ACS central science

http://pubs.acs.org/journal/acscii

Research Article

Discovering Relationships between OSDAs and Zeolites through Data Mining and Generative Neural Networks

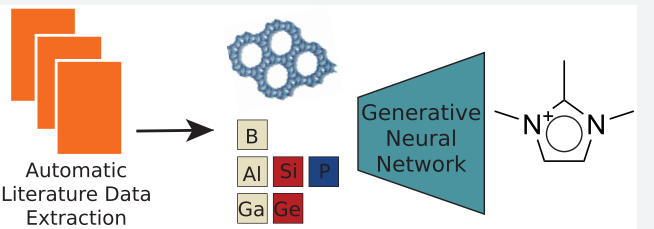
Zach Jensen, Soonhyoung Kwon, Daniel Schwalbe-Koda, Cecilia Paris, Rafael Gómez-Bombarelli, Yuriy Román-Leshkov, Avelino Corma, Manuel Moliner, and Elsa A. Olivetti*

Cite This: *ACS Cent. Sci.* 2021, 7, 858–867

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Organic structure directing agents (OSDAs) play a crucial role in the synthesis of micro- and mesoporous materials especially in the case of zeolites. Despite the wide use of OSDAs, their interaction with zeolite frameworks is poorly understood, with researchers relying on synthesis heuristics or computationally expensive techniques to predict whether an organic molecule can act as an OSDA for a certain zeolite. In this paper, we undertake a data-driven approach to unearth generalized OSDA–zeolite relationships using a comprehensive database comprising of 5,663 synthesis routes for porous materials. To generate this comprehensive database, we use natural language processing and text mining techniques to extract OSDAs, zeolite phases, and gel chemistry from the scientific literature published between 1966 and 2020. Through structural featurization of the OSDAs using weighted holistic invariant molecular (WHIM) descriptors, we relate OSDAs described in the literature to different types of cage-based, small-pore zeolites. Lastly, we adapt a generative neural network capable of suggesting new molecules as potential OSDAs for a given zeolite structure and gel chemistry. We apply this model to CHA and SFW zeolites generating several alternative OSDA candidates to those currently used in practice. These molecules are further vetted with molecular mechanics simulations to show the model generates physically meaningful predictions. Our model can automatically explore the OSDA space, reducing the amount of simulation or experimentation needed to find new OSDA candidates.



INTRODUCTION

Zeolites and related zeotype materials are crystalline, microporous materials extensively used in a variety of industrial applications.[1−3] Among their different physicochemical properties, the crystalline structure and building unit geometry are critical in determining their suitability for target applications based on structure-dependent molecular shape selectivity, diffusivity, and confinement. Although there are over 250 recognized zeolite structures,[4] the exact mechanisms associated with the nucleation and crystallization of zeolites are still not fully understood,[5−8] making the *a priori* prediction of a desired zeolite phase from an initial set of conditions inexact and difficult. For this reason, the discovery of new zeolite structures has historically been based on trial-and-error synthesis methodologies guided by accumulated human knowledge and chemical intuition.[9] Variables known to influence zeolite formation include the types and amounts of framework atoms, mineralizing agents, and inorganic/organic structure directing agents.[1,9,10]

Organic structure directing agent (OSDA) molecules play a crucial role in zeolite synthesis. They can provide different effects within the synthesis from charge balancing and space filling to a templating, lock-and-key relationship.[11] This results in a wide range of OSDA specificity with some OSDAs able to crystallize many different zeolite phases while others can only direct the formation of a limited number of phases. The size, flexibility, hydrophilicity, and charge of the OSDA, among other factors, play an important role in zeolite crystallization kinetics and phase specificity.[12−14] Indeed, experimental heuristics within the zeolite community connects the OSDA size with increasing zeolite pore size and increasing OSDA rigidity with increasing specificity or formation of fewer zeolite phases, although designing OSDAs from these heuristics remains challenging.[12] Researchers have used computational approaches including density functional theory and molecular dynamics to suggest candidate OSDAs for specific zeolite structures,[15−17] but these approaches are typically limited to a single zeolite system, computationally expensive, and focus on pure silica systems. More recently, a strategy involving the "*ab initio*" design of the OSDA to mimic the transition states of industrially relevant catalytic reactions has gained attention,[18,19] but this technique relies on computationally expensive density functional theory calculations, hindering its

Received: January 6, 2021
Published: April 16, 2021

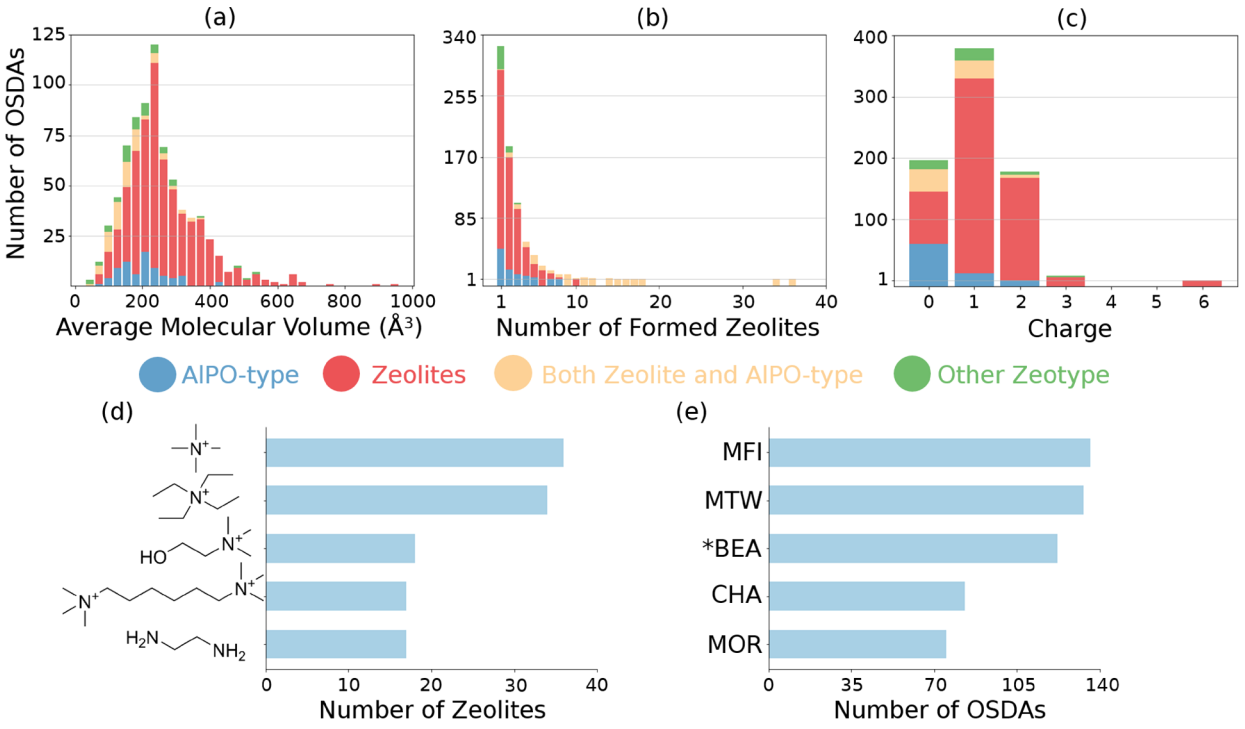


Figure 1. Overview of the automatically extracted data set. (a–c) Average molecular volume, OSDA specificity, and charge distributions for all OSDAs in the data set. (d) Shows the five OSDAs known to make the most zeolite structures. (e) Shows the five zeolites that can be made with the most OSDAs.

widespread implementation. Undoubtedly, we must develop new modeling approaches that are more efficient and comprehensive to advance OSDA design.

