

TEAM 29: ENABLING FAST PREDICTION OF CATALYTIC MATERIAL PROPERTIES WITH GRAPH NEURAL NETWORKS

Alexandria Do, Robert Ramji*, and David A. Najera**

*University of California San Diego, La Jolla, CA 92093-0238

ABSTRACT

Density functional theory (DFT) is a computational method that utilizes quantum calculations to determine the electronic structure of materials. In terms of catalyst design, DFT calculations have been used for assessing the effects of doping [1] and structural optimization [2], as well as determining the adsorption energies of reactants [3]. DFT calculations can achieve high accuracy, which makes them useful for discovering and validating new catalyst materials, a critical task in the fight against climate change. However, these calculations are extremely time consuming and resource intensive, which makes DFT an untenable method for materials discovery at the scale needed to find new efficient and affordable catalysts. In this paper, we explore state-of-the-art graphical neural network (GNN) models used to replace expensive DFT simulations. We propose updates to the DimeNet++ architecture by modifying the way atom’s pair distances are embedded in the computational graph as well as implementing a variational layer that seeks to address aleatoric uncertainty. We demonstrate that the proposed modifications improve upon the state-of-the-art models in at least one error metric when tested on out-of-domain datasets.

Index Terms—graphical neural networks, GNN, density functional theory, DFT

1. INTRODUCTION AND RELATED WORK

Density functional theory (DFT) is a computational method that utilizes quantum calculations to determine the electronic structure of materials. In terms of catalyst design, DFT calculations have been used for assessing the effects of doping [1] and structural optimization [2], as well as determining the adsorption energies of reactants [3]. DFT calculations can achieve high accuracy, which makes them useful for discovering and validating new catalyst materials, a critical task in the fight against climate change. However, these calculations are extremely time consuming and resource intensive, which makes DFT an untenable method for materials discovery at the scale needed to find new efficient and affordable catalysts. In response to this issue, the Open Catalyst Project (OCP) [4] is an open call for creating fast, inexpensive ML-driven models that can predict the electronic properties of catalytic mate-

rials. The scope of the OCP addresses the need for improved and affordable catalytic materials for reactions that produce chemicals for renewable energy storage, fertilizers, and other environmentally impactful industrial processes. The OCP is split into three tasks: calculating per-atom forces (S2EF), relaxation structures (IS2RS), and relaxation energies (IS2RE). The focus of this paper is the IS2RE task. The predicted energies can then be used for determining the structure and energies of the relaxed (energy minimized) forms of the catalytic systems.

The proposed approach for modeling the atomic relaxation is based on state-of-the-art methods for modeling atomic interactions, namely Graph Neural Networks (GNNs) [5]. GNNs are a natural fit for this type of problem as atoms can be represented by nodes in a graph, while the relationships between atoms are represented with edges [4]. Node features will include atom positions and atomic number. The edge features include bond type and distance. The energy of the system is treated as a global graph feature. The OCP challenge provides baseline models coded in PyTorch [6] which we leverage and improve on by expanding upon the notion of model-form uncertainty. We evaluate the model’s performance by computing the mean absolute error (MAE) and comparing it to the challenge leaderboard.

1.1. Dataset

The data for the IS2RE task makes up an almost 100 Gb subset of the total dataset provided for the OCP. For this task, the data consists of graph representations of molecular systems (as pictured below Fig 1) prior to DFT relaxation, and their target relaxed energies (as calculated by DFT). Each entry in the IS2RE data represents a single system of an adsorbate (molecule such as CO) and a catalytic surface (five atomic layers of metal catalyst atoms of either one, two, or three atomic species). Each entry is in the form of a graph that encodes the 3D (x,y,z) atomic positions, as well as relevant elemental data, such as covalent radius and electron affinity.

