Adaptive Pseudo-labeling for Quantum Calculations

Kexin Huang[*]
Department of Computer Science
Stanford University
kexinh@cs.stanford.edu

Vishnu Sresht
Pfizer
vishnu.sresht@pfizer.com

Brajesh Rai
Pfizer
brajesh.rai@pfizer.com

Mykola Bordyuh
Pfizer
mykola.bordyuh@pfizer.com

Abstract

Machine learning models have recently shown promise in predicting molecular quantum chemical properties. However, the path to real-life adoption requires (1) learning under low-resource constraint and (2) out-of-distribution generalization to unseen, structurally diverse molecules. We observe that these two challenges originate from label scarcity issue. We hypothesize that pseudo-labeling on vast array of unlabeled molecules can serve as proxies as gold-label to greatly expand the training labeled data. The challenge in pseudo-labeling is to prevent the bad pseudo-labels from biasing the model. We develop a simple and effective strategy PSEUDσ that can assign pseudo-labels, detect bad pseud-labels through evidential uncertainty, and then prevent them from biasing the model using adaptive weighting. Empirically, PSEUDσ improves quantum calculations accuracy across full data, low data and out-of-distribution settings.

1 Introduction

Accurate quantum mechanical (QM) calculations of drug-like molecules at CCSD(T) or MP2 level of theory, which are essential to characterize biomolecular interactions continue to be prohibitively expensive, despite recent advances in hardware capabilities. Machine learning (ML) models have astonishing performance in approximating these calculations at a fraction of the computational cost [18]. In the absence of large-scale benchmark data sets reporting CCSD(T) or MP2 level calculations, most publications on this topic have relied on QM9, a standard benchmark of DFT-level energy and properties, for training and evaluating QM/ML models. However, it is unclear how the reported architectures (e.g. SchNet) would perform in the regime of low but accurate data (CCSD(T) or MP2).

Two challenges remain in the way of realistic adoption of ML-aided QM calculations. Firstly, training molecules only cover part of the distribution and real-world adoption requires out-of-distribution generalization. Secondly computing QM properties for datasets of the size of QM9 is very expensive. Widespread applicability of ML for QM (e.g. such as to CCSD(T)/MP2) thus requires models to generalize well given low-data constraints.

Both challenges listed above can be attributed to the difficulty and cost to generate experimental data in chemistry and materials science. If we have a large and diverse set of labeled molecules, the ML model can then generalize to larger chemical spaces and achieve better predictive performance.

Our key observation is that state-of-the-art ML models can obtain a low-precision estimation of the properties even given a small set of labeled molecules. This suggests that the ML predicted label for any unlabelled molecule can be used as a low-precision estimation of the true label.

[*] Work done at Pfizer

NeurIPS 2021 AI for Science Workshop.

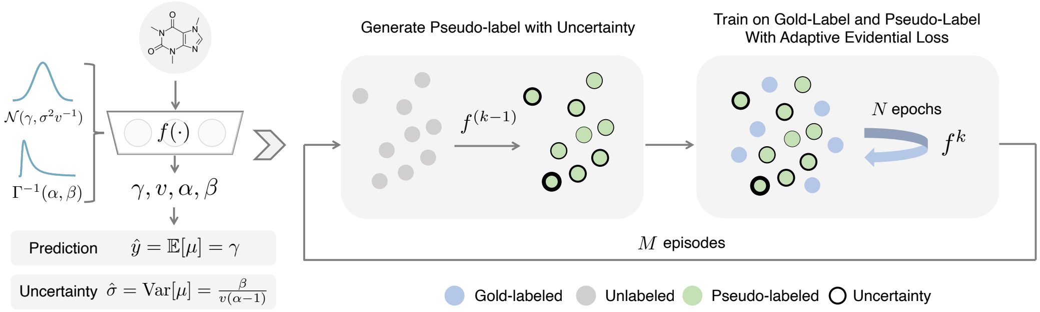
$\mathcal{N}(\gamma, \sigma^2 v^{-1})$
$\Gamma^{-1}(\alpha, \beta)$
$f(\cdot)$
γ, v, α, β
Prediction $\hat{y} = \mathbb{E}[\mu] = \gamma$
Uncertainty $\hat{\sigma} = \mathrm{Var}[\mu] = \frac{\beta}{v(\alpha-1)}$
Generate Pseudo-label with Uncertainty
$f^{(k-1)}$
Train on Gold-Label and Pseudo-Label With Adaptive Evidential Loss
N epochs
f^k
M episodes
Gold-labeled
Unlabeled
Pseudo-labeled
Uncertainty