Data-driven approaches have been used to study porous materials,[20−23] but data-driven zeolite synthesis studies are limited in scope and rely on overly simplified OSDA−zeolite interactions that cannot capture the complexity of the system. Machine learning (ML) and data mining have been used in studies that do not require explicitly modeling the OSDA−zeolite interaction including specific zeolites within limited regions of the chemical space,[24,25] OSDA-free zeolite systems,[26] and interzeolite transformations.[27] Studies that have attempted to model this interaction either simplify the OSDA representation to basic properties such as molecular volume[28] or are limited to a single zeolite structure,[16] suggesting that more advanced ML techniques and larger data sets are needed to better model the OSDA−zeolite relationship.[29] The literature provides a comprehensive data set of the known OSDA−zeolite pairs, and recent studies have provided natural language processing (NLP) frameworks that can be adapted to extract OSDA, zeolite, and chemistry information, including all the elemental species present in the synthesis gel.[28,30−32] Literature-extracted data combined with advanced ML techniques for the OSDA−zeolite relationship could expand the scope of data-driven zeolite studies.

ML also enables the pursuit of inverse design for both porous materials[33−35] and organic molecules. One approach to inverse design is generative neural network models, which have been successful for many applications including drug discovery,[36] property optimization,[37] synthesis prediction,[38] and molecular design.[39,40] These models learn a latent representation of an organic molecule typically by compressing the training data into a multidimensional Gaussian distribution and reconstructing it from sampled vectors. This latent space can then be explored to generate novel organic molecules that resemble the support distribution. These new samples are then converted into standard molecular representations such as the "simplified molecular-input line-entry system" (SMILES) format.[41] Recent models have been trained to generate molecules directly from the molecule's physical and chemical properties.[42] During inference, researchers input in the desired properties and generate molecules that possess them. This model can be adapted to zeolite data by training the model to generate organic molecules from specific zeolite structures and gel chemistry.

In this paper, we use a data-driven approach to examine the relationships between OSDAs, qualitative gel chemistry, and resulting zeolite structures. We present an exhaustive OSDA, zeolite, and qualitative synthesis data set extracted through NLP and text mining techniques. We use structural descriptions of the OSDAs to reduce the dimensionality of the chemical space and visualize trends found in the crystallization of certain zeolites. Finally, we adapt a generative neural network model trained on this extracted data set to suggest potential OSDA molecules conditioned on specific zeolite structures and synthesis conditions. The data, models, and resulting analyses provide research opportunities for the community to further expedite zeolite research and represent an important first step toward developing a high-throughput zeolite research pipeline.

RESULTS AND DISCUSSION

Extracted Data Set. We extract a data set of OSDAs, chemistry, and zeolite phases from across the entire zeolite literature with automated techniques.[28,30,31] This data set consists of articles from over 15 different publishers and 140 journals and spans the year range from 1966 to 2020. It contains 5,663 synthesis routes from 1,384 articles containing

https://doi.org/10.1021/acscentsci.1c00024
ACS Cent. Sci. 2021, 7, 858−867

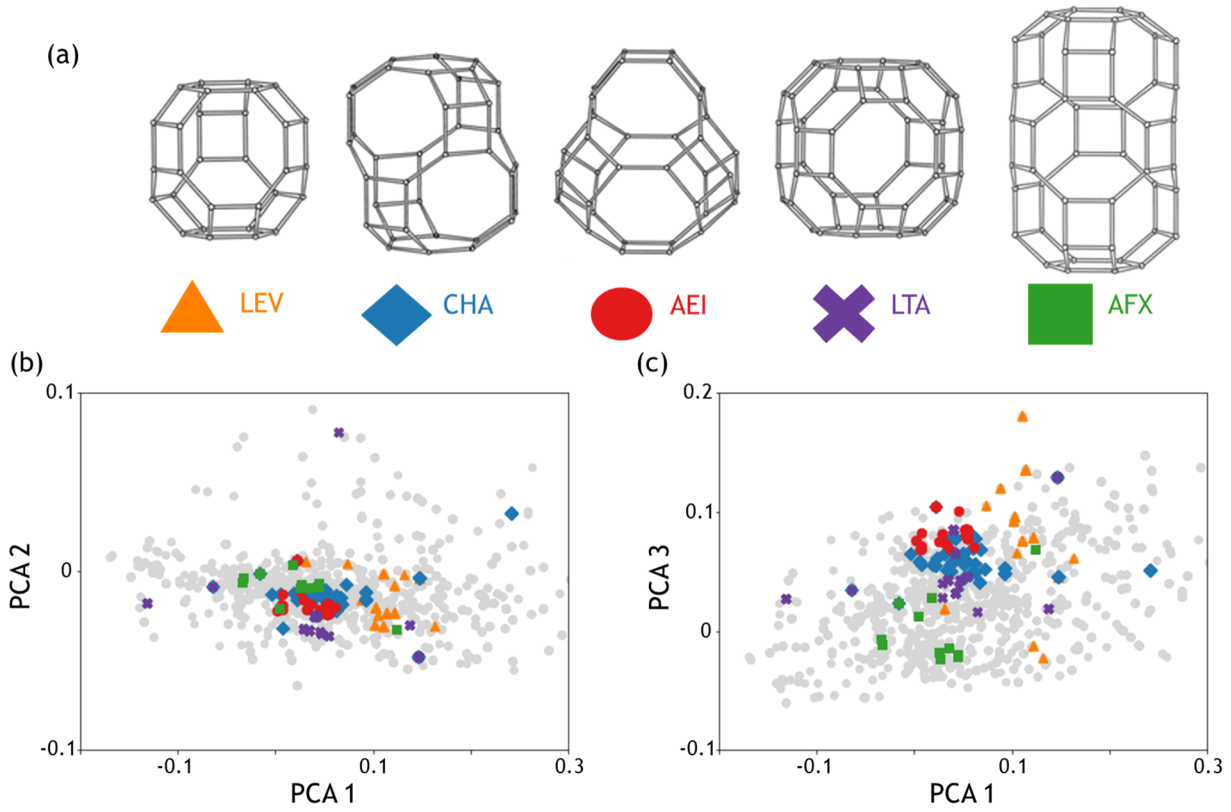


Figure 2. Principal component analysis (PCA) WHIM vector representation of OSDA molecules used in five cage-based small-pore zeolite systems. PCA 1, 2, and 3 represent the first three principal component axes. The gray points represent all of the OSDAs extracted from the literature.

the OSDAs, qualitative synthesis gel components, and the resulting zeolite phases. This data set contains information on 758 distinct OSDA molecules and 205 zeolite phases. Among the different synthesis routes, 3,085 describe traditional zeolites (pure Si, Si/Al, and Si/B frameworks), 1,274 describe aluminophosphate (AlPO)-type materials, while the remaining 1,304 data points describe additional zeotypes including germanium-based and metal-containing (Ti, Sn, or Zr, among others) microporous structures. Distributions of different zeolite structures and synthesis gel chemistry contained in the data set can be seen in Figures S1 and S2.

