Uncertainty Analysis in Predicting Molecular Properties Using Chemical Foundation Models

Siya Kunde[1], Emilio Vital Brazil[2], Priscilla Barreira Avegliano3, Eduardo Soares[2]
[1]IBM Research US, Yorktown Heights, NY, USA
[2]IBM Research Brazil, Rio de Janeiro, RJ, Brazil
[3] IBM Research Brazil, Sao Paulo, Brazil
skunde@ibm.com, evital@br.ibm.com, pba@br.ibm.com, eduardo.soares@ibm.com

Abstract

Large pre-trained foundation models are becoming prevalent and have a high risk impact in domains of the physical sciences. Uncertainty analysis of prediction results can help engender trust in the model outcomes and indicate reliability to decision makers. In this paper, we introduce a method for uncertainty quantification and characterization tailored to chemical foundation models, with a focus on predicting molecular properties. Our approach is tested on a variety of datasets including the widely-used QM9 dataset, ESOL, FreeSolv, Lipophilicity and LD_{50}. We apply our method to a SMILES-based foundation model, comparing the uncertainty profiles between fine-tuned and frozen model versions. We also provide comparison to a conformal prediction method: normalized conformal regressor. Results demonstrate the effectiveness of our approach in identifying and quantifying uncertainties, offering insights into model reliability, the impact of model fine-tuning on prediction results and a comparison to well known method.

Introduction

Foundation models for chemical applications, particularly in predicting molecular and reaction properties, have seen significant advancements, with improvements in the accuracy of predictions for quantum-mechanical properties (Soares et al. 2024; Ross et al. 2022), reaction yields (Jablonka et al. 2024; Boulougouri, Vandergheynst, and Probst 2024), and reaction kinetics (Probst, Schwaller, and Reymond 2022). Additionally, substantial progress has been made in areas such as retrosynthesis (Yang et al. 2024) and forward reaction prediction (Schwaller et al. 2019), further demonstrating the potential of these models in practical chemistry.

Despite these advancements, many foundation models struggle when applied to real-world scenarios (Varshney and Alemzadeh 2017; Angelov et al. 2021). This shortcoming often arises from issues related to generalization, where models fail to accurately predict outcomes outside the narrow domain of the training data (Figueroa et al. 2012). Furthermore, models frequently lack robust mechanisms to identify and filter out erroneous predictions, particularly in edge cases, which can severely impact their reliability in practical applications (Soares and Angelov 2019). A critical factor contributing to these challenges is the misalignment between the training and test datasets and the actual application domain (Angelov and Soares 2021). When the datasets used in model development do not accurately reflect the conditions and variability present in real-world applications, the resulting models may be ill-suited for the intended tasks, leading to significant deviations in performance (Bayram, Ahmed, and Kassler 2022).

The selection of test sets is particularly crucial, as an inappropriate choice can either overestimate or, more commonly, underestimate the model's predictive errors in real-world use (Baumann and Baumann 2014). This underestimation is often more prevalent and problematic, as it can lead to an overconfidence in the model's capabilities, ultimately resulting in poor decision-making in applied contexts.

Furthermore, the unique nature of chemical data and the specific requirements of chemical modeling introduce additional complexities (Baumann and Baumann 2014). Unlike other fields where general guidelines for optimizing foundation models may be broadly applicable, chemical models require careful consideration of factors such as input representations, model architectures, and the intrinsic properties of the datasets used (Horawalavithana et al. 2022; Takeda et al. 2023). These elements differ significantly from those in other domains, making it essential to tailor optimization strategies specifically to the chemical context to achieve accurate and reliable predictions.

Here, we propose a comprehensive investigation into the uncertainties inherent in the predictions of a chemical foundation model, specifically when predicting molecular properties across diverse datasets like QM9, ESOL, FreeSolv, Lipophilicity and LD_{50}. Additionally, we will examine how these uncertainties vary between a fine-tuned version of the model and a version where the model parameters remain frozen during these property predictions. Lastly, we provide a comparison of our method to an existing method: normalized conformal regressor.

Overview of the proposed approach

This section presents an overview of the proposed uncertainty analysis pipeline, detailing its structure and methodology. Additionally, we discuss the method that we compare to : normalized conformal regressor. Lastly we briefly describe the SMILES-based foundation model utilized in this experiment.

Uncertainty Analysis Using Domain & Error Space

In this subsection we describe the method to train and validate uncertainty of predictions. Given a set of disjoint sets for train, validation and test, $X_{train}, X_{val}, X_{test}, y_{train}, y_{val}, y_{test}$, where X's are embeddings from the foundation model and y's are the ground truth for the task. We either use a pre-trained model for prediction or we train a regression model using the X_{train} embeddings as features. From the regression predictions ($y_{val}\_pred, y_{test}\_pred$) we obtain e_{val}, e_{test} which is $abs(y\_pred - y)$. Using these, we run the algorithms to train and validate uncertainty analysis using algorithms 1 and 2.

