A Chemically-Guided Generative Diffusion Model for Materials Synthesis Planning

Elton Pan, Soonhyoung Kwon, Sulin Liu, Mingrou Xie,
Yifei Duan, Killian Sheriff, Yuriy Roman-Leshkov
Rafael Gomez-Bombarelli, Elsa Olivetti*
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
{eltonpan, s1105hk, sulinliu, mrx, duanyf99
ksheriff, yroman, rafagb, elsao}@mit.edu

Thorben Prein
Technische Universität München
Arcisstraße 21, 80333 München, Germany
t.prein@tum.de

Manuel Moliner
ITQ-UPV
Av. dels Tarongers, 46022, Valencia, Spain
mmoliner@itq.upv.es

Abstract

Data-driven synthesis planning is a crucial step in the discovery of novel materials with desirable properties. Zeolites are crystalline nanoporous materials with applications in catalysis, adsorption, and ion exchange. The synthesis of zeolitic materials remains a significant challenge due to its high-dimensional synthesis space and intricate structure-synthesis relationships. Considering the *one-to-many* relationship between structure and synthesis, we propose a generative modeling approach using a chemically-guided diffusion model for materials synthesis planning. Given a target zeolite structure and organic structure-directing agent (OSDA) as inputs, the diffusion model generates probable synthesis routes and achieves state-of-the-art performance compared to regression and deep generative models. The model learns chemically meaningful relationships, generating realistic synthesis routes that closely follow the distribution of literature-reported synthesis routes. As such, this approach could enable the discovery of zeolitic materials beyond domain-specific heuristics and trial-and-error experimentation.

1 Introduction

With materials informatics being largely focused on material property prediction, the synthesis of materials remains a bottleneck with computational approaches for guiding the synthesis process being less explored [1]. Accurate synthesis prediction is crucial for realizing materials in the lab, but materials synthesis is often complex and difficult to model computationally. Conventional ML approaches for modeling synthesis primarily rely on regression-based methods [2, 3], which attempt to deterministically map structure to synthesis conditions. Although useful, these methods often fall short due to their inability to capture the *one-to-many* nature of structure-synthesis relationships [2, 4]. Generative models have shown promise in modeling one-to-many relationships in materials science, such as property-composition (composition generation) [5] and property-structure (crystal structure generation) [6], but their application to synthesis has yet to be explored.

In this work, we propose a generative approach for materials synthesis prediction and contrast it with regression-based approaches. We present empirical evidence (in the case of zeolitic materials) that

38th Conference on Neural Information Processing Systems (NeurIPS 2024) AI4Mat Workshop.

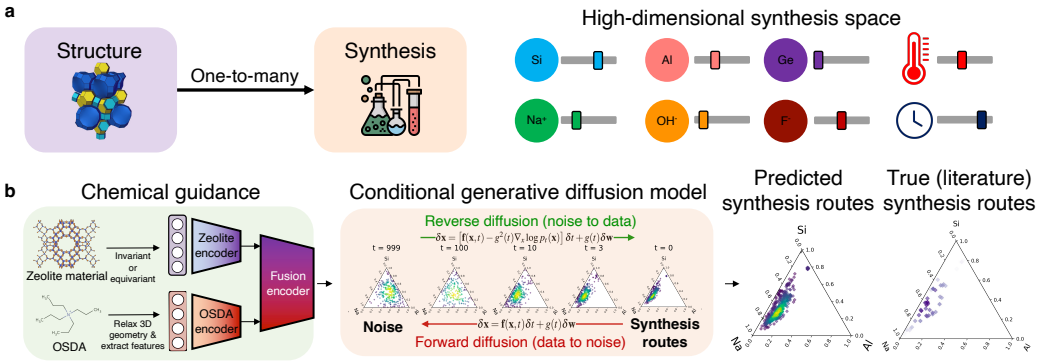


Figure 1: **A diffusion approach for materials synthesis planning**. **(a)** Materials, such as zeolites, have *one-to-many* structure-synthesis relationships i.e. a structure can be realized at a range of synthesis conditions (e.g., precursor ratios, reaction temperatures), which is known as the *synthesis window*. **(b)** Given a desired zeolite material structure, the generative model generates a *distribution* of synthesis routes, capturing the one-to-many structure-synthesis relationship. First, two separate encoders encode the zeolite material and organic structure-directing agent (OSDA) before their latent representations are fused via a fusion encoder. Second, the joint representation serves as chemical guidance for the diffusion model, which samples from noise via reverse diffusion (green arrow) to generate probable synthesis routes (ternary phase diagram) that matches the distribution of ground-truth synthesis routes reported in literature.

generative approaches outperform regression-based approaches, and propose that this is largely due to the *one-to-many* nature of structure-synthesis relationships. Our main contributions are:

1. We analyze why generative approaches are better suited for materials synthesis tasks compared to regression.
2. We demonstrate the first-known application of a chemically-guided denoising diffusion model for materials synthesis prediction, which exhibits state-of-the-art performance compared to other deep generative and regression-based approaches.

2 Methods

2.1 Zeolite and OSDA featurization

Zeolite We adopt two different approaches of encoding the zeolite structure: 1) Invariant features i.e. their structural properties (e.g., ring sizes, largest included sphere) are retrieved from the International Zeolite Association (IZA) database [7]. These serve as inputs into a MLP encoder. 2) Equivariant graph neural network (EGNN) [8] that encodes the zeolite as a graph (see Section A.3). **OSDA** Pose optimizations are performed at constant volume [9] using molecular mechanics GULPy (General Utility Lattice Program) package [10] with the MMFF94 force field [11]. Each OSDA is featurized using its physicochemical descriptors (e.g., molecular volume and 2D shape descriptors) of the organic molecule [12, 13]. The OSDA features are defined in Table 2. The zeolite embedding is concatenated with OSDA embedding and further encoded using the fusion encoder before the joint embedding guides the reverse diffusion process to generate synthesis routes (Fig. 1b). Model performance across denoising diffusion trajectory can be found in Fig. 8.

2.2 Denoising diffusion probabilistic models

Denoising diffusion probabilistic models (DDPMs) [14] are generative models that utilize a diffusion process to generate data by reversing a forward process that incrementally adds noise to the data. The forward process is defined by a Markov chain that gradually corrupts data x_0 into a noisy sample x_t via addition of Gaussian noise: $q\left(x_t \mid x_{t-1}\right) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t} x_{t-1}, \beta_t \mathbf{I}\right)$ with β_t representing the variance at timestep t. The generative process (green arrow in Fig. 1b) learns to reverse this corruption step by step, sampling from a distribution $p_\theta\left(x_{t-1} \mid x_t\right)$ parameterized by a neural network. In this

work, we use a U-Net [15]. This reverse process can also be represented as a Gaussian distribution $p_\theta\left(x_{t-1} \mid x_t\right)=\mathcal{N}\left(x_{t-1} ; \mu_\theta\left(x_t, t\right), \Sigma_\theta\left(x_t, t\right)\right)$. Here, we learn a score function

$$s_\theta\left(x_t, t\right)=\nabla_{x_t} \log p_\theta\left(x_{t-1} \mid x_t\right)=-\frac{1}{\Sigma_\theta\left(x_t, t\right)}\left(x_t-\mu_\theta\left(x_t, t\right)\right) \tag{1}$$

where μ_θ and Σ_θ are the mean and variance of the score model s_θ, respectively.

