CAPTURING NON-LOCAL FEATURES OF CRYSTALS FROM THEIR BOND NETWORKS

Qianxiang Ai
Abstrax Tech, Inc.
qianxiang.ai@Abstraxtech.com

Sartaaj Takrim Khan
Chemical Engineering & Applied Chemistry
University of Toronto
sartaaj.khan@mail.utoronto.ca

Senja Barthel
Mathematics
Vrije Universiteit Amsterdam
s.barthel@vu.nl

Seyed Mohamad Moosavi
Chemical Engineering & Applied Chemistry
University of Toronto
mohamad.moosavi@utoronto.ca

ABSTRACT

Representing crystal structures for machine learning property prediction traditionally relies on either composition-based methods or structure-based graph neural networks (GNNs). While these methods have been successful in predicting certain properties, they fall short in accurately capturing the periodicity of crystal structures, particularly long-range information. In this work, we show that topological features derived from labeled quotient graphs (LQGs)–finite graph representations that encode bond topology without relying on real-space geometric information–can effectively predict non-local properties, i.e., properties that are not solely determined by individual local atomic environments. Using a dataset of 25,000 silica zeolite structures, we demonstrate that XGBoost models trained on LQG-derived topological features (XGB-LQG) outperform conventional GNNs (CGCNN, MEGNet) in predicting non-local properties. Furthermore, hybrid architectures that combine GNN embeddings with LQG features achieve intermediate performance, highlighting the complementary nature of geometric and topological representations. Our results establish LQGs as a powerful representation for incorporating bond topology into crystal property prediction.

1 INTRODUCTION

Crystal property prediction models generally fall into two distinct categories based on their input information: composition-based models (Gupta et al., 2023; Li et al., 2021; Ward et al., 2016; Venkatraman, 2021), which only take unit cell compositions (i.e., atom types and counts) as input, and structure-based models, which incorporate atomic coordinates. Structure-based models, particularly those based on graph neural networks (GNNs), have been extensively explored in recent years for predicting crystal properties. Notable examples include Crystal Graph Convolutional Neural Networks (CGCNN) (Xie & Grossman, 2018), MatErials Graph Network (MEGNet) (Chen et al., 2019), and Atomistic Line Graph Neural Network (ALIGNN) (Choudhary & DeCost, 2021). These models are designed to learn a representation of the crystal with real-space geometric features, including bond lengths and/or bond angles that are derived from atomic coordinates.

In contrast, molecular property prediction models often use molecular graphs without incorporating real-space geometric features, as the properties of interest are typically macroscopic responses of an ensemble of conformers. Thus, for molecules, there are at least three levels of representation, ranging from abstract to concrete: chemical formula (composition-level), molecular graph (graph-level), and conformer (structure-level), as illustrated in Figure 1 (top). These three levels of representation can be applied similarly to categorize crystal representations. Interestingly, crystal property prediction models tend to use either the most abstract or the most concrete representation (Figure 1, bottom). The contrast in the choice of representation between molecular and crystal prediction mod-

els motivated our investigation of graph-level representations for crystal property prediction without real-space geometric features.

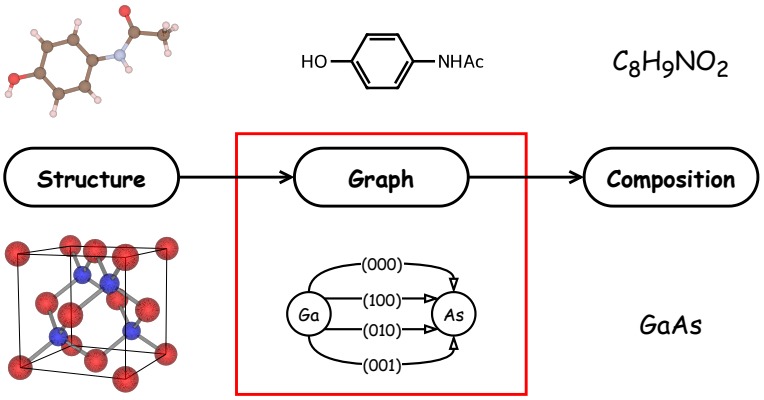

Figure 1: Materials can be represented at multiple levels of abstraction, ranging from their chemical compositions (most abstract), their chemical bond graphs, to their real-space structures (most concrete). These representations for both molecules and crystals are illustrated using acetaminophen and gallium arsenide as examples. The graph-level representation (labeled quotient graph) for crystals, which is the focus of this study, is highlighted in the red box.

