Guided synthesis of EMT zeolites by machine learning

Emmanuel A. Olanrewaju, Santosh Adhikari, Zhiyin Niu, Michael Nikolaou, Jeremy C. Palmer, Jeffrey D. Rimer, and Mingjian Wen[*,†]

William A. Brookshire Department of Chemical and Biomolecular Engineering, University of Houston, Houston, Texas 77204, USA

(Received 25 March 2026; revised 23 June 2026; accepted 20 July 2026; published 4 August 2026)

Zeolites are microporous crystalline materials with diverse frameworks, widely used in industrial applications such as petroleum refining and molecular separation. Unlike most zeolites, EMT can be synthesized under mild conditions (at low temperatures and without the use of organic structure-directing agents), making it attractive for cost-effective and environmentally sustainable production. However, the specific synthesis conditions that selectively produce EMT rather than similar frameworks like FAU are not yet well established. In this work, we develop machine learning (ML) models to guide the discovery of synthesis conditions for EMT zeolites. Our dataset comprises 174 experimental synthesis attempts, recording reaction time, temperature, silica and alumina sources, Si/Al stoichiometric ratio, and other synthesis parameters. We apply both classical ML methods and pretrained foundation models to predict zeolite framework outcomes from these synthesis parameters. Feature importance analysis identifies critical parameters for EMT formation, validating known synthesis principles. Leveraging the ML models, we explore the synthesis space and identify six promising new conditions for EMT formation. Experimental validation confirms EMT crystallization in five cases, including two with Si/Al stoichiometric ratios outside the training dataset's range. Evaluation on independent literature-reported synthesis conditions further demonstrates the generalizability of the model. This work demonstrates a data-driven approach to accelerating zeolite synthesis, closing the loop between ML prediction and experimental validation.

DOI: 10.1103/v2yp-ylxm

I. INTRODUCTION

Zeolites are crystalline materials containing pores and channels arranged in highly ordered three-dimensional frameworks. Their unique pore sizes and channel geometries, which can be chemically tuned, make them invaluable for industrial processes like petroleum refining (where they help convert crude oil into gasoline) and molecular separation (where they selectively filter molecules based on size and shape) [1–4]. The synthesis of a specific zeolite framework depends on a complex interplay of parameters such as atomic composition (e.g., SiO_2 and Al_2O_3 molar amounts and stoichiometric ratio), reaction conditions (e.g., time and temperature), and even precursor sources (e.g., fumed vs colloidal silica sources) [5,6]. No unique synthesis route exists for a given zeolite framework; variations in synthesis parameters can yield the same crystalline phase, demonstrating the flexibility of the synthesis process [7–11]. For example, EMT zeolites, typically synthesized at low temperatures of 30–60 °C with Si/Al ratios around 2 [12,13], have been successfully dealuminated (postprocessed) at elevated temperatures of 450–600 °C to obtain Si/Al ratios ranging from 4.5 to 52 [14]. Despite this flexibility, small changes in synthesis parameters can result in completely different zeolite frameworks [15–17]. For example, at low temperature, minor variations in temperature or reaction time can lead to a transition from EMT to alternative frameworks such as FAU or SOD [13]. This dual behavior reveals fundamental gaps in understanding the factors governing zeolite synthesis.

Computational methods have been extensively employed to study zeolites. Molecular simulations using classical force fields have successfully modeled molecular adsorption and transport processes within zeolites [18–21]. First-principles calculations based on density functional theory (DFT) have provided accurate predictions of electronic structure, framework thermodynamic stability, and zeolite-catalyzed reactions [22–27]. Modeling the zeolite synthesis process, however, still remains infeasible. Classical force fields require extensive system-specific parametrization [28], and are further limited by accessible time and length scales, typically capturing only microsecond-scale processes in complex, multicomponent synthesis environments of synthesis gel. DFT, while more accurate, incurs prohibitive computational costs for the relevant length and time scales involved in zeolite crystallization [6,29].

More recently, machine learning (ML) techniques have been applied to investigate zeolite synthesis. Extensive databases of zeolites have been constructed using data mining and natural language processing to systematically extract synthesis information from the literature [30–32]. Using such databases, ML models have been developed for mapping the synthesis-structure-property relationship [33–41]. These studies have provided insights into the factors governing zeolite synthesis, such as the effects of synthesis parameters on product yield [35] and inorganic structure-directing agents [36]. Despite the progress and success, most existing work

[*]Contact author: mjwen@uestc.edu.cn

[†]Present address: Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu 611731, China.

in this domain employs ML models to categorize synthesis conditions based on the resulting zeolite frameworks. However, the reliability of these predictions often remains unverified, primarily for two reasons. First, the adopted ML methods inherently contain errors and uncertainties, particularly when trained on limited datasets [42–44]. Second, and more critically, the proposed synthesis conditions are rarely experimentally validated [6,45].

In this work, we integrate data analytics, ML methods, and experiments to study the synthesis conditions that govern EMT formation. Using an in-house dataset of 174 experimental synthesis conditions targeting EMT zeolites, we develop ML models to predict crystallization outcomes. Comprehensive evaluation of multiple algorithms, including both classical statistical ML methods [46] and the recently introduced TabPFN foundation model [47], reveals comparable performance across all methods. The ML models successfully identify key synthesis parameters such as temperature, Si/OH stoichiometric ratio, Si/Al stoichiometric ratio, and Si molar amount that play a significant role in the formation of EMT zeolites.

We also leverage the developed ML models to explore synthesis conditions that favor EMT formation. This is accomplished by first generating new synthesis conditions in a dimension-reduced latent space, followed by screening these conditions based on both the predicted likelihood of forming EMT frameworks (as determined by the ML models) and additional criteria, such as the uniqueness of each condition. Subsequent synthesis experiments demonstrate that five of the top six predicted conditions result in a pure EMT phase. Thus, the ML models are found to have a success rate of 83%. This value is significantly larger than the baseline prevalence in the training data set, where only 32% of the conditions yielded pure EMT. Notably, while all synthesis mixture Si/Al ratios in the training data exceeded 2.5, two of the five successful conditions exhibited lower ratios of 1.77 and 2.3. This demonstrates that our ML approach can identify synthesis conditions that extrapolate beyond the parameter space of the training data, uncovering previously unexplored regions of the chemical space.

Finally, we benchmark the model against literature-reported synthesis conditions to assess its external generalizability. This benchmark shows that the model reliably predicts EMT outcomes for literature cases that remain close as well as distant from the training domain, including cases with NaOH far outside the original training range.