2.3 Chemically-guided diffusion model

In standard guided diffusion models, a classifier is used to guide the generation process by adjusting the score to steer the model towards specific target classes. In contrast, classifier-free guidance [16] eliminates the need for a separate classifier by conditioning the diffusion model directly on the desired attributes. During training, the score function $\tilde{s}_\theta$ is trained both with and without conditioning information. Sampling is then performed using the following linear combination of the conditional (concatenation with c) and unconditional (concatenation with null token $\varnothing$) score estimates:

$$\tilde{s}_\theta\left(x_t, t, c\right)=(1+w) s_\theta\left(x_t, t, c\right)-w s_\theta\left(x_t, t, \varnothing\right) \tag{2}$$

where c refers to the chemical guidance from zeolite and OSDA embeddings shown in Fig. 1b, while w is the strength of the chemical guidance. For implementation details, refer to Section A.3.

3 Experimental setup

3.1 Dataset and task

We use the ZeoSyn dataset [17], a large zeolite synthesis dataset comprising 23,961 synthesis routes, 233 zeolite topologies and 921 organic structure-directing agents (OSDAs). For more details on dataset, refer to Section A.1. Given a target zeolite material x_{zeolite} and an OSDA x_{OSDA}, the task is to predict an *ensemble* of synthesis routes consisting of gel compositions $\{y_{\text{gel}}\}$ (e.g. Si/Al) and reaction conditions $\{y_{\text{cond}}\}$ (e.g. crystallization temperature). We refer to these as *synthesis parameters*. See Table 1 for a complete list of synthesis parameters, and Fig. 5 for an example on reaction conditions.

3.2 Metrics

For each test zeolite-OSDA system, we sample 1000 synthesis routes using the model and compute the following metrics with reference to *unseen* synthesis routes reported in literature. **Wasserstein distance** measures the distance between two probability distributions by finding the minimum cost to move probability mass from one distribution to another [18]. **Coverage** Inspired by Xu [6], we use two coverage metrics, COV-R (recall) and COV-P (precision), to measure the similarity between sets of generated and literature-reported synthesis for each zeolite-OSDA system. Intuitively, COV-R measures the fraction of literature synthesis routes being correctly predicted, and COV-P measures the fraction of generated synthesis routes being probable. COV-F1 is computed as the harmonic mean of COV-R and COV-P. Refer to Section A.2 for a detailed justification of the metrics.

3.3 Baselines

Random A random dummy baseline corresponds to picking a random point in synthesis space. **AMD** Schwalbe-Koda et al. proposed a deterministic regression-based approach using average minimum distance (AMD) for zeolite structural featurization for synthesis prediction task [19]. **BNN** Bayesian neural networks [20] extend standard neural networks by incorporating Bayesian inference by treating network weights as probability distributions, enabling a distribution of outputs. We implement a classical generative approach, Gaussian mixture model (**GMM**) [21], which models data probabilistically as a sum of Gaussians (each with its mean and covariance). We also implement standard deep generative baselines: Conditional variational autoencoder (**VAE**) [2, 4], generative adversarial network (**GAN**) [22] and normalizing flow (**NF**). For NF, we use the real-valued non-volume preserving (RealNVP) transformations [23].

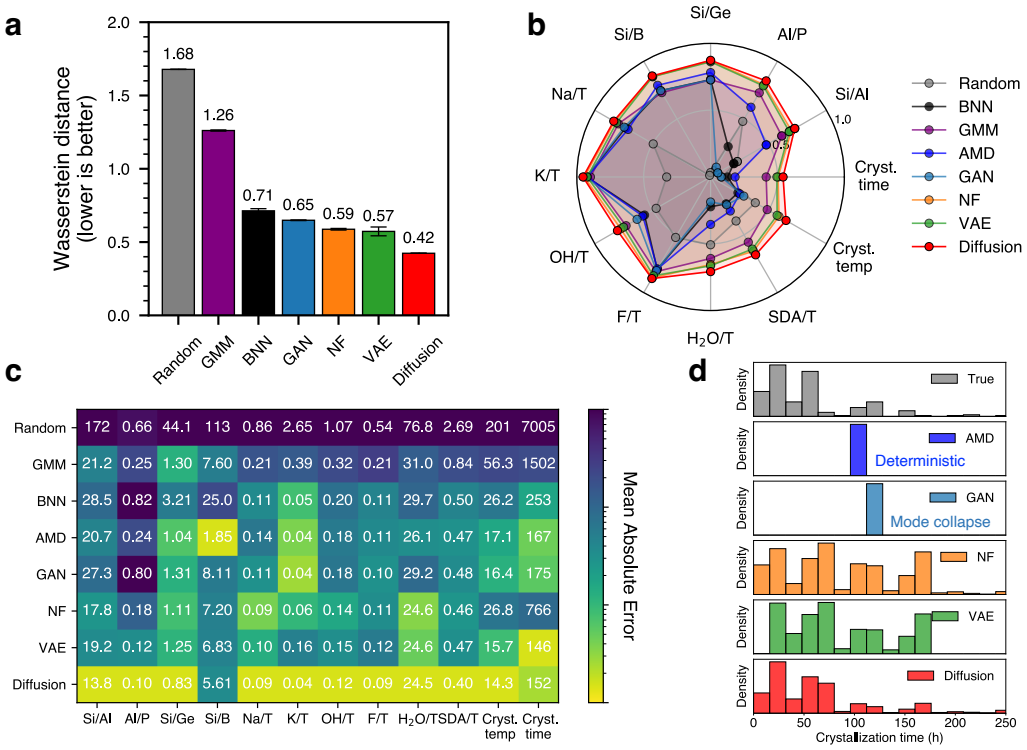


Figure 2: **Diffusion model achieves state-of-the-art performance in materials synthesis prediction**. **(a)** Wasserstein distance (lower is better) between generated and literature synthesis routes **(b)** COV-F1 (higher is better, ranging from 0 to 1) of individual synthesis parameters **(c)** Mean absolute error between the means of distribution of generated and literature synthesis parameters **(d)** Distribution of predicted synthesis parameter across various different modeling approaches. AMD (purple) is regression-based (outputs deterministic, single-point prediction), while GAN (light blue) suffers from mode collapse. NF and VAE output distributions that do not match the ground-truth. In contrast, diffusion model (red) is able to accurately capture the literature distribution (grey) of the synthesis parameter (*one-to-many* structure-synthesis relationship).

4 Results and Discussion

4.1 Generative approaches better model structure-synthesis relationships

Wasserstein distance We compare the performance of different generative approaches of predicting zeolite synthesis route by computing the mean Wasserstein distance (lower is better) between generated routes and those reported in the literature of unseen zeolite-OSDA systems (Fig. 2a). Classical generative approaches like GMM (purple) do not perform much better than a random baseline (grey). Probabilistic regression-based approaches like BNN (black) perform better than GMM. However, deep generative models such as GAN, NF, VAE and diffusion outperform the classical approaches, with the diffusion model (red) significantly outperforming the next best baseline (VAE) by over 25%.

