DEVELOP GRAPH CONVOLUTIONAL NETWORKS TO SCREEN ADSORBATES IN METAL-ORGANIC FRAMEWORKS BASED ON PORE LIMITING DIAMETERS

Sudheesh Kumar Ethirajan †
3^{rd} *yr. Ph.D. candidate*
Department of Chemical Engineering
University of California, Davis
Davis, CA 95616, USA
sethirajan@ucdavis.edu

Kun-Lin Wu †
4^{th} *yr. Ph.D. candidate*
Department of Chemical Engineering
University of California, Davis
Davis, CA 95616, USA
klwwu@ucdavis.edu

† contributed equally

1 INTRODUCTION

Metal-Organic Frameworks (MOFs) distinguish themselves as nano-porous materials, admired for their adaptable active sites, intricate 3-D porous environments, and experimental versatility. In the broad spectrum of MOF applications, the meticulous screening of adsorbates is paramount, with the Pore Limiting Diameter (PLD) emerging as a critical determinant for adsorbate accessibility in gas adsorption and separation processes. However, the computational complexity of computing PLD across the vast chemical design space of MOFs using tools like ZEO++ and Pore-Blazer necessitates a quest for more efficient screening methods.

In response, our project draws inspiration from the innovative work of Mehrdad Jalali et al. We utilize a Graph Convolutional Network (GCN) grounded in social network analysis to predict guest molecule accessibility in MOFs based on Pore Limiting Diameter (PLD) ranges. Beyond this novel approach, our investigation encompasses diverse model architectures, including a baseline Multilayer Perceptron (MLP), a conventional Graph Convolutional Network, and the Graph Sample and Aggregated (GraphSAGE) model. Hyperparameter optimization is conducted through Optuna, and we delve into the study of graph embeddings using t-SNE dimensionality reduction techniques.

This research aims to advance MOF screening methodologies, shedding light on the efficacy of various model architectures in predicting guest molecule accessibility within these intricate nano-porous structures. The subsequent sections detail the dataset curation, methodology, results, and implications of our investigation into the application of GCNs for predicting guest molecule accessibility in MOFs.

2 MOF DATA PREPARATION

Our exploration involves a dataset comprising 1988 Metal-Organic Frameworks (MOFs), derived from Mehrdad Jalali's research. Initially, the dataset, obtained from the paper's GitHub repository, faced issues with errors in SMILES strings. We meticulously cleaned the data, creating a refined dataset, `SMILES_METAL_1988_NoPLD.csv`. These MOFs are characterized by seven key features, including atomic properties and Organic Ligand SMILES representations. The initial six features undergo normalization and vectorization, leading to the computation of cosine similarity between MOF pairs.

Additionally, Morgan Fingerprints are generated for the Organic Ligand SMILES, and Tanimoto similarity is calculated. The structural relationships within MOFs are captured through the construction of an adjacency matrix, considering weighted averages of cosine and Tanimoto similarities. Pruning edges with a threshold of 0.9 refines connections, establishing a robust foundation for graph-based analyses. Importantly, this dataset, sourced from Jalali's paper, underpins our comprehensive exploration of MOF relationships and the replication of the MOFGalaxyNet paper's methodology. Building upon this, we computed a similarity matrix with varying weights using Tanimoto Similarity. Subsequently, we derived an adjacency matrix from this similarity matrix, with a subsequent edge elimination process executed to optimize the graph analysis. The results of this process were stored in the file `EdgesList_1988_0.9_alpha_0.9_omega.csv`, containing the nodes (MOF types) and edges (links) information for further analysis. This rigorous approach ensures the accuracy of our graph-based analysis and sets the stage for evaluating the performance of GCN models in predicting pore limiting diameters.

3 Methods

3.1 Graph Convolutional Network (GCN)

In our exploration of Metal-Organic Frameworks (MOFs), Graph Convolutional Networks (GCNs) serve as a crucial tool for capturing structural relationships within MOFs. Introduced by Kipf and Welling in 2016, GCNs are essential for analyzing graph-structured data, aligning seamlessly with our project on MOFs. By employing a layer-wise propagation scheme, GCNs effectively capture intricate structural patterns and relationships within MOFs. This is particularly relevant to our project's goal of predicting guest molecule accessibility based on Pore Limiting Diameter (PLD) ranges. Notably, our node features are mean-aggregated, further enhancing GCNs' capabilities in capturing nuanced MOF structures. Additionally, the incorporation of social network analysis enriches the GCN methodology by providing insights into the MOF chemical design space through a network lens.

