Developing a novel 3D GNN and Random Forest Regression model for screening and predicting potential oxide electrocatalysts with greater accuracy and computational efficiency

Stanley Chen
Department of Science: The Harker School
San Jose, America
stanley.chen397@gmail.com

Abstract**—** In order to develop green hydrogen into a viable source of renewable energy that can be produced and stored at scale to address climate change, cost effective electrocatalysts (catalysts that facilitate the reaction of electrolysis), are imperative. Current simulation based efforts, such as Density Functional Theory (DFT), to screen such catalysts are too computational intensive and not scalable. To address this, ML models have been developed that predict the total relaxed energy of catalyst structures, a key attribute that drives chemical reactions. Phase 1 of my research evaluated the Dimenet++ model and Graphormer 3D Transformer model on a subset of the OC20 (Open Catalyst 2020) IS2RE dataset to analyze the relationship between a 3D Transformer and a GNN, and establish a baseline model that can be used to compare to Phase 2. In addition, I trained a joint model with both OC20 and OC22 data and compared the results to the independently trained models produced earlier to evaluate joint training on model performance. However, the thermodynamic data provided by OC20 is not enough for accurate screening. Atom charge distribution is an important factor in driving the chemical reactions of electrocatalysts so it can potentially also bolster model performance, and it has not yet been included in previous models because it is calculated through DFT. Using Bader Charges calculated through DFT for atom structures in OC20, I performed feature engineering with 13 additional atomic characteristics, and constructed a Random Forest and GNN model to predict Bader charges for all atom structures. I then modified the Graphormer model to incorporate this new predicted charge attribute. The refined model saw ~3.5% less in loss and ~10% greater in EwT across all the validation splits. With this new model, electrocatalysts can be screened faster with higher accuracy at scale.

Keywords—Machine learning, transformers, Graphormer_3D, electrocatalyst, Dimenet++, OC20, OC22**, green hydrogen**

I. Introduction

Renewable energy is a hot stove for exploration all around the world. New technologies are urgently required to combat the problem of climate change and greenhouse emissions. In fact, by 2050, solar and wind energy is projected to encompass 70% of renewable energy generation [1]. However, with this prediction, a new problem arises; these sources are not stable and are subject to change. Several methods have been proposed to address this issue but the most promising one is HES (Hydrogen Energy Storage) [1]. Hydrogen gas is valuable because of its unique property of being able to store large amounts of energy in a stable, transportable form while also producing no carbon dioxide when incinerated [2]. Hydrogen production is split into two categories, green and gray. Gray hydrogen is produced through carbon-intensive energy sources. What is problematic, however, is currently, only 2% of global hydrogen production originates from green hydrogen synthesized through water electrolysis [3]. This low percentage stems from the expensive noble metals used in the stack of the catalyzer [1][2][3][4]. Currently, only platinum and iridium are viable materials used in catalysts because of the limiting environment; a highly acidic one to encourage proton exchange [5]. Screening potential catalysts is too expensive and time-consuming (computationally intensive) because of traditional methods i.e DFT (Density Functional Theory) which calculates the vibrations and movements of each individual atom within a molecular structure and can only deal with small and local molecular environments [1]. In the above strategy, catalysis is a key component in determining the speed and efficiency of the reaction taking place. However, screening eventual catalysts is a time-consuming process that involves iterating through millions of different possible configurations of atomic structures. Simulations such as DFT are at the forefront of this appraisal of new and better catalysts. However, despite increases in computational efficiency, the computational cost of DFT is still a limiting factor in the deep dive into possible catalysts. A possible solution to this problem exists with the burgeoning AI explosion. Using machine learning models trained on expensive DFT calculation's data, possible patterns that could not be previously detected can now be categorized and referenced for future exploration on similar materials (different models for different subsections; i.e cannot train a model on both oxide and other catalysts). In fact, the application of machine learning and artificial intelligence sees increases in many data-rich fields due to its potential to efficiently expedite simulation processes. Demonstrations from scientists have occurred in reaction network elucidation, thermochemistry prediction, structure optimization, accelerating individual calculations, and integration with characterization. These tasks are all fundamental in the modeling of heterogeneous catalysis and as such are invaluable to evaluating possible electrocatalysts and provide useful guidelines moderating the direction of scientific work in this field.

