Prediction of Interaction Energies of Encoded Molecules via Graph Neural Networks

Utkarsh Saraswat, Technical University of Munich

***Abstract*—Machine learning models have found extensive applications in scientific research. A particular kind of Artificial Neural Network (ANN) known as graph neural network (GNN) has recently emerged as a new technique to learn complex phenomena exhibiting graph-like properties. In the last decade, there has been a huge surge in research papers using GNN and deep neural networks to predict and build theories on molecular properties. However, while building any ML model, feature selection is an important aspect. In this study, we are trying to analyze various GNN approaches used to predict energies of weak interactions between molecules and molecular aggregates and come up with a strategy to select training features accordingly. Our study is largely qualitative with some quantitative aspects such as the accuracy of the prediction model, size of training data, number of features in the training data, and computational time being the key parameters. For qualitative evaluation, we have classified the phenomena based on the structure, environment, and complexity of the interaction.**

***Index Terms*—Interaction energy, Solvation energy, Adsorption, Molecular aggregates, Graph neural network**

I. Introduction

The interaction energy of two chemical entities is widely calculated using Density Functional Theory (DFT), and Molecular Dynamics (MD) simulations [1]. These methods are accurate but often computationally expensive [2]. Most of the ML models have shown more accurate and faster results than ab-initio simulations [2] [3]. However, the conventional ML models face several challenges in representing joint structures of molecules spanning through all the atoms [3]. End to end ML approach is also difficult to transfer and interpret, and often overfits the training data. These drawbacks are very well handled by Graph Neural Networks(GNN), which are a special class of ANN incorporating graphical encoding of features. Using GNN, the features of the dataset or phenomena to study are encoded as graph G(V,E) with node features and edge features. In the case of predicting interaction energy in molecules, they are especially effective due to their ability to incorporate the intrinsic properties of molecules directly into the features. Often, the edge features capture bond properties like bond type, conjugation, and the node features capture atomic properties like atomic mass, partial charge, etc within a molecule [1] [2]. But in some cases, node features may not represent an atom but an entire molecule. In most studies, GNN was accompanied by Message Passing Neural Network (MPNN) [2] and in some cases, variations of GNN such as Graph convolution network and Graph Activation network have also been used [4]

We classify our study into two broad categories, 1) Interaction energies of standard solute-solvent, where an individual molecule is surrounded by a homogeneous medium consisting of a different molecular species, and 2) Interaction of a molecule with a larger, macroscopic aggregate of a different species. In case one, we have included studies that either predict solvation energy or other related properties like activity coefficient. Within this case, we try to understand how i) Elements present in the solvent-solute system, ii) Molecular structure, and iii) the Nature of interaction in (ionic solution, and organic solution) affect the calculation of solvation energy and related parameters. In case two, our studies include organo-metallic interactions and adsorption, where adsorbates are either gas molecules or organic compounds. The comparison is made between i) the size of the adsorbate, ii) the Structure and orientation of the metal crystal iii) how a group of atoms contribute to the adsorption energy as compared to a single atom. In both cases, we distinguish between the types of GNN architecture applied in each case. We have tried to build the complexity of the study incrementally, starting with a simple molecule with a simple GNN implementation to molecular aggregates with a complex GNN design.

II. Discussions

A. *Prediction of solvation energy and molecular properties in solutions*

Solvation energy is the measures of total energy released when a solute molecule dissolves in a liquid solvent [1]. Our first study by Ramin Ansari, and Amirata Ghorbani [2] involves the calculation of solvation energy in non-aqueous solutions. Data from SMILES representation of 5952 solute-solvent pairs has been used to build features. The training data was obtained from experimental values of the partition coefficient of 657 organic solvents and 142 unique solutes. The partition coefficient is defined as the ratio of the concentration of a substance in one medium or phase to the concentration in a second phase [5], and it was used to calculate the solvation energy. To convert the molecule into a graph, there were 8 node features, namely i) categorical representation of the type of atom (Carbon, Nitrogen, Oxygen, Fluorine. Phosphorus, Sulfur, Chlorine. Bromine. Iodine, etc), ii) formal charge of the atom (integer value) iii) categorical hybridization of the atom iv) presence of Hydrogen bond (boolean) v) if an atom belongs to an aromatic ring (bool) vi) number of bonded neighbors (degree of graph node) vii) number of Hydrogen bonded to an atom and viii) partial charge. The edge feature consisted of 4 values which were i) categorical bond type, ii) if the pair of atoms in the node are in the same ring (bool), iii) conjugated or not, and iv) categorical stereochemical property. Molecular graphs were constructed separately for solvents and the solute. The comparison was made between two GNN architechture, Graph attention network (GAN) and MPNN. After the data transformation by GNN or MPNN, the vectors generated from solvent and solute were concatenated as a single vector and then trained through a fully connected neural network. The neural network consisted of 3 hidden layers with 256, 256, and 128 neurons in order, and the dataset was split into training, validation, and test sets in 8:1:1 ratio. The final results were taken as average of over 10 independent runs to minimize the error. It was observed that the MAE was higher in the case of solvents with less presence in the training data, and it was mentioned that more more work neesd to be done to select features. A similar study by Dongdong Zhang, Song Xia et el was on aqueous solutions [1] .