We focus on a graph-level representation known as labeled quotient graphs (LQGs). LQGs were first formally introduced in 1984 (Chung et al., 1984). LQGs are finite graphs that encode the connectivity information of the infinite periodic bond network (underlying net) of a crystalline material. The periodicity is encoded in the labels of the LQG. That is, the LQGs are independent from atomic (or structure subunit) coordinates in real space and only depend on the presence of bonds between atoms (subunits). It is possible to obtain the underlying net of a crystal structure from its LQG but the atomic coordinates are not retrievable. The graph isomorphism type of an underlying net is called the bond topology of the structure.

Given the crystal structure and its underlying net, an LQG can be constructed following the procedure in Section 2.1. LQG edges are labeled to encode periodicity, enabling the original net to be reproduced by unfolding the LQG. As a finite graph representation, LQGs are used to digitally store periodic nets in databases such as the Reticular Chemistry Structure Resource (O'keeffe et al., 2008), EPINET (Ramsden et al., 2009), and the Topological Types Database (Blatov et al., 2014). Similar to molecular conformation generation, theories and tools have been developed to generate real-space crystal structures (net-embeddings) for a given LQG through barycentric placement (Delgado-Friedrichs & O'Keeffe, 2003) or co-lattice methods (Eon, 2011; Xiao et al., 2023). LQG-based string representations have also been proposed to represent the bond topologies of crystal structures (Delgado-Friedrichs et al., 2017; Krenn et al., 2022; Xiao et al., 2023). Pertaining to property predictions, previous studies have shown that some properties of crystal structures are strongly correlated with or determined by features (topological invariants) of their LQGs. For example, the dimensionality of a crystal structure can be derived from its quotient graph's basic cycle sums matrix (Gao et al., 2020). Additionally, the LQG determines both the minimal and maximal possible symmetry of its net-embeddings (Thimm, 2009). These results make LQGs a promising representation for predicting crystal properties, especially long-range properties or properties that are dependent on different regions of the crystal, which are hard tasks for structure-based GNN models (Gong et al., 2023).

While LQGs have been used for crystal structure enumeration/generation, their application to property prediction remains underexplored. Zeolites, with their well-defined bond topologies, various pore geometries, and diverse connectivity patterns, make an interesting test case for evaluating the predictive power of LQG-based models. In this study, we explore predicting zeolite properties using only the crystal's LQG without incorporating real-space coordinate or lattice information. We found that XGBoost models trained using LQG features outperform structured-based GNN models in pre-

dicting non-local properties including pore geometry and accessible surface area. GNN models can be improved by including LQG features for all properties considered in this study. Our findings establish LQGs as a viable approach to include bond topology information in crystal property prediction.

2 METHODS

2.1 LABELED QUOTIENT GRAPHS

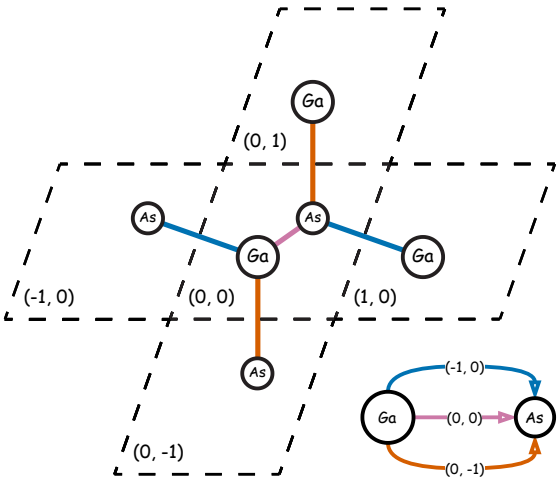

Figure 2: Construction of the LQG for the (100) surface of the primitive cell of gallium arsenide ($\{111\}$ of the conventional cell). Dashed boxes denote unit cells of the surface. The 2-tuple at the bottom of a unit cell denotes the cell index of this unit cell. The resulting LQG of two nodes and three labeled edges is shown at the bottom right of the figure.