Figure 1a shows the average molecular volume distribution of the OSDAs in the data set. The molecular volumes range from about 30 to 1000 Å^3. Figure 1a shows that larger OSDAs are related to the synthesis of zeolites instead of AlPO-type materials. This observation agrees with less correlation between organic molecules and the pores/cages observed experimentally for AlPO-type materials[12] and the limited stability of large-pore AlPO-type materials compared to their aluminosilicate counterparts. This limited stability has mostly precluded heuristic studies using bulky and expensive OSDA molecules in their synthesis.

The majority of the OSDAs have high specificity, producing fewer than 5 zeolite phases, while a few outliers are capable of making more than 20 phases (Figure 1b). These lower-specificity OSDAs are typically small and simple alkylammonium cations, such as tetramethylammonium (TMA) or tetraethylammonium (TEA) shown in Figure 1d. These molecules act as space-filling molecules to provide charge balance to the framework and generally do not provide a true templating effect. Other low-specificity OSDAs feature high flexibility with many rotatable bonds, such as hexamethonium (see Figure 1d). The zeolites that have been experimentally obtained using the most organic molecules are MFI, MTW, *BEA, CHA, and MOR (Figure 1e). These topologies are among the most widely used industrial applications (along with FAU and FER), thereby having more fundamental research efforts to improve their physicochemical properties and cost effectiveness.[43]

The number and distribution of ionic charges within the OSDAs play an important role in the nucleation and crystallization processes and, together with the presence/absence of alkali cations, are crucial for positioning the negatively charged heteroatoms in specific framework positions. Heteroatom location has been shown to drastically alter the catalytic properties of the materials.[44−46] In zeolite synthesis, most OSDAs contain one or two positive charges, generally in the form of mono- or dicationic ammonium species (Figure 1c).[12−14] While the use of neutral amines has also been reported for the synthesis of zeolite-type materials, these molecules mostly act as pore fillers. In contrast, AlPO-type materials are preferentially synthesized using amines as OSDAs (blue bar in 0 charge in Figure 1c), which are protonated in the neutral or acidic media of a typical AlPO-type material synthesis gel.

Literature-Mined OSDA/Zeolite Correlations. Due to the complex interactions between OSDAs and the resulting zeolite framework (Figures S3 and S4), simple descriptors like molecular volumes and/or flexibility parameters (e.g., nConf20)[47] are insufficient to describe links between specific OSDAs and zeolite structures. We also consider nonstructural properties of the OSDAs and their effect on zeolite structure

https://doi.org/10.1021/acscentsci.1c00024
ACS Cent. Sci. 2021, 7, 858−867

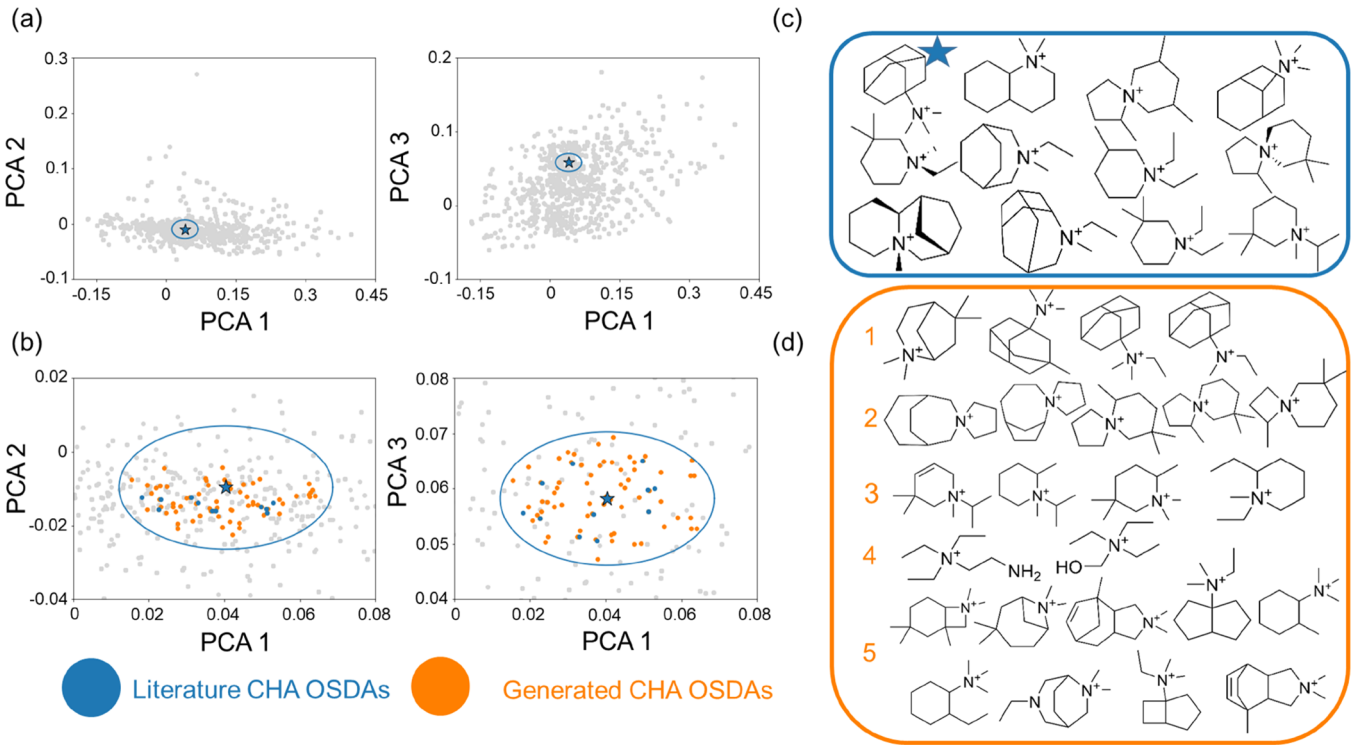


Figure 3. Comparing literature OSDAs and generated OSDAs of a CHA zeolite. (a) Shows the position of TMAda (shown with the blue star) relative to the rest of the OSDAs in the PCA WHIM space. (b) A zoomed in view of the ellipse surrounding it. (c) The blue square contains literature CHA OSDAs that fall within the ellipse. (d) The orange square contains examples of generated OSDAs for CHA that fall within the ellipse.

(Figure S5), but these features also inadequately describe the OSDA−zeolite relationship. To capture molecular shape matching, we need a more informative structural descriptor to capture not only the size of the molecule but also other structural features such as folding and charge distributions. Accordingly, weighted holistic invariant molecular (WHIM)[48] descriptors contain information about the size, shape, symmetry, and atom distribution that is dependent on the three-dimensional conformation of the molecule. Depending on the flexibility of the molecule, different conformations can have drastically different WHIM representations. For example, a long linear molecule can either stretch out or fold, giving two different three-dimensional representations (Figure S6). To address this challenge, we calculate the average conformation WHIM descriptor using geometries obtained with RDkit[49] to capture the varying three-dimensional representation of each molecule based on its different conformations.

