Topological connection between zeolites and their potential interzeolite transformation

Jose Rebolledo-Oyarce
Department of Chemical and Biomolecular Engineering, University of Notre Dame
Notre Dame, Indiana 46556, United States
jrebolle@nd.edu

Abstract

Zeolites have a high industrial value due to their multiple applications in the adsorption of compounds or as catalysts.However, obtaining large zeolites is highly expensive, which is why the interconversion between zeolites is critical for the viability of these processes. This interconversion does not define descriptors that allow relating the zeolites, for which Variational Graph Autoencoding was applied to combine the relationship between the different atoms that make up the zeolite and the different descriptors that it is possible to extract from these atoms.The corresponding latent variable was used to cluster the different graphs that represent the different zeolites, obtaining good interconversion predictions for different zeolites, such as BEA and FAU and it is possible to extend this analysis to other organized structures such as the Metalorganic frameworks (MOFs)

1 Introduction

Zeolites are crystalline microporous solids with diverse framework structures and constructed primaly by $SiO_4{}^4$ and $AlO_4{}^5$ tetrahedral units. Zeolites are widely used in adsorption, catalysis, and ion-exchange processes due to their kinetic stability [1, 2].

Zeolites are commonly synthesized by hydrothermal treatment of amorphous aluminosilicate gel (Si and Al source) in presence of structure directing agents (SDA) whether organic or inorganic [1]. SDA are neccesary to arrange Si and Al in higly organized system with different diverse porous size and multiples type of rings size and for that reason, the cost and the environmental burden increases at the moment of synthesized large-scale zeolite, which are the most used in the industrial process [1].

Much effort has been devoted to the developmoent of SDA-free synthesis protocols to decrease such costs as well the emissions of toxic species in gaseous and water steams generated during the synthesis or subsequent treatments required to remove SDA from zeolite due to it is not necessary in any interesting industrial application. The Assembly-Disassembly-Organization-Reassembly (ADOR) mechanism or seed-assisted hydrothermal synthesi are promising methods to synthesized zeolite without SDA or with low quantity of SDA [3], however these method are limited in application due to the incorporation of other agents and can not applying in all zeolite framework.

Due to these limitations, in the industrial application, where it is necessary large-scale zeolites, these system are produced by the densification or rearrangements of other structures, a process known as zeolite interconversion or interzeolite transformation (see Fig. 1) [1, 4], that means, take small-scale zeolites that are easy to synthesize and reorganize them into large-scale zeolites. However, although the interzeolite transformation is a widely used and versatile approach, its mechanism remains elusive and which zeolite can be transform to other and why, it is still an open question.

A common assumption is to use the number of commom composite building units (CBUs) and framework densities (ΔFD) as descriptors to predict which zeolites can be related [5]. However,

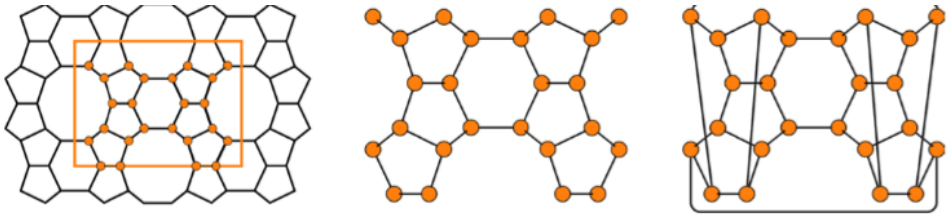

Figure 1: Schematic representation of a zeolite as graph and multigraph or crystal graph, extrated from Schwalbe-Koda et al. [5]

Schwalbe-Koda and co-workers [5] explored 70,000 articles and found not clear correlation between these descriptors and interzeolite transformation. To do this, they took advantage of the possibility of representing a zeolite as a graph (see Fig. 1) and explored new descriptors such as graph similarity D-measure or Smooth Overlap of Atomic Position (SOAP) distance to predict the interzeolite interconversion.

Nevertheless, in their approach they only consider the zeolite topology as another feature but they did not consider how the different chemical species interact with other, that means, they did not consider directly the present of neighbors around each atom. With that in mind, this work, we develop a systematic way to transform atomic representation of zeolite obtaining by Database of Zeolite Structures (IZA database) to periodic graphs, i.e., the classical graph (unit cell) is modified to satisfy periodic boundary conditions by looping bonds back into the unit cell. Using the graph representation is possible to obtaining the interaction between the different species using the adjacency matrix of each graph and combining with some atomic descriptors as SOAP [6].

