Chemistry Insights for Large Pretrained GNNs

Katherine Xu[*]
Massachusetts Institute of Technology
kxu@alum.mit.edu

Janice Lan
Fundamental AI Research at Meta AI
janlan@meta.com

Abstract

There have been many recent advances in leveraging machine learning for chemistry applications. One particular task of interest is using graph neural networks (GNNs) on the Open Catalyst 2020 (OC20) dataset to predict the forces and energies of atoms and systems. While large GNNs have shown good progress in this area, we have little understanding of how or why these models work. In an attempt to gain a better understanding and increase our confidence that the models learn meaningful concepts that align with chemical intuition, we present perturbation analyses of GNN predictions on OC20, where we performed small changes on individual atoms and compared the model predictions before and after the changes. We provide visualizations of individual systems as well as analyses on general trends. We observed evidence that aligns with chemical intuition, including the importance of adsorbate atoms on the overall system, that modifying atomic numbers to neighbors of the same row of the periodic table causes less difference than other elemental changes, and a positive correlation between force magnitudes and energy changes.

1 Introduction

Renewable energy storage is important because energy demand does not always coincide with the periods of time when wind and solar plants generate power (3). Unfortunately, existing methods to store renewable energy at a large scale have a high cost due to the low efficiency of power conversion between electricity and hydrogen (32), and state-of-the-art electrocatalysts for increasing the efficiency of these reactions often use expensive metals (16). As a result, a key challenge of renewable energy storage is discovering efficient and low-cost catalysts for the power conversion reactions (32). There are many potential catalyst materials that can be created, but experimentally testing the billions of possible catalysts is not feasible (2). Density Functional Theory (DFT) (20) is a quantum mechanical method that can model properties for specific catalysts to provide insight into the best materials for testing. However, using DFT is computationally expensive, taking up to tens or hundreds of hours per calculation, which limits the exploration of new catalysts.

Instead of relying solely on DFT for molecular simulations, Machine Learning (ML) models can be trained on large catalysis datasets to predict reaction mechanisms for potential catalysts. One such dataset is the Open Catalyst 2020 (OC20) dataset (2), which contains more than 1.2 million DFT relaxations of adsorbates onto surfaces. The Structure to Energy and Forces (S2EF) task of OC20 predicts the energy and per-atom forces as computed by DFT for a given configuration of the adsorbate and slab atoms. ML models trained on OC20 for the S2EF task can use atom-level information, such as positions and atomic numbers, to predict the energy and forces. One class of ML models that have been used for these atomic simulations is Graph Neural Networks (GNNs) (31), where the nodes of the graph are atoms and the edges represent relationships between nearby atoms.

Recent large GNN models with millions of parameters have demonstrated good progress for the S2EF task. These models include the Crystal Graph Convolutional Neural Network (CGCNN) (25), SchNet

[*] Work done as part of an internship at Meta AI

NeurIPS 2022 AI for Science Workshop.

(18), DimeNet++ (7), SpinConv (21), GemNet (6), GemNet-OC (8), and Spherical Channel Network (SCN) (33). Among these models, GemNet-OC and SCN perform the best in terms of energy Mean Absolute Error (MAE), force MAE, force cosine, and energy and forces within threshold (EFwT) when evaluated on the OC20 test dataset for the S2EF task (33).

However, we have little insight into these large GNN models and their predictions for the OC20 S2EF task. We would like to understand what the models have learned as well as how and why they work. Through this effort, we can determine the types of errors that the models produce and whether their predictions are realistic in terms of chemistry and physics. We provide an approach to gain insight into these models that involves input perturbation analysis: we make slight changes to the atoms that are given as inputs to the model, and we compare the predicted energy and forces after the perturbation with those before the perturbation.

In this paper, we propose perturbation analyses as a useful approach to provide insights into these GNN models. We focus on the GemNet-OC model trained for the S2EF task, and we analyze this model using systems from the OC20 validation dataset. Our main findings are:

- Perturbations on adsorbate atoms affect the predicted energy of a system more than perturbations on surface atoms, which in turn have more of an effect than perturbations on subsurface slab atoms.
- Modifying the element of an atom to another element in the same row of the periodic table, especially closer neighbors, influences the predicted energy less than other elemental changes.
- There is a positive correlation between the original force magnitude on atoms and the changes in predicted energy when perturbing the atoms.

2 Related Work

Previous analyses of GNNs on OC20 have been mainly limited to visualizing the weights and activations of the layers of the models (18; 21; 33). In (21), the learned embeddings for different elements of the periodic table show that neighboring elements with similar properties have comparable weights. Other analyses include ablation studies that examine how different components of a model affect the force MAE (8). However, these analyses are insufficient for understanding what chemistry insights these GNNs have learned and why they generated certain predictions.

More generally, the explainability of GNNs, especially those used in chemistry, has been less explored than other ML models used for image and language tasks (19; 22; 24; 26; 28). One class of methods for interpreting model predictions is perturbation-based approaches, which compare changes in the outputs of a model given different perturbations of the inputs (9). Such techniques have been applied to GNNs to identify features in an input graph that contribute the most to model predictions (5; 12; 13; 17; 27; 29). For example, GNNExplainer finds a subgraph of the GNN and a subset of node features that are most related to a predicted output (27). While these perturbation-based methods have been used for GNNs overall, they have yet to be applied to GNNs on OC20.

There has also been work related to the explainability of models for chemistry applications such as materials and drug discovery (10; 11; 15; 23). Research in this area includes using intrinsically interpretable models rather than complex neural networks, as well as model explanation methods such as perturbation analyses and salience maps (14; 30). For example, there are frameworks for using probabilistic graphical models on chemistry and physics datasets in a way that provides interpretable results (4) and for explaining molecular property prediction models using graphical depictions (1).

