EXPLAINING ZEOLITE SYNTHESIS-STRUCTURE RELATIONSHIPS USING AGGREGATED SHAP ANALYSIS

Elton Pan, Zach Jensen, Rafael Gómez-Bombarelli, & Elsa Olivetti
Department of Materials Science and Engineering
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
{eltonpan,zjensen,rafagb,elsao}@mit.edu

Soonhyoung Kwon, Mingrou Xie, Yuriy Roman
Department of Chemical Engineering
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
{s1105hk,mrx,yroman}@mit.edu

Manuel Moliner
Instituto de Tecnología Química, Universitat Politècnica
de València-Consejo Superior de Investigaciones Científicas
Valencia, Spain
mmoliner@itq.upv.es

ABSTRACT

Zeolites, crystalline aluminosilicate materials with well-defined porous structures, have emerged as versatile materials with applications in carbon capture. Hydrothermal synthesis is a widely used method for zeolite production, offering control over crystallinity and and pore size. However, the intricate interplay of synthesis parameters necessitates a comprehensive understanding to optimize the synthesis process. We train a supervised classification machine learning model on ZeoSyn (a dataset of zeolite synthesis routes) to predict the zeolite framework product given a synthesis route. Subsequently, we leverage SHapley Additive Explanations (SHAP) to reveal key synthesis-structure relationships in zeolites. To that end, we introduce an aggregation SHAP approach to extend such analysis to explain the formation of composite building units (CBUs) of zeolites. Analysis at this unprecedented scale sheds light on key synthesis parameters driving zeolite crystallization.

1 INTRODUCTION

Zeolites are nanoporous, crystalline aluminosilicate materials with a wide range of industrial applications including catalysis, separations, and carbon capture Davis (2002); Boer et al. (2023). While thousands of potential zeolite structures are thought to be thermodynamically accessible Pophale et al. (2011), only about 260 have been synthesized, highlighting a synthesis bottleneck to zeolite discovery and deployment. The synthesis of zeolites synthesis is intricate, with numerous variables influencing the resultant zeolite structure Corma & Davis (2004), such as types and amounts of framework elements, the presence of inorganic and organic cations, structure-directing agents, mineralizing agents and hydrothermal conditions. Davis (2002) However, knowledge of the holistic interplay between these factors across the entire field is lacking. Machine learning has the potential in generalizing some of these relationshipsCorma et al. (2006); Serra et al. (2007); Jensen et al. (2019); Muraoka et al. (2019) but have been limited to subsections of the zeolite design space due to a lack of data.

We leverage ZeoSyn (Fig. S3), a comprehensive dataset of 23,925 zeolite synthesis routes for >200 unique zeolite frameworks (an order of magnitude larger than all previously published zeolite synthesis datasets). We train a supervised classification machine learning model on ZeoSyn to predict zeolite framework products given a synthesis route. We employ SHapley Additive Explanations (SHAP) to reveal key synthesis parameters driving the formation of over 200 zeolite frameworks and their constituent composite building units (CBUs), and show potential applications in phase-selective and intergrowth synthesis. Analysis at this unprecedented level of scale is a step toward an improved understanding of zeolite synthesis-structure relationships, which could potentially guide and accelerate the discovery of new zeolite frameworks.

2 METHODS

2.1 ZEOLITE FRAMEWORK PREDICTION MODEL

We train a supervised classification model using random forest to predict a zeolite framework product (three-letter code eg. **KFI** in Fig. 1a) given a synthesis recipe. The model takes in a 43-dimensional vector as input where each element corresponds to either gel composition (eg. Si, Al, P, etc), reaction condition (eg. crystallization time), or descriptors of organic structure-directing agent (OSDA). The OSDA plays an important role in zeolite synthesis as it as a template, guiding the arrangement of building blocks to form a porous zeolite structure. We featurize the OSDA using its physicochemical descriptors (eg. molecular volume and 2D shape descriptors) of the organic molecule Schwalbe-Koda et al. (2021). The full list of OSDA features and their descriptions can be found in Table S1.

2.2 SHAP ANALYSIS OF ZEOLITE FORMATION

To analyze the outcomes of the classification model (depicted in Fig. 1a, we employ SHAP Lundberg & Lee (2017), which is a generalized measure for the impact of features. SHAP values quantify the relative importance of a specific feature and its impact on classification. We calculate SHAP values at two levels: **1) Framework-level SHAP** quantifies the impact of synthesis parameters on the formation of a zeolite framework, based on the predicted probabilities by the classifier. **2) CBU-level SHAP** quantifies the impact of synthesis parameters on the formation of a composite building unit (CBU). To obtain CBU-level SHAP values, we employ an aggregation approach as follows:

