Scilight

Translational Insights
https://www.sciltp.com/journals/ti

Article

Check for updates

Efficient Adaptation of Chemical Language Models for Molecular Property Prediction

Kavya Agrawal [1,†], Harsha Harod [1,†], Ines Simeone [2], Michele Ceccarelli [3], Halima Bensmail [4], Sukrit Gupta [1] and Raghvendra Mall [4,5,*]

[1] Center for Applied Research in Data Science, Indian Institute of Technology Ropar, Punjab 140001, India
[2] Humanitas Research, 20089 Rozzano, Italy
[3] Sylvester Comprehensive Cancer Center, Miller School of Medicine, University of Miami, Coral Gables, FL 33136, USA
[4] Qatar Center for Artificial Intelligence, Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha P.O. Box 5825, Qatar
[5] Biotechnology Research Center, Technology Innovation Institute, Abu Dhabi P.O. Box 9639, United Arab Emirates
* Correspondence: ramall@hbku.edu.qa or raghvendramall@ieee.org
† These authors contributed equally to this work.

How To Cite: Agrawal, K.; Harod, H.; Simeone, I.; et al. Efficient Adaptation of Chemical Language Models for Molecular Property Prediction. *Translational Insights* **2026**, *1*(1), 13. https://doi.org/10.53941/ti.2026.100013

Received: 5 May 2026
Revised: 20 June 2026
Accepted: 30 June 2026
Published: 14 July 2026

Abstract: Motivation: Chemical language models (CLMs) like ChemBERTa and Molformer enable compound property prediction, but their computational demands limit adoption in resource-constrained settings. We integrate Low-Rank Adapters (LoRA) with CLMs to significantly reduce trainable parameters required for finetuning. Results: We observe 3–5% area under the receiver operating curve (AUC) improvement across classification tasks for molecule toxicity, blood-brain barrier permeability, and flavor prediction over Molformer-XL. Our approach achieves Matthews correlation coefficient (MCC) scores of 0.80–0.90 across three tasks while reducing model parameters by 75–95%. By comparing embeddings from zero-shot and finetuned CLMs combined with molecular physicochemical properties, we attain optimal performance across four datasets. This lightweight adaptation retains performance efficiency while reducing over-parameterization.

Keywords: parameter-efficient finetuning; chemical language models; low-rank adaptation; molecular properties; computational drug discovery

1. Introduction

In recent decades, the study of molecular structures and their properties has been central to advances in drug discovery. Although traditional quantitative structure-activity relationship (QSAR) methods rely on handcrafted descriptors, deep learning techniques, particularly transformer-based architectures [1], enable end-to-end learning of molecular representations from its Simplified Molecular Input Line Entry System (SMILES) sequence. This has catalyzed progress in this field by developing chemical language models (CLM), pretrained on massive molecular corpora, now rivaling experimental approaches in predicting properties such as toxicity, solubility, and bio-activity [2,3]. However, two critical challenges persist: (1) what the added benefits are of learned representations over hand-crafted physico-chemical features for molecular property prediction and (2) the computational cost of fine-tuning $O(10^8)$-parameter CLMs for specialized tasks.

This work addresses these gaps through a systematic study of two state-of-the-art (SOTA) CLMs, Molformer [4] and ChemBERTa-2 model [5], to enhance molecular property prediction tasks. Both models utilize SMILES representation to learn molecular embeddings. Molformer's linear attention mechanism, pretrained on 1.1 billion molecules, captures long-range dependencies in SMILES sequences, achieving strong results on classification and regression tasks, especially for MoleculeNet datasets. In contrast, ChemBERTa-2 is a Bidirectional Encoder Representations from Transformers (BERT)-style transformer that learns molecular fingerprints through semi-supervised

pretraining of the language model that employs two different approaches, i.e., Masked Language Model (MLM) and Multi-Task Regression (MTR) and is pretrained on 77 million SMILES representation of drugs.

Our downstream tasks encompass medicinal chemistry applications, using the ClinTox dataset for toxicity, the blood-brain barrier prediction (BBBP) dataset for permeability prediction, molecular taste prediction using the Flavor Analysis and Recognition Transformer (FART) dataset which involves classifying compounds into different flavor categories [6] and the identification of the Beta-Secretase 1 (BACE) inhibitor. To improve performance and efficiency, we adopt Low Rank Adaptation (LoRA) parameter-efficient finetuning technique, which substantially reduces the number of trainable parameters compared to full finetuning. This is particularly advantageous for large-scale language models, where training from scratch is computationally prohibitive.

