ADAPTIVE PSEUDO-LABELING FOR QUANTUM CALCULATIONS

Anonymous authors
Paper under double-blind review

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 constraints and (2) out-of-distribution generalization to unseen, structurally diverse molecules. We observe that these two challenges can be alleviated via abundant labels, often not the case in quantum chemistry. We hypothesize that pseudo-labeling on a vast array of unlabeled molecules can serve as gold-label proxies to expand the training labeled dataset significantly. The challenge in pseudo-labeling is to prevent the bad pseudo-labels from biasing the model. Motivated by entropy minimization framework, we develop a simple and effective strategy PSEUDσ that can assign pseudo-labels, detect bad pseud-labels through evidential uncertainty, and 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 is 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 (von Lilienfeld & Burke, 2020). Such speedups could accelerate the discovery of new therapeutics.

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. This data set contains molecular geometries and energies that were generated using Density Functional Theory (DFT), which provides a faster (compared to CCSD(T), for example), but less accurate description of molecular energetics and properties. Previous publications have utilized computed geometries and properties from in-distribution molecules in QM9 to test and validation model performance, showing excellent performance of QM/ML models as demonstrated by low mean absolute error (MAE) of predicted energies of other properties relative to the gold-standard DFT calculations. 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 the 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 costly. 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 of generating experimental data in chemistry and materials science. If we have a large and diverse set of labeled molecules, the ML model can generalize to larger chemical spaces and achieve better predictive performance.

Present work. 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. We hypothesize that by considering these low-precision pseudo-labels as the training targets in the training process, we can largely increase the training dataset size and diversity and

alleviate the fundamental label scarcity issue, along with the associated challenges in low-data and out-of-distribution generalization.

In our study, we develop a simple, effective, and model-agnostic pseudo-labeling strategy called PSEUDσ. Notably, we solicit an extensive array of unlabeled molecules from the PC9 dataset (Glavatskikh et al., 2019) where an ML model assigns pseudo-labels for them. This generates a large set of "labeled" training data points, even when only a relatively 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 (Section 6). 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. Theoretically, we analyze the connection between evidential uncertainty and entropy minimization theory and show that our uncertainty measures are excellent proxies to measure pseudo-label quality. Empirically, we show 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. Recently, many ML models have been proposed to improve quantum calculations. They mainly focus on improving the physics-based representation in the full QM9 dataset setting (Schütt et al., 2017; Unke & Meuwly, 2019; Anderson et al., 2019; Lu et al., 2019; Klicpera et al., 2020; Liu et al., 2021; Qiao et al., 2021). In contrast, our work proposes to shift the focus on improving training strategy instead of the model architecture. In addition, PSEUDσ is model-agnostic, and it can improve on any atomistic model. Additionally, we focus on realistic quantum calculations such as learning in the low-data regime and out-of-distribution inference.

Pseudo-labeling. Pseudo-labeling/self-training generates pseudo-labels for unlabeled data. Numerous works exist for how to assign pseudo-labels, notably through trained ML model prediction (Lee et al., 2013), and label propagation(Shi et al., 2018; Iscen et al., 2019). PSEUDσ is different as it focuses on detecting and preventing bad pseudo-label from affecting the model. Also, PSEUDσ adopts a novel episodic pseudo-label strategy with a re-initialized learning rate. (Xie et al., 2020) re-initialize the network as a student when a new pseudo-label set is generated along with noise per epoch (Xie et al., 2020). In contrast, PSEUDσ has no student, and no noise as both are shown to be ineffective for QM in Section 6. In addition, small perturbational noise in 3D molecular geometry could easily lead to a drastic energy difference. Thus, the noise strategy does not work for QM tasks. More related is a concurrent work (Rizve et al., 2021) that develops an uncertainty-aware pseudo-labeling strategy, but they introduce additional hyperparameters to remove pseudo-labels at some uncertainties. In contrast, PSEUDσ uses an effective adaptive weighting scheme, along with an episodic pseudo-labeling training schedule. Additionally, PSEUDσ is the first method that studies pseudo-label in quantum calculations, which present unique challenges and motivations.

