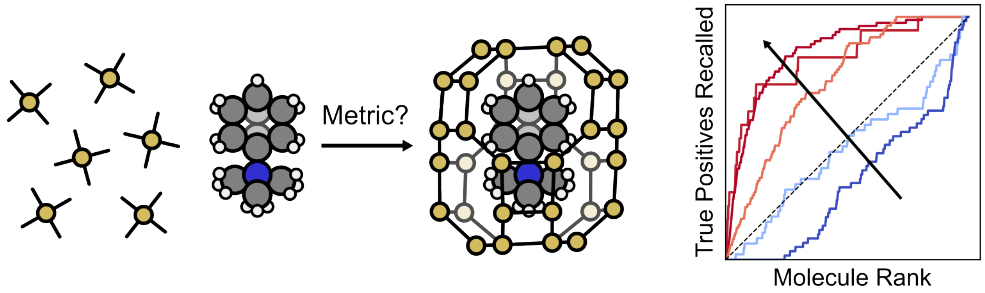
Graphical Abstract

Learning descriptors to predict organic structure-directing agent applicability in zeolite synthesis

Alexander J. Hoffman, Mingrou Xie, Rafael Gómez-Bombarelli

Using symbolic regression and machine learning ...



... to develop new metrics to predict synthesis outcomes

Highlights

Learning descriptors to predict organic structure-directing agent applicability in zeolite synthesis

Alexander J. Hoffman, Mingrou Xie, Rafael Gómez-Bombarelli

- Add new parameters from earlier work that can be considered when trying to predict zeolite synthesis outcomes
- Use machine learning tools to identify a new equation that can be used to rank organic structure-directing agents for zeolite synthesis more accurately than previous equations developed from chemical intuition alone

Learning descriptors to predict organic structure-directing agent applicability in zeolite synthesis

Alexander J. Hoffman[a], Mingrou Xie[b], Rafael Gómez-Bombarelli[a]

[a] *Department of Materials Science and Engineering, Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, 02139, MA, USA*
[b] *Department of Chemical Engineering, Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, 02139, MA, USA*

Abstract

Zeolite synthesis frequently relies on organic structure-directing agents (OSDAs), but the process of identifying the best OSDA to synthesize a given zeolite remains difficult. We use previously gathered binding energy data, in additional to the formation energies of the siliceous zeolite frameworks and approximate binding entropies of OSDAs to develop new descriptors to improve predictions based on known OSDA-zeolite pairs in the literature. Our earlier work used templating energy ($E_{ij,T}$) to rank the most likely OSDA-zeolite pairs to be produced from synthesis. Using literature recall area-under-the-curve (AUC) as a performance metric, we find that computing energies associated with the net transformation that occurs during zeolite synthesis (the sum of the formation energy of the zeolite framework and the OSDA binding energy) provides a modest improvement over $E_{ij,T}$ when predicting the zeolite phase that a given OSDA produces, from 67.5% average literature recall to 72.3%, but negligibly improves predictions for the best OSDA for a given zeolite framework, from 68.3% to 68.8%. We then use machine learning symbolic regression to develop a new descriptor, which we call $\alpha_{ij,T}$, that slightly improves upon $E_{ij,T}$ for predicting an OSDA for a given framework, with an average literature recall of 71.8%. While zeolite synthesis remains difficult to predict *a priori*, the approaches used in this work provide one option for improving these predictions.

Keywords: zeolite synthesis, machine learning, organic structure-directing agent

Preprint submitted to Elsevier *December 4, 2024*

1. Introduction

Zeolites are microporous solids that are often synthesized using organic and inorganic structure-directing agents (OSDAs and ISDAs, respectively). Over 200 unique zeolite frameworks have been synthesized [1], while over 300,000 more have been hypothesized as possible based on SiO_4 tetrahedron connectivity geometries and estimated formation energies [2, 3, 4, 5, 6]. Notably, these hypothetical frameworks are filtered based on having a formation energy below +30 kJ (mol Si)$^{-1}$. More recent work has noted that pure silica frameworks have formation energies well below this threshold (6–19 kJ (mol Si)$^{-1}$) and that the range of formation energies of existing silica materials changes with the density of the framework [7]; as such, many of these materials may be difficult or impossible to synthesize in (alumino)silicate form. More recent work using support vector machines has indicated that formation energies of siliceous frameworks alone are insufficient for identifying synthesis targets from hypothetical frameworks [8]. Instead, accounting for the most likely feasible composition of hypothetical frameworks can increase the likelihood of identifying synthesis targets. OSDAs are molecules that template the void spaces of these zeolite materials during synthesis, thus directing the formation of distinct framework phases. OSDAs are typically composed primarily of C, N, and H, where the N atoms in these molecules are often quaternary cations. While other synthesis parameters can be tuned to affect the final zeolite phase that forms, the selection of an OSDA (when used) is vital in controlling the zeolite framework that forms during synthesis. SHapley Additive exPlanations (SHAP) have been used to identify the most salient parts of synthesis recipes (*e.g.*, temperature, ISDAs, synthesis gel composition)[9]; for many frameworks, OSDA shape and size remain the most important factors influencing the final crystal structure. As such, methods to identify effective OSDAs *a priori* to synthesize frameworks are crucial for guiding the synthesis of proposed hypothetical frameworks.

Recently, our group published work showing that phase selectivity can be guided using predictions from computed binding energies of OSDAs in zeolites [10]. This high-throughput virtual screening (HTVS) approach allowed us to identify an OSDA that produced an intergrowth of the CHA and AEI frameworks by identifying a biselective OSDA [11], whose Cu-exchanged form performed better than pure-phase CHA for NO_x selective catalytic reduction [12], and a CHA/ERI intergrowth [13]. This method to determine phase control of zeolites used a computed templating energy, $E_{ij,T}$ (Eq. 1), to estimate how likely an OSDA would produce a given framework:

$$E_{ij,T} = -k_B T \log \left(C_{ij,OSDA} C_{ij,Si} D_{ij,OSDA} D_{ij,Si}\right)^{\frac{1}{4}} \tag{1}$$

This equation uses a combination of the competition (C_{ij}) and directivity (D_{ij}) of a given OSDA and zeolite, which are defined as

$$C_{ij} = \frac{\exp \frac{-\Delta E_{ij}}{k_B T}}{\sum_{j=zeo} \exp \frac{-\Delta E_{ij}}{k_B T}}, \tag{2}$$

which represents how well an OSDA templates a given zeolite framework compared to all other frameworks, and

$$D_{ij} = \frac{\exp \frac{-\Delta E_{ij}}{k_B T}}{\sum_{i=OSDA} \exp \frac{-\Delta E_{ij}}{k_B T}}, \tag{3}$$

which captures how well an OSDA templates a given framework compared to all other OSDAs. The ΔE_{ij} term in Eq. 2 and 3 is the binding energy of OSDA i in zeolite j

$$\Delta E_{ij} = E_{\mathrm{Z}_j\text{-}\mathrm{OSDA}_i} - E_{\mathrm{Z}_j} - E_{\mathrm{OSDA}_i} \tag{4}$$

where $E_{\text{Z-OSDA}}$ is the energy of the bound OSDA-zeolite complex, E_{Z} is the energy of the empty zeolite, and ΔE_{OSDA} is the energy of the gas-phase zeolite. The binding energies used to compute these values can be normalized either per OSDA in the zeolite unit cell (*i.e.*, the loading) or per tetrahedral Si atom in the zeolite, denoted as $\Delta E_{ij,\mathrm{OSDA}}$ and $\Delta E_{ij,\mathrm{Si}}$, respectively. The competition and directivity metrics can be computed from binding energies normalized per OSDA molecule ($\Delta E_{ij,\mathrm{OSDA}}$)—denoted by an OSDA subscript, $C_{ij,OSDA}$ and $D_{ij,OSDA}$—or per Si atom ($\Delta E_{ij,\mathrm{Si}}$)—$C_{ij,\mathrm{Si}}$ and $D_{ij,\mathrm{Si}}$. These ΔE_{ij} were computed using the DREIDING force field [14], which shows acceptable agreement with higher-accuracy methods for computing energies like density functional theory (DFT) [15]. Such binding energy values reflect the fit of the OSDA within the zeolite framework of interest and are an important piece of information for assessing the utility of an OSDA for synthesizing a given framework, although additional information about molecule and void shape can improve predictions of synthesis outcomes [16].

The approach this previous work used, however, neglected the relative stability of the underlying zeolite framework and the binding entropy of OSDA molecules. During zeolite synthesis with an OSDA, the net transformation that occurs is

$$n \times \mathrm{SiO}_2 + m \times \mathrm{OSDA} \rightarrow (\mathrm{OSDA})_m\mathrm{Si}_n\mathrm{O}_{2n(zeo)} \quad (5)$$

where n SiO_2 moieties and m OSDAs become the OSDA-Zeolite complex. If thermodynamics primarily dictate the formation of a zeolite framework during synthesis, the energy of this net transformation should drive the selection of one zeolite phase over another. Recent work has introduced the use of the net energy of transformation from material sources in the synthesis solution to the final zeolite product [17, 18]. Such net thermodynamic changes are important to include when attempting to guide zeolite synthesis using HTVS approaches like those from our past work, and may improve predictions for OSDA phase selectivity.

Our earlier work also neglected the potential role that binding entropy (ΔS_{ij}) may play in determining how well an OSDA templates a given framework. Molecular adsorption to a material reduces the number of translational and rotational degrees of freedom available [19]. Recent work suggested that the adsorption entropy of alkanes from the gas-phase could be computed using only the properties of a 3D conformation of the molecule (mass, moments of inertia) and a few parameters [20]. Such similar methods may improve predictions of zeolite phase if used to augment earlier HTVS data.

Here, we expound on calculations of the net material transformation to form a zeolite, which we term formation affinities, $\Delta E_{\mathrm{form},ij}$ (based on Eq. 5), and explore other ways of using metrics calculated *in silico* to identify the most promising OSDAs for a zeolite. Using formation affinities alone improves predictions of the framework that forms for a given OSDA for several key zeolites, but worsens predictions for others. Additionally, we incorporate estimations of binding entropy based on gas-phase estimates [20], to see if including such entropic contributions improves literature recall of known zeolite-OSDA pairs. Finally, we use the sure independence screening and sparsifying operation (SISSO) program to learn possible descriptors (*i.e.*, equations) that predict zeolite synthesis outcomes. These new equations more accurately predict synthesis outcomes from literature than the $E_{ij,T}$ equation used in prior work while including additional metrics that could affect synthesis (such as the framework formation energy, $\Delta E_{\mathrm{form},j}$). These new descriptors may help assess OSDAs for more complicated zeolites beyond those that have been accomplished from this theory-first approach so far with a window-cage topology [10, 12, 13].

2. Results

2.1. Zeolite Formation Energy for Phase Prediction

The net transformation that zeolite synthesis produces is described in Eq. 5. This process can be broken into two components, zeolite formation from some silica source:

$$n \times \mathrm{SiO_2} + m \times \mathrm{OSDA} \rightarrow \mathrm{Si}_n\mathrm{O}_{2n(zeo)} + m \times \mathrm{OSDA} \tag{6}$$

and OSDA binding:

$$\mathrm{Si}_n\mathrm{O}_{2n(zeo)} + m \times \mathrm{OSDA} \rightarrow (\mathrm{OSDA})_m\mathrm{Si}_n\mathrm{O}_{2n(zeo)} \tag{7}$$

Historically, zeolite formation energies have been computed relative to α-quartz for comparison to experimental measurements [7, 21], which we also do in this work using DFT (method details in Section 5.1). Earlier work has shown that estimations of the formation energy are sensitive to the method used to optimize the structure and compute the energy [22]. We compute the formation energy with DFT because it provides relatively accurate estimates of formation energy compared to experimentally measured formation enthalpies of pure-silica zeolites (Figure B.2, Supporting Information (SI)) [23]. Formation energies from the DREIDING forcefield do not match well with experimentally measured formation enthalpies nor with the larger set of DFT-calculated formation energies (Figures B.1 and B.2, SI). The step in Equation 7 corresponds to the OSDA binding within the zeolite to form an OSDA-zeolite complex, which has been calculated in our prior work [10, 15] using the DREIDING force field [14]. As such, the total formation affinity is simply the sum of the framework's formation energy and the binding energy of the OSDA (Figure 1):

$$\Delta E_{\mathrm{form},ij} = \Delta E_{\mathrm{form},j} + \Delta E_{ij} \tag{8}$$

We consider how including the formation energy to compute the total thermodynamic transformation affects literature recall. Notably, this assessment excludes other components that can influence zeolite synthesis outcomes and be modeled atomistically, such as ISDAs and framework heteroatoms (Figure 1). Moreover, these high-throughput atomistic simulations cannot account for things like synthesis time or temperature, which may also

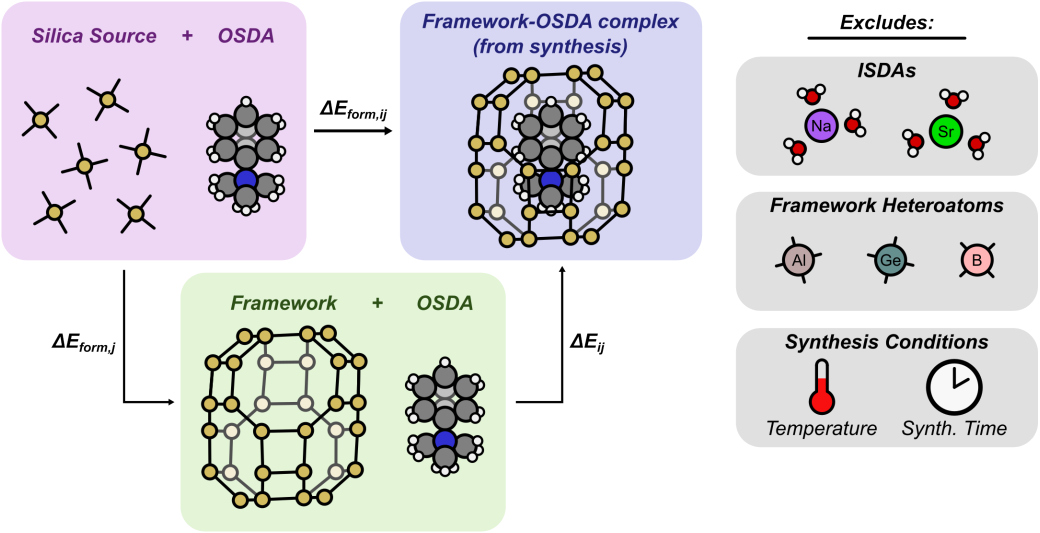


Figure 1: Summary of the approach used in this work to identify OSDAs that produce particular frameworks and the factors that influence zeolite synthesis that were excluded in our analysis.

influence zeolite crystallization. Instead, we focus specifically on the stability of the underlying zeolite framework and the fit of the OSDA within that framework. We use both the formation affinities (normalized per Si atom in the unit cell of the zeolite, $\Delta E_{\mathrm{form},ij,\mathrm{Si}}$) and a recalculated version of $E_{ij,T}$ from Eq. 1. This updated formation templating energy, $E_{\mathrm{form},ij,T}$, is computed from competition and directivity values that are based on the formation affinities in Eq. 8 rather than the original ΔE_{ij} values in Eq. 4. Once they are computed, we can compute the literature recall for these descriptors in a similar approach to our earlier work to indicate how well they rank OSDAs for synthesizing a given zeolite framework [10]. The literature data were extracted from papers and made publicly available in earlier work [24], and the dataset was recently expanded to include additional details [9].

