# An exhaustive mapping of zeolite-template chemical space

Mingrou Xie[1], Daniel Schwalbe-Koda[2,3], Yolanda Marcela Semanate-Esquivel[4], Estefanía Bello-Jurado[4], Alexander Hoffman[2], Omar Santiago-Reyes[2,5], Cecilia Paris[4], Manuel Moliner*[4], Rafael Gómez-Bombarelli*[2]

[1] Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.

[2] Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.

[3] Department of Materials Science and Engineering, University of California, Los Angeles, Los Angeles, CA 90095, USA.

[4] Instituto de Tecnología Química, Universitat Politècnica de València - Consejo Superior de Investigaciones Científicas, Avenida de los Naranjos s/n, 46022, Valencia, Spain.

[5] Division of Chemistry and Chemical Engineering, California Institute of Technology, 1200 E. California Blvd., Pasadena, CA 91125, USA.

## Abstract

Zeolites are industrial catalysts and adsorbents whose synthesis usually employs specific molecules known as organic structure-directing agents (OSDAs). The OSDA's templating effect is pivotal in determining the zeolite polymorph formed and its physicochemical properties. However, *de novo* design of selective OSDAs is challenging because of the diversity and size of the zeolite-OSDA chemical space. Here, a computational workflow powered by machine learning enables an exhaustive exploration of the OSDA space for known zeolites. Models were developed to predict molecule-zeolite binding energies and trained on hundreds of thousands of datapoints, the largest ever library of synthetically accessible, hypothetical OSDA-like molecules was enumerated from commercially available precursors, and nearly 500 million zeolite-molecule pairs were screened. From these, two new OSDAs were identified and validated experimentally to template zeolites with unique compositions. The nearly exhaustive scale of the OSDA library and open-access data are expected to accelerate OSDA design for the entire field.

# Introduction

Zeolites are microporous, crystalline materials that are used in industrial separations[1] and catalysis[2]. While some occur naturally as minerals, the majority are synthesized using hydrothermal processes, often requiring the presence of an organic structure directing agent (OSDA) molecule to facilitate the crystallization of specific zeolite frameworks[3,4]. OSDAs are crucial for selectively driving the synthesis of a particular framework against other polymorphs, and also for influencing the density and placement of aluminum in aluminosilicate zeolites[5,6], which act as Brønsted acid catalytic sites. The aluminum distribution then impacts the reactivity, selectivity and stability of zeolites in industrial and environment applications[5,7].

After decades of experimental efforts, about 1000 molecules have been reported in the synthesis of over 200 known frameworks.[8,9] Finding new OSDAs is key to both synthesizing new zeolites and tuning the properties of known ones such as modulating the concentration or positions of heteroatoms like aluminum or germanium.[4,5] In particular, increasing Si/Al ratios generally improves the hydrothermal stability of zeolites for catalytic applications[10], while lowering Si/Al ratios (i.e. below 5) increases their cationic exchange capacity, which is useful in adsorption and catalytic processes requiring metal-exchanged catalysts[5]. General heuristics have been gathered for relating OSDA properties and the pore topology of the frameworks they template, but fully first-principles understanding of templating remains elusive because of the complexity of zeolite nucleation and growth.

*De novo* design of OSDAs is thus challenging. A common strategy relies on performing relatively affordable force-field calculations of the binding energetics between fully formed zeolite pores and molecules, under the assumption that zeolite-molecule interactions provide a strong thermodynamic driving force during crystallization. Earlier instances of such approaches focused on very small molecular libraries and a single framework[11–15], and often suffered from low predictive power since they ignored the possibility of forming secondary products. Larger-scale studies in recent years have progressed to encompass phase competition by considering multi-molecule and multi-zeolite selectivity metrics and resulted in highly predictive, experimentally validated OSDA-zeolite pairings[9,16–18].

However, these selectivity-oriented studies have been limited in scale. Determining selectivity requires calculating the full interaction matrix of *all* molecules and *all* zeolites under consideration. Thus, existing work has been constrained to targeting known zeolites (~220) and repurposing molecules within the space of *already-known* OSDAs (~1000)[9,17].

In this work, we utilize machine learning (ML) to extend the paradigm of selectivity analysis into uncharted chemical space by evaluating the matching between all known zeolites (~220) and all possible ammonium OSDAs that can be made from commercial amines and halides (~2,000,000) (Fig. 1a). ML models were trained on framework-molecule binding affinities and used to evaluate hundreds of millions of framework-molecule pairs, from which OSDA candidates were filtered, confirmed by high-throughput simulations and validated in the lab. [19,20] (Fig. 1b). The thorough nature of the enumerated library, the scale enabled by ML, and the open-sourced data and tools provide a comprehensive mapping of zeolite-OSDA selectivity for the broader community.

# Results

## Chemical space of hypothetical OSDAs

While different OSDA families have been employed in the synthesis of zeolites[4,21], ammonium cations are by far the most extensively used OSDAs. The scientific and patent literature show that ammonium OSDAs range from 50 to 650 Å$^3$ in volume and must be soluble and stable during synthesis. These constraints allowed us to design a library of OSDA-like molecules using reaction-driven enumeration, which provides a baseline guarantee of chemical validity and synthesizability. We can then utilize high throughput screening to select OSDA candidates for targeted frameworks.

We started by gathering a list of amines and halides from commercial suppliers (Supplementary Section 1.1). 543,174 monoquaternary and 1,793,915 diquaternary ammonium compounds were created using *in silico* nucleophilic aliphatic substitution reactions between the amines and halides (Fig. 2a; Methods). Both pools of molecules cover a wide range of synthesizability, cost, and novelty ranges (Fig. 2b-d; terms defined in Methods). The enumerated molecules are diverse, with the monoquaternary and diquaternary libraries containing 7,407 and 29,357 distinct Murcko scaffolds[22] respectively (Supplementary Figures 1-3).

## Binding affinity prediction model

The binding affinities between molecules and frameworks were estimated through three prediction tasks. The first task classifies whether a molecule can fit (binding, $b = 1$) or not (non-binding, $b = 0$) in a framework's cavities. The second task predicts the difference in energies between the framework-molecule complex and the individual components,

$$E = E_{OSDA-zeo} - E_{zeo} - E_{OSDA(g)} . \quad (1)$$

$E$ measures the strength of the interactions between molecules and frameworks as a proxy for the molecule's ability to template the framework. Combining the binary classifier and energy regressor outputs gives the binding energy, which we denote as $BE$ to differentiate from the predicted energy $E$:

$$BE_i = f(E_i, b_i) . \quad (2)$$

$f$ refers to the method of combining the $b$ and $E$ predictions. We tested all possible $f$ across both single model predictions and predictions from an ensemble of five models ($b$: single model, min/mean/max logits, mode; $E$: single model, min, mean, max; $BE$: single model, min, mean, max) and found that averaging an ensemble of $BE$ predictions achieved the best performance (Fig. 2a-c, Methods and Supplementary Table 6). With Equation 2, both molecular fit and binding energetics are captured by a single scalar $BE$. Using $BE$ over $E$ alone improved the root mean squared error (RMSE) of predicted $BE$ from 1.143 to 1.069 kJ/ mol Si and achieved excellent framework ranking per molecule (Fig. 2c-d, Supplementary Figure 7 and Supplementary Table 3). From $BE$, the competition energy ($CE$) of each framework-molecule pair can be computed to measure phase competition (Methods).

The last task is a multiclass classification of the molecule loading per unit cell, $l_{uc}$. $l_{uc}$ allows for conversion of the binding energy to a per-molecule basis:

$$BE\left(\frac{kJ}{mol\ molecule}\right) = BE\left(\frac{kJ}{mol\ Si}\right) \times \frac{N_{Si}}{l_{uc}} , \quad (3)$$

which can be used as another measure of binding affinity. Separate models were trained for each task, since it was observed that a multitask model afforded lower performance.