Aggregated SHAP Let $S_f \in \mathbb{R}^{n \times m}$ be the framework-level SHAP matrix of framework f with n observations and m features. The CBU-level SHAP matrix S_{CBU} is given by aggregating framework-level SHAP matrices:

$$S_{CBU} = \sum_{f \in F_{CBU}} S_f \tag{1}$$

where F_{CBU} is the set of synthesized frameworks containing a specific CBU according to the IZA databasebae (2021). For example, to obtain CBU-level SHAP matrix S_{sod} corresponding to the *sod* CBU, we determine the set of frameworks containing *sod*, $F_{sod} = \{\text{FAU}, \text{SOD}, \text{LTA}\}$. Subsequently, the CBU-level SHAP is given by $S_{sod} = S_{\text{FAU}} + S_{\text{SOD}} + S_{\text{LTA}}$. Intuitively, by summing up S_f corresponding to frameworks containing the CBU, this aggregation approach amplifies SHAP values corresponding to *common* features that highly impact CBU formation, while suppressing SHAP values corresponding to the features that do not have much impact. This effectively shifts the SHAP analysis from a framework-centric to a CBU-centric view, allowing for an understanding of factors driving the building units that make up zeolites.

3 RESULTS AND DISCUSSION

3.1 ZEOLITE FRAMEWORK PREDICTION MODEL

The model is evaluated on held-out, unseen test syntheses on the framework prediction task, with a model accuracy of 0.73. This performance is surprisingly strong, as our model has the challenging task of predicting 1 out of 220 possible zeolite frameworks. The confusion matrix shown in Fig. 1b shows that the majority of the predictions lie along the diagonal of the matrix, meaning that the majority of predictions (0.68 – 0.88) are correct. As shown by the high-intensity off-diagonal elements, the majority of the errors made by the model are misclassifications as another framework ("Others") or dense/amorphous dense ("Failed"). The model also shows strong performance in discriminating different pore sizes as shown in Fig. S6 with high accuracies of 0.78 – 0.86.

3.2 RATIONALIZATION OF SYNTHESIS-STRUCTURE RELATIONSHIPS IN ZEOLITES

The synthesis knowledge learned by the classification model can be analyzed to rationalize the impact of synthesis parameters on the formation of a specific zeolite framework. For each prediction, SHAP values are calculated to determine the impact of each synthesis parameter on the probability

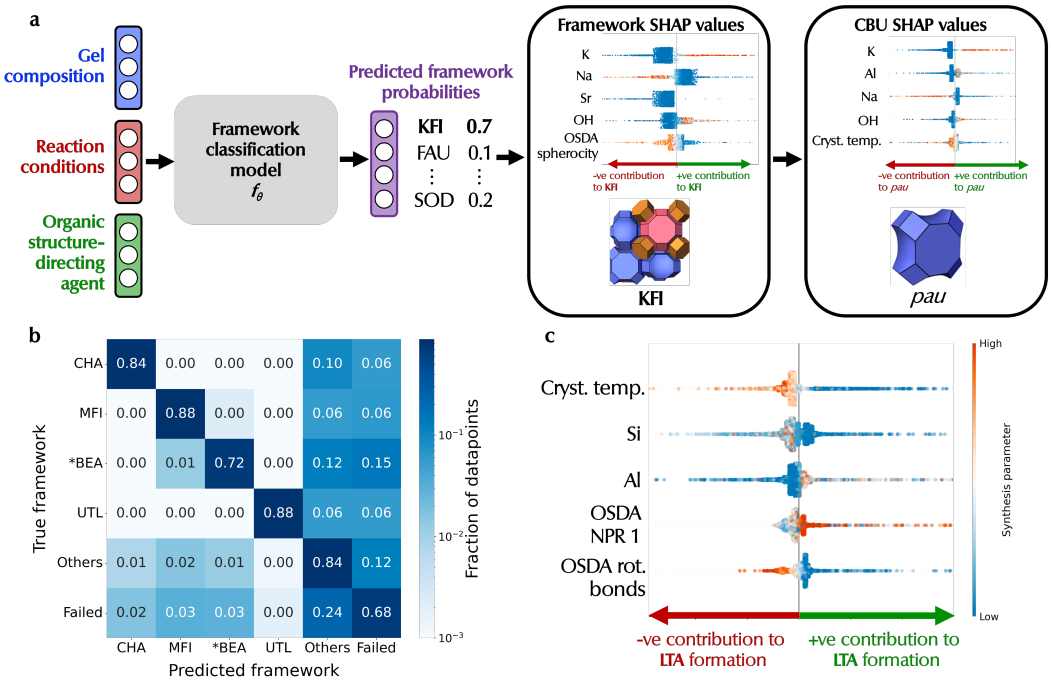

Figure 1: **Interpretable ML framework for explaining synthesis-structure relationships in zeolites (a)** Schematic of zeolite phase predictor model. Given synthesis parameters, the model f_θ predicts the resultant framework (eg. **KFI**). The predicted framework probabilities are used to calculate framework-level SHAP values. In addition, CBU-level SHAP values of composite building units (CBUs) are obtained by aggregated SHAP method. **(b)** Normalized confusion matrix of phase predictor model. Here, we have selected 1 representative small (**CHA**), medium (**MFI**), large (***BEA**), and extra-large pore (**UTL**) framework. "Others" refers to all other frameworks while "Failed" refers to amorphous/dense phases. **(c)** An example of a framework-level SHAP analysis quantifying the positive/negative impact of synthesis parameters on the probability of **LTA** framework formation.