Because WHIM is a high-dimensional descriptor, we use principal component analysis (PCA) to reduce the dimensionality of the WHIM descriptor space and enable visualization of all the OSDAs in the data set. The first principal component (PCA 1) accounts for 58% of the variance and correlates with the volume of the molecule, as it contains WHIM features corresponding to the longest axial and global dimension of the molecule. The second principal component (PCA 2) accounts for 15% of the variance and is composed of global dimension and symmetry features. The third principal component (PCA 3) accounts for 13% of the variance and has many contributing features including all three of the axial dimensions and the global dimensions (Figures S7 and S8). We show the PCA WHIM visualization comparing the OSDAs to a sampling of the entire organic space to highlight the limited chemical space of known OSDAs (Figure S9). These dimensionally reduced WHIM descriptors highlight relationships between OSDAs and zeolite phases.

We select five cage-based small-pore zeolites, LEV, CHA, AEI, LTA, and AFX (Figure 2a), to evaluate the OSDA−zeolite correlations through the WHIM descriptor featurization and PCA analysis (Figure 2b,c). Cage-based zeolites have a strong correlation between the three-dimensional structure of the OSDA and the shape of the cage, making them good candidates for analysis. Since gel composition also affects the relationship between the OSDA and zeolite, we filter the data set to include only conventional zeolite chemistry versions for the selected zeolites using the extracted qualitative synthesis gel information. We also explore this relationship for selected large-pore zeolites to examine the generalization of this approach to other zeolite systems (Figure S10).

For these five zeolites, Figure 2 shows that each zeolite topology is associated with specific and distinct OSDA characteristics. Differences are observed between the locations of the clusters, particularly PCA 2 and PCA 3, likely due to the differences in cage size and shape requiring different molecular structures as OSDAs. The OSDAs for LTA show larger variability among the PCA parameters than those for the other zeolite clusters (purple crosses in Figure 2b). The synthesis of high-silica LTA has been preferentially reported by using large aromatic molecules[50,51] (Figure S11), while small organic molecules have been employed for the synthesis of low-silica LTA, i.e., tetramethylammonium or diethyldimethylammonium (Figure S11), which act as pore fillers in combination with additional alkali cations. The difference in OSDA size for high- and low-silica LTA materials is likely responsible for the large PCA variability observed for the LTA cluster.

https://doi.org/10.1021/acscentsci.1c00024
ACS Cent. Sci. 2021, 7, 858−867

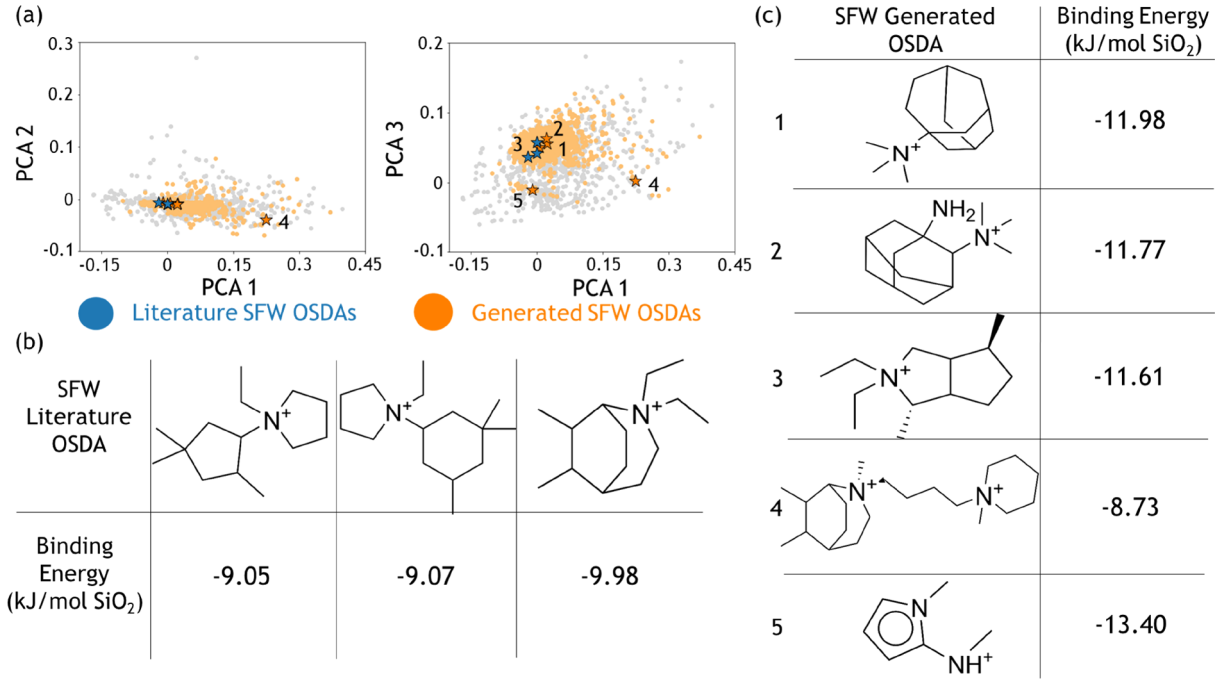


Figure 4. OSDAs for SFW obtained from literature and generated by our model. (a) PCA-reduced WHIM locations for the three OSDAs known to make SFW (blue stars) and five selected molecules generated by our model (orange stars). (b) Minimum conformer binding energy with SFW for the three literature OSDAs. (c) Binding energy with SFW for the five selected generated molecules.

The two clusters representing CHA and AEI are very close in the PCA WHIM vector representation (blue diamonds and red circles in Figure 2b,c) and have significantly reduced variance compared to LTA. The overlapped region of both clusters suggests that the OSDAs used to synthesize these frameworks are structurally similar. In fact, some of these molecules can be used to make either framework by modifying the synthesis conditions (Figure S12). This phenomenon is expected given that both AEI and CHA zeolites have many structural similarities including a cage-like three-dimensional small-pore system and identical framework density (15.1 T/1000 Å^3). However, since these materials present cavities with different shapes (Figure 2a), elongated and symmetrical in the case of CHA (11.7 × 10.2 Å) and basket-cage-type in AEI (12.6 × 11.2 Å), there are specifically shaped OSDA molecules that would selectively fit CHA or AEI cavities, thus guiding their preferential crystallization (Figure S12).

Suggesting New Candidate OSDAs through Generative Modeling. We adapt a generative neural network model published by Kotsias et al.[42] to suggest alternative organic molecules for use as OSDAs. This model is trained on the extracted literature data to output a SMILES string for an OSDA molecule given a zeolite phase and gel chemistry as input (architecture and training procedure is described in the Experimental Section). This model allows us to move beyond mining relationships from the literature toward the process of discovering new OSDAs for particular zeolite structures. This model requires a large quantity of data to train a useful model,[52] which is enabled by the size of our extracted data set. Quantitative performance and benchmarking metrics for the model are discussed in the Supporting Information (see Table S1 and Figure S13).