Uncertainty. Model uncertainty is a well-studied subject (Kendall & Gal, 2017; Lakshminarayanan et al., 2017; Blundell et al., 2015). (Amini et al., 2020) use evidential uncertainty to add a prior over the gaussian parameters to search for higher-order patterns for regression tasks. PSEUDσ leverages evidential uncertainty as the uncertainty measure. Note that PSEUDσ is uncertainty measure-agnostic. We can easily switch to alternative uncertainty measures. Recently, (Soleimany et al., 2021) adapt evidential uncertainty and has shown it can successfully help guide property prediction. In contrast, we leverage evidential uncertainty as a proxy for pseudo-label quality to tackle low-data and out-of-distribution challenges in realistic quantum calculations setup.

3 PROBLEM FORMULATION

Let $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ denote N molecules, where each molecule $\mathbf{x}_i$ is uniquely defined by 3D coordinates $\{(a_j^i, b_j^i, c_j^i)\}_{j=1}^{N_i}$ for N_i atoms in the corresponding molecule. We then denote $\mathcal{Y} = \{y_1, \ldots y_N\}$ a set of quantum mechanical properties for each molecule. The labeled dataset thus consists of a set of pairs of 3D coordinates and scalar labels $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}\}$.

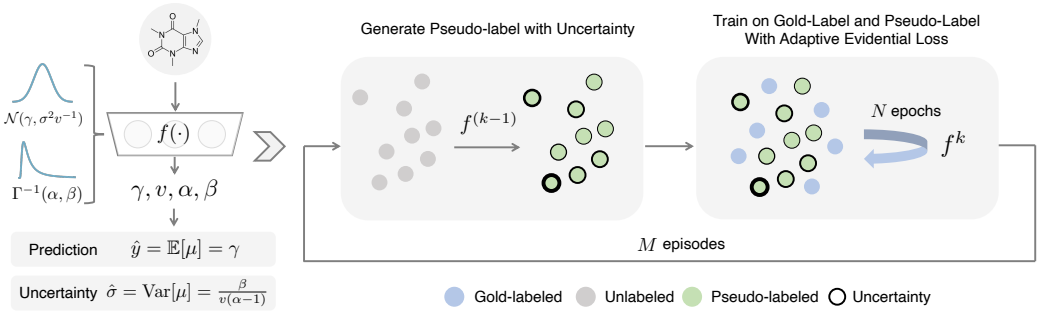

Figure 1: PSEUDσ illustration. In every episode k, PSEUDσ assigns pseudo-labels along with their evidential uncertainty using trained neural network $f^{(k-1)}$ from previous episode. The uncertainty is used as weight to adaptively adjust the loss in this episode's neural network $f^{(k)}$'s training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

In addition to the labeled data, we solicit a large quantity of unlabeled data to generate pseudo-labels. We denote an unlabeled dataset $\mathcal{X}_{\mathcal{U}} = \{\mathbf{x}_1, \ldots, \mathbf{x}_M\}$, where M is the size of the unlabeled dataset. Given an atomistic model $f(\cdot)$, we can generate pseudo-labels $\hat{\mathcal{Y}} = \{\hat{y}_1, \ldots, \hat{y}_M\}$, where $\hat{y}_i = f(\mathbf{x}_i)$ for $\mathbf{x}_i \in \mathcal{X}_{\mathcal{U}}$.

The problem is to find a machine learning-based atomistic model $f : \mathbf{x} \mapsto y$ that can establish an accurate map from 3D coordinates to the quantum mechanical properties of the molecules, with the help of pseudo-labeled dataset $\mathcal{X}_{\mathcal{U}}$.

4 PSEUDσ: ADAPTIVE PSEUDO-LABELING FOR QUANTUM CALCULATIONS

PSEUDσ (Figure 1) is an approach for quantum chemical property prediction. Building on theoretical motivation from Section 5, PSEUDσ solicits pseudo-labels on a vast array of an unlabeled dataset to increase the diversity of the training space via an episodic labeling strategy. Then, it intelligently weights the pseudo-labels using evidential uncertainty to allow a positive transfer. The overview is in Algorithm 1.

Episodic Pseudo-labeling. We devise a pseudo-labeling strategy that can ensure learning from the pseudo-labels to the fullest extent. In the standard pseudo-labeling (Lee et al., 2013; Rizve et al., 2021), a set of labels are regenerated in every epoch. In contrast, we devise an episodic training strategy, where each episode consists of multiple epochs, and pseudo-labels are regenerated in every episode. This is important because we observe that per-epoch pseudo-labeling is far from converging, preventing the model from learning from all useful information in pseudo-labels in one episode. In contrast, our episodic approach gives the model more time to absorb useful information from a given set of pseudo-labels.