Generally, both $\Delta E_{\mathrm{form},ij,\mathrm{Si}}$ and $E_{\mathrm{form},ij,T}$ correlate with $E_{ij,T}$ (Figure 2), albeit with some scatter. This correlation initially indicates that the rankings these metrics produce will be similar. If the thermodynamics of the net transformation during zeolite synthesis determines the final phase produced, then it is possible that $E_{ij,T}$ closely matches the total transformation energy and, thus, produces relatively high literature recall.

Our earlier work [10] illustrated the process of computing literature recall

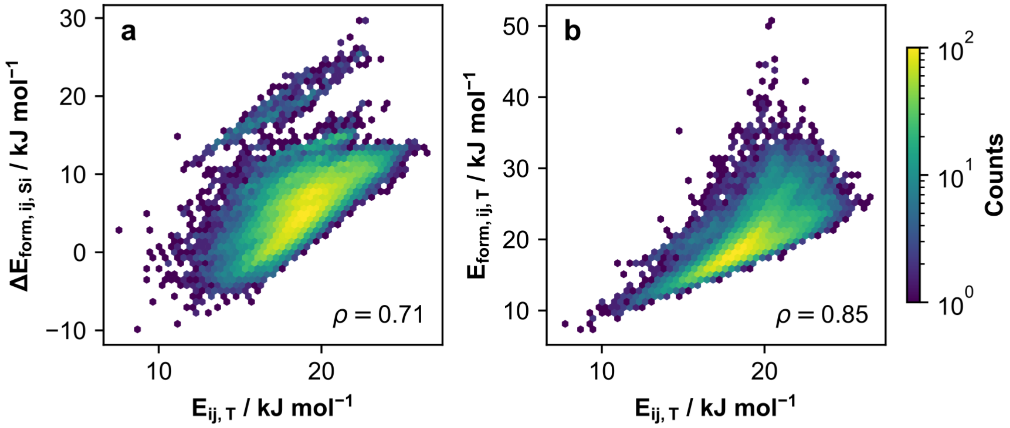


Figure 2: (**a**) Formation affinities per Si atom in the unit cell ($E_{\mathrm{form},ij,\mathrm{Si}}$) and (**b**) formation templating energies ($E_{\mathrm{form},ij,T}$) as functions of the templating energy used in earlier work. Each plot is labeled with the corresponding Spearman correlation coefficient (ρ).

and showed the recall performance for five frameworks (AEI, MFI, CHA, MOR, and MTW) and several OSDAs. Recall for a zeolite is determined by ranking the OSDAs and computing the number of experimentally validated OSDA-zeolite pairs as a function of that ranking. The area under the curve (AUC) in these recall plots is then normalized by the maximum recall (where all known OSDAs are the highest ranked) such that AUCs $\in [0, 1]$. Recall can also be plotted for OSDAs by ranking the zeolites they are predicted to template best. Notably, there are no clear true negatives that we can use to construct recall-precision curves because some OSDAs may work for zeolites but have not yet been attempted under the necessary synthesis conditions. The literature recall predictions from $\Delta E_{\mathrm{form},ij,\mathrm{Si}}$ (Figure 3a) produce similar performance to $E_{ij,T}$ (Figure 3c) for all frameworks except for MOR, for which it makes better predictions for effective OSDAs. Generally, the $E_{\mathrm{form},ij,T}$ metric worsens recall for all five zeolite frameworks relative to $E_{ij,T}$ (Figure 3b). This initial set of examples appears to indicate that the inclusion of formation energy can improve the recall of experimentally validated OSDA-zeolite pairs.

We computed the recall using this approach for all three of these metrics for all frameworks and OSDAs for which they could be computed. Figure B.3 in the Supporting Information (SI) shows the recall AUCs for all zeolites evaluated in this work. Generally, literature recall is better the more

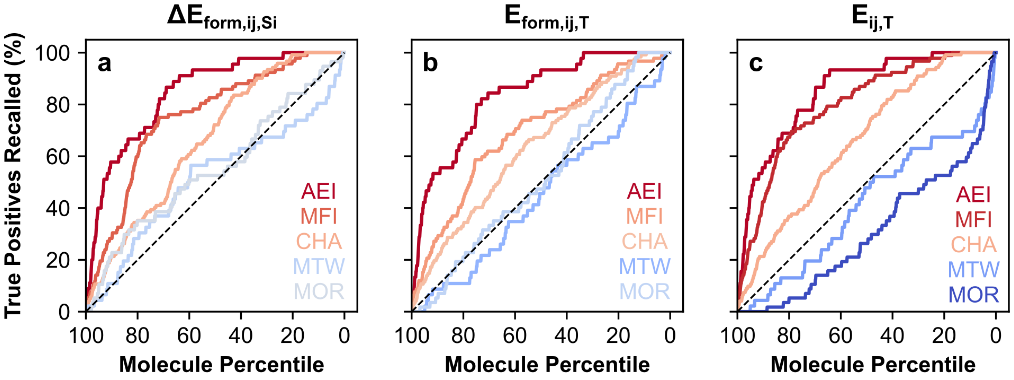


Figure 3: Literature recall curves for AEI, MFI, CHA, MOR, and MTW using OSDA rankings computed with (**a**) formation affinities per Si atom ($\Delta E_{\mathrm{form},ij,\mathrm{Si}}$), (**b**) formation templating energies ($E_{\mathrm{form},ij,T}$), and (**c**) the templating energy used in earlier work ($E_{ij,T}$).

publications have been made about synthesizing a given framework, while materials with very few publications can have recall AUCs from very poor to near unity (Figure B.5, SI). This improvement may arise because frequently studied frameworks are simpler to synthesize and have more OSDAs that work for their formation or because not all synthesis routes have been explored for rarely studied frameworks. Notably, predicting good OSDAs for some naturally occurring frameworks—such as MOR—is particularly difficult. Some naturally occurring frameworks are industrially relevant and can be synthesized with only inorganic SDAs, such as FAU [25, 26, 27], MOR [28, 29], and LTA [26, 30]. As such, their synthesis may rely less heavily on OSDA choice than other artificial frameworks. Because they rely less on OSDA choice, the metrics that perform well in literature recall for synthetic frameworks appear to perform worse for some commonly studied frameworks like MOR and LTA.

Finally, we also computed literature recall AUCs for each OSDA molecule in the dataset to determine how these metrics performed for predicting the final phase produced from synthesis using a given OSDA. We averaged the recall AUCs across all zeolites or across all OSDAs for each of the metrics (Table 1). The metrics that include the formation energies of the siliceous forms of the frameworks ($E_{\mathrm{form},ij,T}$ and $\Delta E_{\mathrm{form},ij,\mathrm{Si}}$) do not always outperform the recall of the original $E_{ij,T}$ metric for ranking OSDAs to synthesize those frameworks; however, $\Delta E_{\mathrm{form},ij,\mathrm{Si}}$ outperforms $E_{ij,T}$ on recall for zeolite phase

Table 1: Average literature recall areas-under-the-curve (AUC) for the templating energy from earlier work ($E_{ij,T}$) compared to similar metrics that include siliceous zeolite formation energies.

Metric	**Average Recall AUC**	
	Framework	OSDA
$E_{ij,T}$	0.684	0.675
$E_{\mathrm{form},ij,T}$	0.680	0.684
$\Delta E_{\mathrm{form},ij,\mathrm{Si}}$	0.688	0.723

outcomes for OSDAs on average. The performance of these metrics suggests that net thermodynamic information is important for determining phase selectivity for an OSDA, but that the selection of a molecule when targeting a framework might be confounded by other factors, such as the absence of true negatives in the available data.

2.2. Entropic contributions to OSDA templating

We estimate the standard adsorption entropies of OSDAs, ΔS°_{ij} based on equations described elsewhere (details in Section 5.2) [20]. This assessment does not directly include conformational entropy contributions ($\Delta S_{\mathrm{ads,conf}}$); however, these equations include an empirical fit to gas-phase adsorption entropies and, as such, should include a $\Delta S_{\mathrm{ads,conf}}$ contribution implicitly. Moreover, $\Delta S_{\mathrm{ads,conf}}$ losses relative to the gas phase should be relatively small compared to rotational and translational entropy losses. We also note that molecules with high gas or aqueous phase conformational entropy would likely make poor OSDAs. Flexible molecules are less likely to adopt the shape required to template a zeolite pore for long enough to stabilize the crystallization of the zeolite at typical solvothermal synthesis temperatures. Once calculated, we use these ΔS°_{ij} values to compute the Helmholtz free energies of binding, ΔA_{ij}, for each zeolite-OSDA pair:

$$\Delta A_{ij} = \Delta E_{ij} - T\Delta S^{\circ}_{ij} \tag{9}$$

We use a constant synthesis temperature of 400 K for these estimations for all zeolites (the same value used to compute C and D throughout this work and in earlier work [10]). While the Gibbs free energy, ΔG_{ij}, is the natural potential for zeolite crystallization systems, the Helmholtz free energy captures the majority of the contributions to the Gibbs free energy and should be sufficiently informative for HTVS work, where fast methods are used to rapidly assess and eliminate large numbers of candidates [31].

Similar to our initial assessment of formation affinities, we compute a templating energy based on the Helmholtz free energies. First, we calculate competition and directivity values for each OSDA-zeolite pair based on these Helmholtz free energies:

$$C_{A,ij} = \frac{\exp \frac{-\Delta A_{ij}}{k_B T}}{\sum_{j=zeo} \exp \frac{-\Delta A_{ij}}{k_B T}}, \tag{10}$$

$$D_{A,ij} = \frac{\exp \frac{-\Delta A_{ij}}{k_B T}}{\sum_{i=OSDA} \exp \frac{-\Delta A_{ij}}{k_B T}}. \tag{11}$$

Next, we compute a Helmholtz templating energy based on these $C_{A,ij}$ and $D_{A,ij}$ values analogously to the earlier $E_{ij,T}$:

$$A_{ij,T} = -k_B T \log \left(C_{A,ij,\mathrm{OSDA}} C_{A,ij,\mathrm{Si}} D_{A,ij,\mathrm{OSDA}} D_{A,ij,\mathrm{Si}}\right)^{\frac{1}{4}} \tag{12}$$

These $A_{ij,T}$ values correlate with $E_{ij,T}$ (Figure 4), although the inclusion of the entropy introduces significant scatter.

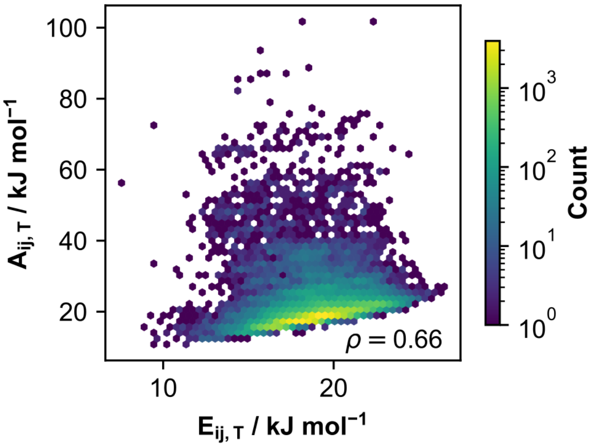


Figure 4: Helmholtz templating energy, $A_{ij,T}$, as a function of $E_{ij,T}$.