With this model, we generate potential OSDA molecules for a cage-based zeolite system featured above, CHA, due to its industrial relevance. A total of 10,000 samples are drawn from the model using different zeolite gel chemistry variations including pure Si, Si−Al, and Si−B while also including Na^+ and K^+ cations and F^- as a mineralizer. This procedure generates 408 unique OSDAs for CHA.

To filter the generated OSDA molecules, we compare them to the OSDA currently used in industry for CHA, *N*,*N*,*N*-trimethyladamantammonium (TMAda). We take the PCA-reduced WHIM coordinates of the TMAda and create an ellipsoid around the point taking 5% of the range along the first three principal component axes (see Figure 3a,b). Of the 408 generated CHA molecules, 57 fall within the TMAda ellipsoid. Another 11 OSDAs previously reported in the literature for CHA and 24 other OSDAs reported for other topologies also fall within this range. Organic molecules within the ellipsoid are expected to be structurally similar to TMAda and therefore may be suitable alternative OSDAs as we explore further below.

Figure 3 shows this information flow and some of the resulting generated organic molecules for CHA (Figure 3d). The highlighted points within the WHIM space represent OSDAs that fall within the ellipsoid in all three PCA dimensions. Looking qualitatively, the generated OSDAs contain many similar features as the OSDAs found in literature used for the synthesis of CHA (Figure 3c). For instance, different adamantyl-type, rigid molecules are predicted (row 1 in Figure 3d), in good agreement with the experimentally described TMAda, considered as the most effective template to stabilize the CHA cavity.[53−55] Beyond adamantyl-type molecules, different alkyl-substituted spiro and piperidinium molecules have been generated by the model as proposed OSDAs for CHA (rows 2 and 3 respectively in Figure 3d), which present similar structural features as some reported CHA OSDAs. In addition, two simple tetraalkylammonium cations have also been generated (row 4 in Figure 3d). We note that tetraethylammonium has been recently reported as

https://doi.org/10.1021/acscentsci.1c00024
ACS Cent. Sci. 2021, 7, 858−867

an OSDA for the synthesis of CHA in its silicoaluminate form.[56] The model also generates other types of molecules not directly seen in the literature (row 5 in Figure 3d) but have commonly observed features including a single positively charged nitrogen atom and cyclic structures. The generated molecules demonstrate the model's ability to add domain and data-informed chemical noise into the OSDA space in a way that allows intelligent prediction of potential OSDA candidates.

We also evaluate the generated OSDA candidates for a zeolite that is less studied than CHA. We choose the SFW framework, which has been synthesized as a Si−Al zeolite using three different OSDA molecules according to our data set and presents high potential interest for its application as an effective catalyst for NO_x abatement.[57,58] SFW is structurally similar to CHA, having the same framework density (15.1 T/1000 Å^3) and being cage-based with the *gme* cage replacing the *cha* cage. Since few OSDAs are known for SFW, we use molecular mechanic simulations to calculate the binding energy of each of the generated molecules with the SFW framework to gauge our model's predictive ability, rather than comparing to known molecules as for CHA. The atomistic simulations follow the procedures laid out by Schwalbe-Koda and Gómez-Bombarelli[59,60] (see also the Experimental Section).

The molecular mechanic simulations show that many of the generated molecules produced by our model are suitable OSDA candidates for SFW. Of the generated molecules, 60% have binding energies within the range of the literature OSDAs (−9.98 to −7.48 kJ/mol SiO2). Interestingly, an additional 7% have lower binding energies than the known OSDAs. Figure 4a shows the results of generating molecules for SFW in the reduced WHIM space. The blue stars represent the OSDAs known to synthesize SFW, *N*-ethyl-*N*-(2,4,4-trimethylcyclopentyl)pyrrolidinium, *N*-ethyl-*N*-(3,3,5-trimethylcyclohexyl)pyrrolidinium, and *N*,*N*-diethyl-5,8-dimethyl-azonium bicyclo[3.2.2]nonane, while the orange points represent generated molecules. Figure 4b shows the binding energy for each of the three literature OSDAs. We select five of the generated molecules, shown in Figure 4c. Molecules 1, 2, and 3 are structurally similar to the known OSDAs and have very low binding energies. These strong binding energies support the relationship between distance in the WHIM space and OSDA potential. Molecules 4 and 5 are chosen for strong binding energies while being structurally different than the known OSDAs. Molecule 4 is significantly larger than the known OSDAs, indicating that a single, well-fitting OSDA per cage could also have a strong templating effect toward SFW, while molecule 5 is significantly smaller than the known OSDAs, requiring packing more molecules into the cage. These two molecules demonstrate the model's ability to suggest molecules that are structurally dissimilar from the known OSDAs.

While the model is able to generate physically meaningful suggestions for the SFW zeolite, it has performance limitations. We probe its ability to provide different distributions of molecules depending on the zeolite and chemistry. First, we compare the generated SFW OSDAs with generated LAU OSDAs. LAU is structurally very different than SFW, having a higher framework density (18.0 T/1000 Å^3), a 1-dimensional, 10-membered ring channel, and no composite building units in common with SFW. Furthermore, LAU is typically synthesized as an M−(Al/Ga)PO (M = Co, Mn, Zn, Fe)-type material,[61,62] while SFW is a conventional zeolite,[57,58] making them chemically different as well. There is a clear difference in the WHIM distributions of the molecules generated for the two systems indicating the model's ability to distinguish between the structures during prediction (Figure S14a). Figure S14b shows the distributions of minimum distance in the WHIM space to one of the known SFW and LAU OSDAs. We also generate LAU OSDAs using the SFW zeolite chemistry to compare the effect chemistry has on the model. As expected, having similar chemistry shifts the generated distributions closer together although they are still distinct. We also compare the SFW binding energies of the generated OSDAs and OSDAs from the entire zeolite literature (Figure S15). Figure S15 shows these distributions are very similar, indicating the model may have a limited ability to predict OSDAs specific to each zeolite system. However, the model is able to match the literature distribution, containing molecules known to be suitable OSDAs. These results taken together demonstrate the model's ability to generate different OSDA suggestions by injecting chemical noise into the OSDA space but still matching the performance of known literature OSDAs. This result indicates that generated molecules may have potential as OSDAs for several structurally similar zeolite systems. Pairing this model with binding energy simulations could help in selecting predicted OSDAs.

CONCLUSION

We have extracted and featurized data on OSDAs, zeolite phases, and gel chemistry from across the zeolite literature, resulting in a large, comprehensive data set of zeolite synthesis parameters. We have then mined this literature data to uncover relationships between the structure of the OSDA and the resulting zeolite phase using a calculated three-dimensional feature called WHIM. Finally, we model the interaction between the OSDA, zeolite, and gel chemistry using a generative neural network. This model can suggest novel organic molecules with binding energies below and comparable with their known literature counterparts.

While all of the chemistry data extracted in this paper is qualitative, a promising avenue for supplemental work is to extract quantitative information about the gel chemistry. This information would allow for more detailed thermodynamic and kinetic studies of zeolite synthesis. Additional atomistic simulations could further aid the selection of OSDAs with greatest potential to experimentally form the target zeolite. This model and data could be combined with more advanced, rapid simulation techniques and experimental optimization to develop a high-throughput zeolite synthesis pipeline.