Figure 2 shows the construction of the LQG for the (100) surface of the primitive cell of gallium arsenide ($\{111\}$ of the conventional cell) using the procedure introduced by Chung et al. Given the coordinate system defined by surface lattice vectors and a selected unit cell as initial cell with an index of (0, 0), other unit cells can be assigned indices based on their relative positions to the origin unit cell in the lattice, e.g. (-1, 0) for the cell next to the initial cell on the left. Using an ionic radii-based distance cutoff, Ga-As bonds can be identified in the surface structure. An LQG can then be constructed as follows:

1. For each atom inside the (0, 0) cell add a node to the LQG. In the GaAs example there are two atoms and therefore two nodes in its LQG;
2. Find all pairs of atoms containing at least one atom in the (0, 0) cell between which bonds are formed. Group them into translationally equivalent bond sets. In the GaAs example there are three bond sets, colored by purple, blue, and orange, respectively.
3. From each bond set, arbitrarily select one bond. Add a directed edge to the LQG, connecting the nodes that correspond to the atoms forming the chosen bond, starting from a node that corresponds to an atom in the (0,0) cell. Label the edge with (0,0) if the bond is contained in the initial cell. If one of the bond-forming atoms lies outside the (0,0) cell in the (n,m) cell, label the edge with (n,m).

In practice, dedicated coordination determination packages are recommended for more complicated bonding situations (Pan et al., 2021). The periodicity of the crystal structure's net is represented by edge labels from the LQG construction. For example, an LQG cycle with a nonzero cycle sum (sum of edge labels in that cycle) determines a crystallographic translation as a path between two

translationally equivalent atoms from neighboring cells. Note that edges are only present between nodes corresponding to atoms that are close enough to interact chemically. This is different from the construction of graphs in some previous GNN models. For example, CGCNN graphs could connect atoms that are very far away compared to the average chemical bond lengths (Xie & Grossman, 2018). Additionally, no geometric information (e.g., atomic positions/distances in real-space) is explicitly present in an LQG: The only node feature is the atomic type, and the only edge feature is the LQG edge label.

Following the procedure in Section A.1, a set of 25,000 all-silica zeolite structures is produced to benchmark models in this study. Because the pure silica zeolites in this study only consist of interconnected SiO_4 tetrahedra, the LQGs are simplified by contracting SiO_4 tetrahedra to individual nodes, and the corner- or edge-sharing relations between two tetrahedra are mapped to edges. This procedure of contracting primary building units (PBUs) is used to create a more compact PBU-contracted graph from the bond-based LQG for each zeolite structure. Graph features (topological invariants) of the PBU-contracted graphs are used to train regression models to predict zeolite properties. A list of 203 graph features, including centrality measures, spectral properties, degree distributions, community measures, cycle properties, can be found in the Supplementary Materials.

2.2 MODELS

The "Dummy" model in Table 1 is a dummy regressor that takes the mean of the training labels and makes the assumption that all of the data points in the test set are equal to the mean of the training set. For more interesting comparisons, GNNs such as a crystal graph convolutional neural network (CGCNN) (Xie & Grossman, 2018) and MatErials Graph Network (MEGNet) (Chen et al., 2019) were chosen. The CGCNN aims to represent the crystal structure by capturing the pairwise interactions between atoms and declaring the node and bond features with atomistic properties (group number, period number, electronegativity, covalent radius, valence electrons, first ionization energy, electron affinity, block, atomic volume) and bond-relevant properties (bond distance) respectively. MEGNet follows a similar approach but includes global state information about the crystal, such as temperature. The hyperparameters chosen for these models are given in the appendix (Table 3 and Table 4).

Given the computed LQG features described previously as input, for every zeolite property, an XGBoost model (XGB-LQG) was trained and evaluated. Extensive hyperparameter tuning was done by performing a stepwise Bayesian Optimization while performing cross-validation across different folds for combinations of hyperparameters. The final hyperparameters for this model are given in the appendix (Table 2). Furthermore, a CGCNN and a neural network accepting LQG descriptors were concatenated to perform property predictions (CGCNN-LQG). This was done by concatenating the learned representation from CGCNN and the latent representation of LQG from a neural network, before passing to the final prediction layer. Similar approaches have been applied to include additional features in GNNs for both molecular (Yang et al., 2019) and crystal (Gong et al., 2023) property predictions. While LQGs can be used directly as input to GNN models, we intentionally avoid using GNNs for LQGs of zeolites investigated in this study since many of the PBU-contracted LQGs are regular graphs, and common GNN architectures are expected to fail in distinguishing them (Xu et al., 2018). Additionally, zeolites' LQGs are almost identical locally – nearly all nodes correspond to SiO_4 tetrahedra connected to other four tetrahedra, which would likely lead to immediate oversmoothing in message-passing.

3 RESULTS

To evaluate LQG features as a crystal representation, we predicted six zeolite properties: diameter of the largest included sphere (D_i), diameter of the largest free sphere (D_f), diameter of the largest included sphere along free path (D_{if}), primitive cell volume, density, and accessible surface area (ASA). Figure 3 and Table 1 present the comparative prediction results across models and representations.