To screen a large chemical space, informative representations of both the framework and molecule are required. We curated a set of physically interpretable features (labelled as

"physical") leveraging the geometric nature of the prediction tasks and consisting mostly of descriptors of the global molecular shape and flexibility (Methods). The information-dense and task-relevant "physical" framework and molecule feature sets gave the most consistent performance across all three tasks, compared to other, more abstract representations (Supplementary Figures 8-9). Based on SHapley Additive exPlanations (SHAP) analysis[23], the models generally relied on molecular volume and correlated features such as surface area, in keeping with empirical observations (Supplementary Figure 11). For a given zeolite, the models were able to generalize accurately to novel, unseen molecules. However, the models did not generalize well across frameworks (Supplementary Figure 10). This is a common observation in the related field of substrate-ligand binding affinity prediction[24] and arises from the difficulty of characterizing the "negative space" of the pore in comparison to featurization of molecular geometry. In addition, the dataset of zeolite frameworks is smaller and more heterogeneous. Further method development will be needed for extending the approach to novel zeolite frameworks.

When using computations to guide experiments, quantifying the prediction uncertainty provides a measure of reliability. Using standard deviation $\sigma_{i,BE}$ as a metric of uncertainty, we found that lower model accuracy generally correlated with higher uncertainty (Fig. 3f). The predictions also captured the opposite trend for dense frameworks, where the ease of predicting $b = 0$ increases with larger, more out-of-distribution (OOD) molecules. We also observed that our model captured meaningful regions of uncertainty. Boundary points exhibited higher epistemic uncertainty where the molecule has a very snug fit in the framework pores, due to dissenting predictions of $b$ (Fig. 3e, CHA). Aleatoric uncertainty was also higher for small molecules (Fig. 3e, LTA), reflective of the fact that each data point is a static snapshot of many possible packing arrangements on a relatively flat energy landscape.

## Zeolite-OSDA genome

With a large dataset of predicted binding energies, we can construct a "binding energy fingerprint" from $BE$s to *known* OSDAs for each framework. Qualitatively, we observed experimental trends when we clustered these fingerprints by their relative similarity (Supplementary Figure 12). For instance, pairs of frameworks that are in phase competition with each other, such as MFI/MEL, ISV/BEC, and CHA/AEI, are close to each other in the dendrogram. KFI, which shares a molecule-occupiable *lta* cage with LTA, UFI and RHO but

contains a smaller molecule-occupiable *pau* cage, is separated from the latter three frameworks in the dendrogram (Supplementary Figure 12).

We also evaluated the distribution of predicted binding energies for the hypothetical OSDA space. Different frameworks have different ranges of $BE$ values, which can bias the model predictions. In the training data, the best $BE$ is -29.83 kJ/ mol Si for IRR, compared to -10.20 kJ/ mol Si for DDR. Consequently, when counting the frequency of frameworks that are predicted to be the best framework for each molecule, only 9 and 22 unique frameworks were observed for the monoquaternary and diquaternary libraries respectively (Supplementary Tables 6 - 10). Unsurprisingly, most of these frameworks have training data points with $BE$ stronger than -20 kJ/ mol Si. Structurally, we observed that many of these frameworks require germanium or other heteroatoms, are zeotypes, or are not stable upon OSDA removal. The presence of specific composite building units (CBUs) did not explain why these frameworks have stronger $BE$s than others in both predictions and ground truth (Supplementary Figures 13-14). As our docking workflow carries out docking by packing multiple copies of a single conformer into a zeolite, we hypothesize that the frameworks with high $BE$ points have pore geometries that allow energetically favorable packing of similar conformers. Furthermore, factors beyond $BE$ influence synthesis outcomes, such as framework stability and framework interactions with inorganic cations[25], which are not considered in this work. Other data nuances are discussed in Supplementary Section 4.1. In subsequent prediction and filtering steps, the down-selection filters are designed to account for these biases in our predictions (Methods).

## Screening OSDAs for known frameworks

We showcase two experimentally validated case studies that leveraged our *de novo* OSDA screening workflow to select OSDA candidates for templating two frameworks, ERI and CHA[26,27], with new chemical compositions that break current synthesis limitations.

### High-silica ERI

ERI is a small pore zeolite with large *eri* cavities (Fig. 4a) that is traditionally prepared using diquaternary OSDAs such as OSDA-ERI-1 and OSDA-ERI-2 (Fig. 4c) in the presence of $K^+$ cations [28–30]. It is well established that the OSDA and $K^+$ cations stabilize the *eri* cavities and the *can* cages respectively[30]. The organic and inorganic cations introduce high positive charges that usually limit the final Si/Al ratios to 5-6[28–30] and thus zeolite stability under harsh reaction conditions. Using monoquaternary OSDAs would help to reduce the positive charges and increase the Si/Al ratio. For instance, Xie has achieved Si/Al ratios of 6.1-7.6[31] using 1,3-dicyclohexylimidazolium (OSDA-ERI-3, Fig. 4c).

With our workflow, we evaluated all hypothetical monoquaternary molecules on ERI and other frameworks. We used $BE$ and $CE$ cutoffs to reduce the initial library of 543,174 molecules by 18-fold to 29,504 molecules (Fig. 4b). Synthesizability, stability, novelty and cost filters were then applied to maximize molecular stability and optimize for diversity and affordability (Methods). We limited the final pool to 50 molecules or less to make docking feasible. The selected pool of molecules (Fig. 4b) was docked and optimized, and re-ranked with the force field $BE$ and $CE$ (Supplementary Figure 17). Mol-ERI-1 (Fig. 4d), which to our knowledge has never been reported for use in zeolite synthesis before, displayed the best metrics for ERI (Supplementary Figure 17). To compare, variants of Mol-ERI-1 were also docked (Supplementary Figure 18). We verified that Mol-ERI-1 was the best performing molecule, implying that the screening workflow was able to discern $BE$ differences from small changes in the molecular structure. We further found that Mol-ERI-1 has stronger energies than OSDA-ERI-3 (Supplementary Figure 18), potentially suggesting a larger synthesis window.

Based on these results, we attempted synthesis of ERI using Mol-ERI-1, as well as Mol-ERI-4 for comparison (Fig. 4d). For both molecules, we obtained well-crystallized ERI materials (Fig. 4e; Methods; Supplementary Section 6-8 and Supplementary Figure 29a), but the synthesis

window was larger for Mol-ERI-1 than Mol-ERI-4 (Supplementary Figure 27). The resultant ERI(Mol-1) zeolite has a Si/Al ratio of 12.3, higher than both traditional diquaternary OSDAs that template ERI (5-6) as well as ERI(Mol-4) (7.7), which corroborated with the computed $BE$s and $CE$s (Supplementary Table 11).

To evaluate how increasing Si/Al ratios can improve stability, we used selective catalytic reduction (SCR) of NOx as a model reaction. Cu-containing, high-silica small pore zeolites are intensively used as catalysts to reduce NOx gases in light-duty diesel applications[32,33]. These catalysts need to be stable at high temperatures in the presence of steam. We compared 3 wt.% Cu-exchanged ERI zeolites made from OSDA-ERI-1 and Mol-ERI-1 for SCR applications. Both fresh Cu/ERI materials initially showed high SCR performance for most reaction temperatures, but Cu/ERI(Mol-1) presented significant resistance against catalyst deactivation compared to Cu/ERI(OSDA-1) when subjected to severe ageing treatments in 10% steam at 750°C for 13 h (Fig. 4g).

**Al-rich, Na-free CHA**

CHA is another small pore zeolite containing *cha* cavities (Fig. 5a). Contrary to ERI, monoquaternary OSDAs are typically used to produce CHA with Si/Al ratios above 10. The signature OSDA is N,N,N-trimethyladamantammonium[34] (OSDA-CHA-1 in Fig. 5c), although simpler molecules have also been used (OSDA-CHA-2 and OSDA-CHA-3 in Fig. 5c) [9,35] . In recent years, Al-rich compositions with low to intermediate Si/Al ratios of 4-6 have been desired as catalysts in NOx emissions control in heavy-duty diesel applications[32,36–38]. However, a Si/Al ratio of 5 indicates two Al species per *cha* cavity, requiring both alkali $Na^+$ cations and monoquaternary OSDAs to balance the $2^+$ charge[36–38]. Post-synthetic treatments are then necessary to remove the alkali cations as they adversely affect catalytic performance in metal-containing zeolites[39,40]. We sought to find a diquaternary OSDA that would eliminate the use of $Na^+$ and enable simpler, one-pot syntheses of metal-containing, Al-rich CHA.

