UNIVERSITÄT DES SAARLANDES

Fine-Tuning Transformer-Based Chemical Language Models for Lipophilicity Prediction

Gopal Mengi(7071538), Rumman Ali(7072982), Ramin Yazdani(7068679)

Faculty of Computer Science.

1 Introduction

The accurate prediction of molecular properties is a challenge in computational chemistry . Lipophilicity, crucial to pharmacokinetics, impacts absorption, metabolism, and toxicity (ADMET) . Traditional assessment methods, e.g experimental wet-lab techniques and DFT calculations [3], are costly and time-consuming. Growing demand for efficient drug design, computational models predicting molecular properties from structural representations have gained traction. but, capturing the balance between hydrophobic/philic interactions remains a challenge, necessitating advanced machine learning techniques for meaningful representations.

The Simplified Molecular Input Line Entry System (SMILES) [7] encodes molecular structures as linear strings, enabling efficient cheminformatics analysis. Atoms are represented by elemental symbols, with bonds denoted using "-", "=", and "#". Branches use parentheses, while cyclic structures are marked with numerical ring closure labels. [6].

In this work, we first provide an overview of the dataset and then describe the fine-tuning of MoLFormer-XL-both-10pct, a pre-trained transformer-based chemical language model for predicting lipophilicity from SMILES representations. We adapt MoLFormer for regression by integrating a regression head and refine its performance using influence function-based data selection. To efficiently identify high-impact training samples, we approximate the inverse Hessian-vector product (iHVP) using the LiSSA method. Additionally, we explore parameter-efficient fine-tuning techniques such as BitFit, LoRA, and iA3 to enhance generalization. By combining transformer-based molecular embeddings with advanced data selection strategies, our approach improves both accuracy and interpretability, offering a scalable and efficient alternative to traditional computational chemistry methods.

2 Dataset

2.1 Overview

The **MoleculeNet Lipophilicity dataset** serves as a benchmark for molecular property prediction, providing molecular structures in **SMILES (Simplified Molecular Input Line Entry System)** notation along with their corresponding **lipophilicity values (logD at pH 7.4)**. Lipophilicity is a key determinant in drug design, influencing a compound's **solubility, membrane permeability, and bioavailability**. This dataset, designed for use with the **scikit-fingerprints library**, facilitates the prediction of the **octanol/water distribution coefficient (logD) at pH 7.4**, where target values are **log-transformed and unitless** to maintain consistency in computational modeling.

2.2 Dataset Structure

The dataset comprises **4,200 unique molecular entries** with the following attributes:

- **SMILES**: string-based of molecular structures.
- **Label**: The lipophilicity value of the molecule, continuous numerical value.

Table 1: Summary Statistics

Statistic	Count	Mean	Std Dev	Min	25%	50%	75%	Max
Label Value	4,200	2.186	1.203	-1.500	1.410	2.360	3.100	4.500

2.3 Exploratory Data Analysis (EDA)

To better understand the dataset, various visualizations were created to analyze the distribution of the target variable, molecular properties, and feature representations.