Similar to our approach in Section 2.1, we also computed the average literature recall AUCs for this new metric and compared it to $E_{ij,T}$. Adding an entropic contribution to OSDA binding energies appears to worsen recall AUC when averaged across frameworks and with a slight improvement across OSDAs for $\Delta A_{\mathrm{form},ij,\mathrm{Si}}$ (Table 2). Importantly, the entropic contributions estimated here are for these molecules in the gas-phase and based on a model for non-polar alkane adsorption in zeolites. A more realistic approach would account for the effects of aqueous solvation on OSDAs and their free energies (and, therefore, entropies) because zeolites are often synthesized hydrothermally. The decline in recall performance for this Helmholtz templating energy relative to $E_{ij,T}$ may be caused by an overestimation of entropic losses upon confinement of the OSDA in the zeolite during synthesis relative to the OSDA in the aqueous solution phase. Using chemical intuition to incorporate zeolite stability and molecular entropic losses within the original mathematical formulation for synthesis predictions did not satisfactorily improved recall of experimentally validated pairs. Therefore, we sought an alternative approach that could encompass more energy terms as well as other factors not included in the templating energy formulation using the SISSO method to develop new descriptors.

Table 2: Average literature recall areas-under-the-curve (AUC) for the templating energy from earlier work ($E_{ij,T}$) compared to the $A_{ij,T}$ metric that include an estimation of OSDA entropy loss upon adsorption.

Metric	**Average Recall AUC**	
	Framework	OSDA
$E_{ij,T}$	0.683	0.675
$A_{ij,T}$	0.629	0.629
$\Delta A_{\mathrm{form},ij,\mathrm{Si}}$	0.619	0.686

2.3. Regressed descriptors for zeolite synthesis

2.3.1. Variables for synthesis prediction

Before performing symbolic regression using SISSO, we analyzed how the variables we might include in such runs correlate with one another. Squared Spearman correlation coefficients (ρ^2) between several of the metrics used to understand binding energies are above 0.7 for several variables (Figure 5; all ρ^2 are provided for the potential energy SISSO run in Figure B.6 in the SI). Perhaps most notably, $\Delta E_{ij,\mathrm{Si}}$ correlates relatively strongly with all formation affinities ($\Delta E_{\mathrm{form},ij,\mathrm{Si}}$, $\Delta E_{\mathrm{form},ij,\mathrm{OSDA,mol}}$, and $\Delta E_{\mathrm{form},ij,\mathrm{OSDA,atom}}$), with $0.73 \leq \rho^2 \leq 0.80$. This correlation indicates that $\Delta E_{ij,\mathrm{Si}}$ may have matched well with the net thermodynamic transformation during synthesis and, as such, led to good literature recall despite excluding zeolite formation energies. These correlations may explain why the metrics discussed above all appear to produce recall AUCs > 0.6: multiple descriptors depend on one another or correlate with one another. As such, the OSDA rankings they provide for a given framework (or the framework rankings for an OSDA) appear similar, even if they describe a different set of physical contributions to synthesis.

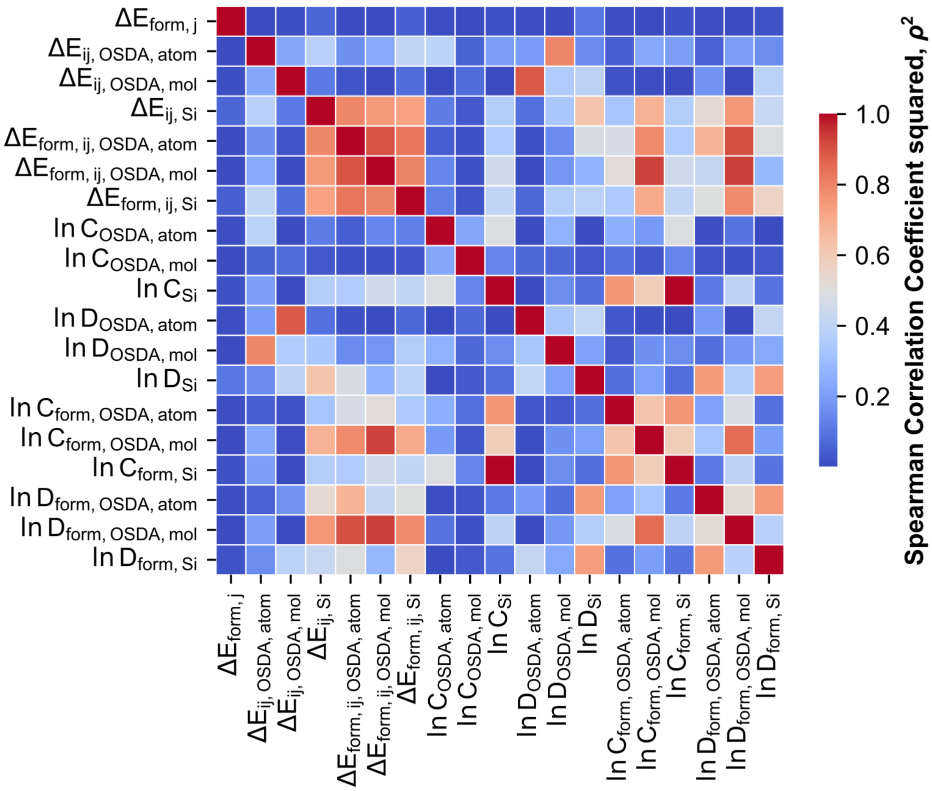


Figure 5: Spearman correlation coefficients squared, ρ^2, for the variables used as possible inputs to SISSO runs for potential energies.

When performing symbolic regression with SISSO, only the units of any input variables are incorporated during descriptor construction. As such, two variables may correlate with one another, but in final descriptors may produce values of significantly different scale if swapped. As such, we include all variables that were part of the original work, such as binding energies $\Delta E_{ij,\mathrm{Si}}$ and $\Delta E_{ij,\mathrm{OSDA}}$, and the logarithms of C and D values from these binding energies. We included $\Delta E_{ij,\mathrm{OSDA}}$ values normalized both per OSDA molecule in the unit cell ($\Delta E_{ij,\mathrm{OSDA,mol}}$) and per OSDA atom in the unit cell ($\Delta E_{ij,\mathrm{OSDA,atom}}$). The earlier $E_{ij,T}$ metric was computed from C and D using $\Delta E_{ij,\mathrm{OSDA,atom}}$, which is similar in scale to $\Delta E_{ij,\mathrm{Si}}$ such that the respective C and D values were also similar in scale. Additionally, we use the logarithms of C and D because their values cover many orders of magnitude, which causes issues during SISSO descriptor construction. Using these variables allows SISSO to recreate the $E_{ij,T}$ metric or augment it with other values if that produces the best predictions. We also include the formation energy $\Delta E_{\mathrm{form},j}$ and formation affinities per OSDA and per Si. We only exclude C and D values derived from formation affinities when they correlate strongly with other variables available during symbolic regression ($\rho^2 > 0.8$). As such, we exclude $\ln C_{\mathrm{form},ij,\mathrm{OSDA,mol}}$, $\ln C_{\mathrm{form},ij,\mathrm{OSDA,atom}}$, and $\ln D_{\mathrm{form},ij,\mathrm{OSDA,mol}}$. In the next section, we assess the equations produced by SISSO from these variables.

2.3.2. Descriptors for synthesis

We saved a total of 50,000 equations from the SISSO assessment we performed on the binding and formation energy data. These equations were assessed and filtered by fitting decision trees and logistic curves as described in Section 5.3. These machine learning (ML) approaches can fit data quickly so that descriptors which have decision boundaries that best classify synthesized zeolite-OSDA pairs can be identified, filtered, and the best descriptors evaluated more thoroughly. We use this SISSO approach to build a physically informed method of ranking OSDAs for synthesis of a given framework to identify promising templates for zeolite phase selectivity. Traditional classifiers like random forests and neural networks struggle to identify zeolite-OSDA pairs (Figure A.1, SI). Once the top-performing SISSO descriptors from the decision tree and logistic curve fits were identified, they were used to rank OSDAs for a given zeolite (or zeolite frameworks for a given OSDA) and the literature recall computed from synthesized zeolite-OSDA pairs (see Section 2.3). Lower values for the descriptor—as with $E_{ij,T}$—indicate that an OSDA is more likely to produce a given framework.

The descriptor that provides the highest average recall AUC across all zeolites, which we call $\alpha_{ij,T}$, is

$$\alpha_{ij,T} = \frac{\ln C_{\text{form},ij,\text{Si}}}{\ln D_{\text{OSDA},ij,\text{mol}} \times \ln D_{\text{form},ij,\text{Si}}} - \frac{\Delta E_{ij,\text{Si}}}{\Delta E_{ij,\text{OSDA, mol}}}. \tag{13}$$

This descriptor provides modest improvement on the recall AUC performance of the templating energy $E_{ij,T}$ when averaged across zeolite frameworks and OSDAs, with an increase from 0.683 to 0.718 and from 0.675 to 0.716, respectively (Table 3). In the case of recall for the zeolites shown earlier in Figure 3, recall improves for the three frameworks for which recall was already relatively good (MFI, AEI, and CHA), but does not improve significantly for MTW or MOR (Figure 6c). The $\alpha_{ij,T}$ descriptor also correlates quite strongly with the $E_{ij,T}$ metric that we assessed earlier in this work (Figure 6a). This correlation further indicates that there are several metrics that capture much of the thermodynamics of zeolite synthesis but which include different arrangements of these contributing variables.

Table 3: Average literature recall areas-under-the-curve (AUC) for the templating energy from earlier work ($E_{ij,T}$) compared to the $\alpha_{ij,T}$ metric.

Metric	**Average Recall AUC**	
	Framework	OSDA
$E_{ij,T}$	0.683	0.675
$\Delta E_{\text{form},ij,\text{Si}}$	0.688	0.723
$\alpha_{ij,T}$	0.718	0.716

After we downselect the descriptors developed by SISSO to the top 600 from decision tree and logistic curve classification, we compute the average recall AUCs across zeolites and use those to rank the performance of each descriptor. Among the top 100 equations with the highest zeolite recall AUCs, the majority contain $\ln D_{\text{OSDA,mol}}$, including the top-performing descriptor that we identified, $\alpha_{ij,T}$ (Figure 6c). Such commonalities indicate that directivity from per-molecule OSDA binding energies plays an outsized role in determining which OSDA best fits a given framework. In 70% of all descriptors where $\ln D_{\text{OSDA,mol}}$ appears, it is in the denominator of the expression. This preference for dividing by $\ln D_{\text{OSDA,mol}}$ shows that favorable binding energies of OSDAs in a framework relative to other OSDAs drive selectivity. Stronger binding energies (more negative) produce D values that are closer

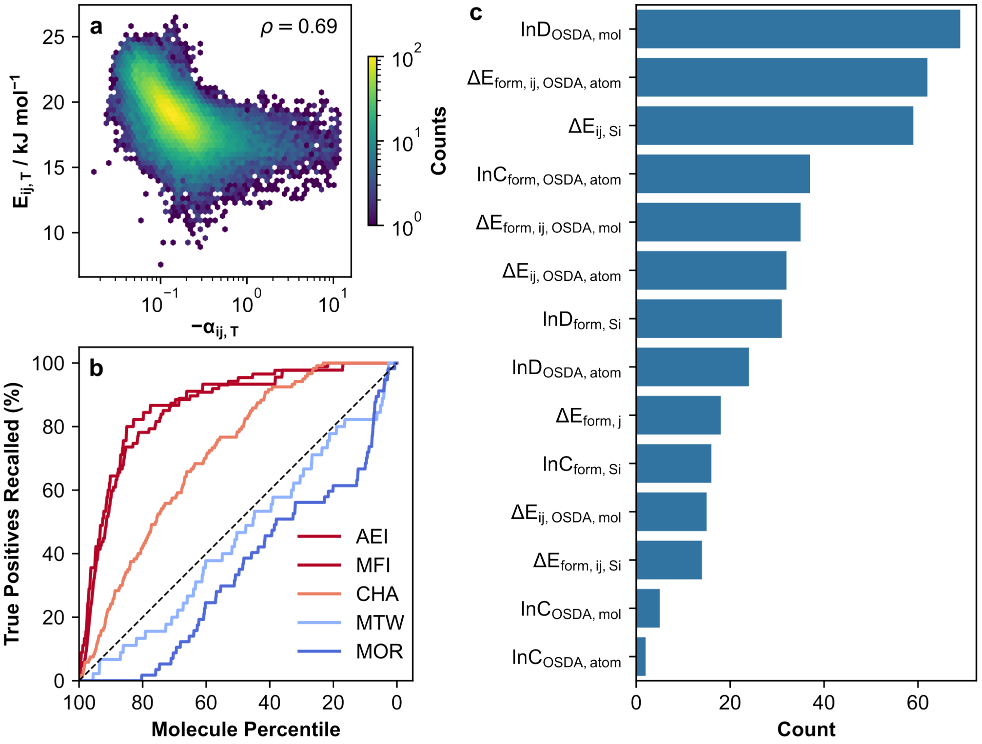


Figure 6: (**a**) Templating energy ($E_{ij,T}$) as a function of the (negative) best SISSO-identified descriptor, $\alpha_{ij,T}$ (with outliers removed). (**b**) Zeolite recall curves from $\alpha_{ij,T}$ for the exemplar frameworks shown in Figure 3. (**c**) The frequency with which each available variable appears in the top 100 equations produced by SISSO (AUC > 0.68).

to unity; the inverse of the logarithm of those values then, is more negative than corresponding D values of OSDAs with weaker binding energies. For example, for a hypothetical group of OSDAs whose binding energies (per molecule) follow an arithmetic series of -100 to -10 kJ mol^{-1} (each separated by 10 kJ mol^{-1}), the lowest value of $(\ln D_{\text{OSDA,mol}})^{-1}$ is an order of magnitude more negative than for a geometric series of binding energies, where many OSDAs are clustered closer to -100 kJ mol^{-1} (Figure D.1, SI). As such, $(\ln D_{\text{OSDA,mol}})^{-1}$ captures the influence of OSDA binding strength relative to other OSDAs more strongly and disproportionately appears in high-performing descriptors.