3 Methods

3.1 OC20 validation dataset

We perform our analysis using a random sample of 10,000 systems from the OC20 validation dataset. For each system, we consider the final frame of the trajectory, as we want relaxed atoms. This set has 75 unique adsorbates, 4,202 unique bulk materials, and all of the 56 elements in OC20 (Figure 1). A system has on average 4.8 adsorbate atoms, 20 surface slab atoms, and 49.8 subsurface slab atoms.

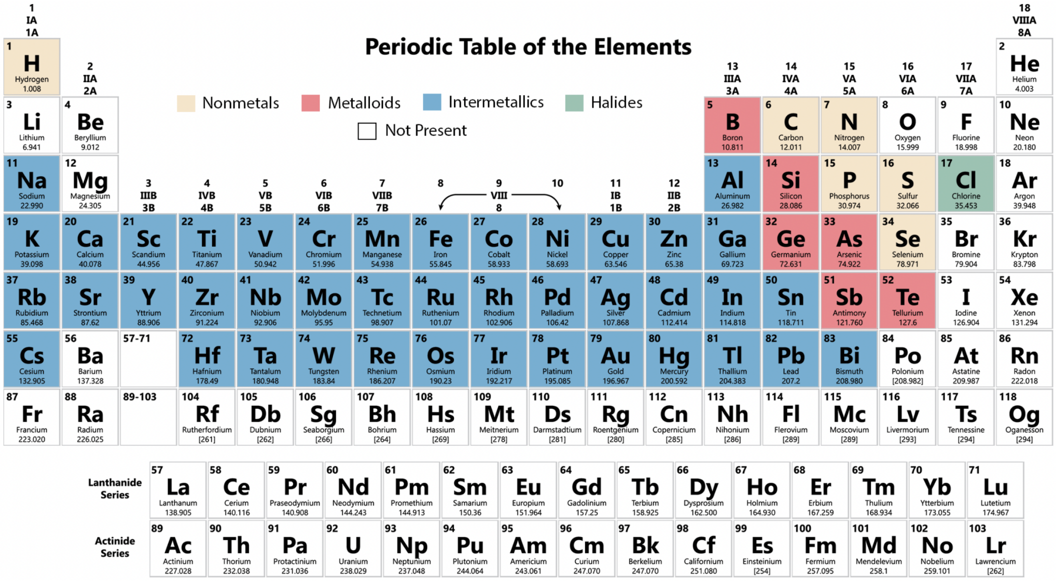

Figure 1: Periodic table highlighting the elements present in the OC20 dataset. Oxygen is also present in OC20 but only for adsorbate atoms.

3.2 Input perturbation analysis

GemNet-OC trained for the OC20 S2EF task uses atom-level information from a system, such as positions and atomic numbers, to make predictions for the system. The GemNet-OC model that we use here obtains 0.2411 eV energy MAE and 0.01901 eV/Å force MAE on the OC20 test set. To provide insights into this model, we perturb atoms in a system and compare the predicted energy and forces for the system before and after the perturbation. As a simplification, we perturb only a single atom in a system at a time, so that we can attribute changes in the model predictions to one specific change in the input. We generate visualizations that display the perturbation effects. Some of these visualizations show the detailed effects of perturbing atoms in an individual system, and others aggregate the effects of perturbing atoms across multiple systems to see more general trends.

Our perturbations are motivated by these questions:

1. How does perturbing adsorbate atoms, surface slab atoms, and subsurface slab atoms affect the predicted energy and forces differently?
2. How does perturbing an atom to another element affect the predicted energy and forces?
3. How do the forces on an atom relate to the change in predicted energy when the position of the atom is perturbed?

Since the space of possible perturbations is large, we answer these questions using straightforward perturbation methods for better model explainability. We can perturb an atom by removing it from the system, changing its spatial position, or changing its atomic number. When we remove an atom and return the model predictions on the perturbed input, we aim to observe how the presence of the atom affects the predictions for the system. We can also modify the position of an atom by shifting the atom in the XYZ directions to discover how the input positions of atoms affect the predictions.

In addition, changing the atomic number of an atom is equivalent to replacing the atom with another element on the periodic table that is present in the OC20 dataset because each element has a unique atomic number. We selectively choose to replace atoms with elements that are 1-hop or 2-hops away from the original element. To define elements that are 1-hop and 2-hops away, let $X_{(r,c)}$ represent an atom of the element in row (period) r and column (group) c of the periodic table. Then, let the k-hop elements of $X_{(r,c)}$ be $X_{(r-k,c)}$, $X_{(r+k,c)}$, $X_{(r,c-k)}$, and $X_{(r,c+k)}$, if they exist. For our perturbations, we consider elements that are only 1-hop ($k = 1$) and 2-hops ($k = 2$) away. We are also interested in distinguishing between the effects of replacing atoms with k-hop elements that are in the same row or same column as the original element. Figure 2 shows an example of 1-hop and 2-hop elements.

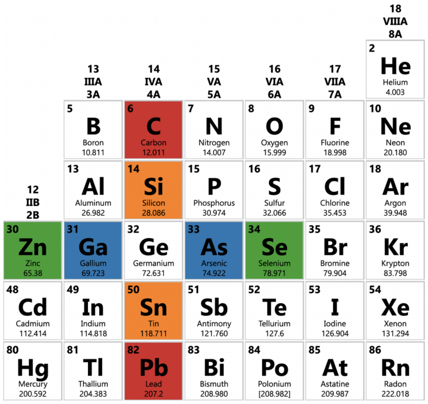

Figure 2: The 1-hop elements in the same row (blue), 1-hop elements in the same column (orange), 2-hop elements in the same row (green), and 2-hop elements in the same column (red) for Germanium.

