Exploration of Data Selection and Fine-Tuning Methods for Lipophilicity Prediction

Aishwarya Hiwrale 7028820
Deepali Pawade 7057787
Pooja Halannavar 7058527

Abstract

Interpreting complex chemical data requires selecting relevant features that accurately represent the properties of molecules, such as lipophilicity, which plays a critical role in drug absorption and efficacy. In this report, we improve a chemical language model's ability to predict lipophilicity by first applying data selection techniques, such as influence functions, cosine similarity, and Tanimoto similarity measures, to ensure proper data alignment. We then fine-tune a pre-trained model using parameter-efficient fine-tuning (PEFT) methods, including BitFit, LoRA, and iA3, which update only a small subset of the model's weights, reducing computational costs. Our results demonstrate that careful data selection combined with efficient fine-tuning strategies significantly improves predictive performance while maintaining computational efficiency, highlighting their importance in machine learning applications for chemistry.

1 Introduction

Modern machine learning models have shown great promise in predicting molecular properties such as lipophilicity, a key factor in drug absorption and efficacy. However, achieving high predictive accuracy requires not only quality training data but also efficient fine-tuning techniques. In this work, we enhance a chemical language model's performance by fine-tuning a pre-trained model on a dedicated lipophilicity dataset while employing advanced data selection and parameter-efficient fine-tuning (PEFT) methods.

To improve the model's performance, we employ influence functions to identify the most significant external data points that contribute to the model's accuracy. In addition to influence functions, we experiment with two other data selection strategies namely cosine similarity and Tanimoto similarity, chosen to ensure that the selected external data align with the structural and chemical properties of the training set. To further optimize the fine-tuning process, we use parameter-efficient fine-tuning (PEFT) methods, including BitFit, LoRA, and iA3, which minimize computational costs by updating only a small subset of the model's weights.

In Section 3, we briefly discuss the fine-tuning process, Section 4 outlines data selection strategies, and in the subsequent sections, we discuss PEFT methods as well as our experiments, results, and challenges.

2 Base Model

2.1 Dataset Description

The dataset used for training and evaluation is the **Lipophilicity dataset** from the MoleculeNet benchmark (Wu, 2018). Lipophilicity, also known as hydrophobicity, measures how well a substance dissolves in nonpolar solvents (e.g., oil) compared to polar solvents (e.g., water). The dataset contains SMILES molecular representations along with their experimentally measured lipophilicity values.

2.2 External Dataset Used for Additional Training

The dataset is sourced from Hugging Face under the name: MoleculeNet Lipophilicity Dataset. The data is split into an 80-20 ratio for training and testing. Tokenization is performed before training to convert SMILES representations into a format suitable for transformers.

2.3 Pre-Trained Model Used

The selected model is **MoLFormer-XL-both-10pct** from IBM, a transformer-based chemical language model. A custom *regression head* is added to adapt the model for the prediction task. This head consists of:

- A linear layer mapping hidden states to a 256-dimensional space.
- ReLU activation.
- Dropout layer (0.1) to prevent overfitting.
- A final linear layer for outputting a single scalar value.

3 Fine-Tuning Process

This section describes the process of fine-tuning the MoLFormer model using Masked Language Modeling (MLM) followed by regression fine-tuning for molecular property prediction. The aim is to enhance the model's ability to understand chemical structures before training it for a downstream regression task.

3.1 Masked Language Modeling (MLM)

The pre-trained MoLFormer-XL-both-10pct model from IBM is used for MLM. The model is initialized along with its tokenizer, and a cross-entropy loss function is employed for training. The AdamW optimizer is chosen with a specified learning rate to ensure stable convergence during training.

To improve the model's ability to learn meaningful representations of chemical structures, a masking strategy is applied. In this strategy, a portion of the input tokens (typically 15%) is randomly masked while ensuring that special tokens such as classification and padding tokens remain unchanged. The model is then trained to predict the masked tokens, encouraging it to develop a deeper understanding of molecular patterns.

The MLM training process is carried out for multiple epochs. During each epoch, batches of input data are processed where the masked inputs are passed through the model, and the loss is computed based on the predicted token probabilities. The optimizer updates the model weights iteratively to minimize the loss. Once training is completed, the fine-tuned MLM model is saved for later use.

3.2 Regression Fine-Tuning

After the MLM step, a regression head is added to the fine-tuned MoLFormer model. This regression head consists of a dropout layer to prevent overfitting and a fully connected linear layer that maps the learned representations to a single scalar value, which corresponds to the predicted molecular property.

For the regression task, the model is further fine-tuned using a dataset of labeled molecular data. A Mean Squared Error (MSE) loss function is used to measure the difference between predicted and actual values, while the AdamW optimizer with a lower learning rate is applied to fine-tune the model effectively. The model undergoes multiple training epochs, where the weights are updated iteratively to improve performance.

At the end of the fine-tuning process, the final regression model is saved and can be used for molecular property prediction tasks. By incorporating an MLM phase before regression fine-tuning, the model gains a better contextual understanding of molecular structures, leading to improved predictive accuracy in the regression task.