The selected properties represent non-local structural characteristics that depend on the extended crystal frameworks – a class of predictions where GNNs have demonstrated limitations (Gong et al., 2023; Khan & Moosavi, 2024). Our results confirm this pattern: both CGCNN and MEGNet exhib-

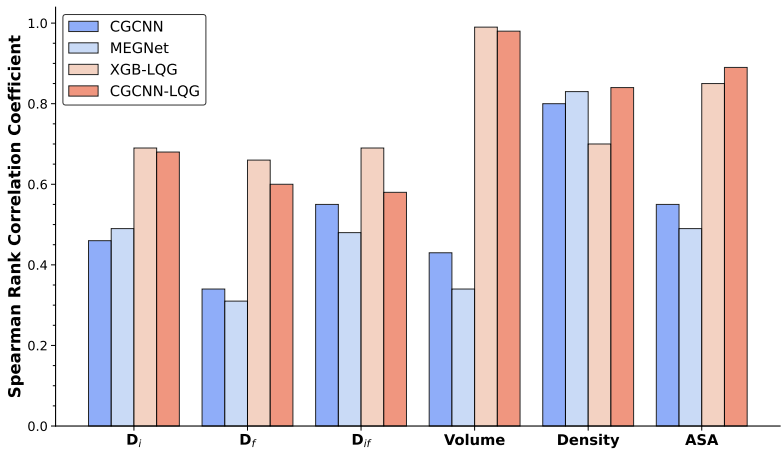

Figure 3: **Regression results on zeolites.** Comparison between GNNs (CGCNN and MEGNet) and XGBoost models using LQG features (XGB-LQG). The figure also includes results from a hybrid architecture combining CGCNN's embeddings with LQG feature encodings (CGCNN-LQG). The values reported in this figure are spearman rank correlation coefficients, and corresponding mean absolute errors (MAE) are shown in Table 1.

ited limited predictive power for pore geometry (D_i, D_f, D_{if}), primitive cell volume, and accessible surface area, though they achieved reasonable performance on density predictions.

Notably, the XGBoost model using only LQG features (XGB-LQG) outperformed both GNNs across all properties with the only exception of density, despite operating without any real-space geometric information, which clearly establishes the capability of LQG features in predicting properties that are not determined by local surroundings. This performance gap was particularly notable for primitive cell volume predictions, highlighting GNNs' challenges in capturing long-range structural information in framework materials like zeolites. The results demonstrate how appropriate representation selection enables simple models (XGB-LQG) to surpass complex models that use more concrete representations (GNNs).

We further developed a hybrid architecture combining CGCNN's embeddings with LQG feature encodings before the final regression layer. This concatenated model consistently outperformed the base CGCNN across all properties, confirming that LQG features provide complementary topological information missing in conventional GNN representations. However, the hybrid model did not surpass XGB-LQG's performance on any property prediction except for density and accessible surface area. This architecture-dependent discrepancy may suggest XGBoost's tree-based architecture exploits these discrete topological features better than neural networks.

4 Conclusion

This study explores the application of labeled quotient graphs (LQGs) in crystal property prediction. We demonstrate that XGBoost models trained on LQG features without real-space geometric information tend to outperform geometry-aware GNNs in predicting non-local zeolite properties. GNNs can also be improved by including LQG features that capture bond topology information that GNNs often struggle to learn. Encouraged by these results, we speculate that further improvements could be achieved by exploring additional LQG features, as well as LQGs in other formats, e.g., as strings (Delgado-Friedrichs et al., 2017; Krenn et al., 2022; Xiao et al., 2023). Other mathematical representations of periodic graphs, such as 2D hyperbolic tilings placed on triply periodic minimal surfaces (Ramsden et al., 2009), could also be valuable tools in crystal property predictions. Finally, while LQG could be a suitable representation in predicting non-local properties for materials of high bond topology diversity, this representation is expected to fall short for other situations where geometric information is crucial, e.g., layered materials with identical intra-layer bond topology but different inter-layer distances.

ACKNOWLEDGMENTS

Q.A. acknowledges the support by the National Institutes of Health under award number U18TR004149. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Q.A. thanks the computation resource generously provided by Dr. Connor W. Coley's group and helpful discussions with Dr. Runzhong Wang. This work received support from the National Research Council of Canada (NRC) under the Materials for Clean Fuels Challenge Program (grant number MCF-146). In addition, the project received support from the University of Toronto's Acceleration Consortium through the Canada First Research Excellence Fund under Grant number CFREF-2022-00042 and from the Data Science Institute (DSI) at the University of Toronto. S.M.M. research program receives financial support from Natural Sciences and Engineering Research Council of Canada (NSERC) through the discovery program.