3.2 GraphSAGE (Graph Sample and Aggregated)

In our MOF exploration, GraphSAGE, introduced by Hamilton et al. in 2017, is pivotal for its inductive learning framework. Tailored to our project, GraphSAGE efficiently generalizes to unseen MOFs during inference. By incorporating neighbor sampling and aggregation, GraphSAGE learns versatile embeddings, capturing both node-level and graph-level patterns within MOF structures. This aligns seamlessly with our project's goal of efficient screening methods for MOFs, with added context that our node features are mean-aggregated. The utilization of social network analysis further enriches the GraphSAGE approach by providing a holistic view of the 1988 MOFs chemical design space.

3.3 Multi-Layer Perceptron (MLP) as Baseline Model

Within MOF screening, the Multi-Layer Perceptron (MLP) holds a foundational role as a baseline model. Despite lacking explicit consideration of graph structure, MLPs provide a straightforward benchmark for evaluating more sophisticated graph-based models like GCN and GraphSAGE. In our project, the choice of an MLP as a baseline allows for a direct comparison, particularly highlighting the impact of mean-aggregated node features. This comparison aids in assessing the effectiveness of graph neural networks in capturing nuanced dependencies within MOF data, with the supplementary insight gained from social network analysis enriching the exploration of MOF chemical design space.

4 Results

4.1 MOF-Social Network

We constructed the MOFGalaxyNet, a network representation of our cleaned data with 1988 MOF samples. The dataset underwent a sparsification process, where only a limited number of MOF labels were displayed, determined by a threshold applied during data cleaning. The network, as shown in Figure 1, was then visualized and analyzed using Gephi software, utilizing the OpenOrd layout for enhanced visualization.

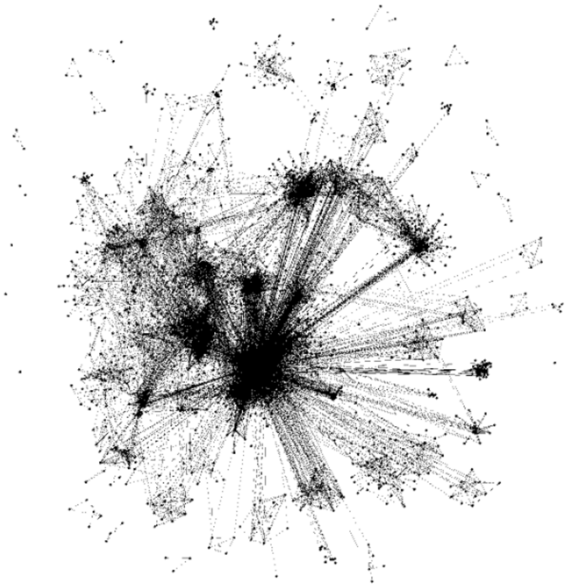


Figure 1: Our MOFGalaxyNet built by Gephi software considering 1988 MOFs samples

4.2 Optuna: Hyper-parameters Tuning

In our experimentation phase, we employed Optuna for hyperparameter tuning across three distinct models: Multi-Layer Perceptron (MLP), Graph Convolutional Network (GCN), and GraphSAGE. The aim was to optimize the models' performance by systematically tuning key hyperparameters. The training dataset consisted of 80% (1590) of the 1988 MOFs, with the remaining 20% (398) allocated for validation. Our hyperparameter tuning were conducted sequentially in two studies (see figure 2).

4.2.1 Study 1: Neurons & Hidden layers Tuning

The first study with 15 trials primarily focused on three aspects: the number of hidden layers (with accepted values of 1, 2, or 3), the number of neurons in each hidden layer (chosen from [16, 32, 64]), and the dropout probability in each layer. Notably, the dropout probability varied in a logarithmic scale, ranging from 1e-3 to 1e-1. To ensure a fair comparison, we maintained certain parameters constant, such as the optimizer (Adam), batch size (16), learning rate (0.02), learning rate step (100), learning decay rate (0.99), and the total number of epochs (5000) in our first study across different models. This standardized configuration ensured a consistent baseline for evaluation.

4.2.2 Study 2: Learning Parameters Tuning

From our first study, the best model was chosen for the second study. The second study was with 20 trials and primarily focused on finding the learning parameters for each model: the mini-batch size varied from 4 to 256, learning rate varied in a logarithmic scale, ranging from 1e-4 to 1e-1, scheduler step varied from 50 to 500 epochs, and the learning rate decay varied from 0.9 to 1.0.

