Electronic Supplementary Material (ESI) for Digital Discovery.

Supporting Information for: Inorganic synthesis-structure maps in zeolites with machine learning and crystallographic distances

Daniel Schwalbe-Koda*,†, Daniel E. Widdowson,‡ Tuan Anh Pham,† and Vitaliy Kurlin*,‡

†*Lawrence Livermore National Laboratory, Livermore, CA, United States*

‡*University of Liverpool, Liverpool, United Kingdom*

E-mail: dskoda@llnl.gov; vitaliy.kurlin@gmail.com

Supporting Text

Outlier zeolite pairs in the minimum spanning tree

Figure 2 of the main text illustrates how a distance matrix between zeolites can be visualized in a discrete map using a minimum spanning tree. Although many methods would be possible to perform this analysis, we decided to adopt this visualization to avoid a multitude of neighbors per zeolite or approaching limits where all frameworks are strongly connected. As the minimum spanning tree minimizes the total length of the edges, it is prone to connecting only nearest-neighbor structures, even when several other structures have nearly the same distance up to a small threshold (say, 10^{-3} Å). This can lead to differences in interpretation of the results for the tree. While the map is not intended to be a definitive clustering of zeolites, its interpretation has proven useful to connect framework pairs that could have

been otherwise overlooked. In addition to the **SOD-LTA** pair described in the main text, other pairs that are intuitively regarded as similar in the zeolite community appear distant in Fig. 2 of the main text. Some of these outliers, along with possible explanations for their perceived distance according to the AMD, are:

- **OFF-LTL**: both zeolites are related by *d6r* and *can* building units, as well as *dsc* chains, exhibiting reasonably similar densities (16.1 and 16.7 T/1000 Å^3 for **OFF** and **LTL**, respectively), ring sizes (12, 8, 6, 4), and channel dimensionality (3D). Despite this similarity, the AMD distance places **OFF** closer to **SWY** — a zeolite that shares exactly the same density (16.1 T/1000 Å^3), building units (*d6r*, *can*, and *gme*), and channel dimensionality (3D) — and to **ERI**, a zeolite with equally similar building units (*d6r* and *can*), same density, and with which it is known to form intergrowths. Analogously, **LTL** is a nearest neighbor of **MOZ**, a zeolite that shares all of its CBUs with **LTL**, including the larger *ltl* CBU, and has similar density. Furthermore, ZSM-10 can also be synthesized as an aluminosilicate in the presence of potassium, similar to the Linde type L structure.

- **GME/AFI** are structurally related, with **GME** sometimes undergoing a reconstructive transformation towards AFI. Although this relationship is captured by graph similarity metrics,[41] graph distances do not necessarily correlate to structural distances (Fig. S1), and may indicate different phenomena. According to the AMD metric, the closest neighbors of zeolite **GME** are **SFW**, **AFX**, and **AFV**, which are known to be related to **GME** and sometimes form integrowths. On the other hand, the closest neighbors to **AFI** are reasonably far in the AMD metric space (0.18 Å), with its closest nearest neighbor being **TON**. While both are zeolites known for their 1D channel, this link cannot be immediately rationalized based on the AMD distance.

- Zeolites **UTL/PCR/OKO** exhibit known structural similarity given the known disassembly of **UTL** towards **PCR** and **OKO** using the ADOR method[54] or inverse sigma

transformation.[73] The product structures, **PCR** and **OKO**, are closely related in the AMD space: they are 9th nearest neighbor, with the 3rd-10th nearest neighbors to **OKO** sharing very similar distances as **PCR**. The tree map, therefore, does not convey the continuity of the distance, though it makes a convenient plot. The first nearest neighbor of **OKO** is the idealized ***PCS** structure, which is the IPC-6 structure that also results from the ADOR method. The most similar structure to **UTL** is the **EWS** framework, possibly due to the large channel intersection and the layered structure of the latter.

- In Fig. S6, the **GIU** zeolite was highlighted as the center of a cluster characterized by six-membered rings. However, the zeolites **MEP**, **DOH**, and **MTN** share five-membered rings which are more rare than the six-membered rings. At first, the connection between **GIU** and **MEP**, as seen from their ring distribution, is not obvious to rationalize. However, **GIU** and **MEP** are both clathrasil zeolites with similar densities. **GIU** is built by *sod* and *can* building units, and forms large cages, while **MEP** is built with *mtn* cages. However, the AMD distance between **GIU** and **MEP** is 0.18 Å. In comparison, **GIU** has 13 other neighbors within that distance (e.g., **FAR**, **AFG**, **CAN**, **LTN**, **LIO**, **VNI**), and **MEP** has two (**DOH** and **MTN**). Thus, the **MEP**-**GIU** edge can be explained by the procedure defining the minimum spanning tree, as this connection is the minimum pathway that connects the **MEP**-**DOH**-**MTN** cluster to the rest of the tree.

Supporting Figures

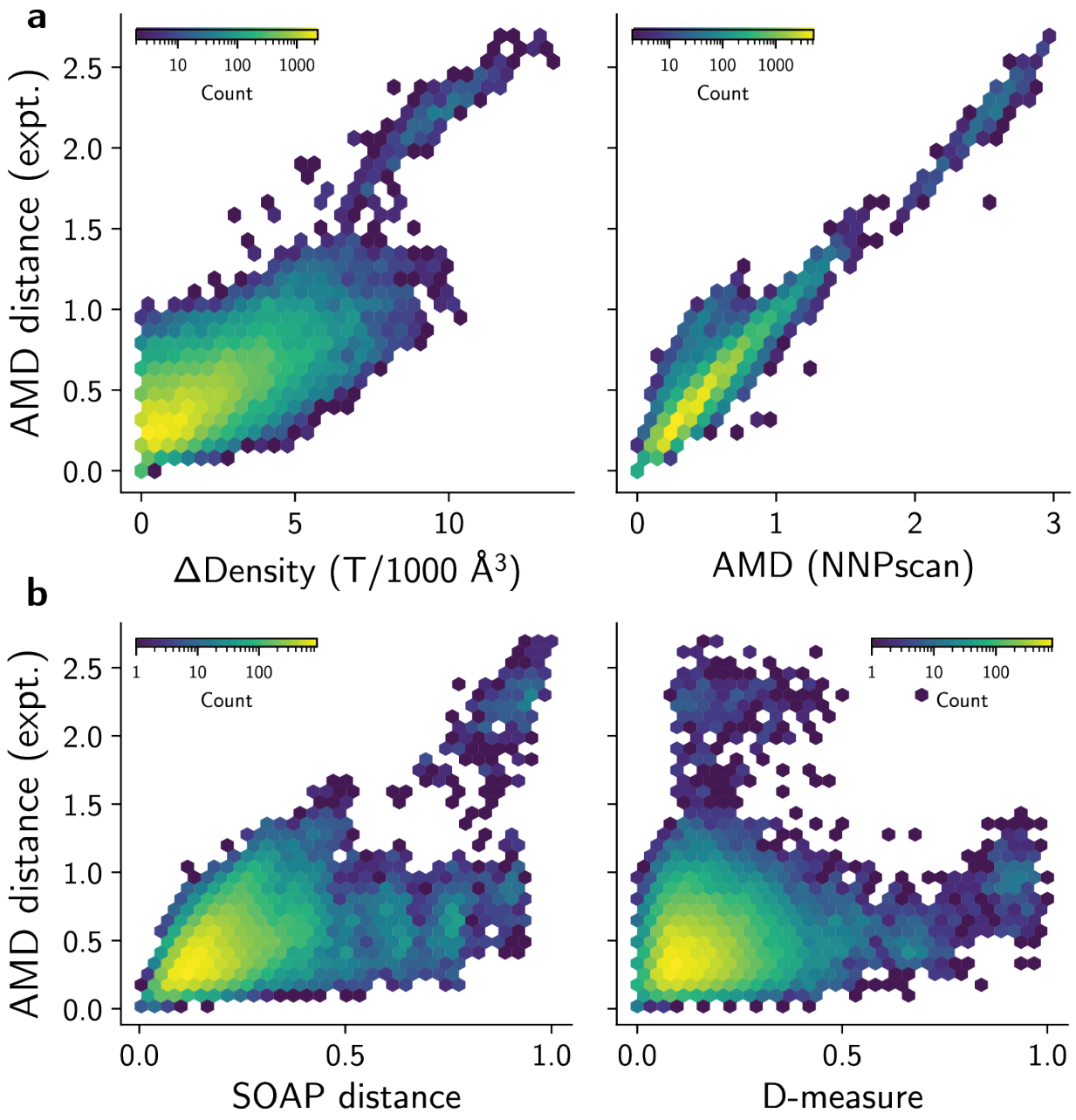

Figure S1: Correlations of AMD distance values between zeolites with experimental structural parameters from the IZA database and **a**, density differences, **b**, AMD distance values for zeolites optimized with NNPscan, **c**, SOAP distance, and **d**, the graph distance D-measure. The values from **c, d** were retrieved from Schwalbe-Koda et al. (Ref. 41 from the main text)

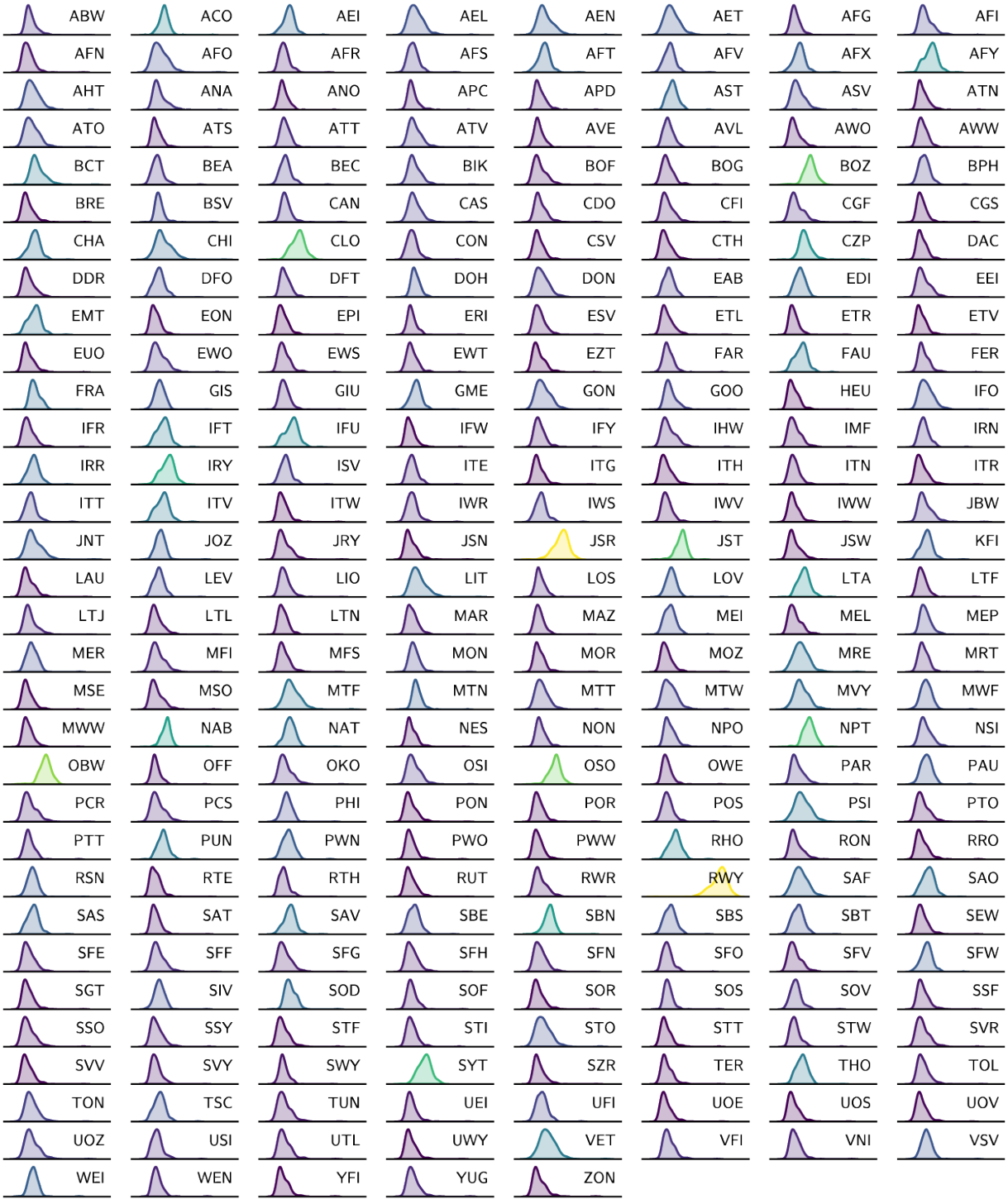

Figure S2: Distributions of AMD distances for each zeolite against all other known zeolites. Brighter colors indicate a higher median. All subfigures share the same x axis.

