Do Better QM9 Models Extrapolate as Better Quantum Chemical Property Predictors?

Yucheng Zhang*
The University of Tokyo
7-3-1 Hongo, Bunkyo City, Tokyo, Japan 113-8654
yuchengzhang2017@gmail.com

Nontawat Charoenphakdee
Preferred Networks, Inc.
Otemachi Bldg. 1-6-1 Otemachi, Chiyoda-ku, Tokyo, Japan 100-0004
nontawat@preferred.jp

So Takamoto
Preferred Networks, Inc.
Otemachi Bldg. 1-6-1 Otemachi, Chiyoda-ku, Tokyo, Japan 100-0004
takamoto@preferred.jp

Abstract

The implicit hypothesis behind benchmarking on the gold standard QM9 dataset is that, model improvement on small and concentrated molecules implies improvement in generalization as better quantum chemical property (QCP) predictors. This extrapolation ability for deep learning (DL) models is highly useful for various real-world applications, yet the related investigation remains quite limited. The goal of this paper is to promote the development of DL models that can extrapolate beyond the in-domain dataset, and can handle larger molecules than that of the training data. To achieve this goal, a cross-dataset benchmark of training models on **QM**9 dataset and testing on **AL**chemy datasets with **L**arger molecular size (**QMALL**) is proposed. Experimental results using recent DL methods are provided to investigate their out-of-distribution (OOD) behavior. Analysis of the overall performance drop, model ranking inconsistency, aggregation method selection, and error patterns created new insights into this OOD extrapolation issue, highlighting its challenge for the research community to tackle.

1 Introduction

Recently, researchers leveraged DL to accelerate prediction of molecular properties that are crucial in physics, chemistry, material science, and biology [1, 2]. Although state-of-the-art (SOTA) DL architectures have achieved remarkable success in QCP prediction benchmarks such as QM9 [3, 4], most of the publications on this topic developed and compared the models with the test mean absolute error (MAE) as the only criterion [5–8]. However, in-domain better accuracy may not imply the realistic adoption as a better QCP predictor [9], since the QCP predictor is often applied in unknown chemical regions during molecule screening and optimization. The importance of OOD extrapolation is obvious considering the large size of the molecular chemical space, which is estimated to be in the order of 10^{60} [10]. If DL models are sensitive to slight distribution change of molecular structures or

*This research was conducted during his internship at Preferred Networks.

Machine Learning and the Physical Sciences workshop, NeurIPS 2022.

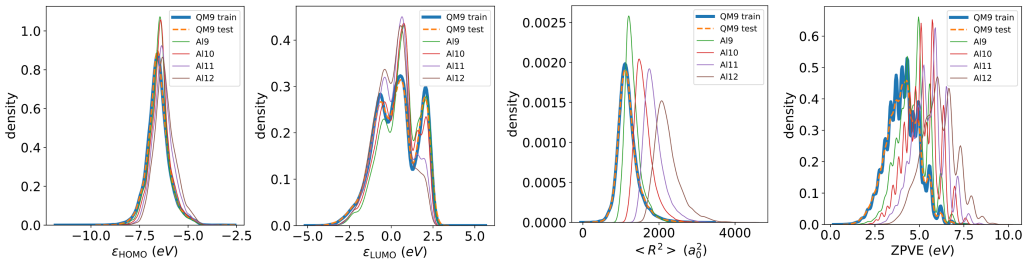

Figure 1: Property distribution comparison between QM9 train/test splits and Alchemy subsets [14] with 9 to 12 heavy atoms after excluding charged molecules and molecules with S and Cl (See Figure A4 in Appendix for entire plots of 12 properties).

properties, the meaning of QCP prediction benchmark will be severely limited to demonstrating the expressiveness of DL models.

The gold standard QM9 benchmark dataset (elements of HCNOF with up to 9 heavy atoms) has several limitations for OOD evaluation. **First**, following the ordinary ML settings, the property distributions between train/valid/test splits are highly similar (see Figure 1). Therefore, the test performance of the trained model might not be able to reveal the OOD performance. **Second**, the structures used in train/valid/test split share the same bias due to the same flowchart for preparing the optimized structures. However, in reality, it is unachievable for different researchers in different institutes to follow the same flowchart (see Figure A5 and Table A3 in Appendix for details). **Third**, although molecules in QM9 are relatively small compared with other well-known datasets (e.g., PubChem [11], ChEMBL [12], ZINC [13]), size-related generalization remains less explored. Extrapolation from small molecules to larger ones is very important from the perspective of calculation cost, where DFT suffers from the $O(N^3)$ scaling bottleneck. Besides, empirically the molecular weight (another indicator of molecular size) of medicinal chemistry compounds rises steadily in recent years [14].