Coverage The optimal generative modeling approach should ideally maximize both COV-P and COV-R simultaneously, hence we consider their harmonic mean (COV-F1) as the metric (see detailed explanation in Section A.2). Fig. 2b shows the COV-F1 of 12 zeolite synthesis parameters (e.g. Si/Al). Overall, the models perform better on synthesis parameters related to heteroatoms (Si/Al, Al/P, Si/Ge, Si/B), cations (Na^+/T, K^+/T) and anions (F^-/T, OH^-/T). They struggle in terms of synthesis conditions such as crystallization temperature and time. The latter is especially evident for regression-based AMD (blue, see right half of the plot). GAN (light blue) follows the same trend as AMD. Two key observations are as follows: **1)** Deep generative approaches (diffusion, VAE and NF)

outperform regression-based (AMD, BNN) approaches. This is due to generative models being better at recalling the unseen literature distribution (Fig. 3 right). **2)** The diffusion model outperforms other deep generative baselines. This is because diffusion is able to generate higher quality/more probable synthesis routes as seen in its high precision (Fig. 3 left). This result is further supported by diffusion achieving the lowest MAE for 10/12 synthesis parameters as shown in Fig. 2c.

Rationalizing failure modes We uncover why generative models are better suited for materials synthesis modeling. For a given target material, there often exists a range of possible synthesis parameters (e.g. reaction temperature) at which a material can be synthesized. This means that the same zeolite material can synthetically realized at a joint *distribution* of synthesis parameters. This renders structure-synthesis relationship *one-to-many* instead of one-to-one. We refer to this as the *synthesis window*, which can be visualized using the kinetic phase diagram shown Fig. 1b. For the sake of simplicity, we consider a synthesis parameter by visualizing the ground-truth literature (grey) crystallization time in Fig. 2d, which exists as a *distribution*. Regression-based AMD (purple) is deterministic, and hence outputs a point prediction (weighted average of the distribution). Unfortunately, generative models like GAN (light blue) also outputs a point prediction due to its propensity to suffer from mode collapse [24]. While models such as NF (orange) and VAE (green) address the mode collapse issue, they fail to accurately capture the ground-truth literature distribution (grey). In contrast, the diffusion model (red) is able to accurately capture the ground-truth distribution. This is further corroborated by Fig. 4, which shows the generated synthesis parameters following closely those reported in literature for all synthesis parameters. Moreover, the model can generalize to *unseen* zeolite-OSDA systems (Fig. 5) and captures phase boundaries between competing phases (Fig. 6). Therefore, this rationalizes why generative approaches are better suited for modeling structure-synthesis relationships for materials.

4.2 Model captures chemically meaningful relationships

For the chemical guidance described in Fig. 1b, the zeolite and OSDA encoders both learn smooth and continuous latent spaces with respect to properties of zeolites and OSDA, respectively (Fig. 9 and 10). Given that the diffusion model learns the joint distribution of synthesis parameters, we inspect two synthesis parameters (crystallization temperature and time) in Fig. 11 for two *unseen* zeolite-OSDA systems. An inverse relationship is observed between generated temperatures and times. This observation aligns well with the Arrhenius equation $k = Ae^{\frac{-E_a}{RT}}$, where rate constant (related to crystallization time) is inversely related to temperature. In addition, the model captures domain-specific heuristics (Fig. 12) and thermodynamics of zeolite formation (Fig. 13), a strong testament to the model successfully learning structure-synthesis relationships.

5 Broader Impact

The ability to propose probable and diverse synthesis pathways in a computationally efficient manner is a crucial step toward bridging computational materials design (*what* to synthesize) and synthesis planning (*how* to synthesize) in order for AI-guided materials design and automated synthesis to realistically make a tangible impact on tackling pressing modern-day challenges.

References

[1] Thorben Prein et al. "MTENCODER: A Multi-task Pretrained Transformer Encoder for Materials Representation Learning". In: *AI for Accelerated Materials Design-NeurIPS 2023 Workshop*. 2023.

[2] Christopher Karpovich et al. "Interpretable machine learning enabled inorganic reaction classification and synthesis condition prediction". In: *Chemistry of Materials* 35.3 (2023), pp. 1062–1079.

[3] Thorben Prein et al. "Reaction Graph Networks for Inorganic Synthesis Condition Prediction". In: *AI for Accelerated Materials Design - NeurIPS 2024*. 2024. URL: `https://openreview.net/forum?id=VGsXQOTs1E`.

[4] Christopher Karpovich et al. "Inorganic synthesis reaction condition prediction with generative machine learning". In: *arXiv preprint arXiv:2112.09612* (2021).

[5] Elton Pan, Christopher Karpovich, and Elsa Olivetti. "Deep reinforcement learning for inverse inorganic materials design". In: *arXiv preprint arXiv:2210.11931* (2022).

[6] Tian Xie et al. "Crystal diffusion variational autoencoder for periodic material generation". In: *arXiv preprint arXiv:2110.06197* (2021).

[7] C Baerlocher and L. B. McCusker. "Database of zeolite structures". In: *http://www.iza-structure.org/databases/* (2021).

[8] Vıctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. "E (n) equivariant graph neural networks". In: *International conference on machine learning*. PMLR. 2021, pp. 9323–9332.

[9] Daniel Schwalbe-Koda and Rafael Gómez-Bombarelli. "Benchmarking binding energy calculations for organic structure-directing agents in pure-silica zeolites". In: *The Journal of Chemical Physics* 154.17 (2021).

[10] Julian D Gale. "GULP: A computer program for the symmetry-adapted simulation of solids". In: *Journal of the Chemical Society, Faraday Transactions* 93.4 (1997), pp. 629–637.

[11] Thomas A Halgren. "MMFF VI. MMFF94s option for energy minimization studies". In: *Journal of computational chemistry* 20.7 (1999), pp. 720–729.

[12] Daniel Schwalbe-Koda et al. "A priori control of zeolite phase competition and intergrowth with high-throughput simulations". In: *Science* (2021), eabh3350.

[13] Zach Jensen et al. "Discovering Relationships between OSDAs and Zeolites through Data Mining and Generative Neural Networks". In: *ACS Central Science* 7.5 (2021), pp. 858–867.

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models". In: *Advances in neural information processing systems* 33 (2020), pp. 6840–6851.

[15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer. 2015, pp. 234–241.

[16] Jonathan Ho and Tim Salimans. "Classifier-free diffusion guidance". In: *arXiv preprint arXiv:2207.12598* (2022).

[17] Elton Pan et al. "ZeoSyn: A Comprehensive Zeolite Synthesis Dataset Enabling Machine-Learning Rationalization of Hydrothermal Parameters". In: *ACS Central Science* 10.3 (2024), pp. 729–743.

[18] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. "A metric for distributions with applications to image databases". In: *Sixth international conference on computer vision (IEEE Cat. No. 98CH36271)*. IEEE. 1998, pp. 59–66.