Formally, PSEUDσ mainly consists of three stages: in the first stage, regular training is conducted on labeled data D, and the output model is the initialized model $f^{(1)}$. In the second stage, the updated model at episode k then conduct inference on the entire unlabeled data $\hat{\mathbf{y}}_i = f^{(k)}(\mathcal{U}_i)$ to generate the pseudo-label set. The per-episode pseudo-label set is then combined with the gold-labeled data to form the training data for the next episode. In the third stage, the model is further trained using the combined dataset to get a new model $f(k+1)$ after N epochs. The second and third stages are then reiterated till the loss stops decreasing.

Note that in contrast to noisy-student training, we do not retrain a separate student model but continue from the checkpoint. 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 the local optimum from the previous set of pseudo-labels.

Evidential uncertainty quantification. Pseudo-labels are noisy. Many are incorrect and can potentially lead to negative transfer. Thus, it is instrumental in detecting the quality of pseudo-labels. However, there is no auxiliary information in the dataset about the pseudo-labels. Thus, we need to quantify it through some proxies that can be assigned without auxiliary information. Our key observation is that low-quality pseudo-labels have high model uncertainty, and high-quality pseudo-labels have low model uncertainty. Another advantage of model uncertainty is that it can be estimated solely from $\mathbf{x}$, if we make the model uncertainty-aware.

Building on the theoretical motivation about the connection between evidential uncertainty and entropy minimization in Section 5, we use evidential uncertainty as to the proxy for label quality. Formally, we can model the label probabilistically as it is drawn from $(y_1, \cdots, y_i) \sim \mathcal{N}(\mu, \sigma^2)$, where the mean μ are variance σ^2 are unknown. To estimate them, we pose a prior

$$\mu \sim \mathcal{N}(\gamma, \sigma^2 v^{-1}), \sigma \sim \Gamma^{-1}(\alpha, \beta), \tag{1}$$

where the parameters $\theta = (\mu, \sigma)$ is an instantiation of the posterior $p(\mu, \sigma^2 | \gamma, v, \alpha, \beta)$. Dependence of the variance for the distribution of labels and variance for the parameter μ allows us to factorize posterior for μ and σ independently. The posterior becomes a $\mathrm{NormalInvGamma}(\gamma, v, \alpha, \beta)$ where the maximum likelihood estimation of θ can be analytically found as

$$\mathbb{E}[\mu] = \gamma, \mathbb{E}[\sigma^2] = \frac{\beta}{\alpha - 1}. \tag{2}$$

The uncertainty of the model prediction can also be calculated, i.e. epistemic uncertainty:

$$\mathrm{Var}[\mu] = \mathbb{E}[\sigma^2]/v = \frac{\beta}{v(\alpha - 1)}. \tag{3}$$

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. The prior is optimized by evidential loss $\mathcal{L}^{\text{evi}}$ (Amini et al., 2020):

$$\mathcal{L}_i^{\text{evi}} = -\log \mathrm{St}\left(y_i; \gamma, \frac{\beta(1+v)}{v\alpha}, 2\alpha\right) + \lambda |y_i - \gamma|(2v + \alpha) \tag{4}$$

where the first term is to maximize the log-likelihood of the posterior predictive, which is derived as the Student's t-distribution, and a second term is a regularizer that encourages lower total evidence (i.e., higher epistemic uncertainty) when the gap between ground truth and prediction is high. The regularization is controlled by a hyperparameter λ.

Adaptive weighting. The evidential uncertainty detects the low-quality pseudo-labels. The next step is to remove the noisy effect from the model training. Naive methods often use removal based on a threshold (Rizve et al., 2021). However, it has two disadvantages: (1) it introduces a new hyperparameter - the threshold; (2) it removes a large set of unlabeled data, which reduces the diversity of the training space, harming the generalization. Instead, we propose an adaptive weighting mechanism that adapts the evidential loss given the inverse epistemic uncertainty. Intuitively, a higher uncertainty data point requires a lower effect in the loss function because it is more likely that the sample has a low pseudo-label quality, and we want to reduce its effect on the model. Conversely, if a pseudo-label has low uncertainty, the label quality is high enough to be used as a proxy of gold-label. Thus, we should assign a higher score in the loss for it. Thus, the final loss becomes

$$\mathcal{L} = \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} \mathcal{L}_i^{\text{evi}} + \sum_{i \in \mathcal{U}} \frac{\mathrm{Var}[\mu]_i^{-1}}{\sum_{i \in \mathcal{U}} \mathrm{Var}[\mu]_i^{-1}} \mathcal{L}_i^{\text{evi}}, \tag{5}$$

where the first part about labeled dataset $\mathcal{D}$ is without weighting because they are all gold-labeled. This adaptive loss solves two disadvantages since it has zero hyperparameters, and it removes the effect of bad pseudo-labels while retaining all training examples to maximize the diversity of the training space.