1.2. ML Approaches

Several iterations of GNN-based models were explored for this application by the creators of the OCP. The creators used

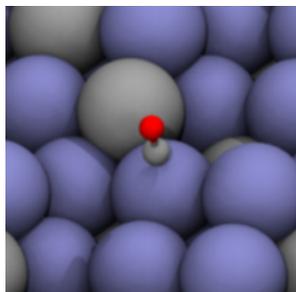


Fig. 1. Carbon monoxide and binary catalyst in initial position

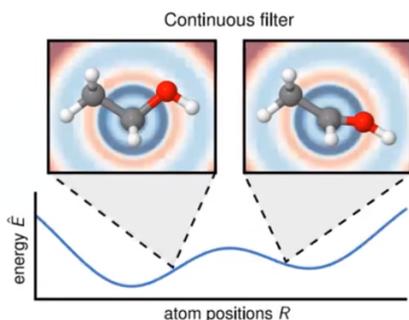


Fig. 2. Continuous radial basis function filter example

the SchNet [7], a model enabled by continuous-filter convolutional layers (Fig 2) to learn the 3D position-dependent interactions between atoms without the need for the discrete grid system used in models trained image and video data. Continuous filters are necessary to compute the forces via a back propagation-like gradient step, which discrete, discontinuous filters would render impossible. The filters used in SchNet are radial basis functions, or series of Gaussian functions distributed between each atom’s 3D position and a specified cut-off distance (approx. 6 Angstroms).

This was a good start, but radial basis functions only encode distance, which neglects important angular components of force that occur in a many-body system such as the adsorbate-on-slab studied here. Dimenet (Directional Message Passing Network) [8] was the next iteration on the GNN for this problem. To account for angular information, Dimenet uses basis functions from a spherical 2D Fourier-Bessel basis as its continuous filters, which it integrates into a message-passing layer (embedding the messages between atoms rather than the atoms themselves, unlike a standard GNN). The network then uses the distance between each pair of atoms and the angles between each triplet of atoms within a set distance of the target atom to learn filters that include radial and angular information. This is an effective but expensive way to define filters, as the computational cost scales as $O(n^3)$. Indeed, Dimenet’s creators found that their embedding strategy increased the computational cost of training the message passing layer by 15x [9] over other parts of

the model that use more standard GNN embeddings. The following section describes our approach which builds upon these established GNNs.

2. METHODOLOGY

The Dimenet++ model introduced many innovative features that resulted in improvements in the average prediction accuracy of relaxed energy. In this work, we explored two modifications to the Dimenet++ model. The first approach consisted in modifying the way in which feature extraction is performed by exploring different ways in which interatomic distances are represented. The second approach consisted in adding a variational layer to account for potential uncertainties in the energy estimates in the training data. The two approaches will be described in detail in the following sections.

2.1. Feature Extraction: Representing Atoms’ Positions

A molecule is uniquely defined by the atomic numbers and positions [8]. As such, these are the sole features considered in this work. Based on the fact that atoms are rotationally invariant [8], the original authors of Dimenet++ came up with a feature transformation scheme to better capture this physical property. To this end, the interatomic distances are represented via radial basis functions. In addition to the distances, it has been found that the angles between atoms also provide essential information for prediction of the molecule’s relaxed state. Hence, the angles are also transformed using spherical Bessel functions and spherical harmonics. The details of these original feature transformations can be found in the Dimenet paper [8] [9]. The radial basis functions are just one way of representing a distance. In this work, we experimented with many forms and found that the following exponential basis representation resulted in good performance.

$$\tilde{e}_{ExpBF,n}(d) = \sqrt{\frac{2}{c}} e^{n\pi d/c} / d \quad (1)$$

Where d is the interatomic distance, c is the cutoff distance that defines the range in which atoms interact with each other, and n is the dimensionality of the basis, which was set to 16 based on experimentation. The first six functions are plotted in Fig. 3. As illustrated, this exponential basis function places a strong emphasis on closer atoms and decays rather quickly after that. It should be noted that the “best” choice of basis function representation is not clear from the physical problem alone, and that our choice is rather empirical.