The combination of these features with the adjacency matrix was using the Variational Graph Autoencoder (VGAE) propsed by Kipf and Welling [7] which it is an unsupervised method to classify graphs that combines the unsupervised method knowning as autoencoder proposed by Kingma and Welling [8] and the graph convolutional networks proposed by Kipf and Welling [9] to obtaining a final latent variable with works as a pseudo-PCA parameters resulting as perfect parameter to clustering the different frameworks in small groups which had good result to predict the interconversion between different zeolites.

2 Methodology

2.1 Extracting Periodic Graph from a Zeolite Framework

The biggest challenge in these types of problems is to systematically extract the information from any zeolite regardless of the number of atoms, the number of rings, or the geometry of the zeolite itself. For that reason, the first step in this project was to generate a database with all zeolites available in the Database of Zeolite Structure (IZA) that contain the name of the zeolite (standardized name consisting of three letters) and the geometric structure of each zeolite.

With this database it is possible to obtain the number of Silicon, Aluminum, and Oxygen atoms available in each zeolite and their respective Cartesian positions. In the zeolite community it is a common practice that in any topological analysis only the positions of silicones and aluminums are considered, since the active sites are eventually located on these atoms in the activation process [5]. With these idea in mind, and taking advantage of the fact that each silicon and aluminum is going to have 4 bonds and each oxygen is going to have 2 bonds, a simple code can determine the neighbors of each atom and all the edges (bonds) between each atom and at the same time ensure the periodicity of our system, i.e., ensure that we generate closed graphs, as can be seen in the graphs generated in Figures 2 and 3.

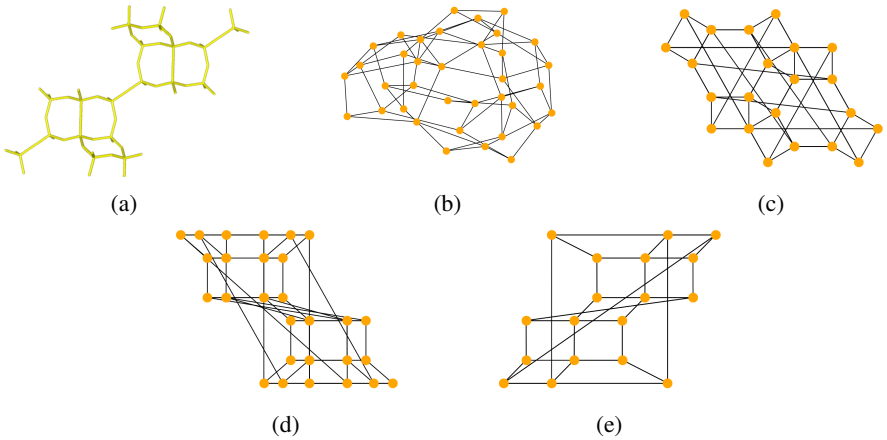

Figure 2: Schematic representation of Chabazite (CHA): (a) Molecular representation from VESTA, (b) zeolite as graph without position of nodes, c) frontal view of zeolite as graph including position of nodes, d) side view of zeolite as graph including position of nodes, and e) top view of zeolite as graph including position of nodes