The first term of the $\alpha_{ij,T}$ descriptor $(\ln C_{\text{form,Si}}/(\ln D_{\text{OSDA,mol}} \times \ln D_{\text{form,Si}}))$ contains only variables from the C and D metrics, while the second term

$(\Delta E_{ij,\mathrm{Si}}/\Delta E_{ij,\mathrm{OSDA,mol}})$ contains only binding energies. The first term appears to capture the relative performance of an OSDA for a given framework. These two terms have similar ranges (with the exception of some outliers in the first term); as such, we expect that they contribute similarly to the computed ranking for AUC calculation. When simplified, the second part of the equation simplifies to:

$$\frac{\Delta E_{ij,\mathrm{Si}}}{\Delta E_{ij,\mathrm{OSDA,mol}}} = \frac{n_{\mathrm{OSDA,mol}}}{n_{\mathrm{Si}}} \tag{14}$$

where $n_{\mathrm{OSDA,mol}}$ is the loading of OSDA molecules per unit cell of the framework and n_{Si} is the number of Si tetrahedral atoms per unit cell. Together, these terms indicate that the best OSDA for a given zeolite is one which has a much stronger binding energy than competing molecules and which can produce a high loading per tetrahedral site.

3. Discussion

This work relies on relatively low-fidelity binding energy data to make predictions about which OSDA best templates a given zeolite framework or to predict the framework that would result from a given OSDA. Because binding energies of OSDAs within frameworks were computed using a DREIDING force field rather than more accurate DFT methods, predictions of synthesis outcomes may be significantly worse than they might otherwise be. Moreover, our binding energy calculations use a simplified model to assess OSDA fit only (Figure 1). While we assess these binding energies in pure silica materials, our past work has shown that this approach works well to identify the fit of these molecules within zeolites and identify promising OSDAs for zeolite materials with a wide range of Si/Al [10, 15, 32]. When we compute literature recall to evaluate the metrics we use, we include all possible frameworks for two reasons. First, this approach appears to work despite its constrained framework composition. Second, when attempting to identify an OSDA to synthesize a given framework—even if targeting a specific composition—one must compare its phase purity across all possible options [10]. By neglecting the influence of ISDA templating of specific composite building units (CBUs) or the role of heteroatoms in stabilizing some CBUs, these data create an incomplete picture of zeolite synthesis. We also neglect the formation of silanol defects, crystallization kinetics, and other unusual synthesis behaviors that could affect OSDA choice. For example, MWW zeolites form in layers around their OSDAs and are calcined at high temperature to both remove the OSDA and form the final crystal phase [33, 34, 35]. In cases such as these, computing binding energies in the final, fully formed crystal do not accurately reflect the synthesis process.

We also attempted to include entropic effects that might affect zeolite crystallization. The metric that we used in this work is based on an empirical fit to alkane adsorption entropies in a small set of zeolites [20]. We used this metric, estimates of the gas-phase entropies of the OSDAs studied here from statistical mechanics, and tabulated data for occupiable volumes of different zeolites [36] to estimate adsorption entropy of OSDAs within zeolites. However, zeolites are often synthesized hydrothermally, as noted in Section 2.2; a more rigorous investigation of OSDA binding entropy would incorporate these solvent effects and would compute the Gibbs free energy of adsorption (ΔG_{ij}) rather than the Helmholtz free energy (ΔA_{ij}). Additionally, more rigorous investigations on the polarity of the material or distribution of acid sites on the entropy of different adsorbates that develop a similar but more extensive descriptor would be useful. Such an approach would require extensive additional calculations, such as molecular dynamics [37, 38], and is beyond the scope of this work.

The growing accuracy of ML broadly, but especially neural network force fields (NFFs) and machine-learned interatomic potentials (MLIPs), has enabled novel computation-driven methods for materials discovery and synthesis [39, 40]. This work uses a combination of different levels of theory—DFT for formation energies and a force field to estimate OSDA binding energies. In the future, we anticipate that MLIPs may enable rapid assessment of zeolite formation energies and more accurate estimations of OSDA binding energies than those computed with the DREIDING force field. For example, there are already some MLIPs specifically for zeolites [7, 41, 42, 43, 44, 45] and a growing number of potentials for general materials chemistry purposes like CHGNet [46] and MACE-MP-0 [47], both of which are trained on data from the Materials Project [48, 49]. Beyond MLIPs, other ML methods have been used to predict materials properties of zeolite frameworks [50, 51]. Even without Achieving DFT-level accuracy for all of the metrics used in this work and for additional compositions besides siliceous frameworks may enable more robust predictions of synthesis outcomes *in silico* that can guide experimental procedures.

Finally, the standard by which we are measuring the performance of our metrics—literature recall AUC—may be an imperfect measure of success.

Some of the OSDAs that we have identified here may work well for some frameworks, but the specific synthesis recipe that could produce such materials requires additional information to develop (*e.g.*, the inclusion of particular ISDAs, synthesis time). Many materials have only a few publications or patents about their syntheses and are rarely studied; in these cases, literature recall proves more difficult because of the scarcity of synthesis data. Moreover, this work treats the absence of a publication using a given OSDA to synthesize a zeolite as a true negative; however, such absences may reflect a dearth of attempts to develop synthesis recipes for a given OSDA-zeolite pair. While there are some cases where OSDA molecules cannot template a given framework, assigning a quantity to these cases remains challenging. For example, the OSDA 1-methyl-4-[3-[1-methyl-1-(phenylmethyl)piperidin-1-ium-4-yl]propyl]-1-(phenylmethyl)piperidin-1-ium (see Figure E.1 in the SI for structure) is extremely large and has only been used to synthesize beta zeolites [52, 53, 54, 55]. Our past work was unable to dock this OSDA within several frameworks—such as CHA and LTA—which may suggest that these frameworks can never be synthesized using this OSDA and may most closely represent a "true negative" OSDA-zeolite pair. Such non-binding OSDA-zeolite pairs are excluded from our assessment of literature recall because no binding energy (and therefore no $E_{ij,T}$) could be calculated. The inability to fit within the zeolite pores is already an indication of the OSDA's inability to template the given zeolite. Additionally, the simple thermodynamic calculations used here may not capture the full complexity of zeolite formation, where nucleation mechanisms remain mysterious and methods to control formation are still nascent [56]. As such, *a priori* zeolite synthesis prediction and guidance remains challenging for some zeolite frameworks that require highly specific OSDAs and synthesis recipes to form, despite recent progress and the growing availability of more accurate tools to drive data-driven synthesis [9, 57].

4. Conclusions

We expand our previous studies of OSDA selection using high-throughput approaches by including additional factors that may influence zeolite synthesis. Specifically, we test whether including formation energies of zeolite frameworks ($\Delta E_{\text{form},j}$) and estimates of OSDA binding entropies in frameworks (ΔS_{ij}) improves predictions of experimentally validated OSDA-zeolite pairs. The net transformation during OSDA-guided zeolite synthesis involves the crystallization of the framework material around an OSDA molecule. We compute the energy associated with this transformation normalized per Si atom in the zeolite ($\Delta E_{\mathrm{form},ij,\mathrm{Si}}$) and compare it to a previous metric that we used to rank the best OSDAs for zeolite synthesis, the templating energy ($E_{ij,T}$) [10]. Using this energy slightly improves predictions for the most likely zeolite to be produced from a given OSDA, but does not markedly improve identification of the best OSDA for a targeted zeolite material.

Chemical intuition alone may not produce the best descriptor to predict OSDA-zeolite pairs. As such, we turn to ML techniques to construct a better equation using SISSO. The descriptor from SISSO that we identified with the best performance for predicting OSDA-zeolite pairs, $\alpha_{ij,T}$ (Eq. 13), only provides a marginal improvement over $E_{ij,T}$ for these predictions. This descriptor has two terms: a term containing only values of competition (C) and directivity (D) that capture differences in how well an OSDA templates a given framework; and a term that depends on the ratio of the number of OSDAs to the number of framework Si atoms in the material. The first term contains $(\ln D_{\mathrm{form},Si})^{-1}$, which appears the most frequently out of any possible variables in the equations from SISSO that have the highest literature recall AUC performance on average. This finding suggests that directivity of an OSDA—specifically derived from the formation affinity of that OSDA-zeolite pair—disproporionately contributes to the likelihood of the OSDA-zeolite pair being experimentally validated. This assessment did identify a descriptor that predicts which OSDA-zeolite pairs are most likely slightly better than descriptors developed based on chemical intuition; however, this new descriptor $\alpha_{ij,T}$ does not improve significantly over those developed from chemical intuition. As such, using chemically reasonable descriptors may remain desirable because they are more readily interpretable.

Predicting the outcomes of zeolite synthesis or the best OSDA to synthesize a given framework remains challenging. The data used here to develop this descriptor were collected using relatively low-accuracy methods within a high-throughput virtual screening framework that aims to filter out unlikely candidates and identify the most promising candidates. Higher fidelity data may provide different recalls and rankings than those produced here, but methods for gathering those data remain costly. Finally, literature recall provides an imperfect metric for assessing the performance of these metrics. Zeolite and zeotype synthesis often involves a complex mixture of a silica source, heteroatoms, an OSDA, ISDAs, and other mineralizing agents, all of which may affect the framework that is produced. Literature data also do not necessarily contain all possible OSDA-zeolite pairs; some OSDAs may be able to synthesize frameworks that these metrics have identified as promising, but the correct recipe has not yet been identified and verified in literature. The absence of an OSDA-zeolite pair from literature is treated here as a true negative, but such true negatives cannot actually exist within the synthesis literature. This work provides a metric to more accurately predict viable OSDA-zeolite pairs, but more accurate data and faster predictions will be necessary to guide zeolite synthesis reliably *a priori*.

5. Methods

5.1. Density Functional Theory Calculations and atomistic zeolite models

Density functional theory (DFT) calculations were done using the Vienna ab initio simulation package (VASP) [58, 59, 60]. Planewaves were created with an energy cutoff of 600 eV using the projector-augmented wave (PAW) method [61]. Calculations were performed using the Perdew–Burke–Ernzerhof (PBE) form of the generalized gradient approximation (GGA) [62], with the DFT-D3 correction to account for van der Waals and dispersive interactions [63]. Monkhorst-Pack k-point meshes for each calculation were constructed to maintain a constant sampling density of 64 k-points Å^{-3}. Hexagonal cells used Γ-centered k-point meshes. Each structure was optimized such that both the atomic positions and unit cell were permitted to relax (ISIF = 3 in VASP). Self-consistent field (SCF) cycles were considered converged once energies varied by $< 10^{-6}\,\text{eV}$. Geometries were optimized until forces on each atom were $< 10^{-2}\,\text{eV}\,\text{Å}^{-1}$. All zeolite structures were retrieved as CIF files from the database of the International Zeolite Association (IZA) [64].

5.2. Entropy corrections

Entropies were estimated based on an empirical fit to gas-phase adsorption data of alkanes in many different zeolite frameworks [20]. This earlier work found that the standard state adsorption entropy of a molecule i in zeolite j (ΔS°_{ij}) could be described by

$$-\Delta S^\circ_{ij} = S^\circ_{\text{1D,trans},i} + \left(F_{\text{rot, slab}} + \frac{1}{7}\left[\left(1 - \frac{V_{\text{crit}}}{2V_{\text{occ},j}}\right)^{-3} - 1\right]\right) S^\circ_{\text{rot},i} \quad (15)$$

In this equation, $S^{\circ}_{\mathrm{1D,trans},i}$ is the standard one-dimensional translational entropy of the molecule as an ideal gas, $F_{\text{rot, slab}}$ is the fractional loss in rotational entropy when a molecule adsorbs to a slab, V_{crit} is a critical volume at which all rotational entropy is lost, $V_{\mathrm{occ},j}$ is the occupiable volume available for zeolite j within a $1000\,\mathrm{\AA}^3$ cube of space in the crystal, and $S^{\circ}_{\mathrm{rot},i}$ is the standard rotational entropy of the molecule as an ideal gas [20]. The values of $F_{\text{rot, slab}}$ and V_{crit} are 0.03 (unitless) and $127.3\,\mathrm{\AA}^3$, and both are constant for all zeolites. The value of $V_{\mathrm{occ},j}$ depends on the zeolite, and was calculated in earlier work by filling the voids of the zeolite with spheres of diameter $2.8\,\mathrm{\AA}$ to approximate packing water molecules within the zeolite voids [36]. Binding entropies were only estimated for zeolite frameworks for which occupiable volumes were previously calculated.