We used the same workflow as with ERI, although the exact $BE$, $CE$ and molecular volume thresholds were determined via CHA's characteristic $BE - V_{mol}$ plot (Fig. 5b). Filtering for strong $BE$ and $CE$ reduced the initial library of 1,793,915 diquaternary molecules by 768-fold to 2,333 molecules. The remaining filters produced 27 molecules (Fig. 5b). Based on force field energies, we first selected Mol-CHA-3 and Mol-CHA-6 as representative molecules

containing 2/3-C aliphatic and piperazine scaffolds respectively (Fig. 5d and Supplementary Figure 19). Mol-CHA-6 in particular has been described for CHA synthesis, but in a dual-OSDA system combined with OSDA-CHA-1, and always with $Na^+$ [41]. Under the single-OSDA synthesis conditions we explored, both molecules were found to degrade during synthesis (Supplementary Figure 28).

We then selected Mol-CHA-13, which not only showed competitive $BE$ and $CE$ (Fig. 5b and Supplementary Figure 18), but also contains a rigid bicyclic octahydropyrrolo[3,4-C]pyrrole scaffold that could prevent molecule degradation under synthesis conditions. With Mol-CHA-13, we were able to synthesize Al-rich CHA under $Na^+$-free conditions (Supplementary Figure 28) with an Si/Al ratio of 4.9. Both zeolite crystallinity (Fig. 5d) and OSDA stability (Fig. 5e and Supplementary Table 11) were confirmed. The successful synthesis then motivated one-pot synthesis of metal-containing, Al-rich CHA zeolite under Na-free conditions. We used Mol-CHA-13 and ~1 wt.% Fe to synthesize Fe/CHA(Mol-13) (Supplementary Section 7.2.3), obtaining tetrahedrally-coordinated Fe species as measured by UV-Vis spectroscopy (Fig. 5g-h). To compare, we prepared Fe/CHA(OSDA-1) with OSDA-CHA-1 and $Na^+$ in a one-pot synthesis with the same Fe content (Supplementary Section 7.2.4). In their fresh forms, both Fe/CHA catalysts worked well for SCR of NOx with $NH_3$, but Na-free Fe/CHA(Mol-13) displayed substantially improved performance over all reaction temperatures after ageing with steam at 600°C for 13 h (Fig. 5g). We surmised that the degraded performance in Fe/CHA(OSDA-1) was due to the presence of $Na^+$, which facilitated excessive Fe oligomerization to form large, inactive iron oxide species, as indicated by a broad shoulder in the UV-Vis spectrum at and above 300 nm for the aged Fe-CHA(OSDA-1) sample[42] (Supplementary Figure 32b).

The success in finding OSDA candidates for CHA and ERI from an enormous library of ammonium compounds demonstrates the strength of the screening workflow. At the same time, the synthesis outcomes of Mol-CHA-3 and Mol-CHA-6 showed that the suitability of OSDA candidates requires consideration of other factors such as their stability under synthesis conditions. We encourage the community to leverage and build on our data and tools to screen OSDA candidates for targeted frameworks (Supplementary Section 5.3).

# Discussion

Computational cost and lack of data have traditionally limited OSDA design. In this work, we have created an exhaustive database of OSDA-like molecules and executed the largest ever computational screening for OSDAs. Our end-to-end computational workflow incorporates domain expertise and a suite of open-source computational tools, and was experimentally validated with two case studies. We have described nuances in our results to guide the community in interpreting and utilizing our open-sourced tools and predictions.

Future development of the workflow will focus on the zeolite representation to extend the predictive models towards more novel zeolites in a reliable manner in a data-efficient manner. Where more data is required, intelligent methods of sampling the unexplored chemical space will help to speed up the development time of improved models.

Currently, the timescale of experiments limits the scope of framework-molecule pairs we can test in the lab. As the community moves towards more high throughput experimentation methods, the data gained from screening multiple OSDAs selected through rational design will be valuable for improving existing tools and informing about the framework-molecule matching problem. Meanwhile, other components of the synthesis recipe also require rational design, which recent work has started to address[8].

With progress in synthesis planning methods and data, we will also move towards better, quantitative understanding of the role OSDAs play in zeolite crystallization. Further data analysis could reveal insights into whether it is factors such as OSDA solubility and cost that hinder synthesizability of a new framework, or if the laws of physics play a greater role. Our work is another step towards tackling the long-standing challenge of finding synthesis recipes for zeolites that have never been made before.

# Methods

## OSDA enumeration

We reacted amines and halides together using SMILES arbitrary target specification (SMARTS) language to form monoquaternary and diquaternary ammonium compounds. Azines were

removed due to instability under synthesis conditions. Molecules with more than 24 heavy atoms were also removed due to solubility limitations. Following previous work[9], we enumerated all possible stereoisomers, since different stereoisomers can have different shapes and volumes.

We then computed both model inputs and molecular properties (synthesizability, novelty and cost) used for down-selection (see Screening section below) for each molecule. To characterize the molecules' synthesizability, we computed the SA Score as proposed by Ertl[43]. The synthesizability score penalizes molecular fragments that are not widely observed in PubChem, thus favoring molecules that have been synthesized before. As a measure of novelty, we computed the maximum Tanimoto similarity between 1024-bit Morgan fingerprints of the hypothetical molecule and all known OSDAs as a similarity score. The lower the similarity score, the higher the novelty. The cost per mol of each molecule was naïvely computed by adding up stoichiometric ratios of the amines and halides used to form them. While the cost can act as a proxy for synthesizability (as does the SA Score), we note that it involves several simplifications. The cost does not take into consideration alternative, cheaper reagents and reaction pathways, other reagents costs, and separation costs. The prices are also not indicative of bulk amount prices nor the country of origin.

## Docking pipeline

The docking pipeline was used to both generate data for training the models, and for refinement of predicted binding affinities from screening of the hypothetical molecules. Following previous work[9,44], we used RDKit[45] to generate molecular conformers, relaxed the conformers with the MMFF94 force field[46], and docked them in frameworks with the VOID library[19]. Two framework-molecule poses for each of the top 5 loadings for each framework-molecule pair were saved and optimized with the Dreiding force field[47] in GULP[48], assisted by the GULPy library[49]. The binding energy between the docked molecules and the framework was then computed using the frozen pose method, which has been described and benchmarked in previous work[20].

## Binding affinity prediction models

### Data

We curated a dataset of 614,263 framework-molecule pairs, using 2,974 molecules from literature[9] and generative models in previous work[16], and 216 experimentally known frameworks. 353,593 pairs are *binding*, which means that 1) the molecule was able to dock within the framework and 2) the binding energy computed was negative but above -35 kJ/mol Si. The lower bound is used to exclude unphysical systems where molecules formed unphysical bonds with the framework atoms during optimization. Preliminary tests found that models trained on only the poses from the Voronoi docking algorithm tended to underestimate the binding affinities. As such, we supplemented the dataset with poses from Monte Carlo docking, choosing the lowest energy pose regardless of docking algorithm for each pair. The molecule loadings were normalized by unit cell, and each discrete value was assigned to a class for classification. We used 46 classes to categorize loadings $l_{uc}$ ranging from 0 to 19 (Supplementary Table 2).

This dataset was split 8:1:1 by stereoisomers into the training, validation and test datasets. The five models in the ensemble were trained by varying the data within each training and validation datasets such that each model sees a completely different validation dataset. The test dataset is held constant across all five models. Supplementary Table 1 and Figures 4-5 describe the distribution of labels for each model and task.

We also curated two other datasets based on random selection of 2,293 hypothetical monoquaternary and 3,692 diquaternary molecules docked in randomly selected frameworks. The test, hypothetical monoquaternary, and hypothetical diquaternary datasets were used to investigate OOD performance of the models.

### Molecule and framework representations

For the molecules, physical, WHIM[50], and GETAWAY[51,52] descriptors as well as Morgan fingerprints were generated using RDKit[45] and averaged across all generated conformers (with a maximum of 20 conformers generated per molecule).