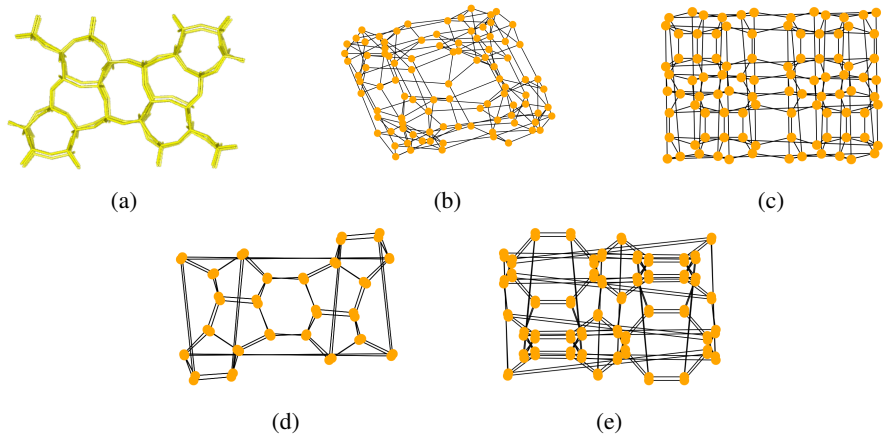

Figure 3: Schematic representation of Zeolite Socony Mobil5 (MFI): (a) Molecular representation from VESTA, (b) zeolite as graph without position of nodes, c) frontal view of zeolite as graph including position of nodes, d) side view of zeolite as graph including position of nodes, and e) top view of zeolite as graph including position of nodes

Using these graphs and the Networkx package [10], it is possible to calculated the adjacency matrix of each graph. In addition to the graph representation and using the DScribe package [6], it is possible to calculate the SOAP descriptors of each zeolite per atom (node) that constitute each graph applying the following expression:

$$p_{nn'l}^{Z_1Z_2} = \pi \sqrt{\frac{8}{2l+1}} \sum_m c_{nlm}^{Z_1} {}^* c_{n'lm}^{Z_2} \tag{1}$$

where $p_{nn'l}^{Z_1Z_2}$ is the partial power spectrum vector. n and n' are indices for the different radial basis functions up to n_{max}, l is the angular degree of the spherical harmonics up to l_{max} and Z_1 and Z_2 are atomic species.

The coefficients c_{nlm}^{Z} are defined as the following inner products:

$$c^Z_{nlm} = \iiint_{\mathcal{R}^3} \mathrm{d}V g_n(r) Y_{lm}(\theta, \phi) \rho^Z(\mathbf{r}) \tag{2}$$

where $\rho^Z(r)$ is the gaussian smoothed atomic density for atoms with atomic number Z defined as:

$$\rho^Z(\mathbf{r}) = \sum_i^{|Z_i|} e^{-1/2\sigma^2 |\mathbf{r}-\mathbf{R}_i|^2} \tag{3}$$

$Y_{lm}(\theta, \phi)$ are the real spherical harmonics, and $g_n(r)$ is the radial basis function. For the radial degree of freedom the selection of the basis function $g_n(r)$is not as trivial and multiple approaches may be used, but in this case it is going to use spherical gaussian type orbitals as radial basis functions, as they allow much faster analytic computation [6]

2.2 Variational Graph Autoencoding (VGAE)

Following the formulation proposed by Kifp and Welling [7], VGAE correspoding to two Graph Convolutional Network (GCN) was used to encoding the adjacency matrix with the feature matrix where the first layer is:

$$H^{(1)} = \mathrm{ReLU}\left(\widetilde{D}^{-\frac{1}{2}} \cdot \hat{A}\widetilde{D}^{-\frac{1}{2}} H^{(0)} W^{(0)} + b^{(0)}\right) \tag{4}$$

where $\hat{A} = A + I_N$ is the adjacency matrix of the undirected graph G with added self-connections. I_N is the identity matrix. $D_{ii} = \sum_j \hat{A}_{ij}$ and $W^{(0)}$ is a layer-specific trainable weight matrix. And $H^{(0)}$ corresponds to the feature matrix. And $b^{(0)}$ corresponds to the bias term.

And the second layer correspond to two GCNs that share the same $H^{(1)}$ and normalized adjacency matrix ($\widetilde{A} = \widetilde{D}^{-\frac{1}{2}} \cdot \hat{A}\widetilde{D}^{-\frac{1}{2}}$) to construct the latent variable Z. These two GCNs are used to calculated the mean ($\mu = \mathrm{GCN}_\mu$) and the standard deviation of the GCN method ($\log \sigma = \mathrm{GCN}_\sigma$). These two GCNs are expressed as following:

$$\mu = \left(\widetilde{D}^{-\frac{1}{2}} \cdot \hat{A}\widetilde{D}^{-\frac{1}{2}} H^{(1)} W^{(1)} +^{(1)}\right) \tag{5}$$

$$\log \sigma = \left(\widetilde{D}^{-\frac{1}{2}} \cdot \hat{A}\widetilde{D}^{-\frac{1}{2}} H^{(1)} W^{(1)} +^{(2)}\right) \tag{6}$$