4 Results

4.1 Perturbations on adsorbate and slab atoms

We are interested in determining how perturbations on individual adsorbate atoms and slab atoms affect the model predictions of a system differently. To understand this, we focus on perturbing individual atoms in a system by removing them. For each atom removed in the system, we obtain a single value for the change in the predicted energy of the system due to the perturbation. We collect these results over all of the adsorbate atoms, surface slab atoms, and subsurface slab atoms from the relaxed state of 10,000 OC20 validation systems. Since interactions primarily happen between the adsorbate atoms and surface slab atoms, we would expect perturbations on adsorbate atoms and surface slab atoms to influence the predictions more than perturbations on subsurface slab atoms.

We can visualize the energy changes due to the perturbation for specific systems as 3D scatter plots using Plotly. Figures 3 and 4 show example plots for OC20 systems. Examining the 3D plot of a system provides an evaluation of which atoms are affected more by the perturbation. These 3D plots include information on each atom, such as whether it is from a subsurface slab, surface slab, or an adsorbate, and more details on mouseover. They also allow users to rotate, translate, and scale the plot for different views of the system.

Table 1 shows the mean, standard deviation, and median for the magnitude of the energy change due to removing individual atoms in 10,000 OC20 systems. Perturbing adsorbate atoms led to the largest mean, standard deviation, and median for the change in energy, whereas perturbing subsurface slab atoms led to the smallest mean, standard deviation, and median. The mean values for subsurface, surface, and adsorbate atoms are greater than their respective medians, suggesting that the distributions are skewed right. Figure 5 shows the distributions of the energy changes when perturbing different atoms. When we use the magnitudes of the change in energy and plot the results on a log scale, we can more easily see that the changes in energy are greater for adsorbate atoms than for surface slab or subsurface slab atoms.

Based on these figures, the change in the predicted energy is greater when removing absorbate atoms than when removing surface slab atoms, and it is greater when removing surface slab atoms than when removing subsurface slab atoms. This observation suggests that perturbing adsorbate atoms influences the model predictions more than perturbing surface slab atoms, which in turn influences the predictions more than perturbing subsurface slab atoms.

	Mean	**St Dev**	**Median**
Subsurface	0.122	0.340	0.036
Surface	0.253	0.459	0.125
Adsorbate	0.890	0.734	0.721

Table 1: Statistics of the magnitude of the change in energy after removing individual atoms.

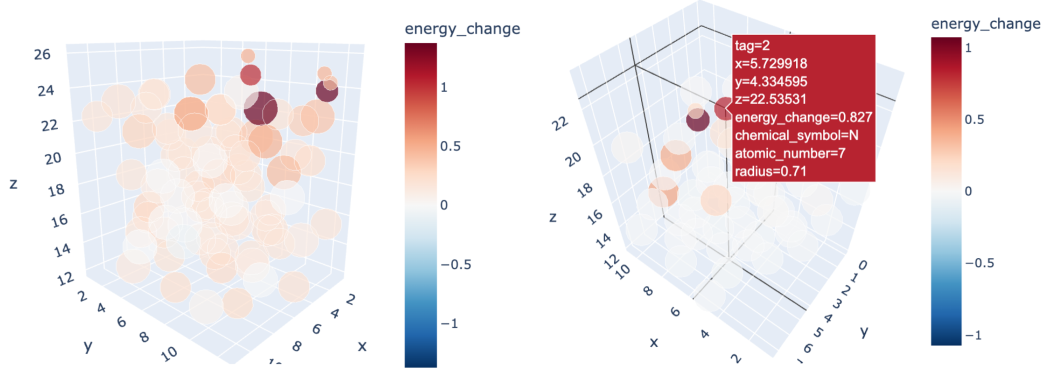

Figure 3: Plotly visualizations of 2 different arbitrary systems. You can drag to rotate the plot and mouse over an atom to get more details (right). An atom's color value indicates the energy change of the system when that atom is removed.

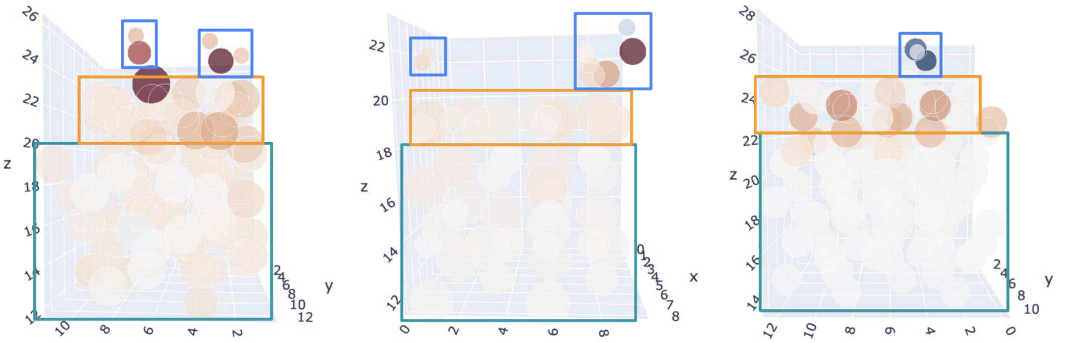

Figure 4: Plotly visualizations of 3 different arbitrary OC20 systems with rectangles indicating the approximate regions of adsorbate (blue), surface slab (orange) and subsurface slab atoms (green). Like Figure 3, the color of each atom indicates the energy change of the system when it is removed.