of forming a specific zeolite framework (Fig. 1c). For instance, the first row uncovers a physically-grounded trend that low crystallization temperatures (blue points) have positive SHAP values (increases probability of **LTA** formation), while high temperatures (orange points) have negative SHAP values (decreases probability of **LTA** formation). We quantify the impact of synthesis parameters at two different levels of zeolite structure: **1) Framework-level SHAP** shows the positive/negative impact of a synthesis parameter on the probability of crystallizing a specific zeolite framework (eg. **KFI** in Fig. 1a) **2) CBU-level SHAP** shows the positive/negative impact of a synthesis parameter on the probability of forming a structure that contains a specific composite building unit (CBU) (eg. *pau* cage in Fig. 1a)

3.2.1 Framework-level SHAP

Framework-level SHAP identifies the most important synthesis parameters driving the formation of a specific zeolite framework. Here, we consider all 43 inputs into the model f_θ and show only the top 10 most important synthesis parameters (in descending order) for specific frameworks as shown in Fig. 2a. We note the two different types of synthesis parameters: 1) inorganic, which relate to composition of the inorganic components of the synthesis gel (eg. Si, Al, OH, F etc) 2) OSDA, which relate to the organic template (eg. OSDA volume, OSDA rotatable bonds etc). Consequently, this allows us to categorize each zeolite framework as a one of the following types: 2a: **1) Gel-dominated 2) OSDA-dominated synthesis**.

Gel-dominated synthesis These frameworks have syntheses where inorganic components play a more crucial role, with few ($\leq$3 out of top 10) OSDA-related parameters. Fig. 2a shows two of such frameworks (**CAN**, **KFI**). In terms of the gel composition, **CAN** and **KFI** share the common