Algorithm 1: Train Uncertainty Characterization

```
Input: $X_{val}, e_{val}$
Output: Clusters of validation data to characterize uncertainty
Construct graph $G = (V, E)$ where $V$ and $E$ are empty sets
Sort $e_{val}$ in ascending order
Let $e_{val_{unique}}$ be unique values in $e_{val}$
for n in range(0,length($e_{val_{unique}}$) do
    $v = e_{val_{unique}}[n]$
    Construct subgraph $H_n = (V_n, E_n)$ where $V_n$ and $E_n$ are empty sets
    Select $V_n = \{e_{val(i)} \dots e_{val(j)}\} \subset e_{val}$ where all values equal $v$
    Add edges to $E_n$ to connect every two distinct vertices in $V_n$
    Add all vertices from $V_n$ to $V$
    Add all edges from $E_n$ to $E$
    if $\exists V_{n-1}$ then
        Add edges to $E$ by connecting all vertices in $V_{n-1}$ to all vertices in $V_n$
    $V_{n-1} = V_n$
Perform hierarchical clustering using $X_{val}$ and the graph $G$ to obtain a set of
    clusters $C$.
```

Algorithm 2: Validation of Fitted Uncertainty

```
Input: $C, X_{test}, y_{test}, alpha$
Output: Uncertainty characterization for all test items
Classify items in $X_{test}$ to one of the clusters $c \in C$.
Compute $C_{c\_p}$, the $1 - alpha^{th}$ percentile value, for validation data in each
    cluster $c \in C$
Compute the lower and upper bounds (lb and ub respectively) of uncertainty for
    each $k \in y_{test}$ with $k \pm C_{c\_p}$.
Compute interval $I_k = ub_k - lb_k$ for each $k \in y_{test}$.
For all values in $y_{test}$ outside of the interval, mark items as *Outside p* (uncertain),
    else as *Inside p*.
For all $y_{val}$, let $I_{y_{val}} = Q_{95}(y_{val}) - Q_5(y_{val})$.
Size of the Interval Average, SI_A $= {}^1/_{|y_{test}|} \sum_{k \in y_{test}} I_k$.
Relative Size of the Interval Average, RSI_A $=$ SI_A$/I_{y_{val}}$.
```

Algorithm 1 takes as inputs the features and errors of validation set, X_{val} and e_{val} respectively. We construct a graph $G = (V, E)$ where V and E are empty sets. We sort the validation errors values in ascending order and obtain a list of all unique values present in there. For each n^{th} unique value, we create a subgraph $H_n = (V_n, E_n)$ by adding as many number of vertices as there are items with that value to V_n, and add edges to E_n such that each vertex in V_n is connected to every other vertex in V_n. Add the vertices V_n and edges E_n to V and E respectively. Next, each subgraph H_n is fully connected to the subgraph created before itself H_{n-1}, such that there exists an edge from every vertex in V_n to every vertex in V_{n-1}. Add these new edges to E. Save the V_n in V_{n-1} for use in the next iteration. In the last step, we perform a hierarchical clustering (the Agglomerative Clustering implemented by scikit-learn) using X_{val} as features and the adjacency matrix of graph G as connectivity parameters, and obtain a set of clusters C.

Algorithm 2 takes C, X_{test}, y_{test}, $alpha$ as inputs. C are the clusters of validation data produced from Algorithm 1. The significance level, is provided using $alpha$. We train a classifier on the data in C and classify each test sample to one of the clusters. For each cluster, we compute the $1 - alpha^{th}$ percentile value (C_{c_p}) for all validation data in the cluster. For each test sample $k \in y_{test}$, we compute the lower (lb_k) and upper (ub_k) bound with $k \pm C_{c\_p}$ and set interval $I_k = ub_k - lb_k$. All y_{test} outside of the interval are marked as items *Outside p* (uncertain), else as *Inside p*. We compute SI_A as the average size of the interval by taking into account the intervals for all test samples. We also compute RSI_A by dividing SI_A by $I_{y_{val}} = Q_{95}(y_{val}) - Q_5(y_{val})$ to allow comparison across datasets.

Enhanced Adaptability Through Normalized Conformal Regressor

In this paper, we investigate the use of normalized conformal regressor to improve the adaptability of prediction intervals by incorporating difficulty estimates (Johansson, Boström, and Löfström 2021). Traditional conformal intervals are uniform in size, failing to account for instance-specific variability. Normalization is achieved through a *DifficultyEstimator* from the `crepes.extras` library (Boström 2022), which estimates difficulty based on k-nearest neighbors using three approaches: (i) Euclidean distances to the nearest neighbors, (ii) standard deviation of the target values among neighbors, and (iii) absolute errors of the neighbors (Boström 2024).

A parameter β (defaulting to 0.01) regulates the normalization process, with optional min-max scaling to standardize difficulty estimates across different estimators. The normalization process employs a leave-one-out protocol to compute scaling bounds for k-nearest neighbor methods. This study focuses on the first method, leveraging distances to the $k = 25$ nearest neighbors to construct normalized prediction intervals. These intervals are adjusted to reflect instance-specific difficulty, offering a robust methodology for enhanced predictive modeling under varying levels of uncertainty.

SMILES foundation model

For our proposed approach, we utilized the SMI-TED$_{289M}$ foundation model as the SMILES encoder (Soares et al. 2024). SMI-TED$_{289M}$ (shown in the Figue 1) is a large-scale, open-source encoder-decoder model pre-trained on a curated dataset of 91 million SMILES strings from PubChem, encompassing 4 billion molecular tokens. This model has demonstrated superior performance compared to state-of-the-art methods across various molecular tasks.