EXPERIMENTAL SECTION

Data Extraction, Processing, and Validation. Over 3.5 million chemistry and materials science journal articles were scanned for keywords relating to zeolite materials including "zeolite", "osda", "aluminophosphate", and "molecular sieve", resulting in a corpus of approximately 90,000 papers. From this corpus, OSDA names, zeolite structures, and synthesis gel components were extracted from the tables and synthesis sections of each paper using regular expression and domain specific keyword matching. While this approach works well for extracting raw zeolite data with very high recall, it is difficult to determine specific OSDA−zeolite−synthesis systems, especially for papers that contain multiple experimental samples.

https://doi.org/10.1021/acscentsci.1c00024
ACS Cent. Sci. 2021, 7, 858−867

Because of this, each extracted paper was manually checked to ensure integrity and accuracy of the extracted synthesis route.

Data Normalization and Featurization. Since authors use a variety of chemical names to describe both OSDA molecules and zeolite structures, the extracted text data needed to be normalized so different naming schemes did not affect the final representation. For OSDAs, the CIRpy (Chemical Identifier Resolver) Python package was used to determine the IUPAC name and SMILES string. If the OSDA name was not given in the paper, a chemistry expert determined the correct IUPAC name and SMILES. Each zeolite material was normalized to its International Zeolite Association (IZA) code through its list of known materials. Materials not in the IZA database were manually assigned the correct three letter code.

RDkit[49] was utilized to featurize the OSDA molecules. In addition to the canonical SMILES representation, physical and chemical properties of the organic molecules were also calculated, including molecular volume, surface area, charge, and WHIM descriptors. A total of 2,000 gas phase conformers for each molecule were generated, embedded, and optimized with the MMFF94 force field.[63] Average WHIM descriptors were calculated from the WHIM descriptors of all conformers. PCA transformations for the WHIM vectors were calculated using scikit-learn after each WHIM feature was standardized to remove the mean and scale to unit variance.

Zeolite structures were featurized with structural data obtained from the IZA database including framework density, maximum ring size, channel dimensionality, maximum included volume of a sphere, accessible volume, maximum channel area, and minimum channel area. Qualitative gel chemistry was one-hot encoded to describe the important components of zeolite synthesis. One-hot categories were the presence of Si, Al, Ge, P, Ti, B, Ga, Fe, Na, K, F, additional framework elements, additional cations, extra solvents in addition/instead of water, acid used in the synthesis, and other synthesis components.

Generative OSDA Model. The generative neural network borrowed heavily in both architecture and training protocol from Kotsias et al.,[42] but instead of using organic molecular descriptors, the model used zeolite and synthesis gel features as inputs. For each extracted synthesis route, the zeolite and synthesis are featurized and concatenated into the input vector, while the SMILES string of the OSDA is the output. To augment the training data, up to 100 different noncanonical versions of each OSDA's SMILES string are generated, resulting in training sets of approximately 150,000 points for the different train/test splits. This data augmentation has been shown to increase the accuracy of generative models for organic molecules.[64] The input is fed through 6 dense layers of 256 units with ReLU activation. Then, the data is fed through three unidirectional LSTM layers consisting of 256 units. Finally this output goes through a feedforward dense layer with 35 units having a softmax activation. Batch normalization is used on the first dense layers and LSTM layers. The model was implemented in Keras v2.2.4 with TensorFlowGPU v2.0.0 backend and trained using two NVIDIA Titan Xp GPUs.

The model was trained for 100 epochs on a variety of train/test splits to test various aspects of the generative model using the "teacher's forcing method".[65] The Adam optimizer with default parameters was used with a batch size of 128. A custom learning rate scheduler was used with an initial rate of 10^{-3} for 50 epochs, and then each epoch was exponentially decayed down to 10^{-6}. Four different training and testing splits were used to train the model. (1) The training and test split was chosen at random with 80% of the data used for training and 20% used for testing. (2) The training and test split was chosen so all data points resulting in CHA were isolated in the test set. This results in 5,398 training points and 265 (5%) testing points. (3) Data was split in the same manner as (2) but using AEI. This results in 5,555 training points and 108 (2%) testing points. (4) The final model was trained on the entire data set with no held out test set. Splits 1, 2, and 3 are used to evaluate the model's performance, while split 4 is used to look at specific zeolite systems CHA and SFW. Holding out an entire zeolite structure from the training tests the model's capability of suggesting new OSDA candidates for previously unseen zeolites and can confirm that the model is not memorizing pairs of OSDAs and zeolites, which can occur when randomly splitting. CHA and AEI were chosen due to their cage-like structure, industrial relevance, and presence of enough data to construct a large-enough test set for benchmarking.

OSDA generation followed the procedure outlined in Kotsias et al.[42] very closely. All generation occurred with multinomial sampling with the temperature parameter set equal to 1. Specific zeolite phases were manually chosen and paired with the appropriate chemistry conditions. For example, when looking at CHA zeolites, the CHA phase is paired with Si/F, Si/B, Si/Na, Si/K, Si/Al/Na, Si/Al/K, and Si/Al/F. For each zeolite/chemistry pair, 10,000 molecules are generated along with the negative log-likelihoods of generating that molecule.

Atomistic Simulations. Molecular mechanics simulations were performed using the General Utility Lattice Program (GULP),[66,67] version 5.1.1, through the GULPy package.[59] The Dreiding force field[68] was used to model interactions between the zeolite and the OSDA. The initial structure for the SFW zeolite was retrieved from the International Zeolite Association database and optimized using the Sanders–Leslie–Catlow force field.[69] Docking of OSDAs in SFW was performed using the VOID package using the default parameters.[60] Pose optimizations were performed at constant volume, and binding energies were calculated following the frozen pose method.[59]

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acscentsci.1c00024.

Additional figures and tables describing the data set, relationships between the OSDAs and zeolite structures, and metrics of model performance (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Elsa A. Olivetti − *Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States;* orcid.org/0000-0002-8043-2385; Phone: +1 617 2530877; Email: elsao@mit.edu

Authors

Zach Jensen − *Department of Materials Science and Engineering, Massachusetts Institute of Technology,*

https://doi.org/10.1021/acscentsci.1c00024
ACS Cent. Sci. 2021, 7, 858−867

Cambridge, Massachusetts 02139, United States; orcid.org/0000-0001-7635-5711

Soonhyoung Kwon − *Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States*

Daniel Schwalbe-Koda − *Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States;* orcid.org/0000-0001-9176-0854

Cecilia Paris − *Instituto de Tecnología Química, Universitat Politècnica de València-Consejo Superior de Investigaciones Científicas, 46022 Valencia, Spain*

Rafael Gómez-Bombarelli − *Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States;* orcid.org/0000-0002-9495-8599

Yuriy Román-Leshkov − *Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States;* orcid.org/0000-0002-0025-4233

Avelino Corma − *Instituto de Tecnología Química, Universitat Politècnica de València-Consejo Superior de Investigaciones Científicas, 46022 Valencia, Spain;* orcid.org/0000-0002-2232-3527

Manuel Moliner − *Instituto de Tecnología Química, Universitat Politècnica de València-Consejo Superior de Investigaciones Científicas, 46022 Valencia, Spain;* orcid.org/0000-0002-5440-716X

Complete contact information is available at: https://pubs.acs.org/10.1021/acscentsci.1c00024

Notes

The authors declare no competing financial interest.