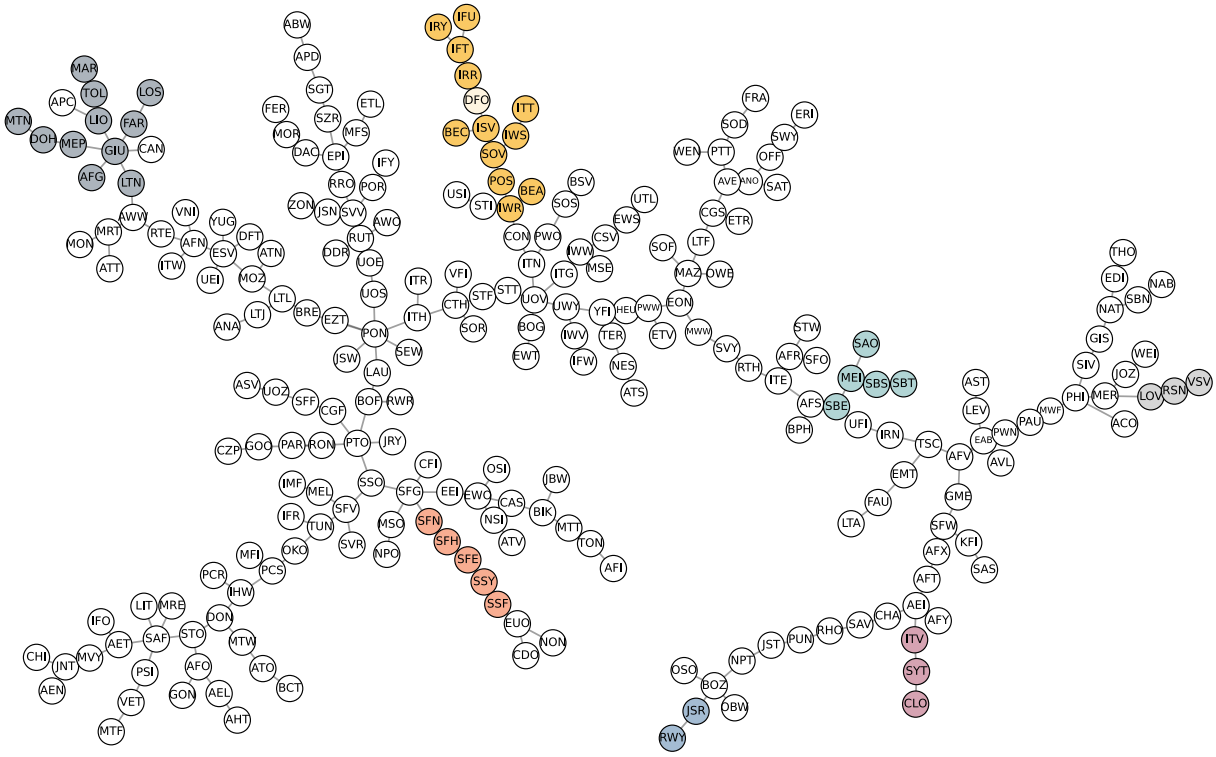

Figure S3: Minimum spanning tree of known zeolites with qualitatively chosen labels for a visual guide. The tree was created using the AMD distances between zeolites. Although several clusters could be highlighted, only some of those discussed in the main text are shown here. Within the cluster centered at **GIU**, only zeolites with zero accessible area are colored, though **APC** and **CAN** may still be considered part of the same group of structures.

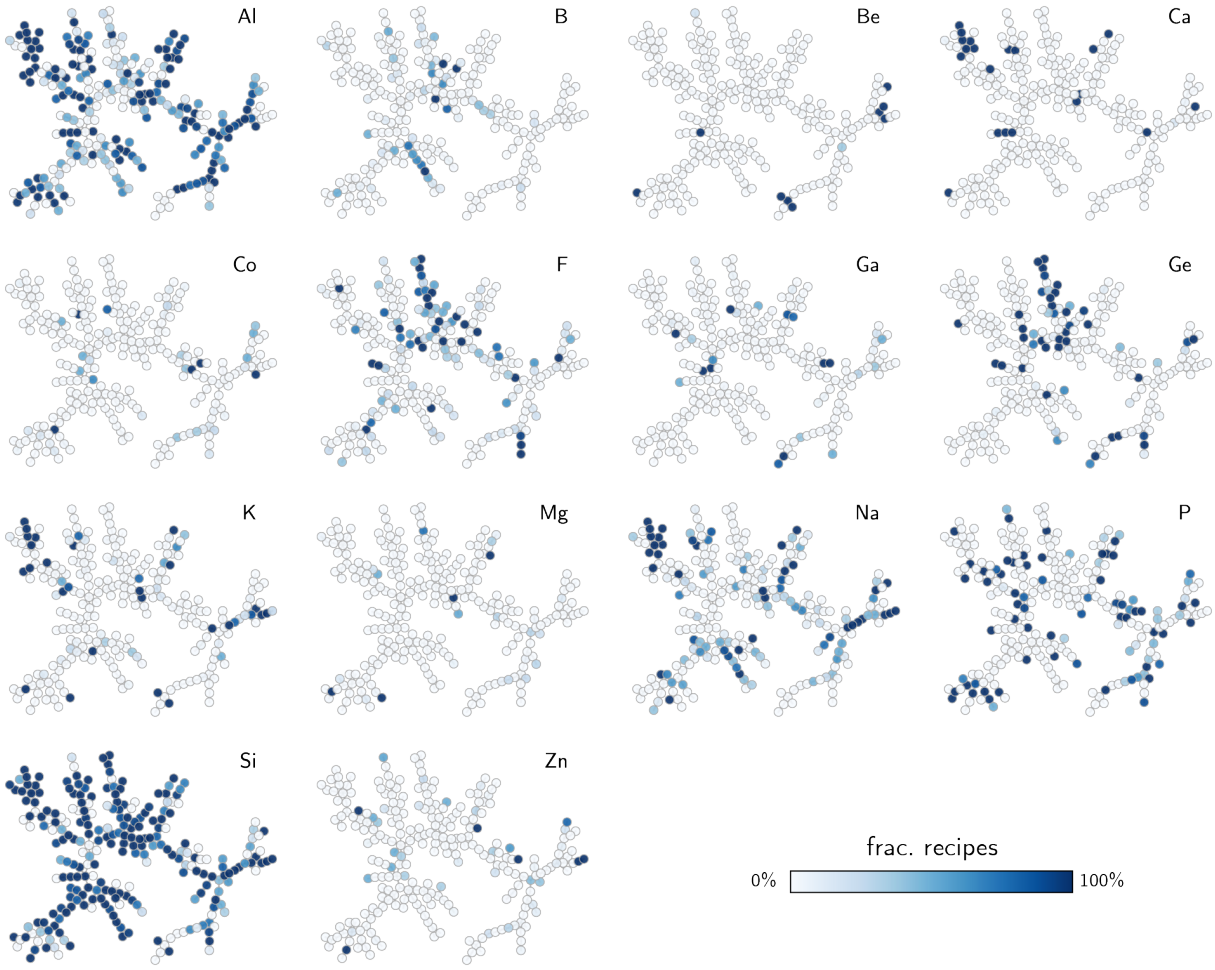

Figure S4: Minimum spanning tree of known zeolites labeled according to the fraction of recipes per element. The tree was created using the AMD distances between zeolites.

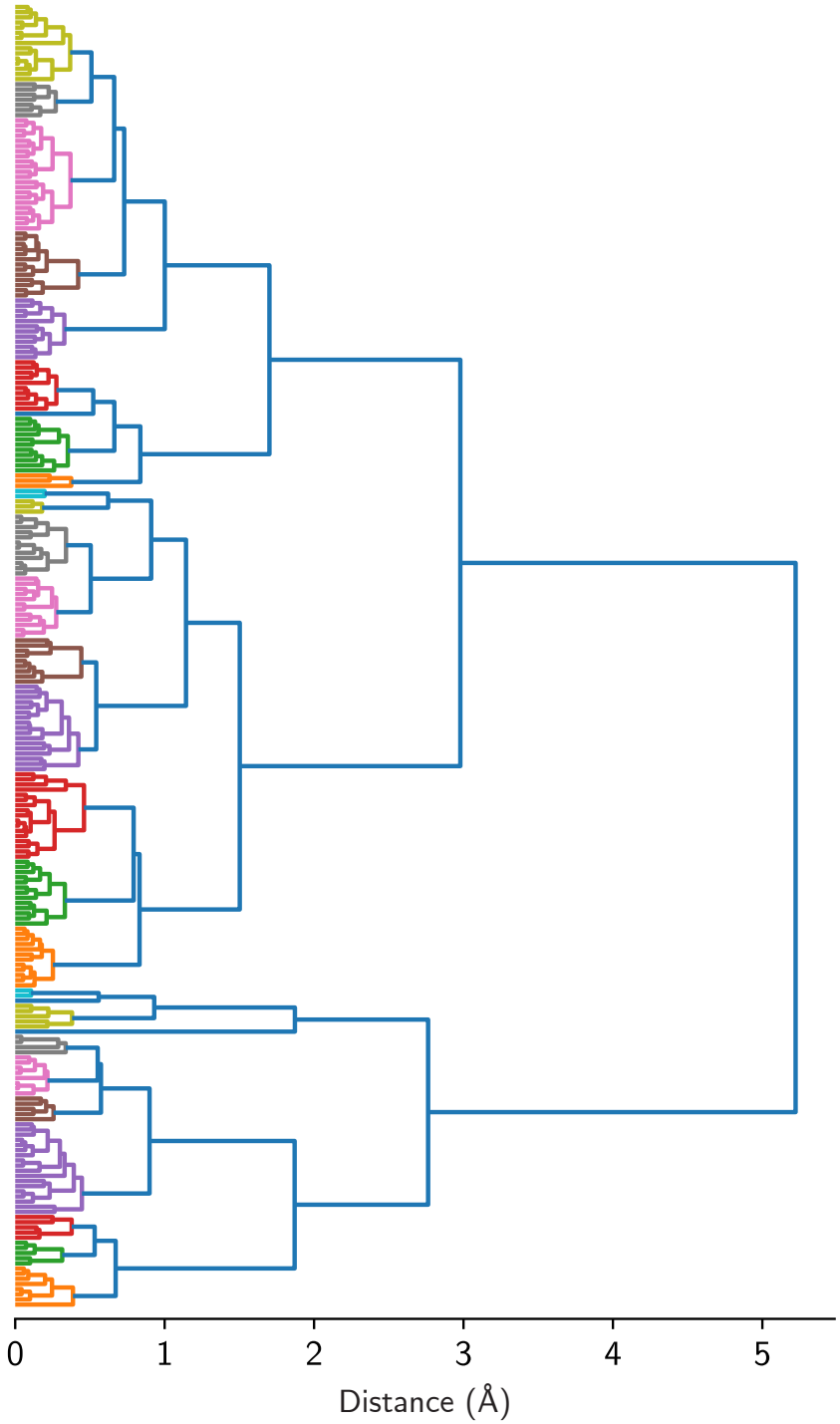

Figure S5: Complete dendrogram of zeolites created using the AMD values as features.