4.2 Perturbations on the element of a specific atom

In addition to perturbing atoms by removing them from systems, we are curious about how replacing an atom with another element on the periodic table affects the model predictions. We test this by changing the atomic numbers of individual atoms in a system to those of neighboring elements that are 1-hop or 2-hops away, and we record the mean change in energy for a given element over all atoms of the element in the system. We would expect that replacing an atom with adjacent elements on the periodic table has less of an effect on the predictions than replacing the atom with elements that are farther away due to changes in atomic properties. For example, the atoms of elements within the same row of the periodic table would likely have similar atomic sizes but different electron configurations, and the atoms of elements within the same column would likely have similar electron configurations but different atomic sizes.

Figure 6 shows the distributions of the mean change in predicted energy when we replace individual atoms of 2 selected elements with elements that are 1-hop away in the same row of the periodic table (1_hop_row), 1-hop away in the same column (1_hop_column), 2-hops away in the same row (2_hop_row), or 2-hops away in the same column (2_hop_column). The mean energy change when replacing atoms with 1-hop elements tends to be lower and closer to 0 than when replacing atoms with 2-hop elements, and the distributions of the mean energy change for 1-hop elements appear to have larger spreads than for 2-hop elements. Also, replacing atoms with k-hop elements in the same row generally led to lower mean energy changes than replacing them with elements in the same column. While we can't currently explain the root cause of the difference between row changes and column changes, we hope that future work can shed more light into this phenomenon, and at the very least this can be used as a debugging tool for any new ML models in chemistry.

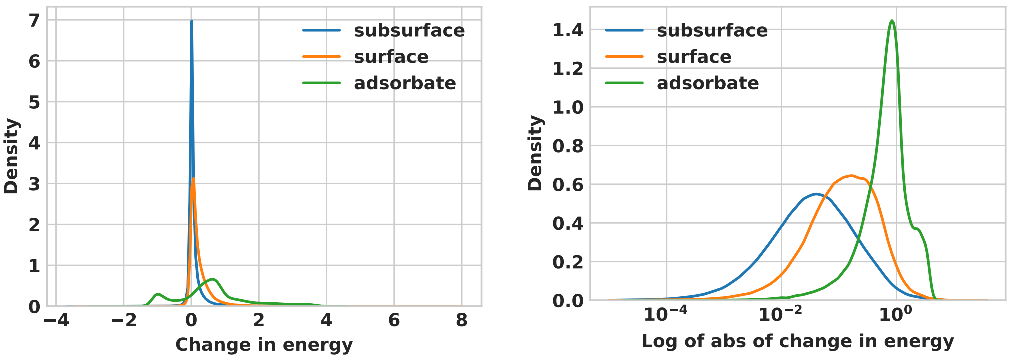

Figure 5: Density plots of the change in predicted energy when removing atoms. The plot on the right has the same results as the one on the left, but it log-scales the absolute values of the energy changes.

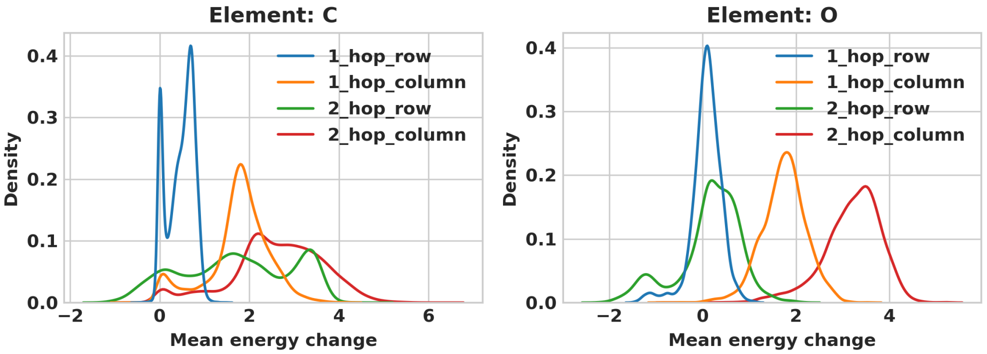

Figure 6: Mean change in the predicted energy of systems when replacing atoms of the specified elements with their 1-hop and 2-hop elements. 1_hop_row and 2_hop_row refer to replacing atoms with elements in the same row but different column on the periodic table. 1_hop_column and 2_hop_column refer to replacing atoms with elements in the same column but different row. For example, the 1_hop_row elements for carbon are boron and nitrogen, and the 1_hop_column element is silicon. See Figures 9-11 in the appendix for more examples of other elements.

Figure 7 compares more elements besides the 1-hop and 2-hop elements. We see similar trends in that replacing an atom with neighboring elements results in a lower energy change than replacing it with elements farther away. Overall, these observations suggest that the model learns some information about how nearby elements on the periodic table are likely to have similar properties.

4.3 How energy changes relate to forces

Since force is the gradient of energy, we want to examine the relationship between the forces on atoms and the change in predicted energy when we perturb the atoms. To explore this, we change the positions of individual atoms and observe the effect of the perturbation on the energy of the system. Generally, we would expect atoms with larger forces to increase the energy more if those atoms are moved toward those forces. However, this analysis is complicated by the fact that there are many atoms interacting with each other and there is not just a single force. Also, slightly moving any atom in these systems in any direction should generally increase the energy because we perturb atoms from a relaxed state. Consequently, we would not expect a clear correlation between the force magnitude on an atom and the energy change due to moving the atom, only general trends. To simplify our analysis somewhat, we focus on moving atoms toward their forces, in XYZ directions independently, and look for general trends. In particular, we expect a positive correlation between the force magnitude on an atom and the difference in the energy of a system due to changing its position.