2.2. Variational Output: Capturing Aleatory Uncertainty

There are a vast number of non-equilibrium states reachable in high-energy molecular simulations [9] which are relevant to catalysis but may not be captured by a limited training dataset. Furthermore, perturbations to the initial conditions

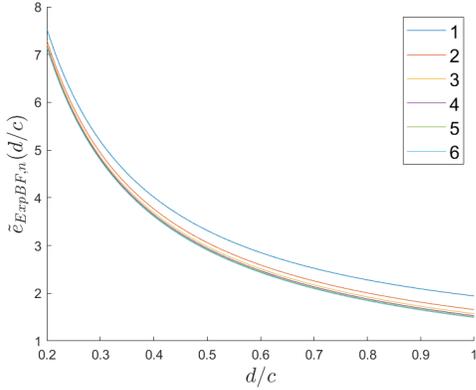


Fig. 3. Modified exponential basis functions meant to emphasize nearby atoms more strongly than the traditional radial basis function basis.

of these systems may introduce noise in the energy predictions. These two aspects are fundamentally tied to the physics of the problem at hand and demonstrate the need to account for uncertainty in the energy predictions. In this work we implement a variational output block inspired by the techniques used in Variational Autoencoders (VAE). Much like a VAE, we implement an output block composed of a series of dense layers using the “swish” activation function to predict the mean and log-variance of the relaxed energy. A schematic showing the output block architecture is shown in Fig 4. As shown, both the mean and the variance are parameterized by the transformed interatomic distances. We recognize that it is possible that there exists epistemic uncertainty that could be better parameterized by other physical or chemical properties of the atoms, but we did not explore such alternatives in this report. The end result of this approach is that the model predicts a parameterized normal distribution representing the relaxed energy instead of a single scalar value. To perform backpropagation with conventional stochastic gradient descent methods, we then sample from the random distribution by using the reparameterization trick as follows:

$$p(E|\tilde{e}_{ExpBF,n}) \sim \mathcal{N}(\mu_E, \sigma) \quad (2)$$

$$E = \mu_E + \sigma\epsilon, \epsilon \sim \mathcal{N}(0, 1) \quad (3)$$

As shown in Fig. 5, the output of the network is a random sample from this normal distribution that represents the possible values that the relaxed energy can take, given the system’s initial state and the atom’s atomic numbers.

3. RESULTS

The primary metrics used to evaluate the models were the Mean Absolute Error and the Energy within Threshold (EwT) which is the percentage of predicted energies within 0.02 (eV) of the ground truth energies. These metrics were based on

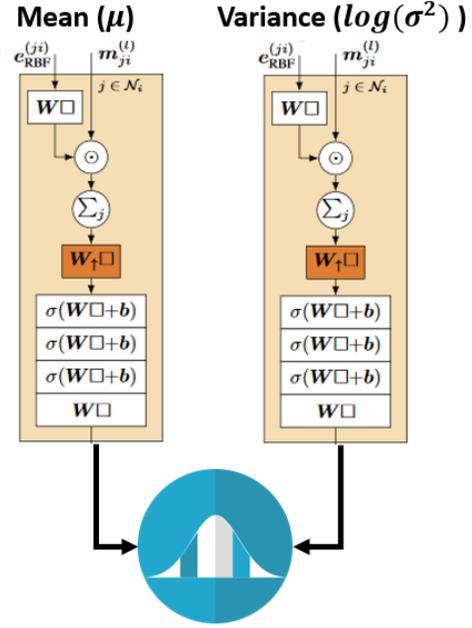


Fig. 4. Schematic showing the output blocks used for the energy prediction.