[19] Daniel Schwalbe-Koda et al. "Inorganic synthesis-structure maps in zeolites with machine learning and crystallographic distances". In: *Digital Discovery* 2.6 (2023), pp. 1911–1924.

[20] Laurent Valentin Jospin et al. "Hands-on Bayesian neural networks—A tutorial for deep learning users". In: *IEEE Computational Intelligence Magazine* 17.2 (2022), pp. 29–48.

[21] Douglas A Reynolds et al. "Gaussian mixture models." In: *Encyclopedia of biometrics* 741.659-663 (2009).

[22] Mehdi Mirza and Simon Osindero. "Conditional generative adversarial nets". In: *arXiv preprint arXiv:1411.1784* (2014).

[23] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. "Density estimation using real nvp". In: *arXiv preprint arXiv:1605.08803* (2016).

[24] Zhaoyu Zhang, Mengyan Li, and Jun Yu. "On the convergence and mode collapse of GAN". In: *SIGGRAPH Asia 2018 Technical Briefs*. 2018, pp. 1–4.

[25] Miguel A Camblor, Luis A Villaescusa, and MJ Diaz-Cabanas. "Synthesis of all-silica and high-silica molecular sieves in fluoride media". In: *Topics in Catalysis* 9 (1999), pp. 59–76.

[26] Thiquynhxuan Le et al. "Process regulation of microwave intensified synthesis of Y-type zeolite". In: *Microporous and Mesoporous Materials* 284 (2019), pp. 476–485.

[27] Killian Sheriff, Yifan Cao, and Rodrigo Freitas. "Chemical-motif characterization of short-range order with E (3)-equivariant graph neural networks". In: *arXiv preprint arXiv:2405.08628* (2024).

A Appendix

A.1 ZeoSyn dataset and preprocessing

The ZeoSyn dataset [17] contains comprehensive synthesis information on zeolites including gel composition, reaction conditions (crystallization time/temperature), precursors, and OSDAs. The dataset also includes the resulting zeolite structures formed (or lack thereof e.g., dense/amorphous phases) for each synthesis route. The dataset consists of 23,961 synthesis routes from 3,096 journal articles spanning the years 1966–2021. It contains data on 921 unique OSDA molecules, 233 zeolite structures, and 1,022 unique materials. The gel compositions are a combination of 51 different gel components including Si, Al, P, Na^+, K^+, OH^-, F^-, Ge, Ti, B, Ga, V, OSDA, H_2O, and additional solvents. In the context of zeolite synthesis, the OSDA refers to an organic molecule that act as a 'template' that guides the arrangement of inorganic building blocks (e.g. Si, Al) to form the zeolite material.

For data splits, a 80/20 train/test split via zeolite-OSDA systems. This is to prevent data leakage across systems and ensure the model generalizes to unseen chemical systems. This results in 1856 and 464 systems in the training and test sets, respectively. Furthermore, the training set is further partitioned via a 87.5/12.5 train/validation random split for hyperparameter tuning. Given most synthesis parameters do not follow the Gaussian distribution (see blue distributions in Fig. 4), we perform quantile transformation using `sklearn.preprocessing.QuantileTransformer` to map the raw synthesis parameters to a uniform distribution.

A.2 Coverage metric

We define 2 metrics to compare two distributions of synthesis parameters: Synthesis routes generated by a model $\{\boldsymbol{S}_k\}_{k\in[1\ldots K]}$ and synthesis routes reported in literature $\{\boldsymbol{S}_l\}_{l\in[1\ldots L]}$. This method is inspired by [6].

For each synthesis parameter, we calculate the Euclidean distance between generated and literature synthesis parameter, which we call $D_{\text{syn}}(\boldsymbol{S}_k, \boldsymbol{S}_l)$. Following the well-known classification metrics of precision and recall, we define the coverage metrics as:

$$\text{COV-R (Recall)} = \frac{1}{L}\Big|\{l \in [1..L] : \exists k \in [1..K], D_{\text{syn}}(\boldsymbol{S}_k, \boldsymbol{S}_l) < \delta_{\text{syn}}\}\Big| \tag{3}$$

$$\text{COV-P (Precision)} = \frac{1}{K}\Big|\{k \in [1..K] : \exists l \in [1..L], D_{\text{syn}}(\boldsymbol{S}_l, \boldsymbol{S}_k) < \delta_{\text{syn}}\}\Big| \tag{4}$$

where COV is Coverage. The recall metrics measure how many ground synthesis routes are correctly recalled by the generated synthesis routes, while the precision metrics measure how many generated synthesis routes are probable (close to literature-reported routes). The threshold δ_{syn} is defined by a domain expert in zeolite synthesis in Table 1. Finally, the overall metric is COV-F1, which is defined as the harmonic mean of COV-R and COV-P i.e. $\text{COV-F1} = \frac{2}{\frac{1}{\text{COV-R}} + \frac{1}{\text{COV-P}}}$, which ranges from $0-1$ (higher is better). An ideal generative model for materials synthesis planning should maximize its quality of generated synthesis routes (high precision) while maximizing their diversity (high recall). As such, COV-F1 is used as the metric to evaluate the models.

A.3 Model implementation

Diffusion model training and sampling We train a conditional DDPM with U-Net [15] as the score function, with an input dimension of 12, followed by subsequent layers dimensions multipliers of 128, 64 and 32 for downsampling and the opposite for upsampling. The score function is trained with $T = 1000$, with a $p_{\text{uncond}} = 0.1$ for classifer-free guidance with exponential moving average ($\beta = 0.99$), batch size of 32 at a constant learning rate of 4×10^{-4} for 1M epochs. For sampling, $w = 1.0$ was found to be optimal (see Fig. 7).

Equivariant graph encoder We implement an equivariant graph neural network (EGNN) using the `e3nn` package (similar to [27]) to encode the structure of the zeolite as a graph. Here, we use a single layer EGNN with $l_{\text{max}} = 2$, $n_{\text{neighbors}} = 10$ and $n_{\text{nodes}} = 140$. In this case, we assume a

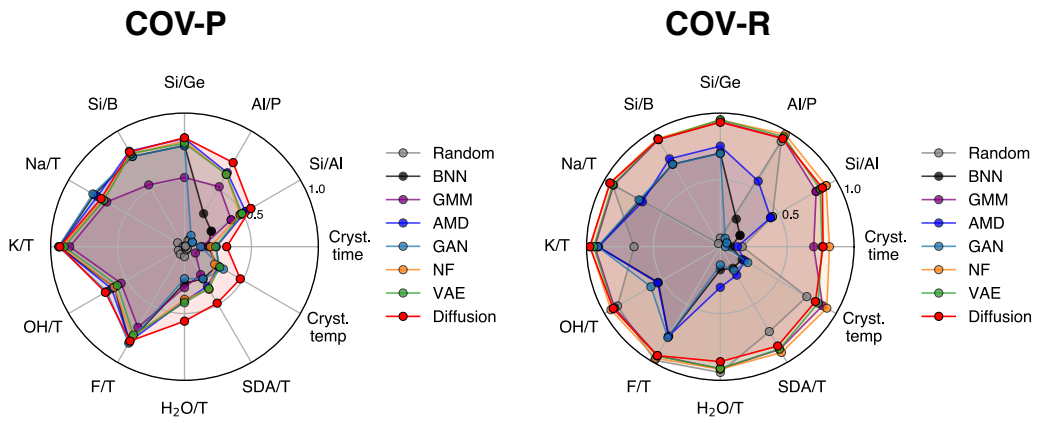


Figure 3: **COV-P (precision) and COV-R (recall) for all models.** COV-F1 is the mean of COV-P and COV-R. Notice that high COV-F1 for generative models such as diffusion and VAE (vs. AMD) largely driven by higher COV-R, while high COV-F1 for diffusion (vs. other generative models) largely driven by higher COV-P.