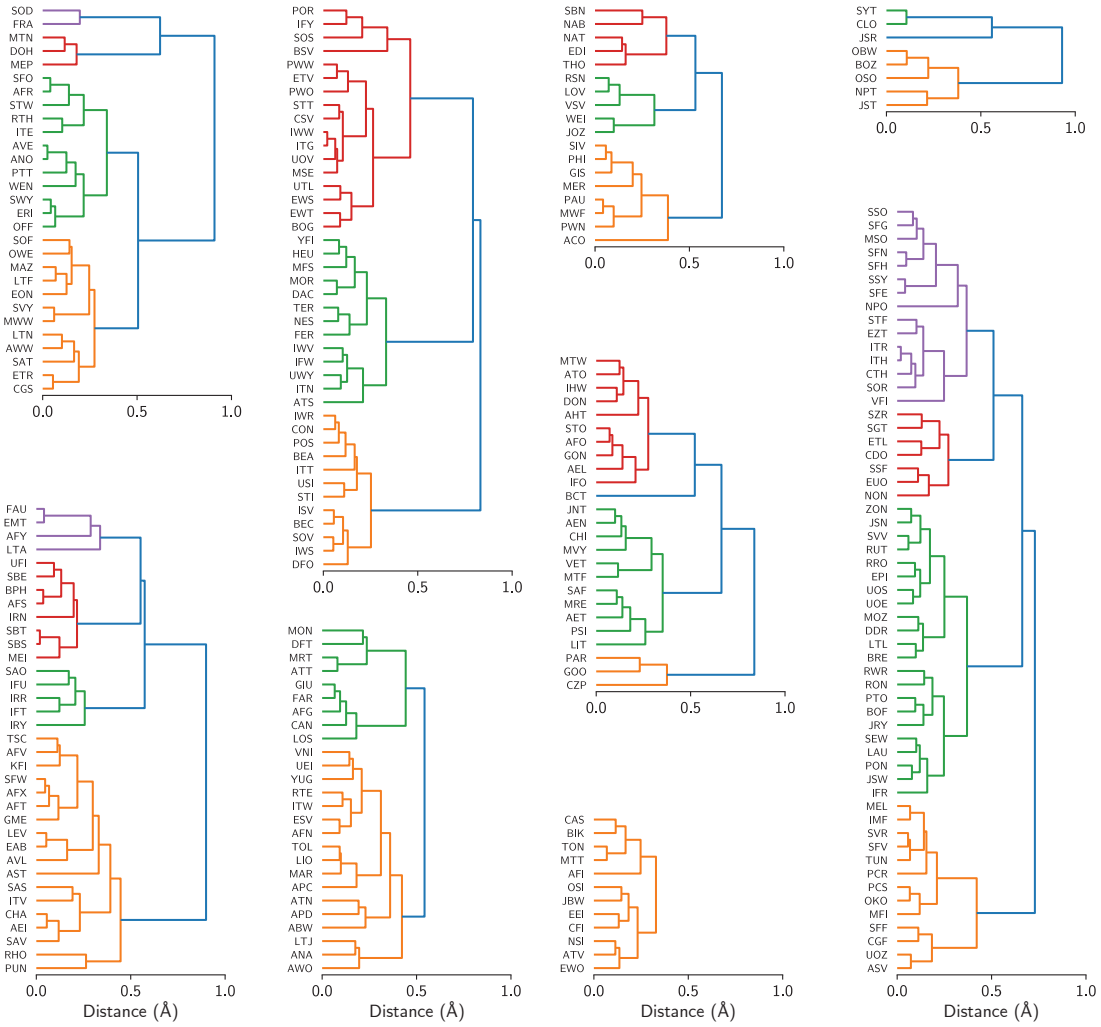

Figure S6: Subclusters of the full dendrogram in Fig. S5 with labels.

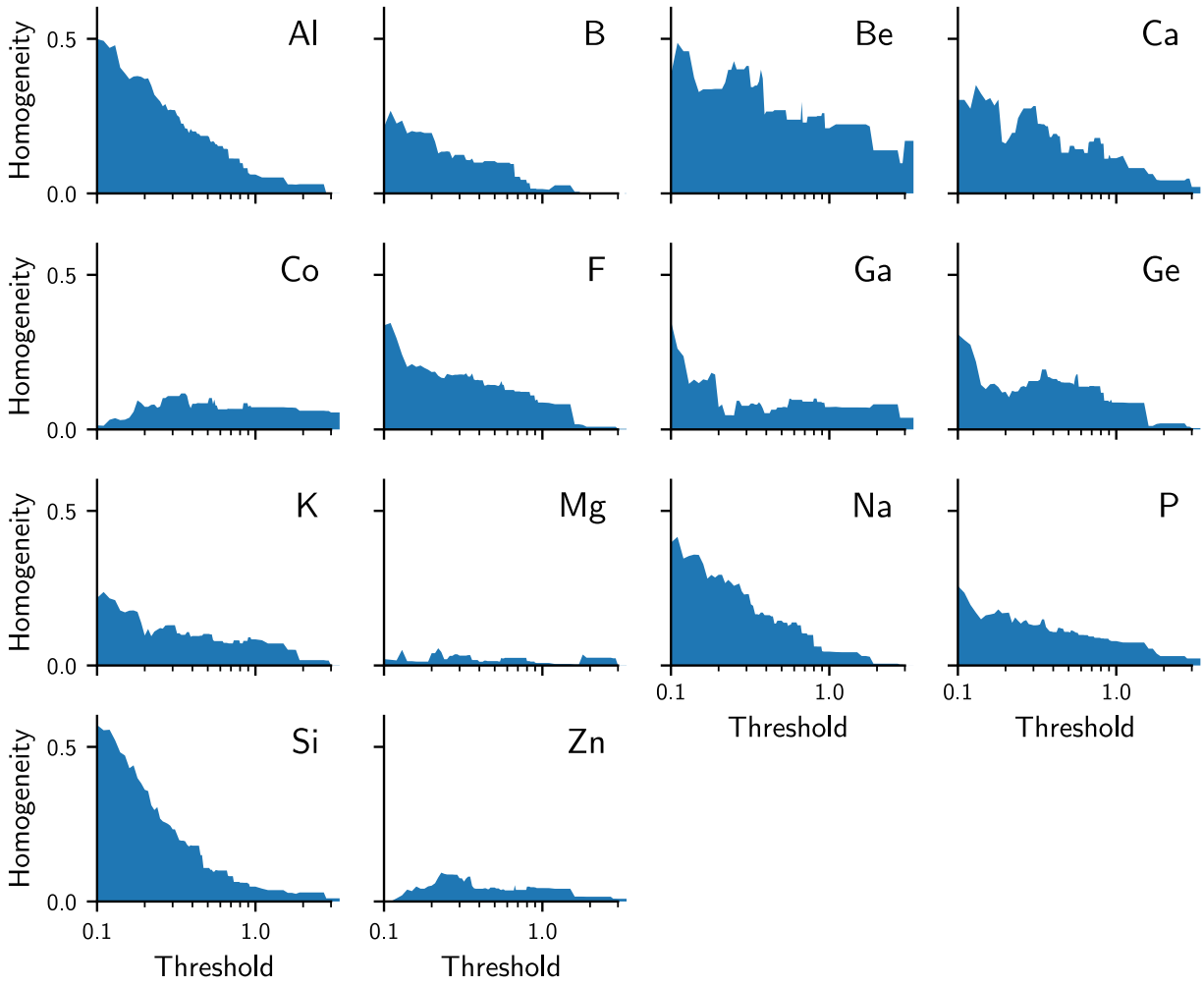

Figure S7: Clustering homogeneity for different synthesis conditions when flat clusters are created with the given threshold between AMD distances. A homogeneity of zero indicates a perfect mixing between positive and negative labels in the same clusters.

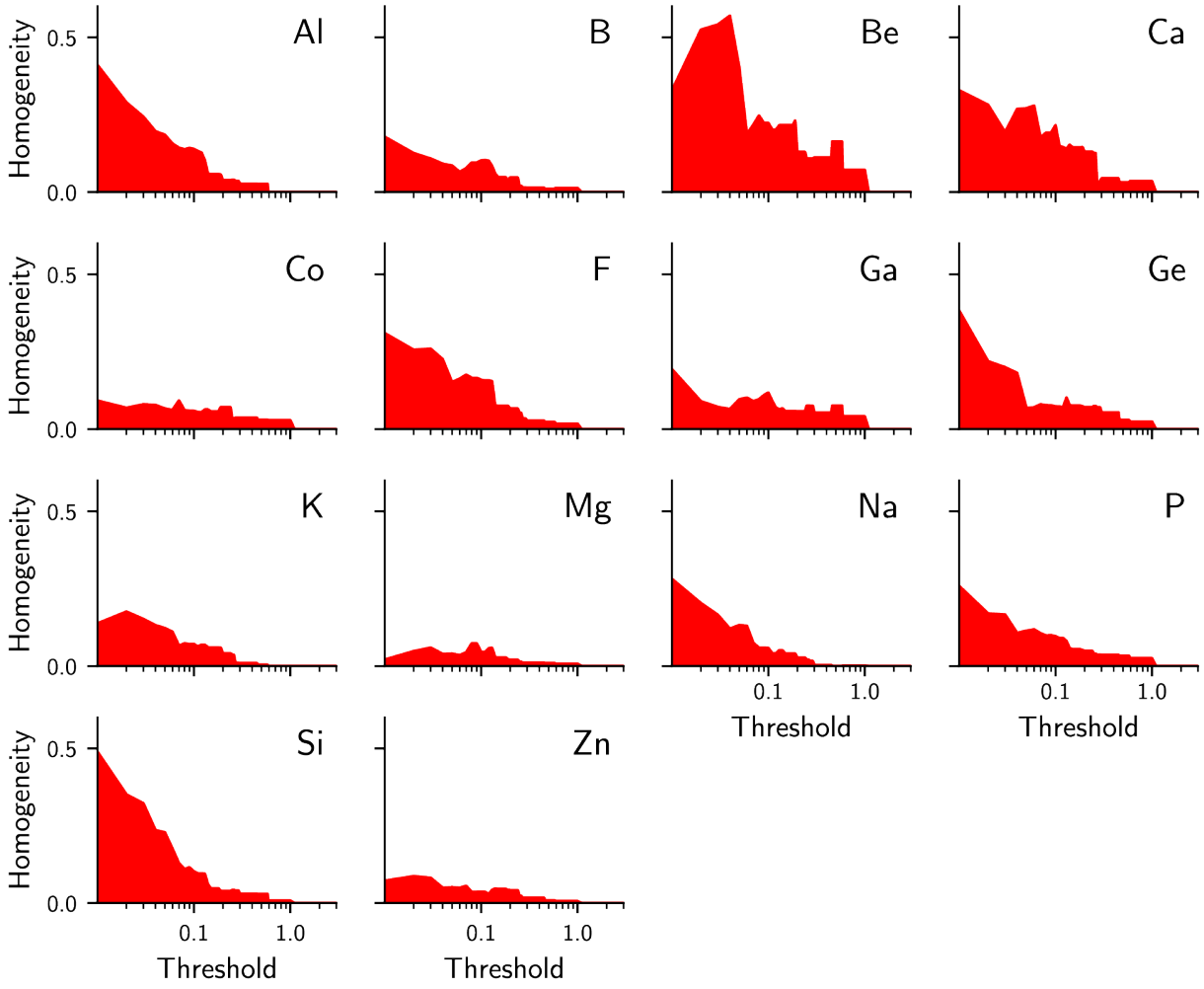

Figure S8: Clustering homogeneity for different synthesis conditions when flat clusters are created with the given threshold between SOAP distances. Because the distance metrics are different, the thresholds are not the same as the ones in Fig. S7 for the AMD vectors. A homogeneity of zero indicates a perfect mixing between positive and negative labels in the same clusters.