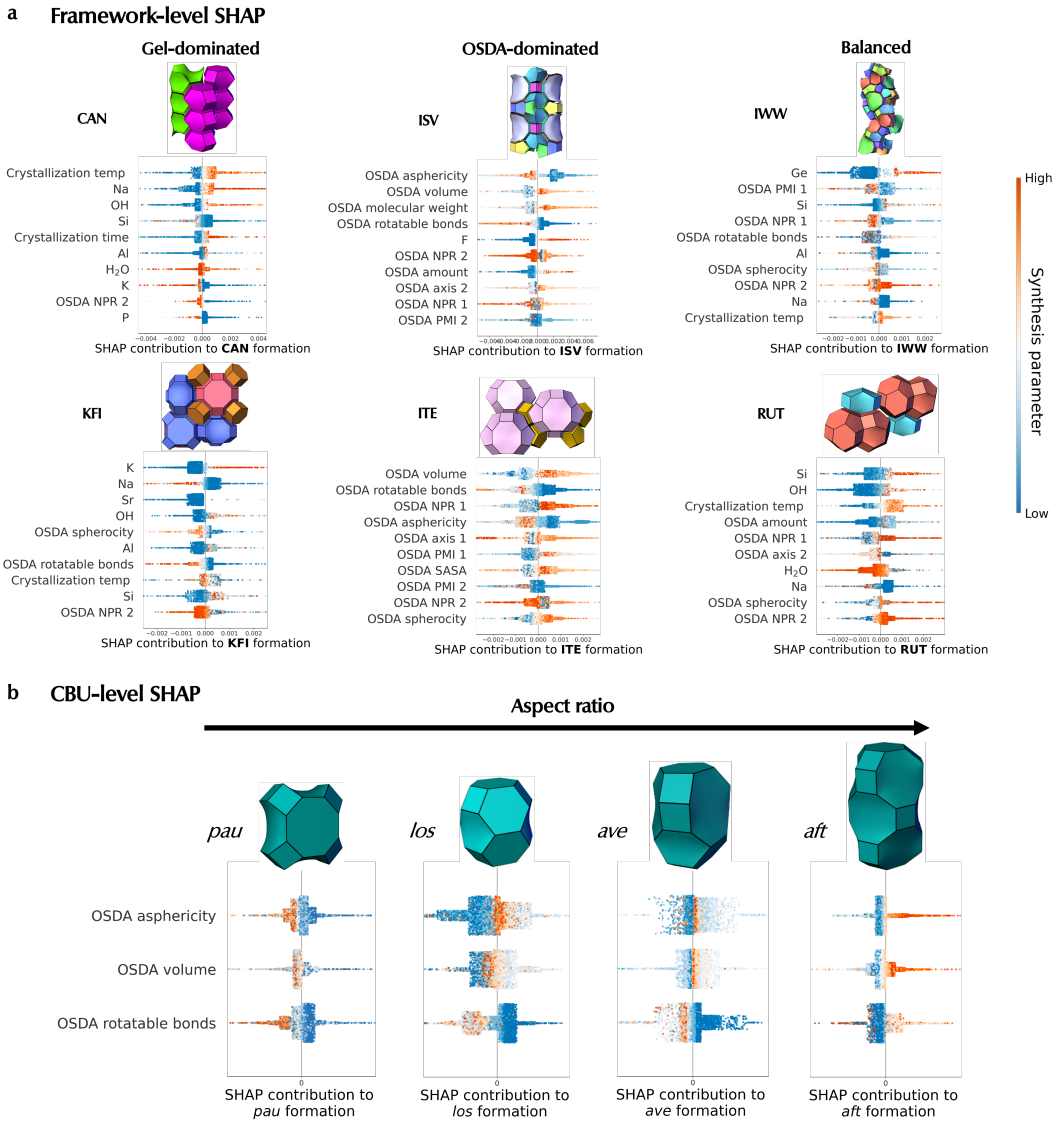

Figure 2: **(a) Framework-level SHAP analysis** revealing the top 10 (out of 43) most important synthesis parameters favoring the formation of specific frameworks. Each framework belongs to one of the following types of synthesis: 1) Gel-dominated synthesis (**CAN**, **KFI**) where most top parameters are inorganic-related, 2) OSDA-dominated synthesis (**ISV**, **ITE**) where most top parameters are OSDA-related, and 3) balanced synthesis (**IWW**, **RUT**) where even attribution is given to inorganic and OSDA parameters. Every point is an individual synthesis colored by the value of synthesis parameter. **(b) CBU-level SHAP analysis** (obtained from aggregated SHAP) of large CBUs showing OSDA parameters favoring their formation.

trend that both are favored by high levels of mineralizing agent OH. **CAN** formation seems to be favored by high Na and low K Barnes et al. (1999). Conversely, **KFI** formation follows the opposite trend, where it appears to be favored by low Na and high K Han et al. (2021). In terms of reaction conditions, high and low crystallization temperatures favor **CAN** (due to high framework density) and **KFI**, respectively Dusselier & Davis (2018).

OSDA-dominated synthesis These frameworks have syntheses where OSDA features are more important. As shown in Fig. 2a, both **ISV** and **ITE** have all of their top synthesis parameters related to the OSDA. One can immediately observe that OSDAs favoring these two frameworks have low asphericity (indicating the need for a spherical OSDA), high volume, and few number of rotatable

bonds (indicating rigidity). However, differences do exist; **ITE** formation is associated with high values of OSDA NPR 1 (first normalized principal moment of inertia ratio) with the orange points clearly on the right hand side, while this effect is not present in **ISV** formation where orange and blue points overlap one another. Moreover, unlike **ITE**, **ISV** requires higher amounts of OSDA. We hypothesize that physicochemical OSDA insights could be used to guide the design of optimal OSDAs that target a specific framework.

3.2.2 CBU-LEVEL SHAP

Zeolites adopt a hierarchical structure where CBUs combine to form the zeolite frameworks themselves. The formation of large CBUs are influenced by OSDA parameters due to the need for a structure-directing effect by OSDAs. Fig. 2b shows a series of large CBUs with an increasing aspect ratio (*pau* $<$ *los* $<$ *ave* $<$ *aft*). In the first row, we discover a clear relationship between aspect ratio of the CBU and OSDA asphericity (a measure of the deviation from sphere). For *pau*, low OSDA asphericity gives rise to positive SHAP values, indicating the need for a spherical OSDA. Indeed, this is due to the symmetrical shape of the *pau* cage. As we transition to higher aspect ratios (*aft*), higher levels of OSDA asphericity (orange) are needed to drive its formation, indicating the increasing need for longer, asymmetric molecules to template CBUs.

3.2.3 APPLICATIONS OF SHAP ANALYSIS

We showcase the utility of the aforementioned SHAP analysis on an important application in zeolite synthesis: **Competing phases** We consider the most common pair of competing phases in the ZeoSyn dataset, **TON** & **MFI** (Fig. S5), where these 2 frameworks are frequently formed in the same synthesis. **MFI** is a framework that often appears as a competing phase due to its ease of synthesis and wide synthesis window. Here, we consider achieving phase-selective of **TON** in the absence of **MFI**. Fig. S9a shows the framework-level SHAP for **TON** and **MFI** frameworks. In order to achieve a phase-selective synthesis of **TON**, one may inspect the impact of OSDA spherocity (first row) on the two frameworks, which reveals *opposing* effects on the frameworks: Clearly, an OSDA with low spherocity promotes **TON** formation while suppressing **MFI** as indicated by the rightmost column. As such, this showcases framework-level SHAP as a powerful tool for identifying promising synthesis "knobs" and recommends the appropriate direction to tune these "knobs" for phase-selective synthesis.