The data, models, and code are provided in the following GitHub repository: www.github.com/olivettigroup/OSDA_Generator.

■ ACKNOWLEDGMENTS

The authors thank the Spanish Goverment under Awards "Severo Ochoa" (SEV-2016-0683) and RTI2018-101033-B-I00 (MCIU/AEI/FEDER, UE) and Generalitat Valenciana under Award AICO/2019/060 for support. We would like to acknowledge partial funding from the National Science Foundation DMREF Awards 1922311, 1922372, and 1922090, the Office of Naval Research (ONR) under contract N00014-20-1-2280, the MIT Energy Initiative, and MIT International Science and Technology Initiatives (MISTI) Seed Funds. Z.J. was supported by the Department of Defense (DoD) through the National Defense Science & Engineering Graduate (NDSEG) Fellowship Program. D.S.-K. was additionally funded by the MIT Energy Fellowship.

■ REFERENCES

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

(2) Martínez, C.; Corma, A. Inorganic molecular sieves: Preparation, modification and industrial application in catalytic processes. *Coord. Chem. Rev.* **2011**, *255*, 1558−1580.

(3) Li, Y.; Li, L.; Yu, J. Applications of zeolites in sustainable chemistry. *Chem.* **2017**, *3*, 928−949.

(4) Baerlocher, Ch.; McCusker, L. B. *Database of Zeolite Structures*; 2018, http://www.iza-structure.org/databases/.

(5) Kumar, M.; Choudhary, M. K.; Rimer, J. D. Transient modes of zeolite surface growth from 3D gel-like islands to 2D single layers. *Nat. Commun.* **2018**, *9*, 2129.

(6) Burkett, S. L.; Davis, M. E. Mechanisms of structure direction in the synthesis of pure-silica zeolites. 1. Synthesis of TPA/Si-ZSM-5. *Chem. Mater.* **1995**, *7*, 920−928.

(7) Shevlin, S. Looking deeper into zeolites. *Nat. Mater.* **2020**, *19*, 1038−1039.

(8) 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.

(9) Cundy, C. S.; Cox, P. A. The hydrothermal synthesis of zeolites: Precursors, intermediates and reaction mechanism. *Microporous Mesoporous Mater.* **2005**, *82*, 1−78.

(10) Corma, A.; Davis, M. E. Issues in the Synthesis of Crystalline Molecular Sieves: Towards the Crystallization of Low Framework-Density Structures. *ChemPhysChem* **2004**, *5*, 304−313.

(11) Lok, B.; Cannan, T.; Messina, C. The role of organic molecules in molecular sieve synthesis. *Zeolites* **1983**, *3*, 282−291.

(12) Lobo, R. F.; Zones, S. I.; Davis, M. E. Structure-direction in zeolite synthesis. *Top. Inclusion Sci.* **1995**, *6*, 47−78.

(13) Moliner, M.; Rey, F.; Corma, A. Towards the Rational Design of Efficient Organic Structure-Directing Agents for Zeolite Synthesis. *Angew. Chem., Int. Ed.* **2013**, *52*, 13880−13889.

(14) Burton, A. Recent trends in the synthesis of high-silica zeolites. *Catal. Rev.: Sci. Eng.* **2018**, *60*, 132−175.

(15) 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.

(16) 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.

(17) Moliner, M.; Serna, P.; Cantín, Á.; 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.

(18) 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.

(19) 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. *Nature Catalysis* **2018**, *1*, 547−554.

(20) Boyd, P. G.; Lee, Y.; Smit, B. Computational development of the nanoporous materials genome. *Nature Reviews Materials* **2017**, *2*, 17037.

(21) Chong, S.; Lee, S.; Kim, B.; Kim, J. Applications of machine learning in metal-organic frameworks. *Coord. Chem. Rev.* **2020**, *423*, 213487.

(22) Jablonka, K. M.; Ongari, D.; Moosavi, S. M.; Smit, B. Big-Data Science in Porous Materials: Materials Genomics and Machine Learning. *Chem. Rev.* **2020**, *120*, 8066.

(23) Xie, Y.; Zhang, C.; Hu, X.; Zhang, C.; Kelley, S. P.; Atwood, J. L.; Lin, J. Machine Learning Assisted Synthesis of Metal−Organic Nanocapsules. *J. Am. Chem. Soc.* **2020**, *142*, 1475−1481.

(24) 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.

(25) Serra, J. M.; Baumes, L. A.; Moliner, M.; Serna, P.; Corma, A. Zeolite synthesis modelling with support vector machines: a combinatorial approach. *Comb. Chem. High Throughput Screening* **2007**, *10*, 13−24.

(26) Muraoka, K.; Sada, Y.; Miyazaki, D.; Chaikittisilp, W.; Okubo, T. Linking synthesis and structure descriptors from a large collection of synthetic records of zeolite materials. *Nat. Commun.* **2019**, *10*, 4459.