And the latent variable is define as following:

$$Z = \mu + \epsilon \cdot \sigma \tag{7}$$

where $\epsilon \sim \mathcal{N}(0, 1)$. After this encoding process, it is necessary to apply an decoding process to reconstruct $\widetilde{A}$ and perform the optimization process. For that, the simplest way to reconstruct the modified adjacency matrix ($\widetilde{A}$) is using the following expression:

$$\hat{A}_{\text{reconstructed}} = \sigma\left(Z \cdot Z^T\right) \tag{8}$$

where σ is the logistic sigmoid function. Finally, in the optimization process, the loss function with which better results were obtained is Binary Cross-Entropy (BCE):

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\left(\hat{A}_{\text{true}} \cdot \log \hat{A}_{\text{reconstructed}} + (1 - \hat{A}_{\text{true}}) \cdot \log(1 - \hat{A}_{\text{reconstructed}})\right) \tag{9}$$

where $\widetilde{A}_{\text{true}}$ and $\widetilde{A}_{\text{reconstructed}}$ were previously reshaped to have a single dimension. This process can be summarized in the following Figure:

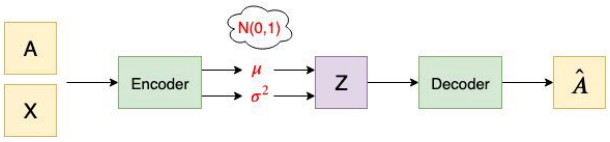

Figure 4: Schematic Representation of Variational Graph Autoenconding

2.3 Treatment of Latent Variable and K-Clustering

The latent variables are usually called as pseudo-PCA of the graph representation. However in this problem, there are graph with different amount of nodes so it is more complicated to compared directly the resulting latent variables. For that, a proposed idea is simple pooling layer of these latent variables using the average as the pooling rule to obtaining two unique variable per graph.

With these two-dimensional vector per graph, it is possible to apply a k-clustering method, which consisted in minimize the distance between each point and the cluster centroid that corresponds to that point and it can be formulated as following

$$\arg\min \sum_{i=1}^{k} \sum_{x \in S_i} ||x - C_i||^2 \tag{10}$$

where S_i is the set of point in the cluster i.

3 Results and Discussion

We test this approach using a dataset extracted from the zeolite topology applying a combination of MAZE and Networkx package [10, 11] resulting in the extraction of the adjacency matrix of each zeolite framework with dimension $N \times N$ where N is the number of atoms that constitute each zeolite and vary in 5 and 1440. And the feature matrix of each graph as generate obtaining the SOAP vector of each node using the Eq. 1 which has dimension of $N \times M$ where N is the number of node of each graph and M corresponding to the number of features generated that in this case is equal to 252.

It is important to mention that due to the small amount of framework that share the same amount of nodes, it was impossible to generate two type of dataset (training and validation dataset) and at the same time this method can not deal with dataset that it has not seen before in training which leads a total of 254 graph that constitute this dataset.

For that reason, as a proof of concept of the Variation Graph Autoencoding method, QM9 dataset was used to testing the implemented method to ensure that it is possible to obtaining the latent variable (Z) consistently independently of the dataset and it is possible to applying the training result in a validation dataset. The evolution of this proof of concept is presented in the Fig. 5.

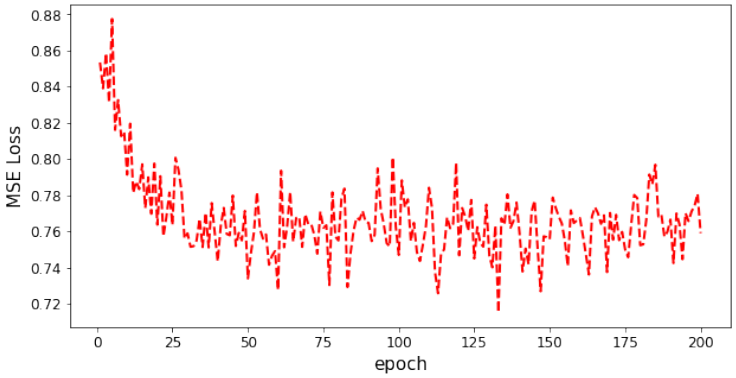

Figure 5: Evolution of the loss function applying VGAE using 5000 points from the QM9 dataset