II. METHODS

A. Dataset

The in-house dataset is built around canonical OSDA-free (organic structure-directing agent free) EMT synthesis routes reported in the literature [12,13,48]. We retained the key features that define these routes (alkaline, OSDA-free synthesis from sodium-based silica and aluminum precursors) and then deliberately broadened the synthesis space around them. Compared to the literature routes, our dataset spans a much wider range of compositions, crystallization temperatures and times, and precursor choices, so that the ML model can learn how EMT formation depends on synthesis parameters beyond the narrow window of the literature conditions. A side-by-side comparison of the literature routes and the in-house parameter ranges is provided in Table S1 of the Supplemental Material (SM) [49].

The raw experimental data was cleaned by checking for duplicates, estimating missing values, and removing information that can only be determined postsynthesis, such as crystal morphology and pH. See SM for detailed data cleaning procedures. After cleaning, the final dataset consisted of 174 samples with 13 synthesis parameters (i.e., features) and a label indicating the final synthesis outcome. The features include composition information (e.g., Si and Al molar amounts), precursor type (e.g., fumed silica or Ludox-AS40 as Si source), synthesis time, and temperature, among others. Table I lists all features, along with their data type. The synthesis outcome label consists of three distinct values: ‘pure-EMT’ for zeolites with a pure EMT phase; ‘non-EMT’ for outcomes in amorphous or solution forms where no EMT phase was observed; and ‘hybrid-EMT’ for zeolites exhibiting intergrowth of EMT with other phases such as FAU. Out of the 174 samples, there are 57 pure-EMT, 69 hybrid-EMT, and 48 non-EMT. A pair plot showing the distributions of the features is provided in Fig. S1 in the SM .

B. Model training

Given the relatively small dataset, we employed classical statistical ML algorithms (Random Forest [50], XGBoost [51], and CatBoost [52]) to develop classification models to map the features to the labels. We also used TabPFN [47], a recently proposed foundation model pretrained on massive synthetic data to address classification problems with small data, which is well suited to our case.

The features were preprocessed before being fed to the ML algorithms. Numerical features with continuous values (e.g., synthesis time) were normalized to have a mean of zero and a standard deviation of one to ensure that the scale of different features does not bias the learning process. Categorical features that take discrete values (e.g., the reagent that provides SiO_2) were encoded using a one-hot scheme.

For all models trained in this work, the dataset was randomly split into three subsets for training, validation, and testing, with a ratio of 8:1:1. The model parameters were optimized using the training set, hyperparameters were determined based on model performance on the validation set, and performance metrics are reported on the test set in the main text. Results on the training and validation sets are provided in the SM. To reduce the dependency of the model on a particular data split, we trained 35 models, each with a different data split. The reported performance metrics represent the average across all models, with error bars indicating the standard deviation of these measurements.

III. RESULTS AND DISCUSSION

A. Importance of synthesis parameters

We first benchmark XGBoost [51], Random Forest [50], CatBoost [52], and TabPFN [47] by developing ternary

TABLE I. Experimental synthesis parameters for EMT zeolites. X is one of: fumed silica, Ludox-AS40, Ludox-SM30, Na_2SiO_3 and tetraethyl orthosilicate (TEOS); Y is either $Al(OH)_3$ or $NaAlO_2$. Two types of features are used: numerical (N), which can have values in a continuous range, and categorical (C), which takes discrete values.

Feature	Explanation	Type
Time	Synthesis time (hour)	N
Temperature	Synthesis temperature (°C)	N
Si	Molar amount of SiO_2	N
Al	Molar amount of Al cations	N
NaOH	Molar amount of NaOH	N
H_2O	Molar amount of H_2O	N
Si/OH	Stoichiometric ratio between SiO_2 and NaOH	N
Si/Al	Stoichiometric ratio between SiO_2 and Al cation	N
Si size	SiO_2 particle size (nm) in the silica source	N
Si:X	Reagent (X) that provides SiO_2 to the synthesis mixture	C
Al:Y	Reagent (Y) that provides Al cation to the synthesis mixture	C
Pd:Si	Whether the SiO_2 reagent is predissolved before adding to the synthesis mixture	C
Pd:Al	Whether the Al reagent is predissolved before adding to the synthesis mixture	C

classifiers that predict synthesis outcomes (pure-EMT, hybrid-EMT, or non-EMT) from synthesis parameters. Details of model training are provided above in Sec. II B, and their performance metrics, including accuracy, precision, recall, and F_1 score, on the test set are summarized in Table II. All four ML models achieve an F_1 score of about 0.75, and no model performs significantly better than the others. This is also true for accuracy, precision, and recall. XGBoost is selected for subsequent analysis.

Using the XGBoost model, we perform feature importance analysis to identify key synthesis parameters that control EMT formation. Feature importance analysis was conducted using both the mean decrease in impurity (MDI) method [46] and the Shapley Additive Explanations (SHAP) method [53]. MDI quantifies feature importance by measuring the total reduction in impurity (e.g., entropy) brought by each feature across all trees in the ensemble [46], while SHAP employs concepts from cooperative game theory to attribute the contribution of each feature to the model's prediction [53]. Moreover, the sign of a SHAP value (positive or negative) indicates whether a feature pushed the model's specific prediction above or below the average baseline prediction, which can provide deeper insight into the model's decision-making process. Both SHAP analysis (Fig. 1) and MDI (Fig. S2, SM) identify the Si/OH ratio, Si/Al ratio, temperature, and molar amounts of Si, NaOH, and Al as the most influential features. This agreement underscores the importance of these parameters in EMT synthesis, despite differences in their relative rankings between the two methods.

Although feature importance analysis confirms that synthesis outcomes are governed by multiple parameters, distinct regions in the parameter space correspond to specific zeolite types. This relationship is illustrated in Fig. 2, which shows the outcome labels plotted against the three most influential features: Si molar amount, Si/OH ratio, and temperature. Although pure-EMT samples coexist with hybrid-EMT and non-EMT, they are found exclusively within a specific regime of Si molar amounts and Si/OH ratios (dashed box in Fig. 2). The absence of pure-EMT samples outside this region demonstrates that its crystallization is confined to certain synthesis conditions.

TABLE II. Performance of various models on the ternary classification. The error is obtained as the standard deviation of predictions from multiple models, each trained on a different data split.