4 Data Selection Strategies

Motivation Machine learning model performance heavily depends on the quality and relevance of the training data. Selecting appropriate data can improve generalization and efficiency, while irrelevant data may degrade performance. This study explores various data selection techniques for improving lipophilicity prediction, focusing on selecting relevant samples from an external dataset of molecular SMILES strings and their lipophilicity values to enhance the model's accuracy.

4.1 Influence Functions

Influence functions, introduced by Koh and Liang (Koh and Liang, 2020), quantify the impact of individual training samples on model predictions by approximating the influence of a training example on the test loss. Direct computation of the Hessian matrix is expensive, so we use the LiSSA approximation (Agarwal et al., 2016) to iteratively estimate the inverse Hessian-vector product (iHVP). The influence score for a training point z_i on a test point z_t is given by:

$$I(z_i, z_t) = -\nabla_\theta L(z_t, \theta)^T H_\theta^{-1} \nabla_\theta L(z_i, \theta), \quad (1)$$

where H_θ is the Hessian and $\nabla_\theta L$ represents the gradient. Influence scores are computed for all samples in the external dataset, and the top-k samples are selected based on their absolute influence scores. These selected samples are then added to the original lipophilicity dataset for fine-tuning.

4.2 Cosine Similarity

Cosine similarity (Salton, 1983) measures the angle between two vectors in a high-dimensional space, commonly used for comparing molecular embeddings. In our case, molecular embeddings are derived from Morgan fingerprints (ECFP) (Rogers and Hahn, 2010), which encode the structural and chemical properties of compounds. Given two vectors A and B, cosine similarity is computed as:

$$\text{cosine\_sim}(A, B) = \frac{A \cdot B}{\|A\|\|B\|} \tag{2}$$

For data selection, we compute cosine similarity between each external dataset sample and the Lipophilicity training data in the embedding space, selecting the most similar samples for fine-tuning. Using the RDKit (Landrum, 2013) library for generating Morgan fingerprints and computing similarities, we found that a threshold of 0.7 resulted in the selection of 132 relevant samples from an original 300.

4.3 Tanimoto Similarity

Tanimoto similarity, also known as the Jaccard index for binary data, is widely used for comparing molecular fingerprints in cheminformatics (Tanimoto, 1963). It is defined as:

$$\text{Tanimoto}(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{3}$$

We apply Tanimoto similarity, computed using the RDKit (Landrum, 2013) library, to select relevant samples from the external dataset. After experimenting with thresholds ranging from 0.5 to 0.9, we found that a threshold of 0.7 provided an optimal balance between specificity and coverage, resulting in the selection of 127 samples from the external dataset, down from 300.

5 Fine-Tuning with PEFT Methods

Motivation Fine-tuning large pre-trained models is computationally expensive due to the large number of parameters. Parameter-efficient fine-tuning (PEFT) methods alleviate this issue by updating only a small subset of parameters, reducing computational costs while maintaining performance. These methods are especially beneficial when computational resources are limited or when adapting a model for multiple tasks.

In this work, we apply several PEFT methods: BitFit (Ben-Zaken et al., 2021), LoRA (Hu et al., 2021), and iA3 (Liu et al., 2022), to fine-tune a pre-trained model for the lipophilicity prediction task. BitFit updates only bias terms, LoRA introduces low-rank matrices, and iA3 scales activations. These methods were selected for their ability to reduce computational requirements while effectively adapting the model. We compare their performance and evaluate their effectiveness when combined with the data selection strategies discussed earlier.

6 Experiments and Results

6.1 Data Selection Strategy Comparison

In this study, we employed a cosine similarity based data selection strategy to filter external data that is relevant to the target task. The original training set contained 3360 samples, and an additional 132 samples were selected from the external dataset, resulting in a combined training set of 3492 samples. The test set comprised 840 samples. This approach was chosen over alternative methods (e.g., influence-based functions) to ensure that only the most similar examples were used for fine-tuning, although due to time constraints, other strategies were not evaluated.

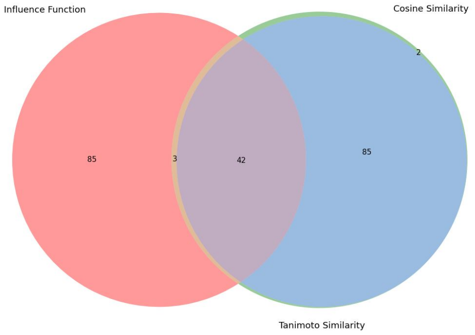

Figure 1: Venn diagram illustrating the overlap among data selected using three methods: influence selection (pink; top 100 samples), cosine similarity (blue; 127 samples), and Tanimoto similarity (green; 132 samples).

This Venn diagram 1 illustrates how three data selection methods—Influence Function (top 100 samples), Cosine Similarity, and Tanimoto Similarity, overlap in choosing relevant data. 42 samples are identified as relevant by all methods.

We also observed that the subset selected by Cosine similarity and Tanimoto similarity were overlapping with each other hence they also performed similar while fine-tuning, so we mainly focus on comparison between influence function and cosine similarity.