REFERENCES

Ch. Baerlocher, Darren Brouwer, Bernd Marler, and L.B. McCusker. Database of zeolite structures. `https://www.iza-structure.org/databases/`. Accessed: 2023-01-01.

Vladislav A Blatov, Alexander P Shevchenko, and Davide M Proserpio. Applied topological analysis of crystal structures with the program package topospro. *Crystal Growth & Design*, 14(7): 3576–3586, 2014.

Chi Chen, Weike Ye, Yunxing Zuo, Chen Zheng, and Shyue Ping Ong. Graph networks as a universal machine learning framework for molecules and crystals. *Chemistry of Materials*, 31(9):3564–3572, 2019.

Kamal Choudhary and Brian DeCost. Atomistic line graph neural network for improved materials property predictions. *npj Computational Materials*, 7(1):185, 2021.

Sui Jin Chung, Th Hahn, and WE Klee. Nomenclature and generation of three-periodic nets: the vector method. *Acta Crystallographica Section A: Foundations of Crystallography*, 40(1):42–50, 1984.

Michael W Deem, Ramdas Pophale, Phillip A Cheeseman, and David J Earl. Computational discovery of new zeolite-like materials. *The Journal of Physical Chemistry C*, 113(51):21353–21360, 2009.

Olaf Delgado-Friedrichs and Michael O'Keeffe. Identification of and symmetry computation for crystal nets. *Acta Crystallographica Section A: Foundations of Crystallography*, 59(4):351–360, 2003.

Olaf Delgado-Friedrichs, Stephen T Hyde, Michael O'Keeffe, and Omar M Yaghi. Crystal structures as periodic graphs: the topological genome and graph databases. *Structural Chemistry*, 28:39–44, 2017.

J-G Eon. Euclidian embeddings of periodic nets: definition of a topologically induced complete set of geometric descriptors for crystal structures. *Acta Crystallographica Section A: Foundations of Crystallography*, 67(1):68–86, 2011.

Hao Gao, Junjie Wang, Zhaopeng Guo, and Jian Sun. Determining dimensionalities and multiplicities of crystal nets. *npj Computational Materials*, 6(1):143, 2020.

Sheng Gong, Keqiang Yan, Tian Xie, Yang Shao-Horn, Rafael Gomez-Bombarelli, Shuiwang Ji, and Jeffrey C Grossman. Examining graph neural networks for crystal structures: limitations and opportunities for capturing periodicity. *Science Advances*, 9(45):eadi3245, 2023.

Vishu Gupta, Kamal Choudhary, Yuwei Mao, Kewei Wang, Francesca Tavazza, Carelyn Campbell, Wei-keng Liao, Alok Choudhary, and Ankit Agrawal. Mppredictor: An artificial intelligence-driven web tool for composition-based material property prediction. *Journal of Chemical Information and Modeling*, 63(7):1865–1871, 2023.

Sartaaj Takrim Khan and Seyed Mohamad Moosavi. Connecting metal-organic framework synthesis to applications with a self-supervised multimodal model. 2024.

Mario Krenn, Qianxiang Ai, Senja Barthel, Nessa Carson, Angelo Frei, Nathan C Frey, Pascal Friederich, Théophile Gaudin, Alberto Alexander Gayle, Kevin Maik Jablonka, et al. Selfies and the future of molecular string representations. *Patterns*, 3(10), 2022.

Yuxin Li, Rongzhi Dong, Wenhui Yang, and Jianjun Hu. Composition based crystal materials symmetry prediction using machine learning with enhanced descriptors. *Computational Materials Science*, 198:110686, 2021.

Michael O'keeffe, Maxim A Peskov, Stuart J Ramsden, and Omar M Yaghi. The reticular chemistry structure resource (rcsr) database of, and symbols for, crystal nets. *Accounts of chemical research*, 41(12):1782–1789, 2008.

Hillary Pan, Alex M. Ganose, Matthew Horton, Muratahan Aykol, Kristin A. Persson, Nils E. R. Zimmermann, and Anubhav Jain. Benchmarking coordination number prediction algorithms on inorganic crystal structures. *Inorganic Chemistry*, 60(3):1590–1603, 2021. doi: 10.1021/acs.inorgchem.0c02996. URL https://doi.org/10.1021/acs.inorgchem.0c02996. PMID: 33417450.

SJ Ramsden, Vanessa Robins, and ST Hyde. Three-dimensional euclidean nets from two-dimensional hyperbolic tilings: kaleidoscopic examples. *Acta Crystallographica Section A: Foundations of Crystallography*, 65(2):81–108, 2009.