For the frameworks, Zeo++[53] was used to compute the physical descriptors as well as the pore size distribution (PSD) and stochastic ray tracing histograms ("Ray"). Some of these features are sensitive to the probe radius. As the prediction tasks concern pores and channels large enough to fit molecules, computations requiring a probe radius as input followed Jones' recommendation of probe radius = 1.0 Å to exclude molecule-inaccessible spaces but capture subtleties in pore geometries sufficiently[54]. Observing that no peaks appear in the PSD above 30.0 Å for all considered frameworks, all PSDs were truncated at 30.0 Å and normalized to the same total number of counts across frameworks.

**Model development**

XGBoost[55] and multilayer perceptron (MLP) models (feedforward neural networks) were considered during preliminary studies to compare tree-based and deep learning methods. As MLPs both performed better and allow continuous training for future applications, they were used in subsequent model training (Supplementary Table 3 and Figure 6). Separate MLP models were trained on each of the 3 prediction tasks, with the loss functions described below. Both binding energies and input features were scaled using standard scaling.

A preliminary hyperparameter tuning with baseline physical feature sets for both molecules and frameworks was used to determine suitable architectures for each task. The hyperparameters were then fixed during feature set selection. Hyperparameter tuning was done with Bayesian optimization using SigOpt[56]. With the "physical" molecular and "physical" zeolite feature sets selected from feature selection, inference over and analysis of the hypothetical molecules were performed with an ensemble of five models for each prediction task using the splits described above.

The loss function for the energy regression task used a mean squared error (MSE):

$$L_E = \frac{1}{N_{bp}} \sum_{i=1}^{N_{bp}} \left(E_{i,pred} - E_{i,true}\right)^2 \quad , \tag{4}$$

where $N_{bp}$ is the number of framework-molecule binding pairs and $i$ is a specific framework-molecule pair. The binary classification task used binary cross-entropy loss as a loss function:

$$L_{binary} = \frac{1}{N_{pairs}} \sum_{i=1}^{N_{pairs}} b_i \log\left(p_{binding}\right) + (1 - b_i) \log\left(1 - p_{binding}\right), \tag{5}$$

where $b = 0$ for non-binding pairs and $b = 1$ for binding pairs. For the loading multiclassification task, the labels were a one-hot encoding of the classes. Hence, a mean squared error ($L_{MSE}$) was added to the usual cross-entropy loss ($L_{CE}$) to train the model to predict loadings closer to the real loading than further away:

$$L_{loading} = L_{CE} + \alpha L_{MSE}, \tag{6}$$

where

$$L_{CE} = \frac{1}{N_{bp}} \left( \sum_{i=1}^{N_{bp}} \sum_{j=1}^{M_{classes}} l_{ij,class} \log(p_{ij}) \right) \tag{7}$$

and

$$L_{MSE} = \frac{1}{N_{bp}} \sum_{i=1}^{N_{bp}} \left( l_{ij,class} - \underset{j}{\operatorname{argmax}}(p_{ij}) \right)^2, \tag{8}$$

where $l_{ij,class}$ is the true class label and equates 1 for a molecule-framework pair $i$ with loading class $j$ and 0 otherwise. $\alpha$ is a hyperparameter, and a value of 0.1 was found to work well for stable training and loss minimization.

The competition energy $CE$ was computed by subtracting the $BE$ of the second-best framework from the $BE$ of each framework for a given molecule. Hence, for a given molecule, the best performing framework has $CE < 0$ kJ/ mol Si, the second-best framework has a $CE = 0$ kJ/ mol Si, and all other frameworks have a positive $CE$. We computed $CE$ over the 216 frameworks considered in this work.

After investigating ensembling methods (see Results and Supplementary Table 6), we chose to use Equation 4 with an ensemble of five models to predict $BE$:

$$\mu_{BE_i} = \frac{1}{N_{models}} \sum_{j=1}^{N_{models}} \left( E_{ij} \times b_{ij} \right), \qquad j = 1, \dots, N_{models}, \tag{4}$$

where $i$ is a framework-molecule pair and $j$ is the $j$-th model. The ensemble uncertainty of a framework-molecule pair can be approximated by the standard deviation of $BE_i$ as

$$\sigma_{i,BE} = \sqrt{\frac{1}{N_{models}} \sum_{j=1}^{N_{models}} \left( BE_{ij} - \mu_{BE_i} \right)^2}, \qquad j = 1, \dots, N_{models}. \tag{5}$$

## Screening

The following filters were applied for both case studies: an upper threshold of -10 kJ/ mol Si for $BE$ and 4 kJ/ mol Si for $CE$, molecular volume cutoffs based on minima in the framework-characteristic $BE - V_{mol}$ plots, a maximum number of 5 rotatable bonds, a similarity score of 0.8 or lower (lower scores indicate greater novelty), a synthesizability score of 5 or lower (lower scores indicate higher synthesizability), and exclusion of molecules with functional groups that are unstable under synthesis conditions or difficult to synthesize (such as 3- and 4-membered rings, aromatic nitrogens, double bonds and stereocenters). An additional price filter of 50,000 USD/ mol was used to reduce the pool of molecules to a size amenable to docking for CHA, while a more stringent filter of 100 USD/mol was needed for ERI to achieve a sufficiently small pool of molecules.

## Experiments

### High-silica ERI

Synthesis of ERI with Mol-ERI-1 and Mol-ERI-4 were carried out under conditions described in Supplementary Sections 6-7. The formation of well-crystallized ERI was confirmed with powder X-ray diffraction (PXRD) (Fig. 4d and Supplementary Figure 29a). Si/Al ratios were determined through inductively coupled plasma (ICP) spectroscopy. Elemental analysis of ERI(Mol-1) indicated a C/N molar ratio of ~15, suggesting occluded Mol-ERI-1 molecules within the as-prepared ERI(Mol-1) sample remained intact after crystallization (Supplementary Table 11); this is further confirmed by $^{13}$C MAS NMR spectroscopy (Fig. 4f). The $^{27}$Al MAS NMR spectra of the as-prepared ERI(Mol-1) sample showed the exclusive presence of tetrahedrally-connected Al species, as indicated by the sole signal centered at ~50 ppm (Supplementary Figure 30). The quantification of the occluded Mol-ERI-1 molecules by combining thermogravimetric and elemental analyses indicated an average encapsulation of ~1 Mol-ERI-1 molecule per *eri* cage (Supplementary Table 11).

### Al-rich CHA

Synthesis of CHA with Mol-CHA-13 was carried out under conditions described in Supplementary Sections 6-8. The formation of well-crystallized CHA was confirmed with PXRD (Fig. 5e). Si/Al ratios were determined through ICP spectroscopy. Elemental analysis

of CHA(Mol-13) indicated a C/N molar ratio of 5.1 (Supplementary Table 11); this is further confirmed by $^{13}C$ MAS NMR spectroscopy (Fig. 5f). The $^{27}Al$ MAS NMR spectra of the as-prepared CHA(Mol-13) sample showed the exclusive presence of tetrahedrally-connected Al species, as indicated by the sole signal centered at ~50 ppm (Supplementary Figure 30). The quantification of the occluded Mol-CHA-13 molecules by combining thermogravimetric and elemental analyses indicated the average encapsulation of ~1 Mol-CHA-13 molecule per *cha* cage (Supplementary Table 11).

## Data availability

The code for developing the machine learning models, analyzing predictions and screening can be found at https://github.com/learningmatter-mit/zeobind. All training and validation data, experimental data, hypothetical molecule data as well as predictions made over the entire known zeolite – hypothetical molecule space, can be found at Materials Data Facility[57].

## Acknowledgments

The authors acknowledge financial support by the Spanish Government through PID2021-122755OB-I00 (funded by MCIN/AEI/10.13039/501100011033) and TED2021-130739B–I00 (funded by MCIN/AEI/10.13039/501100011033/EU/PRTR), and by the Generalitat Valenciana through the Prometeo Program (CIPROM/2023/34). The authors are also thankful for the Severo Ochoa financial support by the Spanish Ministry of Science and Innovation (CEX2021-001230-S/funding by MCIN/AEI/10.13039/501100011033).