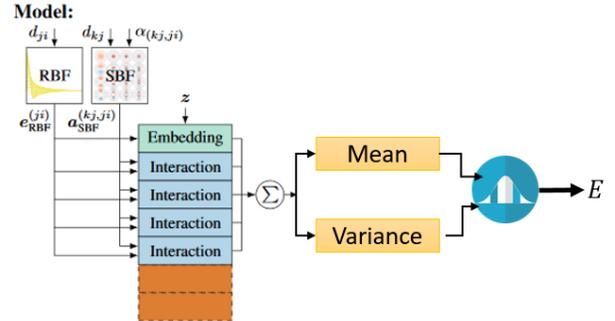


Fig. 5. Schematic of the modified Dimenet++ architecture with variational output.

the guidelines provided by the OCP. Each version of the Dimenet++ model (base, variational, and exponential) was trained on the IS2RE data subset consisting of 10k relaxation trajectories. The MAE, MSE, EwT, and loss were calculated, with the MAE shown in Fig . 6 and the EwT results shown in Fig. 7.

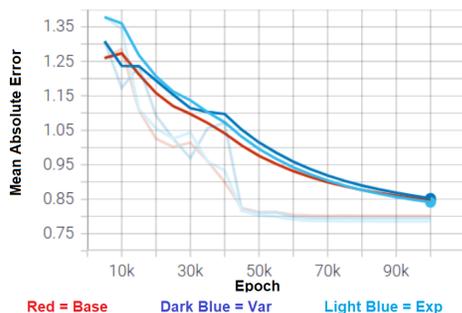
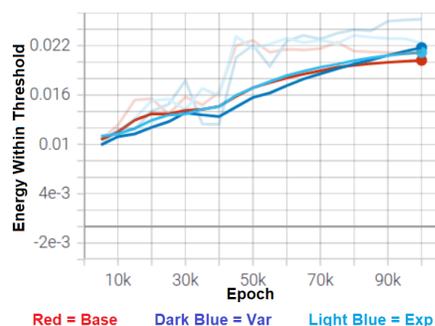
For the EwT metric, we wanted to see if our edited models could maximize the percentage of predicted values within the designated 0.02 (eV) energy range from the ground truth. Fig. 7 shows that both the exponential and variational versions outperformed the baseline DimeNet++ model. Fig. 6 shows the Mean Absolute Error of the training results and there is no apparent difference in the end results of the altered versions when compared to the base version. This trend also applied to the MSE and loss functions. We evaluated our models by testing them on the full IS2RE dataset, which consisted of DFT

Table 1. Mean Absolute Error with percent improvement from baseline model shown right of MAE results.

Model	In Domain		Out of Domain					
	Both	%	Catalyst	%	Adsorbate	%	Both	%
Base	0.798		0.786		0.856		0.801	
Var	0.790	0.972	0.783	0.412	0.921	-7.67	0.848	-5.95
Exp	0.780	2.18	0.771	1.95	0.905	-5.76	0.857	-7.10

Table 2. Energy Within Threshold with percent improvement from baseline model shown right of EwT results.

Model	In Domain		Out of Domain					
	Both	%	Catalyst	%	Adsorbate	%	Both	%
Base	0.025		0.024		0.020		0.019	
Var	0.023	-8.35	0.021	-12.0	0.018	-10.2	0.017	-9.42
Exp	0.023	-7.70	0.023	-5.83	0.017	-15.9	0.016	-12.2

**Fig. 6.** Validation Plot for Mean Absolute Error**Fig. 7.** Validation Plot for Energies within Threshold Predictions

data for approximately 100k structures. The results of these test runs are shown in Tables 1-2. Additionally, we posted our results to the Open Catalyst Project Leaderboard to see how our DimeNet++ versions fared relative to the other teams.

4. DISCUSSION

The results in Table 1 show that our adjusted versions of the DimeNet++ model yielded slight improvements for the energy predictions of in domain systems and out of domain cat-

alysts. It is likely that the improved performance for the exponential model was due to the periodic nature of the catalyst structures, and this property would have made it easier to capture the atomic distances with the Bessel function. Although improvements were shown for the average energy calculations, the base version of DimeNet++ was better at generating more energy values within the 0.02 (eV) threshold. These results suggest that on average, the adjusted models were better at performing the energy predictions, but there were fewer instances of it performing exceedingly well. In the leaderboard, our variational and exponential versions of DimeNet++ outperformed the CGCNN-5M-All model [10] in the MAE for OOD systems containing adsorbate and catalyst and OOD adsorbates, respectively. This is a promising result since our DimeNet++ versions were only trained on less than 10% of the training data available, while the other leaderboard groups had trained on the entire dataset. This shows that despite the limitations to our training data, our model was able to generalize better for these out of domain systems.