We computed the gas-phase translational and rotational entropies of all OSDA molecules in the Organic Structure-directing agent DataBase (OSDB) [10]. Translational entropies were estimated using the Sackur-Tetrode equation [65], which was also used in the work that discovered Eq. 15:

$$S_{\mathrm{trans},i}(T) = S^{\circ}_{\mathrm{Ar,298\,K}} + R\ln\left[\left(\frac{m_i}{m_{\mathrm{Ar}}}\right)^{\frac{3}{2}}\left(\frac{T}{298\,\mathrm{K}}\right)^{\frac{5}{2}}\right] \tag{16}$$

where $S^{\circ}_{\mathrm{Ar,298\,K}}$ is the gas-phase entropy of Ar at 298 K and 1 bar (or 154.8 $\mathrm{J\,(mol\,K)^{-1}}$), R is the universal gas constant, $\frac{m_i}{m_{\mathrm{Ar}}}$ is the ratio of the mass of the molecule to the mass of an Ar atom, and T is the temperature in K. The rotational entropy of each molecule was calculated according to ideal-gas statistical mechanics used in the same work:

$$S_{\mathrm{rot},i}(T) = R\left(\ln\left[\frac{\sqrt{\pi I_A I_B I_C}}{\sigma}\left(\frac{8\pi^2 k_B T}{h^2}\right)^{\frac{3}{2}}\right] + \frac{3}{2}\right) \tag{17}$$

where I_A, I_B, and I_C are the principle moments of inertia of a molecule; σ is the symmetry number, k_B is Boltzmann's constant, and h is Planck's constant. Previous work has generated three-dimensional molecular conformers for all of the OSDA molecules that were studied in this work [10]. Up to five conformers were generated using the `RDKit` software package for each molecule (v. 2024.3.1) [66]. The energies of each conformer were estimated using the semi-empirical tight binding scheme (GFN-xTB) [67]. Moments of inertia and symmetry numbers were estimated based on the conformer with the lowest GFN-xTB energy. Symmetry numbers were computed using the `pymatgen` package (v. 2024.5.1) [68].

5.3. Descriptor fitting

The sure independence screening and sparsifying operation (SISSO) code [69] was used to generate a one-dimensional descriptor (*i.e.*, a single equation) to describe zeolite synthesis outcomes based on zeolite formation energies, binding energies, and the C and D metrics described in Eq. 2 and 3. We used SISSO to develop a descriptor for classification that could distinguish between positive and negative synthesis outcomes but which can also be used to rank OSDAs for a given framework or frameworks for a given OSDA. As such, we constrain our primary study to one-dimensional descriptors that produce a single numeric value for a given OSDA-zeolite pair. While SISSO can develop multi-dimensional descriptors (that is, descriptors composed of multiple equations), such equations may not be useful for ranking OSDA-zeolite pairs. Users could fit multi-dimensional logistic curves to multi-dimensional equations, but such efforts would only serve for classification purposes. Alternatively, linear combinations of equations produced by SISSO could be used—where each equation was multiplied by a fitted parameter; however, the combinatorial space of such an effort would be enormous and would risk overfitting. Therefore, we pursue only a one-dimensional descriptor from SISSO in this work.

SISSO constructs features by iteratively performing functional or algebraic operations on a set of variables. These variables are provided as inputs to the primary feature space, $\mathbf{\Phi_0}$, and have the corresponding data against which fitting is performed provided in an external file. A set of operations on $\mathbf{\Phi_0}$ produces the $\mathbf{\Phi_1}$ feature space; for example, if $\mathbf{\Phi_0}$ contains variables a and b with the same units, $\mathbf{\Phi_1}$ may contain $a+b$, $a-b$, a/b, a^2, and so on. Subsequent spaces $\mathbf{\Phi_n}$ are composed of items from previous feature spaces $\mathbf{\Phi_{m<n}}$. The maximum permitted feature complexity used in this work was 4 (*i.e.,* up to $\mathbf{\Phi_4}$). A total of 50,000 descriptors were saved in the output of the SISSO runs (the `SISSO.in` file is provided in the Supporting Information). Importantly, this approach and the operators used with SISSO would allow SISSO to reproduce the earlier equation that we used to rank zeolite-OSDA pairs (Eq. 1) [10].

The binding energy data were taken from earlier work [10], while formation energies were computed here using the DFT methods described above in Section 5.1. The binding energies were calculated for OSDAs in pure-silica zeolite frameworks with the DREIDING force field [14]—an approach that was shown to correlate well with DFT-calculated binding energies for a subset of these zeolite-OSDA pairs [15]. This dataset contains zeolite frameworks that have been synthesized successfully in labs and all known OSDAs to template these frameworks (at the time of its publication).

The dataset for binding energies contains a total of 112,426 OSDA-zeolite pairs—far too large to perform SISSO. These pairs were generated from a list of OSDAs that were compiled from previous publications and patents for zeolite synthesis using natural language processing [70]. Where possible, any molecules used as OSDAs within this synthesis dataset were docked within siliceous forms of all zeolite frameworks. Because of the size of the dataset, we take a stratified sample of 5000 zeolite-OSDA pairs that have not been identified in literature and 500 zeolite-OSDA pairs that have produced zeolite frameworks to perform the SISSO analysis. After removing highly correlated C and D variables derived from formation affinities (see Section 2.3.1), we performed sparse regression using the remaining 12 variables shown in Figure 5 on this sample of the full dataset.

Once the descriptors were generated with SISSO, we fit each equation to the outcomes in the sample dataset using decision trees and logistic regression with the scikit-learn package [71]. Decision trees for classification were fit to all 50,000 equations with a depth of two, similar to an earlier approach used to evaluate SISSO equations to predict perovskite synthesis outcomes [72]. Logistic curves were fit with the descriptor as the independent variable and the synthesis outcome for the framework-OSDA pair, where 0 meant the zeolite-OSDA pair had not been found in literature and 1 meant that it had, also similar to earlier work studying ion ordering in perovskites [73]. Once all equations had been assessed on the training sample using this method, they were ranked based on their relative performances for classification with each method. The top 200 equations based on decision trees, 200 equations from logistic regression (excluding any whose decision tree classifiers had already been selected), and 200 with the highest average ranking between the two (again excluding redundant descriptors) were selected for further filtering. The descriptors were then computed for all data points in the full dataset for which all datapoints were available. These descriptors—alongside $E_{ij,T}$, $E_{\mathrm{form},ij,T}$, $\Delta E_{\mathrm{form},ij,\mathrm{Si}}$, and $A_{ij,T}$—were then evaluated by computing literature recall area-under-the-curve (AUC), where OSDAs are ranked for a given framework (or frameworks ranked for a given OSDA) by their descriptor values from low to high and the number of experimentally validated zeolite-OSDA pairs counted for each ranking (see Section 2.1 for examples). Zeolites and OSDAs for which binding entropies could not be estimated using the equations from Ref. [20] from tabulated values in Ref. [36] were excluded from this analysis. High AUCs indicate that rankings provided by the descriptors tend to place known zeolite-OSDA pairs at a low value. Some of the filtered descriptors whose decision tree or logistic curve classifications performed well had AUCs < 0.5, which indicated that they instead ranked zeolite-OSDA pairs in the opposite order of how AUC rankings were computed for the other descriptors. In these cases, the descriptors were multiplied by -1 and the AUCs recomputed. The best descriptor from Eq. 13 was determined as the equation that had the best AUC.

5.4. Machine Learning Classifier Tests

We fit two ML models to our data to predict the presence of OSDA-zeolite pairs in literature. These models were trained as a benchmark for our SISSO descriptor and to contrast classifier models with a physically informed descriptor to rank OSDAs. Both models were trained using the same data and variables that were used to evaluate the SISSO descriptors. The code for these models has been made publicly available with the rest of our code and data (see Code Availability Statement).

First, we fit a random forest classifier from the scikit-learn package [71]. A grid search for the best hyperparameters suggested that a model with 100 decision tree estimators, each with a maximum depth of 30, had the best five-fold cross-validation score. The confusion matrix showing this models performance is shown in Figure A.1a in the SI.

Finally, we fit a neural network (NN) to classify zeolite-OSDA literature pairs using PyTorch (v. 2.4.1) [74]. The NN model had three hidden layers, each of which used the rectified linear unit (ReLU) activation function, until passed through a final sigmoidal layer to predict the probability a zeolite-OSDA pair was in the literature. The model was trained for 20 epochs and used a binary cross entropy loss function. The confusion matrix for this model is shown in Figure A.1b in the SI.

6. Acknowledgments

This work was funded by BASF and ExxonMobil. We thank Elton Pan (MIT), Elsa Olivetti (MIT), Anthony Debellis (BASF), Vivek Vattipalli (BASF), Ahmad Moini (BASF), Di (Daniel) Du (ExxonMobil), and Allen Burton (ExxonMobil) for helpful discussions and suggestions.

We acknowledge the MIT Office of Research Computing and Data for providing high performance computing resources that contributed to the research results reported within this paper.

7. CRediT Author Statement

A.J.H.: conceptualization, investigation, formal analysis, data curation, visualization, writing - original draft. **M.X.**: investigation, data curation, writing - review & editing. **R.G-B.**: conceptualization, supervision, funding acquisition, writing - review & editing.

8. Code Availability Statement

The code and data to reproduce the analysis and figures from this work are available at https://github.com/learningmatter-mit/ZeoliteSynMetrics.

References

[1] C. Baerlocher, L. B. McCusker, D. H. Olson, Atlas of Zeolite Framework Types, 6th Edition, Elsevier Science, Amsterdam, 2007, pages: 404.

[2] R. Pophale, P. A. Cheeseman, M. W. Deem, A database of new zeolite-like materials, Physical Chemistry Chemical Physics 13 (27) (2011) 12407–12412, publisher: The Royal Society of Chemistry. doi:10.1039/c0cp02255a.

[3] M. W. Deem, R. Pophale, P. A. Cheeseman, D. J. Earl, Computational Discovery of New Zeolite-Like Materials, J. Phys. Chem. C 113 (51) (2009) 21353–21360, publisher: American Chemical Society. doi:10.1021/jp906984z.

[4] Atlas of Prospective Zeolite Structures.
URL http://www.hypotheticalzeolites.net/

[5] M. M. J. Treacy, I. Rivin, E. Balkovsky, K. H. Randall, M. D. Foster, Enumeration of periodic tetrahedral frameworks. II. Polynodal graphs, Microporous and Mesoporous Materials 74 (1) (2004) 121–132. doi:10.1016/j.micromeso.2004.06.013.

[6] Y. Li, X. Li, J. Liu, F. Duan, J. Yu, In silico prediction and screening of modular crystal structures via a high-throughput genomic approach, Nat Commun 6 (1) (2015) 8328, publisher: Nature Publishing Group. doi:10.1038/ncomms9328.

[7] A. Erlebach, P. Nachtigall, L. Grajciar, Accurate large-scale simulations of siliceous zeolites by neural network potentials, arXiv preprint arXiv:2102.12404 (2021) 1–33.

[8] B. A. Helfrecht, G. Pireddu, R. Semino, S. M. Auerbach, M. Ceriotti, Ranking the synthesizability of hypothetical zeolites with the sorting hat, Digital Discovery 1 (6) (2022) 779–789, publisher: Royal Society of Chemistry. doi:10.1039/D2DD00056C.

[9] E. Pan, S. Kwon, Z. Jensen, M. Xie, R. Gómez-Bombarelli, M. Moliner, Y. Román-Leshkov, E. Olivetti, ZeoSyn: A Comprehensive Zeolite Synthesis Dataset Enabling Machine-Learning Rationalization of Hydrothermal Parameters, ACS Cent. Sci. 10 (3) (2024) 729–743, publisher: American Chemical Society. doi:10.1021/acscentsci.3c01615.

[10] D. Schwalbe-Koda, S. Kwon, C. Paris, E. Bello-Jurado, Z. Jensen, E. Olivetti, T. Willhammar, A. Corma, Y. Román-Leshkov, M. Moliner, R. Gómez-Bombarelli, E. Olivett, T. Willhammar, A. Corma, Y. Román-Leshkov, M. Moliner, R. Gómez-Bombarelli, A priori control of zeolite phase competition and intergrowth with high-throughput simulations, Science 374 (6565) (2021) 308–315, publisher: American Association for the Advancement of Science. doi:10.1126/science.abh3350.

[11] D. Schwalbe-Koda, A. Corma, Y. Román-Leshkov, M. Moliner, R. Gómez-Bombarelli, Data-Driven Design of Biselective Templates for Intergrowth Zeolites, The Journal of Physical Chemistry Letters 12 (43) (2021) 10689–10694, publisher: American Chemical Society. doi:10.1021/acs.jpclett.1c03132.

[12] E. Bello-Jurado, D. Schwalbe-Koda, M. Nero, C. Paris, T. Uusimäki, Y. Román-Leshkov, A. Corma, T. Willhammar, R. Gómez-Bombarelli, M. Moliner, Tunable CHA/AEI Zeolite Intergrowths with A Priori Biselective Organic Structure-Directing Agents: Controlling Enrichment and Implications for Selective Catalytic Reduction of NOx, Angewandte Chemie International Edition 61 (28) (Jul. 2022). doi:10.1002/anie.202201837.

[13] S. Kwon, E. Bello-Jurado, E. Ikonnikova, H. Lee, D. Schwalbe-Koda, A. Corma, T. Willhammar, E. A. Olivetti, R. Gomez-Bombarelli, M. Moliner, Y. Román-Leshkov, One-Pot Synthesis of CHA/ERI-Type Zeolite Intergrowth from a Single Multiselective Organic Structure-Directing Agent, ACS Appl. Mater. Interfaces 16 (12) (2024) 14661–14668, publisher: American Chemical Society. doi:10.1021/acsami.3c15810.

[14] S. L. Mayo, B. D. Olafson, W. A. Goddard, DREIDING: A generic force field for molecular simulations, Journal of Physical Chemistry 94 (26) (1990) 8897–8909, publisher: BioDesign, Inc. doi:10.1021/j100389a010.

[15] D. Schwalbe-Koda, R. Gomez-Bombarelli, Benchmarking binding energy calculations for organic structure-directing agents in pure-silica zeolites, Journal of Chemical Physics 154 (17) (2021) 174109–174109, pub-

lisher: AIP Publishing LLCAIP Publishing. doi:10.26434/chemrxiv.13270184.v2.