II. Methodology

A. Phase 1 Database Introduction

In this research, two databases were used. One is the OC20 database (1,281,040 DFT relaxations across wide range of materials, surfaces and adsorbates and contains inorganic and organic materials) [5] and the other one is the OC22 database (an extension of the OC20 database that consists of 62,331 DFT relaxations across a variety of oxide materials, and adsorbates) [6]. There are three reasons why this research employs these specific datasets. The first reason is because both are part of the largest dataset possessing the largest sample size in the field of electrocatalyst screening.

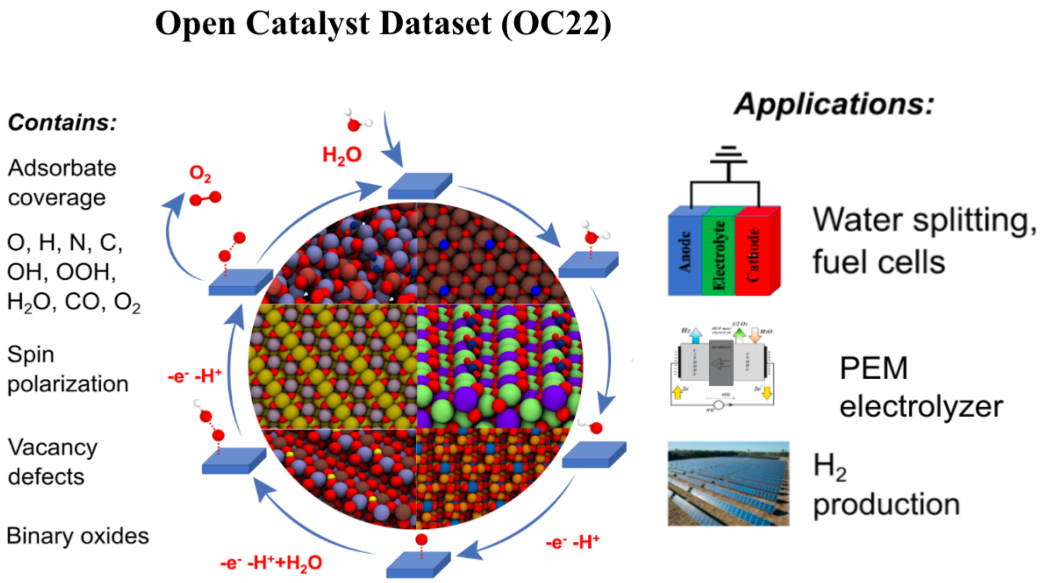


Fig.1. Overview of the contents and application of the OC20 dataset.

The second reason is because the data in these datasets are of the highest quality due to them being produced by intensive DFT calculations. Lastly, the dataset includes a github repository which contains results on all the models and efforts that were conducted on these datasets meaning that all previous work and data is publicly available. The OC22 Database is split into 3 tasks which are listed below: S2EF (structure to energy and forces) takes a structure and predicts the energy and per-atom forces. S2EF-Total is similar to S2EF but differs in the aspect that it instead finds the DFT total energy. In fact, the two tasks are related in that the energy output of S2EF is the S2EF-Total DFT total energy minus the DFT energy of a clean surface and the gas phase adsorbate energy. The IS2RE task takes an initial atom structure and predicts the relaxed energy. Similar to the relationship between S2EF-Total and S2EF, IS2RE-Total is the IS2RE relaxed energy plus the DFT energy of a clean surface and the gas phase adsorbate energy (the relaxed DFT total energy). Lastly, the IS2RS task takes an initial structure and predicts the relaxed structure. All the tasks specified above contain a train and validation split. The training split is separated into the following sections: Adslabs (Adsorbate+Slabs), Slabs. The validation split is split into two categories: ID which comprises about 47% of the dataset (in domain: for structures from the same distribution as training) and OOD which comprises about 53% of the dataset (Out of Domain: for unseen catalyst compositions) [6]. Similar to the OC22 database, the OC20 database consists of the same tasks with the exclusion of the S2EF-Total and IS2RE-Total tasks [6].