Related Work: PC9 (elements of HCNOF with up to 9 heavy atoms) [15] are sampled from PubChemQC [16] for cross-dataset extrapolation. However, PubChemQC only shares 3/12 properties with QM9, while the systematic difference of interatomic distances is observed owing to DFT optimization at different accuracy levels [15]. Besides, the comparison methods are kernel ridge regression, elastic net, Gaussian process regression and SchNet, which are no longer SOTA models.

Changes in model trends between datasets have already been witnessed in fields such as computer vision [17]. In the domain of DL and quantum chemistry, the correlation of model force MAEs between datasets is studied [18], where model choices are found not consistent between datasets. However, datasets consisting of different molecular properties are not focused on.

2 QMALL benchmark

To mimic the OOD working condition of the practical QCP predictor, a cross-dataset extrapolation task is proposed[2]. In this QMALL benchmark, the training data is QM9 but the test data are different splits of the Alchemy dataset according to the molecular size.

Originally, the shared 12/12 properties in QM9 and Alchemy are calculated by density functional theory (DFT) at the same accuracy level of B3LYP/6-31G(2df,p). However, differences caused by software (QM9 used Gaussian [19] but Alchemy used PySCF [20]) are still too big for studying extrapolation of some properties [14], and thus modifications are performed as follows.

Step 1: The Alchemy dataset is stratified into Al9, Al10, Al11 and Al12 subsets based on the number of heavy atoms, while charged molecules and molecules with S and Cl are excluded to conform with QM9. **Step 2**: The 12 QCPs in Alchemy are recomputed by Gaussian [19] at the same accuracy level with QM9 for compensating the DFT systematic difference. **Step 3**: For each property, a linear transformation is applied to eliminate the systematic differences, where the weight and bias are calculated based on 500 data points per dataset (See Tables A1 and A2 in Appendix for the coefficients). **Step 4**: The DL models of interest are trained on QM9. **Step 5**: The models with the best performance on the test split of QM9 are picked for inference on Al10, Al11 and Al12.

[2]The re-computed 12 properties of Al9-12 by Gaussian for QMALL benchmark is provided at `https://github.com/YZHANG1996/QMALL.git`

Table 1: OOD extrapolation performance of models trained on QM9 and inferred on Alchemy10-12. SchNet, MEGNet and DimeNet++ models are taken from the original repository, where MEGNet is the MEGNet-simple model without auxiliary information. Boldfaced value indicates that the model obtains the best performance with respect to one dataset and property.

Target (Unit)		μ (D)	α (a_0^3)	ϵ_{HOMO} (meV)	ϵ_{LUMO} (meV)	$\Delta\epsilon$ (meV)	$\langle R^2 \rangle$ (a_0^2)	ZPVE (meV)	$U0$ (meV)	U (meV)	H (meV)	G (meV)	Cv $(\frac{cal}{molK})$
SchNet	QM9	0.033	0.235	41	34	63	0.073	1.7	14	19	14	14	0.033
	Al10	0.054	0.302	77.62	76.58	126.90	**0.744**	3.68	216.4	213.3	214.7	212.7	0.161
	Al11	0.099	0.436	102.51	105.55	171.61	5.018	5.30	223.2	219.9	219.8	221.4	0.160
	Al12	0.131	0.644	123.77	139.41	220.19	6.586	6.882	239.6	235.7	232.7	244.3	0.224
MEGNet	QM9	0.050	0.081	43	44	66	0.302	1.43	12	13	12	12	0.029
	Al10	0.109	0.29	72.10	113.41	116.05	1.842	3.616	229.6	232.0	221.6	221.8	0.180
	Al11	0.226	0.47	97.51	233.37	157.16	6.138	6.535	255.6	264.2	242.0	253.2	0.350
	Al12	0.359	0.77	132.45	385.22	**201.69**	6.821	13.267	291.5	321.8	291.1	304.2	0.855
EGNN	QM9	0.029	0.071	29	25	47	0.107	1.55	12.7	10.8	12.7	12.56	0.031
	Al10	0.061	7.395	**57.07**	70.44	**94.53**	64.342	5.61	6062.3	5764.5	6298.5	4363.3	0.169
	Al11	0.108	14.713	**79.25**	105.32	**139.42**	139.20	14.24	13084.2	11811.9	13072.4	9111.7	0.169
	Al12	0.136	21.710	**99.98**	168.57	214.36	213.28	29.06	20702.8	18274.7	20151.4	14086.8	0.231
DimeNet++	QM9	0.0297	**0.0435**	24.6	19.5	**32.6**	0.331	**1.21**	6.32	6.28	6.53	7.56	0.023
	Al10	0.060	**0.230**	680.67	57.62	719.49	62.327	**2.86**	**202.6**	**203.0**	**202.6**	**201.9**	**0.150**
	Al11	0.104	**0.301**	1366.34	85.57	1424.09	141.192	**4.48**	**208.9**	**208.9**	**207.4**	**204.9**	**0.150**
	Al12	0.115	**0.371**	2070.23	112.44	2145.81	252.42	**6.22**	**230.2**	233.8	**226.1**	**222.4**	**0.224**
ET	QM9	**0.010**	0.044	**23.2**	**17.3**	38.4	**0.034**	1.64	**6.15**	**6.14**	**6.04**	**7.21**	**0.022**
	Al10	**0.036**	0.242	211.13	**51.21**	238.19	0.788	3.85	205.6	205.2	205.7	206.7	0.254
	Al11	**0.066**	0.325	592.94	**71.23**	633.03	**4.767**	7.38	212.9	211.5	212.4	212.4	0.569
	Al12	**0.065**	0.416	1023.32	**93.61**	1083.53	**4.734**	15.93	232.3	**232.3**	234.1	232.3	0.948