(27) Schwalbe-Koda, D.; Jensen, Z.; Olivetti, E.; Gómez-Bombarelli, R. Graph similarity drives zeolite diffusionless transformations and intergrowth. *Nat. Mater.* **2019**, *18*, 1177−1181.
(28) 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 Cent. Sci.* **2019**, *5*, 892−899.
(29) Moliner, M.; Román-Leshkov, Y.; Corma, A. Machine Learning Applied to Zeolite Synthesis: The Missing Link for Realizing High-Throughput Discovery. *Acc. Chem. Res.* **2019**, *52*, 2971−2980.
(30) Kim, E.; Huang, K.; Saunders, A.; McCallum, A.; Ceder, G.; Olivetti, E. Materials synthesis insights from scientific literature via text extraction and machine learning. *Chem. Mater.* **2017**, *29*, 9436−9444.
(31) Kim, E.; Jensen, Z.; van Grootel, A.; Huang, K.; Staib, M.; Mysore, S.; Chang, H.-S.; Strubell, E.; McCallum, A.; Jegelka, S.; et al. Inorganic materials synthesis planning with literature-trained neural networks. *J. Chem. Inf. Model.* **2020**, *60*, 1194−1201.
(32) Mahbub, R.; Huang, K.; Jensen, Z.; Hood, Z. D.; Rupp, J. L.; Olivetti, E. Text mining for processing conditions of solid-state battery electrolytes. *Electrochem. Commun.* **2020**, *121*, 106860.
(33) Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science* **2018**, *361*, 360−365.
(34) Kim, B.; Lee, S.; Kim, J. Inverse design of porous materials using artificial neural networks. *Science advances* **2020**, *6*, eaax9324.
(35) Yao, Z.; Sanchez-Lengeling, B.; Bobbitt, N. S.; Bucior, B. J.; Kumar, S. G. H.; Collins, S. P.; Burns, T.; Woo, T. K.; Farha, O. K.; Snurr, R. Q.; Aspuru-Guzik, A. Inverse Design of Nanoporous Crystalline Reticular Materials with Deep Generative Models. *Nature Machine Intelligence* **2021**, *3*, 76−86.
(36) Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The rise of deep learning in drug discovery. *Drug Discovery Today* **2018**, *23*, 1241−1250.
(37) 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 Cent. Sci.* **2018**, *4*, 268−276.
(38) Segler, M. H.; Preuss, M.; Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **2018**, *555*, 604−610.
(39) Elton, D. C.; Boukouvalas, Z.; Fuge, M. D.; Chung, P. W. Deep learning for molecular design−a review of the state of the art. *Molecular Systems Design & Engineering* **2019**, *4*, 828−849.
(40) Schwalbe-Koda, D.; Gómez-Bombarelli, R. Generative Models for Automatic Chemical Design. *Lect. Notes Phys.* **2020**, *968*, 445−467.
(41) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Model.* **1988**, *28*, 31−36.
(42) Kotsias, P.-C.; Arús-Pous, J.; Chen, H.; Engkvist, O.; Tyrchan, C.; Bjerrum, E. J. Direct steering of de novo molecular generation with descriptor conditional recurrent neural networks. *Nature Machine Intelligence* **2020**, *2*, 254−265.
(43) Zones, S. Translating new materials discoveries in zeolite research to commercial manufacture. *Microporous Mesoporous Mater.* **2011**, *144*, 1−8.
(44) Dědeček, J.; Tabor, E.; Sklenak, S. Tuning the Aluminum Distribution in Zeolites to Increase their Performance in Acid-Catalyzed Reactions. *ChemSusChem* **2019**, *12*, 556−576.
(45) Knott, B. C.; Nimlos, C. T.; Robichaud, D. J.; Nimlos, M. R.; Kim, S.; Gounder, R. Consideration of the aluminum distribution in zeolites in theoretical and experimental catalysis research. *ACS Catal.* **2018**, *8*, 770−784.
(46) Li, C.; Vidal-Moya, A.; Miguel, P. J.; Dedecek, J.; Boronat, M.; Corma, A. Selective introduction of acid sites in different confined positions in ZSM-5 and its catalytic implications. *ACS Catal.* **2018**, *8*, 7688−7697.
(47) Wicker, J. G.; Cooper, R. I. Beyond rotatable bond counts: capturing 3D conformational flexibility in a single descriptor. *J. Chem. Inf. Model.* **2016**, *56*, 2347−2352.
(48) Todeschini, R.; Gramatica, P. The WHIM theory: New 3D molecular descriptors for QSAR in environmental modelling. *SAR and QSAR in Environmental Research* **1997**, *7*, 89−115.
(49) *RDKit: Open-source cheminformatics*; http://www.rdkit.org, Release 2013.
(50) Corma, A.; Rey, F.; Rius, J.; Sabater, M. J.; Valencia, S. Supramolecular self-assembled molecules as organic directing agent for synthesis of zeolites. *Nature* **2004**, *431*, 287−290.
(51) Boal, B. W.; Schmidt, J. E.; Deimund, M. A.; Deem, M. W.; Henling, L. M.; Brand, S. K.; Zones, S. I.; Davis, M. E. Facile synthesis and catalysis of pure-silica and heteroatom LTA. *Chem. Mater.* **2015**, *27*, 7774−7779.
(52) Mohapatra, S.; Yang, T.; Gómez-Bombarelli, R. Reusability report: Designing organic photoelectronic molecules with descriptor conditional recurrent neural networks. *Nature Machine Intelligence* **2020**, *2*, 749−752.
(53) Zones, S. I. Zeolite SSZ-13 and its method of preparation. US Patent US4,544,538, 1985.
(54) Gallego, E. M.; Li, C.; Paris, C.; Martín, N.; Martínez-Triguero, J.; Boronat, M.; Moliner, M.; Corma, A. Making Nanosized CHA Zeolites with Controlled Al Distribution for Optimizing Methanol-to-Olefin Performance. *Chem. - Eur. J.* **2018**, *24*, 14631−14635.
(55) Di Iorio, J. R.; Li, S.; Jones, C. B.; Nimlos, C. T.; Wang, Y.; Kunkes, E.; Vattipalli, V.; Prasad, S.; Moini, A.; Schneider, W. F.; et al. Cooperative and Competitive Occlusion of Organic and Inorganic Structure-Directing Agents within Chabazite Zeolites Influences Their Aluminum Arrangement. *J. Am. Chem. Soc.* **2020**, *142*, 4807−4819.
(56) Martín, N.; Moliner, M.; Corma, A. High yield synthesis of high-silica chabazite by combining the role of zeolite precursors and tetraethylammonium: SCR of NO x. *Chem. Commun.* **2015**, *51*, 9965−9968.
(57) 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.
(58) Xie, D.; McCusker, L. B.; Baerlocher, C.; Zones, S. I.; Wan, W.; Zou, X. SSZ-52, a zeolite with an 18-layer aluminosilicate framework structure related to that of the DeNOx catalyst Cu-SSZ-13. *J. Am. Chem. Soc.* **2013**, *135*, 10519−10524.
(59) Schwalbe-Koda, D.; Gomez-Bombarelli, R. Benchmarking binding energy calculations for organic structure-directing agents in pure-silica zeolites. *ChemRxiv. 13270184.v2*, 2020.
(60) Schwalbe-Koda, D.; Gomez-Bombarelli, R. Supramolecular Recognition in Crystalline Nanocavities Through Monte Carlo and Voronoi Network Algorithms. *J. Phys. Chem. C* **2021**, *125*, 3009−3017.
(61) Song, X.; Li, J.; Guo, Y.; Pan, Q.; Gan, L.; Yu, J.; Xu, R. Syntheses and Characterizations of Transition-Metal-Substituted Aluminophosphate Molecular Sieves—(C3N2H5)8—[M8Al16P24O96](M= Co, Mn, Zn) with Zeotype LAU Topology. *Inorg. Chem.* **2009**, *48*, 198−203.
(62) Gaslain, F. O.; White, K. E.; Cowley, A. R.; Chippindale, A. M. Control of framework stoichiometry in MeGaPO laumontites using 1-methylimidazole as structure-directing agent. *Microporous Mesoporous Mater.* **2008**, *112*, 368−376.
(63) Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490−519.
(64) Arús-Pous, J.; Johansson, S. V.; Prykhodko, O.; Bjerrum, E. J.; Tyrchan, C.; Reymond, J.-L.; Chen, H.; Engkvist, O. Randomized SMILES strings improve the quality of molecular generative models. *J. Cheminf.* **2019**, *11*, 71.
(65) Williams, R. J.; Zipser, D. A learning algorithm for continually running fully recurrent neural networks. *Neural computation* **1989**, *1*, 270−280.

https://doi.org/10.1021/acscentsci.1c00024
ACS Cent. Sci. 2021, 7, 858−867

(66) Gale, J. D. GULP: A computer program for the symmetry-adapted simulation of solids. *J. Chem. Soc., Faraday Trans.* **1997**, *93*, 629−637.
(67) Gale, J. D.; Rohl, A. L. The General Utility Lattice Program (GULP). *Mol. Simul.* **2003**, *29*, 291−341.
(68) Mayo, S. L.; Olafson, B. D.; Goddard, W. A. DREIDING: A generic force field for molecular simulations. *J. Phys. Chem.* **1990**, *94*, 8897−8909.
(69) Sanders, M. J.; Leslie, M.; Catlow, C. R. Interatomic potentials for SiO2. *J. Chem. Soc., Chem. Commun.* **1984**, 1271−1273.

https://doi.org/10.1021/acscentsci.1c00024
ACS Cent. Sci. 2021, 7, 858−867