[16] C. Waitt, X. Gao, R. Gounder, A. Debellis, S. Prasad, A. Moini, W. F. Schneider, Analysis and Augmentation of Guest–Host Interaction Energy Models as CHA and AEI Zeolite Crystallization Phase Predictors, J. Phys. Chem. C 127 (46) (2023) 22740–22751, publisher: American Chemical Society. doi:10.1021/acs.jpcc.3c05421.

[17] O. F. Altundal, S. Leon, G. Sastre, Different Zeolite Phases Obtained with the Same Organic Structure Directing Agent in the Presence and Absence of Aluminum: The Directing Role of Aluminum in the Synthesis of Zeolites, J. Phys. Chem. C 127 (22) (2023) 10797–10805, publisher: American Chemical Society. doi:10.1021/acs.jpcc.3c01567.

[18] O. F. Altundal, G. Sastre, The Directing Role of Aluminum in the Synthesis of PST-21 (PWO), PST-22 (PWW), and ERS-7 (ESV) Zeolites, J. Phys. Chem. C 127 (31) (2023) 15648–15656, publisher: American Chemical Society. doi:10.1021/acs.jpcc.3c03640.

[19] C. T. Campbell, J. R. V. Sellers, The Entropies of Adsorbed Molecules, J. Am. Chem. Soc. 134 (43) (2012) 18109–18115, publisher: American Chemical Society. doi:10.1021/ja3080117.

[20] P. J. Dauenhauer, O. A. Abdelrahman, A Universal Descriptor for the Entropy of Adsorbed Molecules in Confined Spaces, ACS Cent. Sci. 4 (9) (2018) 1235–1243, publisher: American Chemical Society. doi:10.1021/acscentsci.8b00419.

[21] N. J. Henson, A. K. Cheetham, J. D. Gale, Theoretical Calculations on Silica Frameworks and Their Correlation with Experiment, Chem. Mater. 6 (10) (1994) 1647–1650, publisher: American Chemical Society. doi:10.1021/cm00046a015.

[22] Y. G. Bushuev, G. Sastre, Feasibility of Pure Silica Zeolites, J. Phys. Chem. C 114 (45) (2010) 19157–19168, publisher: American Chemical Society. doi:10.1021/jp107296e.

[23] P. M. Piccione, C. Laberty, S. Yang, M. A. Camblor, A. Navrotsky, M. E. Davis, Thermochemistry of Pure-Silica Zeolites, J. Phys. Chem.

B 104 (43) (2000) 10001–10011, publisher: American Chemical Society. doi:10.1021/jp002148a.

[24] Z. Jensen, S. Kwon, D. Schwalbe-Koda, C. Paris, R. Gómez-Bombarelli, Y. Román-Leshkov, A. Corma, M. Moliner, E. A. Olivetti, Discovering Relationships between OSDAs and Zeolites through Data Mining and Generative Neural Networks, ACS Cent. Sci. 7 (5) (2021) 858–867, publisher: American Chemical Society. doi:10.1021/acscentsci.1c00024.

[25] S. Ferdov, K. Tsuchiya, N. Tsunoji, T. Sano, Comparative study between high-silica faujasites (FAU) from organic-free system and the commercial zeolite Y, Microporous and Mesoporous Materials 276 (2019) 154–159. doi:10.1016/j.micromeso.2018.09.036.

[26] M. Maldonado, M. D. Oleksiak, S. Chinta, J. D. Rimer, Controlling Crystal Polymorphism in Organic-Free Synthesis of Na-Zeolites, J. Am. Chem. Soc. 135 (7) (2013) 2641–2652, publisher: American Chemical Society. doi:10.1021/ja3105939.

[27] J. Wang, P. Liu, M. Boronat, P. Ferri, Z. Xu, P. Liu, B. Shen, Z. Wang, J. Yu, Organic-Free Synthesis of Zeolite Y with High Si/Al Ratios: Combined Strategy of In Situ Hydroxyl Radical Assistance and Post-Synthesis Treatment, Angewandte Chemie 132 (39) (2020) 17378–17381, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ange.202005715. doi:10.1002/ange.202005715.

[28] L. Zhang, S. Xie, W. Xin, X. Li, S. Liu, L. Xu, Crystallization and morphology of mordenite zeolite influenced by various parameters in organic-free synthesis, Materials Research Bulletin 46 (6) (2011) 894–900. doi:10.1016/j.materresbull.2011.02.018.

[29] T. Xiao, M. Yabushita, T. Nishitoba, R. Osuga, M. Yoshida, M. Matsubara, S. Maki, K. Kanie, T. Yokoi, W. Cao, A. Muramatsu, Organic Structure-Directing Agent-Free Synthesis of Mordenite-Type Zeolites Driven by Al-Rich Amorphous Aluminosilicates, ACS Omega 6 (8) (2021) 5176–5182, publisher: American Chemical Society. doi:10.1021/acsomega.0c05059.

[30] M. T. Conato, M. D. Oleksiak, B. P. McGrail, R. K. Motkuri, J. D. Rimer, Framework stabilization of Si-rich LTA zeolite prepared in organic-free media, Chemical Communications 51 (2) (2015) 269–272, publisher: Royal Society of Chemistry. doi:10.1039/C4CC07396G.

[31] E. O. Pyzer-Knapp, C. Suh, R. Gómez-Bombarelli, J. Aguilera-Iparraguirre, A. Aspuru-Guzik, D. R. Clarke, What Is High-Throughput Virtual Screening? A Perspective from Organic Materials Discovery, Annual Review of Materials Research 45 (1) (2015) 195–216–195–216, publisher: Annual Reviews. doi:10.1146/annurev-matsci-070214-020823.

[32] M. Xie, D. Schwalbe-Koda, Y. M. Semanate-Esquivel, E. Bello-Jurado, A. Hoffman, O. Santiago-Reyes, C. Paris, M. Moliner, R. Gómez-Bombarelli, An exhaustive mapping of zeolite-template chemical space (Oct. 2024). doi:10.26434/chemrxiv-2024-d74sw.

[33] M. E. Leonowicz, J. A. Lawton, S. L. Lawton, M. K. Rubin, MCM-22: A Molecular Sieve with Two Independent Multidimensional Channel Systems, Science 264 (5167) (1994) 1910–1913, publisher: American Association for the Advancement of Science. doi:10.1126/science.264.5167.1910.

[34] A. Corma, V. Fornes, S. B. Pergher, T. L. M. Maesen, J. G. Buglass, Delaminated zeolite precursors as selective acidic catalysts, Nature 396 (6709) (1998) 353–356, publisher: Nature Publishing Group. doi:10.1038/24592.

[35] A. Corma, C. Corell, J. Pérez-Pariente, Synthesis and characterization of the MCM-22 zeolite, Zeolites 15 (1) (1995) 2–8. doi:10.1016/0144-2449(94)00013-I.

[36] M. M. J. Treacy, M. D. Foster, Packing sticky hard spheres into rigid zeolite frameworks, Microporous and Mesoporous Materials 118 (1) (2009) 106–114. doi:10.1016/j.micromeso.2008.08.039.

[37] K. Alexopoulos, M.-S. Lee, Y. Liu, Y. Zhi, Y. Liu, M.-F. Reyniers, G. B. Marin, V.-A. Glezakou, R. Rousseau, J. A. Lercher, Anharmonicity and Confinement in Zeolites: Structure, Spectroscopy, and Adsorption Free

Energy of Ethanol in H-ZSM-5, J. Phys. Chem. C 120 (13) (2016) 7172–7182, publisher: American Chemical Society. doi:10.1021/acs.jpcc.6b00923.

[38] D. R. Galimberti, J. Sauer, Chemically Accurate Vibrational Free Energies of Adsorption from Density Functional Theory Molecular Dynamics: Alkanes in Zeolites, J. Chem. Theory Comput. 17 (9) (2021) 5849–5862, publisher: American Chemical Society. doi:10.1021/acs.jctc.1c00519.

[39] S. Ma, Z.-P. Liu, Machine learning potential era of zeolite simulation, Chemical Science 13 (18) (2022) 5055–5068, publisher: Royal Society of Chemistry. doi:10.1039/D2SC01225A.

[40] M. Moliner, Y. Román-Leshkov, A. Corma, Machine Learning Applied to Zeolite Synthesis: The Missing Link for Realizing High-Throughput Discovery, Acc. Chem. Res. 52 (10) (2019) 2971–2980, publisher: American Chemical Society. doi:10.1021/acs.accounts.9b00399.

[41] A. Erlebach, M. Šípka, I. Saha, P. Nachtigall, C. J. Heard, L. Grajciar, A reactive neural network framework for water-loaded acidic zeolites, Nat Commun 15 (1) (2024) 4215, publisher: Nature Publishing Group. doi:10.1038/s41467-024-48609-2.

[42] M. Bocus, R. Goeminne, A. Lamaire, M. Cools-Ceuppens, T. Verstraelen, V. Van Speybroeck, Nuclear quantum effects on zeolite proton hopping kinetics explored with machine learning potentials and path integral molecular dynamics, Nat Commun 14 (1) (2023) 1008, publisher: Nature Publishing Group. doi:10.1038/s41467-023-36666-y.

[43] R. Millan, E. Bello-Jurado, M. Moliner, M. Boronat, R. Gomez-Bombarelli, Effect of Framework Composition and NH3 on the Diffusion of Cu+ in Cu-CHA Catalysts Predicted by Machine-Learning Accelerated Molecular Dynamics, ACS Cent. Sci. 9 (11) (2023) 2044–2056, publisher: American Chemical Society. doi:10.1021/acscentsci.3c00870.

[44] I. Saha, A. Erlebach, P. Nachtigall, C. J. Heard, L. Grajciar, Germanium Distributions in Zeolites Derived from Neural Network Potentials. (Jun. 2024). doi:10.26434/chemrxiv-2024-qp5bb.

[45] M. Zheng, B. C. Bukowski, Probing the Role of Acid Site Distribution on the Water Structure in Aluminosilicate Zeolites: Insights from Molecular Dynamics, J. Phys. Chem. C 128 (18) (2024) 7549–7559, publisher: American Chemical Society. doi:10.1021/acs.jpcc.4c01087.

[46] B. Deng, P. Zhong, K. Jun, J. Riebesell, K. Han, C. J. Bartel, G. Ceder, CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling, Nat Mach Intell 5 (9) (2023) 1031–1041, publisher: Nature Publishing Group. doi:10.1038/s42256-023-00716-3.

[47] I. Batatia, P. Benner, Y. Chiang, A. M. Elena, D. P. Kovács, J. Riebesell, X. R. Advincula, M. Asta, W. J. Baldwin, N. Bernstein, A. Bhowmik, S. M. Blau, V. Cărare, J. P. Darby, S. De, F. Della Pia, V. L. Deringer, R. Elijošius, Z. El-Machachi, E. Fako, A. C. Ferrari, A. Genreith-Schriever, J. George, R. E. A. Goodall, C. P. Grey, S. Han, W. Handley, H. H. Heenen, K. Hermansson, C. Holm, J. Jaafar, S. Hofmann, K. S. Jakob, H. Jung, V. Kapil, A. D. Kaplan, N. Karimitari, N. Kroupa, J. Kullgren, M. C. Kuner, D. Kuryla, G. Liepuoniute, J. T. Margraf, I.-B. Magdău, A. Michaelides, J. H. Moore, A. A. Naik, S. P. Niblett, S. W. Norwood, N. O'Neill, C. Ortner, K. A. Persson, K. Reuter, A. S. Rosen, L. L. Schaaf, C. Schran, E. Sivonxay, T. K. Stenczel, V. Svahn, C. Sutton, C. van der Oord, E. Varga-Umbrich, T. Vegge, M. Vondrák, Y. Wang, W. C. Witt, F. Zills, G. Csányi, A foundation model for atomistic materials chemistry, arXiv:2401.00096 [physics] (Dec. 2023).

[48] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K. A. Persson, Commentary: The Materials Project: A materials genome approach to accelerating materials innovation, APL Materials 1 (1) (2013) 11002–11002, publisher: {AIP} Publishing. doi:10.1063/1.4812323.

[49] A. Jain, G. Hautier, C. J. Moore, S. Ping Ong, C. C. Fischer, T. Mueller, K. A. Persson, G. Ceder, A high-throughput infrastructure for density functional theory calculations, Computational Materials Science 50 (8) (2011) 2295–2310. doi:10.1016/j.commatsci.2011.02.023.

[50] R. Gaillac, S. Chibani, F.-X. Coudert, Speeding Up Discovery of Auxetic Zeolite Frameworks by Machine Learning, Chem. Mater. 32 (6) (2020)

2653–2663, publisher: American Chemical Society. doi:10.1021/acs.chemmater.0c00434.

[51] M. Ducamp, F.-X. Coudert, Prediction of Thermal Properties of Zeolites through Machine Learning, J. Phys. Chem. C 126 (3) (2022) 1651–1660, publisher: American Chemical Society. doi:10.1021/acs.jpcc.1c09737.

[52] K. Tsuji, M. E. Davis, Further investigations on the synthesis of pure-silica molecular sieves via the use of organic structure-directing agents, Microporous Materials 11 (1) (1997) 53–64. doi:10.1016/S0927-6513(97)00024-2.

[53] R. Kore, R. Srivastava, Synthesis of zeolite Beta, MFI, and MTW using imidazole, piperidine, and pyridine based quaternary ammonium salts as structure directing agents, RSC Adv. 2 (26) (2012) 10072–10084, publisher: The Royal Society of Chemistry. doi:10.1039/C2RA20437A.