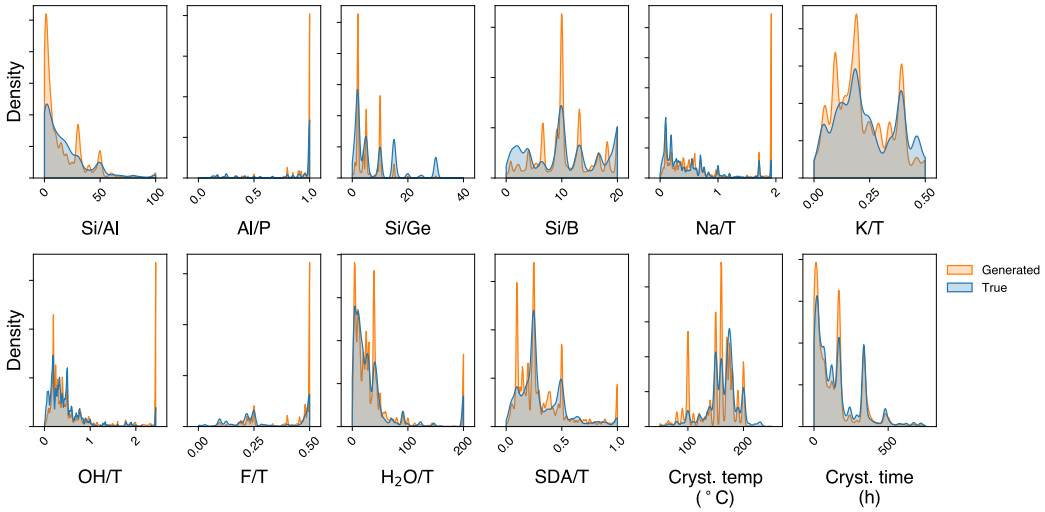


Figure 4: **Model accurately captures unseen synthesis routes reported in literature.** Note synthesis parameters are aggregated across all *unseen* zeolite-OSDA systems. The generated distribution (orange) follows closely the true distribution (blue). However, generated distribution tends to be more multi-modal and 'spiky' (e.g. Si/Al, crystallization temperature).

pure-silica zeolite structure obtained from the IZA database [7], and initialize node embeddings as one-hot according to the element identity (Si or O).

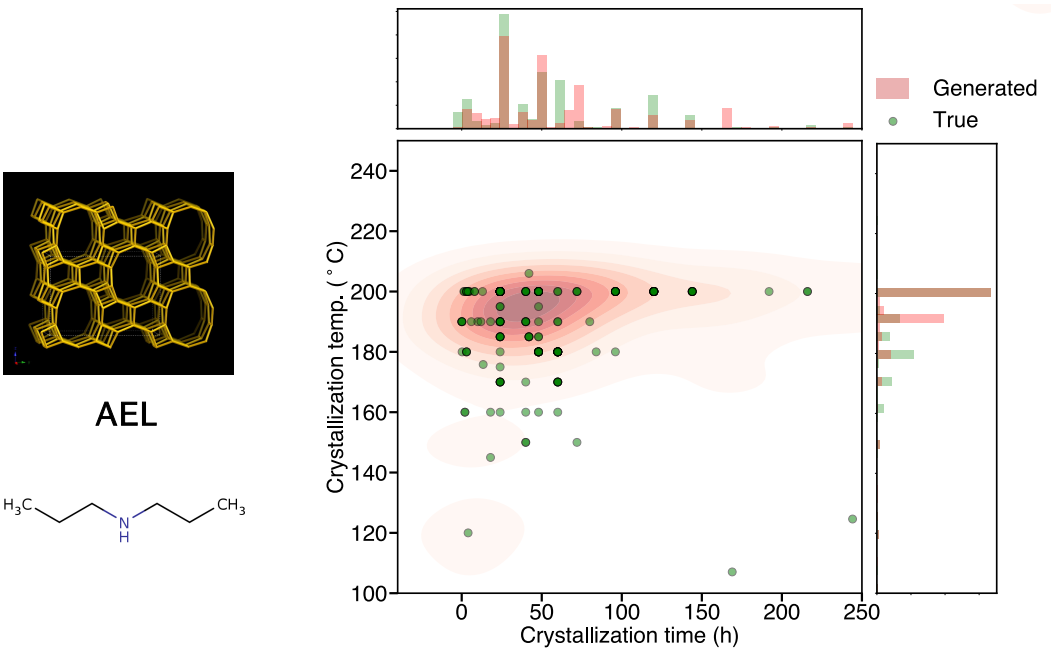


Figure 5: **Diffusion-generated $\{y^{\text{gen}}_{\text{cond}}\}$ vs. ground-truth $\{y^{\text{true}}_{\text{cond}}\}$ reaction conditions for an unseen zeolite-OSDA system** (AEL zeolite with dipropylamine OSDA). Here, the reaction conditions are crystallization temperature and time. The generated reaction conditions (red) accurately captures the distribution of literature-reported conditions. However, a minority (see 2 points at bottom right) are not captured by the diffusion model. This two points are outliers (very low crystallization temperature and very long crystallization time).

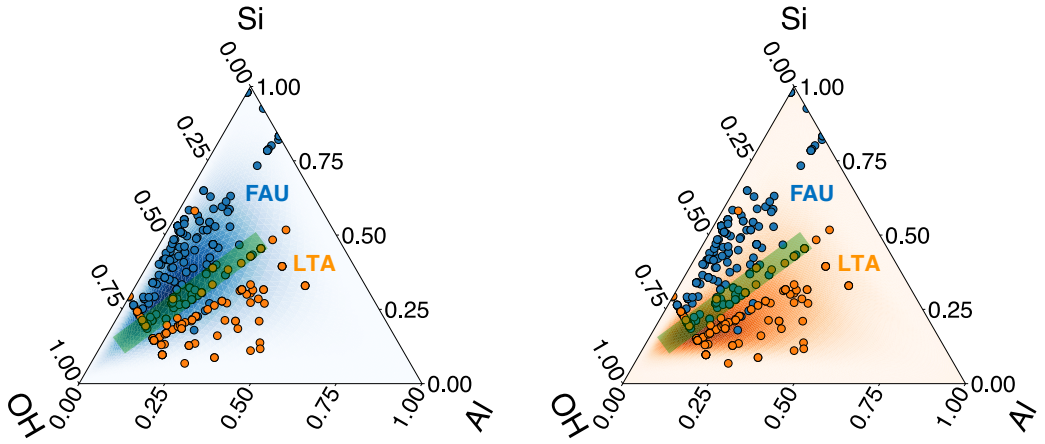


Figure 6: **Diffusion-generated gel compositions captures boundaries between competing phases.** Circles refer to literature-reported OSDA-free syntheses of two competing phases, FAU and LTA. Heatmaps show the generated distribution of OSDA-free synthesis recipes, where blue and orange heatmaps refer to FAU and LTA phases respectively. We observe that the model accurately captures the phase boundary (green shaded region) between FAU and LTA.