The study by Yeji Kim, Yoonho Jeong et al [6] was focused on the prediction of solvation energy of pharmaceutical compounds with consideration of molecular interaction between solute and solvent. Graphical transformation of the molecule was performed using CIGIN (Chemically Interpretable Graph Interaction Network). MPNN was used in the same manner as the previous case but with an additional phase to calculate molecular attraction between the solvent and the solute using interaction matrix. The molecular graph was also similar to the previous case, but the nodal features represented covalent bonds by including parameters such as radical electrons, donor-acceptor electrons, etc instead of ionic properties. Acidic or basic nature was also one of the nodal features. Similarly, edge features contained parameters relevant to covalent bonds. Overall there were 9 nodes and 4 edge features. The article also explicitly mentioned not including Hydrogen atoms or covalent Hydrogen bonds in the features due to their normalization effect on molecular interaction and hence avoiding unnecessary complexity in the computation. For the interaction phase part, the intermolecular attraction was expressed as a function of solute and solvent.

$$I_n m = f(A_n, B_m)$$

$$f(A_n, B_m) = tanh(A_n B_m)$$

$$A^{'} = I.B$$

$$B^{'} = I^T.A$$

In the equation above, A and B represent the solute and solvent feature vectors, respectively, and I represents the interaction matrix. The influence of B on A is calculated as $I.A$ and the influence of A on B is calculated as $I^t.B$. The data was split into training and test data with a ratio of 9:1. The key aspect of this study was validation on an external set of 642 solvents from the FreeSolv dataset, which were not part of the data used in training. The result yielded An RMSE of 0.73 ± 0.01

A more advanced study by Shiyi Qin, Shengli Jiang et al [4] involves calculation of the activity coefficient at infinite dilution instead of interaction energy. Since the activity coefficient is an intrisnsic property of solutions, it can give insights into feature selection and Neural network architecture to retrieve intermolecular properties. A realtively larger and diverse database with 200,000 entries for binary mixtures and 160,000 entries for ternary mixtures was used for training. The molecular graphs were generated using RDKit version 2019.03.02. Additional augmented data was generated using via COnductor-like Screening MOdel for Real Solvation (COSMO-RS). The key aspect of this study was the creation of a secondary graph representing intermolecular attraction. To compare it's effect, there were three GNN architectures used i)SolvCAT: isolated interaction network built at the local level for each molecule using atoms as node. Local graph convolution applied to nodes and the features were concatenated into a single vector to be trained through a connected Neural Network (FCNN) ii) SolvGCN: a network simulating molecular interactions by treating a molecule (instead of an atom) as a node. It constituted a global level convolution and was combined with atomic level local convolution iii)SolvGNN: in this case, edge features in the global interaction graph were specified as a Hydrogen bond between the solvent-solvent as well as solute-solvent. A layer of MPNN was added before FCNN. In all three cases, refined representations of intermolecular attraction were sent to FCNN at the final stage. The schema of the neural network architecture is presented in fig 1. The hyperparameters used were i) the number of graph convolution layers (1,2), ii) the number of fully connected readout layers (1,2,3), iii) the number of hidden neurons (128,256), and iv) the learning rate (0.0005,0.001). It was found that the SolveGNN model achieves significant accuracy concerning activity coefficient measured by DFT or COSMO-RS estimations, with a maximum MAE of around 0.032

The overall conclusion based on these studies is as follows:

- Graph neural networks work as feature extractors/refiners and have shown promising results in the prediction of molecular properties. Though just accounting for a graph representing molecular structure might not be enough to measure emergent properties like solvation energy. Inclusion of inter-molecular features can show significantly better results.
- The feature selection strategy combined with neural network architecture affects the final results. In most cases, the raw data was retrieved from SMILES, which was later transformed into molecular graphs. Most GNN architectures were used in conjunction with MPNN, which makes it an integral part of GNN-based learning.
- The amount of data GNN needed is significantly lesser (in order of 10^5) than data fed into traditional NN (in order of millions). This gives GNN a significant advantage in Chemistry due to the lack of availability of big data.