	XGBoost	Random Forest	CatBoost	TabPFN
Accuracy	0.76 ± 0.07	0.75 ± 0.09	0.76 ± 0.08	0.74 ± 0.09
Precision	0.79 ± 0.07	0.78 ± 0.09	0.79 ± 0.07	0.76 ± 0.09
Recall	0.76 ± 0.07	0.75 ± 0.09	0.76 ± 0.08	0.74 ± 0.09
F_1	0.76 ± 0.07	0.75 ± 0.09	0.76 ± 0.08	0.74 ± 0.09

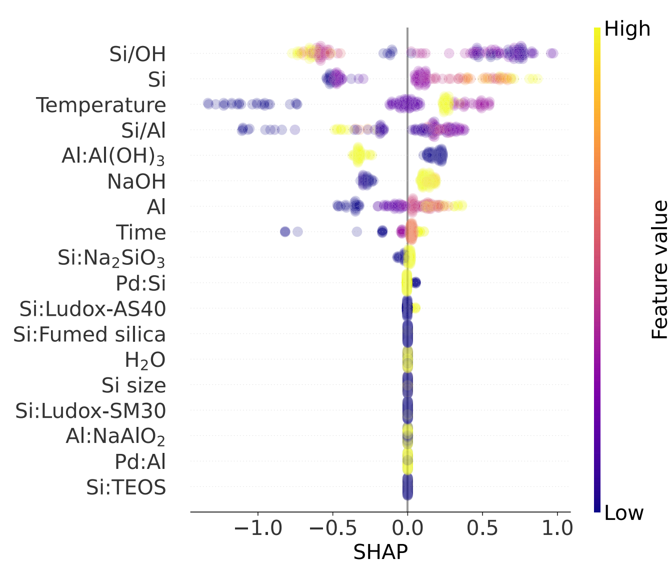

FIG. 1. Feature importance analysis using SHAP for the ternary classification model. The horizontal spread of SHAP values indicates feature importance: A wider distribution signifies greater influence. The 18 features displayed (compared to 13 in Table I) result from one-hot encoding, which expands categorical features into multiple dimensions.

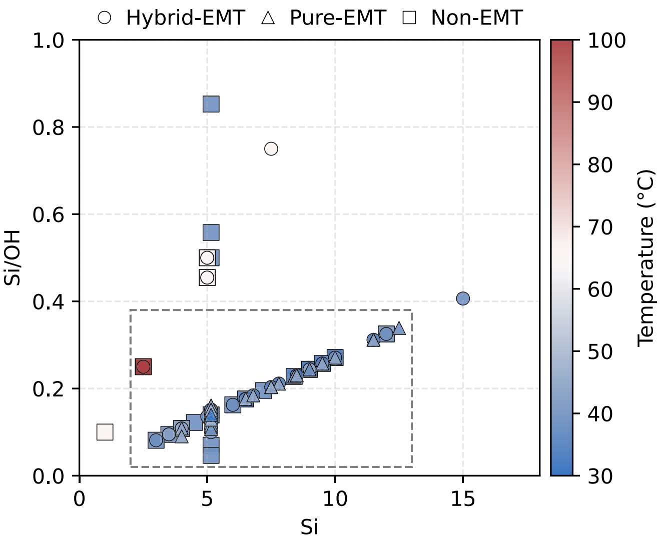

FIG. 2. Distribution of synthesis outcome labels with respect to the three most important features. These features are SiO_2 molar amount, Si/OH stoichiometric ratio, and temperature. All pure-EMT samples lie inside the dashed box.

B. A simplified and focused model

A closer examination of the ternary classification model in Sec. III A reveals that prediction errors are primarily associated with the hybrid-EMT class. As shown in Fig. 3, only three pure-EMT samples are misclassified as non-EMT, and only one non-EMT sample is misclassified as pure-EMT. This stands in stark contrast to the hybrid-EMT class, which exhibits substantial misclassifications with both pure-EMT and non-EMT. We posit that because the hybrid-EMT class represents mixed-phase products, its synthesis conditions have a larger overlap with those of both pure-EMT and non-EMT, leading to a higher rate of misclassification. Since our primary objective is to identify conditions that yield pure-EMT zeolites, we can simplify the problem by merging the hybrid-EMT and non-EMT classes into a single class, called 'Others'. This results in a binary classification of 'pure-EMT' versus 'Others'.

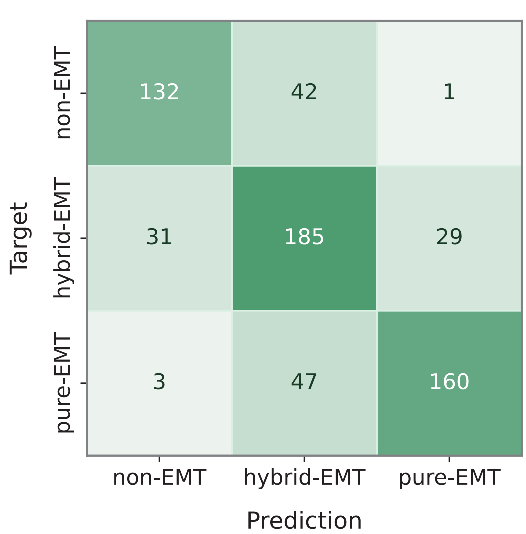

FIG. 3. Confusion matrix for the XGBoost ternary classification model.

TABLE III. Binary classification results using the reduced dataset of 101 samples. The reported values are averaged over predictions from multiple models, each trained on a different data split. Therefore, the mean F_1 score is not generally equal to the harmonic mean of the average precision and average recall, nor is it guaranteed to lie between them. The error is obtained as the standard deviation of predictions from multiple models.

	XGBoost	Random Forest	CatBoost	TabPFN
Accuracy	0.81 ± 0.12	0.81 ± 0.13	0.79 ± 0.13	0.80 ± 0.13
Precision	0.81 ± 0.13	0.81 ± 0.15	0.79 ± 0.14	0.80 ± 0.14
Recall	0.79 ± 0.13	0.80 ± 0.15	0.79 ± 0.14	0.79 ± 0.14
F_1	0.79 ± 0.13	0.79 ± 0.15	0.78 ± 0.14	0.79 ± 0.14

The task can be further simplified by focusing on the region of the parameter space most relevant to EMT formation. As discussed above (see Fig. 2), pure-EMT crystallizes only within specific regions in the parameter space. For the binary classification task, we identified that the Si/Al ratio and NaOH molar amount are effective at delineating the pure-EMT region (Fig. 4). All 57 pure-EMT samples are located within the bounds of $2 \leqslant \mathrm{Si/Al} \leqslant 10$ and $30 \leqslant \mathrm{NaOH} \leqslant 50$. This region contains a total of 101 samples with 57 pure-EMT and 44 Others. Using these two features for filtering is more effective than using the top two features from the importance analysis (Si molar amount and Si/OH ratio in Fig. 2), which define a larger region containing 147 samples. Owing to feature correlations, constraining Si/Al and NaOH also implicitly restricts other parameters. For example, the temperature within this focused region is confined to 30–45 °C, a significant reduction from the full dataset's range of 30–100 °C.