It is worth mentioning that, the molecular size in the proposed benchmark is slightly larger than QM9, which behaves as a challenging yet reasonable task for DL models compared with inference on other molecular systems [21]. Thus, this benchmark might not be suitable if the goal of the user is to evaluate the extrapolation ability on molecules that are much larger than those in the training dataset.

3 Results and analysis

In this section, we report the performance of QMALL benchmark for five well-known DL models: SchNet [5], MEGNet [22], DimeNet++ [7], Equivariant graph neural networks (EGNN) [6], and Equivariant transformer (ET) [23]. The experiment codes were run on NVIDIA Tesla V100 GPUs. QM9 is under CC BY-NC SA 4.0 license and Alchemy is under MIT license.

3.1 Analysis of overall performance

Table 1 shows the test MAE with respect to different properties and datasets for each model.

The performance drops as the molecular size increases. This phenomenon can be observed in most cases for all models and properties. For instance, all models' performance for μ drops as the molecular size gets larger (from QM9 to Al12). The results suggest the difficulty of QMALL. Notably, for $U0$, U, H, G, the OOD MAEs of all models become at least 11 times larger (e.g., 19 meV to 213.3 meV), which suggests that further reducing the in-domain MAE to be lower than 6 meV might not effectively improve the extrapolation accuracy.

Better in-domain (QM9) prediction accuracy cannot always imply a better QCP predictor for OOD extrapolation. While ET gives the best performance for all datasets in μ and ϵ_{LUMO}, its superior in-domain accuracy on many other properties can no longer retain with larger molecules. For instance, DimeNet++ becomes more effective than ET in $U0$, U, H, G, and C_v although its performance is worse than ET in QM9. ET is also outperformed by EGNN in ϵ_{HOMO} on OOD datasets. Another example is for EGNN, although its prediction MAE of α on QM9 is three times smaller than that of SchNet, its OOD MAE is over 20 times larger on Al10-12.

The performance ranking for the same property is not consistent for different test datasets. For $\Delta\epsilon$, the best model is DimeNet++ on QM9, EGNN on Al10 and Al11, and MEGNet on Al12. This suggests model selection for extrapolation by using QM9 training dataset might not be easy.

With recent progress, the in-domain MAE of ϵ_{HOMO} on QM9 dropped from 41 meV (SchNet) to 24.6 meV (DimeNet++), and was further reduced to 23.2 meV (ET). However, if considering the OOD MAE on Al10/11/12, DimeNet++ is 8.8/13.3/16.7 times larger than SchNet, while ET is 2.7/5.8/8.3 times larger than SchNet, respectively. On Al12, MAE of DimeNet++ is over 2000 meV while MAE of ET is 1023 meV, which are broken for practical use. This observation calls for caution when researchers want to employ the SOTA DL models for inference on their datasets. Note that although ϵ_{HOMO} and ϵ_{LUMO} are both energies of the molecular orbitals, the above collapsed predictions never appear for ϵ_{LUMO}. This might hint how black box DL models treat the two properties differently.

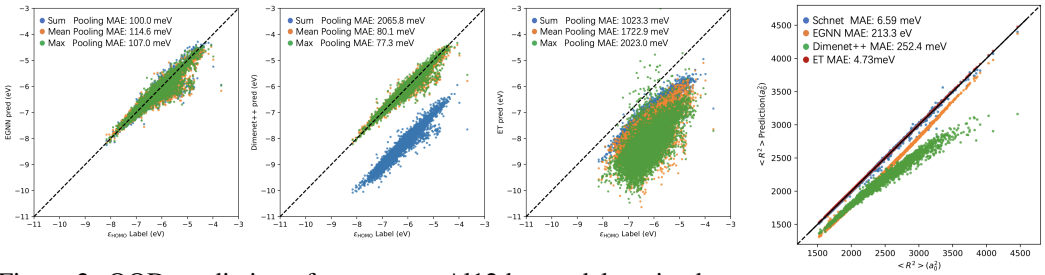

Figure 2: OOD prediction of ϵ_{HOMO} on Al12 by models trained on QM9. Selection of suitable aggregation method depends both on characteristics of the physical property and the model architecture, while ET with the highest in-domain prediction accuracy extrapolates worst.