Experiments

We utilize the proposed method to analyze uncertainty in molecular property prediction. Our approach is tested on the widely-used QM9 dataset, targeting key quantum properties such as the highest occupied molecular orbital (ϵ_{homo}) energy, the lowest unoccupied molecular orbital (ϵ_{lumo}) energy, and the dipole moment ($\langle R^2 \rangle$). Additionally we utilize other datasets like water solubility ($ESOL$), hydration free energy of small molecules in water ($FreeSolv$), Octanol/water distribution coefficient of molecules ($Lipophilicity$) and the median lethal dose (LD_{50}). While the LD_{50} dataset is obtained from (Feinstein et al. 2021), the others were taken from the benchmark dataset MoleculeNet (Wu et al. 2018). Due to the huge size and hence increased processing time, we utilized only 15% of data sampled from each set from the QM9 dataset. We adopt the original train/valid/test set splits for all the tasks to ensure an unbiased assessment. We evaluate the predictions from two versions of the SMI-TED$_{289M}$ model: one with frozen weights and another fine-tuned for the tasks. We repeated the experiments for our method with 10 different random seeds. For the comparison method of normalized conformal regressor we were able to use the only available output from finetuned model that was from a single random seed.

Results and Discussion

Tables 1 and 2 highlight the primary results of this paper. The regression results from the fine-tuned versions of the model produced superior results compared to the one with frozen weights.

Impact of fine tuning model

Normalized conformal regressor By applying Conformal Prediction techniques to estimate the confidence interval of predicted molecular properties, we can see that overall the finetuned model outperforms the frozen model for this purpose. The finetuned model presents a more similar coverage in relation to the to the 95% target. Despite sometimes the frozen model achieving a better coverage, exceeding the 95% for properties (e.g., 98.23% for $ESOL$ and 95.981% for ϵ_{lumo}), it is achieved at the cost of generating larger and over-conservative intervals. For some of the properties (ϵ_{homo}, ϵ_{lumo} and $\langle R^2 \rangle$), the finetuned model presents comparable coverage with smaller intervals.

The property LD_{50} is the only exception in which the finetuned model produces larger interval compared to the frozen model, despite also providing a slightly higher coverage. This indicates that fine-tuning may increase uncertainty for certain properties and this behaviour requires a deeper analysis of some circumstances.

Uncertainty analysis using domain & error space The experiments conducted with our proposed method for uncertainty analysis using domain and error space revealed that overall the finetuned models performed significantly better than the respective frozen models (check Table 3). For properties ϵ_{homo}, ϵ_{lumo}, $Lipophilicity$ and LD_{50} the coverage was significantly larger. The higher coverage for $Lipophilicity$ however came at the expense of significantly larger values for interval sizes, while the same for LD_{50} was not significant. Interestingly, these properties were the ones for which we has larger amount of data compared to $ESOL$ and $FreeSolv$.

$\langle R^2 \rangle$ and $ESOL$ were an exception to this where the frozen model performed significantly better. These observations for $\langle R^2 \rangle$ combined with the results shown in Figure 4 exhibit the difficulty of task in comparison to ϵ_{lumo} for example. There were no significant differences in the performance for the two versions of the model for $FreeSolv$.

Overall, for both the methods tested, the finetuned model offers more precise confidence intervals without compromising coverage, demonstrating improved performance over the frozen model in most cases. This suggests that finetuning can be more suitable for the task of uncertainty estimation.

A comparison of the two methods

The normalized conformal regressor method and the method we proposed both detected superior performance from the finetuned versions of the model for majority of the tasks. While the first provided superior coverage (close to 95%) for both models and across tasks, our method's performance suffered whenever the dataset size was very small. Our method was able to provide smaller intervals for all tasks with the frozen model and a few tasks with the finetuned model. In comparison the other method worked best when the model was superior (finetuned).

Conclusion

In conclusion, this paper introduces a novel method for quantifying and characterizing uncertainties in chemical foundation models. We conduct experiments across a variety of datasets. Our approach, applied to the SMILES-based SMI-TED$_{289M}$ model, effectively differentiates uncertainty profiles between fine-tuned and frozen model versions. The results demonstrate that fine-tuning substantially improves the model's predictions and reduces uncertainty for most properties, particularly for ϵ_{homo}, ϵ_{lumo}, $Lipophilicity$ and LD_{50}. However, the uncertainty values for $\langle R^2 \rangle$ between model versions highlight the complexities involved in predicting various quantum properties. This emphasizes the necessity of robust uncertainty quantification methods to ensure the reliability and trustworthiness of predictions in chemical modeling.

One promising direction as a future work, is to refine the cluster calculation process, exploring more sophisticated methods that can better capture the underlying structure of the data. Additionally, we aim to improve the interval size estimation by incorporating proximity-based measures, which can provide a more nuanced understanding of the relationships between samples. By focusing on the local neighborhood of each sample, we can develop more accurate and informative interval estimates that reflect the inherent uncertainty of the clustering process. By pursuing these advancements, we envision a more robust and reliable uncertainty quantification framework that can be applied to a wide range of real-world applications, ultimately leading to more informed decision-making and improved outcomes.