Using the filtered dataset of 101 samples, we developed binary classification models to distinguish pure-EMT from Others. Similar to the ternary classification, the XGBoost, Random Forest, CatBoost, and TabPFN models demonstrate comparable performance on the test set across all classification metrics, as summarized in Table III (results on the training and validation sets are provided in the SM). At first glance, the performance of the binary classification models may appear only marginally better than that of the ternary classification models presented in Table II. However, it is important to note that in this binary classification task, 73 of the 174 samples in the original dataset were excluded because they can be trivially classified as 'Others'. This is because their synthesis conditions lie outside the defined ranges for the Si/Al ratio and NaOH molar amount. Taking this into account, the effective performance of the binary classification model is significantly improved. For example, it achieves an accuracy of 89%, compared to 76% for the ternary classification.

SHAP feature importance analysis was also performed for the binary classification model, as shown in Fig. 5. Compared to the ternary classification (Fig. 1), the top four most influential features remain the same: temperature, Si/OH and Si/Al stoichiometric ratios, and Si molar amount. This further confirms their critical role in determining EMT zeolite

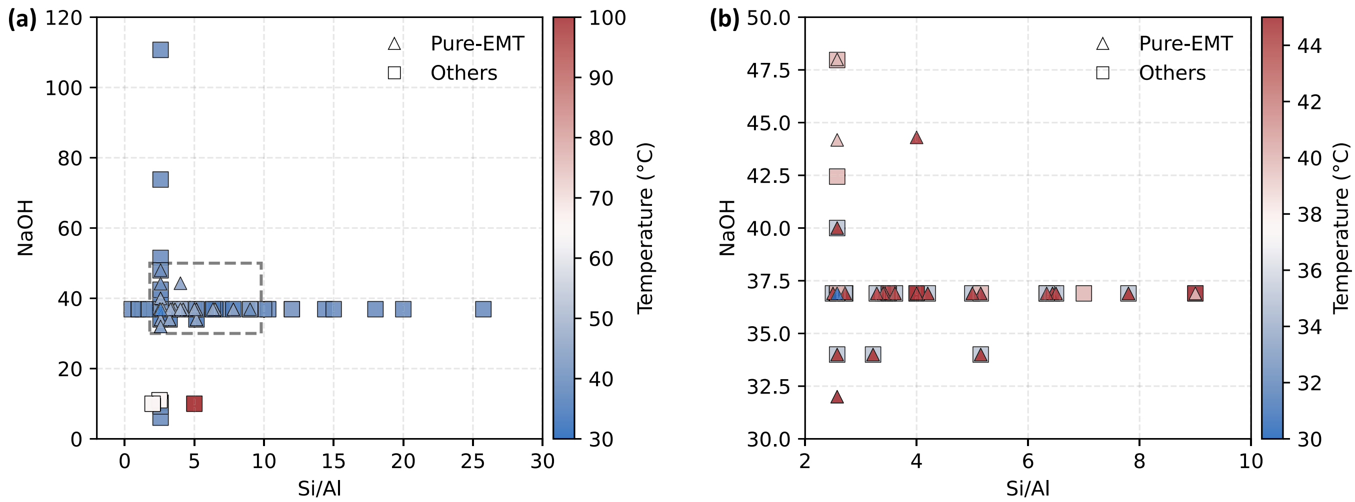

FIG. 4. Distribution of binary synthesis outcome labels with respect to Si/Al stoichiometric ratio, NaOH molar amount, and temperature. (a) Distribution of all 174 samples in the dataset. (b) Magnified view of the dashed region in (a), which contains all pure-EMT samples.

formation. For a binary classification model, the SHAP values can be directly interpreted as the contribution of each feature to predicting the classification outcome [53]. In this context, positive SHAP values indicate that a feature value favors pure-EMT formation, while negative SHAP values suggest the opposite. Figure 5 shows that, in general, positive SHAP values are associated with high temperature and high Si/OH ratio, suggesting that larger values of the two features can promote pure-EMT formation. For Si molar amount and Si/Al ratio, high SHAP values are associated with both high and low feature values, indicating a more complex relationship with EMT formation.

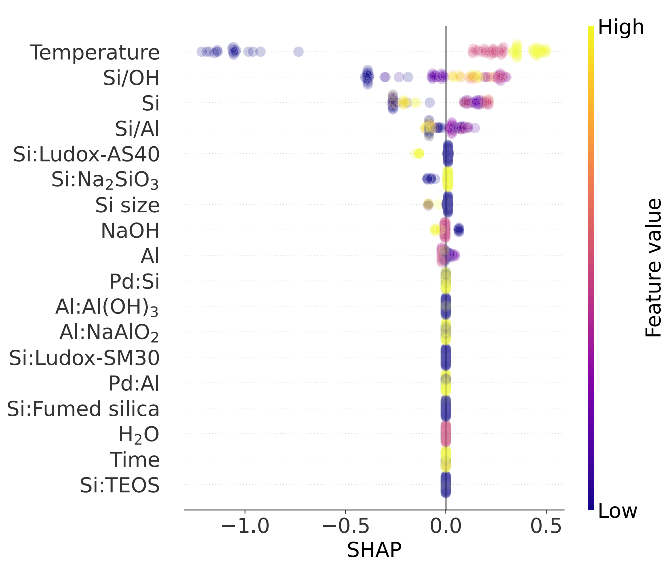

FIG. 5. Feature importance analysis using SHAP for the binary classification model. The 18 features displayed (compared to 13 in Table I) result from one-hot encoding, which expands categorical features into multiple dimensions.

C. Discovery of new synthesis conditions

With the trained binary ML classifier, we aim to discover new synthesis conditions for synthesizing pure EMT zeolites. This is achieved through a two-stage pipeline consisting of a "synthesis parameters proposer" and a "synthesis feasibility evaluator", as illustrated in Fig. 6. First, the proposer generates new synthesis conditions that could potentially yield pure-EMT zeolites. Next, the evaluator examines the feasibility of the conditions using the trained ML classifier, supplemented by additional validation checks. Finally, the most promising candidate synthesis conditions are selected for experimental synthesis and validation.

The synthesis parameters proposer aims to generate new synthesis conditions both within and beyond the parameter bounds of the original dataset. This can be accomplished through methods such as grid search or random sampling of features beyond their observed ranges. However, this approach faces two challenges. A grid search across all 13 features produces an exponentially large number of candidates; subsequent evaluation of these candidates becomes computationally intensive (see Sec. S5 in SM for a time estimate). More critically, the features exhibit strong correlations (see Fig. S1 in SM), and thus most candidates from grid search or random sampling may be chemically unrealistic or thermodynamically infeasible.