In Table 1, the supervised fine-tuned model outperforms all other methods, achieving the lowest test loss and highest R2 score, indicating effective

Method	Test Loss	MAE	R^2 Score
Supervised	0.4293	0.4932	0.6937
Unsupervised	0.7458	0.6707	0.4669
Influence	0.7221	0.6648	0.4804
Cosine	0.8163	0.8545	0.3818

Table 1: Performance comparison of full-finetuned models from all Tasks.

task-specific learning. Unsupervised fine-tuning (MLM) resulted in a noticeable performance drop, suggesting that it may not be ideal for regression tasks. Both influence-based and cosine similarity-based data selection methods improved upon MLM, with influence-based selection showing a more favorable performance. Overall, supervised fine-tuning remains the most effective approach for this regression task.

6.2 PEFT Comparison

To adapt a pre-trained Masked Language Model (MLM) for our regression task, we explored three parameter-efficient fine-tuning (PEFT) methods: BitFit, LoRA (Low-Rank Adaptation), and iA3 (Implicit Adapter). The models were evaluated using Test Loss, Mean Absolute Error (MAE), and R^2 Score. Table 2 summarizes the performance of each method for Cosine Similarity:

Method	Test Loss	MAE	R^2 Score
BitFit	2.0002	1.2043	-0.4110
LoRA	0.9668	0.7675	0.2974
iA3	1.0022	0.7900	0.2733

Table 2: Performance comparison of PEFT methods using cosine similarity based data selection.

Among the PEFT methods using cosine similarity, LoRA clearly outperformed the others by achieving the lowest test loss and the most favorable R^2 score. In contrast, BitFit demonstrated the worst performance, while iA3 achieved results very close to LoRA, with only marginal differences in test loss and R^2 score.

Method	Test Loss	MAE	R^2 Score
BitFit	1.7207	1.1068	-0.2193
LoRA	0.9468	0.7648	0.3127
iA3	0.9882	0.7674	0.2840

Table 3: Performance comparison of PEFT methods using influence based data selection.

Table 3 shows that LoRA and iA3 performed better than BitFit when using influence-based data selection, consistent with the trend observed in Table 2 for cosine similarity-based selection. LoRA remained slightly better than iA3, indicating its consistent ability to generalize well across different data selection strategies.

7 Discussion

This study examined data selection strategies and PEFT methods for predicting lipophilicity using a pre-trained chemical model. The influence function-based selection showed slightly better alignment with the task than cosine similarity, while Tanimoto similarity, though relevant, didn't outperform the others.

Among the PEFT methods, LoRA outperformed BitFit and iA3, especially with cosine similarity and influence-based selections. LoRA efficiently adapts large models for specific tasks, making it ideal for molecular property predictions, where data is limited.

Despite these advantages, fine-tuning large models with limited resources remains challenging, and there are trade-offs in performance compared to full model fine-tuning. Future research should focus on optimizing PEFT hyperparameters to better balance efficiency and accuracy.

8 Conclusion and Future Work

This study shows that combining data selection strategies with PEFT methods can enhance pre-trained models for predicting molecular properties like lipophilicity. Influence functions provided the best task alignment, and LoRA proved most efficient in boosting predictive accuracy.

Future work will focus on refining data selection with diverse molecular data, improving PEFT adaptability across tasks, and exploring techniques to enhance model generalization. Expanding the dataset and testing larger molecular representations may further improve real-world performance.

References

Naman Agarwal, Brian Bullins, and Elad Hazan. 2016. Second-order stochastic optimization in linear time. In *Conference on Machine Learning (COLT)*.

Lior Ben-Zaken, Alon Recanati, Shai Oren, and Shai Shalev-Shwartz. 2021. Bitfit: Simple parameter-efficient fine-tuning for transfer learning. *arXiv preprint arXiv:2106.10165*.

Edward Hu, Yiming Shen, Zekun Lee, Denny Wang, Xian Zhang, Le Song, Karan Raj, and Luke Zettlemoyer. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Pang Wei Koh and Percy Liang. 2020. Understanding black-box predictions via influence functions. *arXiv preprint arXiv:1703.04730*, v3.

Greg Landrum. 2013. Rdkit: Open-source cheminformatics. `https://www.rdkit.org/`. Accessed: 2025-03-10.

Yichao Liu, Yifan Zhang, Zhiqiang Yang, Zhiwei Li, Yang Yang, Qiao Yu, and Chia-Hsiu Zhang. 2022. ia3: Implicit adapter for efficient fine-tuning. *arXiv preprint arXiv:2202.06388*.

R. R. Rogers and J. H. Hahn. 2010. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754.

Gerard Salton. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill.

T. T. Tanimoto. 1963. An n-dimensional vector space for chemical structure comparison. *Journal of Chemical Information and Computer Sciences*, 3(3):232–235.

Ramsundar B. Feinberg E.N. et al. Wu, Z. 2018. Moleculenet: A benchmark for molecular machine learning. *Chemical Science*.