Figure 3: OOD prediction of $\langle R^2 \rangle$ on Al12 by models trained on QM9.

3.2 Analysis of the importance of aggregation methods

From the perspective of physics, properties can be divided into intensive (e.g., ϵ_{HOMO}, ϵ_{LUMO}, $\Delta\epsilon$) and extensive ones (e.g., U, $U0$, H and G). It has been suggested by the existing works [5, 24] that average pooling should be applied for the intensive properties while the sum pooling should be applied for the extensive properties. On the other hand, sum pooling gives better in-domain accuracy for some GNN-based models. For instance, EGNN used sum pooling as the only default aggregation method for the QM9 tasks. Another example is the provided DimeNet++ model in the original GitHub[3]. It was trained with sum pooling for all 12 properties. Here, as a showcase, using an intensive property ϵ_{HOMO}, which is the energy of the **highest** occupied molecular orbital, we reported the effect of aggregation after retraining EGNN, DimeNet++ and ET with sum/mean/max pooling on QM9. Figure 2 shows the performance of different models and pooling methods on Al12.

For ET, which ranks best in QM9, all aggregation methods are far from making the OOD inference useful since the MAE is beyond 1000 meV. For EGNN, sum pooling gives the best performance while OOD accuracy of all aggregation methods are within 100-115 meV.

For DimeNet++, the in-domain MAE with mean pooling (25.53 ± 0.51 meV) and max pooling (28.06 ± 0.84 meV) are slightly worse than that with sum pooling (25.04 ± 0.42 meV) according to five repeated experiments. If the in-domain MAE on the test set is the only criterion of model development, sum pooling will be determined as the more suitable aggregation method. However, if the trained models are adopted for inference on larger molecules (Al10-12), the sum pooling scheme of DimeNet++ renders the worst ranking among all models and aggregation methods, as shown in Table 1. Thus, it is surprising to observe that by only changing the aggregation to mean/max pooling (where MAE becomes 77-81 meV), DimeNet++ outperforms other models.

From the results, the best aggregation method does not only depend on the characteristics of the property, but also on the model architecture. This finding challenges the current wisdom that mean pooling should be applied to the intensive property. Besides, the OOD behavior of DimeNet++ with sum pooling calls for necessity of examining the extrapolation ability of proposed DL models for practical adoption as better QCP predictors.

3.3 Analysis of error patterns

Although MAEs for the Al12 are larger than QM9, it can be observed from the error patterns that a simple modification of prediction output can improve the performance. For instance, in Figure 2, the large error of DimeNet++ with sum pooling can be reduced by adding an appropriate constant. It is worth mentioning that the error pattern of ϵ_{HOMO} for DimeNet++ with sum pooling is not the unique case, e.g., α for EGNN also shares this pattern. It is also found that changing the aggregation method of EGNN cannot eliminate such errors. Although calculating constants for the target dataset actively may improve the accuracy, it can be cumbersome since different constants should be calculated for different datasets (e.g., molecules with different numbers of heavy atoms, see A2 and A3).

Adding constant values is not sufficient for all properties and models. Taking $\langle R^2 \rangle$ (the electronic spatial extent) as an example in Figure 3. Although the linear transformation can improve the performance of EGNN, it might not improve the performance of DimeNet++. This result emphasizes

[3] `https://github.com/gasteigerjo/dimenet`

the different model behaviors under OOD extrapolation, and also the further study on useful prediction adjustments towards an effective OOD QCP predictor.

4 Conclusions

The QMALL benchmark was proposed to evaluate the extrapolation ability of machine learning models in predicting quantum chemical properties of organic molecules. Experimental results with five well-known DL models show that the test performance drops significantly as the molecular size increases in most cases, better in-domain accuracy may not imply better OOD accuracy, and the performance ranking can be inconsistent between the in-domain and OOD prediction tasks. Overall, it can be observed that the QMALL benchmark is currently very challenging for the current SOTA DL methods compared with the benchmark on QM9 dataset. Furthermore, the importance of aggregation methods and different error patterns were analyzed to suggest several future directions of improving the DL methods for the extrapolation task. We believe that the proposed QMALL benchmark can be useful for further development of DL methods that can extrapolate well in QCP prediction.