We address these challenges by proposing candidate synthesis conditions in a dimension-reduced latent space obtained via principal component analysis (PCA). This was carried out as follows [Fig. 6(a)]. First, we transformed the 13-dimensional feature vectors of all data into a three-dimensional latent space using PCA by retaining the first three principal components. Next, for each of the three principal components we sampled 19 evenly spaced values spanning from 30% below the observed minimum to 30% above the observed maximum. Altogether, 6859 new candidate synthesis conditions were created. Finally, each candidate in the latent space was mapped back to the original 13-dimensional synthesis parameter space via the inverse PCA transform.

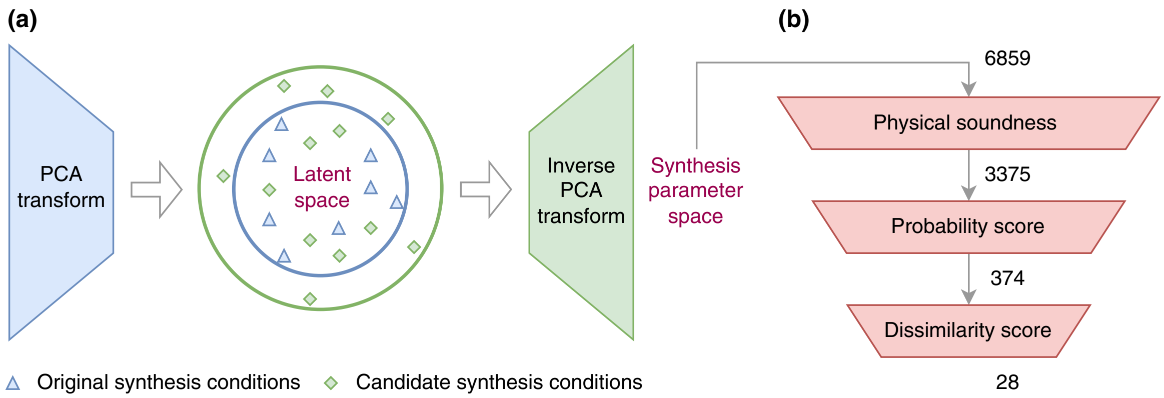

FIG. 6. A two-stage pipeline for discovering new synthesis conditions for EMT zeolites. (a) New candidate synthesis conditions are proposed in a dimension-reduced latent space obtained via principal component analysis (PCA). The blue circle indicates the boundary of the parameter space covered by the dataset, whereas the green circle represents the expanded parameter space explored by the proposer. (b) The proposed candidate synthesis conditions are screened to identify the most promising ones.

Three filtering criteria were then applied to the PCA-generated populate to identify the most promising candidate synthesis conditions: physical soundness, likelihood of producing pure EMT zeolites, and distinctiveness from known conditions in the dataset [Fig. 6(b)]. Some candidate conditions contained unphysical parameters, such as negative Si/Al ratios. After removing these invalid cases, the remaining 3375 candidate conditions were evaluated using the XGBoost classifier developed in Sec. III B to predict their likelihood of yielding pure EMT zeolites. Rather than using the binary yes/no classification, we examined the continuous probability score predicted by the model, which provides more granular information. A higher probability score indicates a greater likelihood of forming pure EMT zeolite. We retained candidates with probability scores greater than 0.85, which reduced the total to 374 candidates. We then computed Euclidean distances of each of the 374 candidates to all samples in the original dataset and used the minimum distance as the "dissimilarity score" $\tilde{d}$. This score quantifies the novelty of each candidate relative to known conditions; a higher value indicates a more novel synthesis condition. Of the 374 conditions, 346 were found to have dissimilarity score below 0.5 and were discarded, yielding a final pool of 28 promising candidates (see Sec. S6 in SM for postprocessing details).

From the 28 candidate synthesis conditions, we selected the six candidates with the highest predicted probabilities for experimental synthesis (see Sec. S7 in SM for experimental details). The predicted synthesis parameters for these six candidates are listed in Table IV. Remarkably, five of the six candidates successfully produced pure-EMT zeolites, corresponding to an 83% success rate. The remaining candidate yielded a hybrid-EMT phase. This success rate represents a 2.6-fold improvement over the 32% success rate achieved by the experience-guided methods used to generate the original dataset. The structural identity of the resulting EMT samples was confirmed by powder x-ray diffraction against the IZA reference EMT pattern, and scanning electron microscopy of a representative sample showed submicron aggregates of EMT nanocrystallites, consistent with what is typically observed for EMT prepared under OSDA-free conditions [12,13,48] (Sec. S8 in SM).

Upon closer examination, we found that two of the five pure-EMT samples were synthesized under extrapolated conditions. Although all proposed synthesis parameters differ from those in the dataset, most fall within the dataset's observed ranges. For example, the Si/OH ratios for the six experimentally verified conditions range from 0.11 to 0.21, compared to 0.09 to 0.34 in the original dataset (Fig. 7). Nevertheless, two conditions have Si/Al ratios of 1.77 and 2.3, which fall below the dataset's lower bound of 2.5. This observation indicates that our ML-based pipeline can successfully extrapolate beyond the training data to discover novel synthesis conditions. Since Si/Al ratio is a key determinant of

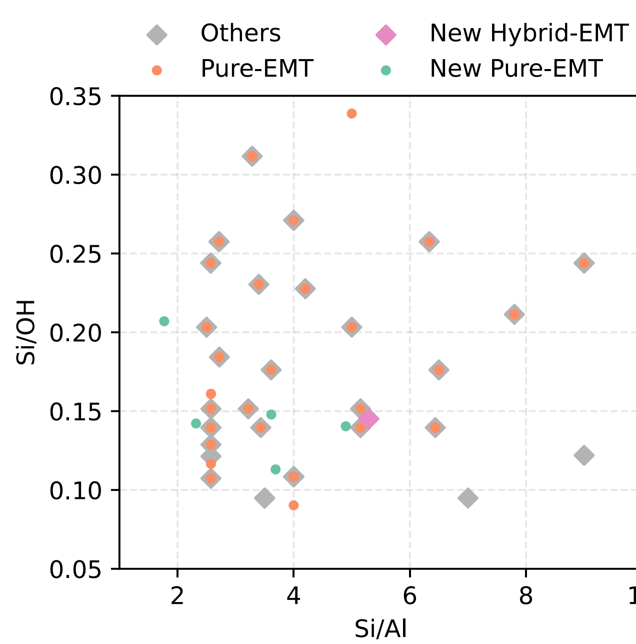

FIG. 7. Distribution of the six experimentally verified synthesis conditions proposed by the ML model. They correspond to the new pure-EMT and new hybrid-EMT (purple diamond and green circles). Samples from the training data are also included (orange circles and gray diamonds). The number of training data appears to be less than 101 since many of them overlap in this two-dimensional plot.