B. Phase 2 Database Introduction

The new model's input data was from the same superset as Phase 1. Only the sid, edge_index, atomic_numbers, pos, and natoms attributes were used, however. The new model's output data (Bader Charges) was extracted from DFT calculated attributes within the OC20 dataset. Since all the features were aggregated together, such as position and other characteristics, a lot of data parsing was needed. Furthermore, since each system did not have a corresponding Bader Charge, this research had to locate the ID of each system and pair them up together. The systems that did not have a Bader Charge were not utilized in this experiment. Each system, on average contains ~100 atoms, and since the Bader Charges of each atom is what we are calculating, a system is also multiple inputs. (For example, a 5k subsplit of our dataset is around 500k atoms corresponding to 500k Bader Charges)

C. Replicated Model Introduction

In this experiment, the two machine learning models trained on the IS2RE 10k split of the Open Catalyst Project database were Dimenet++ (a GNN-based model), and Graphormer (a 3D transformer model). GNN models operate on a graph structure consisting of nodes and edges. The edges serve as paths where messages are passed along. Based on these messages, atom embeddings at the node are iteratively updated. These node embeddings are initialized with the atom's properties, such as the atomic number and structure etc. [5]. 3D transformer models rely entirely on attention mechanisms and are effective at capturing spatial interactions in 3D data. The Graphormer model employs a standard transformer with 3 structural encodings: centrality, spatial, and edge encodings [7]. These specific models were chosen to be replicated because of their results for the IS2RE being reported in previous papers [5][6][7]. Therefore, we could compare our replicated model results to the original ones. The rationale behind these two models is not only for this reason; however, another reason was to compare a peak GNN model with a peak 3D transformer model in order to identify the areas where one model might perform better than the other (tradeoff between training time and accuracy). The Dimenet++ model possessed the best metrics in the OC20 paper [5] while the Graphormer 3D model won the Open Catalyst Challenge.

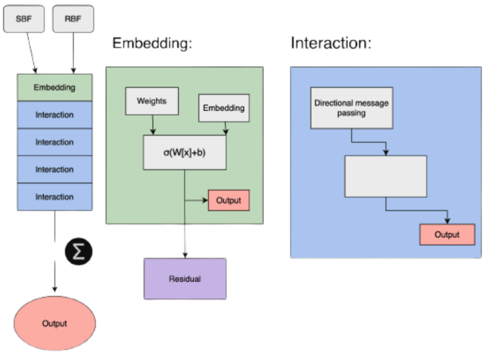


Fig. 2. Architecture of the Dimenet++ model

$$Attn_{3D}(q_s, k_t, v_t) = \left(Softmax\left(\frac{q_s W^Q (k_t W^K)^T}{\sqrt{d}} \right) \cdot \frac{\vec{r}_{st}}{d_{st}} \right) v_t W^V$$

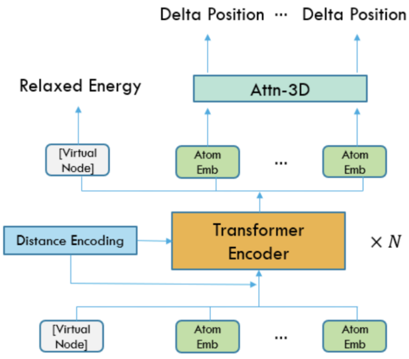


Fig. 3. Architecture of the Graphormer-3D model
Source: Adapted from Microsoft Research Lab - Asia