Table 1: Results of applying the Normalized Conformal Regressor method on two versions of the SMI-TED$_{289M}$ model for the ϵ_{homo}, ϵ_{lumo}, $\langle R^2 \rangle$, $ESOL$, $FreeSolv$, $Lipophilicity$ and LD_{50} tasks.

Experiment Type	Target	InsideP	OutsideP	Size Of Interval Average	Relative Size Of Interval Average
Finetuned	ϵ_{homo}	95.418	4.582	0.016	0.235
	ϵ_{lumo}	95.717	4.283	0.017	0.116
	$\langle R^2 \rangle$	95.120	4.880	90.525	0.111
	ESOL	96.460	3.540	3.080	0.384
	FreeSolv	95.385	4.615	6.018	0.528
	Lipophilicity	94.524	5.476	1.937	0.561
	LD_{50}	96.286	3.714	6.914	2.510
Frozen	ϵ_{homo}	95.269 ± 0.278	4.731 ± 0.278	0.052 ± 0.001	0.754 ± 0.009
	ϵ_{lumo}	95.981 ± 0.205	4.019 ± 0.205	0.090 ± 0.001	0.616 ± 0.008
	$\langle R^2 \rangle$	94.781 ± 0.325	5.219 ± 0.325	700.315 ± 10.831	0.859 ± 0.013
	ESOL	98.230 ± 1.533	1.770 ± 1.533	5.011 ± 0.455	0.624 ± 0.057
	FreeSolv	94.308 ± 1.202	5.692 ± 1.202	8.689 ± 0.517	0.763 ± 0.045
	Lipophilicity	95.738 ± 0.419	4.262 ± 0.419	3.293 ± 0.084	0.953 ± 0.024
	LD_{50}	95.533 ± 0.117	4.467 ± 0.117	3.014 ± 0.019	1.094 ± 0.007

Table 2: Results of applying the proposed method on two versions of the SMI-TED$_{289M}$ model for the ϵ_{homo}, ϵ_{lumo}, $\langle R^2 \rangle$, $ESOL$, $FreeSolv$, $Lipophilicity$ and LD_{50} tasks.

Experiment Type	Target	InsideP	OutsideP	Size Of Interval Average	Relative Size Of Interval Average
Finetuned	ϵ_{homo}	93.137 ± 0.037	6.863 ± 0.037	0.013 ± 0.000	0.190 ± 0.002
	ϵ_{lumo}	97.669 ± 0.030	2.331 ± 0.030	0.020 ± 0.000	0.140 ± 0.000
	$\langle R^2 \rangle$	62.196 ± 0.286	37.804 ± 0.286	264.903 ± 6.476	0.325 ± 0.008
	ESOL	46.814 ± 2.356	53.186 ± 2.356	0.783 ± 0.054	0.098 ± 0.007
	FreeSolv	60.462 ± 1.692	39.538 ± 1.692	1.907 ± 0.054	0.167 ± 0.005
	Lipophilicity	95.881 ± 0.543	4.119 ± 0.543	4.269 ± 0.031	1.235 ± 0.009
	ld50	81.925 ± 0.267	18.075 ± 0.267	2.154 ± 0.010	0.782 ± 0.004
Frozen	ϵ_{homo}	64.417 ± 10.129	35.583 ± 10.129	0.022 ± 0.004	0.317 ± 0.062
	ϵ_{lumo}	64.313 ± 11.622	35.687 ± 11.622	0.035 ± 0.010	0.242 ± 0.070
	$\langle R^2 \rangle$	74.726 ± 10.822	25.274 ± 10.822	577.405 ± 272.818	0.709 ± 0.335
	ESOL	64.513 ± 6.803	35.487 ± 6.803	2.106 ± 0.821	0.262 ± 0.102
	FreeSolv	63.538 ± 8.286	36.462 ± 8.286	2.655 ± 0.484	0.233 ± 0.042
	Lipophilicity	57.119 ± 10.201	42.881 ± 10.201	1.359 ± 0.252	0.393 ± 0.073
	ld50	61.621 ± 9.371	38.379 ± 9.371	1.836 ± 1.257	0.666 ± 0.456

References

Angelov, P.; and Soares, E. 2021. Detecting and learning from unknown by extremely weak supervision: exploratory classifier (xClass). *Neural Computing and Applications*, 33(22): 15145–15157.

Angelov, P. P.; Soares, E. A.; Jiang, R.; Arnold, N. I.; and Atkinson, P. M. 2021. Explainable artificial intelligence: an analytical review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11(5): e1424.

Baumann, D.; and Baumann, K. 2014. Reliable estimation of prediction errors for QSAR models under model uncertainty using double cross-validation. *Journal of cheminformatics*, 6: 1–19.

Bayram, F.; Ahmed, B. S.; and Kassler, A. 2022. From concept drift to model degradation: An overview on performance-aware drift detectors. *Knowledge-Based Systems*, 245: 108632.

Boström, H. 2022. crepes: a Python package for generating conformal regressors and predictive systems. In *Conformal and Probabilistic Prediction with Applications*, 24–41. PMLR.

Boström, H. 2024. Conformal Prediction in Python with crepes. *Proceedings of Machine Learning Research*, 230: 1–14.

Boulougouri, M.; Vandergheynst, P.; and Probst, D. 2024. Molecular set representation learning. *Nature Machine Intelligence*, 1–10.