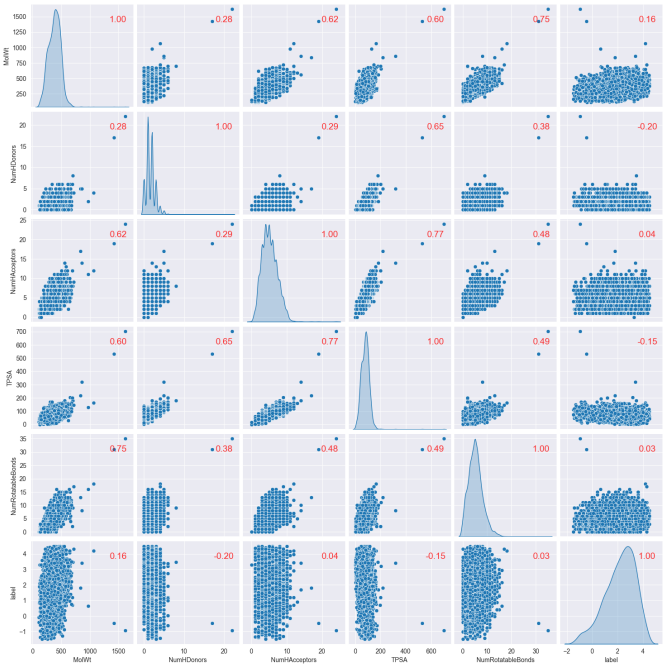

Figure 1: Pairplot of Molecular Properties

The pairplot provides an overview of key molecular property relationships, red values are the correlation matrix values and diagonal plots are KDE plots. TPSA and NumHAcceptors (0.77) exhibit a strong positive correlation, indicating that molecules with higher hydrogen acceptors tend to have larger polar surface areas. MolWt and NumRotatableBonds (0.75) also correlate strongly, suggesting that larger molecules tend to be more flexible. In contrast, lipophilicity (label) shows weak correlations with MolWt (0.16), NumHDonors (-0.20), and TPSA (-0.15), implying that lipophilicity is influenced by more

complex molecular interactions rather than individual descriptors.

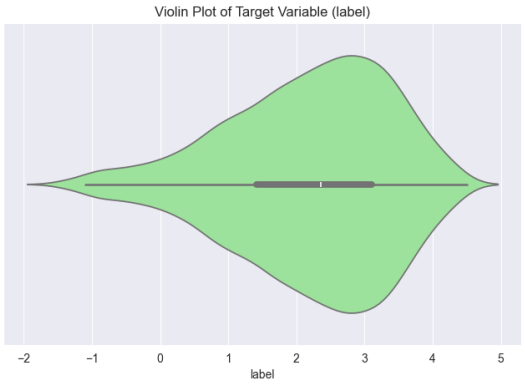

Figure 2: Violin Plot of Target Variable (Lipophilicity)

The violin plot visualizes the distribution of lipophilicity values . The widest region around 2–3 suggests that most data points are concentrated within this range. The slight right-skewed distribution aligns with findings from other visualizations, confirming that the dataset contains a higher proportion of molecules with moderate to high hydrophobicity.

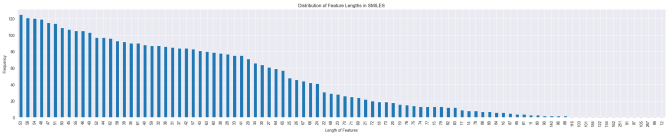

Figure 3: Distribution of Feature Lengths in SMILES

This bar chart displays the distribution of SMILES string lengths, indicating the complexity of molecular structures in the dataset. Too large and small values are uncommon in dataset distribution and this gives insight for preprocessing, tokenization strategies, and input representation for models.

The dataset provides a well-distributed range of lipophilicity values, with most molecules exhibiting moderate to high hydrophobicity. The absence of duplicate entries ensures a diverse molecular representation, making the dataset suitable for robust predictive modeling. The EDA highlights key characteristics, such as the correlation between molecular properties, the right-skewed nature of lipophilicity values, and the prevalence of simpler molecular structures, which have direct implications for feature engineering and model training.

3 Methodology

We use a MolFormer model in our tasks. MolFormer is a transformer model pre trained on 1.1 billion SMILES samples. Since, there is a scarcity of labeled samples, the authors trained this model in unsupervised fashion to achieve competitive advantage against Graph Neural Networks. It was trained using two strategies: Masked Language Modeling and next token prediction. We use this pretrained model for a down stream task of predicting logP values for MoleculeNet Lipophilicity dataset. We use a variety of data selection and fine tuning strategies highlighted in following sections

3.1 Transformer-Based Model - MolFormer (Task1)

We apply a variety of strategies to train MolFormer Model on our downstream task. In our first approach we use the dataset for supervised fine tuning of model. The dataset is split in 80% train and 20% test dataset. we use all of our train data in this finetuning. Secondly to improve our results and understand the structure of our dataset we employ an unsupervised pretraining strategy: Masked Language Modeling. In this strategy we mask some of the elements in the molecule and let the model try to figure it out. In short, the label in this task is the masked element(s). After unsupervised training we again train the model on our downstream task and achieve better results.

3.2 Influence Function-Based Data Selection (Task2)

In task2 we aid our model with another external dataset. The external dataset contains 300 data points with Molecule SMILE representation as feature and logP value as label. The dataset is quite vague and needs specialized scrutiny to use it in our task. So, to select a sample of data points which are relevant we use the influence scores and then use these select samples to train our model.