5 PSEUDσ MOTIVATION: CONNECTION TO ENTROPY MINIMIZATION

We derive motivations about why evidential uncertainty and the weighting mechanism could be beneficial to pseudo-labeling based on the entropy minimization framework for semi-supervised

Algorithm 1: PSEUDσ Algorithm.

Input: Labeled data $\mathcal{D} = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_N, y_N)\}$, unlabled data $\mathcal{U} = \{\mathbf{x}_1, \cdots, \mathbf{x}_{N_U}\}$

$\hat{\mathcal{U}} \leftarrow \{\}$ // Initialize with empty unlabeled data

for $k \in \{1, \cdots, M\}$ // Outer-loop with k episodes

do

  $\mathcal{T} \leftarrow \mathcal{D} \cup \hat{\mathcal{U}}$ // Join updated pseudo-labels

  for $(\mathcal{X}_i, \mathbf{y}_i) \in \mathcal{T}$ // Inner-loop with N epochs

  do

    $\theta = (\gamma, v, \alpha, \beta) = f_i^{(k-1)}(\mathcal{X}_i)$ // Evidental parameters

    $\hat{\mathbf{y}}_i = \mathbb{E}[\mu] = \gamma$ // Posterior prediction

    $\mathrm{Var}[\mu] = \mathbb{E}[\sigma^2]/v = \frac{\beta}{v(\alpha-1)}$ // Epistemic uncertainty

    $\mathcal{L} = \mathrm{L}(\hat{\mathbf{y}}_i, \mathbf{y}_i, \mathrm{Var}[\mu], \theta)$ // Adaptive evidential loss via Eq. 5

    $f_{i+1}^{(k-1)} = \mathrm{Update}(f_i^{(k-1)}, \mathcal{L})$ // Inner-loop update

  end

  for $\mathcal{X}_i \in \mathcal{U}$ **do**

    $\hat{\mathbf{y}}_i = f^{(k)}(\mathcal{X}_i)$ // Infer a new set of pseudo-labels

    $\hat{\mathcal{U}}_i \leftarrow (\mathcal{X}_i, \hat{\mathbf{y}}_i)$ // Update pseudo-labels

  end

end

learning from Yves Grandvalet (2004; 2006); Lee et al. (2013). Notably, our use of Bayesian modeling enables us to analytically derive a conditional entropy for pseudo-labeled data. We find that evidential loss strongly relates to conditional entropy, and minimizing evidential loss directly minimizes entropy. Secondly, we find the conditional entropy could be decomposed into the inverse epistemic uncertainty and the log-likelihood, which motivates our weighting mechanism.

In the standard regression setting, one seeks to maximize the likelihood of the model $p_\theta(\mathcal{Y}|\mathcal{X})$ on the labeled dataset $\mathcal{D}$. To utilize an unlabeled dataset, we need to extract some useful information on how the model behaves on the unlabeled dataset and inject this information to improve the model. To measure the utility, entropy $\mathcal{H}\left(Y|X_{\mathcal{U}}\right)$ is introduced (Yves Grandvalet, 2006) as a proxy to measure the amount of information in unlabeled data:

$$\mathcal{H}(\mathcal{Y}|\mathcal{X}_{\mathcal{U}}) = \sum_{\mathbf{x}_i \in \mathcal{X}_U} \mathrm{E}_{y \sim \mathrm{p}_\theta(y|\mathbf{x}_i)}[-\log \mathrm{p}_\theta(y|\mathbf{x}_i)]. \tag{6}$$

Intuitively, high entropy is indicative of the poor model confidence on the unlabeled data point, while the low entropy can be interpreted as high confidence of the model predictions. It then casts the regression as the following optimization problem:

$$\arg\max_\theta \left[\log \mathrm{p}_\theta(\mathcal{Y}|\mathcal{X}) \ - c\,\mathcal{H}(\mathcal{Y}|\mathcal{X}_{\mathcal{U}})\right] \tag{7}$$

where c is the proportionality constant. Intuitively, here, the objective tends to maximize the log-likelihood on the labeled dataset while minimizing the entropy on the unlabeled dataset at the same time to transfer knowledge from unlabeled data.