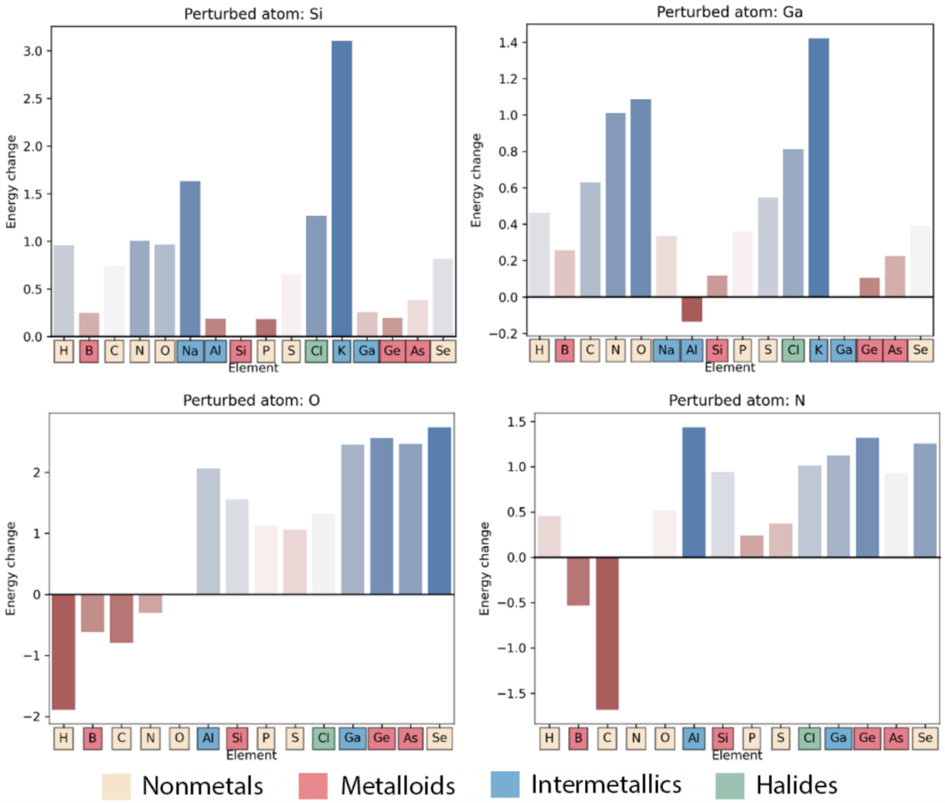

Figure 7: Bar plot of the change in predicted energy when changing one specific atom of one system to several elements. Darker blue bars indicate a higher energy change, and darker red bars indicate a lower energy change. The colors of the element symbols on the x-axis indicate the element type.

In fact, this is what we observe in Figure 8. For each atom, and for each of the XYZ directions, we move the atom 0.1Å in the opposite direction of the force. (We acknowledge that 0.1 was arbitrarily chosen as a small fraction of the typical distance between atoms, and we should try different distances in follow-up work.) For example, if a force on an atom is (-0.03, 0.2, -0.05), then we would perturb it by (0.1, 0, 0), (0, -0.1, 0), and (0, 0, 0.1) independently, getting respective energy differences $\Delta_x, \Delta_y, \Delta_z$. Next, we measure the correlations of $(0.03, \Delta_x), (0.2, \Delta_y), (0.05, \Delta_z), \ldots$ and similar values for all the atoms. Each correlation value is for one system, of which there are $3N$ data points, where N is the number of adsorbate and surface atoms in the system. We ignore subsurface atoms because the model is not trained on their forces.

The histogram in Figure 8 shows the distribution of the correlation coefficients over the 10,000 systems. This large number of positive correlations is impressive because the model we are using, GemNet-OC, is trained to predict forces directly rather than as a gradient of energy. Despite that, the model seems to have learned that relationship on its own.

As a side note, we also looked at the correlation between an atom's original force magnitude and the energy difference of removing that atom, but we did not see a strong correlation there, presumably because removing an atom is such a large change that it cannot be explained by forces as a gradient.

5 Conclusion

Large GNNs have demonstrated good progress on molecular datasets such as OC20, but we have little insight into whether these models learn relevant concepts that align with chemical intuition. Using GNNs for molecules is particularly difficult for interpretability because we do not have as much human intuition on the predictions as we do for image and language tasks. In this paper, we perform perturbations on individual atoms in 10,000 systems from the OC20 validation dataset, and we present

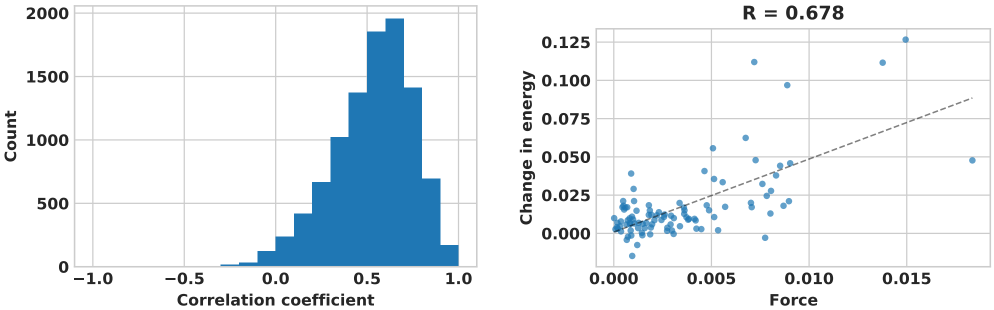

Figure 8: Left: Distribution of Pearson correlation coefficients over 10,000 OC20 validation systems. Each correlation value is calculated over each system, with points representing the force and energy change for each atom and each of the XYZ directions. The correlation is positive as expected. Right: A scatter plot of a typical system, where each point represents one atom's perturbation in one of the XYZ directions.

analyses of how the perturbations affect the GNN predictions of the systems using GemNet-OC. Our results so far increase our confidence that GemNet-OC is learning reasonable concepts.