5 Broader impact

In practical molecule design or optimization, a generative model is often coupled with a QCP predictor or DFT evaluator. As the molecule generator learns to create better molecules, the target property is gradually optimized. Under such circumstance, our expectation of molecules with better properties will continuously "push" the property distribution towards regions of chemical space beyond the training set. If QCP predictors cannot work well with slightly different property distributions, the error will accumulate when guiding the molecule generation and thus resulting in failure of continuous optimization. Although Al10-12 and QM9 are both datasets of organic molecules, since the property distributions between Al10-12 and QM9 are more different than those between train/valid/test splits of QM9 (entire plots are shown in Figure A4 in Appendix), QMALL can be used as one crucial test for revealing the OOD extrapolation capability of the proposed DL models.

ϵ_{HOMO}, ϵ_{LUMO} and $\Delta\epsilon$ are important for applications such as batteries, semiconductors, alloys, electronic devices, and photovoltaic materials. With progress in recent five years, the in-domain MAE of ϵ_{HOMO} on QM9 dropped from 41 meV (SchNet) to 24.6 meV (DimeNet++), and was further reduced to 23.2 meV (ET). However, not so much attention has been paid to the capability of DL methods to extrapolate beyond the in-domain dataset, which is undoubtedly important for real-world applications. This paper proposes an extrapolation benchmark QMALL, which reveals that there is a gap between having preferable performance for the in-domain dataset and OOD dataset. Also, it exhibits huge room for improvement of DL methods to excel in the extrapolation task.

One direction towards better generalization is to pre-train models on huge unlabeled datasets, and fine-tune for downstream tasks [25–27]. It is interesting to see if OOD prediction accuracy will increase a lot after the pre-trained models are fine-tuned on QM9, since the molecular representation might be better after grasping more information of the chemical space.

Overall, we believe that the proposed benchmark QMALL, which takes the extrapolation capability into account, will promote the research direction towards developing machine learning methods that can perform OOD generalization effectively.

For negative societal impact, in our understanding, this paper does not contain any information that is harmful to anyone in the sense that it does not contain personally identifiable information or offensive content.

References

[1] Michael I Jordan and Tom M Mitchell. Machine learning: Trends, perspectives, and prospects. Science, 349(6245):255–260, 2015.

[2] Keith T Butler, Daniel W Davies, Hugh Cartwright, Olexandr Isayev, and Aron Walsh. Machine learning for molecular and materials science. Nature, 559(7715):547–555, 2018.

[3] Lars Ruddigkeit, Ruud Van Deursen, Lorenz C Blum, and Jean-Louis Reymond. Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17. Journal of chemical information and modeling, 52(11):2864–2875, 2012.

[4] Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole Von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. Scientific data, 1(1):1–7, 2014.

[5] Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. Advances in neural information processing systems, 30, 2017.

[6] Vıctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E (n) equivariant graph neural networks. In International conference on machine learning, pages 9323–9332. PMLR, 2021.

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

[8] Kristof Schütt, Oliver Unke, and Michael Gastegger. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In International Conference on Machine Learning, pages 9377–9388. PMLR, 2021.

[9] Masashi Tsubaki and Teruyasu Mizoguchi. Quantum deep field: data-driven wave function, electron density generation, and atomization energy prediction and extrapolation with machine learning. Physical Review Letters, 125(20):206401, 2020.

[10] Jean-Louis Reymond. The chemical space project. Accounts of Chemical Research, 48(3): 722–730, 2015.

[11] Sunghwan Kim, Paul A Thiessen, Evan E Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A Shoemaker, et al. Pubchem substance and compound databases. Nucleic acids research, 44(D1):D1202–D1213, 2016.

[12] Anna Gaulton, Louisa J Bellis, A Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, et al. Chembl: a large-scale bioactivity database for drug discovery. Nucleic acids research, 40(D1):D1100–D1107, 2012.

[13] John J Irwin and Brian K Shoichet. Zinc- a free database of commercially available compounds for virtual screening. Journal of chemical information and modeling, 45(1):177–182, 2005.

[14] Guangyong Chen, Pengfei Chen, Chang-Yu Hsieh, Chee-Kong Lee, Benben Liao, Renjie Liao, Weiwen Liu, Jiezhong Qiu, Qiming Sun, Jie Tang, et al. Alchemy: A quantum chemistry dataset for benchmarking ai models. International conference on learning representations, 2019.

[15] Marta Glavatskikh, Jules Leguy, Gilles Hunault, Thomas Cauchy, and Benoit Da Mota. Dataset's chemical diversity limits the generalizability of machine learning predictions. Journal of cheminformatics, 11(1):1–15, 2019.

[16] Maho Nakata and Tomomi Shimazaki. Pubchemqc project: a large-scale first-principles electronic structure database for data-driven chemistry. Journal of chemical information and modeling, 57(6):1300–1308, 2017.