TABLE IV. Experimentally verified synthesis conditions proposed by the ML pipeline. "Probability" is the prediction confidence score from the ML model, and "Dissimilarity" is the Euclidean distance of the synthesis condition to the dataset. "Outcome" indicates the observed zeolite framework after synthesis, but not the label predicted by the ML model. Seven features are listed explicitly. The remaining six features are the same for all six samples: Pd:Si (yes), Pd:Al (yes), Si size (2), Si:Na_2SiO_3, Al:$NaAlO_2$, H_2O (450).

NaOH	Al	Si	Si/OH	Si/Al	Time (hours)	Temperature (°C)	Probability	Dissimilarity	Outcome
37.39	2.08	5.27	0.14	2.32	72.11	41.39	0.90	4.98	pure EMT
36.76	1.12	5.14	0.14	4.90	71.85	41.43	0.88	4.65	pure EMT
36.65	1.65	5.40	0.15	3.61	72.58	41.56	0.88	0.67	pure EMT
36.93	1.26	4.15	0.11	3.69	72.32	42.21	0.87	3.90	pure EMT
36.59	2.93	7.58	0.21	1.77	73.18	40.45	0.86	1.79	pure EMT
36.99	1.04	5.35	0.15	5.29	71.29	41.06	0.93	4.65	hybrid EMT

synthesis outcome (Figs. 1 and 5), correctly predicting out-of-range Si/Al ratios represents a nontrivial and noteworthy achievement.

D. External validation on literature-reported syntheses

The experimental validation in Sec. III C confirms the model on new in-house syntheses, including two recipes beyond the training Si/Al range. A more stringent test of generalizability is whether the model transfers to syntheses performed by other groups, where synthesis conditions such as precursor choices and reaction protocols can differ from those in our in-house dataset. To probe this, we compiled 16 EMT zeolite synthesis conditions from the literature [12,48,54–57]. The full dataset, literature compilation strategy, the procedure for imputing descriptors not reported in the original papers, and the model predictions are described in Sec. S9 of the SM. Because some descriptors had to be imputed, the literature validation should be read as a partially imputed external evaluation across reported EMT-forming conditions rather than a fully independent test across all model inputs.

It is shown in Fig. 4 that pure-EMT samples in the in-house dataset fall within a narrow range of NaOH and Si/Al values. Here, we use these two descriptors to organize the literature EMT samples, which fall both within and beyond the NaOH and Si/Al ranges of the in-house dataset (Fig. 8). The gray region marks the in-house NaOH–Si/Al envelope, where the literature EMT samples have a median dissimilarity score of $\tilde{d} = 2.49$, which lies within the $\tilde{d} \approx 0.7 - 5$ range of the in-house cases we validated experimentally. In this region, the model correctly predicts 8 of 12 EMT cases (67%). The blue region marks literature EMT samples that fall outside the in-house NaOH–Si/Al envelope, with NaOH values below 10. This is considered an extrapolation regime, as the median dissimilarity score of these cases is $\tilde{d} = 39.19$, which is approximately 14 times larger than that of the literature cases in the in-house region. The model correctly predicts three of four EMT cases (75%) in this region.

The higher success rate in the extrapolation region than in the in-house region may seem counterintuitive. A closer look at the four low-NaOH cases shows that the differences between the correctly and incorrectly classified ones do not appear to align with NaOH or Si/Al, but rather with temperature and crystallization time: The three correct cases use $90 - 115\,^{\circ}$C and 168–624 h, whereas the misclassified Dougnier case [56] uses 72 °C and only 24 h (Table S7, SM). This is broadly consistent with the feature importance analysis of the binary classifier (Fig. 5), in which temperature is the most important feature and higher temperature tends to favor EMT formation. That said, the small sample size (four cases) makes the 75% rate sensitive to individual predictions and should not be overinterpreted. Even so, the model correctly identifies the majority of literature-reported EMT cases in both regions, with success rates above the 32% EMT incidence observed during the manual generation of the in-house dataset.

Overall, the model's performance across both regions suggests that it does not simply memorize the in-house recipes but captures broader relationships among synthesis descriptors, such as the temperature dependence noted above. A more definitive assessment, however, will require evaluation on a larger set of literature cases.

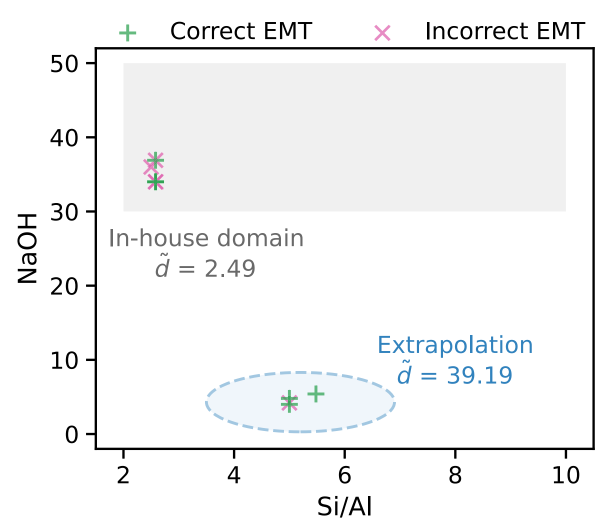

FIG. 8. Model predictions for literature-reported EMT synthesis conditions in the NaOH–Si/Al phase space. The gray region marks the NaOH and Si/Al envelope of the in-house dataset; the blue region marks extrapolation beyond that envelope. The median dissimilarity score $\tilde{d}$ of the literature-reported EMT zeolites in each region is annotated. $\tilde{d}$ measures the distance from a synthesis condition to the in-house dataset (see Sec. III C for its definition).

IV. CONCLUSION

We developed a closed-loop framework integrating ML and experimental synthesis to accelerate the discovery of conditions for EMT zeolite formation. By collecting and systematically analyzing 174 OSDA-free synthesis attempts, we identified temperature, Si/OH ratio, Si/Al ratio, and Si amount as the most critical parameters governing EMT crystallization. Our binary ML classification model, trained on a refined subset of the data, successfully guided the experimental synthesis of five (out of six) EMT zeolites under new conditions, achieving a much higher success rate than the experience-guided approach. Notably, two successful conditions extrapolate beyond the knowledge contained in the dataset. Evaluation against external literature further confirms that the model captures transferable relationships among synthesis descriptors beyond the in-house recipes, correctly predicting the majority of reported EMT cases both within and beyond the training stoichiometric range.