Figure 1: PSEUDσ illustration. In every episode k, PSEUDσ assigns pseudo-labels along with their evidential uncertainty using trained neural network $f^{(k-1)}$ from previous episode. The uncertainty is used as weight to adaptively adjust the loss in this episode's neural network $f^{(k)}$'s training to reduce the effect of bad pseudo-labels in an inner-loop training with N epochs.

In our study, we develop a simple, effective, and model-agnostic pseudo-labeling strategy called PSEUDσ. Particularly, we solicit a large array of unlabeled molecules from PC9 dataset [4] where a ML model assigns pseudo-labels for them. This generates a large set of "labeled" training set, even when only a relative small number of reference QM properties are available for training. As the unlabeled dataset is diverse and unseen in QM9, it also helps generalization to unseen data. One crucial issue in pseudo-labeling is the introduced bias from low-quality pseudo-labels. To resolve this, we rely on a key observation that a data point with less evidence/higher model uncertainty is more likely to be of low-quality pseudo-label (Section4). Thus, we use model-generated evidential uncertainty to quantify each unlabeled data and use it to adaptively lower the weight of bad pseudo-labels in the training loss to reduce the bias effect. Empirically, PSEUDσ can improve QM accuracy for any atomistic model across full-data, low-data, and out-of-distribution settings.

2 Related Works

ML-aided quantum calculations. Prior work on improving accuracy on QM9 mainly focus on improving the physics-based representation in the full QM9 dataset setting [14, 17, 2, 11, 7, 10, 12]. In contrast, our work approaches this problem by improving the training strategy. PSEUDσ is model-agnostic and it can improve on any atomistic model. Additionally, we focus on realistic application scenarios such as learning in low-data regimes and out-of-distribution inference.

Pseudo-labeling. Previous works have generated pseudo-labels for unlabeled data through trained ML model prediction [9] and label propagation[15, 5]. PSEUDσ is different as it focuses on how to detect and prevent bad pseudo-labels from affecting the model. More related is a concurrent work [13] that develops an uncertainty-aware pseudo-labeling strategy, but they remove pseudo-labels at some uncertainties. In contrast, PSEUDσ uses an effective adaptive weighting scheme. Additionally, PSEUDσ is the first method that studies pseudo-labeling in quantum calculations.

Uncertainty. Model uncertainty is a well-studied subject [6, 8, 3]. [1] use evidential uncertainty to add a prior over the gaussian parameters to search for higher-order patterns for regression task. PSEUDσ leverages evidential uncertainty as the uncertainty measure. Recently, [16] adapt evidential uncertainty and have shown it can successfully help guide property prediction. In contrast, we leverage evidential uncertainty as a proxy for pseudo-label quality.

3 PSEUDσ: Adaptive Pseudo-labeling with Uncertainty

Problem Formulation. In this work, we focus on the regression task by establishing a map from the space of molecules $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ to a space of scalar properties $\mathcal{Y} = \{y_1, \ldots y_N\}$. Our dataset $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}\}$ of size $|\mathcal{X}|$ consists of pairs of molecules with atomic positional coordinates and calculated quantum mechanical properties. Every molecule $\mathbf{x}_i$ is uniquely defined by 3d coordinates $\{(a^i_j, b^i_j, c^i_j)\}_{j=1}^{N_i}$ for N_i atoms in the molecule. Let $\mathcal{X}_\mathcal{U} = \{\mathbf{x}_i\}_{i=1}^{N_U}$ be the unlabeled dataset. Denote an atomistic model $f_\theta(\cdot)$ parametrized by a set of learnable parameters θ where it can map input molecule to a predicted property label $f(\mathbf{x}_i) = \hat{y}_i$. Given $\mathcal{D} \cup \mathcal{U}$, we aim to model parameters θ that can maximize the likelihood $p_\theta(\mathcal{Y}|\mathcal{X})$ of the labeled training data. For low-data learning, the number of labeled data $|\mathcal{D}|$ is kept minimal.

