Supporting Information to Symmetry-Informed Graph Neural Networks for Carbon Dioxide Isotherm and Adsorption Prediction in Aluminum-Substituted Zeolites

Marko Petković[ac], José-Manuel Vicent Luna[*a], Elīza Beate Dinne[a], Vlado Menkovski[bc] and Sofía Calero[*ac]

[a] *Materials Simulation and Modelling, Department of Applied Physics and Science Education, Eindhoven University of Technology, Eindhoven*

[b] *Data and AI, Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven*

[c] *Eindhoven Artificial Intelligence Systems Institute, Eindhoven University of Technology, Eindhoven*

[*] E-mail: j.vicent.luna@tue.nl, s.calero@tue.nl

1 Zeolite structures

1.1 Aluminium placement algorithms

To generate the dataset used in this project, the ZEORAN program [1] was used to place aluminium atoms in all-silica zeolites, using four algorithms for distributing the atoms throughout the structure. For some zeolite structures, the aluminium placement was done using PORRAN [2], which is a Python extension of ZEORAN. The four algorithms include *clusters*, *chains*, *maximum entropy* and *random*. These algorithms make use of a graph representation of the zeolite, where edges between T-atoms are drawn if they are part of the same T-O-T bond.

The *clusters* algorithm initially places an aluminium atom in a random position in the structure. Following this, the neighbours of aluminium atoms are recursively substituted with aluminium atoms, until the desired amount of substitutions is reached. Structures generated using this algorithm contain a high amount of non-Löwenstein bonds.

In the *chains* algorithm, a user-defined number of chains (with a user-defined length per chain) of aluminium atoms is placed throughout the zeolite struc-

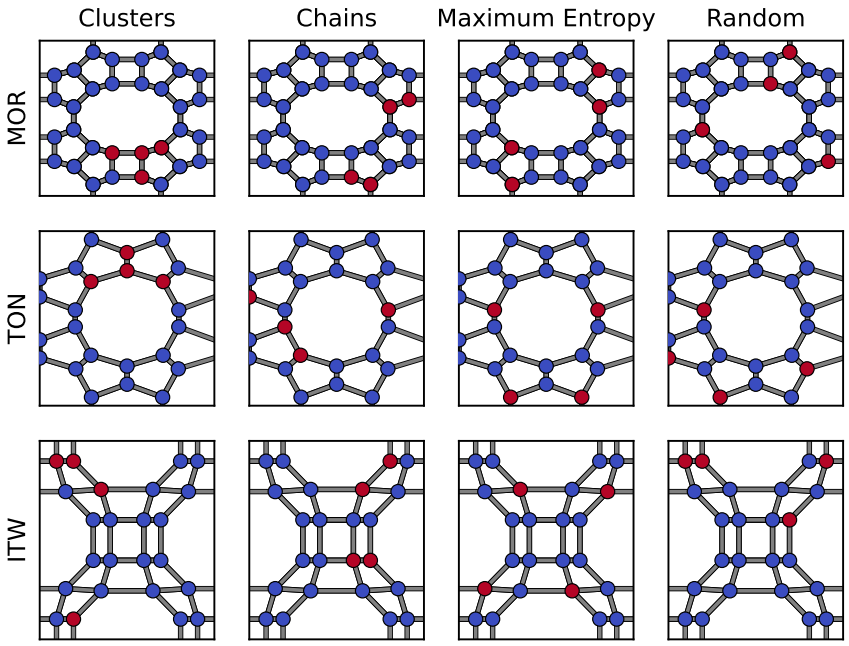

Figure S1: Example of structures generated with 4 aluminium substitutions for MOR, TON and ITW. In case of the *chains* algorithm, 2 chains of length 2 were placed. Note that *clusters* and *chains* algorithms might place aluminium patterns crossing the periodic boundary. All structures are viewed along the z-axis.

tures. Chains are placed in such a way that two separate chains do not connect. Furthermore, chains do not have branches, meaning that each aluminium atom in the chain will have one (if at the end) or two neighbours.

