for large-scale text-mining is available at [`www.github.com/olivettigroup/article-downloader`]. The code used to variationally compare crystal structures as supercell graphs is available from the corresponding author upon request.

[31] Kim, E. *et al.* Machine-learned and codified synthesis parameters of oxide materials. *Sci. Data* **4**, 170127 (2017).

[32] Jensen, Z. *et al.* A machine learning approach to zeolite synthesis enabled by automatic literature data extraction. *ACS Cent. Sci.* **5**, 892–899 (2019).

[33] Ch. Baerlocher and L.B. McCusker. Database of Zeolite Structures. *http://www.iza-structure.org/databases/* (2019).

[34] Schröder, K. P. *et al.* Bridging hydrodyl groups in zeolitic catalysts: a computer simulation of their structure, vibrational properties and acidity in protonated faujasites (HY zeolites). *Chem. Phys. Lett.* **188**, 320–325 (1992).

[35] Pophale, R., Cheeseman, P. A. & Deem, M. W. A database of new zeolite-like materials. *Phys. Chem. Chem. Phys.* **13**, 12407–12412 (2011).

[36] Xie, T. & Grossman, J. C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Phys. Rev. Lett.* **120**, 145301 (2018).

[37] Cordella, L. P., Foggia, P., Sansone, C. & Vento, M. A (sub)graph isomorphism algorithm for matching large graphs. *IEEE Trans. Pattern Anal. Mach. Intell.* **26**, 1367–1372 (2004).

[38] Hagberg, A. A., Schult, D. A. & Swart, P. J. Exploring Network Structure, Dynamics, and Function using NetworkX. In Varoquaux, G., Vaught, T. & Millman, J. (eds.) *Proceedings of the 7th Python in Science Conference*, 11–15 (Pasadena, CA USA, 2008).

[39] Koda, D. S., Bechstedt, F., Marques, M. & Teles, L. K. Coincidence Lattices of 2D Crystals: Heterostructure Predictions and Applications. *J. Phys. Chem. C* **120**, 10895–10908 (2016).

[40] Jäger, M. O. J., Morooka, E. V., Federici Canova, F., Himanen, L. & Foster, A. S. Machine learning hydrogen adsorption on nanoclusters through structural descriptors. *npj Comput. Mater.* **4**, 37 (2018).

[41] De, S., Bartók, A. P., Csányi, G. & Ceriotti, M. Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.* **18**, 13754–13769 (2016).