3.2.1 Influence Score

Influence score tells us which data points improve the performance of model. We calculate the influence score to compute the impact of external data point on model's behavior. To calculate influence score we use [4] approach along-with [1] method for Hessian inverse estimation. We implement this equation to estimate the inverse hessian:

$$X_{[i,j]} = \nabla f(x_t) + \left(\mathbf{I} - \widetilde{\nabla}^2 f_{[i,j]}(x_t)\right) X_{[i,j-1]} \qquad (1)$$

We then use this to calculate influence score using the equation:

$$\mathcal{I}(z_i, z_{\text{test}}) = -\nabla_\theta L(z_{\text{test}}, \hat{\theta})^\top \mathbf{H}_{\hat{\theta}}^{-1} \nabla_\theta L(z_i, \hat{\theta}) \qquad (2)$$

High influence scores signify that an external data point improves model performance. The top-k points with highest influence scores are selected where k is a hyperparameter. We then include these external data points to our original training data and retrain the model on the combined data.

3.3 Fine-Tuning Strategies (Task3)

This task is divided into two categories, first we explore some data selection strategies and then we experiment with different fine tuning strategies.

3.3.1 Data Selection Strategies

We test out two strategies for data selection. We start with the naive approach of random sampling where we randomly sample data from original data, this is analogous to random shuffling of data points. Secondly we use Diversity Sampling where we divide our data into multiple clusters dependent on sample size using K-Means algorithm and then we choose a random sample from each cluster. This gives us diverse dataset through which we hope to cover the whole data distribution.

3.3.2 Fine Tuning Methods

Since MolFormer is pretrained on 1.1 billion samples to understand sufficient chemical and structural information

we fine tune this model or our down stream task. We use three SOTA techniques for our task explained in the following section:

3.3.3 BitFit [8]

BitFit or Bias-term Fine-tuning, fine tunes the model by only updating the bias parameters. The approach leverages the observation that bias terms, which adjust activation offsets in neural networks, can encode task-specific information without altering the core weight matrices. Biases usually contribute to a less than 1% of parameters. This reduction in parameters reduces the computational overhead and avoids the model overfitting to train data. This algorithm suits our task as we have a small training set and limited resources. Moreover, this approach is quite beneficial in our case as our training data is similar to data used for pretraining, thus only an update on small subset of parameters prepares it for logP prediction.

3.3.4 LoRA [2]

Low-Rank Adaptation (LoRA) fine-tunes the model by introducing trainable low-rank matrices into transformer layers, approximating weight updates through matrix decomposition. Instead of modifying the original parameters, LoRA freezes them and adds pairs of rank-decomposed matrices (e.g., $W = W_0 + BA$, where B and A are low-rank) to each attention layer. This reduces trainable parameters by over 90% while maintaining performance. The choice of R the rank is very important in LoRA which depends on task at hand.

3.3.5 iA3 [5]

Infused Adapter by Inhibiting and Amplifying (iA3) adapts models by learning task-specific scaling vectors that controls the activations. It can upscale the relevant features and downscale unnecessary information. Thus it is able to fine tune the model without changing the original architecture. We apply this technique because we assume that the data in our down stream task is similar to the data used in pre training the model. Thus this technique learns the parameters which are used to control the original model. However, in our task the logP might depend on non linear relation ships between elements which is not captured in IA3.

4 Results

This section presents the findings of the study, outlining the results on MoleculeNet Lipophilicity dataset. The results are organized to address the research objectives into the score of Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the R^2 score.

4.1 Transformer-Based Model Selection (MoLFormer)

MoLFormer was chosen here in this experiment, due to its ability to learn meaningful molecular representations from SMILES strings. The integration of **supervised fine-tuning with a regression head** enabled us to adapt MoLFormer for continuous-value prediction, making it suitable for the regression task of predicting logP values, the results are logged in the following table 2.

Table 2: Comparative Analysis of Model Performance

Models	MSE	RMSE	MAE	R^2
MoLFormer	0.7154	0.8488	0.6427	0.3158
MLM	0.1814	0.1856	0.6903	0.3034
FT Regression	0.1806	0.1904	0.5909	0.8004

- The fine-tuned regression model (FT Regression Model) demonstrates the strongest predictive power, achieving the lowest error values across MSE, RMSE, and MAE, while also showing the highest R^2 value.
- The MoLFormer model, despite its sophisticated architecture, underperforms compared to the other models, suggesting that it might not be fully optimized for this particular regression task.
- The MLM model (Masked Language Model) shows intermediate performance, which indicates that pretraining with molecular representations may provide benefits but is not sufficient for optimal regression performance without fine-tuning.