When using the *maximum entropy* algorithm, aluminium atoms are placed approximately uniformly distributed throughout the structure. In ZEORAN, this is achieved using a random walk. In PORRAN, aluminium atoms are iteratively placed, where silicon atoms which are the furthest from their closest aluminium atom have a proportionally higher chance to be selected. As such, there should be no non-Löwenstein bonds present in these structures.

Finally, the *random* algorithm randomly places aluminium atoms in the structure. Structures generated using this algorithm do not follow any particular distribution.

In Figure S1, each algorithm was used to generate an aluminium substitued strcuture for the MOR, TON and ITW zeolites. For each structure, 4 aluminium atoms were placed using their respective algorithms.

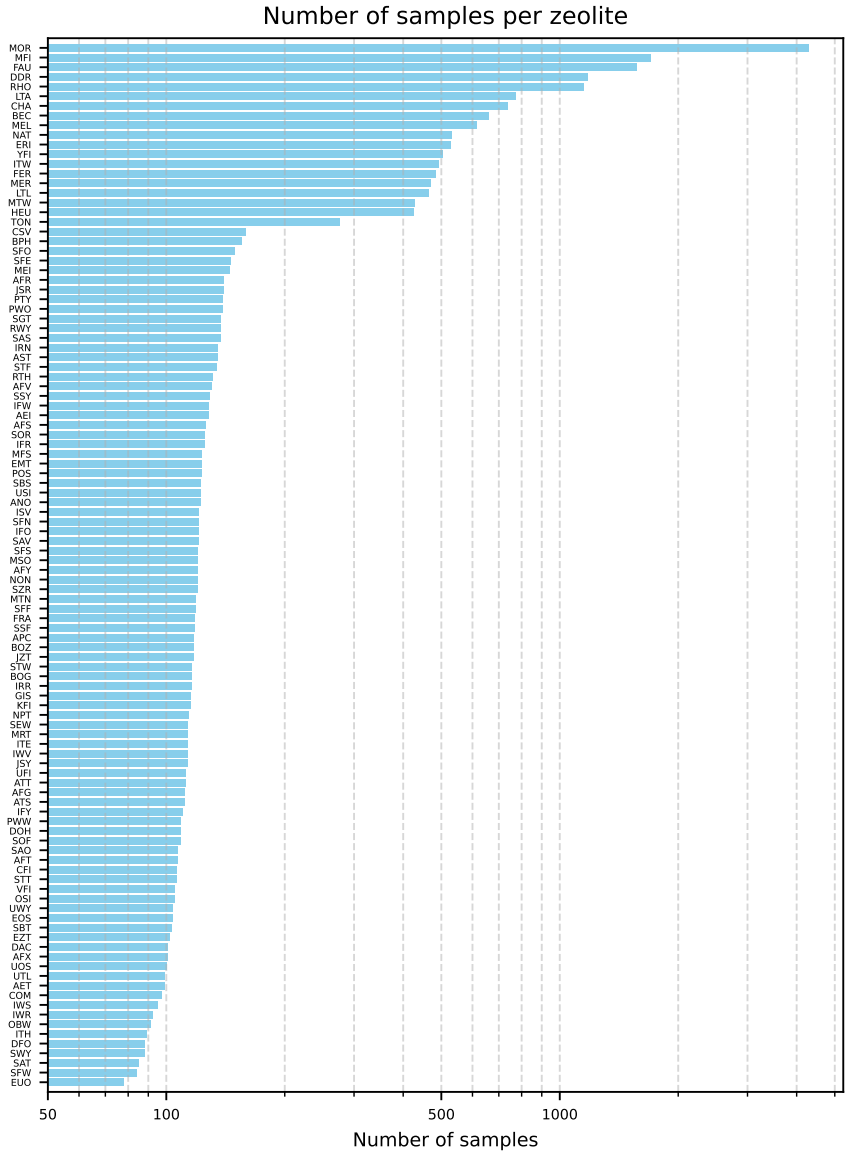

Figure S2: Number of samples for each zeolite topology. Note that the x-axis is in log-scale.