We have only scratched the surface of the possible analyses that can be performed, but we hope that these contributions inspire additional work in the future to improve our understanding of GNNs for molecules. Future work should focus on the results from specific systems rather than aggregating results over many systems. If a system gives unexpected results, then we can probe whether the observations are due to the model learning something spurious or some characteristic of the system. Also, we can compare systems where ML predictions do well vs. don't do well (in terms of energy MAE, force MAE, and similar metrics in the OC20 S2EF task). Perhaps these perturbation analyses align better with chemical intuition in systems with ML predictions closer to ground truth, whereas erroneous ML predictions might show that the model did not learn the right chemical concepts.

Another interesting direction is performing these analyses on different models to observe how the perturbations affect the model predictions differently. For example, do certain models result in a greater change in the predictions when perturbing the atomic numbers of atoms, or do they place more weight on the spatial positions of atoms when making predictions? Furthermore, it would be beneficial to run DFT on the perturbed states of atoms so that we can compare the ML predictions with the DFT results, but this approach would be computationally expensive.

Acknowledgments and Disclosure of Funding

We would like to thank Adeesh Kolluru for helpful ideas and discussions for this work. Additionally, we'd like to thank the Open Catalyst Team for feedback in general and suggestions on the paper, including Abhishek Das, Siddharth Goyal, Anuroop Sriram, Zachary Ulissi, Minkai Xu, and Larry Zitnick.

References

[1] Marco Bertolini, Linlin Zhao, Djork-Arné Clevert, and Floriane Montanari. Beyond atoms and bonds: Contextual explainability via molecular graphical depictions. *ChemRxiv*, 2022.

[2] Lowik Chanussot, Abhishek Das, Siddharth Goyal, Thibaut Lavril, Muhammed Shuaibi, Morgane Riviere, Kevin Tran, Javier Heras-Domingo, Caleb Ho, Weihua Hu, Aini Palizhati, Anuroop Sriram, Brandon Wood, Junwoong Yoon, Devi Parikh, C. Lawrence Zitnick, and Zachary Ulissi. Open catalyst 2020 (OC20) dataset and community challenges. *ACS Catalysis*, 11(10):6059–6072, may 2021.

[3] Paul Denholm, Erik Ela, Brendan Kirby, and Michael Milligan. The role of energy storage with renewable electricity generation. 01 2010.

[4] Jinchao Feng, Joshua L. Lansford, Markos A. Katsoulakis, and Dionisios G. Vlachos. Explainable and trustworthy artificial intelligence for correctable modeling in chemical sciences. *Science Advances*, 6(42):eabc3204, 2020.

[5] Thorben Funke, Megha Khosla, Mandeep Rathee, and Avishek Anand. Zorro: Valid, sparse, and stable explanations in graph neural networks, 2021.

[6] Johannes Gasteiger, Florian Becker, and Stephan Günnemann. Gemnet: Universal directional graph neural networks for molecules, 2021.

[7] Johannes Gasteiger, Shankari Giri, Johannes T. Margraf, and Stephan Günnemann. Fast and uncertainty-aware directional message passing for non-equilibrium molecules, 2020.

[8] Johannes Gasteiger, Muhammed Shuaibi, Anuroop Sriram, Stephan Günnemann, Zachary Ulissi, C. Lawrence Zitnick, and Abhishek Das. How do graph networks generalize to large and diverse molecular systems?, 2022.

[9] Maksims Ivanovs, Roberts Kadikis, and Kaspars Ozols. Perturbation-based methods for explaining deep neural networks: A survey. *Pattern Recognition Letters*, 150:228–234, 2021.

[10] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Multi-objective molecule generation using interpretable substructures, 2020.

[11] Ritesh Kumar and Abhishek K. Singh. Chemical hardness-driven interpretable machine learning approach for rapid search of photocatalysts. *npj Comput. Mater.*, 7(1):197, dec 2021.

[12] Dongsheng Luo, Wei Cheng, Dongkuan Xu, Wenchao Yu, Bo Zong, Haifeng Chen, and Xiang Zhang. Parameterized explainer for graph neural network, 2020.

[13] Lucie Charlotte Magister, Dmitry Kazhdan, Vikash Singh, and Pietro Liò. Gcexplainer: Human-in-the-loop concept-based explanations for graph neural networks, 2021.

[14] Felipe Oviedo, Juan Lavista Ferres, Tonio Buonassisi, and Keith T. Butler. Interpretable and explainable machine learning for materials science and chemistry. *Accounts of Materials Research*, 3(6):597–607, 2022.

[15] Raquel Rodríguez-Pérez and Jürgen Bajorath. Explainable machine learning for property predictions in compound optimization. *Journal of Medicinal Chemistry*, 64(24):17744–17752, 2021. PMID: 34902252.

[16] Foteini M. Sapountzi, Jose M. Gracia, C.J. (Kees-Jan) Weststrate, Hans O.A. Fredriksson, and J.W. (Hans) Niemantsverdriet. Electrocatalysts for the generation of hydrogen, oxygen and synthesis gas. *Progress in Energy and Combustion Science*, 58:1–35, 2017.

[17] Michael Sejr Schlichtkrull, Nicola De Cao, and Ivan Titov. Interpreting graph neural networks for nlp with differentiable edge masking, 2020.

[18] Kristof T. Schütt, Pieter-Jan Kindermans, Huziel E. Sauceda, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. 2017.