This study demonstrates that data-driven approaches, when coupled with rigorous experimental validation, can significantly improve the effectiveness of zeolite synthesis optimization. The approach of proposing synthesis conditions in a dimension-reduced space followed by a multicriteria screening provides a generalizable workflow that can be applied to other zeolite frameworks and, more broadly, to other crystalline materials whose synthesis depends on a complex interplay of compositional and reaction variables. Looking forward, an iterative active learning strategy could systematically improve model accuracy with minimal experimental effort. This process would involve incorporating newly validated outcomes and using them to retrain models and guide subsequent experiments. We anticipate that this iterative approach will enable efficient navigation of the vast synthesis parameter space and accelerate the discovery of new zeolite frameworks. While this work focuses on classifying synthesis outcomes, future efforts could explore the use of regression models to predict quantities such as crystal size and yield, thus enabling multiobjective optimization toward economical and environmentally sustainable synthesis routes. Beyond synthesis itself, the long-term phase stability of metastable frameworks like EMT, including their resistance to phase transformation (e.g., to FAU) and amorphization under aging or thermal treatment, is a critical practical consideration that warrants dedicated investigation in future work.

ACKNOWLEDGMENTS

J.D.R. and J.C.P. gratefully acknowledge funding from Department of Energy, Office of Basic Energy Sciences, Materials Science Division (Award No. DE-SC0021384) and additional support from the Welch Foundation (Grants No. E-1882, No. E-1794, and Award No. V-E-0001). Computational resources were provided by the Research Computing Data Core at the University of Houston and the Hefei Advanced Computing Center.

E.A.O. and S.A.: model development, data analysis, writing - original draft, and writing - review. Z.N. and J.D.R.: data curation and writing - review. M.N. and J.C.P.: writing - review. M.W.: project conceptualization, data analysis, writing - original draft, writing - review, and supervision.

There are no conflicts of interest to declare.

DATA AVAILABILITY

The raw and cleaned experimental synthesis data, along with the code for model training and data analysis are available at the GitHub repository [58].

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

[2] A. Primo and H. Garcia, Zeolites as catalysts in oil refining, Chem. Soc. Rev. **43**, 7548 (2014).

[3] S. Mintova, M. Jaber, and V. Valtchev, Nanosized microporous crystals: Emerging applications, Chem. Soc. Rev. **44**, 7207 (2015).

[4] Y. Li, L. Li, and J. Yu, Applications of zeolites in sustainable chemistry, Chem. **3**, 928 (2017).

[5] J. Li, M. Gao, W. Yan, and J. Yu, Regulation of the Si/Al ratios and Al distributions of zeolites and their impact on properties, Chem. Sci. **14**, 1935 (2023).

[6] S. Ma and Z.-P. Liu, Machine learning potential era of zeolite simulation, Chem. Sci. **13**, 5055 (2022).

[7] R. E. Fletcher, S. A. Wells, K. M. Leung, P. P. Edwards, and A. Sartbaeva, Intrinsic flexibility of porous materials; theory, modelling and the flexibility window of the EMT zeolite framework, Acta Crystallogr. B **71**, 641 (2015).

[8] S. Ghojavand, E. Dib, and S. Mintova, Flexibility in zeolites: Origin, limits, and evaluation, Chem. Sci. **14**, 12430 (2023).

[9] T. Ohsuna, B. Slater, F. Gao, J. Yu, Y. Sakamoto, G. Zhu, O. Terasaki, D. E. Vaughan, S. Qiu, and C. R. A. Catlow, Fine structures of zeolite-Linde-L (LTL): Surface structures, growth unit and defects, Chem. Eur. J. **10**, 5031 (2004).

[10] S. Yang, M. Lach-hab, E. Blaisten-Barojas, X. Li, and V. L. Karen, Machine learning study of the heulandite family of zeolites, Micropor. Mesopor. Mat. **130**, 309 (2010).

[11] E. Dib, E. B. Clatworthy, A. A. Paecklar, J. Grand, N. Barrier, and S. Mintova, Role of Al distribution in CO_2 adsorption capacity in RHO zeolites, J. Phys. Chem. C **127**, 3 (2023).

[12] E.-P. Ng, D. Chateigner, T. Bein, V. Valtchev, and S. Mintova, Capturing ultrasmall EMT zeolite from template-free systems, Science **335**, 70 (2012).

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

[14] S. Morin, A. Berreghis, P. Ayrault, N. Gnep, and M. Guisnet, Dealumination of zeolites part VIII acidity and catalytic properties of HEMT zeolites dealuminated by steaming, J. Chem. Soc. Faraday Trans. **93**, 3269 (1997).

[15] M. Moliner, F. Rey, and A. Corma, Towards the rational design of efficient organic structure-directing agents for zeolite synthesis, Angew. Chem. Int. Ed. **52**, 13880 (2013).

[16] M. D. Oleksiak and J. D. Rimer, Synthesis of zeolites in the absence of organic structure-directing agents: Factors governing crystal selection and polymorphism, Rev. Chem. Eng. **30**, 1 (2014).

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

[18] B. Slater, J. Titiloye, F. Higgins, and S. Parker, Atomistic simulation of zeolite surfaces, Curr. Opin. Solid State Mater. Sci. **5**, 417 (2001).

[19] G. M. Psofogiannakis, J. F. McCleerey, E. Jaramillo, and A. C. Van Duin, ReaxFF reactive molecular dynamics simulation of the hydration of Cu-SSZ-13 zeolite and the formation of Cu dimers, J. Phys. Chem. C **119**, 6678 (2015).

[20] S. Tesson, W. Louisfrema, M. Salanne, A. Boutin, E. Ferrage, B. Rotenberg, and V. Marry, Classical polarizable force field to study hydrated charged clays and zeolites, J. Phys. Chem. C **122**, 24690 (2018).

[21] A. Misturini, F. Rey, and G. Sastre, Molecular simulation of biobutanol recovery using LTA and CHA zeolite nanosheets with an external surface, J. Phys. Chem. C **126**, 17680 (2022).

[22] F. Filippone, F. Buda, S. Iarlori, G. Moretti, and P. Porta, Structural and electronic properties of sodalite: An *ab initio* molecular dynamics study, J. Phys. Chem. **99**, 12883 (1995).

[23] R. Astala, S. M. Auerbach, and P. Monson, Density functional theory study of silica zeolite structures: Stabilities and mechanical properties of SOD, LTA, CHA, MOR, and MFI, J. Phys. Chem. B **108**, 9208 (2004).

[24] H. Shi, A. N. Migues, and S. M. Auerbach, *Ab initio* and classical simulations of the temperature dependence of zeolite pore sizes, Green Chem. **16**, 875 (2014).

[25] J. Rey, P. Raybaud, and C. Chizallet, *Ab initio* simulation of the acid sites at the external surface of zeolite beta, ChemCatChem **9**, 2176 (2017).

[26] N. Kim and K. Min, Accelerated discovery of zeolite structures with superior mechanical properties via active learning, J. Phys. Chem. Lett. **12**, 2334 (2021).