Georg Thimm. Crystal topologies–the achievable and inevitable symmetries. *Acta Crystallographica Section A: Foundations of Crystallography*, 65(3):213–226, 2009.

Vishwesh Venkatraman. The utility of composition-based machine learning models for band gap prediction. *Computational Materials Science*, 197:110637, 2021.

Logan Ward, Ankit Agrawal, Alok Choudhary, and Christopher Wolverton. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials*, 2(1):1–7, 2016.

Thomas F Willems, Chris H Rycroft, Michaeel Kazi, Juan C Meza, and Maciej Haranczyk. Algorithms and tools for high-throughput geometry-based analysis of crystalline porous materials. *Microporous and Mesoporous Materials*, 149(1):134–141, 2012.

Hang Xiao, Rong Li, Xiaoyang Shi, Yan Chen, Liangliang Zhu, Xi Chen, and Lei Wang. An invertible, invariant crystal representation for inverse design of solid-state materials using generative deep learning. *Nature Communications*, 14(1):7027, 2023.

Tian Xie and Jeffrey C Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical review letters*, 120(14):145301, 2018.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, et al. Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling*, 59(8):3370–3388, 2019.

A APPENDIX

A.1 DATA DISTRIBUTIONS

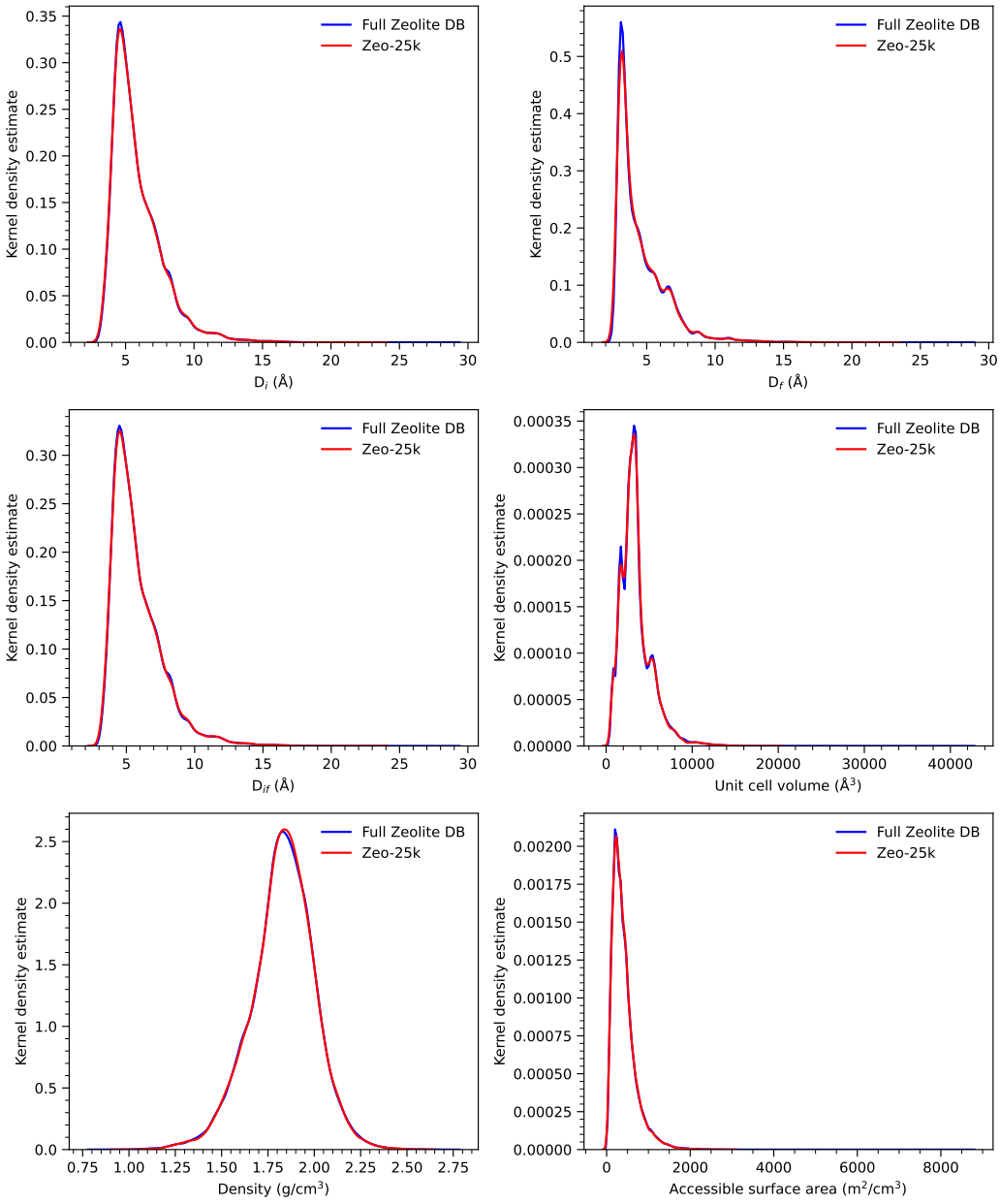

Figure 4: **Distributions of dataset** This figure shows the kernel density estimates (KDE) of each property predicted - including largest included sphere (D_i), largest free sphere (D_f), largest included sphere along free path (D_{if}), density, volume and ASA. The blue curve is obtained using the full zeolite database (DB) with 137k entries, and the red curve the 25k entries that are used for property predictions across all models.

We use a collection of 137,948 all-silica zeolite crystal structures, extracted from the International Zeolite Association (IZA) database (Baerlocher et al.) and computationally generated using a systematic Monte Carlo search (Deem et al., 2009). 25,000 zeolite structures were sampled randomly from the full zeolite dataset, and were randomly divided in an 80/10/10 split to form the train, vali-

dation and test sets, respectively, giving the zeolite IDs for each set and avoiding data leakage from the training sets into the validation or test sets. Crystal properties including the largest included sphere (D_i), largest free sphere (D_f), largest included sphere along free path (D_{if}), unit cell volume, density, and accessible surface area (ASA) were calculated using Zeo++ version 0.3 (Willems et al., 2012). These properties are non-local structural features defined beyond local atomic environments and are empirically challenging for GNN models to predict (Gong et al., 2023). For example, D_i can be as large as 10 Å, D_f and D_{if} are defined based on the extended channels in zeolite structures (Figure 5). Property distributions can be found in Figure 4.