[17] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 2661–2671, 2019.

[18] 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? Transactions on Machine Learning Research, 2022.

[19] MJ ea Frisch, GW Trucks, HB Schlegel, GE Scuseria, MA Robb, JR Cheeseman, G Scalmani, VPGA Barone, GA Petersson, HJRA Nakatsuji, et al. Gaussian 16, 2016.

[20] Qiming Sun, Timothy C Berkelbach, Nick S Blunt, George H Booth, Sheng Guo, Zhendong Li, Junzi Liu, James D McClain, Elvira R Sayfutyarova, Sandeep Sharma, et al. Pyscf: the python-based simulations of chemistry framework. Wiley Interdisciplinary Reviews: Computational Molecular Science, 8(1):e1340, 2018.

[21] Bing Huang and O Anatole von Lilienfeld. Quantum machine learning using atom-in-molecule-based fragments selected on the fly. Nature Chemistry, 12(10):945–951, 2020.

[22] Chi Chen, Weike Ye, Yunxing Zuo, Chen Zheng, and Shyue Ping Ong. Graph networks as a universal machine learning framework for molecules and crystals. Chemistry of Materials, 31 (9):3564–3572, 2019.

[23] Philipp Thölke and Gianni De Fabritiis. Torchmd-net: Equivariant transformers for neural network based molecular potentials. International Conference on Learning Representations, 2022.

[24] Ziteng Liu, Liqiang Lin, Qingqing Jia, Zheng Cheng, Yanyan Jiang, Yanwen Guo, and Jing Ma. Transferable multilevel attention neural network for accurate prediction of quantum chemistry properties via multitask learning. Journal of Chemical Information and Modeling, 61(3):1066–1082, 2021.

[25] Xiaomin Fang, Lihang Liu, Jieqiong Lei, Donglong He, Shanzhuo Zhang, Jingbo Zhou, Fan Wang, Hua Wu, and Haifeng Wang. Geometry-enhanced molecular representation learning for property prediction. Nature Machine Intelligence, 4(2):127–134, 2022.

[26] Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-mol: A universal 3d molecular representation learning framework. ChemRxiv, 2022.

[27] Sheheryar Zaidi, Michael Schaarschmidt, James Martens, Hyunjik Kim, Yee Whye Teh, Alvaro Sanchez-Gonzalez, Peter Battaglia, Razvan Pascanu, and Jonathan Godwin. Pre-training via denoising for molecular property prediction. arXiv preprint arXiv:2206.00133, 2022.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See the end of Section 2.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See broader impact section.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See the beginning of Section 3.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]

(b) Did you mention the license of the assets? [Yes] See the beginning of Section 3.
(c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Appendix

In Appendix, we provide supplementary tables and figures to support the main part of the paper, including (1) Linear regression coefficients for compensating DFT differences (Tables A1 and A2), (2) Error pattern of R2 (Figure A1), (3) Error pattern of α for EGNN (Figure A2), (4) Error pattern of HOMO for DimeNet++ (Figure A3), (5) Property distribution comparison between QM9 and Alchemy datasets (Figure A4), (6) Performance drop due to bias of preparing the optimized molecular geometries (Table A3), and (7) Visualization of different local optima for the molecule with the same SMILES (simplified molecular input line entry specification) (Figure A5).

Table A1: Linear regression coefficients for compensating the DFT difference for μ, α, ϵ_{HOMO}, ϵ_{LUMO}, $\Delta\epsilon$, and $\langle R^2 \rangle$ properties.

Target		μ	α	ϵ_{HOMO}	ϵ_{LUMO}	$\Delta\epsilon$	$\langle R^2 \rangle$
Weight	Al10	0.99792	0.94633	0.99488	0.99915	0.99777	0.99998
	Al11	0.99711	0.94704	0.98807	0.99919	0.99607	0.99893
	Al12	0.9976	0.95114	0.99517	1.00059	0.99898	0.99997
Bias	Al10	−0.00648	4.54114	−0.10756	−0.07546	0.01406	−0.14517
	Al11	−0.00509	4.94485	−0.15179	−0.07515	0.02704	1.84113
	Al12	−0.00358	4.96377	−0.1064	−0.07524	0.00811	−0.16335

Table A2: Linear regression coefficients for compensating the DFT difference for ZPVE, $U0$, U, H, G, and Cv properties.