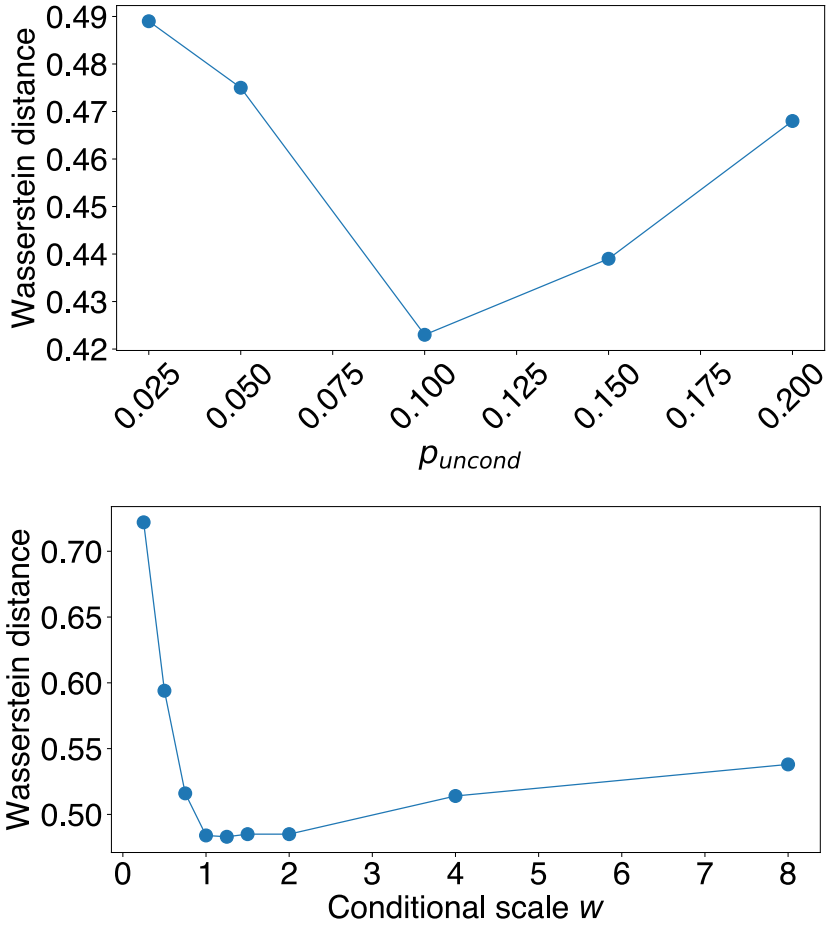
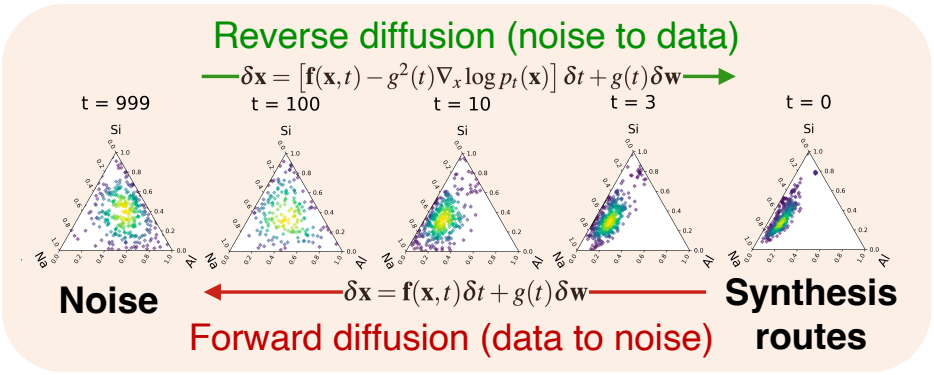


Figure 7: **Influence of DDPM hyperparameters on model performance** in terms of p_{uncond} and conditional generation scale w.



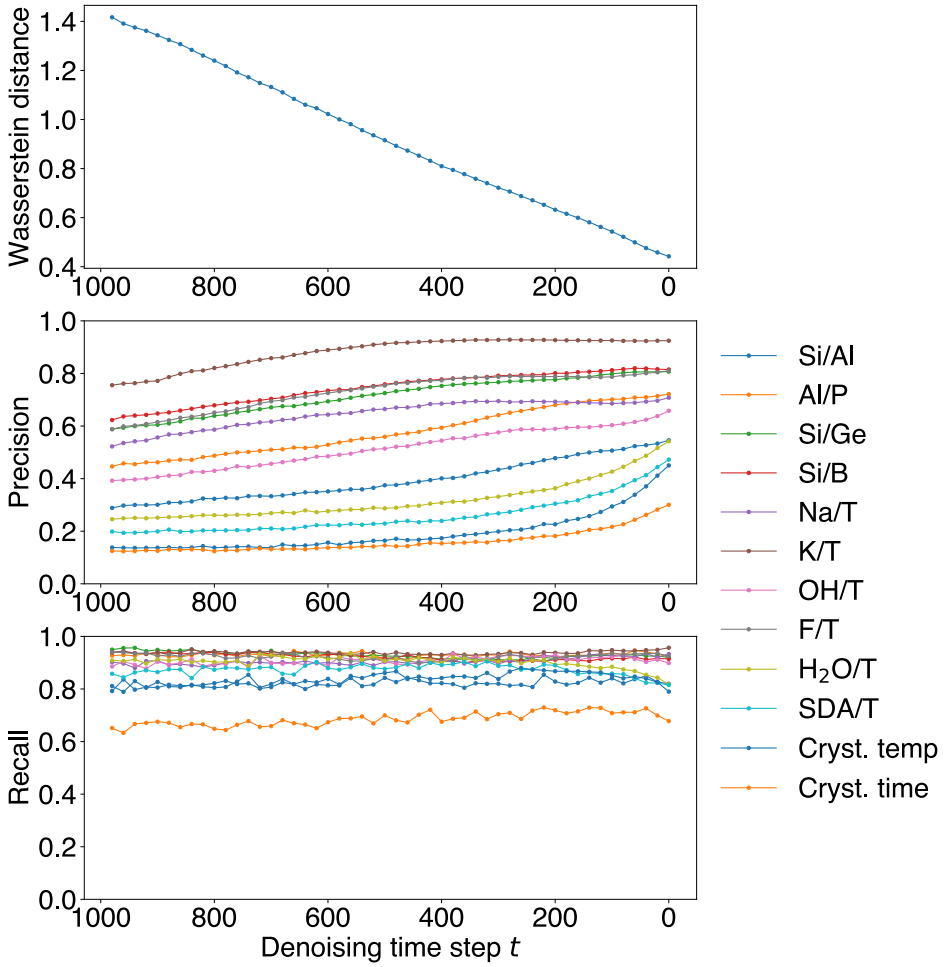


Figure 8: **Model performance across denoising diffusion trajectory** measured by Wasserstein distance (↓), precision (↑) and recall (↑). At inference, reverse diffusion (green arrow) maps noise (at $t = T$) to probable synthesis routes (at $t = 0$), leading to a monotonic improvement in Wasserstein distance (between generated and ground truth distributions) and precision. In contrast, recall remains relatively constant across the trajectory. This shows that, in reverse time, the diffusion model is able to generate higher quality synthesis recipes while maintaining their diversity, hence capturing the *one-to-many* structure-synthesis relationship for zeolite materials.