This comprehensive exploration of hyperparameter space (summarized in the below table 1) aimed to identify the most effective configurations for each model, ensuring optimal performance on the task of predicting guest molecule accessibility based on Pore Limiting Diameter (PLD) ranges within MOFs.

4.3 Comparison between MLP, GCN, and GraphSAGE models

Utilizing the insights gained from Optuna, we compared the metrics (as shown in the appendix, Figure 5) across the best architecture for each model. We evaluated each of the model with optimizer

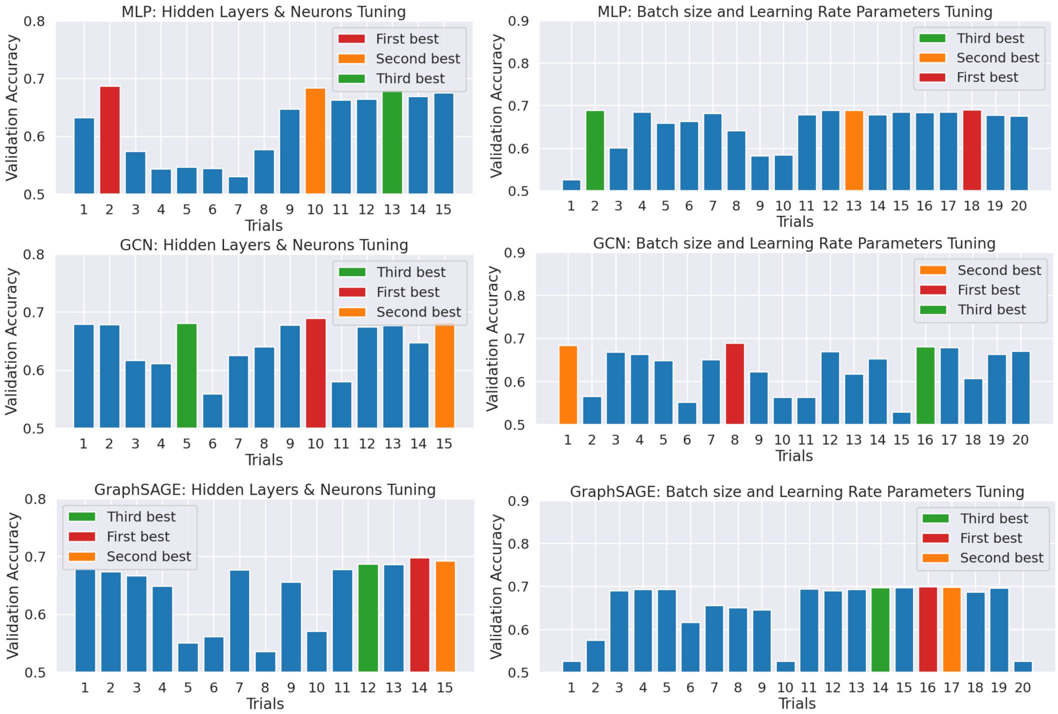


Figure 2: Optuna Hyperparameter Tuning

Model	Hidden layers	Hidden neurons	Dropout probability	Batch size	Learning rate	Epochs step	Decay rate
MLP	3	[32,64,64]	2.499e-03	228	5.281e-03	297	9.646e-01
GCN	3	[32,64,16]	2.848e-03	66	9.115e-03	438	9.186e-01
GraphSAGE	3	[32,64,64]	1.349e-03	202	1.499e-02	290	9.029e-01

Table 1: Best Hyper-parameters from Optuna experimentation

(Adam), batch size (202), learning rate (1.499e-02), learning rate step (500), learning decay rate (9.029e-01), and the total number of epochs (10000). Based on the hyperparameters tested, we find all the three models to have about 70% accuracy on the validation dataset. The confusion matrix is shown in figure 3 for all the three models.