In previous works (Schütt et al., 2017; Liu et al., 2021), molecule properties are not modeled probabilistically such that entropy calculation is infeasible. In contrast, PSEUDσ uses Bayesian modeling approaches that allow us to analytically calculate the entropy. For every molecule $\mathbf{x}_i$ the machine learning model outputs four parameters $f(\mathbf{x}_i) = (\alpha_i, \beta_i, \gamma_i, \nu_i)$. Based on these parameters, the likelihood of label y given the input molecule $\mathbf{x}_i$ is given by the Student-t distribution in the context of evidential regression

$$p_\theta(y|\mathbf{x}_i) = \mathrm{St}(y; \gamma_i, \sigma_{st,i}^2, 2\alpha_i) \tag{8}$$

evaluated at location parameter γ_i, Student t-distribution scale parameter $\sigma_{st,i}^2 = \frac{\beta_i(1+\nu_i)}{\nu_i \alpha_i}$ and $2\alpha_i$ degrees of freedom. The Shannon entropy of the Student t-distribution given in terms of evidential parameters is readily available (Appendix A):

$$\mathcal{H}(y|\mathbf{x}_i) = \frac{2\alpha_i + 1}{2}\left(\Psi(\frac{2\alpha_i + 1}{2}) - \Psi(\alpha_i)\right) + \log\sqrt{2\alpha_i}\,\mathrm{B}(\alpha_i, \frac{1}{2}) + \frac{1}{2}\log\sigma_i^2 \tag{9}$$

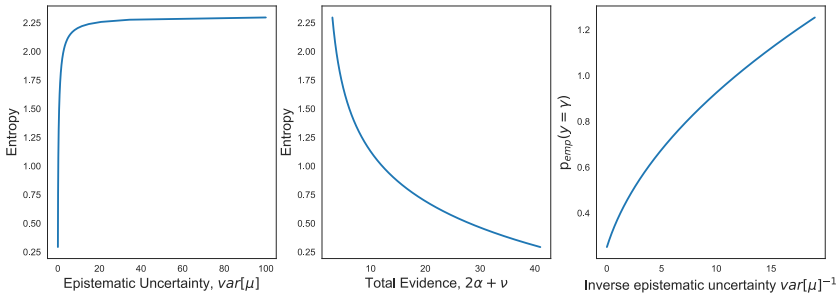

Figure 2: Dependence of evidential entropy on epistemic uncertainty (left panel) and total evidence (middle panel). The entropy increases as uncertainty increases and entropy tends to decrease as the total evidence grows. Right panel shows the dependence of empirical estimate of Student t-probability on inverse model uncertainty. Empirical probability tends to increase as model uncertainty decreases. Evidential parameters are fixed $\beta = 1$ and $\nu = 1$ as α is varied from $(1 \ldots 10]$.

Table 1: Dataset statistics.

Setting	Training Set	Validation Set	Testing Set	Unlabeled Set	OOD Set
Full-data	110,000 (QM9)	10,000 (QM9)	10,831 (QM9)	99,234 (PC9)	-
Low-data-1%	1,100 (QM9)	10,000 (QM9)	10,831 (QM9)	108,900 (QM9)	-
Low-data-10%	11,000 (QM9)	10,000 (QM9)	10,831 (QM9)	99,000 (QM9)	-
Out-of-distribution	110,000 (QM9)	10,000 (QM9)	10,831 (QM9)	99,234 (PC9)	99,234 (PC9)

where Ψ is a digamma function and B is a beta function. We plot the relation between the entropy and total evidence, which is the second term in the evidential loss (Equation 4), and we find that they are highly correlated. Intuitively, we show that minimizing evidential loss can implicitly minimize entropy on the unlabeled data such that it encourages positive information transfer.