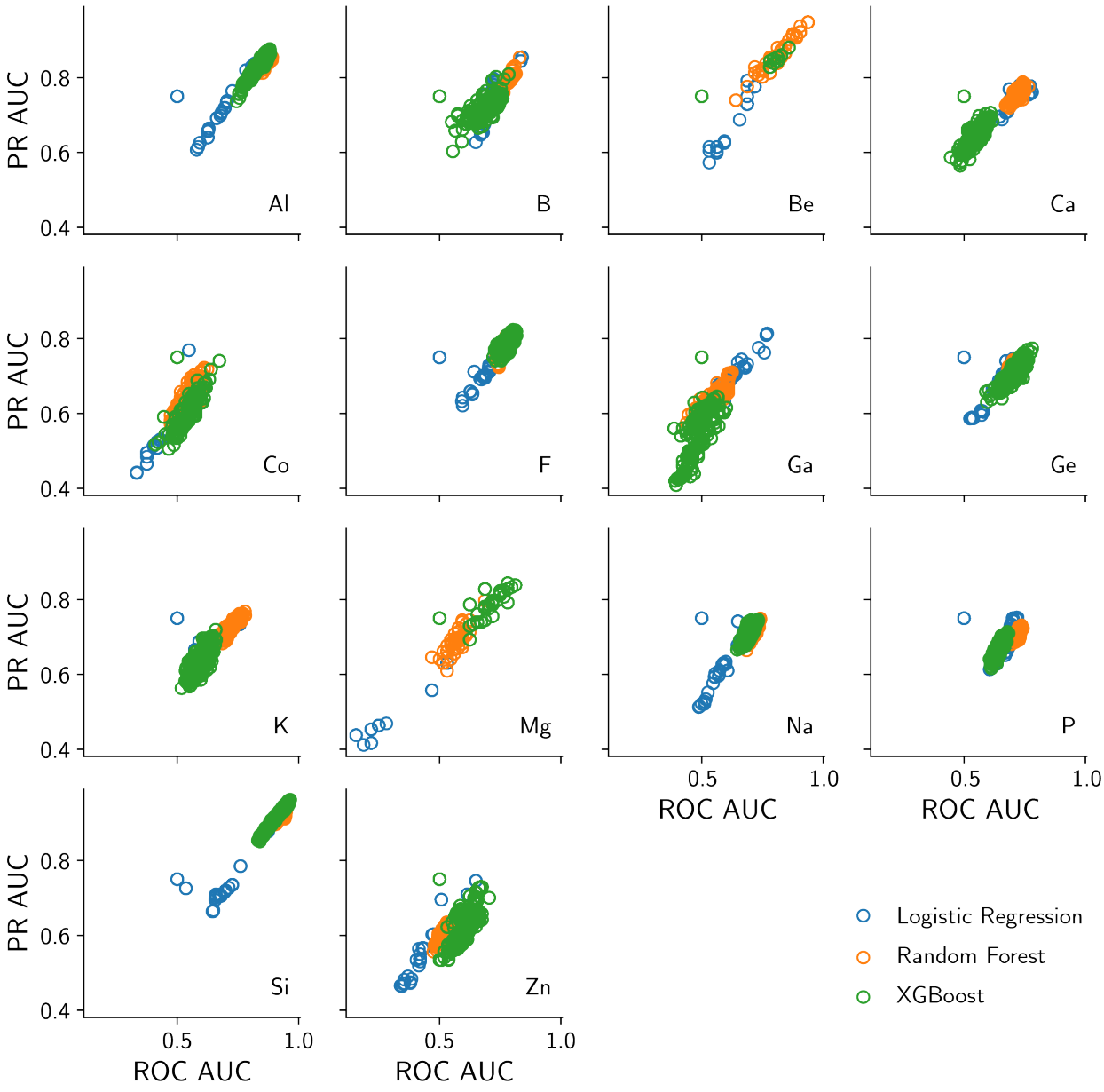

Figure S9: Precision-recall (PR) and Receiver Operating Characteristic (ROC) areas under the curve (AUC) for different hyperparameters of three different classifiers: logistic regression, random forest, and XGBoost. The classifiers were trained using the **AMD distances between IZA zeolites as features**. The PR AUC and ROC AUC are adopted as the main figures of merit for evaluating these classifiers.

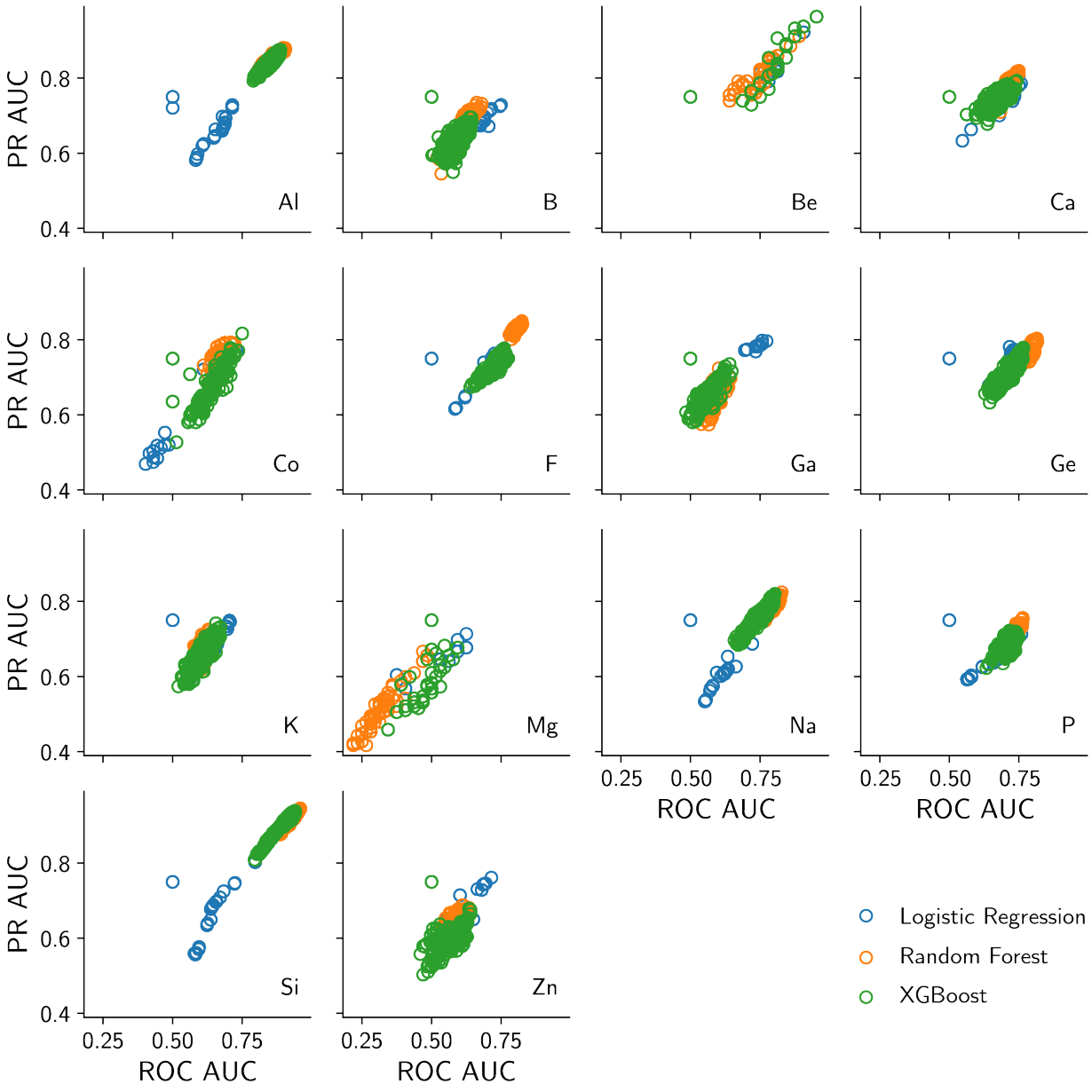

Figure S10: Precision-recall (PR) and Receiver Operating Characteristic (ROC) areas under the curve (AUC) for different hyperparameters of three different classifiers: logistic regression, random forest, and XGBoost. The classifiers were trained using the **SOAP distances between IZA zeolites as features**. The PR AUC and ROC AUC are adopted as the main figures of merit for evaluating these classifiers.

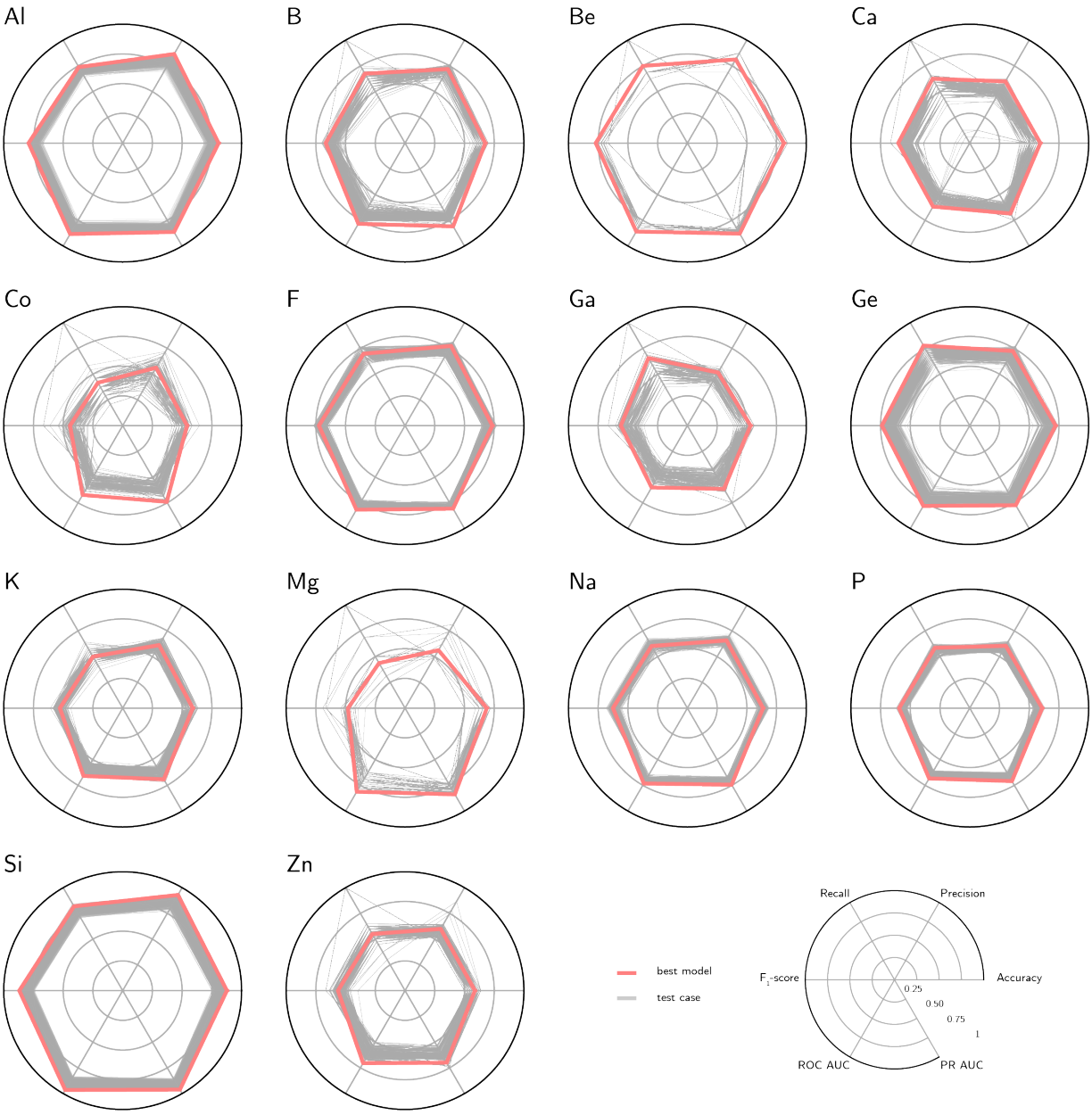

Figure S11: All figures of merit for different hyperparameters of the XGBoost classifiers trained with AMD distances. Each gray line represents the average metric of five runs at each set of hyperparameters. The figures of merit of interest, along with the values of each line in the circle, are shown in the lower right of the plot.