D. Bader Charge Model Introduction

In this phase of the experiment, three model architectures were trained on the 5k split of my created dataset One was a 2D Random Forest Regression Model, while the other two were both GNN's. For the rest of this paper, the first GNN will be referred to as Bader and the second one as Bader++. The 2D RF model used 100 estimators and had no specified max depth. The research implemented these parameters because the dataset did not contain so many features that it necessitated a max depth in order to limit overfitting. For Bader, a batch size of 32 and 64 hidden dimensions were used. Bader++ had the same hyperparameters as Bader except it contained further optimzations such as Gaussian smearing to normalize the input and 16 attention heads to capture long-range interactions within the data. These models were chosen because of the comparison that could be drawn between the 2D and 3D models, paralleling the conclusion that this research delved into in Phase 1 of the experiment, furthermore, it evaluates the efficacy of the ML optimizations stated previously. Moreover, the data was very well suited for a GNN model as it contained distances between each atom. So, by representing each atomic system as a fully connected graph where each node represents an atom and each edge as the distance between said atoms, the GNN's are able to process the data a lot better than other types of models. To supplement the input to these models, feature engineering was implemented, adding 13 additional characteristics from the Mendeleev python package such as boiling point to the node characteristics for each atom.

E. Phase 1 Evaluation Metrics

In this experiment, two primary metrics were used, Energy Mae and EwT. Energy Mae is the mean average error of the relaxed energy (output of the model) as compared to the ground truth energy while EwT is a metric that measures the percentage of systems whose relaxed energy is within 0.02 eV of the ground truth energy (an input) [6]. Energy Mae is a broader metric where the evaluation on each system is a number while EwT has only two options for evaluating on a system, yes or no: either it satisfies the conditions or it does not.

F. Phase 2 Evaluation Metrics

In Phase 2 of this experiment, there were two metrics for assessing model performance. The RMSE (Root Mean Square error) and the loss of the model. The Loss function was the Mean Squared Error.

G. Model development

The models this research will use are the Dimenet++ model and the 3D Graphormer model. 3D models are part of the "new" wave of machine learning algorithms, and are especially useful in this experiment because "compared with the classic message-passing-based GNNs (MPGNNs), Graphormer enjoys two unique characteristics: a global receptive field and an adaptive aggregation strategy" [4]. This paper will first confirm the previous models mentioned on the IS2RE task in order to cross-examine the experimental environment. The IS2RE task is used due to previous results for this task's metrics being reported in the literature for the two models already and furthermore because of the low computational cost of this task. The IS2RE output is valuable because "relaxed energies are often correlated with catalyst activity and selectivity, and the energies are important parameters for detailed microkinetic models" [1]. The original models were trained on the entirety of the OC20 dataset whereas our replicated models were only trained on the 10k split. The sensitivity of each model to dataset size can then be gauged.

The second part of the first phase of the experiment consists of using joint training on the Dimenet++ model. Looking at the results in the previous OC22 paper, it seems that joint training significantly improves the energy metrics for the OC22 dataset but not the force metrics. The results reported in the paper were joint trained on the entire OC20 IS2RE dataset and the OC22 dataset whereas this project only uses the 10k split of the OC20 IS2RE dataset with the same

OC22 dataset. Again, this allows the effect of dataset size during joint training on model performance to be determined. Our replicated Dimenet++ model is tested and validated on the OC22 dataset. This database was used because we wish to explore the reason behind the significant improvement of the energy metrics between the joint trained model and the one trained just on OC20 when evaluated on the OC22 dataset.