Table 1: Pseudσ improves on full data setting. Reported metric is MAE. The lower the better.

Property	Unit	SchNet	PhysNet	Cormorant	MGCN	DimeNet++	SphereNet	Pseudσ-S	Pseudσ-D
ϵ_{HOMO}	meV	41	32.9	36	42.1	24.6	23.6	32.9	**20.4**
ϵ_{LUMO}	meV	34	24.7	36	57.4	19.5	18.9	24.7	**18.2**

Table 2: Pseudσ improves on low-data regime. Reported metric is MAE. The lower the better.

Low-Data Setting		1% QM9 (1,100)		10% QM9 (11,000)	
Property	Unit	SchNet → Pseudσ	DimeNet++ → Pseudσ	SchNet → Pseudσ	DimeNet++ → Pseudσ
ϵ_{HOMO}	meV	$265.4 \xrightarrow{+10.8} 276.2$	$248.9 \xrightarrow{-18.7} 230.2$	$119.0 \xrightarrow{-30.2} 88.8$	$81.1 \xrightarrow{-13.7} 67.4$
ϵ_{LUMO}	meV	$290.6 \xrightarrow{-57.8} 232.8$	$229.3 \xrightarrow{-5.2} 224.1$	$93.3 \xrightarrow{-15.0} 78.3$	$60.8 \xrightarrow{-1.6} 59.2$

Episodic Pseudo-labeling. Pseudo-labeling mainly consists of three stages. In first stage, regular training is conducted on labeled data D. In the second stage, the updated model conduct inference on the unlabeled data and combine the pseudo-labeled dataset with the labeled data. In the third stage, the model is further trained using the combined dataset. The second and third stage forms an episode and is reiterated till the loss stops decreasing or reaches a predefined maximum episode number.

Note that in contrast to noisy-student training, we do not retrain a separate student model but to continue from the checkpoint. In comparison to standard pseudo-labeling where pseudo-labels are regenerated every epoch, we devise a episodic training strategy, where each episode consists of multiple epochs and pseudo-labels are regenerated once an episode ends. This is important since it gives the model more time to absorb useful information from a given set of pseudo-labels. For each episode, we also reinitialize the learning rate with a small step-wise decay strategy to allow the model a chance to jump out of local optimum from the previous set of pseudo-labels.

Evidential uncertainty quantification. Pseudo-labels are noisy. Many are incorrect and can potentially bias the model. Thus, it is the key to first detect the low-quality pseudo-labels in each episode. Our key observation is that low-quality pseudo-labels have high model uncertainty. Thus, we can use uncertainty as a proxy to detect these bad labels. We can model the label probabilistically as it is drawn from a Gaussian $(y_1, \cdots, y_i) \sim \mathcal{N}(\mu, \sigma^2)$, where the parameters are unknown. To estimate them, we pose a normal prior $\mathcal{N}(\gamma, \sigma^2 v^{-1})$ for μ, and inverse-gamma prior $\Gamma^{-1}(\alpha, \beta)$ for σ, where the parameters θ are an instantiation of the posterior $p(\mu, \sigma^2|\gamma, v, \alpha, \beta)$. Assuming the posterior can be factorized independently, the posterior becomes a $\text{NormalInvGamma}(\gamma, v, \alpha, \beta)$ where the maximum likelihood estimation of θ can be analytically found as $\mathbb{E}[\mu] = \gamma$ and $\mathbb{E}[\sigma^2] = \frac{\beta}{\alpha-1}$. The uncertainty of the model prediction, i.e. epistemic uncertainty, becomes $\text{Var}[\mu] = \mathbb{E}[\sigma^2]/v = \frac{\beta}{v(\alpha-1)}$ and the uncertainty of the data, i.e. aleatoric uncertainty is $\mathbb{E}[\sigma^2]$. As the MLE is deterministic, the model can output the four prior parameters $\{\gamma, v, \alpha, \beta\}$ directly where the prediction and uncertainty can be derived from them analytically [1]. The prior is optimized by evidential loss $\mathcal{L}^{\text{evi}}$, where one term is to fit the training label and a second term is a regularizer that encourages higher uncertainty when the prediction gap is high [1].