[19] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, oct 2019.

[20] David S. Sholl and Janice A. Steckel. *Density Functional Theory: A Practical Introduction*. Wiley, 2009.

[21] Muhammed Shuaibi, Adeesh Kolluru, Abhishek Das, Aditya Grover, Anuroop Sriram, Zachary Ulissi, and C. Lawrence Zitnick. Rotation invariant graph neural networks using spin convolutions, 2021.

[22] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2013.

[23] Yuming Su, Yiheng Dai, Yifan Zeng, Caiyun Wei, Yangtao Chen, Fuchun Ge, Peikun Zheng, Da Zhou, Pavlo O. Dral, Cheng Wang, and et al. Interpretable machine learning of two-photon absorption. *ChemRxiv*, 2022.

[24] Zhengyang Wang, Xia Hu, and Shuiwang Ji. icapsnets: Towards interpretable capsule networks for text classification, 2020.

[25] 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.

[26] Fan Yang, Shiva K. Pentyala, Sina Mohseni, Mengnan Du, Hao Yuan, Rhema Linder, Eric D. Ragan, Shuiwang Ji, and Xia (Ben) Hu. XFake: Explainable fake news detector with visualizations. In *The World Wide Web Conference on - WWW '19*. ACM Press, 2019.

[27] Rex Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. Gnnexplainer: Generating explanations for graph neural networks, 2019.

[28] Hao Yuan, Haiyang Yu, Shurui Gui, and Shuiwang Ji. Explainability in graph neural networks: A taxonomic survey, 2020.

[29] Hao Yuan, Haiyang Yu, Jie Wang, Kang Li, and Shuiwang Ji. On explainability of graph neural networks via subgraph explorations, 2021.

[30] Xiaoting Zhong, Brian Gallagher, Shusen Liu, Bhavya Kailkhura, Anna Hiszpanski, and T. Yong-Jin Han. Explainable machine learning in materials science. *npj Computational Materials*, 8(1):204, 2022.

[31] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *AI Open*, 1:57–81, 2020.

[32] C. Lawrence Zitnick, Lowik Chanussot, Abhishek Das, Siddharth Goyal, Javier Heras-Domingo, Caleb Ho, Weihua Hu, Thibaut Lavril, Aini Palizhati, Morgane Riviere, Muhammed Shuaibi, Anuroop Sriram, Kevin Tran, Brandon Wood, Junwoong Yoon, Devi Parikh, and Zachary Ulissi. An introduction to electrocatalyst design using machine learning for renewable energy storage, 2020.

[33] C. Lawrence Zitnick, Abhishek Das, Adeesh Kolluru, Janice Lan, Muhammed Shuaibi, Anuroop Sriram, Zachary Ulissi, and Brandon Wood. Spherical channels for modeling atomic interactions, 2022.

A Appendix

We present additional figures that show the distributions of the mean energy change when we replace an atom with nearby elements that are 1-hop or 2-hops away.