M.X. acknowledges funding from the Agency of Science, Technology and Research (A*STAR) scholarship. D.S.-K. acknowledges funding from the MIT Energy Fellowship. E.B.-J. and M.S. acknowledge the Spanish Government for an FPI scholarship (PRE2019-088360) and a Severo Ochoa FPI scholarship (PRE2020-092319), respectively. R. G.-B. and A.H. acknowledge funding from MIT Deshpande Center and MISTI Inditex Fund.

The Electron Microscopy Service of the UPV is also acknowledged for their help in sample characterization. Computer calculations were executed on the MIT Engaging and Supercloud supercomputers.

## Bibliography

1. Pérez-Botella, E., Valencia, S. & Rey, F. Zeolites in Adsorption Processes: State of the Art and Future Prospects. *Chem. Rev.* **122**, 17647–17695 (2022).
2. Li, Y., Li, L. & Yu, J. Applications of Zeolites in Sustainable Chemistry. *Chem* **3**, 928–949 (2017).
3. Lobo, R. F., Zones, S. I. & Davis, M. E. Structure-direction in zeolite synthesis. in *Inclusion Chemistry with Zeolites: Nanoscale Materials by Design* 47–78 (Springer Science+Business Media Dordrecht, 1995).
4. Moliner, M., Rey, F. & Corma, A. Towards the Rational Design of Efficient Organic Structure-Directing Agents for Zeolite Synthesis. *Angew. Chem. Int. Ed.* **52**, 13880–13889 (2013).
5. Li, J., Gao, M., Yan, W. & Yu, J. Regulation of the Si/Al ratios and Al distributions of zeolites and their impact on properties. *Chem. Sci.* **14**, 1935–1959 (2023).
6. Le, T. T., Chawla, A. & Rimer, J. D. Impact of acid site speciation and spatial gradients on zeolite catalysis. *J. Catal.* **391**, 56–68 (2020).
7. Hoffman, A. J. *et al.* Rigid Arrangements of Ionic Charge in Zeolite Frameworks Conferred by Specific Aluminum Distributions Preferentially Stabilize Alkanol Dehydration Transition States. *Angew. Chem. - Int. Ed.* **59**, 18686–18694 (2020).
8. Pan, E. *et al.* ZeoSyn: A Comprehensive Zeolite Synthesis Dataset Enabling Machine-Learning Rationalization of Hydrothermal Parameters. *ACS Cent. Sci.* acscentsci.3c01615 (2024) doi:10.1021/acscentsci.3c01615.

9. Schwalbe-Koda, D. *et al.* A priori control of zeolite phase competition and intergrowth with high-throughput simulations. *Science* **374**, 308–315 (2021).

10. Simancas, R. *et al.* Recent progress in the improvement of hydrothermal stability of zeolites. *Chem. Sci.* **12**, 7677–7695 (2021).

11. *AI-Guided Design and Property Prediction for Zeolites and Nanoporous Materials*. (2023).

12. Ito, S., Muraoka, K. & Nakayama, A. De Novo Design of Organic Structure-Directing Agents for Zeolites Using a General-Purpose Large Language Model. Preprint at https://doi.org/10.26434/chemrxiv-2024-wqxl7 (2024).

13. Daeyaert, F., Ye, F. & Deem, M. W. Machine-learning approach to the design of OSDAs for zeolite beta. *Proc. Natl. Acad. Sci.* **116**, 3413–3418 (2019).

14. Muraoka, K., Chaikittisilp, W. & Okubo, T. Multi-objective: De novo molecular design of organic structure-directing agents for zeolites using nature-inspired ant colony optimization. *Chem. Sci.* **11**, 8214–8223 (2020).

15. Schmidt, J. E., Deem, M. W., Lew, C. & Davis, T. M. Computationally-Guided Synthesis of the 8-Ring Zeolite AEI. *Top. Catal.* **58**, 410–415 (2015).

16. Jensen, Z. *et al.* Discovering Relationships between OSDAs and Zeolites through Data Mining and Generative Neural Networks. *ACS Cent. Sci.* **7**, 858–867 (2021).

17. Schwalbe-Koda, D. *et al.* Repurposing Templates for Zeolite Synthesis from Simulations and Data Mining. *Chem. Mater.* **34**, 5366–5376 (2022).

18. Schwalbe-Koda, D. First-principles control of zeolite synthesis, transformations, and intergrowth. (Massachusetts Institute of technology, 2023).

19. Schwalbe-Koda, D. & Gómez-Bombarelli, R. Supramolecular Recognition in Crystalline Nanocavities through Monte Carlo and Voronoi Network Algorithms. *J. Phys. Chem. C* **125**, 3009–3017 (2021).

20. Schwalbe-Koda, D. & Gómez-Bombarelli, R. Benchmarking binding energy calculations for organic structure-directing agents in pure-silica zeolites. *J. Chem. Phys.* **154**, 174109 (2021).

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

22. Bemis, G. W. & Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **39**, 2887–2893 (1996).

23. Lundberg, S. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. Preprint at http://arxiv.org/abs/1705.07874 (2017).

24. Goldman, S., Das, R., Yang, K. K. & Coley, C. W. Machine learning modeling of family wide enzyme-substrate specificity screens. *PLOS Comput. Biol.* **18**, e1009853 (2022).

25. Lee, H., Shin, J. & Hong, S. B. Tetraethylammonium-Mediated Zeolite Synthesis via a Multiple Inorganic Cation Approach. *ACS Mater. Lett.* **3**, 308–312 (2021).

26. Moliner, M., Martínez, C. & Corma, A. Synthesis Strategies for Preparing Useful Small Pore Zeolites and Zeotypes for Gas Separations and Catalysis. *Chem. Mater.* **26**, 246–258 (2014).

27. Dusselier, M. & Davis, M. E. Small-Pore Zeolites: Synthesis and Catalysis. *Chem. Rev.* **118**, 5265–5329 (2018).

28. Miller, M. A. *et al.* Synthesis and catalytic activity of UZM-12. in *Studies in Surface Science and Catalysis* vol. 170 487–492 (Elsevier, 2007).

29. Martín, N. *et al.* Cage-based small-pore catalysts for NH3-SCR prepared by combining bulky organic structure directing agents with modified zeolites as reagents. *Appl. Catal. B Environ.* **217**, 125–136 (2017).

30. Lee, J. H. *et al.* Synthesis and Characterization of ERI-Type UZM-12 Zeolites and Their Methanol-to-Olefin Performance. *J. Am. Chem. Soc.* **132**, 12971–12982 (2010).

31. Xie, Dan & Lew, Christopher. Method for preparing zeolite SSZ-98. US20170O88432A1 (2017).

32. Vennestrøm, P. N. R. *et al.* Advances and perspectives from a decade of collaborative efforts on zeolites for selective catalytic reduction of NOx. *Microporous Mesoporous Mater.* **358**, 112336 (2023).

33. Shan, Y. *et al.* Selective catalytic reduction of NO *x* with NH3: opportunities and challenges of Cu-based small-pore zeolites. *Natl. Sci. Rev.* **8**, nwab010 (2021).

34. Zones, S. I. Conversion of faujasites to high-silica chabazite SSZ-13 in the presence of N,N,N-trimethyl-1-adamantammonium iodide. *J. Chem. Soc. Faraday Trans.* **87**, 3709 (1991).

35. Cao, G. *et al.* Synthesis of chabazite-containing molecular sieves and their use in the conversion of oxygenates to olefins. US7094389B2 (2006).

36. Shan, Y. *et al.* A comparative study of the activity and hydrothermal stability of Al-rich Cu-SSZ-39 and Cu-SSZ-13. *Appl. Catal. B Environ.* **264**, 118511 (2020).

37. Bello, E. *et al.* NH3-SCR catalysts for heavy-duty diesel vehicles: Preparation of CHA-type zeolites with low-cost templates. *Appl. Catal. B Environ.* **303**, 120928 (2022).

38. Yang, Sanyyuan, Turrina, Alessandro, Youngner, Logan, & Gilleland, Daniel. A method of synthesizing a low SAR chabazite zeolite and the zeolite obtained thereby. GB2609550A (2023).

39. Xie, L. *et al.* Excellent Performance of One-Pot Synthesized Cu-SSZ-13 Catalyst for the Selective Catalytic Reduction of $NO_x$ with $NH_3$. *Environ. Sci. Technol.* **48**, 566–572 (2014).