4 CONCLUSION

In this work, we leverage SHAP analysis to uncover the impact of the key synthesis parameters for a zeolites. Furthermore, we introduce an aggregated SHAP approach to extend this analysis to the building unit level, allowing understanding of synthesis parameters at a more fundamental level. Furthermore, this approach has been shown to be useful for the rational design of synthesis parameters for phase-selective synthesis. It is hoped that such rationalization would pave the way for data-driven discovery of zeolitic materials for potential applications in carbon capture.

REFERENCES

Database of zeolite structures. *http://www. iza-structure. org/databases/*, 2021.

Mark C Barnes, Jonas Addai-Mensah, and Andrea R Gerson. The mechanism of the sodalite-to-cancrinite phase transformation in synthetic spent bayer liquor. *Microporous and Mesoporous Materials*, 31(3):287–302, 1999.

Dina G Boer, Jort Langerak, and Paolo P Pescarmona. Zeolites as selective adsorbents for co2 separation. *ACS Applied Energy Materials*, 6(5):2634–2656, 2023.

Avelino Corma and Mark E Davis. Issues in the synthesis of crystalline molecular sieves: Towards the crystallization of low framework-density structures. *ChemPhysChem*, 5(3):304–313, 2004.

Avelino Corma, Manuel Moliner, Jose M Serra, Pedro Serna, María J Díaz-Cabañas, and Laurent A Baumes. A new mapping/exploration approach for ht synthesis of zeolites. *Chemistry of materials*, 18(14):3287–3296, 2006.

Mark E Davis. Ordered porous materials for emerging applications. *Nature*, 417(6891):813–821, 2002.

Michiel Dusselier and Mark E Davis. Small-pore zeolites: synthesis and catalysis. *Chemical reviews*, 118(11):5265–5329, 2018.

Shichao Han, Xiaomin Tang, Lijin Wang, Yanhang Ma, Wei Chen, Qinming Wu, Ling Zhang, Qiuyan Zhu, Xiangju Meng, Anmin Zheng, et al. Potassium-directed sustainable synthesis of new high silica small-pore zeolite with kfi structure (zjm-7) as an efficient catalyst for nh3-scr reaction. *Applied Catalysis B: Environmental*, 281:119480, 2021.

Zach Jensen, Edward Kim, Soonhyoung Kwon, Terry ZH Gani, Yuriy Roman-Leshkov, Manuel Moliner, Avelino Corma, and Elsa Olivetti. A machine learning approach to zeolite synthesis enabled by automatic literature data extraction. *ACS central science*, 5(5):892–899, 2019.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Koki Muraoka, Yuki Sada, Daiki Miyazaki, Watcharop Chaikittisilp, and Tatsuya Okubo. Linking synthesis and structure descriptors from a large collection of synthetic records of zeolite materials. *Nature communications*, 10(1):1–11, 2019.

Ramdas Pophale, Phillip A Cheeseman, and Michael W Deem. A database of new zeolite-like materials. *Physical Chemistry Chemical Physics*, 13(27):12407–12412, 2011.

Daniel Schwalbe-Koda, Soonhyoung Kwon, Cecilia Paris, Estefania Bello-Jurado, Zach Jensen, Elsa Olivetti, Tom Willhammar, Avelino Corma, Yuriy Román-Leshkov, Manuel Moliner, et al. A priori control of zeolite phase competition and intergrowth with high-throughput simulations. *Science*, pp. eabh3350, 2021.

Jose Manuel Serra, Laurent Allen Baumes, Manuel Moliner, Pedro Serna, and Avelino Corma. Zeolite synthesis modelling with support vector machines: a combinatorial approach. *Combinatorial chemistry & high throughput screening*, 10(1):13–24, 2007.

A Appendix

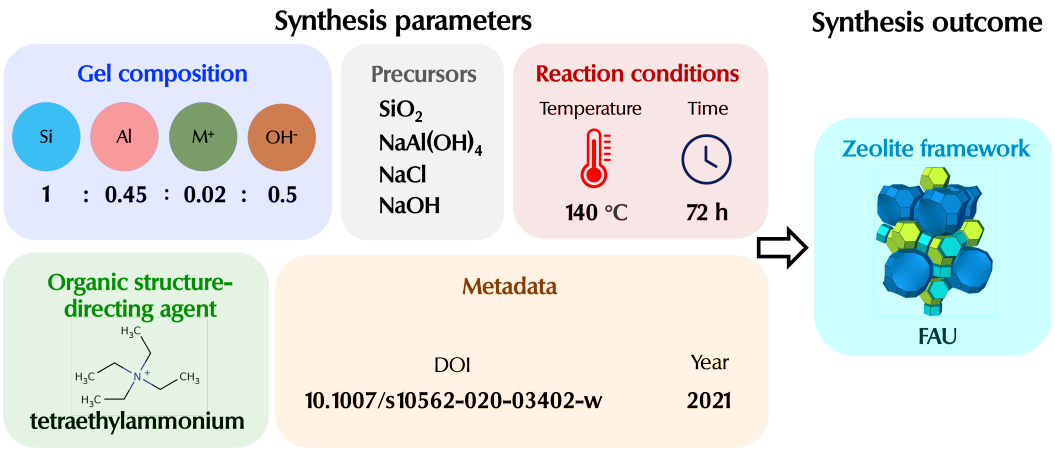

Figure 3: The ZeoSyn dataset.