1.2 Zeolite topologies

In Figure S2, the number of samples for each topology used in this work can be found. While more structures were originally simulated, structures with a

heat of adsorption error of higher than 1.5 kJ/mol or an error higher than 5% of the range of the heat of adsorption for that topology were dropped. Topologies for which too much data was dropped (less than 75 structures) as part of this filtering were not considered in this work.

2 Isotherm simulation validation

2.1 Reduced simulation settings

Due to the large number of simulations needed to obtain an isotherm for a single structure, generating a large and varied dataset can be relatively time consuming. To speed up these simulations, some approximations can be made, such as reducing the number of unit cells used. As a result of the unit cell reduction, the super cell used in the simulation might not have a size twice the cutoff range for the Lennard-Jones potential. Depending on the zeolite topology, this can lead to interactions between atoms not being modeled properly. Therefore, it is necessary to verify that a simulation with the reduced number of unit cells produces results that are in agreement with the simulation with the correct number of unit cells.

Table S1: Adsorption isotherm simulation settings for each topology

	Full simulation box (#unit cells)	Reduced simulation box (#unit cells)
MOR	2x2x4	1x1x2
MFI	2x2x2	1x1x2
MEL	2x2x2	1x1x2
TON	2x2x5	2x2x5
ITW	2x2x3	2x2x3

The number of unit cells used for each zeolite topology can be found in Table S1. For the MOR, MFI and MEL zeolite, the number of unit cells used was reduced. To verify that the full and reduced simulations are in agreement, we selected five configurations of each zeolite, with varying Si/Al ratios. For each of these structures, we carried out a simulation using both the full and reduced settings. A comparison between the two can be found in Figure S3. For all different structures, we notice that there is a near-perfect agreement between the two simulation settings, with only minor fluctuations in the reduced simulation. Therefore, we can shorten these simulations using the reduced settings without any significant sacrifices in accuracy.

2.2 Isotherm Fitting

Since we used reduced simulation settings for some of the isotherms simulations, certain fluctuations might occur in the simulated loading. In order to minimize the effect of these fluctuations, we fit the 2-site Langmuir-Freundlich equation

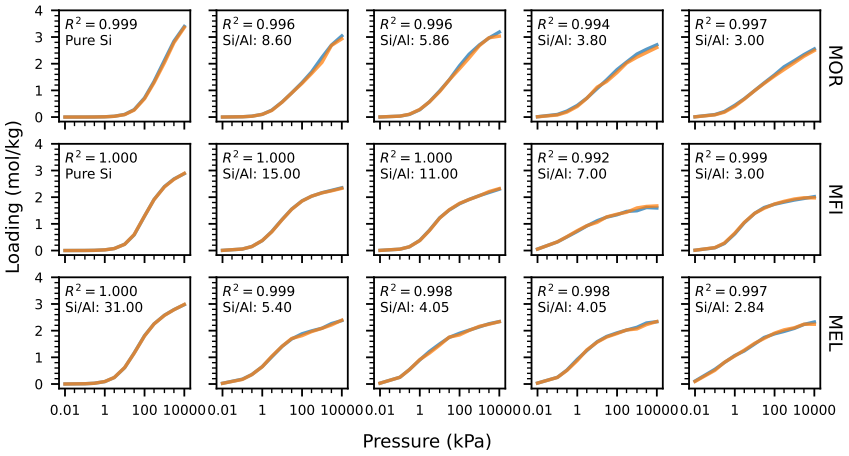

Figure S3: Isotherms for MOR, MFI and MEL structures with various Si/Al ratios, as predicted by the simulations using the full simulation box (blue), and the simulations using the reduced box (orange).

on the simulated loadings, using RUPTURA[3]. We compare the simulated loading with the fitted loading in Figure S4. As can be seen, there are some minor deviations from the diagonal in the parity plots, which suggests some outliers have been smoothed out, while the overall correlation between the fitted and simulated isotherms is excellent. As such, we can conclude that using RUPTURA to fit the 2-site Langmuir-Freundlich does not affect the overall shape of the isotherm while smoothing out fluctuations.