Electrocatalyst screening is not an easy task that can be evaluated on just one characteristic. Factors such as chemical stability also play a large part in determining the viability of each catalyst. The previous proposed model's input is only thermodynamic characteristics. However, this cannot by itself, serve as a good indicator of how good a catalyst will perform. The charge attribute of each atom is also instrumental in the behavior of the system. A good metric for the charge attribute of an atomic system is Bader charges, which are the atomic charges each atom possesses. However, these charges are calculated through DFT which is highly intensive computationally as stated earlier. So, in order to allow mainstream access to these charges, the second phase of this experiment revolves around constructing machine learning models to predict these Bader Charges. Currently, we have tested 3 different models. One is a 2D Random Forest Regressor Model, while the other 2 are GNN's. See Subsection D for more details.

III. Results and Discussion

A. Replication Results

From the Dimenet++ results (Figure 3), the OC20 paper's Table 4 was verified with the Ood metrics (Out of domain) performing significantly worse than that of the ID (in domain) metrics. In these results, a scalar smoothing factor of 0.99 was used to reduce noise in the data. This general trend was captured in these results; however, the metrics differed from the previous literature ones by ~20%. This discrepancy can be explained through the size of the dataset that was used in this experiment. The original paper used 294k systems while our

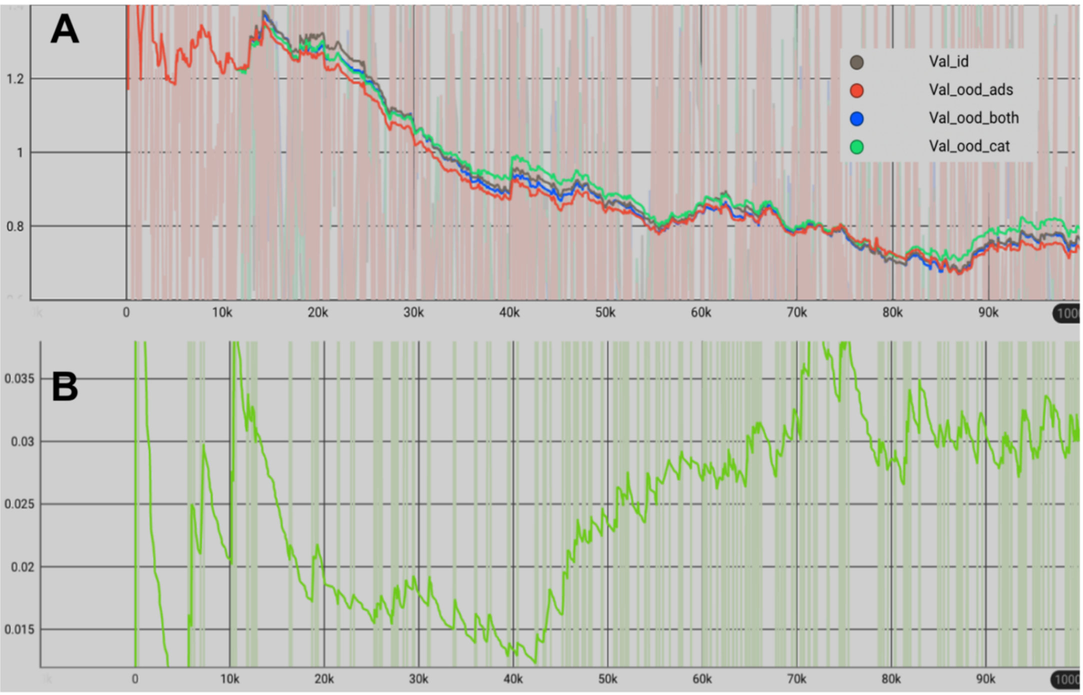


Fig. 3. Graphs of the Energy Mae (A) and average EwT (B) while training on IS2RE 10k direct