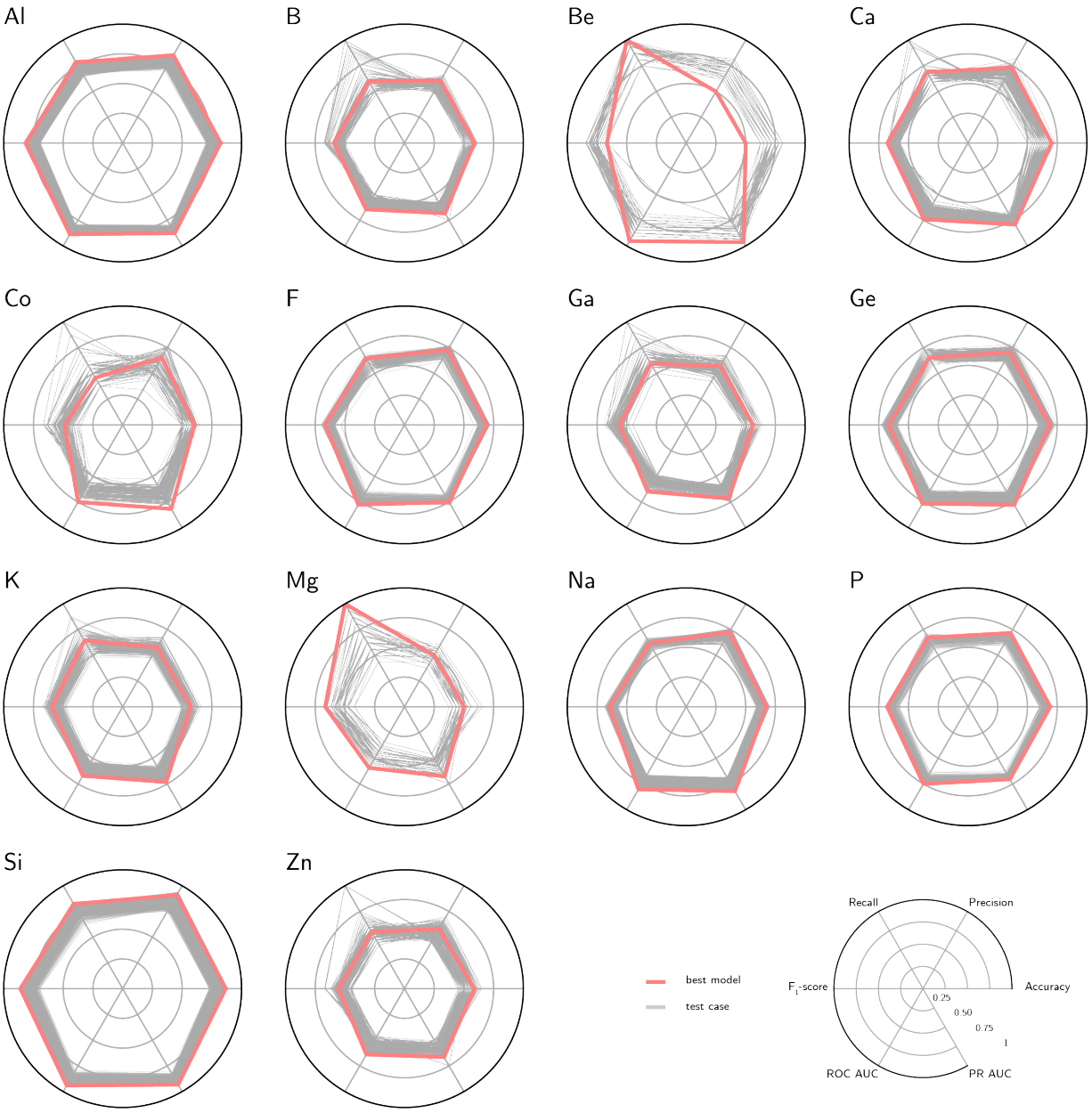

Figure S12: All figures of merit for different hyperparameters of the XGBoost classifiers trained with SOAP distances. Each gray line represents the average metric of five runs at each set of hyperparameters. The figures of merit of interest, along with the values of each line in the circle, are shown in the lower right of the plot.

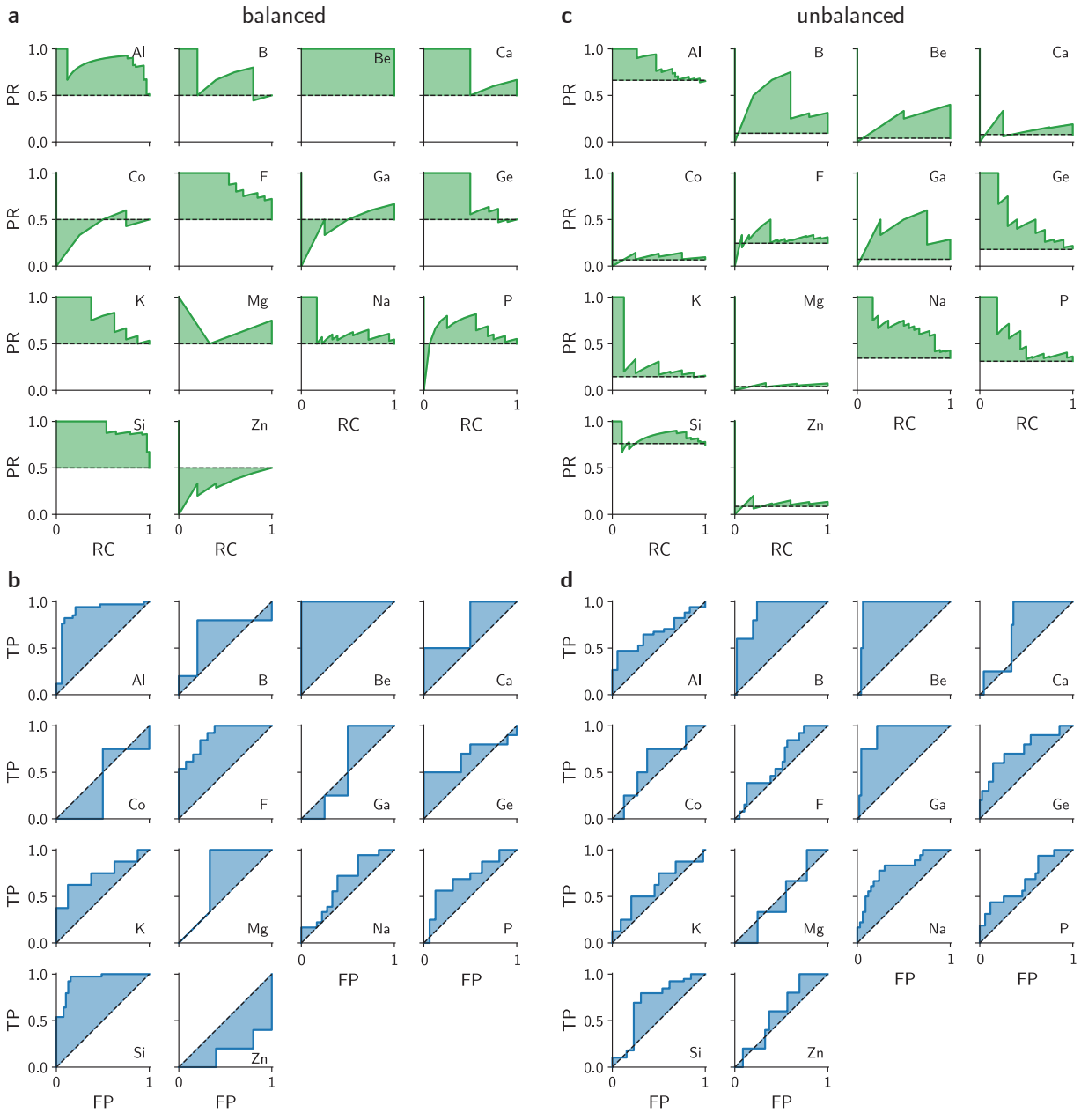

Figure S13: Precision (PR) and recall (RC) curves for the best **a**, balanced and **b**, unbalanced XGBoost classifier. The receiver operating characteristic curve, shown with the true positive (TP) and false positive (FP) ratios, are also depicted for a **c**, balanced and **d**, unbalanced XGBoost classifier. The best hyperparameters are selected according to the validation results. The curves in this figure are computed for a held-out test split.

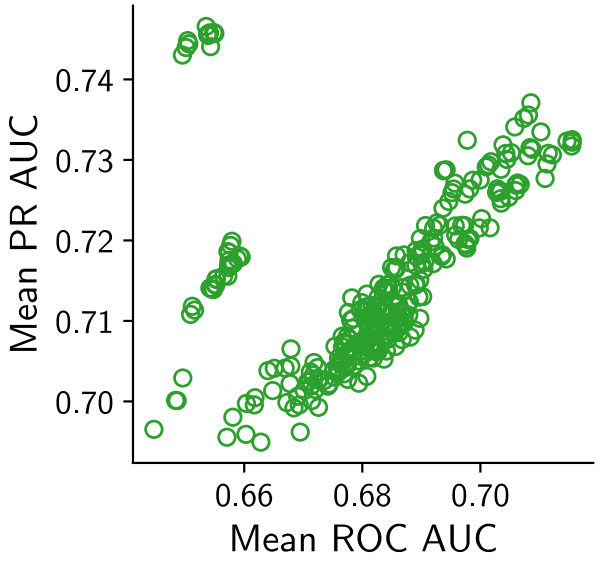

Figure S14: Mean area under the ROC and PR curves for the XGBoost classifiers trained with AMD distances. The figures of merit are computed with respect to all synthesis predictions at once. The best models maximize both the PR and ROC curves.

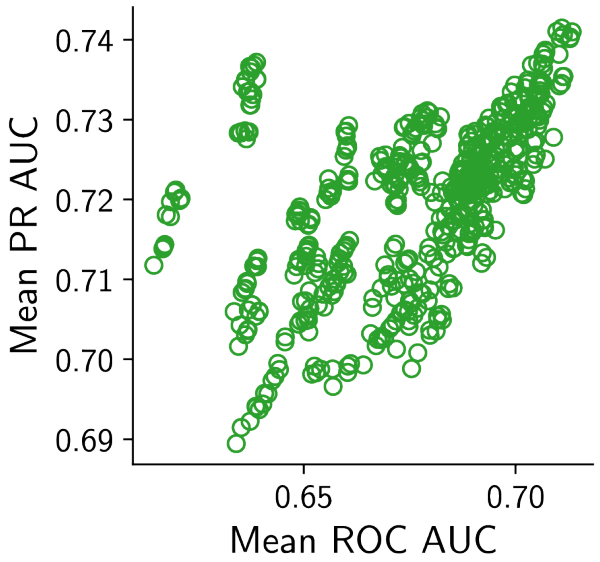

Figure S15: Mean area under the ROC and PR curves for the XGBoost classifiers trained with SOAP distances. The figures of merit are computed with respect to all synthesis predictions at once. The best models maximize both the PR and ROC curves.