[27] M. Ducamp and F.-X. Coudert, Prediction of thermal properties of zeolites through machine learning, J. Phys. Chem. C **126**, 1651 (2022).

[28] R. Gaillac, S. Chibani, and F.-X. Coudert, Speeding up discovery of auxetic zeolite frameworks by machine learning, Chem. Mater. **32**, 2653 (2020).

[29] M. Wu, S. Zhang, and J. Ren, AI-empowered digital design of zeolites: Progress, challenges, and perspectives, APL Mater. **13**, 020601 (2025).

[30] Y. Yan, J. Li, M. Qi, X. Zhang, J. Yu, and R. Xu, Database of open-framework aluminophosphate syntheses: Introduction and application (I), Sci. China Ser. B **52**, 1734 (2009).

[31] Z. Jensen, E. Kim, S. Kwon, T. Z. Gani, Y. Román-Leshkov, M. Moliner, A. Corma, and E. Olivetti, A machine learning approach to zeolite synthesis enabled by automatic literature data extraction, ACS Cent. Sci. **5**, 892 (2019).

[32] E. Pan, S. Kwon, Z. Jensen, M. Xie, R. Gómez-Bombarelli, M. Moliner, Y. Román-Leshkov, and E. Olivetti, ZeoSyn: A comprehensive zeolite synthesis dataset enabling machine-learning rationalization of hydrothermal parameters, ACS Cent. Sci. **10**, 729 (2024).

[33] J. D. Evans and F.-X. Coudert, Predicting the mechanical properties of zeolite frameworks by machine learning, Chem. Mater. **29**, 7833 (2017).

[34] M. Moliner, Y. Román-Leshkov, and A. Corma, Machine learning applied to zeolite synthesis: The missing link for realizing high-throughput discovery, Acc. Chem. Res. **52**, 2971 (2019).

[35] B. Conroy, R. Nayak, A. L. R. Hidalgo, and G. J. Millar, Evaluation and application of machine learning principles to zeolite LTA synthesis, Micropor. Mesopor. Mat. **335**, 111802 (2022).

[36] D. Schwalbe-Koda, D. E. Widdowson, T. A. Pham, and V. A. Kurlin, Inorganic synthesis-structure maps in zeolites with machine learning and crystallographic distances, Digit. Discov. **2**, 1911 (2023).

[37] D. A. Carr, M. Lach-hab, S. Yang, I. I. Vaisman, and E. Blaisten-Barojas, Machine learning approach for structure-based zeolite classification, Micropor. Mesopor. Mat. **117**, 339 (2009).

[38] S. Yang, M. Lach-hab, I. I. Vaisman, and E. Blaisten-Barojas, Identifying zeolite frameworks with a machine learning approach, J. Phys. Chem. C **113**, 21721 (2009).

[39] G. Raman, Forecasting low framework density zeolites from synthesis descriptors using machine learning, J. Solid State Chem. **327**, 124290 (2023).

[40] X. Li, H. Han, N. Evangelou, N. J. Wichrowski, P. Lu, W. Xu, S.-J. Hwang, W. Zhao, C. Song, X. Guo, *et al.*, Machine learning-assisted crystal engineering of a zeolite, Nat. Commun. **14**, 3152 (2023).

[41] X. Peng, R. Pan, X. Li, W. Zhong, and F. Qian, Molecular descriptor-assisted interpretable machine learning: A scheme for guiding the synthesis of zeolites with target structures, Chem. Eng. Sci. **308**, 121378 (2025).

[42] A. J. Cohen, P. Mori-Sánchez, and W. Yang, Insights into current limitations of density functional theory, Science **321**, 792 (2008).

[43] M. Wen, S. M. Blau, X. Xie, S. Dwaraknath, and K. A. Persson, Improving machine learning performance on small chemical reaction data with unsupervised contrastive pretraining, Chem. Sci. **13**, 1446 (2022).

[44] J. Dai, S. Adhikari, and M. Wen, Uncertainty quantification and propagation in atomistic machine learning, Rev. Chem. Eng. **41**, 333 (2025).

[45] A. Gandhi and M. F. Hasan, Machine learning for the design and discovery of zeolites and porous crystalline materials, Curr. Opin. Chem. Eng. **35**, 100739 (2022).

[46] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer, New York, NY, USA, 2009).

[47] N. Hollmann, S. Müller, L. Purucker, A. Krishnakumar, M. Körfer, S. B. Hoo, R. T. Schirrmeister, and F. Hutter, Accurate predictions on small data with a tabular foundation model, Nature (London) **637**, 319 (2025).

[48] E.-P. Ng, H. Awala, K.-H. Tan, F. Adam, R. Retoux, and S. Mintova, EMT-type zeolite nanocrystals synthesized from rice husk, Micropor. Mesopor. Mat. **204**, 204 (2015).

[49] See Supplemental Material at http://link.aps.org/supplemental/10.1103/v2yp-ylxm for data processing procedures for the machine learning models, additional results, and experimental synthesis protocols (2026).

[50] L. Breiman, Random forests, Mach. Learn. **45**, 5 (2001).

[51] T. Chen and C. Guestrin, XGBoost: A scalable tree boosting system, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery, New York, NY, USA, 2016), pp. 785–794.

[52] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, CatBoost: Unbiased boosting with categorical features, in *Advances in Neural Information Processing Systems 31*, edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Curran Associates, Inc., Red Hook, NY, 2018), pp. 6639–6649.

[53] S. M. Lundberg and S.-I. Lee, A unified approach to interpreting model predictions, in *Advances in Neural Information Processing Systems 30*, edited by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc., Red Hook, NY, 2017), pp. 4765–4774.

[54] R. Wendelbo, M. Stöcker, H. Junggreen, H. B. Mostad, and D. E. Akporiaye, Tumbling approach towards template-free synthesis of EMT zeolite, Micropor. Mesopor. Mat. **28**, 361 (1999).

[55] C.-N. Wu and K.-J. Chao, Synthesis of faujasite zeolites with crown-ether templates, J. Chem. Soc. Faraday Trans. **91**, 167 (1995).

[56] F. Dougnier and J.-L. Guth, EMT zeolite synthesis: Na^+ vs. OH^- effect, J. Chem. Soc. Chem. Commun. **31**, 951 (1995).

[57] V. Georgieva, A. Vicente, C. Fernandez, R. Retoux, A. Palčić, V. Valtchev, and S. Mintova, Control of Na-EMT zeolite synthesis by organic additives, Cryst. Growth Des. **15**, 1898 (2015).

[58] E. A. Olanrewaju, Source data for machine learning EMT synthesis, 2026, https://github.com/Olanrewajuemmanuelabiodun/ml_EMT_synthesis.