5. CONCLUSION

We have tested two versions of the DimeNet++ model, one version utilized a variational layer reducing uncertainty in the energy predictions, and the other version represented atomic distances with an exponential Bessel function. During training, both versions outperformed the base DimeNet++ model for predicting energies within the 0.02 (eV) threshold. During testing, improvements for the MAE were found for the in-domain systems and out-of-domain catalysts. On average, the adjusted versions were better at predicting the ground truth energy values, but they were less consistent than the base DimeNet++ model. Additionally, we have published our results on the Open Catalyst Project Leaderboard and found that our adjusted DimeNet++ models were able to outperform another state-of-the-art model in two of the evaluation metrics. Due to the limitations of our training set size, we would be interested in seeing how our adjusted models would fare after being trained with the full dataset in the future.

6. CONTRIBUTIONS

Robert Ramji focused on exploring dataset and physics-inspired improvements to ML models. Alexandria Do focused on hyperparameter tuning and running the model. David Najera focused on implementing model improvements.

7. REFERENCES

- [1] P. Wu, P. Du, Hui Zhang, and C. Cai. Graphene as a promising metal-free electrocatalyst for oxygen reduction reactions in acidic fuel cells: A dft study. *The Journal of Physical Chemistry*, 116(38):20472–20479, 2012.
- [2] Zhansheng Lu, Shuo Li, Chuang Liu, Chaozheng He, Xinwei Yang, Dongwei Ma, Guoliang Xu, and Zongxian Yang. Sulfur doped graphene as a promising metal-free electrocatalyst for oxygen reduction reaction: a dft study. *RSC Adv.*, 7:20398–20405, 2017.
- [3] Mitsuru Wakisaka, Hirokazu Suzuki, Satoshi Mitsui, Hiroyuki Uchida, and Masahiro Watanabe. Increased oxygen coverage at ptfе alloy cathode for the enhanced oxygen reduction reaction studied by expts. *The Journal of Physical Chemistry C*, 112(7):2750–2755, 2008.
- [4] Lowik Chanussot, Abhishek Das, Siddharth Goyal, Thibaut Lavril, Muhammed Shuaibi, Morgane Riviere, Kevin Tran, Javier Heras-Domingo, Caleb Ho, Weihua Hu, and et al. Open catalyst 2020 (oc20) dataset and community challenges. *ACS Catalysis*, 11(10):6059–6072, May 2021.
- [5] William L. Hamilton, Rex Ying, and Jure Leskovec. Representation learning on graphs: Methods and applications, 2018.
- [6] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.
- [7] Kristof T. Schütt, Pieter-Jan Kindermans, Huziel E. Saucedo, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions, 2017.
- [8] Johannes Klicpera, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs, 2020.
- [9] Johannes Klicpera, Shankari Giri, Johannes T. Margraf, and Stephan Günnemann. Fast and uncertainty-aware directional message passing for non-equilibrium molecules, 2020.
- [10] Tian Xie and Jeffrey C. Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical Review Letters*, 120(14), Apr 2018.

8. REPLY TO REVIEW

Critique of group 29 presentation – Enabling Fast Prediction of Catalytic Material Properties with Graph Neural Networks Critiques by group 18.

• Overall: A complete background research help this group to strongly explain the reason of applying ML on training catalytic material Properties in detail. Using angle data to predict the molecule's relaxed state suited the physical meaning well. Adopting Graphical Neural Networks, which is kind of analogous to the molecular topology, is a valid approach for feature extraction.