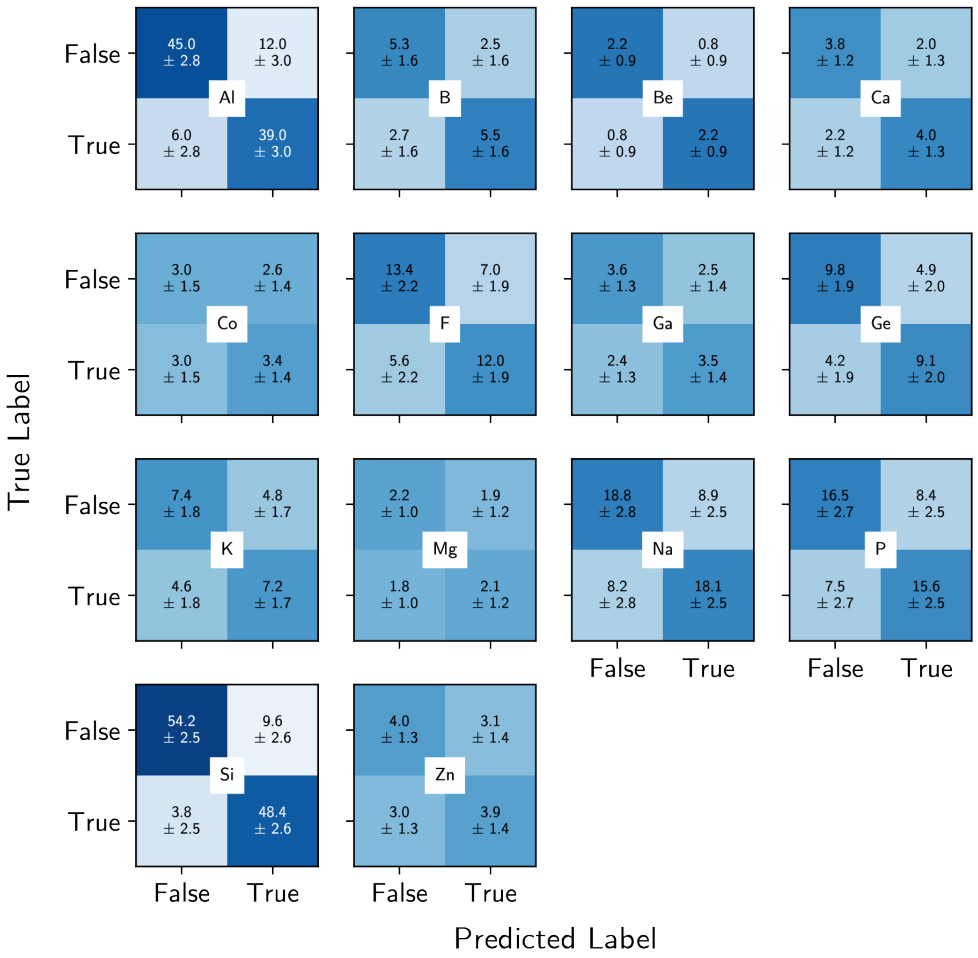

Figure S16: Confusion matrices for XGBoost classifiers trained with AMD distances. The confusion matrix was obtained by performing the prediction on held-out test sets. Each quadrant of the matrix shows the absolute number of structures in the test set with that predicted/true label. The reported error is the standard deviation of 100 runs. Colors offer a visual guide to the fraction of elements in each quadrant, with darker colors indicating higher fraction of elements.

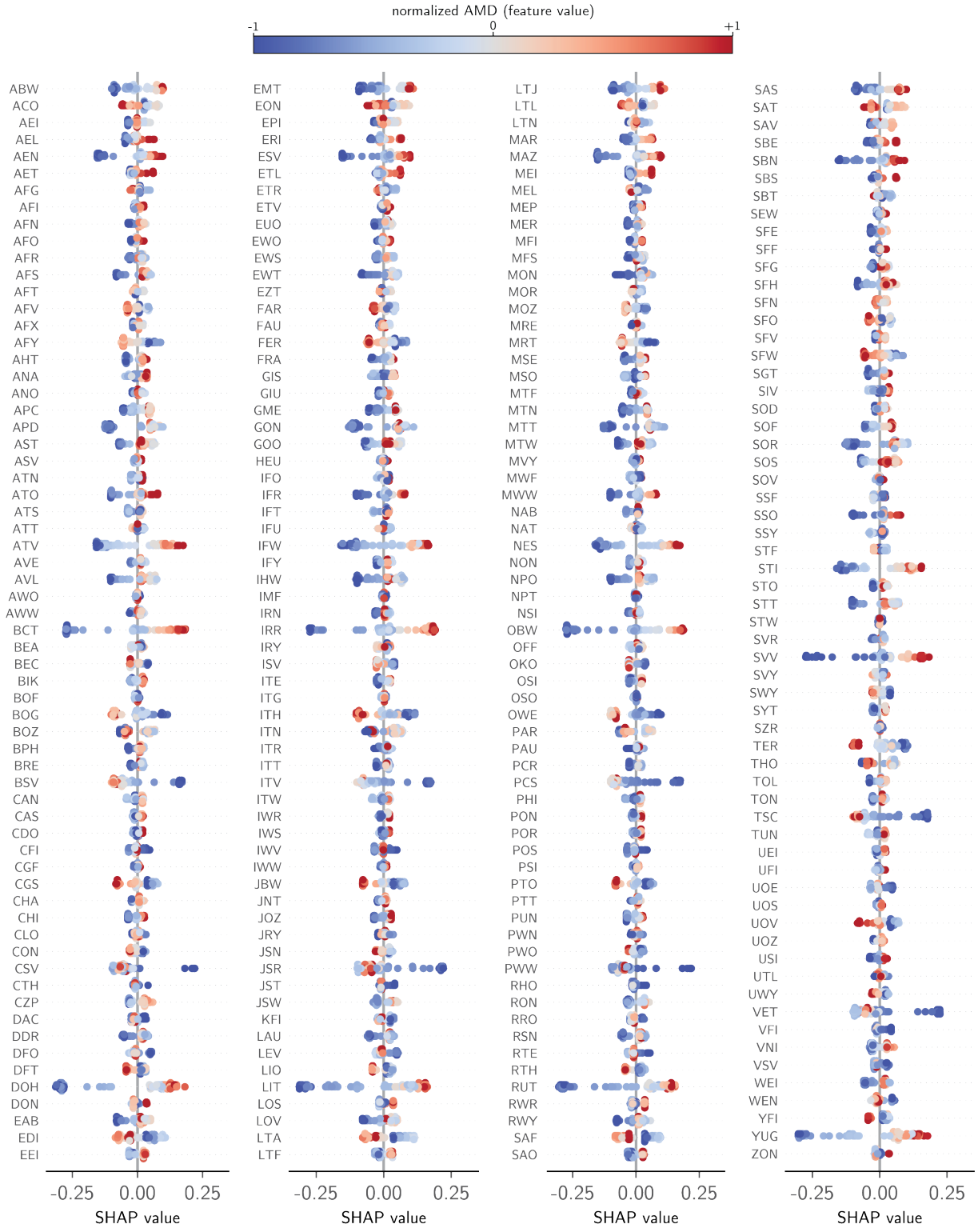

Figure S17: Visualization of the SHAP values for a single XGBoost classifier predicting whether Ge should be used in the synthesis of known zeolites. The classifier was trained with the AMD distance matrix. To plot the results in this graph, each feature (AMD distance to a given zeolite) was normalized in a per-feature basis. Larger SHAP values indicate higher impact in a positive classification.

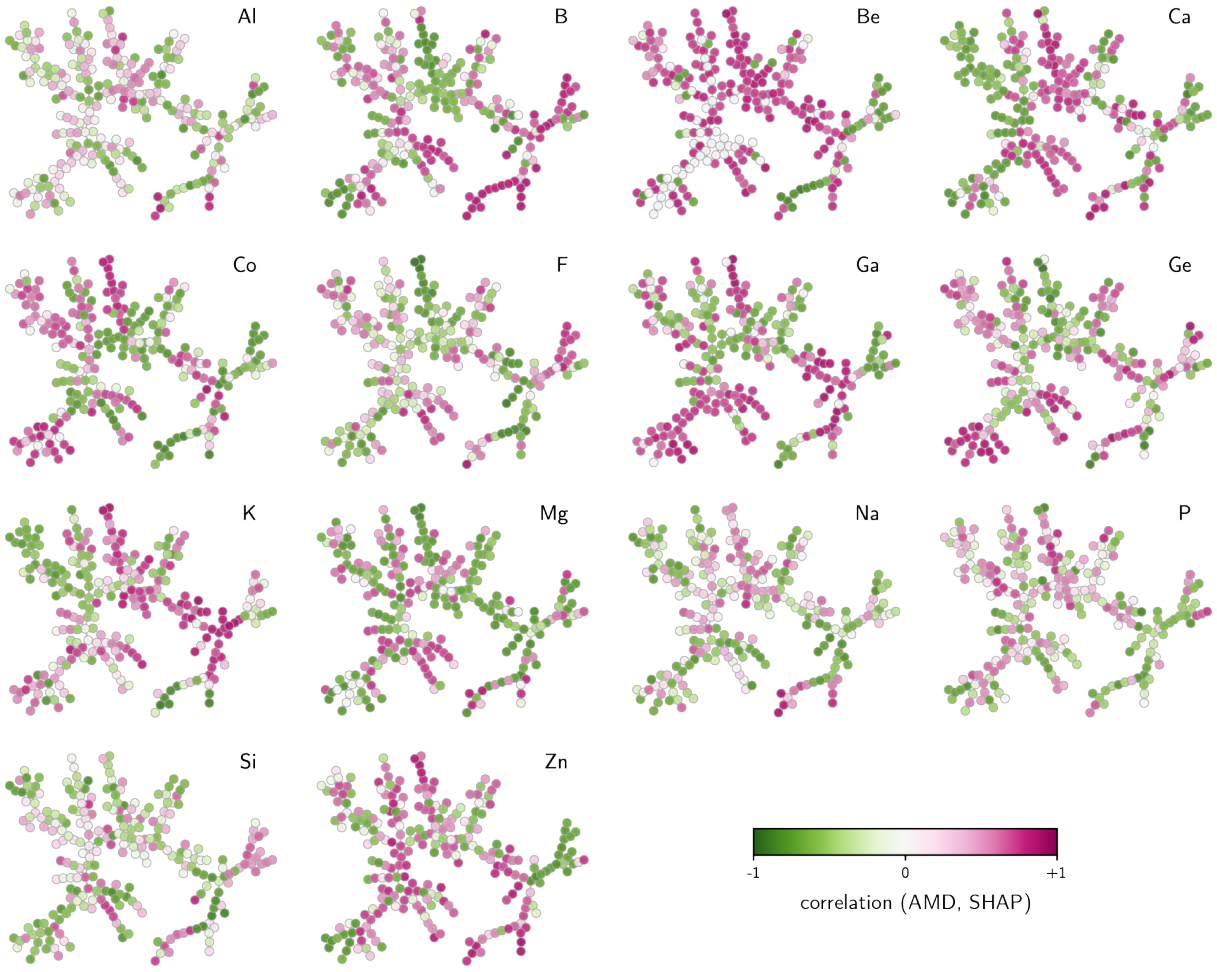

Figure S18: Zeolite tree map labeled with the Pearson correlation coefficient between AMD and SHAP values per synthesis conditions. The tree map was created with the AMD distance matrix. A negative correlation (shown in green) indicates that smaller AMD distances (i.e., high similarity) lead to higher SHAP values (i.e., higher likelihood of a positive classification). The correlation coefficient is the average correlation of AMD and SHAP values for 100 XGBoost models for each synthesis condition.

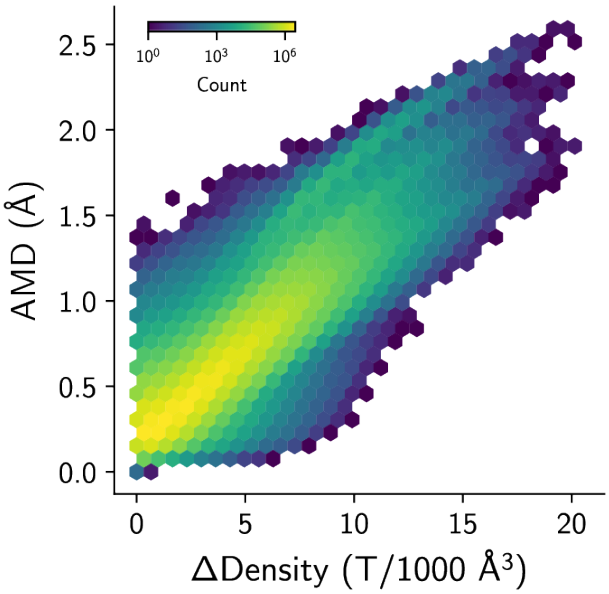

Figure S19: Relationship between AMD distances computed between known and hypothetical frameworks, and their density difference. Both known and hypothetical zeolites were optimized with the NNPscan method by Erlenbach et al.