Table 3. Predicting Initial Structure to Relaxed Energy (IS2RE) with Dimenet++

	Energy_Mae ↓ (10k split)				EwT↑ (energy within threshold) (10k split)				Time to run (s)↓			
	Id	Ood_ads	Ood_cat	Ood_both	Id	Ood_ads	Ood_cat	Ood_both	Id	Ood_ads	Ood_ads	Ood_both
Validation	0.8818	0.9356	0.8488	0.8505	0.0199	0.0162	0.0183	0.0174	28789	20908	27050	22151
Test	0.8786	0.9432	0.8642	0.8804	0.0196	0.0181	0.0194	0.0181	1.7477	1.7341	0.6779	0.6382

reported metrics are predicted from a 10k split. In addition to the 2 metrics that were reported in the OC20 paper (e.g Energy_Mae, EwT), a 3rd metric that measures the inference time of the model was evaluated. For the 10k split trained on one Tesla T4 GPU, the average train time of the Dimenet++ model was ~24500 seconds or 6.8 hours. One of the findings that resulted in using a smaller subsplit was an overfitting issue of around ~0.13 eV for the Energy_Mae and ~0.055 for the EwT. The model replicated above trained for 20 epochs and a total of 100k training steps. In the previous paper, their Dimenet++ model was trained for 200 epochs so our model ran for around 1/10 of the total steps. Around ~86k steps in all four iterations with different validation subsplits, the metrics started decaying, a possible reason being overfitting. This overfitting issue can be fixed by the early stopping of the training process.

The Graphormer 3D model was trained for 100k training steps and 41 epochs on the same data set as the Dimenet++ model. The train time of this model was 32952 seconds or 9 hours. However, this training time only took into account the actual training and not the time taken to validate the model as it was being trained. The total time was ~20 hours. The loss function in the Graphormer documentation is similar to that of the Energy Mae function in the OCP documentation (Dimenet++), therefore the two results can be listed under one metric. Again, in these results, a scalar smoothing factor of 0.99 was utilized. In general, the results of this experiment followed the general trend that was reported in the Benchmarking Graphormer paper but the numbers are ~60% off from the paper's results. This result can be explained through the fact that the model in the paper used the entire IS2RE dataset, while this experiment only used its 10k split, and also further through the fact that the paper's model was trained for 1 million steps while the replicated model only was trained for 100k steps (again, 1/10 of the steps of the previous literature's model). Graphormer has better metrics (both the Energy Mae and EwT) compared to Dimenet++ which is a result that the previous literature also reported. One caveat is that the Graphormer model took around 20 hours to fully train and validate which is much more than the time reported for Dimenet++. This tradeoff of better metrics for slower training times is expected as Graphormer is much more complicated than the Dimenet++ model because of its 12 layers and transformer capabilities [7].