• Just a few comments and suggestions: 1. You showed the brilliant code of your project in the video. But you forgot to run code and display the test results. So I don't know whether your code can work out the image fitting those in your slide. Thank you for your comments. We are unable to run the code live because of the time it takes to run. None of the data is in human-readable format so we decided not to focus too much on trying to show that.

2. I am kind of confused by your result. In your result page, the energy mae curve is the same as loss curve except the label value. I tried to find your loss related code in the video. Maybe I missed, or you updated the loss result after finishing code video, so I didn't find it. The existing code (from the OCP) had the loss directly based on the MAE. We did not have a custom loss code and just relied on the MAE loss function from PyTorch directly.

In general, this final report is detailed and very excellent.

Critique of group 29 presentation - Prediction of catalytic material properties using Neural Networks. Critiques by group 23.

Your team implemented a very complex model and system, this is impressive. You did a good job explaining such a complicated system in the limited amount of time you had available to you. The following critiques are mostly based on advice on how to best present your model and results in your final paper. The subtleties of the model architecture are difficult to understand. Can you better explain these spherical harmonic functions used to represent angles between atoms? Why was this used? Were there any other approaches you considered?

Thank you for your comments. This is all much more detailed in the paper. Due to the time constraints of the presentation, we could not go into too much detail.

What are these different energy metrics, can you elaborate on what they mean and how they differ from each other? For the IS2RE project, the energy metrics used were Mean Absolute Error and Energy within Threshold. The MAE is a measure of the Mean Absolute Error from the ground truth energies (generated through DFT relaxation calculations). The Energy within Threshold measures what percentage of the predicted values were within a 0.02 (eV) threshold from the ground truth.

If the other open catalyst project entries are open source, look into what differentiates your model from the others, and suggest what elements of their approach leads to significant improvements over your model We looked at all the different models prior to choosing a baseline. Each model is highly complex by itself which is why we did not talk about them in the presentation. But we consider the advantages of other models in our paper. Model/Results How did you select the 10K subset of the dataset? Did you randomly sample from the entire dataset to ensure you have a proper distribution? The dataset was prepared by the challenge organizers. Train on a larger portion of the dataset, I think using 10K out of 400K dataset really hinders your ability to compare the results of your model with other models. We definitely agree. This is one aspect where we are at a disadvantage with respect to other teams (for example, Facebook or Deepmind) that have access to much more computational resources than we had. Try to incorporate some of the approaches used by other groups into your model and report on how your results changed. If not possible try comparing your results to theirs via plots or graphs. Thank you for the suggestion. We will look for better ways to compare to other groups.

Critique of group 29 presentation - Enabling Fast Prediction of Catalytic Material Properties with Graph Neural Networks Critiques by group 26.

Background Concise explanation of background and why machine learning will be able to advance the field Maybe briefly state some applications/ industries where your research topic will make a big impact (if you present at a conference this would be a waste but for this class students have many different research backgrounds) Thanks for the comments and suggestions. This is definitely explained in more detail in the paper. Dataset/ ML model You state the source of your dataset but it is unclear what the (physical) input and output to the ML model is as well as they shape of input/output We aimed to clarify this in the paper. The inputs are the atoms' coordinates (vector) and their respective atomic number (integer). Deep learning model seems involved with lots of concepts not learned in the course We had to find solutions outside of what is taught in the course because of the very specific nature of the problem's physics.

Unclear how graphical neural network works Graph Neural Networks are definitely not a very straightforward topic (there are dedicated courses to them) but we try our best to demystify them in the paper. We did not have too much time to do so in the presentation unfortunately. Unclear how and where variational autoencoder is implemented The VAE only served as an inspiration for how to model the output. Only the variational portion was leveraged, not the AE part. Results reference and comparison to other research that investigates same topic is interesting Can you notice effects of hyperparameter tuning and variational AE in results? We noticed a slight improvement in the results after performing hyperparameter tuning and adding the variational component. Slides

Well organized with helpful illustrations Adding slide numbers would help Thanks, we apologize for not including slide numbers.