Target		ZPVE	$U0$	U	H	G	Cv
Weight	Al10	0.99907	1.00018	1.00018	1.00018	1.00018	0.98531
	Al11	0.99908	1.00018	1.00018	1.00018	1.00018	0.98901
	Al12	0.99928	1.00019	1.00019	1.00019	1.00019	0.99403
Bias	Al10	0.00201	−5.58779	−5.58524	−5.58524	−5.58895	0.57454
	Al11	0.00201	−6.13298	−6.12833	−6.12833	−6.13749	0.48288
	Al12	0.00052	−6.52096	−6.51658	−6.51659	−6.52961	0.27717

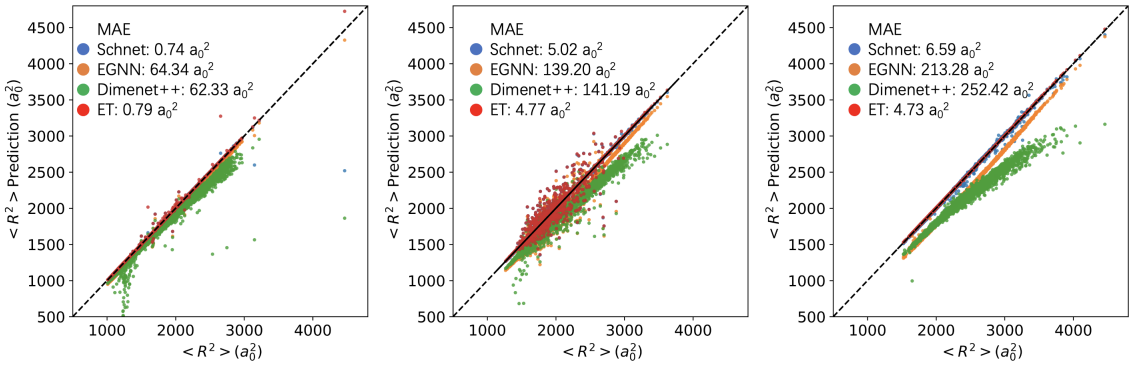

Figure A1: Error patterns of $\langle R^2 \rangle$ are different for different models, where linear transformation can improve the performance of EGNN, but might not improve the performance of DimeNet++.

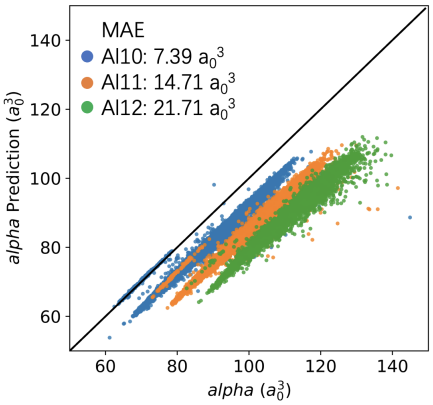

Figure A2: Error pattern of α for EGNN, where changing the aggregation method of EGNN cannot eliminate such errors.

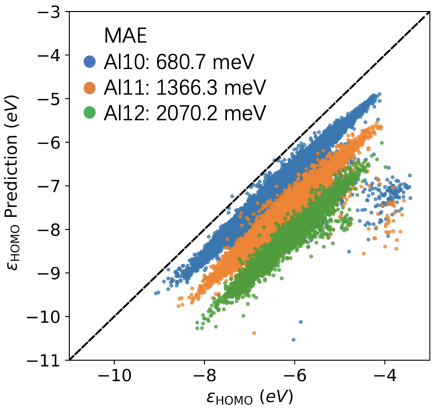

Figure A3: Error pattern of ϵ_{HOMO} for DimeNet++, where simply using mean/max pooling can improve the performance significantly.