A.2 Types of diameter in zeolites

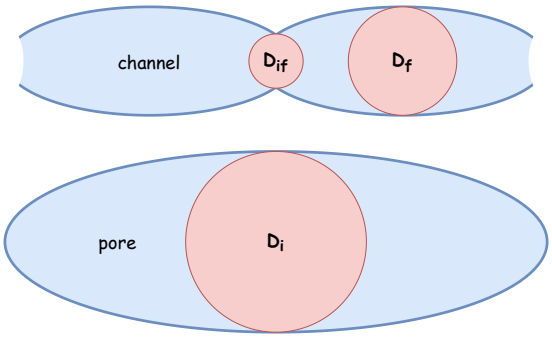

Figure 5: **Different types of diameter in zeolites** Illustration of diameters defined by the largest free sphere (D_f), the largest included sphere along free path (D_{if}), and the largest included sphere (D_i).

A.3 Other metrics for regression results

Table 1: Comparison between GNNs (CGCNN and MEGNet), CGCNN-LQG, and XGB-LQG across various zeolite properties. The "dummy" results are obtained from a model taking the average of the training set labels. The values reported are mean absolute errors (MAE). D_i stands for the diameter of the largest included sphere in Å, D_f the diameter of the largest free sphere in Å, and D_{if} the diameter of the largest included sphere along free path in Å. Values for primitive cell volume, density, and accessible surface area (ASA) are reported in Å^3/cell, g/cm^3, m^2/cm^3, respectively.

Model	**D_i**	**D_f**	**D_{if}**	**Volume**	**Density**	**ASA**
Dummy	1.39	1.35	1.40	1309	0.13	206.94
CGCNN	1.17	1.16	1.08	1250	0.08	166.74
MEGNet	1.18	1.19	1.17	1236	0.07	171.18
XGB-LQG	0.88	0.81	0.88	176	0.08	97.81
CGCNN-LQG	0.93	0.93	1.04	186	0.06	79.70

A.4 Model hyperparameters

Table 2 shows the hyperparameters of the XGBoost model across the different geometric properties. They were obtained by performing a Bayesian Optimization-driven search across the optimal possible combinations of hyperparameters to minimize the negative MAE. When evaluating a set of hyperparameters, the training set was split into 5 folds, with one left for scoring the model with that hyperparameter set. The search space was defined as:

1. Learning rate, ranging from 0.0001 to 1.0 (step-wise being log-uniform);
2. Max depth ranging from 3 to 10 (accepting only integers);
3. Subsample ranging from 0.1 to 1.0 (uniformly increasing);

4. Subsample ratio of columns at each level when constructing a tree, ranging from 0.1 to 1.0 (uniformly increasing);
5. Alpha (L1 regularization term), ranging from 1e-6 to 1.0 (increasing by log-uniform order)
6. Lambda (L2 regularization term), ranging from 1e-6 to 1000 (increasing by log-uniform order)

Table 2: Optimized hyperparameter values for XGB-LQG model on 25,000 zeolites for different geometric properties. This was achieved through a Bayesian Optimization-driven search to minimize the negative MAE based on the parameters presented in the table.