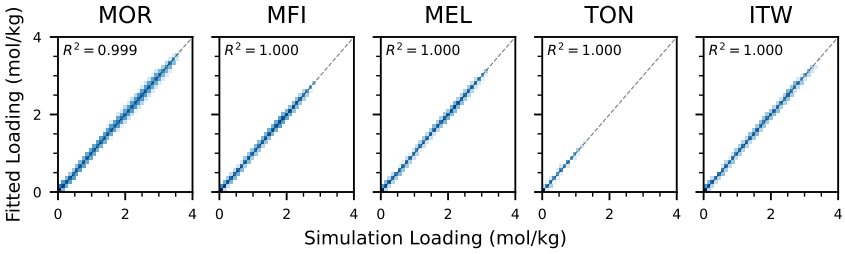

Figure S4: Parity plots for the isotherm fitting using RUPTURA. Darker blue indicates a higher count. Note that the color gets darker in log-scale.

3 Model Hyperparameters

In Table S2, the hyperparameters for the various models used in this work are listed. All models were trained for 400 epochs using the AdamW optimizer with default settings and Mean-Squared Error (MSE) loss for both heat of adsorption and isotherm predictions. The loading loss is computed over a random window of 25 points, selected from 100 logarithmically spaced pressure values. This is formalized in Equation S1, where $\mathcal{P}_i$ denotes the randomly selected subset of pressure points for structure i, N is the batch size, and $q_i'(p)$ and $\hat{q}_i'(p)$ are the true and predicted loading derivative values, respectively. The total loss, shown in Equation S2, combines the MSE on heat of adsorption with a weighted loading loss. During the first 100 epochs, β is 0, and linearly increases to 1 in the following 25 epochs.

$$\mathcal{L}_{\text{loading}} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{25}\sum_{p\in\mathcal{P}_i}\left(q_i'(p) - \hat{q}_i'(p)\right)^2 \tag{S1}$$

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\left(h_i - \hat{h}_i\right)^2 + \beta\cdot\mathcal{L}_{\text{loading}} \tag{S2}$$

Table S2: Hyperparameters for the models used in this work.

	GNN	SymGNN	ALIGNN	Matformer
Hidden features	64	64	128	64
Hidden features (Output)	64	64	128	64
Hidden features (Hypernet)	-	32	-	-
Layers	5	5	3+3[1]	5
Attention heads	-	-	-	4
Batch size	128	128	32	64
Edge dropout	0.5	0.5	-	-

4 Additional Results

In addition to evaluating generalization on CHA and ITW, we also test the models on MEL, MFI, TON and MOR. The results are presented in Table S3. SymGNN generally performs better for isotherm prediction, while the regular GNN shows slightly lower errors for the heat of adsorption.

Figure S5 shows the true and predicted isotherm distributions. SymGNN performs well for MEL and MFI but tends to underestimate the variance near saturation pressures. The models struggle more with TON and MOR. The

[1]The ALIGNN model contains 3 ALIGNN layers, followed by 3 edge-gated graph convolution layers.

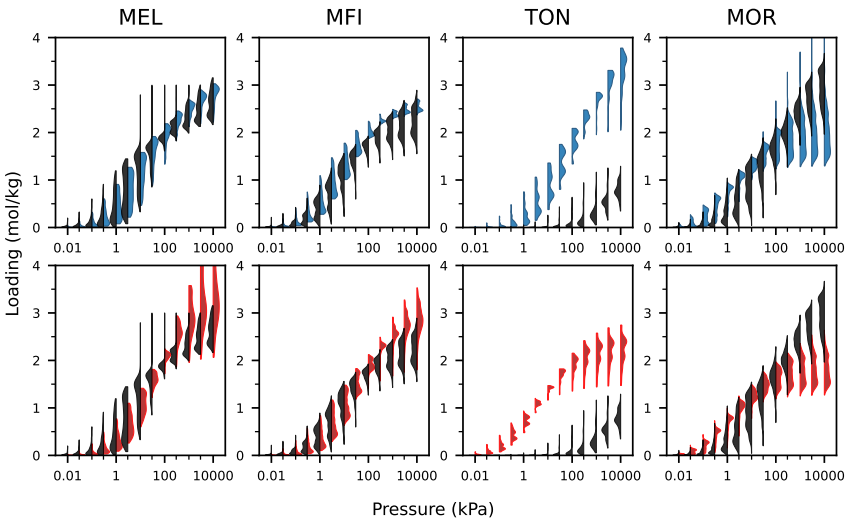

Figure S5: Comparison of SymGNN and regular GNN in isotherm prediction for the generalization experiment carried out with various zeolites. True loading distribution (black) and loading predicted by SymGNN (blue, top row) and GNN (red, bottom row) at all simulated pressures.