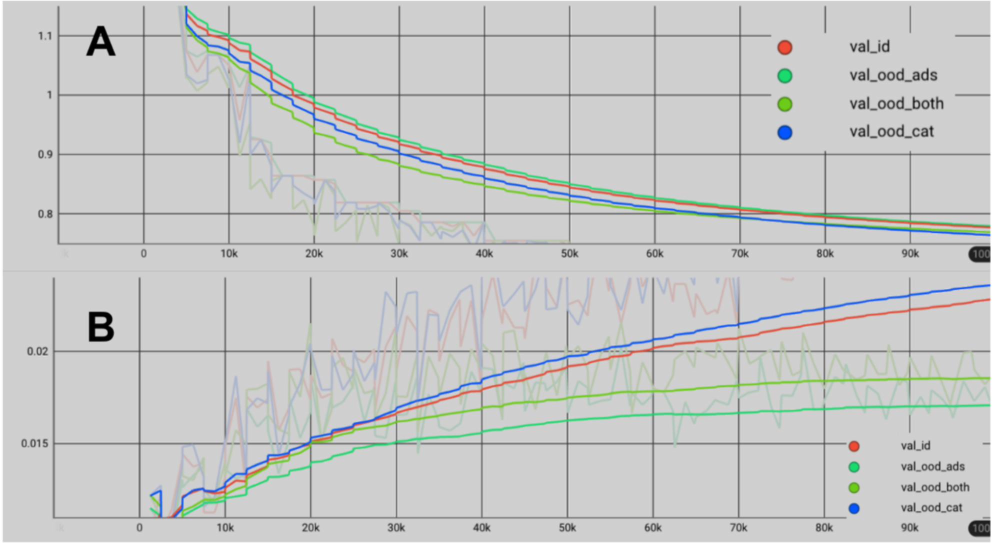


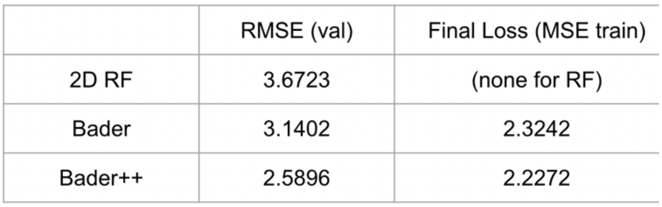
Fig. 4. Graphs of the average loss (A) and EwT (B) while training on IS2RE 10k direct

Table 4. Predicting Initial Structure to Relaxed Energy (IS2RE) with Graphormer 3D

	Energy_Mae ↓ (10k split)				EwT↑ (energy within threshold) (10k split)				Time to run (s)↓
	Id	Ood_ads	Ood_cat	Ood_both	Id	Ood_ads	Ood_cat	Ood_both	Avg,
Validation	0.7338	0.7338	0.7209	0.7338	0.0265	0.0165	0.0263	0.0184	32952

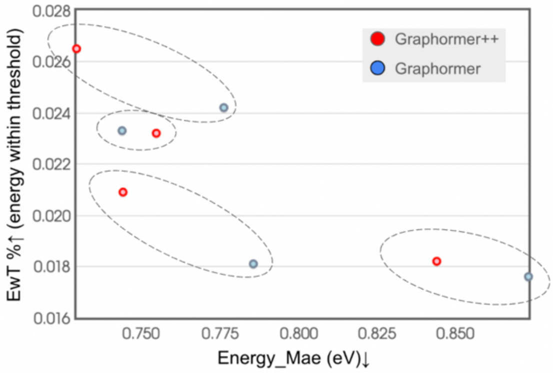
B. *New Model Results*

Table 6. Bader predicting models trained on 5k subsplit



	RMSE (val)	Final Loss (MSE train)
2D RF	3.6723	(none for RF)
Bader	3.1402	2.3242
Bader++	2.5896	2.2272

Fig. 5. Graphormer vs. Graphormer++ (Graphormer with the novel Bader charge model) where each pairing represents a data split



In this phase of the research, we trained a 2D Random Forest Regression Model, and two Graph Neural Networks on a 20k subsplit of the dataset we created using the raw Bader charges provided in the OC20 dataset. These models were validated on a 10k subsplit. As expected, the GNN models performed better than the Random Forest Regression Model, achieving significantly better RMSE's. Furthermore, the Bader++ performed much better than the Bader model, not a surprise as it included many ML optimizations. After developing this model, we then modified the Graphormer model to accommodate this new machine learning predicted charge attribute. For this improved model, we found that the metrics (Energy Within Threshold and the MAE) generally were better, achieving ~3.5% less in loss and ~10% greater in EwT across all the validation splits.

IV. Limitations

In this research, several limitations played a big factor in model development including dataset size and the number of training steps. In both the Graphormer and Dimenet++ models, the 10k split was preferred to the entire IS2RE dataset due to a scarcity of resources (only one GPU, no cluster) and also to further investigate the correlation between dataset size and model performance. As reported earlier, the Dimenet++ model replicated in this experiment was off by around ~20%. The hyperparameters used by the previous literature was conserved in this experiment, so the dataset size was the contributing factor leading to this decline in accuracy. Similarly, for the Graphormer 3D model, the metrics in this experiment differed from the previous papers by ~60%. However, in this case, dataset size was not the only limiting factor. The number of training steps also contributed greatly to this delta as the previous paper trained the Graphormer model for 1 million steps while this replication only utilized 100k steps in order to determine model performance using less intensive computations.