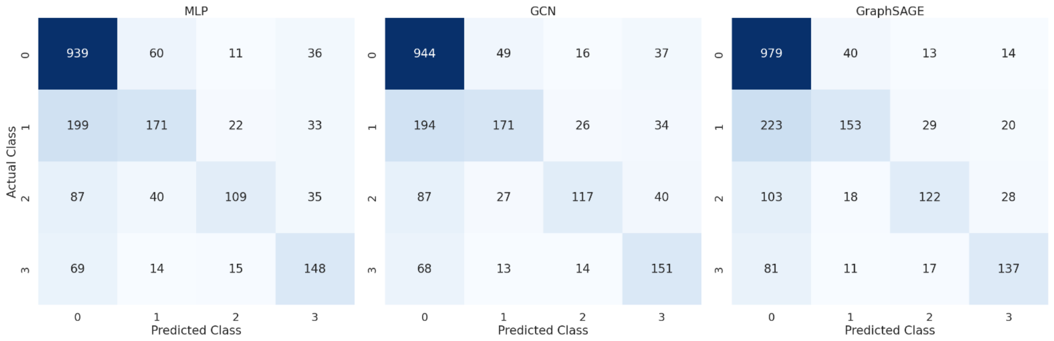


Figure 3: Confusion matrices for different models

We expected the GraphSAGE and GCN models to outperform the MLP baseline model by a considerable margin due to their sophistication but our experimentation indicates that the graph models are only marginally better for this dataset. However, one can expect to see considerable improvement with more advanced graph architectures such as Graph Attention Network (GAT), introduced by Veličković et al. in 2018.

4.4 Node Embeddings

Node embeddings serve as crucial representations in Graph Convolutional Networks (GCNs), encapsulating essential features of nodes within a graph. In the context of our MOF exploration, these embeddings are pivotal for the GCN to comprehend and predict relationships within the complex chemical design space of Metal-Organic Frameworks. They provide a compact yet comprehensive representation of each MOF, enabling the model to discern patterns and similarities crucial for accurate predictions. To enhance the interpretability of these embeddings, t-Distributed Stochastic Neighbor Embedding (t-SNE) is employed. t-SNE, a dimensionality reduction technique, transforms high-dimensional node embeddings into a lower-dimensional space while preserving local relationships.

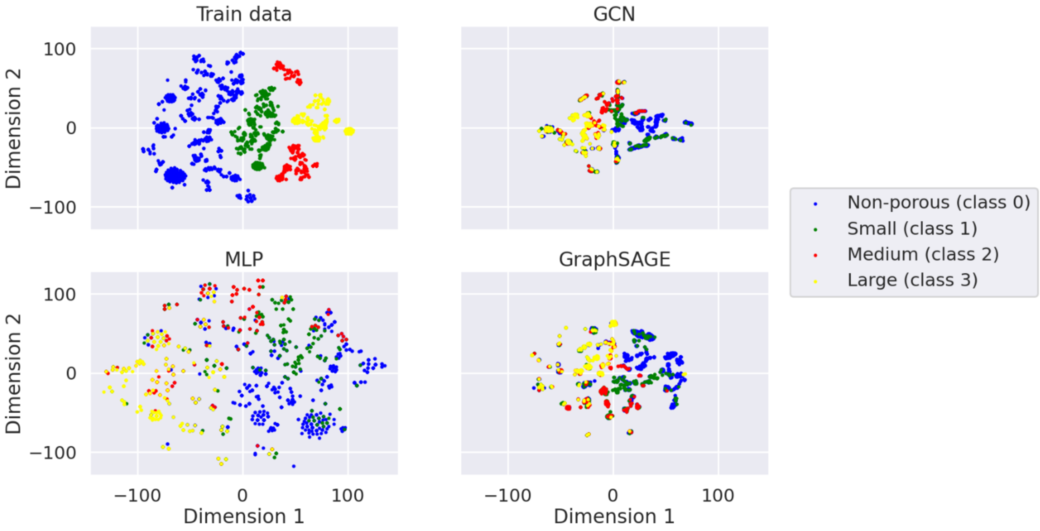


Figure 4: t-SNE visualization of node features & learned embeddings of different models

In our MOF study, we employed t-SNE to visualize complex MOF datasets in two dimensions, facilitating a deeper understanding of underlying patterns and relationships. We find from figure 4 that all the models considered here are having overlapping points that corresponds to different classes. This provides a pictorial description into why all the models accuracy are not very high.