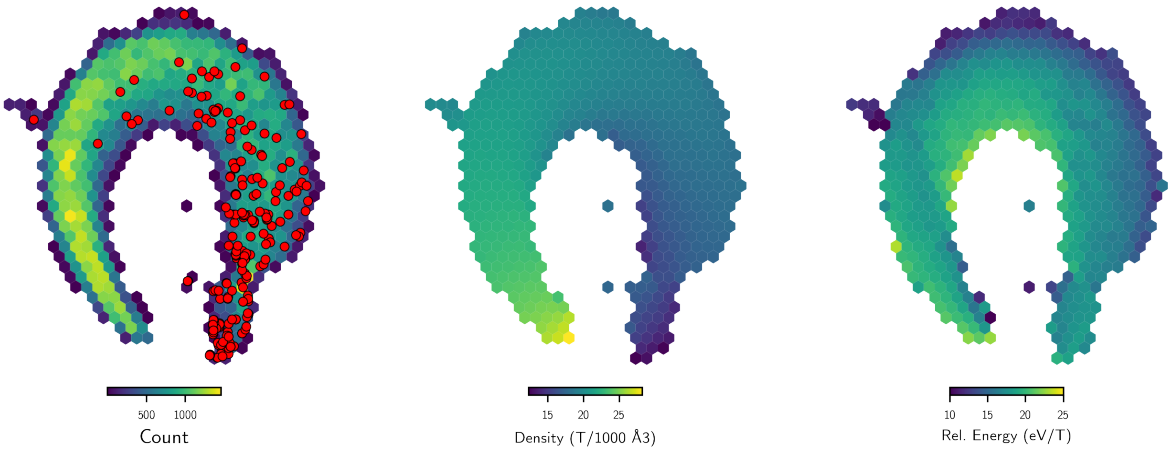

Figure S20: Low-dimensional projection of the hypothetical zeolite space using their AMD distance towards known zeolites as features. The red dots indicate zeolites present in the IZA database. The low-dimensional plot was obtained using UMAP (see Methods), and recovers the density and energy of the frameworks simply by comparing them against known structures.

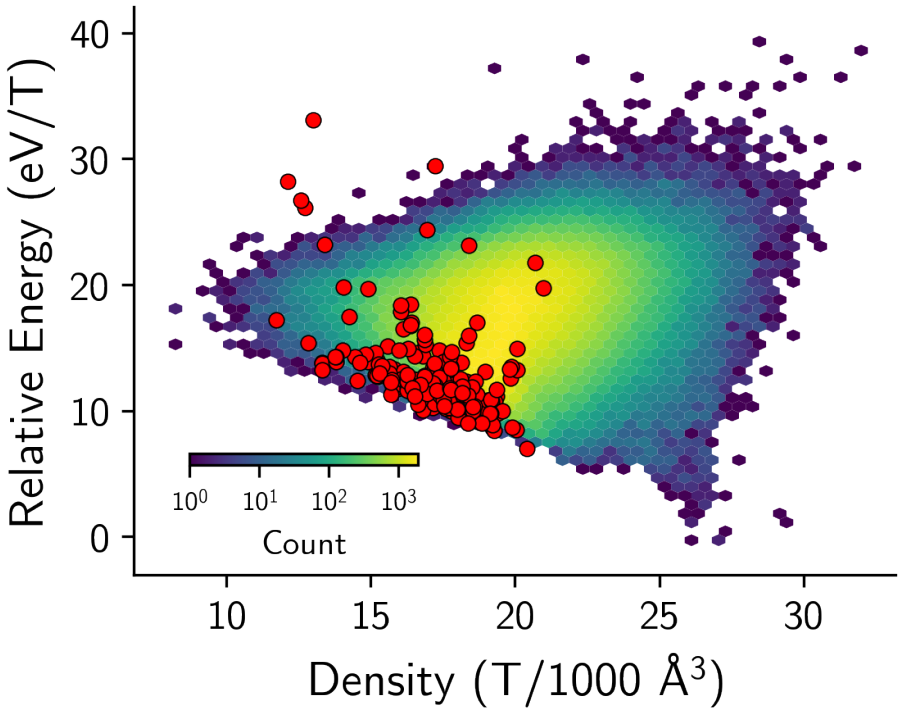

Figure S21: Energy-density plots for zeolites, reproducing the plot from Erlenbach et al. The analysis of the data using these two variables shows a high concentration of mid-energy, mid-density zeolites in the hypothetical structures database.

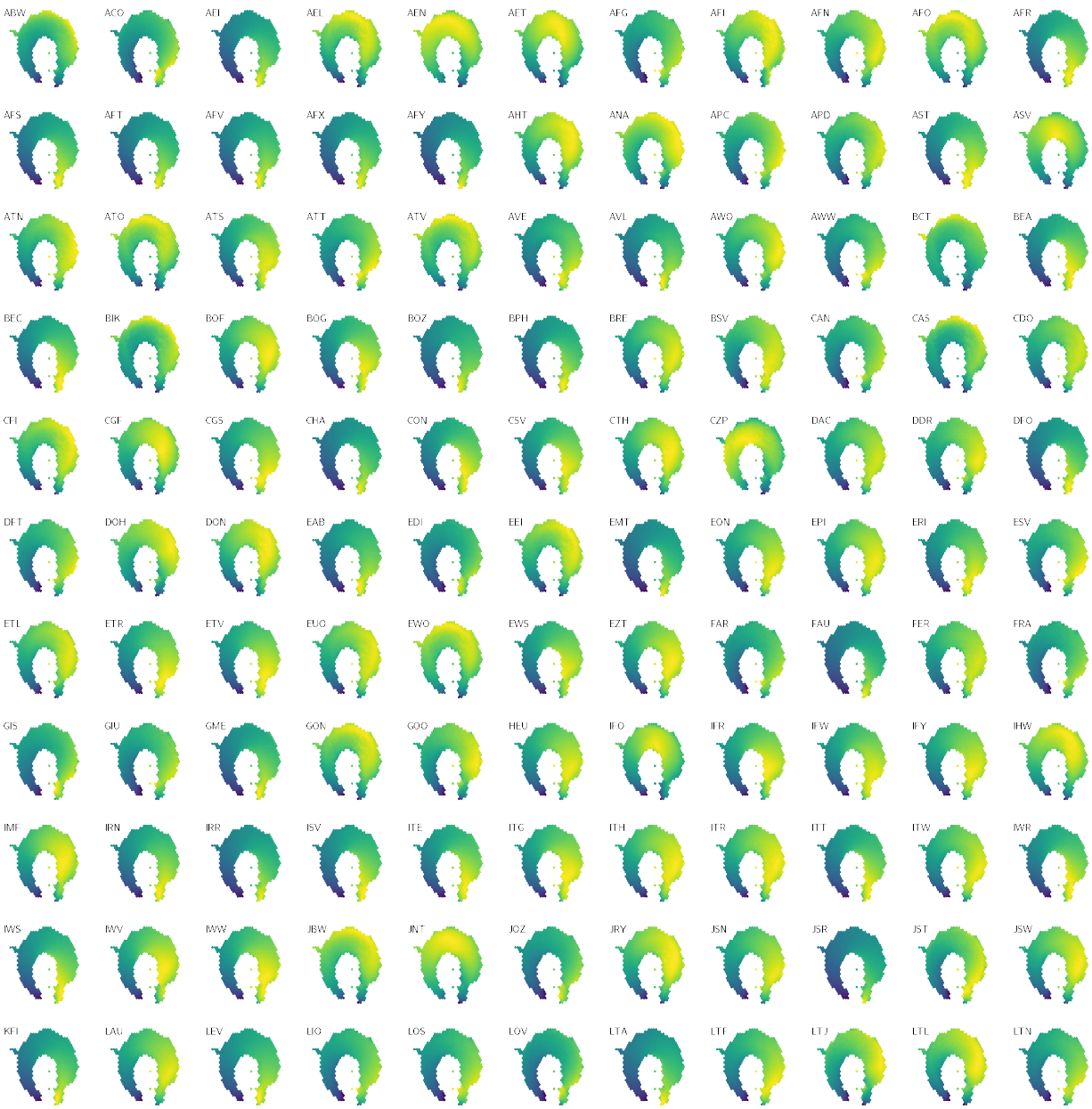

Figure S22: Average AMD distance of hypothetical zeolites towards known zeolites, visualized using the UMAP plot from Fig. S20. Brighter colors indicate lower distances. (continues in Fig. S23).

Figure S23: Average AMD distance of hypothetical zeolites towards known zeolites, visualized using the UMAP plot from Fig. S20. Brighter colors indicate lower distances. (continued from Fig. S22).

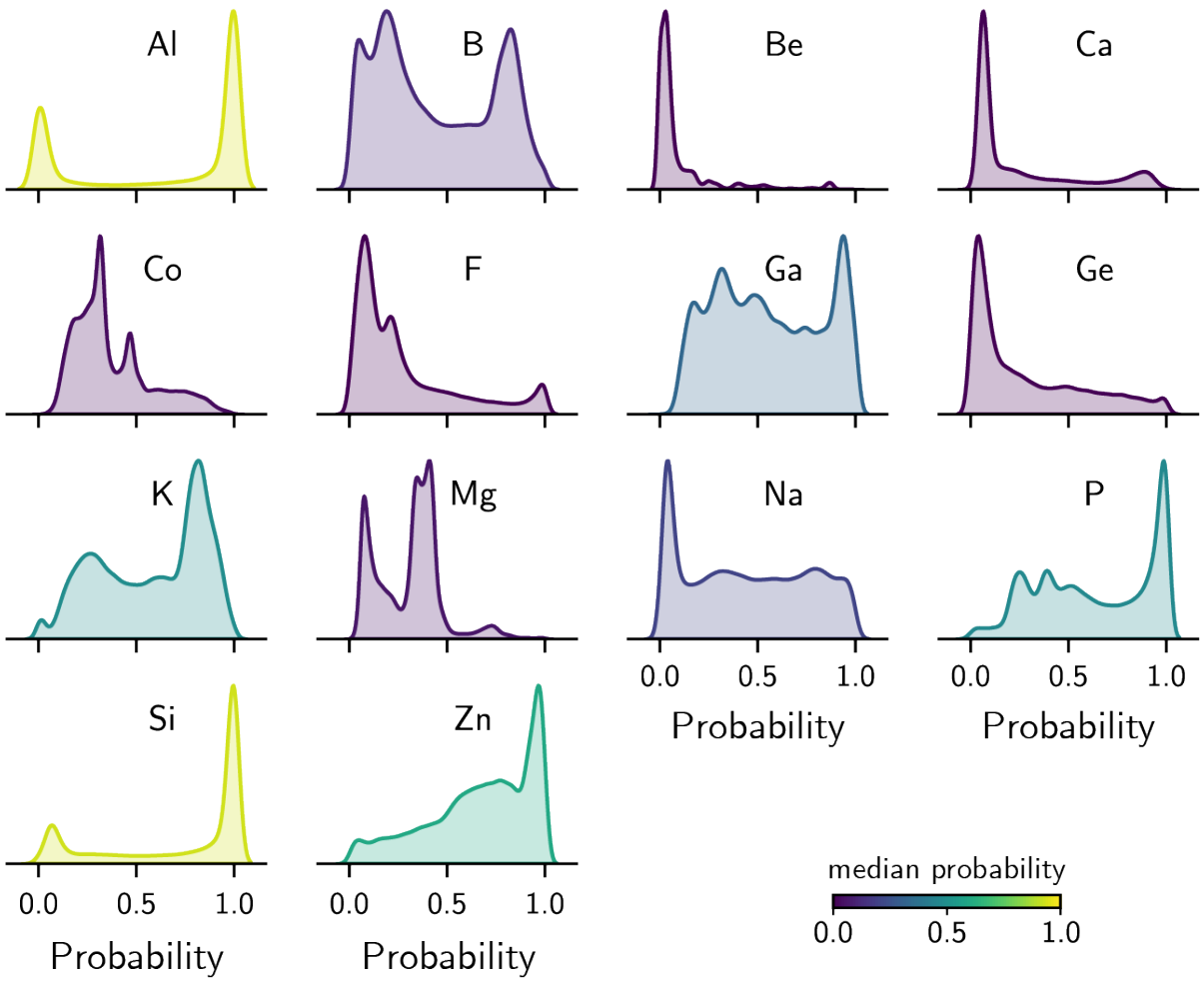

Figure S24: Distributions of synthesis probabilities, as predicted by 100 XGBoost models, for all hypothetical zeolites in the "Deem dataset". Brighter colors indicate higher medians of the distributions. The classifiers were trained on AMD distances.