The LoRA technique on Molformer and ChemBERTa-2 helps preserve pretrained knowledge while specializing for downstream tasks. In addition, we perform comprehensive ablation to show how the combination of finetuned model embeddings with molecular descriptors, fingerprints and graphical features positively influenced predictive accuracy. By unifying efficiency and accuracy, this work establishes a blueprint for deploying large CLMs in resource-constrained environments while maintaining scientific rigor, a prerequisite for industrial adoption. Our key contributions are:

- Introduce EffiChem, a two-stage hybrid framework that *decouples* feature extraction from prediction: LoRA-adapted CLMs generate task-specific embeddings, which are fused with physicochemical descriptors and graph-based features for tree-based ensemble prediction. Unlike adapter-only CLM methods, graph-based models (GROVER [7], GraphMVP [8]), or 3D-aware models (Uni-Mol [9]), EffiChem operates from SMILES alone with no graph construction or conformer generation.
- Employ LoRA for parameter-efficient finetuning of pretrained CLMs, reducing computational cost, and demonstrate that end-to-end LoRA-adapted CLMs perform best across multiple tasks.
- Comprehensive ablation combining task-specific embeddings with molecular descriptors, graphical features and fingerprints, demonstrating whether hand-crafted features enhance model predictive capability or not.
- Achieve consistent improvements across four molecular property prediction tasks for two CLM architectures, Molformer and ChemBERTa, surpassing SOTA baselines.
- We have made all our code available at https://github.com/kavyagl2/EffiChem-2.0 for reproducibility and allowing the community to utilize the models for their tasks.

Figure 1 highlights the workflow of EffiChem architecture.

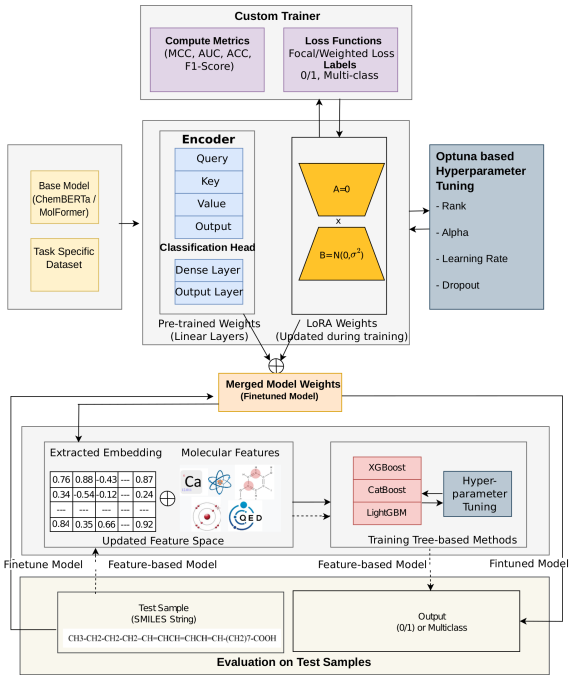

Figure 1. EffiChem end-to-end pipeline (system level). A SMILES string enters Stage 1, where a molecular transformer (MolFormer/ChemBERTa) is efficiently adapted via LoRA using focal or weighted loss with Bayesian hyperparameter tuning (Optuna [10]). Stage 2 extracts the task-adapted embedding from the classification head and concatenates it with hand-crafted molecular features (physicochemical descriptors, graph features, fingerprints) to form an enriched feature vector, which is passed to an ensemble classifier (XGBoost/CatBoost/LightGBM) for binary or multi-class prediction. The internal LoRA weight-update mechanism is detailed separately in Figure 2.

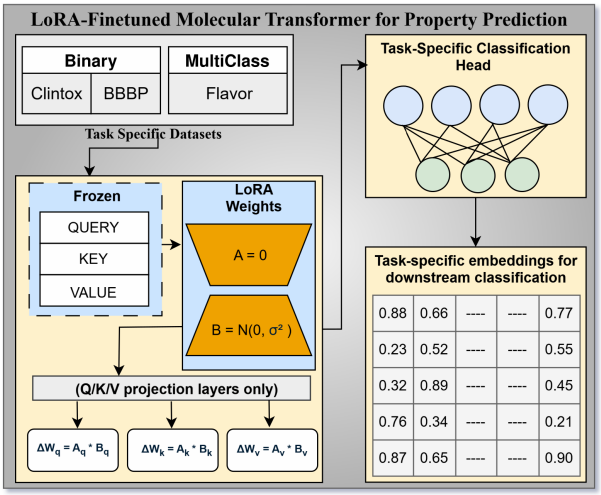

Figure 2. LoRA weight-update mechanism (component level). Inside each transformer layer, the pretrained weight matrix W is frozen; only two low-rank adapter matrices A (down-projection) and B (up-projection) are trained, injected into the Q/K/V projections as W' = W + BA. A lightweight task-specific classification head completes the adaptation for binary or multi-class prediction on ClinTox, BBBP, BACE, and Flavor. How this adapted encoder fits into the full two-stage EffiChem pipeline (embedding extraction, feature fusion, ensemble prediction) is shown in Figure 1.