Though having data rich in diversity and quantity still has the potential to give better and variance-free results.

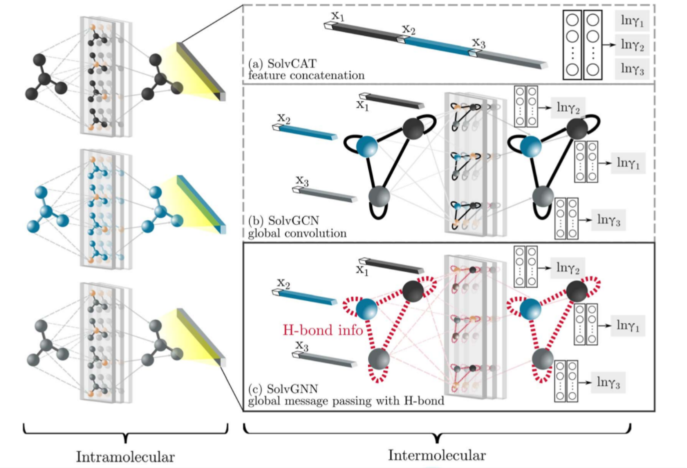

Fig. 1. (quoted from [4]) From top to bottom (1) conducts a simple concatenation of the mole fraction and locally embedded features; SolvGCN (2) constructs an intermediate molecular interaction network followed by global convolution without explicit edge information; SolvGNN (3) explicitly incorporates H-bond information as the edge feature in the interaction network, which undergoes global message passing for "intermolecular"-level feature embedding

B. Prediction of interaction energies of multi-molecular systems

In order to understand complexity of molecular aggregates, we will have a look at the study by Ian Rouse, and Vladimir Lobaskin [7], which used a combination of CNN and FCNN to predict the Potentials of Mean Force or (PMF). It first discusses the difficulty of experimentally measuring interaction forces in adsorption due to the high dimensionality of chemical space of adsorbates and adsorbent surfaces, and Molecular Dynamics and DFT-based simulations would be even more expensive than solvation in this case. The complexity in the system arises due to the contribution of all atoms on the surface as well as the medium. To simplify the process, the researchers tried to predict coefficients of PMF (Potential Mean Force), which is a convenient way to define interaction energy as a function of coordinates along the interacting surface. Therefore the prediction task is not limited to a single value of energy but rather coefficients of an expression. After training the model, it obtained a good fit with the training data but suffered from overfitting. Or in other words, it worked well only on the training data or data similar to it, which was expected due to the complexity of the system and the limited amount of data.

Now we examine the study by Anshul Gupta et al [8], where graph Convolutional Networks (GCN) were used to predict the adsorption of CO_2 in crystalline metallic surface. It consisted of Enhanced Crystell graph neural network or ECGCNN, an advanced version of the previously built CGCNN model by Xie and Grossman [**?**]. ECGCNN encoded atoms of crystal as the nodes and interaction as edges but allowed multiple edges between the same pair of nodes in order to represent the periodic crystal structure. It can be interpreted that an atom can be a part of multiple unit cells, therefore one edge between a pair of atoms will represent a relationship within a unit cell, and the other edge will represent two different unit cells. However, the number of edges per node was not fixed. Only the strong interactions constituted the edge features, and the rest were ignored. This was accomplished by using edge convolution, i.e in the convolution layer, along with node features, edge features could also change. Interestingly, no molecular properties of CO_2 was used as feature in this study. The training data was obtained as Crystallographic Information File (CIF) from Molecular Orbital Framework(MOF) database of around 9525 materials. The adsorption properties of the materials were calculated using Grand Canonical Monte Carlo (GCMC) simulations. The GCN consisted of 4 layers, and the data was randomly split into training, validation, and testing over 5 times. An average of all the results was taken at the end. The ECGCNN showed significant improvement over traditional CGCNN for the prediction of interaction energy. In conclusion

- Use of GNN with dynamics edge features to model the crystal structure of the adsorbent surface
- Multiple edges were allowed between a pair of crystal structures due to the presence of multiple unit cells within the lattice
- The molecular properties of the gas involved were ignored