40. Wang, J. *et al.* One-pot synthesis of $Na^+$-free Cu-SSZ-13 and its application in the $NH_3$–SCR reaction. *Chem. Commun.* **57**, 4898–4901 (2021).

41. Moini, A., Kunkes, E. L., THOMAS, J. C., VATTIPALLI, V. & Castellano, C. R. Zeolite structure synthesized using mixtures of organic structure directing agents. EP4065513A1 (2022).

42. Perez-Ramirez, J. *et al.* Evolution of isomorphously substituted iron zeolites during activation: comparison of Fe-beta and Fe-ZSM-5. *J. Catal.* **232**, 318–334 (2005).

43. Gao, W. & Coley, C. W. The Synthesizability of Molecules Proposed by Generative Models. *J. Chem. Inf. Model.* **60**, 5714–5723 (2020).

44. Ferri, P. *et al.* Approaching enzymatic catalysis with zeolites or how to select one reaction mechanism competing with others. *Nat. Commun.* **14**, 2878 (2023).

45. RDKit version 2021.03.4: Open-source cheminformatics; http://www.rdkit.org.

46. Tosco, P., Stiefl, N. & Landrum, G. Bringing the MMFF force field to the RDKit: implementation and validation. *J. Cheminformatics* **6**, 37 (2014).

47. Mayo, S. L., Olafson, B. D. & Goddard, W. A. DREIDING: A generic force field for molecular simulations. *J. Phys. Chem.* **94**, 8897–8909 (1990).

48. Gale, J. D. & Rohl, A. L. The General Utility Lattice Program (GULP). *Mol. Simul.* **29**, 291–341 (2003).

49. Schwalbe-Koda, D. learningmatter-mit/gulpy: GULPy 1.0. *Zenodo* (2021) doi:10.5281/zenodo.5260056.

50. Todeschini, R. & Gramatica, P. SD-modelling and Prediction by WHIM Descriptors. Part 5. Theory Development and Chemical Meaning of WHIM Descriptors. *Quant. Struct.-Act. Relatsh.* **16**, 113–119 (1997).

51. Consonni, V., Todeschini, R. & Pavan, M. Structure/Response Correlations and Similarity/Diversity Analysis by GETAWAY Descriptors. 1. Theory of the Novel 3D Molecular Descriptors. *J. Chem. Inf. Comput. Sci.* **42**, 682–692 (2002).

52. Consonni, V., Todeschini, R., Pavan, M. & Gramatica, P. Structure/Response Correlations and Similarity/Diversity Analysis by GETAWAY Descriptors. 2. Application of the Novel 3D Molecular Descriptors to QSAR/QSPR Studies. *J. Chem. Inf. Comput. Sci.* **42**, 693–705 (2002).

53. Willems, T. F., Rycroft, C. H., Kazi, M., Meza, J. C. & Haranczyk, M. Algorithms and tools for high-throughput geometry-based analysis of crystalline porous materials. *Microporous Mesoporous Mater.* **149**, 134–141 (2012).

54. Jones, A. J., Ostrouchov, C., Haranczyk, M. & Iglesia, E. From rays to structures: Representation and selection of void structures in zeolites using stochastic methods. *Microporous Mesoporous Mater.* **181**, 208–216 (2013).

55. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. 785–794 (2016) doi:10.1145/2939672.2939785.

56. Dewancker, I. *et al.* Evaluation System for a Bayesian Optimization Service. Preprint at http://arxiv.org/abs/1605.06170 (2016).

57. Xie, M. *et al.* An exhaustive mapping of zeolite-template chemical space. Materials Data Facility https://doi.org/10.18126/TRPY-ER69.

# Extended data figures

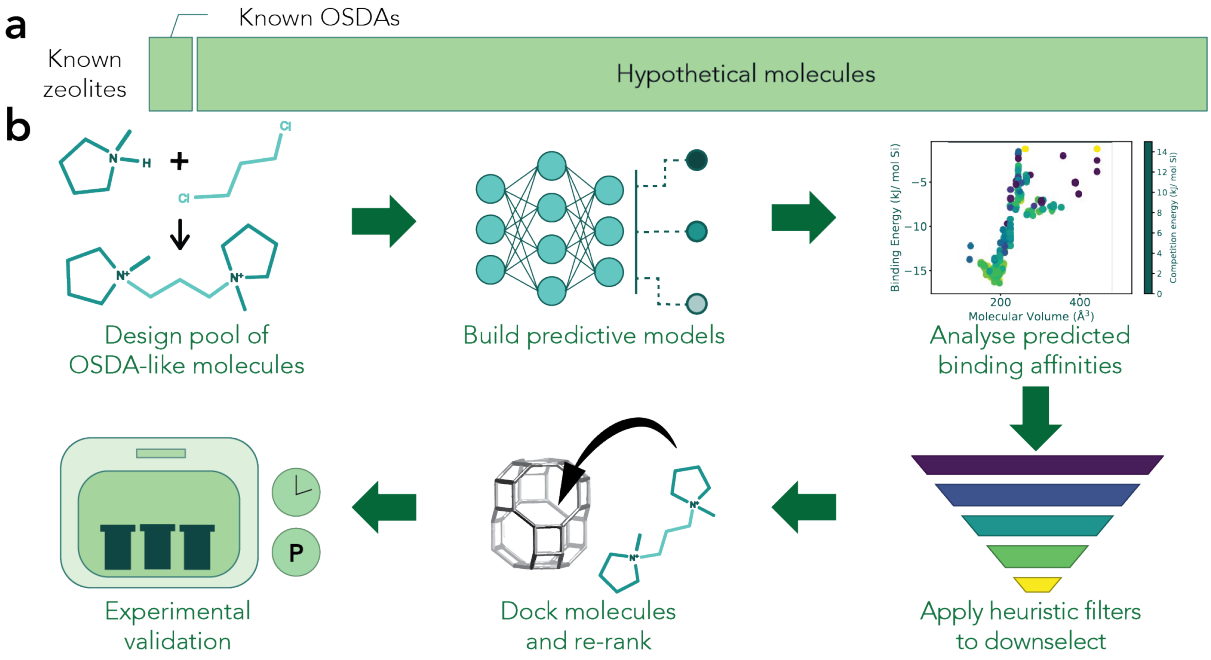


Fig. 1: a) Matrix of zeolite frameworks – molecule pairs (not drawn to scale) showing the difference in chemical space size between known OSDAs and unexplored hypothetical ammonium compounds. b) Schematic of computational workflow for OSDA screening.