5 Conclusion & Future directions

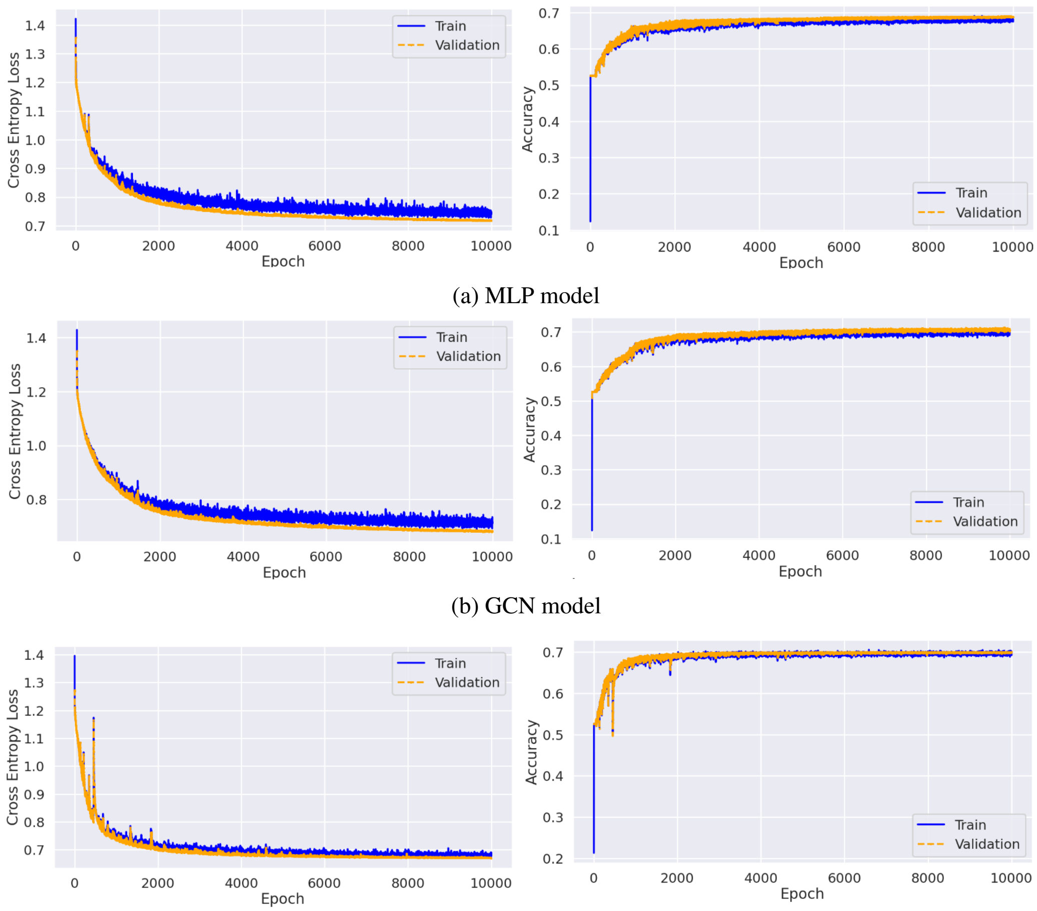
In this project, we successfully implemented an innovative Graph Convolutional Network (GCN) approach to predict guest accessibility in Metal-Organic Frameworks (MOFs). This method holds the potential to significantly streamline high-throughput screening, expediting the development of high-performance MOFs for a diverse range of host–guest interaction applications. The methodology involves two key steps. First, the establishment of MOFGalaxyNet, a social network built on MOFs' similarities, employs social network analysis to yield valuable insights into MOF properties. The second step entails utilizing the GCN model to predict guest accessibility across four distinct categories in MOFs. This pioneering approach accelerates the analysis of MOF structures and improves screening efficiency for various design criteria, extending its utility to predict additional properties such as stability and methane storage.

Furthermore, adopting the strategy of constructing graph structures for each MOF from its Crystallographic Information File (CIF) stands out as a superior method according to existing literature. We posit that this technique could enhance the quality of our input features, thereby contributing to the overall effectiveness of our predictive model. Verification of this hypothesis is slated for exploration in our future work.

6 References

1. Jalali, M., Wonanke, A.D.D. & Wöll, C. MOFGalaxyNet: a social network analysis for predicting guest accessibility in metal–organic frameworks utilizing graph convolutional networks. J Cheminform 15, 94 (2023).
2. Kipf, T. N., & Welling, M. (2017). Semi-Supervised Classification with Graph Convolutional Networks. arXiv [Cs.LG].
3. Hamilton, W. L., Ying, R., & Leskovec, J. (2018). Inductive Representation Learning on Large Graphs. arXiv [Cs.SI].
4. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph Attention Networks. arXiv [Stat.ML].

7 Appendix



(a) MLP model

(b) GCN model

(c) GraphSAGE model

Figure 5: Comparison between different model architectures