With that consideration in mind, and after applying the Variational Graph Autoencoder process (see Fig. 4) of each set of dataset generated it is possible to obtain individual evolution of the loss function along each epoch (see Fig. 6) and an accumulative loss function along the variation of epoch (see Fig. 7). In Fig. 6 is possible to appreciate that the system tends to convergence after 50 epochs. At the same time, if we see the Fig. 7 is possible to say that the whole dataset convergence after 75 epochs but the result tends to oscillate due to the presence of the term ϵ in the Eq. 7.

At the same time, it is important to mention that due to the lack of labels any classification method applying in graph presented relative low accuracy (around 40 to 80%) [12, 9], therefore, high values are expected in the objective function and there are consistent with the values found using a well known dataset as QM9.

Figure 6: Example of the Evolution of Loss function of one set of graph that have the same amount of nodes

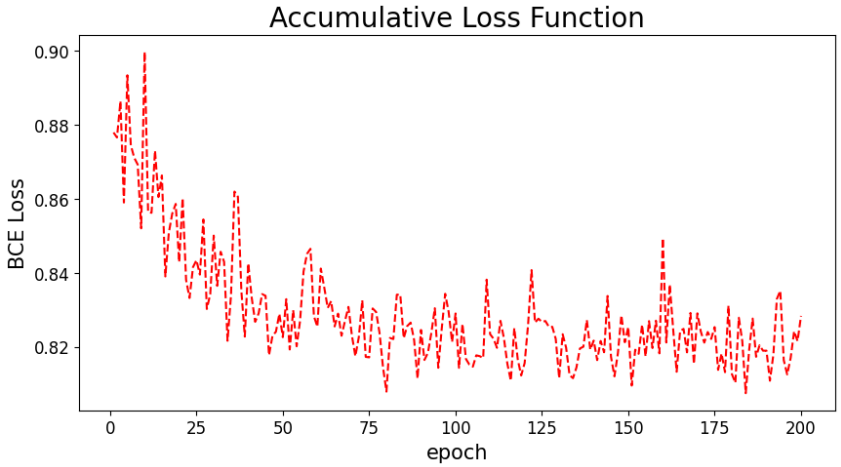

Figure 7: Accumulative Evolution of Loss function of zeolites frameworks

After the application of VGAE in the different zeolite framework and using a pooling layer that calculated the average of in the latent variable resulting in a two dimensional vector per graph which can be plotting in a simple graph as it is shown in Fig. 8.

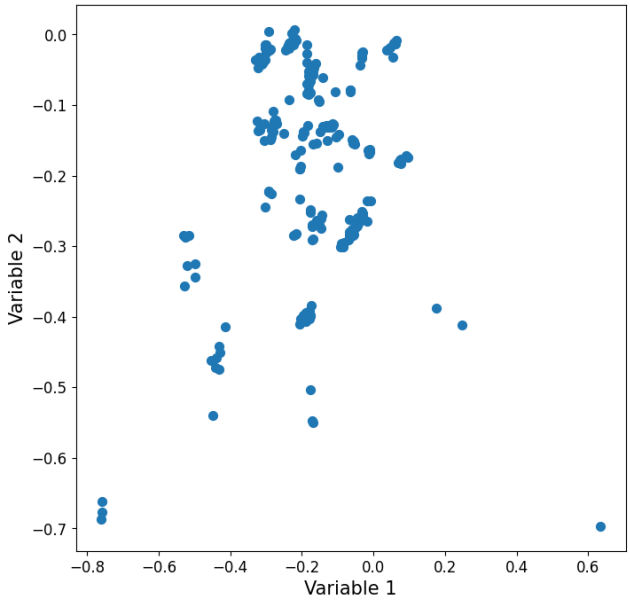

Figure 8: Visualization of Resulting Pooling Latent Variables where each point represent a one zeolite

Due to the lack of dispersion of these point a Elbow analysis is necessary to determine the amount of clusters need to minimize the total distance between each point and its respective centroid. The idea of the Elbow analysis is determine the minima amount of clusters necessary to minimize the total distance between each point and its respective centroid and the minima does not change if the amount of cluster increases. In this analysis it is possible to appreciate that 15 clusters are necessary to have a change in the sum of the total distance equal to 10^{-2}.