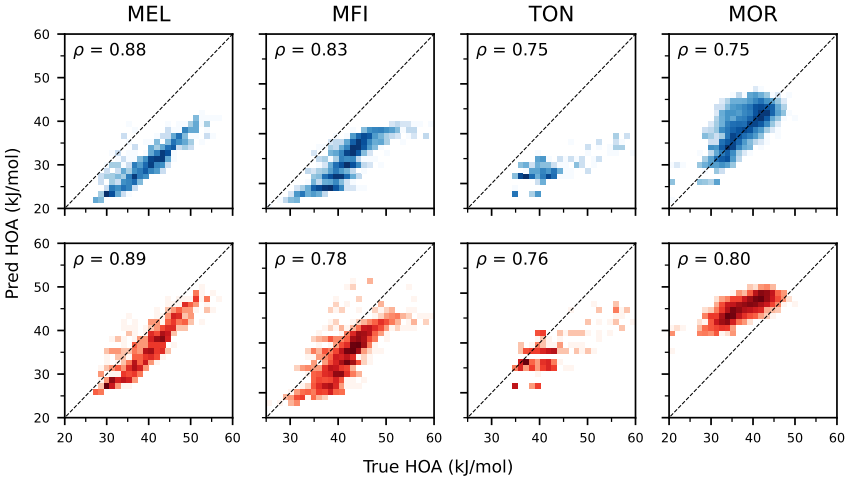

Figure S6: Parity plots for the heat of adsorption prediction for SymGNN (blue, top row) and the regular GNN (red, bottom row) for the generalization experiment carried out with various zeolites. prediction.

Table S3: Results for the zeolite test set.

		Heat of adsorption		Isotherm		Isotherm sat.	
Zeolite	Model	MAE	MSE	MAE	MSE	MAE	MSE
MEL	SymGNN	9.44	97.89	0.15	0.04	0.21	0.06
	GNN	4.21	22.05	0.25	0.13	0.65	0.54
MFI	SymGNN	7.32	60.33	0.17	0.06	0.30	0.13
	GNN	4.60	26.68	0.21	0.08	0.54	0.31
TON	SymGNN	9.17	101.43	0.98	1.84	2.37	5.75
	GNN	4.29	29.50	1.03	1.57	1.51	2.42
MOR	SymGNN	2.24	9.15	0.33	0.19	0.84	0.74
	GNN	7.48	63.13	0.36	0.25	1.02	1.07

adsorption behavior in TON is characterized by a delayed uptake and lower maximum loading (in the range of simulated pressures), which are not well represented in the training set. On the other hand, MOR exhibits a higher maximum loading than the other zeolites. These differences in adsorption behavior are difficult to predict accurately without sufficient examples in the training data. Improving performance on these cases would likely require additional data for a broader range of topologies or pre-training focused on capturing topological effects.

Parity plots for the heat of adsorption in Figure S6 show that both models capture trends related to different Si/Al configurations, as can been seen in the high correlation between the true and predicted values. However, they often underestimate or overestimate the absolute values. This indicates that while the models learn local composition effects, capturing the full influence of topology remains a challenge.

References

1 P. Romero-Marimon, J. J. Gutiérrez-Sevillano and S. Calero, *Chemistry of Materials*, 2023, **35**, 5222–5231.

2 M. Petkovic and K. Wortelboer, *marko-petkovic/porran: Porran*, 2025, `https://doi.org/10.5281/zenodo.15050436`.

3 S. Sharma, S. R. Balestra, R. Baur, U. Agarwal, E. Zuidema, M. S. Rigutto, S. Calero, T. J. Vlugt and D. Dubbeldam, *Molecular Simulation*, 2023, 1–61.