Feinstein, J.; Sivaraman, G.; Picel, K.; Peters, B.; Vázquez-Mayagoitia, Á.; Ramanathan, A.; MacDonell, M.; Foster, I.; and Yan, E. 2021. Uncertainty-informed deep transfer learning of perfluoroalkyl and polyfluoroalkyl substance toxicity. *Journal of chemical information and modeling*, 61(12): 5793–5803.

Figueroa, R. L.; Zeng-Treitler, Q.; Kandula, S.; and Ngo, L. H. 2012. Predicting sample size required for classification performance. *BMC medical informatics and decision making*, 12: 1–10.

Horawalavithana, S.; Ayton, E.; Sharma, S.; Howland, S.; Subramanian, M.; Vasquez, S.; Cosbey, R.; Glenski, M.; and Volkova, S. 2022. Foundation models of scientific knowledge for chemistry: Opportunities, challenges and lessons learned. In *Proceedings of BigScience Episode# 5–Workshop on Challenges & Perspectives in Creating Large Language Models*, 160–172.

Jablonka, K. M.; Schwaller, P.; Ortega-Guerrero, A.; and Smit, B. 2024. Leveraging large language models for predictive chemistry. *Nature Machine Intelligence*, 6(2): 161–169.

Johansson, U.; Boström, H.; and Löfström, T. 2021. Investigating normalized conformal regressors. In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, 01–08. IEEE.

Probst, D.; Schwaller, P.; and Reymond, J.-L. 2022. Reaction classification and yield prediction using the differential reaction fingerprint DRFP. *Digital discovery*, 1(2): 91–97.

Ross, J.; Belgodere, B.; Chenthamarakshan, V.; Padhi, I.; Mroueh, Y.; and Das, P. 2022. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12): 1256–1264.

Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; and Lee, A. A. 2019. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS central science*, 5(9): 1572–1583.

Soares, E.; and Angelov, P. 2019. Novelty detection and learning from extremely weak supervision. *arXiv preprint arXiv:1911.00616*.

Soares, E.; Shirasuna, V.; Brazil, E. V.; Cerqueira, R.; Zubarev, D.; and Schmidt, K. 2024. A Large Encoder-Decoder Family of Foundation Models For Chemical Language. *arXiv preprint arXiv:2407.20267*.

Takeda, S.; Kishimoto, A.; Hamada, L.; Nakano, D.; and Smith, J. R. 2023. Foundation model for material science. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 15376–15383.

Varshney, K. R.; and Alemzadeh, H. 2017. On the safety of machine learning: Cyber-physical systems, decision sciences, and data products. *Big data*, 5(3): 246–255.

Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; and Pande, V. 2018. MoleculeNet: a benchmark for molecular machine learning. *Chemical science*, 9(2): 513–530.

Yang, Y.; Shi, R.; Li, Z.; Jiang, S.; Lu, B.-L.; Yang, Y.; and Zhao, H. 2024. BatGPT-Chem: A Foundation Large Model For Retrosynthesis Prediction. *arXiv preprint arXiv:2408.10285*.

Appendix

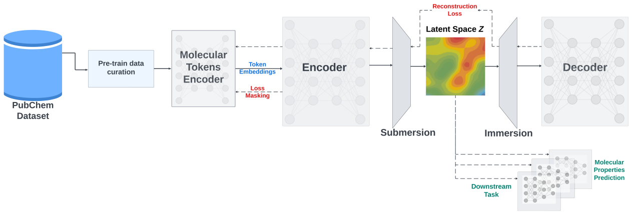
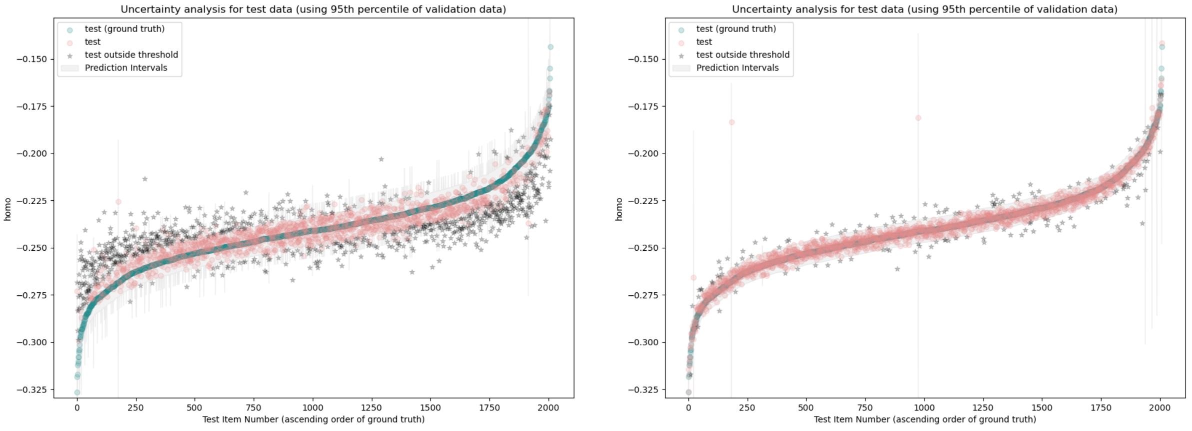


Figure 1: This figure illustrates the architecture of SMI-TED$_{289M}$.