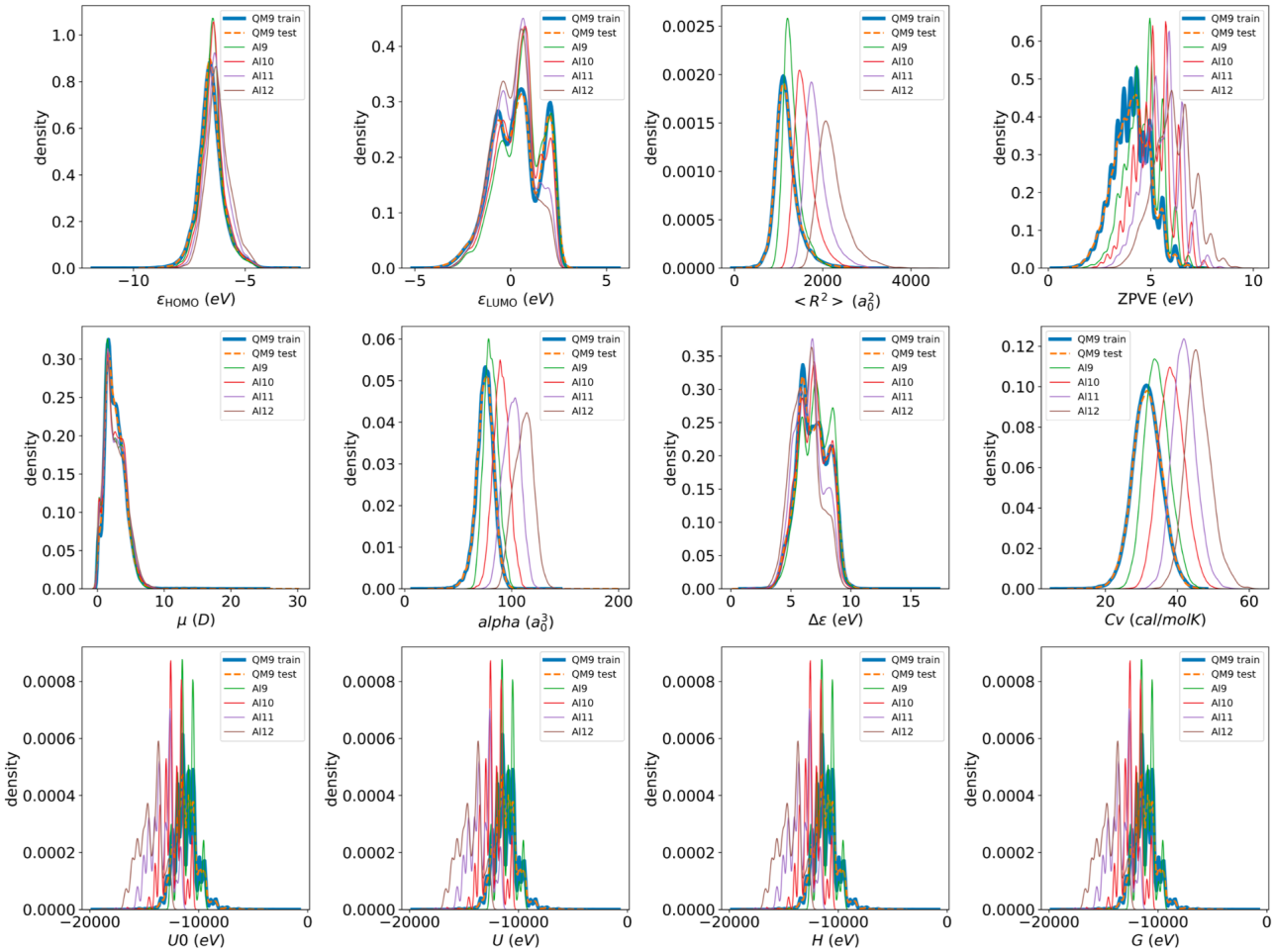

Figure A4: Property distribution comparison between QM9 and Alchemy.

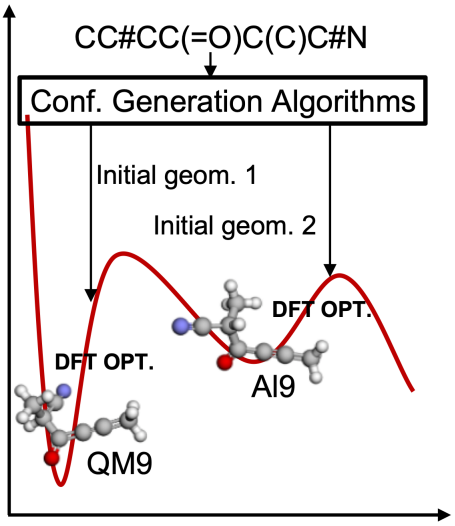

Figure A5: Even for a small molecule consisting of 9 heavy atoms with the same SMILES expressions in QM9 and Al9, the optimized geometry falls in different local optima due to bias of the structure generation algorithm and the DFT optimization scheme.

Table A3: Inference MAE by ET for 1706 molecules with same SMILES strings both in QM9 test split and Al9, where bias for the flowchart of preparing the optimized molecular geometries results in the accuracy decline. In row of QM9 test, prediction is obtained with optimized geometry in QM9 as input; in row of Al9, prediction is obtained with optimized geometry in Al9 as input where the 12 properties in Al9 are recomputed by Gaussian to compensate the DFT systematic difference.

Target (Unit)	μ (D)	α (a_0^3)	ϵ_{HOMO} (meV)	ϵ_{LUMO} (meV)	$\Delta\epsilon$ (meV)	$\langle R^2 \rangle$ (a_0^2)	ZPVE (meV)	$U0$ (meV)	U (meV)	H (meV)	G (meV)	Cv $(\frac{cal}{molK})$
QM9 test	0.070	0.062	27.1	20.2	42.3	0.0826	1.65	9.76	9.68	9.68	10.7	0.032
Al9	0.020	0.081	36.3	34.3	52.6	0.0935	2.55	28.1	27.6	27.3	31.1	0.069