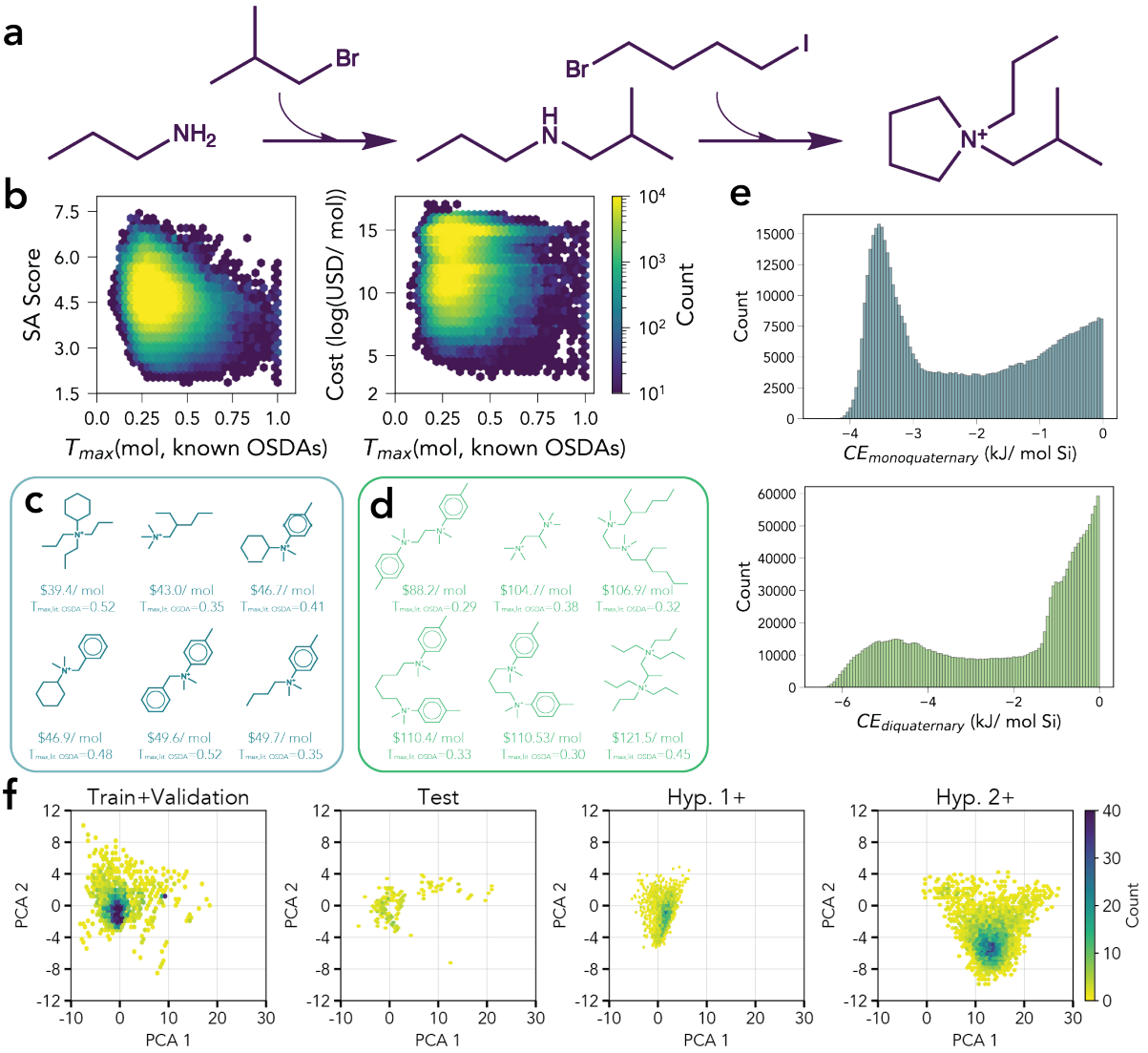


Fig. 2: a) Example of reaction scheme to form an ammonium compound. b) Plots of synthesizability score (SA Score) (left) and cost of hypothetical molecules (right) against similarity score (lower similarity scores indicate greater novelty). c) Six cheapest monoquaternary molecules with similarity score < 0.55. d) Six cheapest diquaternary molecules with similarity score < 0.55. e) Distributions of predicted *CE*s of best-performing framework per molecule for hypothetical monoquaternary (above) and diquaternary molecules (below). f) Principal component analysis (PCA) plots of different sets of molecules. The train, validation and test sets are split by stereoisomers and scaffolds. Hyp.: Hypothetical; $1^+$: Monoquaternary; $2^+$: Diquaternary.

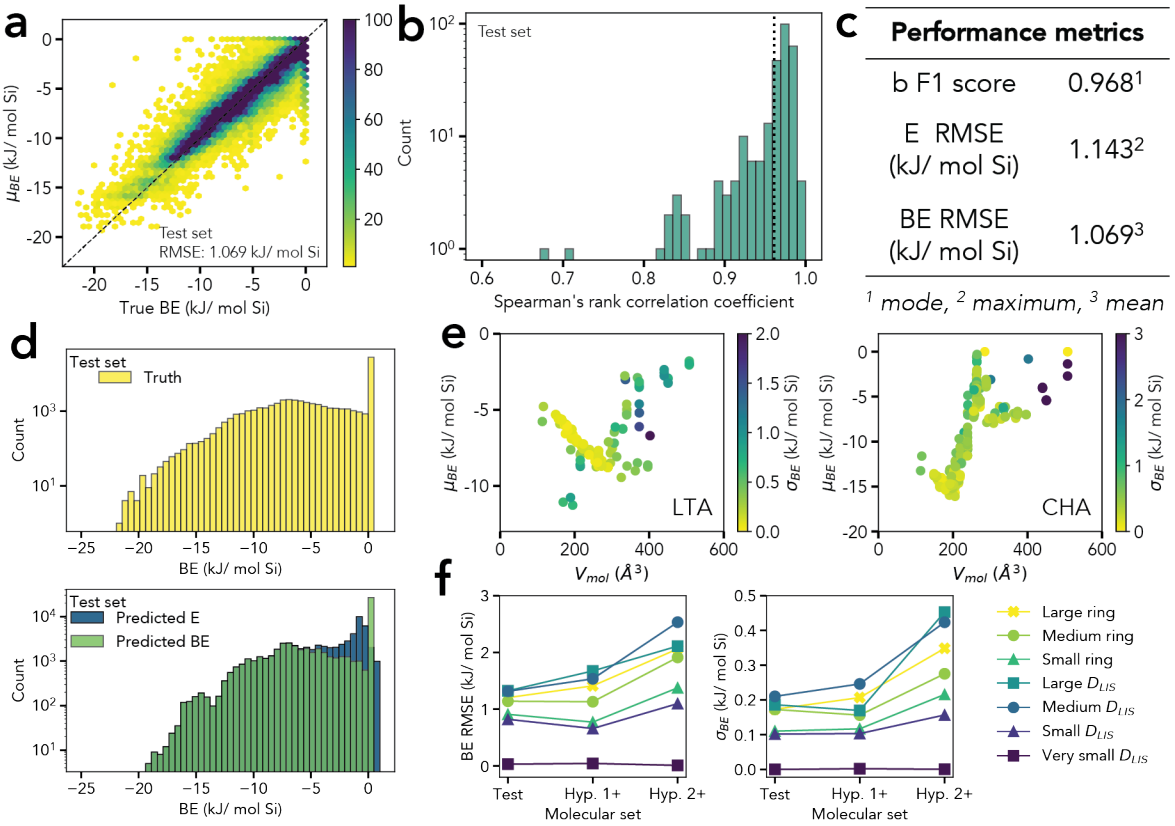


Fig. 3: Performance of the ensemble of NNs. a) Parity plot of predicted against true $BE$ for the test set. b) Distribution of per-molecule Spearman rank correlation coefficients, which measures how well the model ranks frameworks for a given molecule. The dotted line indicates the mean value. c) Performance metrics of best ensembling method on the test set (see Results and Supplementary Table 6 for comparison of ensembling methods). d) (Above) True distribution of $BE$. (Below) Distributions of predicted $E$ and $BE$. e) $BE - V_{mol}$ plots for the LTA (left) and CHA (right) frameworks, colored by the prediction uncertainty ($\sigma_{BE}$). f) Error (left) and uncertainty (right) across increasingly out-of-distribution (OOD) sets of molecules, categorized by framework pore sizes in two ways. The first method is through the largest included sphere diameter ($D_{LIS}$): very small sphere: < 4 Å; small sphere: (4 Å, 6 Å); medium sphere: (6 Å, 9 Å); large sphere: > 9 Å. The second method is through largest ring size of the framework: small ring: < 8 T sites; medium ring: 8 – 10 T sites; large ring: > 10 T sites.