Table 1: **Expert-defined thresholds δ_{syn}.** Values are based on practical utility in zeolite synthesis. Note $T = \sum_i n_i$ where n_i is the amount of the ith heteroatom (e.g., Si, Al, Ge, B) present in synthesis.

Synthesis parameter	δ_{syn} (unitless unless otherwise specified)
Si/Al	1.5
Al/P	0.05
Si/Ge	0.05
Si/B	0.05
Na^+/T	0.05
K^+/T	0.05
OH^-/T	0.05
F^-/T	0.05
H_2O/T	5
SDA/T	0.05
Crystallization temperature	5 °C
Crystallization time	12 h

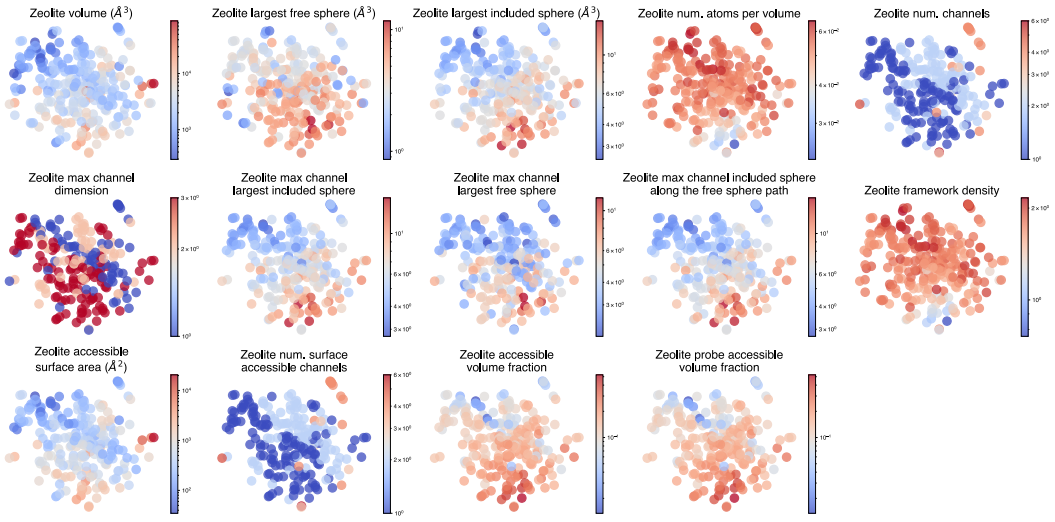


Figure 9: **Learned zeolite representations are chemically meaningful.** Latent space is smooth and continuous with respect to zeolite physical properties (e.g., accessible surface area).

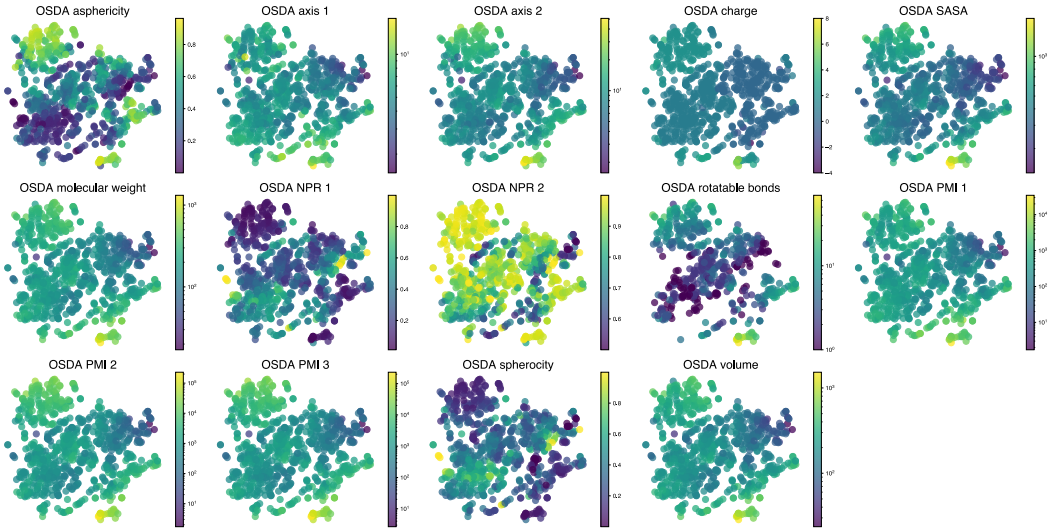


Figure 10: **Learned OSDA representations are chemically meaningful.** Latent space is smooth and continuous with respect to OSDA physical properties (e.g., volume).

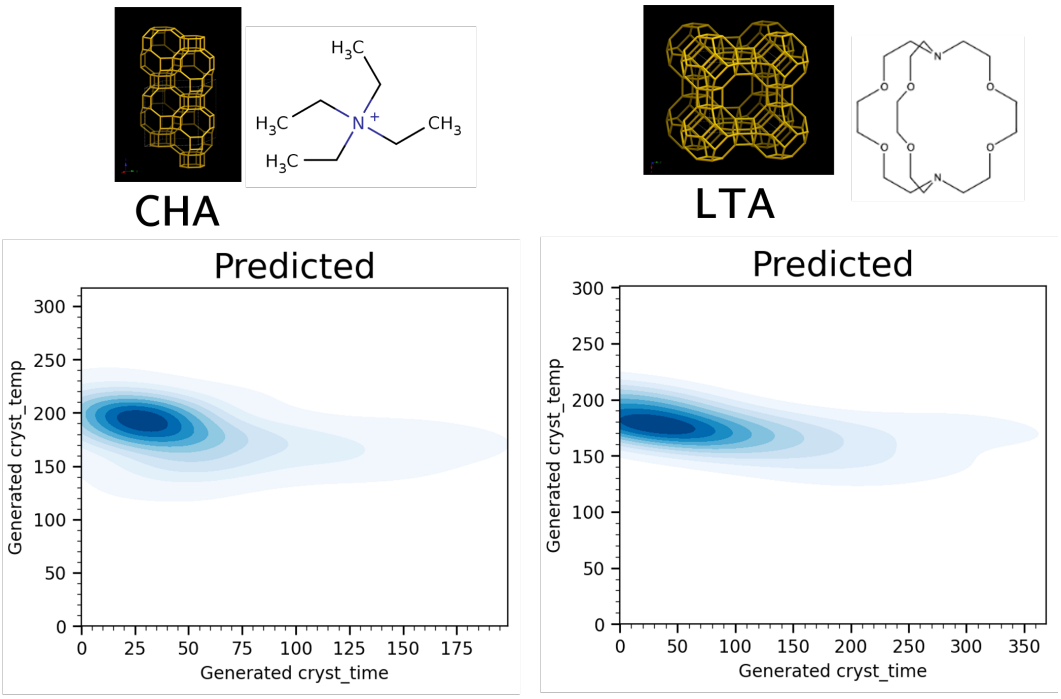


Figure 11: **Model captures physical relationships between synthesis parameters.** In the above 2 examples, generated crystallization temperatures and times follow an inverse relationship, which is aligns with the Arrhenius equation $k = Ae^{\frac{-E_a}{RT}}$, where rate constant k (determines crystallization time) is inversely related to synthesis temperature. A, E_a, R and T refer to the pre-exponential factor, activation energy, gas constant and temperature, respectively.