[54] N. Hould, M. Haouas, V. Nikolakis, F. Taulelle, R. Lobo, Mechanisms of Quick Zeolite Beta Crystallization, Chem. Mater. 24 (18) (2012) 3621–3632, publisher: American Chemical Society. doi:10.1021/cm3020995.

[55] O. Larlus, S. Mintova, S. T. Wilson, R. R. Willis, H. Abrevaya, T. Bein, A powerful structure-directing agent for the synthesis of nanosized Al- and high-silica zeolite Beta in alkaline medium, Microporous and Mesoporous Materials 142 (1) (2011) 17–25. doi:10.1016/j.micromeso.2010.08.025.

[56] R. Jain, A. J. Mallette, J. D. Rimer, Controlling Nucleation Pathways in Zeolite Crystallization: Seeding Conceptual Methodologies for Advanced Materials Design, J. Am. Chem. Soc. 143 (51) (2021) 21446–21460, publisher: American Chemical Society. doi:10.1021/jacs.1c11014.

[57] D. Schwalbe-Koda, D. E. Widdowson, T. Anh Pham, V. A. Kurlin, Inorganic synthesis-structure maps in zeolites with machine learning and crystallographic distances, Digital Discovery 2 (6) (2023) 1911–1924, publisher: Royal Society of Chemistry. doi:10.1039/D3DD00134B.

[58] G. Kresse, J. Furthmüller, Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set, Computational Materials Science 6 (1) (1996) 15–50, publisher: Elsevier. doi:10.1016/0927-0256(96)00008-0.

[59] G. Kresse, J. Furthmüller, Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set, Physical Review B 54 (16) (1996) 11169–11186, publisher: American Physical Society. doi:10.1103/PhysRevB.54.11169.

[60] G. Kresse, D. Joubert, From ultrasoft pseudopotentials to the projector augmented-wave method, Physical Review B 59 (3) (1999) 1758–1775, publisher: American Physical Society. doi:10.1103/PhysRevB.59.1758.

[61] P. E. Blöchl, Projector augmented-wave method, Physical Review B 50 (24) (1994) 17953–17979. doi:10.1103/PhysRevB.50.17953.

[62] J. P. Perdew, K. Burke, M. Ernzerhof, Generalized Gradient Approximation Made Simple, Physical Review Letters 77 (18) (1996) 3865–3868, publisher: American Physical Society. doi:10.1103/PhysRevLett.77.3865.

[63] S. Grimme, J. Antony, S. Ehrlich, H. Krieg, A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu, The Journal of Chemical Physics 132 (15) (2010) 154104–154104. doi:10.1063/1.3382344.

[64] IZA Structure Commission (2018).
URL http://www.iza-structure.org/

[65] D. A. McQuarrie, Statistical Mechanics, 1st Edition, University Science Books, Sausalito, CA, USA, 2000.

[66] G. Landrum, RDKit: Open-source cheminformatics (2006).
URL www.rdkit.org

[67] S. Grimme, C. Bannwarth, P. Shushkov, A Robust and Accurate Tight-Binding Quantum Chemical Method for Structures, Vibrational Frequencies, and Noncovalent Interactions of Large Molecular Systems Parametrized for All spd-Block Elements (Z = 1–86), Journal of

Chemical Theory and Computation 13 (5) (2017) 1989–2009. doi:10.1021/acs.jctc.7b00118.

[68] S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, G. Ceder, Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis, Computational Materials Science 68 (2013) 314–319, publisher: Elsevier. doi:10.1016/J.COMMATSCI.2012.10.028.

[69] R. Ouyang, S. Curtarolo, E. Ahmetcik, M. Scheffler, L. M. Ghiringhelli, SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates, Phys. Rev. Mater. 2 (8) (2018) 083802, publisher: American Physical Society. doi:10.1103/PhysRevMaterials.2.083802.

[70] Z. Jensen, E. Kim, S. Kwon, T. Z. H. Gani, Y. Román-Leshkov, M. Moliner, A. Corma, E. Olivetti, A Machine Learning Approach to Zeolite Synthesis Enabled by Automatic Literature Data Extraction, ACS Cent. Sci. 5 (5) (2019) 892–899, publisher: American Chemical Society. doi:10.1021/acscentsci.9b00193.

[71] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[72] C. J. Bartel, C. Sutton, B. R. Goldsmith, R. Ouyang, C. B. Musgrave, L. M. Ghiringhelli, M. Scheffler, New tolerance factor to predict the stability of perovskite oxides and halides, Science Advances 5 (2) (2019) eaav0693, publisher: American Association for the Advancement of Science. doi:10.1126/sciadv.aav0693.

[73] J. Peng, J. Damewood, R. Gómez-Bombarelli, Data-driven physics-informed descriptors of cation ordering in multicomponent perovskite oxides, CR-PHYS-SC 5 (5), publisher: Elsevier (May 2024). doi:10.1016/j.xcrp.2024.101942.

[74] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf,

E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, PyTorch: An Imperative Style, High-Performance Deep Learning Library, in: Advances in Neural Information Processing Systems, Vol. 32, Curran Associates, Inc., 2019.

Appendix to: Learning descriptors to predict organic structure-directing agent applicability in zeolite synthesis

Appendix A. Classifier Test for Phase Prediction

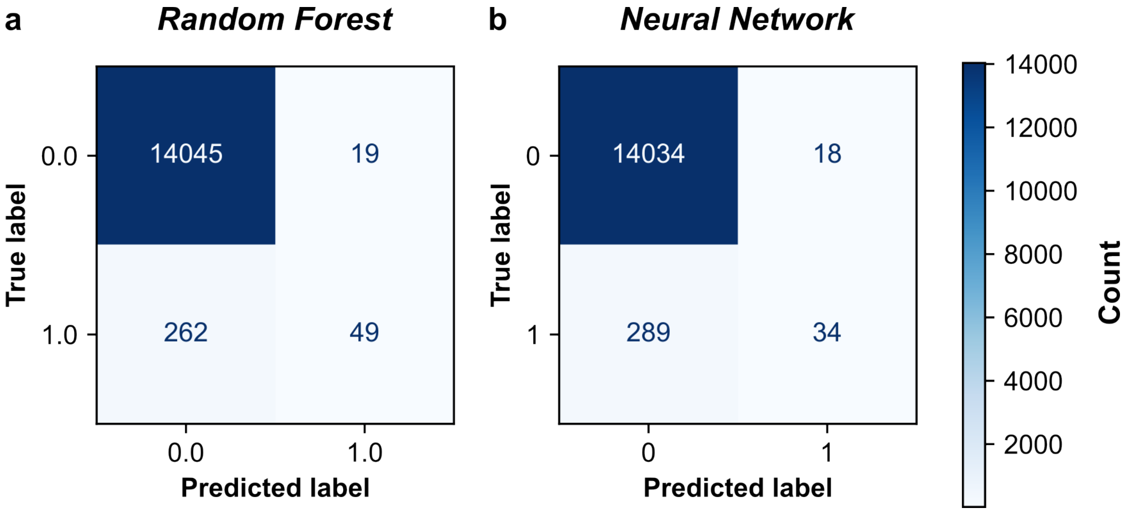


Figure A.1: Confusion matrices on a validation set for (**a**) a random forest and (**b**) neural network classifiers trained on the binding energy from Ref [10].

Appendix B. Zeolite Formation Energy for Phase Prediction

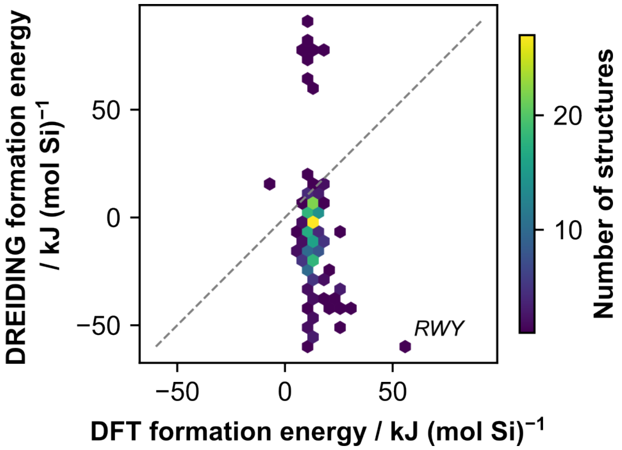


Figure B.1: Comparison of pure-silica zeolite formation energies relative to α-quartz computed with DFT (PBE-D3) and the DREIDING forcefield in kJ (mol Si)$^{-1}$. The point corresponding to the RWY zeolite is highlighted.

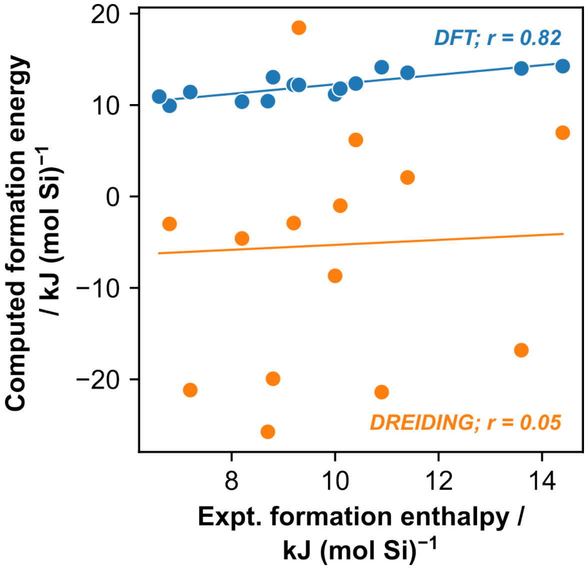


Figure B.2: Comparison of pure-silica zeolite formation energies relative to α-quartz determined experimentally [23] and those computed with DFT (PBE-D3) and the DREIDING forcefield in kJ (mol Si)$^{-1}$. Each series is labeled with the Pearson's correlation coefficient, r.

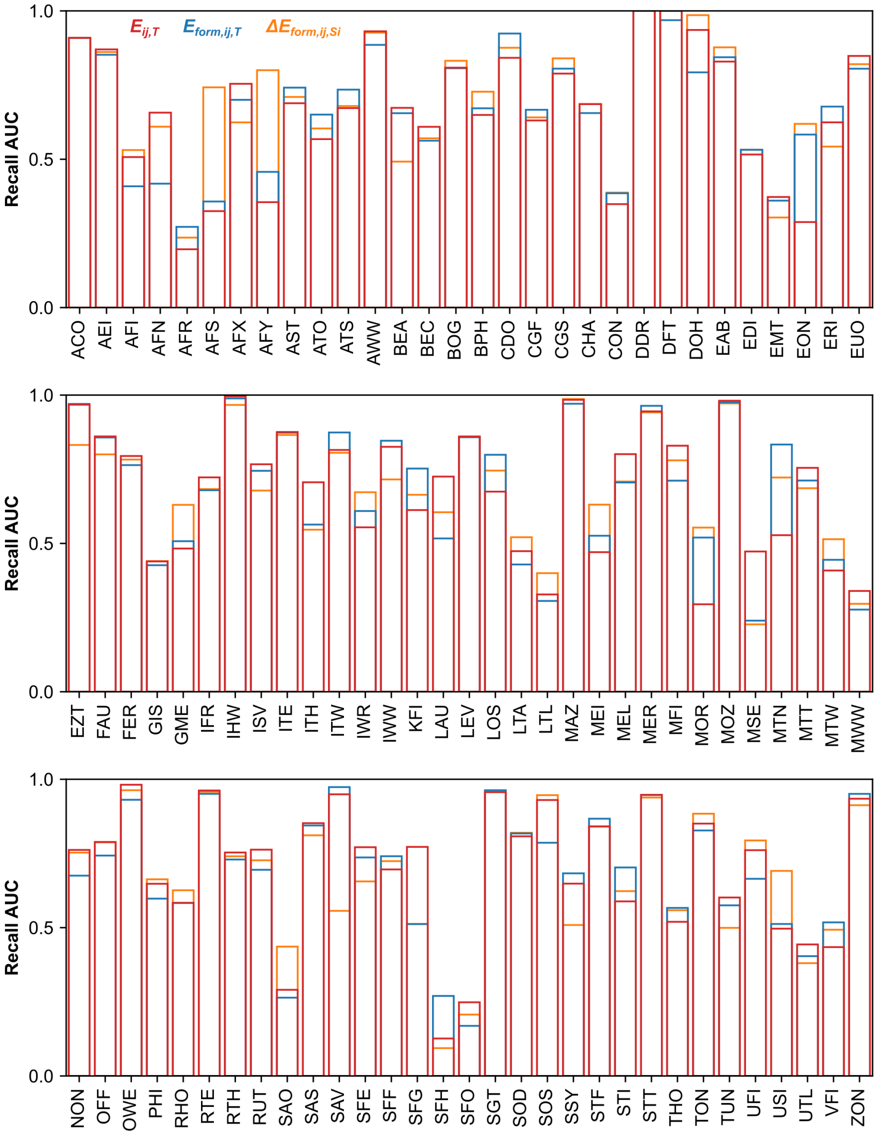


Figure B.3: Recall AUCs for all frameworks considered in this work computed with templating energy ($E_{ij,T}$, red), formation templating energy ($E_{\mathrm{form},ij,T}$, blue), and formation affinity ($\Delta E_{\mathrm{form},ij,\mathrm{Si}}$, orange).