Table 1: Physicochemical descriptors of OSDAs.

OSDA descriptor	Description
Asphericity	An anisometry descriptor for the deviation from the spherical shape
Axis 1	Two-dimensional (2D) shape descriptors of molecule calculated by projecting the atomic coordinates into a 2D space based on a principal component analysis (PCA) of the positions. The range of the distribution of points in each principal component is reported as the axis of the conformer. Axis 1 is reported as the larger axis, whereas Axis 2 is the smaller axis
Axis 2	See above
Charge	Formal charge of molecule
SASA	Solvent-accessible surface area (SASA) is the surface area of a molecule that is accessible to a solvent
Molecular weight	Molecular mass of molecule
NPR 1	Normalized principal moments ratio (I1/I3) where I is principal moment of inertia
NPR 2	Normalized principal moments ratio (I2/I3) where I is principal moment of inertia
Rotatable bonds	Number of rotatable bonds in the molecule. A measure of molecular flexibility.
PMI 1	Principal moments of inertia (PMI) are physical quantities related to the rotational dynamics of a molecule. $$\mathrm{I} = \sum_{i=1}^{\mathrm{A}} m_i \cdot r_i^2 \quad (2)$$ where A is the number of atoms, and m_i is the atomic mass and r_i is the perpendicular distance from the chosen axis of the ith atom of the molecule
PMI 2	See above
PMI 3	See above
Spherocity	Spherocity index of molecule
Volume	Molecular volume calculated by using a grid-encoding of the molecular shape using a grid spacing of 0.2 Å and 2.0 Å of margin for the boxes

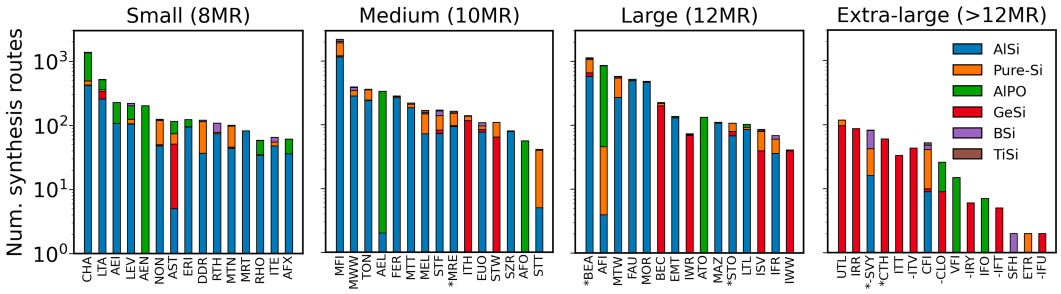

Figure 4: Number of synthetic routes for small, medium, large and extra-large pore frameworks in the dataset. Each framework is further broken down into its constituent zeotypes by color.