Table 3: Results of statistical testing (t-test) for the proposed method to compare the frozen and finetuned versions of the SMI-TED$_{289M}$ model for the ϵ_{homo}, ϵ_{lumo}, $\langle R^2 \rangle$, $ESOL$, $FreeSolv$, $Lipophilicity$ and LD_{50} tasks.

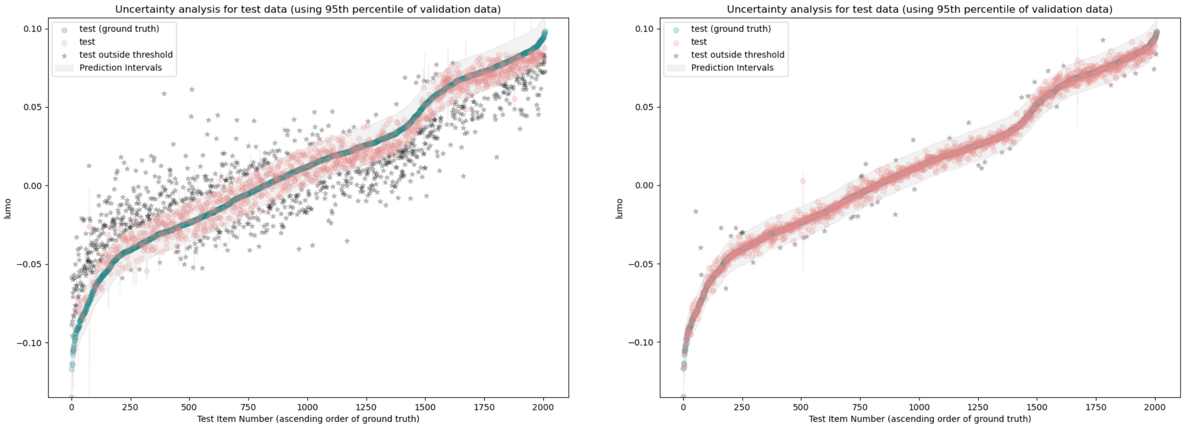
Target	Metric	FrozenMean	FrozenStd	FinetunedMean	FinetunedStd	t	p
ϵ_{homo}	InsideP	64.417300	10.129100	93.137500	0.037300	-8.498298	0.000014
ϵ_{lumo}	InsideP	64.312700	11.622500	97.669300	0.029900	-8.599189	0.000012
$\langle R^2 \rangle$	InsideP	74.726100	10.822000	62.196200	0.285600	3.486434	0.006869
ESOL	InsideP	64.513300	6.802700	46.814200	2.356400	6.819943	0.000077
FreeSolv	InsideP	63.538500	8.286300	60.461500	1.692300	1.150447	0.279609
Lipophilicity	InsideP	57.119000	10.200600	95.881000	0.543500	-11.685147	0.000001
LD_{50}	InsideP	61.620800	9.371300	81.924700	0.266900	-6.579965	0.000102
ϵ_{homo}	OutsideP	35.582700	10.129100	6.862500	0.037300	8.498298	0.000014
ϵ_{lumo}	OutsideP	35.687300	11.622500	2.330700	0.029900	8.599189	0.000012
$\langle R^2 \rangle$	OutsideP	25.273900	10.822000	37.803800	0.285600	-3.486434	0.006869
ESO	OutsideP	35.486700	6.802700	53.185800	2.356400	-6.819943	0.000077
FreeSolv	OutsideP	36.461500	8.286300	39.538500	1.692300	-1.150447	0.279609
Lipophilicity	OutsideP	42.881000	10.200600	4.119000	0.543500	11.685147	0.000001
LD_{50}	OutsideP	38.379200	9.371300	18.075300	0.266900	6.579965	0.000102
ϵ_{homo}	SizeOfIntervalAverage	0.021900	0.004300	0.013100	0.000100	6.106024	0.000178
ϵ_{lumo}	SizeOfIntervalAverage	0.035100	0.010200	0.020300	0.000000	4.367917	0.001803
$\langle R^2 \rangle$	SizeOfIntervalAverage	577.404700	272.818500	264.903400	6.475600	3.430510	0.007502
ESO	SizeOfIntervalAverage	2.106300	0.820700	0.782900	0.054000	4.693416	0.001131
FreeSolv	SizeOfIntervalAverage	2.655400	0.483700	1.907000	0.054000	4.671083	0.001167
Lipophilicity	SizeOfIntervalAverage	1.359200	0.251800	4.268800	0.031100	-36.388550	0.000000
LD_{50}	SizeOfIntervalAverage	1.835700	1.256800	2.154400	0.009900	-0.758642	0.467468
ϵ_{homo}	RelativeSizeOfIntervalAverage	0.317400	0.062100	0.190400	0.001500	6.106024	0.000178
ϵ_{lumo}	RelativeSizeOfIntervalAverage	0.241600	0.069900	0.139700	0.000100	4.367917	0.001803
$\langle R^2 \rangle$	RelativeSizeOfIntervalAverage	0.708600	0.334800	0.325100	0.007900	3.430510	0.007502
ESO	RelativeSizeOfIntervalAverage	0.262400	0.102200	0.097500	0.006700	4.693416	0.001131
FreeSolv	RelativeSizeOfIntervalAverage	0.233100	0.042500	0.167400	0.004700	4.671083	0.001167
Lipophilicity	RelativeSizeOfIntervalAverage	0.393400	0.072900	1.235500	0.009000	-36.388550	0.000000
LD_{50}	RelativeSizeOfIntervalAverage	0.666400	0.456200	0.782100	0.003600	-0.758642	0.467468