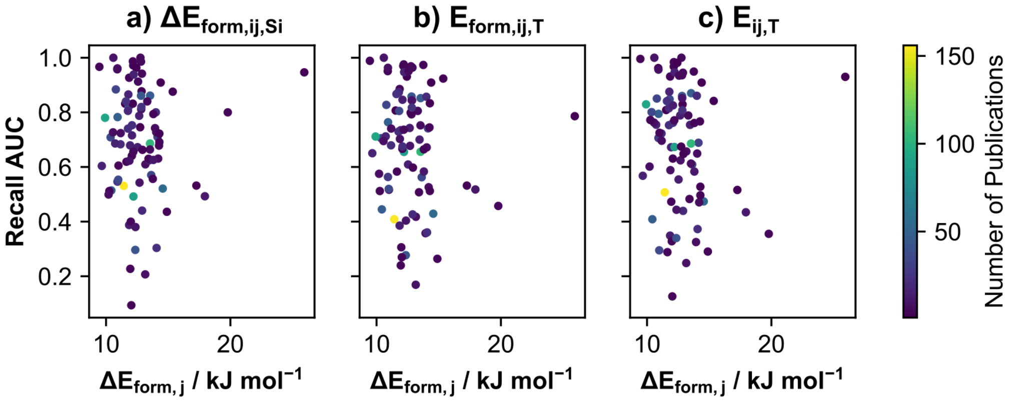


Figure B.4: Literature recall AUC for all frameworks from (**a**) formation affinities per Si atom ($\Delta E_{\text{form},ij,\text{Si}}$), (**b**) formation templating energies ($E_{\text{form},ij,T}$), and (**c**) the templating energy used in earlier work ($E_{ij,T}$), plotted as a function of the formation energy of the underlying zeolite relative to α-quartz ($\Delta E_{\text{form},j}$). Each point is colored by the number of synthesis publications about the framework.

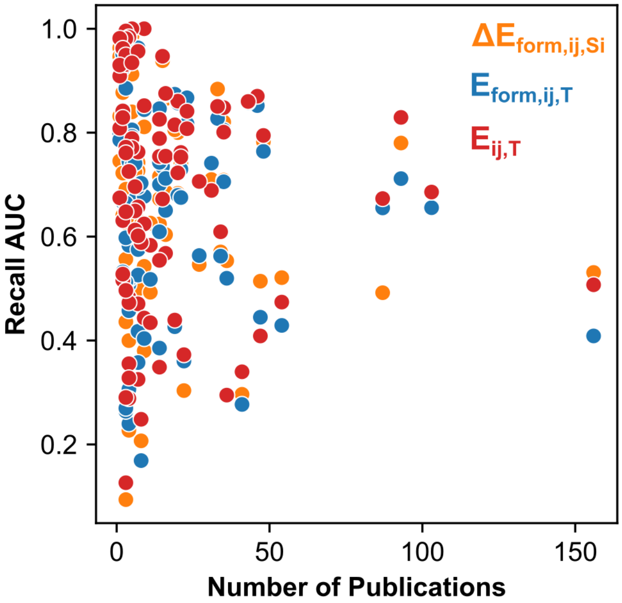


Figure B.5: Literature recall AUC for all frameworks computed using templating energy ($E_{ij,T}$, red), formation templating energy ($E_{\text{form},ij,T}$, blue), and formation affinity ($\Delta E_{\text{form},ij,\text{Si}}$, orange) as a function of the number of publications for each framework.

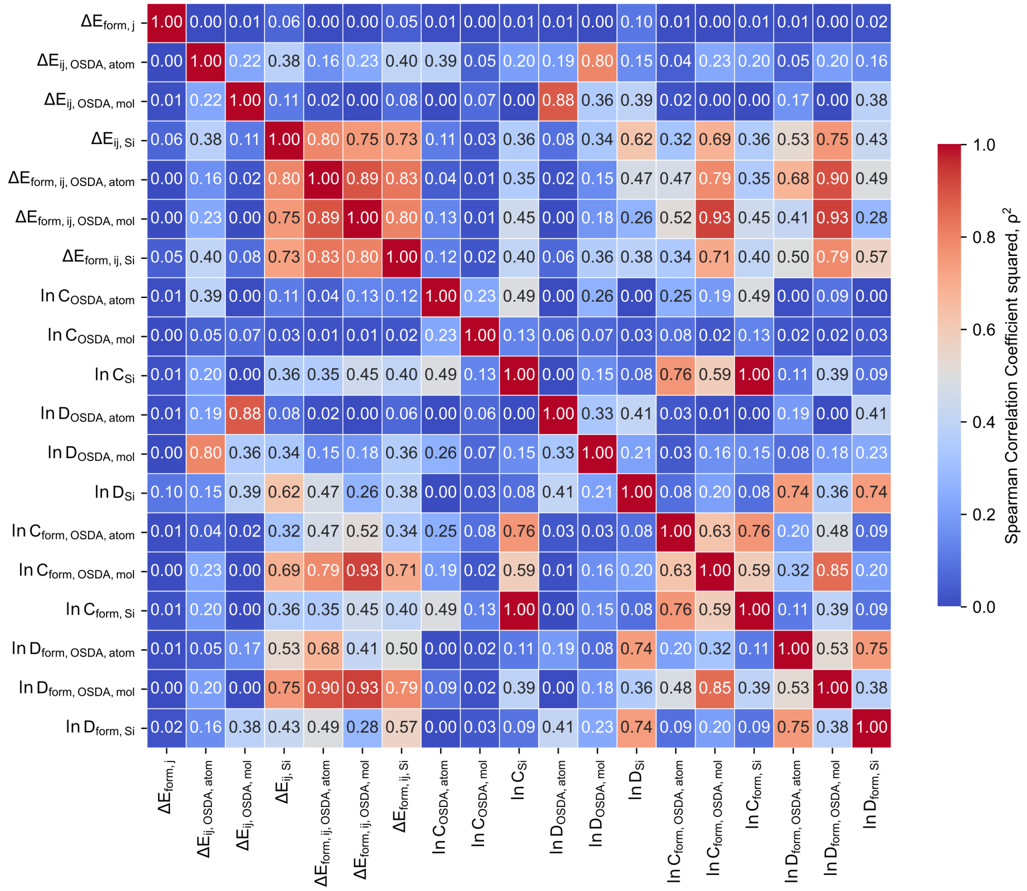


Figure B.6: Heatmap of Spearman correlation coefficients squared, ρ^2, for all of the variables that could be included for symbolic regression based on potential energies (DREIDING binding energies and DFT formation energies).

Appendix C. SISSO equation analysis

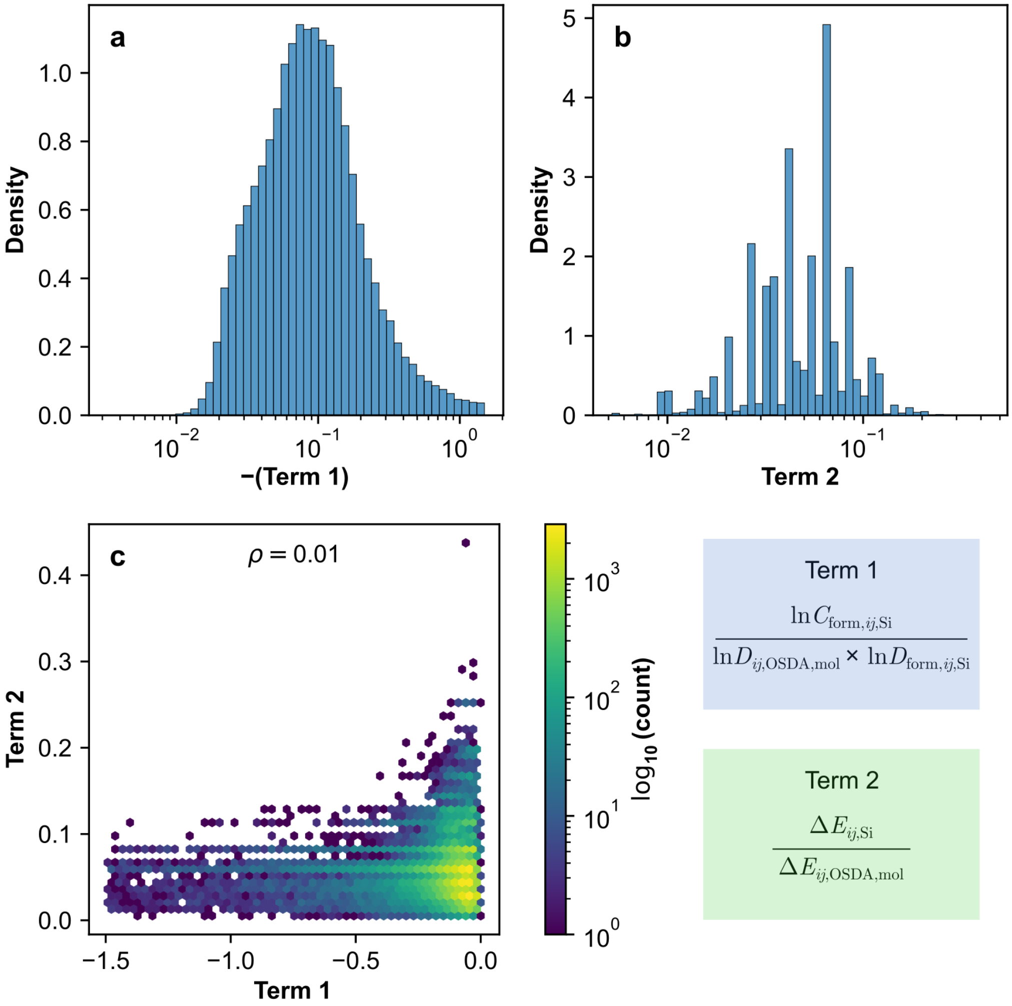


Figure C.1: Distributions of (**a**) the negative of the first term and (**b**) second term in the $\alpha_{ij,T}$ expression. (**c**) The second term as a function of the first term, with the Spearman correlation coefficient (ρ) of the two terms.(**d**) The two terms within $\alpha_{ij,T}$ plotted here.

Appendix D. Binding energy distribution effects on directivity

We illustrate the behavior of the most frequently occurring term in high-performing equations generated by SISSO by constructing two sequences of binding energies. The first is an arithmetic sequence where

$$\Delta E_{ij,\mathrm{arith},n} = -100 + 10(n-1)\ \forall n \in \{0, 1, 2, ..., 9\} \tag{D.1}$$

The second is a geometric sequence where

$$\Delta E_{ij,\mathrm{geom},n} = -101 + 10^{\frac{9-n}{5}}\ \forall n \in \{0, 1, 2, ..., 9\} \tag{D.2}$$

Table D.1: Values for the arithmetic and geometric series of example bining energies per OSDA molecule ($\Delta E_{ij,\mathrm{OSDA,mol}}$) in kJ mol^{-1}.

Arithmetic	**Geometric**
-100	-100
-90	-99.42
-80	-98.49
-70	-97.02
-60	-94.69
-50	-91.00
-40	-85.15
-30	-75.88
-20	-61.19
-10	-37.90

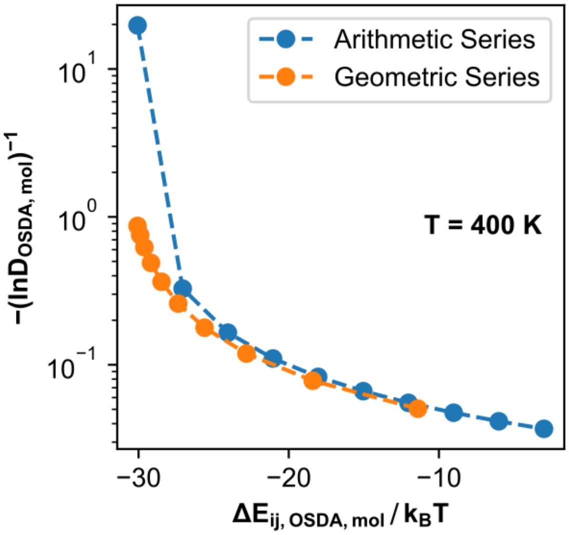


Figure D.1: Dependence of the inverse of the logarithm of the directivity derived from binding energies per OSDA molecule ($(\ln D_{\mathrm{OSDA,mol}})^{-1}$) as a function of binding energies. Two series of arbitrary binding energies are shown: an arithmetic sequence and a geometric sequence, as defined by Eqs. D.1 and D.2, respectively.

Appendix E. Example of large OSDA

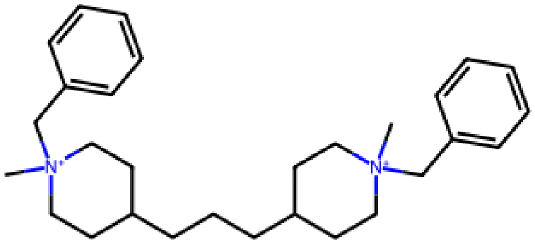


Figure E.1: Structure of 1-methyl-4-[3-[1-methyl-1-(phenylmethyl)piperidin-1-ium-4-yl]propyl]-1-(phenylmethyl)piperidin-1-ium (SMILES: `C[N+]1(Cc2ccccc2)CCC(CCCC2CC[N+](C)(Cc3ccccc3)CC2)CC1`).

Appendix F. SISSO Input

Below are the contents of the `SISSO.in` file for the successful SISSO run performed in this work.

```
ptype=2
ntask=1
desc_dim=1
nsample=(5000,500)
restart=0
nsf=16
ops=‘(+)(-)(*)(/)(exp)(exp-)(^-1)(^2)(log)’
fcomplexity=4
funit=(1:7)
fmax_min=1e-3
fmax_max=1e8
nf_sis=50000
method_so=’L0’
nmodels=50000
isconvex=(1,1)
bwidth=0.001
```