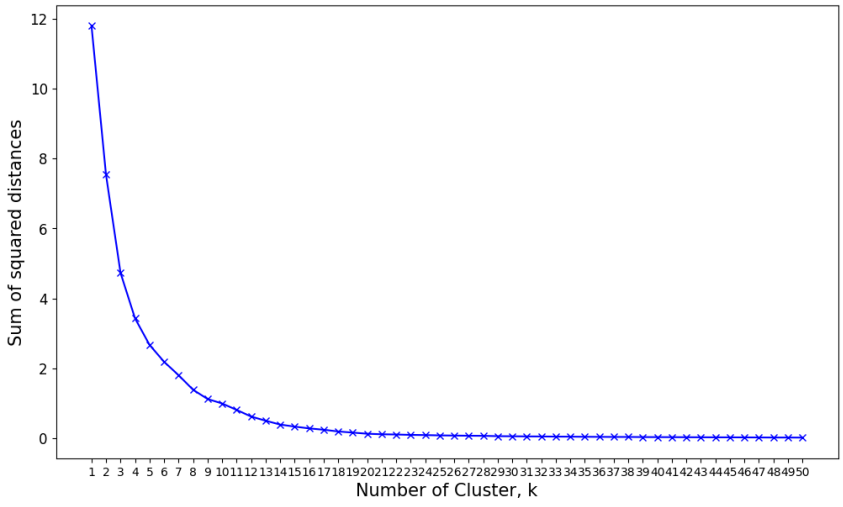

Figure 9: Elbow analysis of Pooling latent variable

Using the same k-Means analysis applied in the Elbow analysis it is possible to obtaining the label of each zeolite framework as plotting the resulting labels in the Fig. 10.

It is important to mention that independent that this approach is an purely unsupervised machine learning model and does not include any label unknown from the literature it is possible to important relation between zeolites. For example, FAU and BEA are in the same cluster which can be

interpreted as the possible to interconverted these two zeolites as shown some experimental results [1, 5]. At the same time, recently, Liu and coworkers [13] were carried out the interconversion of MRE to EUO which, in this model, was labelled in the same cluster. And the same case occur between GIS and ANA as shows by de Lima and coworkers [14].

However, this approach has problem to predicted relation between other well known zeolites as it is the case of CHA and FAU [1] which can be due to the high concetration of the resulted point obtained from the VGAE and the pooling layer which can be solved by the implementation of a semisupervised model with can complemented the unsupervised method as the Teacher-Student approach.

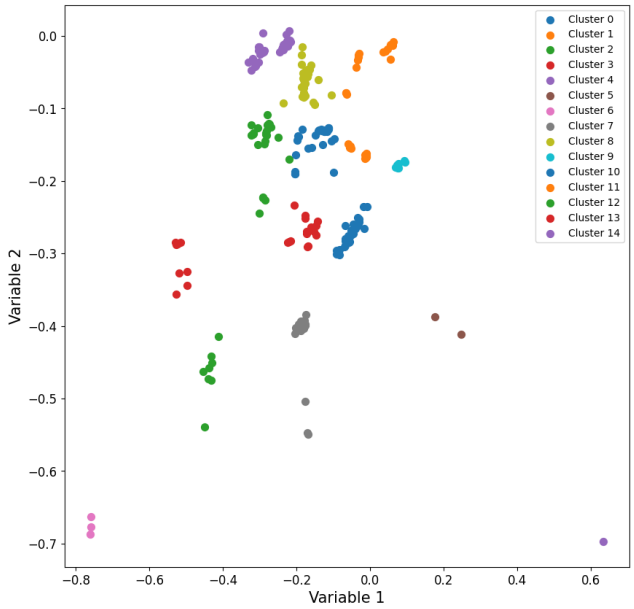

Figure 10: k-Clustering of each graph using k = 15 as the number of clustering resulting from the Elbow analysis

4 Conclusion

We have been define a systematic method to extract respective graph a any zeolite topology and with that representation it is possible to generate any feature matrix. Using this systematic approach is possible to apply any machine learning method to predict a well define label or an undefined label as is shown in this work incorporate the relation that exists between the different atoms that make up the zeolite using the adjacency matrix.