Adaptive weighting. The evidential uncertainty detects the low-quality pseudo-labels. The next step is to remove the noisy effect of them from the model training. Naive method include removal based on a threshold [13]. However, it has the following two disadvantages: (1) it introduces a new hyperparameter; (2) it removes a large set of unlabeled data which reduces the diversity of the training space, harming the OOD generalization. Instead, we propose an adaptive weighting mechanism that adapts the evidential loss given the inverse epistemic uncertainty. Intuitively, a higher uncertainty requires lower effect in the loss function. Thus, the final loss becomes $\mathcal{L} = \frac{1}{|\mathcal{D}|}\sum_{i\in\mathcal{D}}\mathcal{L}_i^{\text{evi}} + \sum_{i\in\mathcal{U}}\frac{\text{Var}[\mu]_i^{-1}}{\sum_{i\in\mathcal{U}}\text{Var}[\mu]_i^{-1}}\mathcal{L}_i^{\text{evi}}$ where the first part is without weighting because they are all gold-labeled. This adaptive loss solves the two disadvantages since it has zero hyperparameter and it removes the effect of bad pseudolabels while retaining all training examples to maximize the diversity of the training space.

4 Experiments

Dataset and Setups. We evaluate Pseudσ using QM9 dataset under two settings. *(A) Full-data*: We follow the previous works [10, 7] where a 110,000/10,000/10,831 training/validation/testing set is obtained. For the unlabeled data, we solicit to PC9, a dataset of 99,234 molecules of similar

Table 3: Out-of-distribution best validation MAE.

Property	Unit	SchNet	DimeNet++	Pseudσ-D
σ_{HOMO}	meV	243.4	230.4	**214.4**
σ_{LUMO}	meV	225.0	184.2	**175.8**

Table 4: Ablation using SchNet as backbone.

Property	Unit	Pseudσ	-pseudo-label	-uncertainty	-student	-uniform
ϵ_{HOMO}	meV	**32.9**	38.9	47.7	41.4	37.2
ϵ_{LUMO}	meV	**24.7**	27.2	32.1	31.4	28.8

characteristics as QM9, curated by [4]. *(B) Low-data*: we set $k\%$ of QM9 full training set as the training set (i.e. $k \times 110{,}000$) and we remove the label of (1-$k\%$) QM9 full training set and make it as the unlabeled set. We evaluate in two k values, 1 and 10, which means only 1,100/11,000 data points is trained respectively.

Pseudσ is model-agnostic. We evaluate it with two model backbones SchNet [14] (Pseudσ-S) and DimeNet++ [7] (Pseudσ-D). We conduct two hyperparameter tunings on σ_{HOMO} with SchNet backbone on the validation MAE with full data/low-data setting respectively. The optimal hyperparameter is then used for all targets. Our preliminary result is conducted on four targets $\mu, \alpha, \sigma_{\text{HOMO}}, \sigma_{\text{LUMO}}$.

Pseudσ improves on fully supervised QM calculations. We compare Pseudσ against 7 state-of-the-art models [14, 17, 2, 11, 7, 10] in Table 1. Pseudσ-D surpasses all baselines in $\mu, \sigma_{\text{HOMO}}, \sigma_{\text{LUMO}}$. Particularly, comparing Pseudσ-S with SchNet and Pseudσ-D with DimeNet++, we find Pseudσ can consistently improve even on fully supervised setting, highlighting the utility of Pseudσ and the high quality of PC9 as unlabeled data.

Pseudσ significantly improves on low-data QM calculations. In Table 2, we investigate how Pseudσ can improve on low-data regime with only 1%, 10% training data point. We observe Pseudσ can consistently and significantly improve the prediction in $\mu, \sigma_{\text{HOMO}}, \sigma_{\text{LUMO}}$ across both low-data setting, and both model backbones, suggesting Pseudσ can help prediction in realistic low-data quantum calculations. Notably, the performance on α is not ideal across the board, requiring further investigation of Pseudσ's behavior across targets as our hyperparameter is tuned on σ_{HOMO}.