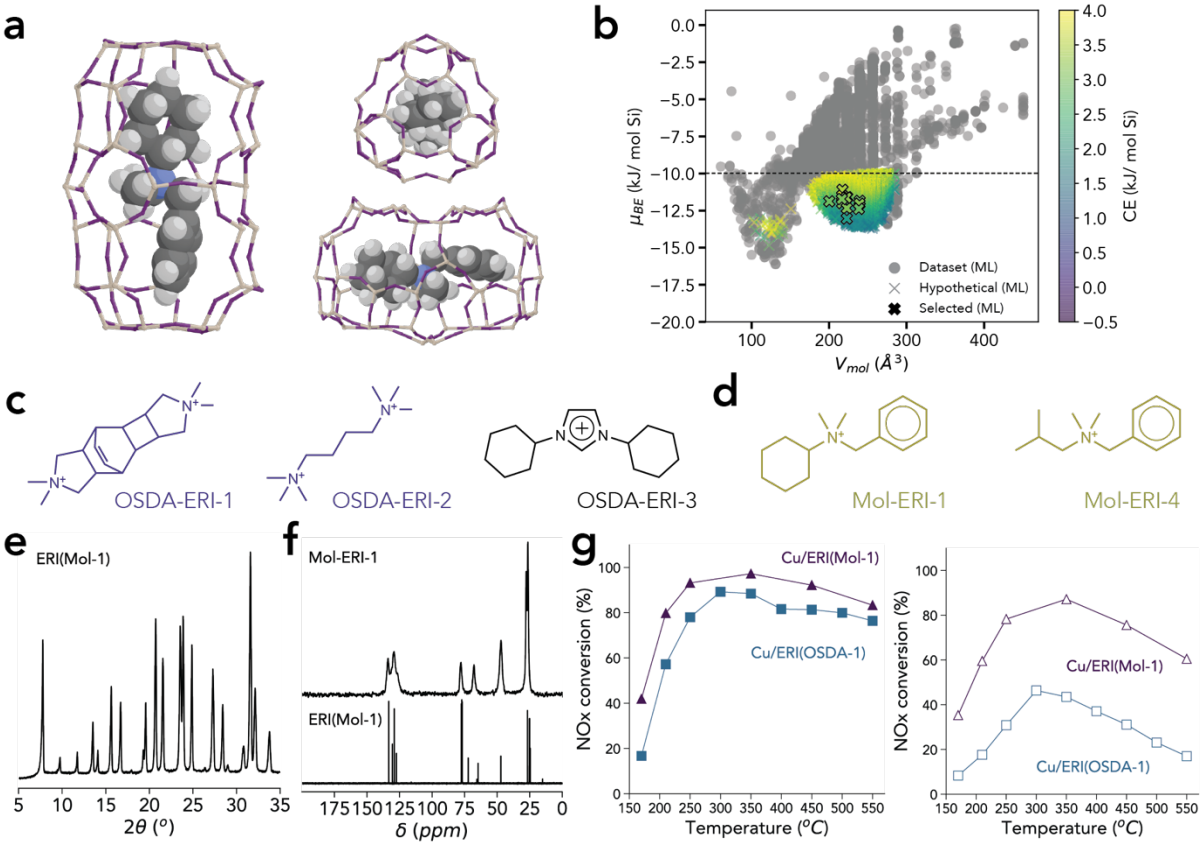


Fig. 4: Schematic for OSDA screening for ERI. a) 3-D diagrams of the validated OSDA (Mol-ERI-1) occluded inside the *eri* cage. b) Predicted $BE - V_{mol}$ plot, colored by predicted $CE$. ML = Machine learning predictions. Dataset = training, validation and test datasets. Hypothetical = hypothetical monoquaternary molecules with predicted $BE$ < -10 kJ/ mol Si and predicted $CE$ < 4 kJ/ mol Si. Selected = Hypothetical monoquaternary molecules selected for docking and optimization. c) Known OSDAs that template ERI. d) Experimentally tested hypothetical molecules for the synthesis of ERI. Molecules from both c) and d) are colored by common scaffold (see Supplementary Figure 17 for the full pool of ML-selected molecules and their labels). e) PXRD diffractogram of as-synthesized ERI prepared with Mol-ERI-1 as OSDA. f) $^{13}C$ MAS NMR and liquid $^{13}C$ NMR of the as-prepared ERI(Mol-1) zeolite and Mol-ERI-1 molecule. g) NO conversion for the $NH_3$-SCR of NOx reaction using Cu/ERI(Mol-1) (purple triangles) and Cu/ERI(OSDA-1) (blue squares) catalysts in their fresh form (filled symbols) and after ageing at 750°C in steam (empty symbols).

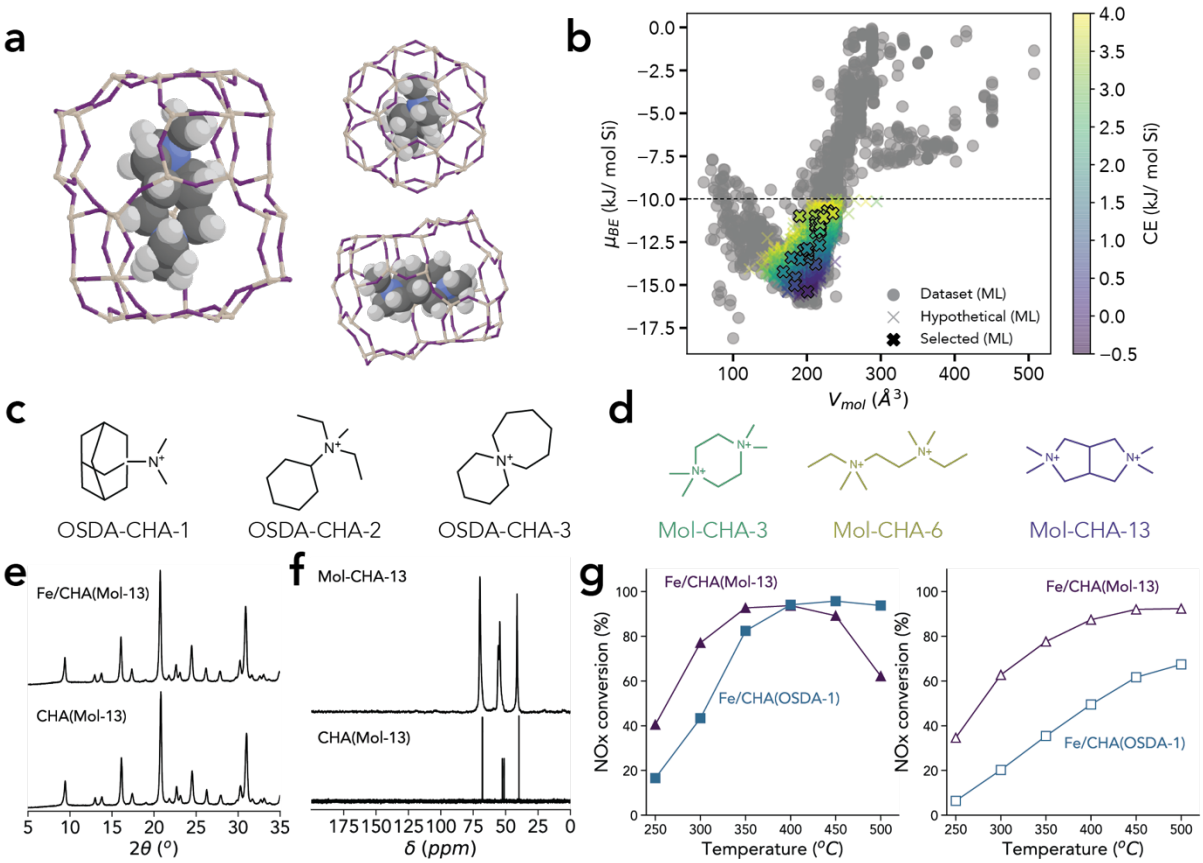


Fig. 5: Schematic for OSDA screening for CHA. a) 3-D diagrams of the validated OSDA (Mol-CHA-13) occluded inside the *cha* cage. b) Predicted $BE - V_{mol}$ plot, colored by predicted $CE$. ML = Machine learning predictions. Dataset = training, validation and test datasets. Hypothetical = hypothetical diquaternary molecules with predicted $BE$ < -10 kJ/ mol Si and predicted $CE$ < 4 kJ/ mol Si. Selected = Hypothetical diquaternary molecules selected for docking and optimization. c) Known OSDAs that template CHA. d) Experimentally tested hypothetical molecules for the synthesis of CHA. Molecules from both c) and d) are colored by charge and common scaffold (see Supplementary Figure 19 for the full pool of ML-selected molecules and their labels). e) PXRD diffractogram of as-synthesized CHA and Fe-CHA prepared with Mol-CHA-13 as OSDA. f) $^{13}C$ MAS NMR and liquid $^{13}C$ NMR of the as-prepared CHA(Mol-13) zeolite and Mol-CHA-13 molecule. g) NO conversion for the $NH_3$-SCR of NOx reaction using Fe-CHA(Mol-13) (purple triangles) and Fe-CHA(OSDA-1) (blue squares) catalysts in their fresh form (filled symbols) and after ageing at 600°C in steam (empty symbols).