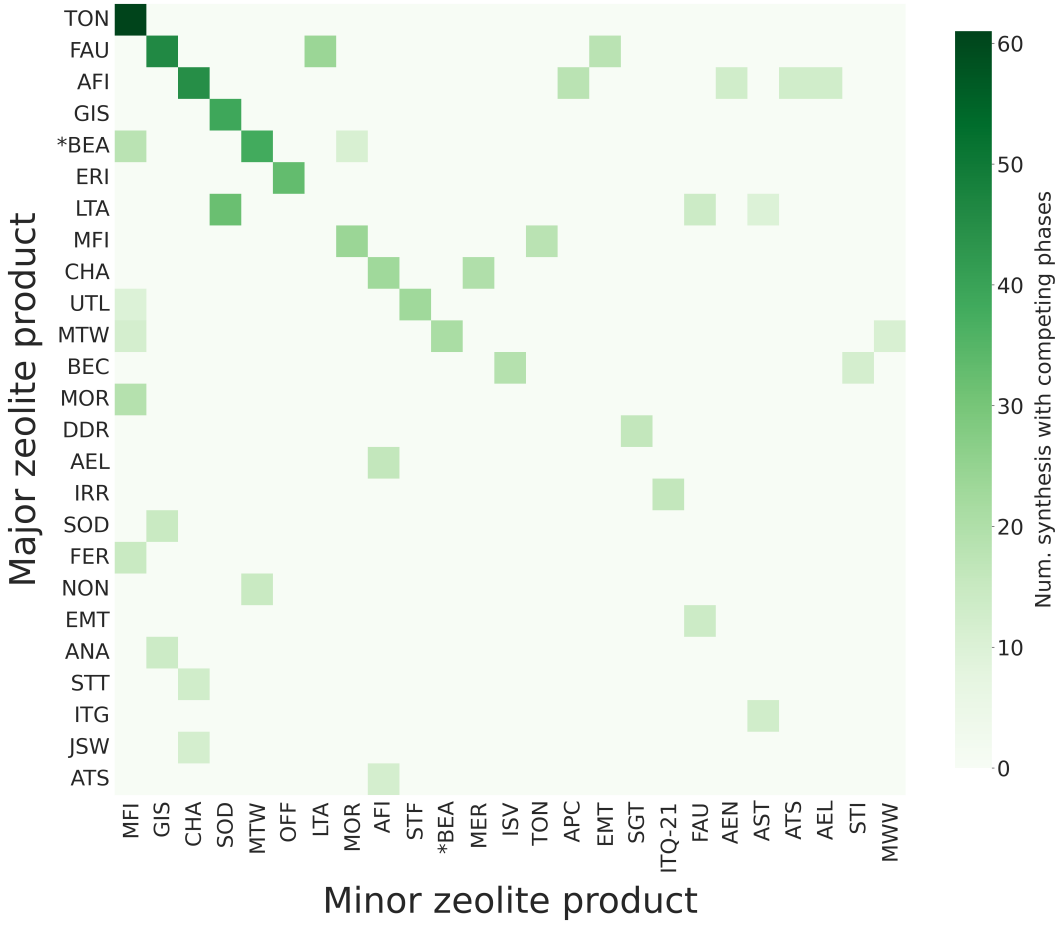

Figure 5: Number of synthesis routes with competing phases. y-axis is the major zeolite product, x-axis is the minor zeolite product.

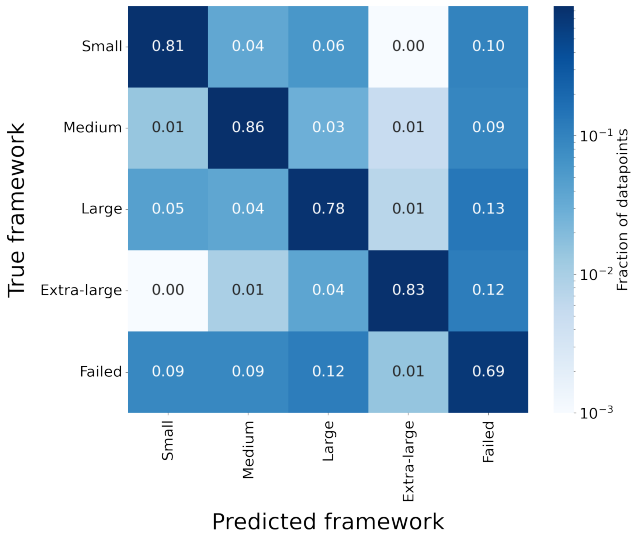

Figure 6: Normalized confusion matrix of phase predictor model. Here, we aggregate frameworks according to small, medium, large and extra-large pore frameworks. "Failed" refers to amorphous/dense phases.

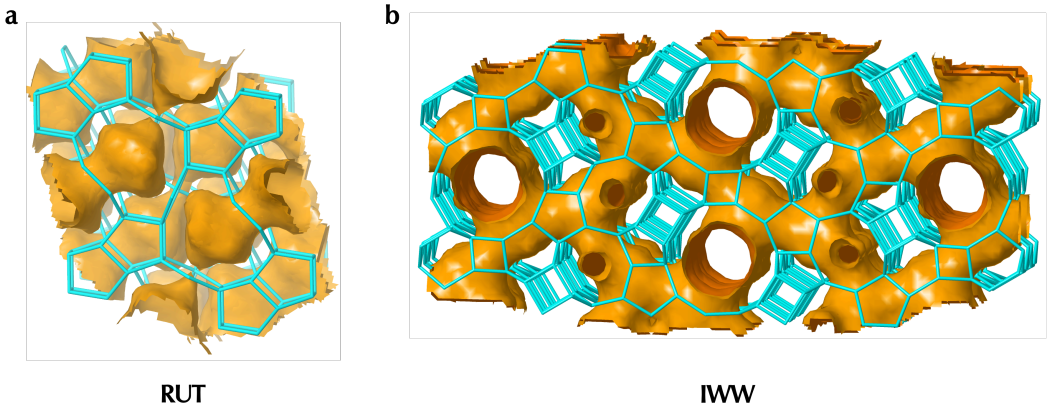

Figure 7: The **(a)** spherical cavities in **RUT (b)** long channels in **IWW** explain the difference in OSDA spherocity favoring the two frameworks: **IWW** is favored by spherical OSDAs while **RUT** is favored by longer OSDAs.