Our final study is by Sergio Pablo-García, Santiago Morandi et al [9] to predict the adsorption of a diverse set of organic molecules containing C 1–4, C 6–10 aromatic rings with functional groups including N, O, S on transition metals. The framework is named as GAME-Net (Graph-based Adsorption on Metal Energy-neural Network), which is a GNN-trained model on a large number of DFT datasets composed of 3,315 closed-shell organic molecules entries and commonly occurring functional groups) adsorbed on metal surfaces. Starting with the geometric representation based on DFT, an atomic ensemble was extracted which consisted of the adsorbate molecule and the metal atoms in its immediate vicinity. In other words, instead of taking contributions from all the metal atoms, an algorithm was developed which could filter the metal atoms contributing to the adsorption of given organic molecules and the rest were discarded. The remaining atoms (metals and organic combined) were transformed into a graph in the usual way. However, in this case, the edge features only had static categorical features, which just included information on the "type of bond" (metal-organic, organic-organic, or metal-metal). To handle the complexity in molecular aggregates as mentioned by Ian Rouse and Vladimir Lobaskin [7], the study excluded the energy from the rest of the system by substracting total energy of the system before interaction from the net energy after adsorption. The calculation is shown as

$$E_{ads} = E_{AM} - E_M - E_A$$

Here E_A was taken as zero and the expression becomes

$$E_{idft} = E_{AM} - E_M$$

Thus the given expression gives only the energy between the adsorbate and the adsorbent. The transformation is visualized in the fig 2. The GNN architecture consisted of i) fully connected layers, ii) convolutional layers, and iii)pooling layer with a total of 285,761 parameters. Different metal facet orientations ((100) and (110) in Miller indices) with 1,776 points in each facet were also used in training data. The MAE with respect to the DFT values was up to 0.18 eV. The error was found to be asymmetrically more in aromatic compounds. The suspected reason is not taking the complexity of molecular bonds into account while building features. In the results, it was claimed that the training time of the model is of the order 10^7 seconds, but once trained, it can immediately predict scaled system energy, which can be converted to actual adsorption energy using reverse transformation. The DFT simulations take time in a similar order but in GAME-Net, it is only a one-time process (as long as one doesn't need to train the model over and over again!). On a more positive note, the model was designed to even work for industrially relevant, larger organic molecules, and the model worked with reasonable accuracy for the case of large multifunctional molecules, even with lesser training data and the presence of aromatic rings. It is optimistic as very large molecules are even harder to simulate with DFT, and if a framework can demonstrate the capability to accurately calculate the adsorption energy of large molecules such as proteins, polyurethanes, and polymeric molecules, it can significantly contribute to industrial applications. The suggested ways to further enhance this approach were i) inclusion of molecular bonds in the graph, ii) inclusion of metal atoms which are not in the immediate vicinity of gas molecules but close enough to influence adsorption, iii) Use of geometric deep learning to create features based on molecule-metal interaction from first principle.

Conclusions of the sub-study:

- The first study used conventional neural networks to come up coefficient for a quantity that can describe interaction energy based on adsorption surface coordinates
- The second study explored a simple case of CO_2 adsorption on metals and use GNN to model metallic lattice but entirely excluded the gas molecule in training
- The third study dealt with a complex case of large organic molecules with metallic lattices of different orientations and tried to build GNN based model with adsorbate molecules and metallic atoms directly connected to the adsorbent
- It becomes particularly tricky when trying to predict molecular properties using GNN in cases where one of the entities is a large aggregate of molecules that constitutes a macroscopic entity (like adsorption on the surface of a catalyst). Since the GNN involves transforming atoms and molecules into nodes, ideally, one needs to take into account all atoms present in the lattice, which is impractical. But the same issue exists even with other approaches of ML as well as DFT

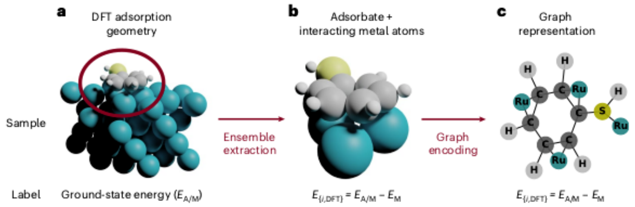
a
DFT adsorption geometry
b
Adsorbate + interacting metal atoms
c
Graph representation
Sample
Ensemble extraction
Graph encoding
H
H
H
C
C
Ru
Ru
C
S
H
Ru
H
C
C
C
Ru
H
H
Label
Ground-state energy ($E_{A/M}$)
$E_{\{i,DFT\}} = E_{A/M} - E_M$
$E_{\{i,DFT\}} = E_{A/M} - E_M$