(a) ϵ_{homo} frozen model

(b) ϵ_{homo} finetuned model

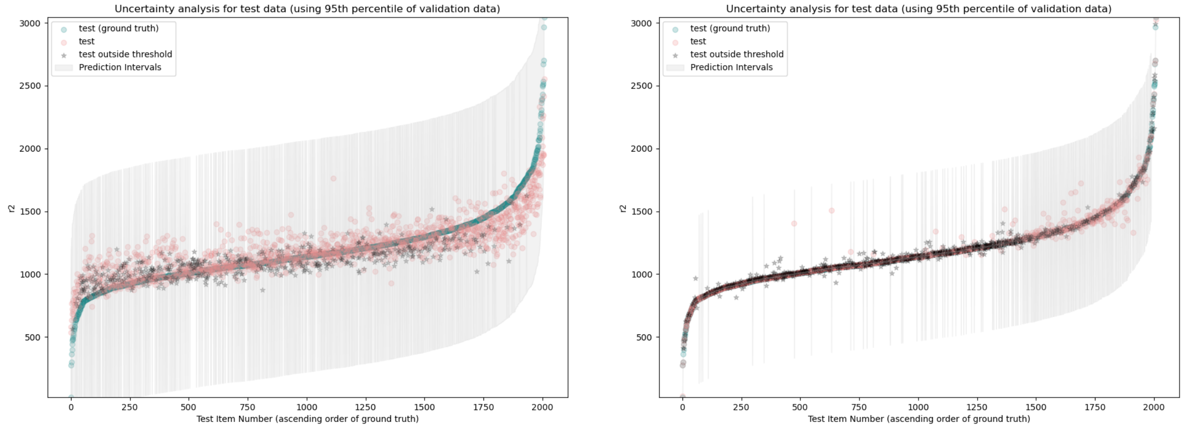
Figure 2: Results for ϵ_{homo} obtained with SMI-TED$_{289M}$ frozen and fine-tuned versions, with our proposed method.



(a) ϵ_{lumo} frozen model

(b) ϵ_{lumo} finetuned model

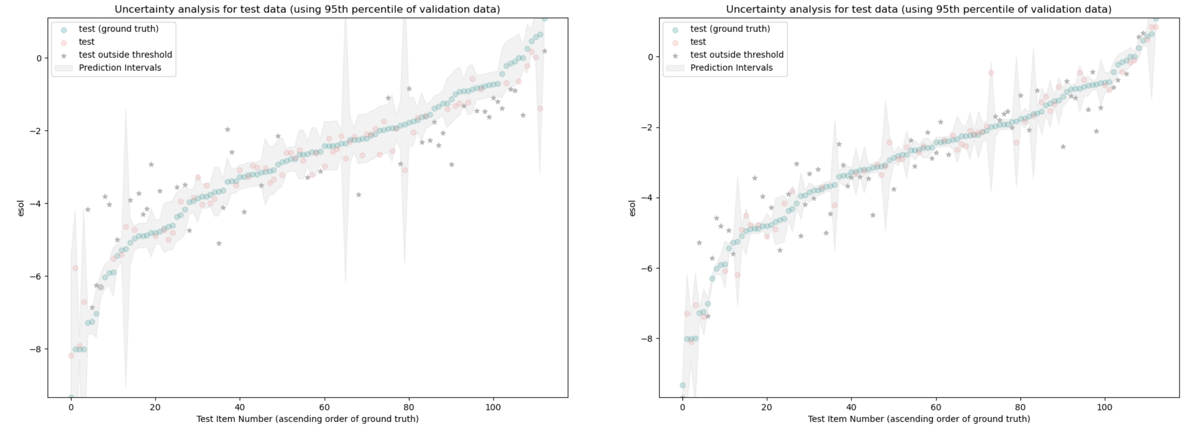
Figure 3: Results for ϵ_{lumo} obtained with SMI-TED$_{289M}$ frozen and fine-tuned versions, with our proposed method.