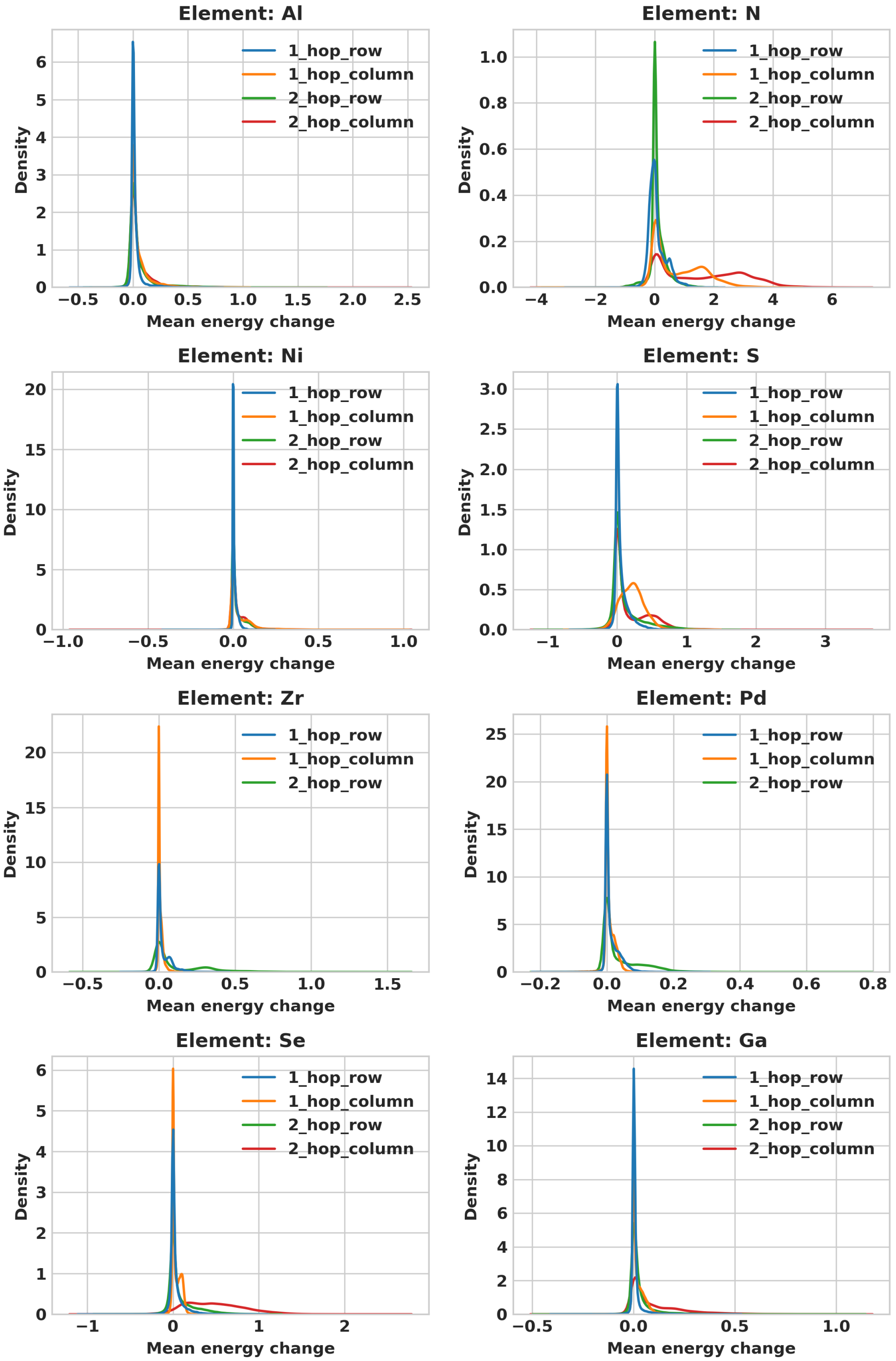

Figure 9: 1-hop and 2-hop distributions for eight of the most common elements in our dataset.

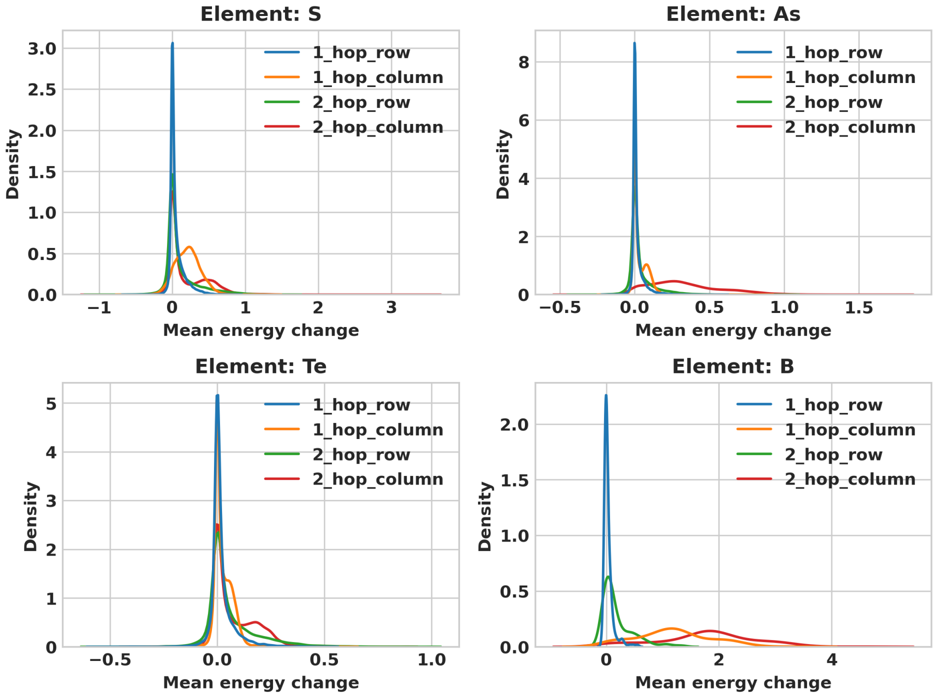

Figure 10: Examples of elements where the column hops change energy more than the row hops.

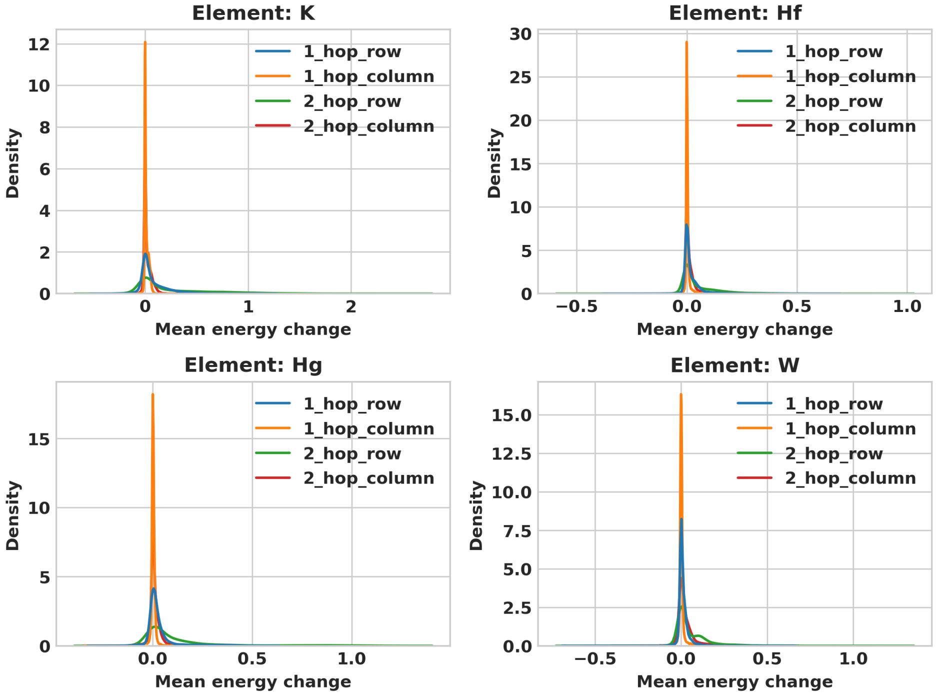

Figure 11: Examples of elements where the row hops change energy more than the column hops.