Fig. 2. (quoted from [9]), Reduced 3D structure considering the bonds within the adsorbate (A–A) and between the metal atoms and the adsorbate (M–A), and the subtracted metal DFT energy, $E_{i,DFT} = E_{A/M}E_M$

III. Conclusion

It can be concluded that GNNs are emerging tools for learning chemical interaction due to their ability to transform molecular structure as a learning feature. They are particularly useful in the simulation of interaction energies in solutions. Their success can be attributed to intuitive feature building, lesser need for training data, and fast learning time. It can be interpreted that in dilute solutions, the behavior of a solute molecule can be determined just by its own chemical structure due to isolation from the other molecules, and homogeneity and inertness of solvent molecules. However, in some cases, a more precise model was built while taking solvent interaction into account by creating separate graphs for inter-molecular forces, such as hydrogen bonds.

For nanoparticle and/or molecular aggregates, ML with GNN faces major obstruction due to the presence of large number of interactions, which would lead to large and dense graphs, and consequently high computational costs. This isn't the case when trying to predict solvation energy. It would nullify the purpose of developing fast models, which is not possible with conventional approaches like DFT and conventional deep neural networks. While it is already cumbersome to propose a theory accounting for all possible or at least measurable environmental effects in molecular interactions, it would nevertheless result in hypergraphs and graphs of graphs, which would again form a computational barrier.

But at the same time, GNN has proven to possess high levels of transferability and interpretability. Since the GNN models are based on molecular structure and bonds, they are helpful in first principle model building. Researchers who have good knowledge about molecular interaction in their respective domains can find them especially useful. There are many ways in which GNN can be improved for nanoparticle interaction, like in the study, where GNN trained on organic functional groups could be transferred for predicting interactions in Biomolecules. Additionally, the selection of features is highly important to come up with a good ML model. So in order to reduce the time complexity of calculations, the selection of the right features can significantly affect the results. But overall, it can be inferred that GNNs are not data hungry like traditional neural networks, thereby serving appropriately to the cases where the data is not available is huge amounts. It can be further noted that the performance of GNN was significantly improved after data augmentation(i.e. by using data of existing molecules to create data of new molecules), training over large and diverse datasets, and using the right features.

Based on the studies in this paper, a proposed methodology to come up with a faster and more accurate modeling for nanoparticle interactions is

- Selection of diverse and large enough data set: Diverse and large data set will increase training time but reduce variance and increase transferrability
- Developing a GNN architecture that incorporates i) molecular structure of adsorbate, ii) adsorbate-adsorbent bonds, and utilize edge convolution to filter out the atoms/bonds which are not contributing to the net value of interaction energy
- A fair division of dev, test, and training set, and the ability to tune hyperparameters such that the trained model is transferable to new data just by modifying the dev set and re-tuning hyperparameters.

REFERENCES

[1] Dongdong Zhang, Song Xia and Yingkai Zhang ”Accurate Prediction of Aqueous Free Solvation Energies Using 3D Atomic Feature-Based Graph Neural Network with Transfer Learning”, Journal of Chemical Information and Modeling, 2022

[2] Ramin Ansari, Amirata Ghorbani, ”Accurate Prediction of Aqueous Free Solvation Energies Using 3D Atomic Feature-Based Graph Neural Network with Transfer Learning” University of Michigan Ann Arbor, Stanford Universtiy, arxiv Chemical Physics, 2021

[3] ASGN: An Active Semi-supervised Graph Neural Network for Molecular Property Prediction Zhongkai Hao 1 , Chengqiang Lu 1 ,Zhenya Huang et el

[4] Shiyi Qin, a Shengli Jiang, et el ”Capturing molecular interactions in graph neural networks: a case study in multi-component phase equilibrium”, Association for Computing Machinery, 2020

[5] :Partition Coefficient - an overview, ScienceDirect Topics”, Journal of Hazardous Materials, p. 1

[6] Yeji Kim, Yoonho Jeong et el, ”MolNet: A Chemically Intuitive Graph Neural Network for Prediction of Molecular Properties”, Asian Chemical Editorial Society, 2022

[7] Ian Rouse and Vladimir Lobaskin, ”Machine-learning based prediction of small molecule–surface interaction potentials”, Faraday Discussion, 2022

[8] Guojing Cong Anshul Gupta, et el ”Prediction of CO 2 Adsorption in Nano-Pores with Graph Neural Networks”, IBM Research, 2022

[9] Sergio Pablo-García, Santiago Morandi et el, ”Fast evaluation of the adsorption energy of organic molecules on metals via graph neural networks”, nature computational science, 2023