Property	**colsample_bytree**	**Learning rate**	**Max depth**	**Alpha**	**Lambda**	**Subsample**
Di (Å)	1.00	0.097	10	1.00	2.73	0.98
Df (Å)	0.84	0.119	10	1.00	0.0047	1.00
Dif (Å)	0.81	0.096	10	9.2×10^{-5}	2.02	1.00
V_{cell} (Å^3)	0.46	0.096	10	1.0×10^{-6}	0.27	1.00
Density (g/cm^3)	1.00	0.122	10	1.00	3.04	1.00
ASA (m^2/cm^3)	0.46	0.097	10	1.0×10^{-6}	0.27	1.00

Table 3 and Table 4 show the hyperparameters for the CGCNN and MEGNet respectively. Both models were trained on a batch size of 64 over 100 epochs. Throughout the 100 epochs, the best model was saved based on the iteration with the lowest average mean absolute error on the validation set.

Some notes regarding the CGCNN:

1. Number of maximum neighbours per atom while constructing the graph is set as 12
2. Cutoff radius for searching neighbours is set as 8 Å
3. The minimum distance required for the expanded distance matrix is set as 0

Some notes regarding MEGNet:

1. The embedding layer for the node attributes had a dimension of (N, 16);
2. The cutoff radius for graph representation was arbitrarily set at 4.0 Å;

Table 3: Optimizer settings and model hyperparameters for CGCNN (batch size of 64 used).

Parameter	**Value**
Optimizer Settings	
Optimizer	Adam
Learning Rate (lr)	0.01
Momentum	0.9
Weight Decay	1×10^{-6}
Model Hyperparameters	
Atom Feature Length ($atom\_fea\_len$)	64
Hidden Feature Length (h_fea_len)	512
Number of Convolution Layers (n_conv)	3
Number of Hidden Layers (n_h)	1

The concatenated model (CGCNN–LQG) concatenates the embeddings of the CGCNN and a neural network accepting LQG descriptors, giving a final embedding shape of (N, 640), where N is the

Table 4: Bond expansion and MEGNet model hyperparameters.

Parameter	**Value**
Bond Expansion Settings	
RBF Type	Gaussian
Initial Value	0.0
Final Value	5.0
Number of Centers	100
Width	0.5
MEGNet Model Hyperparameters	
Node Embedding Dimension ($dim\_node\_embedding$)	16
Edge Embedding Dimension ($dim\_edge\_embedding$)	100
State Embedding Dimension ($dim\_state\_embedding$)	2
Number of Blocks ($nblocks$)	3
Hidden Layer Sizes (Input)	(64, 32)
Hidden Layer Sizes (Conv)	(64, 64, 32)
Number of Set2Set Layers ($nlayers\_set2set$)	1
Number of Set2Set Iterations ($niters\_set2set$)	2
Hidden Layer Sizes (Output)	(32, 16)
Activation Function ($activation\_type$)	softplus2
Gaussian Width ($gauss\_width$)	0.5

specified batch size. The hyperparameters of the LQG neural network model up until the concatenation step are given in Table 5. Upon getting the concatenated tensor, a sequential regression head is used, in which it:

1. Accepts a 640 size tensor and passes it through a linear layer to shape it down to 128;
2. Performs batch normalization and then a ReLU transformation is performed;
3. Reshapes the 128 shape tensor down to 64; batch normalization and ReLU are performed in series again;
4. The final step is a linear layer that transforms the 64 shape tensor down to a tensor containing one element (the output). Softplus is used to ensure positive values in the output.

Table 5: Model architecture and hyperparameters for the neural network concatenated with a CGCNN (with the same parameters prior to the regression head in Table 3). The learning rate configuration was for the concatenated model - not the neural network isolated by itself.

Parameter	**Value**
Model Architecture	
Input Features	203
Number of Layers (n_layers)	10
Neurons per Layer	128
Dropout Rate	0.0
Training Configuration	
Optimizer	Adam
Initial Learning Rate (lr)	0.001
Learning Rate Decay	Cosine scheduler
Decay Steps	50
Minimum Learning Rate	0.00001