(a) $\langle R^2 \rangle$ frozen model

(b) $\langle R^2 \rangle$ finetuned model

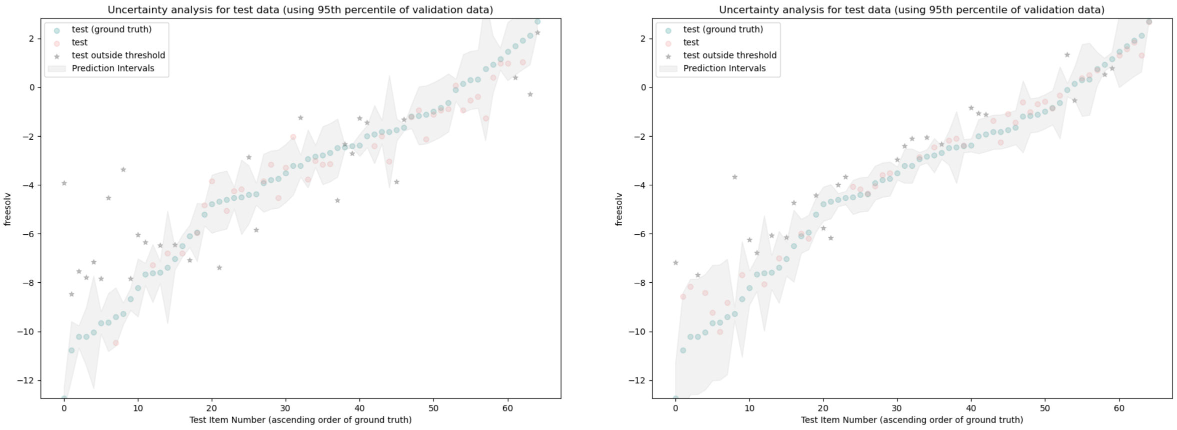
Figure 4: Results for $\langle R^2 \rangle$ obtained with SMI-TED$_{289M}$ frozen and fine-tuned versions, with our proposed method.



(a) ESOL frozen model

(b) ESOL finetuned model

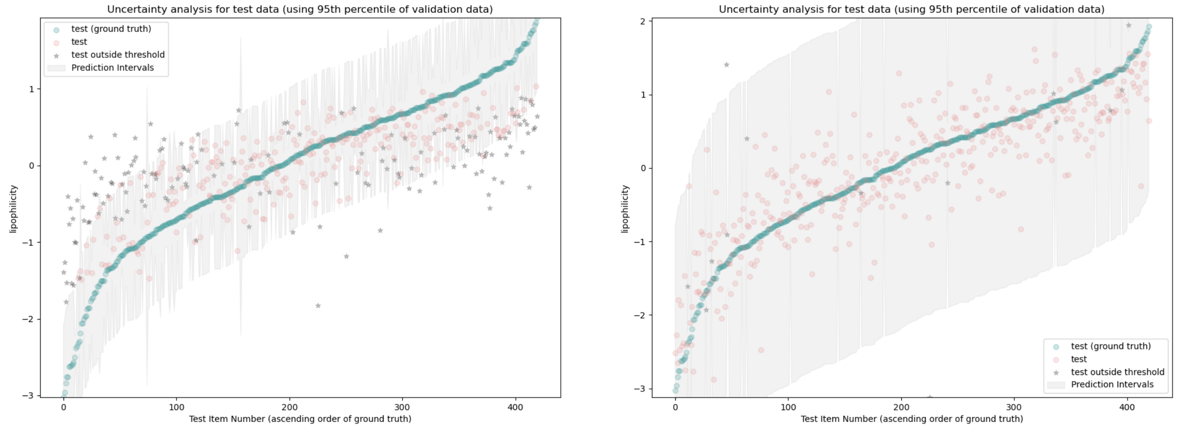
Figure 5: Results for ESOL obtained with SMI-TED$_{289M}$ frozen and fine-tuned versions, with our proposed method.



(a) FreeSolv frozen model

(b) FreeSolv finetuned model

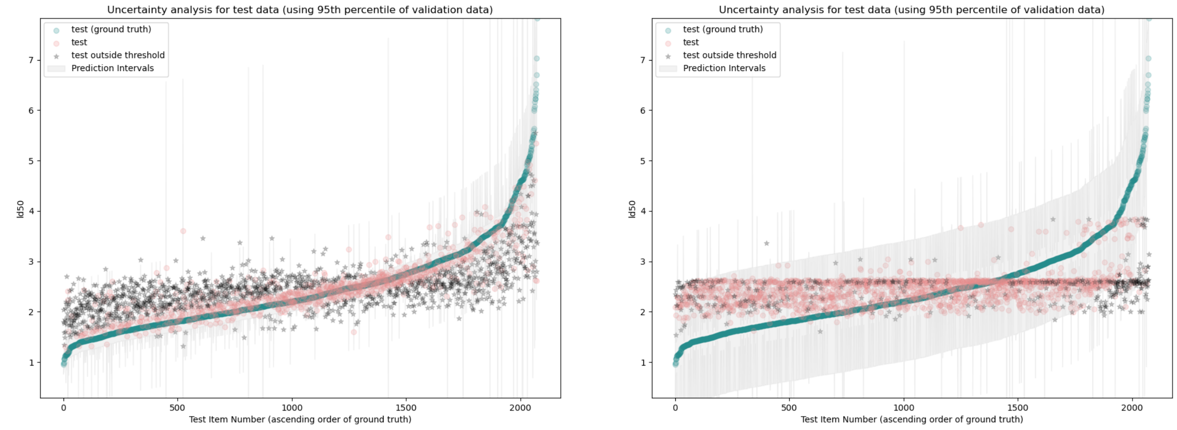
Figure 6: Results for FreeSolv obtained with SMI-TED$_{289M}$ frozen and fine-tuned versions, with our proposed method.



(a) Lipophilicity frozen model

(b) Lipophilicity finetuned model

Figure 7: Results for Lipophilicity obtained with SMI-TED$_{289M}$ frozen and fine-tuned versions, with our proposed method.



(a) LD_{50} frozen model

(b) LD_{50} finetuned model

Figure 8: Results for LD_{50} obtained with SMI-TED$_{289M}$ frozen and fine-tuned versions, with our proposed method.