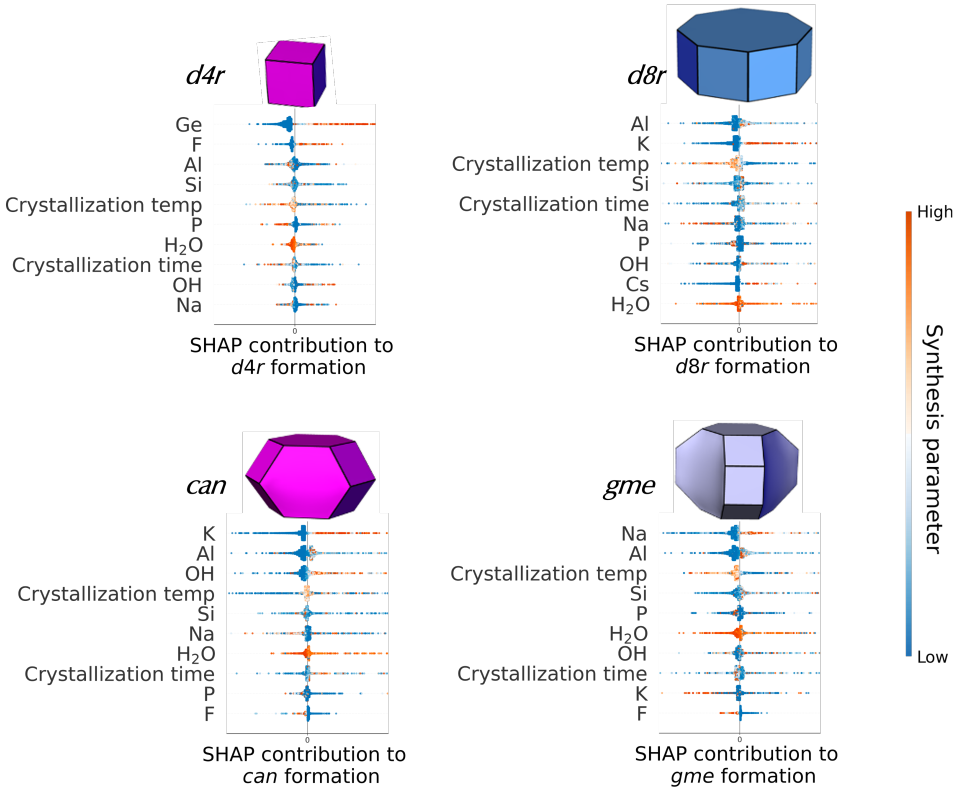

Figure 8: CBU-level SHAP analysis of small CBUs showing top 10 most important inorganic parameters (y-axis) contributing to their formation.

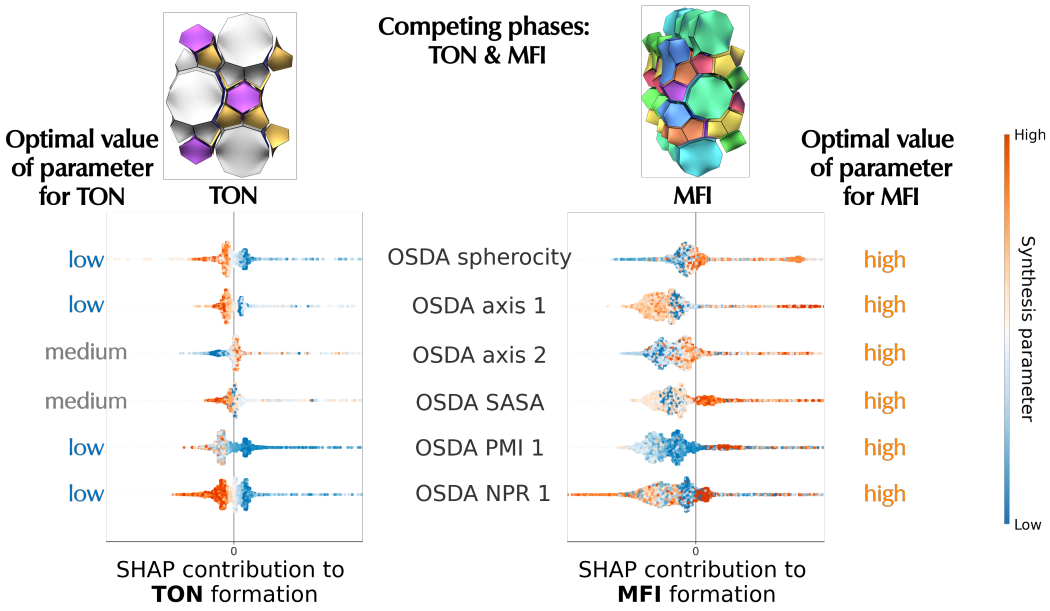

Figure 9: Application of framework-level SHAP on competing phases (**TON** and **MFI**). The left- and right-most columns describe the optimal value of OSDA parameter for maximizing formation probability of **TON** and **MFI**, respectively. For example, the first row shows *opposing* effects of OSDA spherocity: High OSDA spherocity promotes **MFI** formation while suppressing **TON** (and vice versa).