Next, to connect it to the adaptive pseudo-labeling approach, we make further simplifications. We evaluate the entropy using iterative expectation-maximization algorithm, where the predictions y are generated from the iteration step $t-1$ and log-likelihood is evaluated at iterative step t:

$$\mathcal{H}(\mathcal{Y}|\mathcal{X}_{\mathcal{U}}) = \sum_{\mathbf{x}_i \in \mathcal{X}_U} \mathbb{E}_{y \sim p_{\theta(t-1)}(y|\mathbf{x}_i)} \left[-\log \mathrm{p}_{\theta(t)}(y|\mathbf{x}_i) \right]. \tag{10}$$

Resorting to empirical estimate of the entropy, we select y as the highest mode of probability distribution $\mathrm{p}_{\theta(t-1)}(y|\mathbf{x}_i)$ (Yves Grandvalet, 2004), which corresponds to $y = \gamma_i^{t-1}$, we obtain the following:

$$\mathcal{H}(\mathcal{Y}|\mathcal{X}_{\mathcal{U}}) \approx \mathcal{H}_{emp}(\mathcal{Y}|\mathcal{X}_{\mathcal{U}}) = \sum_{\mathbf{x}_i \in \mathcal{X}_U} \mathrm{p}_{emp}(y = \gamma_i^{t-1}, \sigma^2_{st,i}, 2\alpha_i) \log \ \mathrm{p}_{\theta(t)}(y_i^{t-1}|\mathbf{x}_i) \tag{11}$$

where the log probabilities are weighted by empirical probabilities

$$\mathrm{p}_{emp}(y = \gamma_i^{t-1}, \sigma^2_{st,i}, 2\alpha_i) = \mathrm{St}(y = \gamma_i^{t-1}, \sigma^2_{st,i}, 2\alpha_i) = \frac{1}{\sqrt{2\pi\alpha_i\sigma^2_{st,i}}\mathrm{B}(\frac{1}{2}, \alpha)}. \tag{12}$$

We can observe that the objective on the unlabeled dataset introduces a weighted likelihood of pseudo-labels y_i^{t-1}. We then plot the distribution of the weights against our empirical inverse epistemic uncertainty weight in the loss to reduce the influence of low-quality pseudo-labels (Equation 5). We find these derived weights qualitatively behave like the inverse uncertainty weights in our loss function (Figure 2). Thus, we show that the adaptive weighting mechanism could be viewed as parts of the entropy minimization framework, or in another perspective, entropy minimization implicitly uses adaptive weighting.

Table 2: 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 3: 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 $\rightarrow$ PSEUDσ	DimeNet++ $\rightarrow$ PSEUDσ	SchNet $\rightarrow$ PSEUDσ	DimeNet++ $\rightarrow$ 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	$290.6 \xrightarrow{-57.8} 232.8$	$229.3 \xrightarrow{-5.2} 224.1$	$93.3 \xrightarrow{-15.0} 78.3$	$60.8 \xrightarrow{-1.6} 59.2$

6 EXPERIMENTS

6.1 DATASET AND EXPERIMENTAL SETUPS

We evaluate PSEUDσ using QM9 dataset (Wu et al., 2018) under two settings. *(A) Full-data*: We follow the previous works (Liu et al., 2021; Klicpera et al., 2020) 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 characteristics as QM9, curated by (Glavatskikh et al., 2019). *(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. Dataset statistics summary is in Table 1.

PSEUDσ is model-agnostic. We evaluate it with two model backbones SchNet (Schütt et al., 2017) (PSEUDσ-S) and DimeNet++ (Klicpera et al., 2020) (PSEUDσ-D). We do not experiment with the SOTA atomistic model SphereNet (Liu et al., 2021) because it is highly computationally expensive. Our result is conducted on two targets $\sigma_{\text{HOMO}}, \sigma_{\text{LUMO}}$, because the PC9 dataset only has these two targets. We use mean absolute error as the evaluation metric.

For baselines, we compare with 6 state-of-the-art baselines, including SchNet (Schütt et al., 2017), PhysNet (Unke & Meuwly, 2019), Cormorant (Anderson et al., 2019), MGCN (Lu et al., 2019), DimeNet++ (Klicpera et al., 2020), and SphereNet (Liu et al., 2021). We report the best results taken from the original authors' paper while using the same fraction of data split in the full data setting. For PSEUDσ, 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 both targets. Note that the atomistic model itself has the same hyperparameter as the original authors. Code will be released upon anonymous review period.

6.2 RESULTS

Overview of results. We report performances of PSEUDσ in full data (Table 2), low-data (Table 3), out-of-distribution (Table 4) settings and find PSEUDσ achieves the best performance across all settings, suggesting the robustness of the pseudo-labeling strategy. Systematic ablation study (Table 5) also show the importance of each module in PSEUDσ.