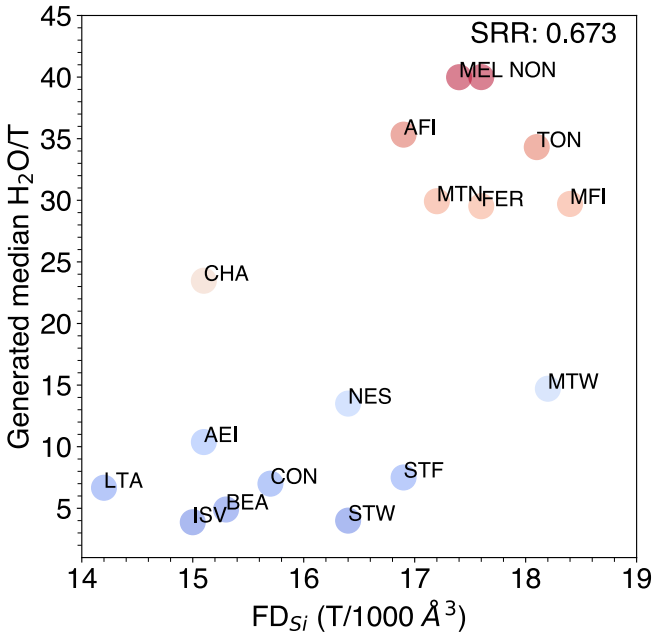


Figure 12: **Model captures domain-specific heuristics.** A positive correlation (Spearman's rank coefficient: 0.673) exists between median diffusion-generated H_2O/T and framework density (FD_{Si}) of zeolite structure for fluoride-mediated synthesis of high silica (Si/Al > 30) aluminosilicates. This agrees with Villaescusa's rule [25], which states that denser phases (higher FD_{Si}) are favored at the less concentrated conditions (higher H_2O/T), showing that the model has learned domain-specific rules in zeolite synthesis.

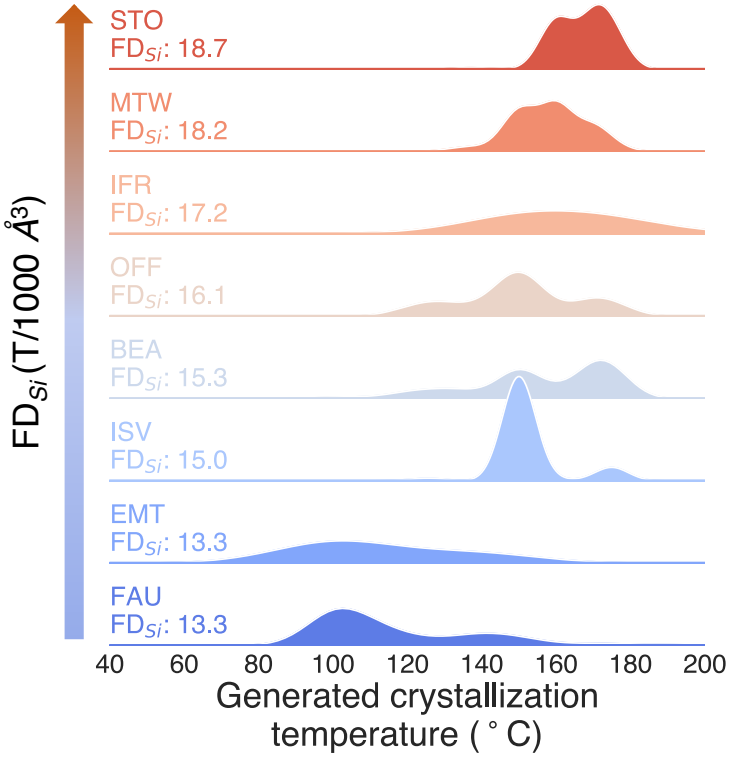


Figure 13: **Model predictions follow thermodynamics of zeolite formation.** A positive correlation (Spearman's rank coefficient: 0.931) exists between median diffusion-generated crystallization temperature and framework density (FD_{Si}) of zeolite structure for large-pore aluminosilicates. This can be seen in an rightward shift in distribution of generated temperatures as FD_{Si} increases. This agrees with the thermodynamic argument that higher crystallization temperatures enable the synthesis to overcome the energy activation barrier to form more stable structures with higher framework density [26, 17]. Furthermore, this observation aligns with Ostwald's rule of stages, where the zeolite passes through metastable states before reaching the most thermodynamically favorable framework.

Table 2: **Physicochemical descriptors of OSDAs.** (Source: [17]).

OSDA descriptor	Description
Asphericity	An anisometry descriptor for the deviation from the spherical shape
Axis 1	Two-dimensional (2D) shape descriptors of molecule calculated by projecting the atomic coordinates into a 2D space based on a principal component analysis (PCA) of the positions. The range of the distribution of points in each principal component is reported as the axis of the conformer. Axis 1 is reported as the larger axis, whereas Axis 2 is the smaller axis
Axis 2	See above
Charge	Formal charge of molecule
SASA	Solvent-accessible surface area (SASA) is the surface area of a molecule that is accessible to a solvent
Molecular weight	Molecular mass of molecule
NPR 1	Normalized principal moments ratio ($\frac{I_1}{I_3}$) where I is principal moment of inertia
NPR 2	Normalized principal moments ratio ($\frac{I_2}{I_3}$) where I is principal moment of inertia
Rotatable bonds	Number of rotatable bonds in the molecule. A measure of molecular flexibility.
PMI 1	Principal moments of inertia (PMI) are physical quantities related to the rotational dynamics of a molecule $$\mathrm{I} = \sum_{i=1}^{\mathrm{A}} m_i \cdot r_i^2 \quad (5)$$ where A is the number of atoms, and m_i is the atomic mass and r_i is the perpendicular distance from the chosen axis of the ith atom of the molecule
PMI 2	See above
PMI 3	See above
Sphericity	Sphericity index of molecule. A measure of how closely the shape of an object resembles that of a perfect sphere
Volume	Molecular volume calculated by using a grid-encoding of the molecular shape using a grid spacing of 0.2 Å and 2.0 Å of margin for the boxes