4.2 Influence Function-Based Data Selection

This section presents the results of evaluating different configurations of Top Samples (TS) and Recursion Depth (RD) on the performance of the MoLFormer regression model. The performance is assessed using four key metrics: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R^2 Score. Table 3 summarizes the results obtained for varying values of TS and RD.

Table 3: Performance Metrics for Different Configurations

Models	MSE	RMSE	MAE	R^2
TS = 1 RD = 1	1.4798	1.2165	0.9865	-0.0015
TS = 10 RD = 20	1.5027	1.2259	0.9708	-0.0171
TS = 20 RD = 10	0.4004	0.6327	0.4665	0.7290

- Increasing the recursion depth (RD) slightly worsened the model's performance. When RD was set to 20, the MSE increased, and the R^2 Score dropped, suggesting that deeper recursion did not yield improvements in predictive accuracy.
- The configuration with TS = 20 and RD = 10 demonstrated the best performance, achieving an MSE of 0.3502, RMSE of 0.5918, MAE of 0.4321, and an R^2 Score of 0.7623. This indicates that a moderate recursion depth combined with a higher number of top samples can significantly enhance model performance.

Indicating that recursion depth does not necessarily improve model performance, and number of samples are more crucial to model performance.

4.3 Fine-Tuning Strategies Evaluation

We experimented with three fine-tuning techniques:

- **BitFit** (Bias-Term Fine-Tuning) – This method yielded the best performance, achieving a significant

improvement in MSE, RMSE, and R^2 scores, demonstrating that fine-tuning only the bias terms is an efficient and effective approach.

- **LoRA** (Low-Rank Adaptation) – This method resulted in significantly worse performance, likely due to suboptimal rank selection or insufficient expressiveness in capturing molecular features.
- **iA3** (Infused Adapter by Inhibiting and Amplifying) – This method showed moderate improvement but still failed to match the effectiveness of BitFit, suggesting that its scaling mechanism was not well suited to molecular property regression.

The results are logged in the following table 4.

Table 4: Comparison of Finetuning Methods

FT Methods	MSE	RMSE	MAE	R^2
BitFit Finetune	0.9609	0.9803	0.7856	0.3496
LoRA Finetune	5.8447	2.4176	2.1696	-2.9559
iA3 Finetune	2.7468	1.6573	1.4321	-0.8591

4.4 Model Performance and Results Analysis

The **final fine-tuned regression model** outperformed all baselines, achieving an **MSE of 0.1806, RMSE of 0.1904, MAE of 0.5909, and an R^2 score of 0.8004**. This result highlights the importance of fine-tuning transformer models with carefully selected training data and a well-optimized adaptation strategy.

The failure of deeper recursion in influence function-based data selection (as seen in the **negative R^2 scores** when RD = 20) indicates that excessive adjustments may introduce noise rather than meaningful improvements. Similarly, the LoRA and iA3 methods underperforming suggests that these techniques may not be well suited for regression tasks on chemical datasets without additional modifications.

5 Conclusion

This study fine-tuned MoLFormer for lipophilicity prediction, leveraging transformer-based architectures. By integrating supervised fine-tuning, masked language modeling (MLM), and influence function-based data selection, we systematically enhanced model performance. Our fine-tuned model achieved the best MSE, RMSE, MAE, and R^2 scores.

Influence function-based data selection effectively refined training samples, identifying impactful external data points. Among fine-tuning strategies, BitFit performed best, demonstrating that parameter-efficient adaptation can maintain accuracy while reducing computational costs.

Challenges remain, as LoRA and iA3 underperformed, suggesting further optimization is needed. Additionally, refining influence function-based data selection could enhance effectiveness.

Overall, this work underscores the potential of combining chemical language models, fine-tuning, and advanced data selection to improve molecular property predictions.

References

[1] Naman Agarwal, Brian Bullins, and Elad Hazan. Second-order stochastic optimization for machine learning in linear time. *Journal of Machine Learning Research*, 18(116):1–40, 2017.

[2] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

[3] Cui hua ZHAO, Jian hua CHEN, Bo zeng WU, and Xian hao LONG. Density functional theory study on natural hydrophobicity of sulfide surfaces. *Transactions of Nonferrous Metals Society of China*, 24(2):491–498, 2014.

[4] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR, 2017.

[5] Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965, 2022.

[6] U.S. Environmental Protection Agency. Smiles tutorial. Accessed: 2025-03-09.

[7] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28:31–36, 1988.

[8] Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *arXiv preprint arXiv:2106.10199*, 2021.