At the same time, VGAE is a simple approach to obtain compact variable as latent variables (Z) and it was tested using a well know dataset and a constructed dataset that come from the Zeolite database. These latent variable can be used to categorized our graph using different approach as the K-clustering methods proposed in the work. In this clustering process, it is possible to predict multiple type of relation between different zeolite which have been proof experimentally. However, this model has limitation in its ability to predict the entire relation between the different zeolite due to the high concentration of the resulting point. Due to that, it is important to include a complemented semisupervised model that can incorporated a few well knowing label and maximize the accuracy of the model as is presented in a Teacher-Student approach which consisted in two model: one unsupervised model and one supervised model with only a few amount of label and transfer that knowledge between the two models.

Finally, trough this work it is possible to appreciate the extraction of data from a graph, how this information can be transferred to a graph neural network and predict interesting parameters for the industry or from an experimental point of view as we saw in class. At the same time, it is possible to appreciate the limitations of these methods due to the lack of data and through the convolution of the

data it is possible to analyze and compare graphs of different sizes and find relationships between them through different methods such as k-clustering.

References

[1] Sarika Goel, Stacey I. Zones, and Enrique Iglesia. Synthesis of zeolites via interzeolite transformations without organic structure-directing agents. *Chemistry of Materials*, 27(6):2056–2066, 2015.

[2] Pablo Del Campo, Cristina Martínez, and Avelino Corma. Activation and conversion of alkanes in the confined space of zeolite-type materials. *Chem. Soc. Rev.*, 50(15):8511–8595, August 2021.

[3] Wieslaw J. Roth, Petr Nachtigall, Russell E. Morris, Paul S. Wheatley, Valerie R. Seymour, Sharon E. Ashbrook, Pavla Chlubná, Lukáš Grajciar, Miroslav Položij, Arnošt Zukal, Oleksiy Shvets, and Jiří Čejka. A family of zeolites with controlled pore size prepared using a top-down method. *Nature Chemistry*, 5(7):628–633, June 2013.

[4] Monica J. Mendoza-Castro, Erika De Oliveira-Jardim, Nelcari-Trinidad Ramírez-Marquez, Carlos-Alexander Trujillo, Noemi Linares, and Javier García-Martínez. Hierarchical catalysts prepared by interzeolite transformation. *Journal of the American Chemical Society*, 144(11):5163–5171, 2022.

[5] Daniel Schwalbe-Koda, Zach Jensen, Elsa Olivetti, and Rafael Gómez-Bombarelli. Graph similarity drives zeolite diffusionless transformations and intergrowth. *Nature Materials*, 18(11):1177–1181, October 2019.

[6] Lauri Himanen, Marc O.J. Jäger, Eiaki V. Morooka, Filippo Federici Canova, Yashasvi S. Ranawat, David Z. Gao, Patrick Rinke, and Adam S. Foster. DScribe: Library of descriptors for machine learning in materials science. *Computer Physics Communications*, 247:106949, February 2020.

[7] Thomas N. Kipf and Max Welling. Variational graph auto-encoders, 2016.

[8] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2013.

[9] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks, 2016.

[10] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using networkx. In Gaël Varoquaux, Travis Vaught, and Jarrod Millman, editors, *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, Pasadena, CA USA, 2008.

[11] Dexter D. Antonio, Jiawei Guo, Sam J. Holton, and Ambarish R. Kulkarni. Simplifying computational workflows with the multiscale atomic zeolite simulation environment (maze). *SoftwareX*, 16:100797, 2021.

[12] Fan-Yun Sun, Jordan Hoffmann, Vikas Verma, and Jian Tang. Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization, 2019.

[13] Wen Liu, Pengfei Wei, Junjie Li, Yanan Wang, Shirui Xu, Zhiqiang Yang, Xuebin Liu, Longya Xu, Xiujie Li, and Xiangxue Zhu. Inter-zeolite transformation from *MRE to EUO: A new synthesis route for EUO zeolite. *Catalysis Today*, 405-406:321–328, December 2022.

[14] Renata C. F. de Lima, Daniele da Silva Oliveira, and Sibele B. C. Pergher. Interzeolitic transformation of clinoptilolite into GIS and LTA zeolite. *Minerals*, 11(12):1313, November 2021.