V. Conclusion

This research confirmed that the Graphormer 3D model performed better over all the metrics than the Dimenet++ model but it took ~20 hours to train while the Dimenet++ model only took ~12 hours. The Dimenet++ model reported results differing from the previous literature by ~20% while the Graphormer 3D model reported results differing from previous literature by ~60%. In Phase 2 of this research, we addressed the problem of previous models only taking in account thermodynamic properties of atomic systems by developing three models for the prediction of Bader Charges: one 2D RF model and two GNN models. By comparing the two latter models, we determined the optimal architecture for this task. Furthermore, by contrasting the RF model with the other models, we reinforced the previous conclusion that 3D models are better suited to the prediction of atomic characteristics than 2D ones. Finally, by combining our Bader Charge model with the previously evaluated Graphormer model, we see a 3.5% decrease in loss and a 10% increase in the EwT.

References

[1] Zitnick, C. Lawrence, Lowik Chanussot, Abhishek Das, Siddharth Goyal, Javier Heras-Domingo, Caleb Ho, Weihua Hu, et al. 2020. "An Introduction to Electrocatalyst Design Using Machine Learning for Renewable Energy Storage." arXiv. http://arxiv.org/abs/2010.09435.

[2] Patonia, Aliaksei, and Rahmatallah Poudineh. 2022. "Cost-Competitive Green Hydrogen: How to Lower the Cost of Electrolysers?" Working Paper 47. OIES Paper: EL. https://www.econstor.eu/handle/10419/253279.

[3] Terlouw, Tom, Christian Bauer, Russell McKenna, and Marco Mazzotti. 2022. "Large-Scale Hydrogen Production via Water Electrolysis: A Techno-Economic and Environmental Assessment." Energy & Environmental Science 15 (9): 3583–3602. https://doi.org/10.1039/D2EE01023B.

[4] Sun, Xinwei, Kaiqi Xu, Christian Fleischer, Xin Liu, Mathieu Grandcolas, Ragnar Strandbakke, Tor Bjørheim, Truls Norby, and Athanasios Chatzitakis. 2018. "Earth-Abundant Electrocatalysts in Proton Exchange Membrane Electrolyzers." Catalysts 8 (December): 657. https://doi.org/10.3390/catal8120657.

[5] Tran, Richard, Janice Lan, Muhammed Shuaibi, Brandon M. Wood, Siddharth Goyal, Abhishek Das, Javier Heras-Domingo, et al. 2023. "The Open Catalyst 2022 (OC22) Dataset and Challenges for Oxide Electrocatalysts." ACS Catalysis 13 (5): 3066–84. https://doi.org/10.1021/acscatal.2c05426.

[6] Chanussot, Lowik, Abhishek Das, Siddharth Goyal, Thibaut Lavril, Muhammed Shuaibi, Morgane Riviere, Kevin Tran, et al. 2021. "Open Catalyst 2020 (OC20) Dataset and Community Challenges." ACS Catalysis 11 (10): 6059–72. https://doi.org/10.1021/acscatal.0c04525.

[7] Shi, Yu, Shuxin Zheng, Guolin Ke, Yifei Shen, Jiacheng You, Jiyan He, Shengjie Luo, Chang Liu, Di He, and Tie-Yan Liu. 2023. "Benchmarking Graphormer on Large-Scale Molecular Modeling Datasets." arXiv. https://doi.org/10.48550/arXiv.2203.04810.

[8] Gasteiger, Johannes, Shankari Giri, Johannes T. Margraf, and Stephan Günnemann. 2022. "Fast and Uncertainty-Aware Directional Message Passing for Non-Equilibrium Molecules." arXiv. https://doi.org/10.48550/arXiv.2011.14115.