Supporting Tables

Table S1: Closest pairs of IZA zeolites according to AMDs. The pairs shown are sorted according to distances between their AMDs, and have maximum AMD distance of 0.05 Å. As multiple observations/discussions of the pairs are available in the literature, we provide a single reference per pair that either reports the phase competition/integrowth of the zeolites or discuss their related structures. The reference is provided using its DOI code, or through a well-known group of zeolites (e.g., ABC-6 or cancrinite-like minerals). The group **ITG-UOV-IWW-UWY** can be rationalized as similar due to their channel topologies, similar densities, and synthesis using germanium.

Zeolite 1	Zeolite 2	AMD distance (Å)	Example reference
ITH	ITR	0.016	10.1524/zkri.2012.1564
ITG	UOV	0.021	
SBS	SBT	0.022	10.1126/science.abi7208
ITG	IWW	0.025	10.1021/ja301082n
MEL	*SFV	0.032	10.1126/science.1207466
MWF	PAU	0.033	10.1038/nature14575
IMF	*SFV	0.035	10.1126/science.1137920
IMF	TUN	0.035	10.1126/science.1137920
AFG	TOL	0.037	cancrinite-group minerals
FAR	MAR	0.037	cancrinite-group minerals
ERI	SWY	0.039	ABC-6 zeolites
OFF	SWY	0.040	ABC-6 zeolites
AFT	AFX	0.040	ABC-6 zeolites
IWW	UOV	0.041	
AFS	BPH	0.042	10.1524/zkri.1992.201.1-2.113
SFH	SFN	0.042	10.1002/chem.200305238
AFX	SFW	0.044	ABC-6 zeolites
EMT	FAU	0.045	10.1524/zkri.2012.1564
LIO	TOL	0.045	cancrinite-group minerals
PTY	PWO	0.046	10.1002/anie.201909336
IWS	SOV	0.047	10.1002/chem.201805187
AWO	UEI	0.047	10.1021/cm001199h
ITG	UWY	0.050	

Table S2: Hyperparameters explored for the logistic regression models. For the L_1 loss, only the `saga` solver was used. The `l1_ratio` parameter is used only in the case of the L_1 loss.

Parameter	Choices
`penalty`	[`l2`, `l1`, `none`]
`C`	[0.001, 0.01, 0.1, 1, 10, 100]
`solver`	[`lbfgs`, `liblinear`, `sag`, `saga`]
`l1_ratio`	[0.25, 0.5, 0.75, 1.0]

Table S3: Hyperparameters explored for the random forest classifiers.

Parameter	Choices
`n_estimators`	[50, 100, 200]
`max_depth`	[None, 10, 20]
`min_samples_split`	[2, 5, 10]
`min_samples_leaf`	[1, 2, 4]
`bootstrap`	[True, False]

Table S4: Hyperparameters explored for the XGBoost classifiers.

Parameter	Choices
`n_estimators`	[50, 100, 200]
`learning_rate`	[0.01, 0.1, 0.2]
`max_depth`	[3, 4, 5, 6]
`min_child_weight`	[1, 2, 3]
`subsample`	[0.5, 0.75, 1]
`colsample_bytree`	[0.5, 0.75, 1]

Table S5: Performance of the selected XGBoost model trained using AMD distances. The figures of merit are computed against a held-out test set. The standard deviation is computed for five different dataset splits, as described in the Methods, all of which have the same number of positive/negative labels. The number of data points (n) is the total number of training data points, sampled at 50:50 positive:negative labels to keep the classifier balanced.

Element	n	Accuracy	Precision	Recall	F_1 score	ROC AUC	PR AUC
Al	200	0.82 ± 0.03	0.84 ± 0.05	0.79 ± 0.09	0.81 ± 0.04	0.90 ± 0.02	0.89 ± 0.02
B	28	0.76 ± 0.15	0.79 ± 0.17	0.70 ± 0.24	0.73 ± 0.21	0.78 ± 0.16	0.80 ± 0.17
Be	12	0.81 ± 0.22	0.81 ± 0.37	0.69 ± 0.37	0.73 ± 0.36	0.78 ± 0.25	0.85 ± 0.17
Ca	24	0.53 ± 0.17	0.53 ± 0.18	0.53 ± 0.21	0.53 ± 0.19	0.65 ± 0.19	0.71 ± 0.15
Co	20	0.48 ± 0.16	0.43 ± 0.24	0.41 ± 0.27	0.41 ± 0.25	0.52 ± 0.14	0.54 ± 0.10
F	74	0.66 ± 0.16	0.67 ± 0.17	0.67 ± 0.14	0.67 ± 0.15	0.74 ± 0.14	0.75 ± 0.16
Ga	22	0.58 ± 0.25	0.58 ± 0.24	0.69 ± 0.29	0.62 ± 0.24	0.59 ± 0.25	0.58 ± 0.22
Ge	54	0.65 ± 0.04	0.66 ± 0.03	0.62 ± 0.14	0.63 ± 0.08	0.74 ± 0.04	0.73 ± 0.10
K	44	0.62 ± 0.05	0.63 ± 0.05	0.62 ± 0.12	0.62 ± 0.07	0.68 ± 0.11	0.70 ± 0.11
Mg	12	0.56 ± 0.20	0.54 ± 0.33	0.37 ± 0.28	0.43 ± 0.27	0.69 ± 0.21	0.73 ± 0.17
Na	104	0.67 ± 0.05	0.70 ± 0.10	0.67 ± 0.17	0.67 ± 0.06	0.74 ± 0.04	0.71 ± 0.06
P	94	0.64 ± 0.06	0.65 ± 0.06	0.60 ± 0.09	0.62 ± 0.07	0.70 ± 0.05	0.70 ± 0.08
Si	230	0.88 ± 0.04	0.95 ± 0.04	0.80 ± 0.06	0.87 ± 0.04	0.97 ± 0.02	0.97 ± 0.02
Zn	26	0.52 ± 0.28	0.53 ± 0.30	0.50 ± 0.28	0.51 ± 0.29	0.57 ± 0.32	0.63 ± 0.24

Table S6: Performance of the selected XGBoost model trained using SOAP distances. The figures of merit are computed against a held-out test set. The standard deviation is computed for five different dataset splits, as described in the Methods, all of which have the same number of positive/negative labels. The number of data points (n) is the total number of training data points, sampled at 50:50 positive:negative labels to keep the classifier balanced.

Element	n	Accuracy	Precision	Recall	F_1 score	ROC AUC	PR AUC
Al	200	0.75 ± 0.03	0.77 ± 0.04	0.70 ± 0.05	0.73 ± 0.04	0.83 ± 0.02	0.81 ± 0.04
B	28	0.68 ± 0.17	0.65 ± 0.15	0.82 ± 0.17	0.72 ± 0.14	0.74 ± 0.21	0.78 ± 0.19
Be	12	0.69 ± 0.18	0.65 ± 0.17	1.00 ± 0.00	0.78 ± 0.11	0.80 ± 0.21	0.83 ± 0.20
Ca	24	0.59 ± 0.17	0.64 ± 0.20	0.62 ± 0.13	0.61 ± 0.12	0.69 ± 0.18	0.79 ± 0.11
Co	20	0.55 ± 0.25	0.57 ± 0.22	0.53 ± 0.25	0.54 ± 0.23	0.54 ± 0.25	0.64 ± 0.20
F	74	0.70 ± 0.14	0.77 ± 0.19	0.61 ± 0.10	0.67 ± 0.12	0.73 ± 0.13	0.77 ± 0.10
Ga	22	0.58 ± 0.09	0.61 ± 0.17	0.69 ± 0.22	0.61 ± 0.08	0.66 ± 0.17	0.76 ± 0.12
Ge	54	0.68 ± 0.13	0.67 ± 0.14	0.71 ± 0.18	0.69 ± 0.14	0.71 ± 0.13	0.68 ± 0.13
K	44	0.53 ± 0.07	0.53 ± 0.06	0.61 ± 0.19	0.55 ± 0.10	0.60 ± 0.08	0.67 ± 0.07
Mg	12	0.48 ± 0.06	0.49 ± 0.04	0.96 ± 0.12	0.65 ± 0.06	0.38 ± 0.24	0.49 ± 0.18
Na	104	0.68 ± 0.11	0.69 ± 0.12	0.65 ± 0.15	0.66 ± 0.13	0.74 ± 0.08	0.73 ± 0.08
P	94	0.69 ± 0.06	0.71 ± 0.09	0.66 ± 0.09	0.68 ± 0.05	0.69 ± 0.06	0.69 ± 0.09
Si	230	0.85 ± 0.03	0.89 ± 0.07	0.80 ± 0.06	0.84 ± 0.03	0.91 ± 0.04	0.91 ± 0.05
Zn	26	0.55 ± 0.12	0.57 ± 0.11	0.65 ± 0.23	0.58 ± 0.11	0.59 ± 0.23	0.65 ± 0.20

Table S7: Comparison between the best-performing XGBoost classifiers trained on AMD and SOAP distances. Each set of hyperparameters was selected according to the best validation result for ROC AUC. The results indicate the performance on the test set. Error bars indicate the standard deviation of 100 independent training/testing runs for this experiment.

	ROC AUC		PR AUC	
Element	AMD	SOAP	AMD	SOAP
Al	0.90 ± 0.04	0.83 ± 0.02	0.87 ± 0.04	0.82 ± 0.04
B	0.75 ± 0.18	0.74 ± 0.20	0.79 ± 0.18	0.78 ± 0.18
Be	0.80 ± 0.23	0.80 ± 0.21	0.86 ± 0.15	0.83 ± 0.20
Ca	0.61 ± 0.18	0.68 ± 0.20	0.69 ± 0.12	0.76 ± 0.14
Co	0.49 ± 0.25	0.49 ± 0.20	0.55 ± 0.20	0.58 ± 0.19
F	0.77 ± 0.12	0.73 ± 0.12	0.81 ± 0.11	0.76 ± 0.10
Ga	0.59 ± 0.25	0.77 ± 0.16	0.58 ± 0.22	0.81 ± 0.12
Ge	0.73 ± 0.08	0.73 ± 0.13	0.72 ± 0.09	0.73 ± 0.13
K	0.71 ± 0.15	0.67 ± 0.09	0.74 ± 0.15	0.66 ± 0.09
Mg	0.76 ± 0.14	0.49 ± 0.30	0.76 ± 0.18	0.55 ± 0.24
Na	0.70 ± 0.06	0.74 ± 0.08	0.66 ± 0.07	0.71 ± 0.09
P	0.68 ± 0.07	0.70 ± 0.08	0.65 ± 0.08	0.67 ± 0.08
Si	0.97 ± 0.02	0.93 ± 0.03	0.97 ± 0.02	0.91 ± 0.06
Zn	0.58 ± 0.29	0.56 ± 0.25	0.63 ± 0.23	0.64 ± 0.18