PSEUDσ improves on fully supervised QM calculations. We report PSEUDσ against 6 state-of-the-art models in Table 2. PSEUDσ-D surpasses all baselines in both targets $\sigma_{\text{HOMO}}, \sigma_{\text{LUMO}}$. Notably, PSEUD$\sigma$-D improves the SOTA by 3.2 meV, a significant margin. Particularly, comparing PSEUDσ-S with SchNet and PSEUDσ-D with DimeNet++, we find PSEUDσ can consistently improve even on the fully supervised setting by a large margin (8.1 meV for SchNet and 4.2 meV for DimeNet++), highlighting the utility of PSEUDσ and the high quality of PC9 as unlabeled data. It also shows that this direction of improving learning strategy instead of improving physics-based representation has potentials.

Table 4: 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 5: 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

PSEUDσ significantly improves on low-data QM calculations. In Table 3, we investigate how PSEUDσ can improve on the low-data regime with only 1%, 10% training data point. That is to do QM using only 1,100 and 11,000 known data. This is to simulate the more expensive QM levels such as CCSD(T)/MP2. We observe PSEUDσ can consistently and significantly improve the prediction in $\sigma_{\text{HOMO}}, \sigma_{\text{LUMO}}$ across both low-data settings and both model backbones, suggesting PSEUDσ can help prediction in realistic low-data quantum calculations. Notably, in σ_{LUMO} with 1% of QM9 data, PSEUDσ improves upon SchNet by 57.8 meV, a considerable margin. We also observe that the gain margin is much more significant when the number of training data is smaller. This showcases the utility of PSEUDσ in extremely low-resource settings, such as QM.

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 also find PSEUD$\sigma$ can again significantly improve OOD accuracy over DimeNet++, a SOTA method, with over 16.0 meV improvement on σ_{HOMO} and 8.4 meV improvement on σ_{LUMO}, highlighting the robustness of PSEUDσ.

Evidential uncertainty highly correlates to label quality. PSEUDσ utilizes uncertainty as a proxy of label quality because they are highly correlated for unseen molecules. In this experiment, we want to validate this hypothesis. We train on the complete QM9 training set with evidential uncertainty and then infer on the QM9 testing set. We find that the non-parametric Spearman correlation between MAE and epistemic uncertainty is 0.42 with a 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.

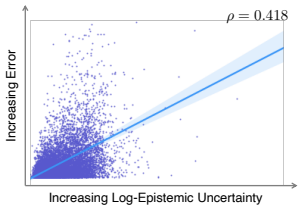

Figure 3: Uncertainty highly correlates to label quality.

Ablations. In Table 5, we conduct a systematic ablation study using SchNet as the backbone architecture. We show that each component in PSEUDσ is indispensable for PSEUDσ. In -pseudo-label, we remove pseudo-labeling and show the usefulness of the general pseudo-labeling strategy. The -uncertainty uses a vanilla per-epoch pseudo-labeling strategy, and the comparison 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 (Xie et al., 2020) is not ideal in QM calculations. The -uniform test, which uses the same weight for all pseudo-labels, shows the importance of detection and adaptive removal of bad pseudo-labels.

7 CONCLUSION

We introduce PSEUDσ, a simple, effective, model-agnostic pseudo-labeling strategy that can improve quantum calculations accuracy in abundant data, low data and out-of-distribution settings. PSEUDσ learns from vast unlabeled data by assigning uncertainty-aware pseudo-labels. These pseudo-labels are adaptively selected to be absorbed into the model via an episodic schedule. Unlike eariler methods in QM that focuses on physics-based representation, we show the potential of data-centric approach.

REFERENCES

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

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

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

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.

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

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

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

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

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, pp. 896, 2013.

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

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, pp. 1052–1060, 2019.

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.

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.

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.

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

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.

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.

O. Anatole von Lilienfeld and Kieron Burke. Retrospective on a decade of machine learning for chemical discovery. *Nature Communications*, 11(1):4895, dec 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-18556-9. URL https://www.nature.com/articles/s41467-020-18556-9http://www.nature.com/articles/s41467-020-18556-9.

Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical Science*, 9(2):513–530, 2018.

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

Yoshua Bengio Yves Grandvalet. Semi-supervised learning by entropy minimization. *NeurIPS*, 2004.

Yoshua Bengio Yves Grandvalet. Entropy regularization. *Semi-Supervised Learning, pages 151–168*, 2006.