Pseudσ improves out-of-distribution QM calculations. Another realistic challenge is to infer accurately on unseen data distribution away from QM9. We conduct inference on the PC9 dataset where it has calculated $\sigma_{\text{HOMO}}, \sigma_{\text{LUMO}}$. We find Pseud$\sigma$ can again significantly improve OOD accuracy in DimeNet++ backbone, highlighting the robustness of Pseudσ.

Evidential uncertainty highly correlates to label quality. In this experiment, we train on the full QM9 training set with evidential uncertainty and then infer on the QM9 testing set. We find that non-parametric Spearman correlation between MAE and epistemic uncertainty is 0.42 with p-value < 1e-16. Additionally, we evaluate on PC9 out-of-distribution set and the Spearman correlation is 0.35 with p-value < 1e-16, suggesting our uncertainty is a robust measure of label quality.

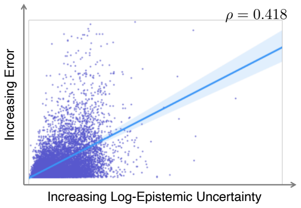

Figure 2: Uncertainty highly correlates to label quality.

Ablations. Table 4 shows that each component is indispensable for Pseudσ. The comparison with -pseudo-label shows the usefulness of the general pseudo-labeling strategy. The -comparison test shows that vanilla pseudo-labeling does not work well for QM calculations, calling for specialized strategy design. The -student test shows that self-distillation to retrain a model in every episode as in [19] is the key. Experiments with adding noise to the 3D positions degrade performance. The -uniform test, which uses the same weight for all pseudo-labels, shows the importance of detection and adaptive removal of bad pseudo-labels.

5 Conclusion

Pseudσ is a simple, effective, model-agnostic pseudo-labeling strategy that can improve quantum calculations accuracy in abundant data, low data and out-of-distribution settings.

References

[1] Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. *NeurIPS*, 2020.

[2] Brandon Anderson, Truong-Son Hy, and Risi Kondor. Cormorant: Covariant molecular neural networks. *NeurIPS*, 2019.

[3] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *ICML*, pages 1613–1622. PMLR, 2015.

[4] 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.

[5] Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. Label propagation for deep semi-supervised learning. In *CVPR*, pages 5070–5079, 2019.

[6] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *NeurIPS*, 2017.

[7] Johannes Klicpera, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs. *ICLR*, 2020.

[8] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *NeurIPS*, 2017.

[9] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896, 2013.

[10] Yi Liu, Limei Wang, Meng Liu, Xuan Zhang, Bora Oztekin, and Shuiwang Ji. Spherical message passing for 3d graph networks. *arXiv:2102.05013*, 2021.

[11] Chengqiang Lu, Qi Liu, Chao Wang, Zhenya Huang, Peize Lin, and Lixin He. Molecular property prediction: A multilevel quantum interactions modeling perspective. In *AAAI*, volume 33, pages 1052–1060, 2019.

[12] Zhuoran Qiao, Anders S Christensen, Frederick R Manby, Matthew Welborn, Anima Anandkumar, and Thomas F Miller III. Unite: Unitary n-body tensor equivariant network with applications to quantum chemistry. *arXiv:2105.14655*, 2021.

[13] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. *ICLR*, 2021.

[14] 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. *NeurIPS*, 2017.

[15] Weiwei Shi, Yihong Gong, Chris Ding, Zhiheng MaXiaoyu Tao, and Nanning Zheng. Transductive semi-supervised deep learning using min-max features. In *ECCV*, pages 299–315, 2018.

[16] Ava P Soleimany, Alexander Amini, Samuel Goldman, Daniela Rus, Sangeeta N Bhatia, and Connor W Coley. Evidential deep learning for guided molecular property prediction and discovery. *ACS Central Science*, 2021.

[17] Oliver T Unke and Markus Meuwly. Physnet: a neural network for predicting energies, forces, dipole moments, and partial charges. *Journal of Chemical Theory and Computation*, 15(6):3678–3693, 2019.

[18] O. Anatole von Lilienfeld and Kieron Burke. Retrospective on a decade of machine learning for chemical discovery. *Nature Communications*, 11(1):4895, dec 2020.

[19] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *CVPR*, pages 10687–10698, 2020.