2. Materials and Methods

2.1. Data Collection and Curation

- The Blood-Brain Barrier Penetration (BBBP) dataset was originally curated in [11] for predicting barrier permeability and contains over 2000 compounds with binary labels, originally integrated into the MoleculeNet [12] database. The blood–brain barrier, formed by brain capillary endothelial cells, blocks all large-molecule drugs and over 98% of small-molecule drugs from entering the brain. This makes delivering treatments to targeted brain regions a major challenge in treating brain disorders.
- ClinTox is part of the MoleculeNet benchmark [12]. The ClinTox dataset contains drugs that either failed clinical trials due to toxicity or succeeded, along with their FDA approval status for 1476 compounds. FDA-approved drugs were derived from the SWEETLEAD database [13], while toxicity-failed candidates originate from the Aggregate Analysis of Clinical Trials (AACT).
- The FART dataset, introduced in [14], is a multi-class taste prediction dataset consisting of 15,025 compounds annotated as 'bitter', 'sour', 'sweet', 'umami' and 'undefined' classes. The original curation applied RDKit SMILES canonicalization, exact-duplicate removal, stereoisomer deduplication, and salt-form desalting prior to releasing the dataset and its splits. This dataset was made available through their GitHub repository at https://github.com/fart-lab/fart (accessed on 2 February, 2026) .
- The BACE dataset is part of the MoleculeNet benchmark [12] and was originally curated for evaluating structure-based virtual screening. It contains experimentally measured binding results for inhibitors of human β-secretase 1 (BACE-1). The dataset comprises 1513 compounds with binary activity labels indicating active or inactive inhibition, along with regression targets for binding affinity. The dataset splits used in this work were curated using the DeepChem reference implementation.

The BBBP and ClinTox datasets used in our experiments are sourced from the IBM MolFormer repository. For all datasets, checks have been performed for validity of drugs, removal of duplicate and class-ambiguous molecules. These curation steps address the concerns usually associated with the MoleculeNet dataset in https://practicalcheminformatics.blogspot.com/2023/08/we-need-better-benchmarks-for-machine.html (accessed on 4 April 2026) .

2.2. Data Splits and Class Distribution

The BBBP, ClinTox, and BACE datasets employ scaffold-based splits, based on molecular substructures to enforce chemical dissimilarity between subsets. This provides a more rigorous and realistic evaluation of the model's ability to generalize to unseen chemical series, compared to random splitting. The datasets are split in a ratio of 80:10:10 for train, validation, and test sets, respectively, as suggested in DeepChem [15]. The Flavor dataset is sourced from [14], which provides pre-split data whose curation explicitly addressed SMILES canonicalization, duplicate removal, stereoisomer deduplication, and salt-form stripping; structural similarity between train and test

splits is therefore controlled at the source.

- BBBP exhibits moderate imbalance between target classes, with 78.8% permeable compounds in the train set. However, test and validation splits show distribution shifts (63.4% and 70.7% permeable), indicating scaffold-driven variability.
- ClinTox faces severe imbalance across both tasks, for 'CT_TOX', non-toxic compounds dominate, 92.3% in training, 90.9% in test and 94.5% for validation. For 'FDA_APPROVED' dataset, approved drugs represent 93.2% of training data, rising to 95.1% and 95.9% in test and validation respectively.
- Flavor dataset displays pronounced multi-class skew, with sweet (41%) and bitter (32%) categories vastly outnumbering sour (11%), umami (4%), and undefined (12%) classes. This imbalance complicates recognition of under-represented flavors like 'umami'. We discuss how to handle class-imbalance for each task in the Supplementary.
- BACE contains 1513 compounds with binary activity labels, comprising 54.3% inactive (822) and 45.7% active (691) molecules prior to splitting. No samples were removed during pre-processing. Following the scaffold-based split [15], the dataset was divided in an $80{:}10{:}10$ ratio into train (1210), validation (151), and test (152) sets.

2.3. Pretrained Chemical Language Models

ChemBERTa-2: ChemBERTa-2 [2] builds upon ChemBERTa [5] by adopting a Robustly Optimized BERT Pretraining Approach (RoBERTa)-style architecture trained on 77 million canonical SMILES from PubChem - one of the largest datasets available for CLMs [16]. It uses two self-supervised learning strategies: Masked Language Modeling (MLM) and Multi-Task Regression (MTR). MLM masks 15% of input tokens, while MTR trains on 200 molecular property prediction tasks in parallel, promoting semantically rich, interpretable representations. In our experiment, we used the two largest ChemBERTa-2 variants (77M) with difference in architecture (MLM and MTR). These models are henceforth referred to as ChemBERTa-77M MTR and ChemBERTa-77M MLM.

On MoleculeNet datasets (e.g., BACE, HIV, Tox21) [12], ChemBERTa-2 outperformed strong baselines like D-MPNN [17] in most tasks after finetuning. For the BBBP and ClinTox datasets, MTR performed better than MLM (see [2] for results).

MolFormer: MolFormer is a scalable CLM that learns molecular representations from SMILES without relying on 3D geometry [18]. It uses a transformer with linear attention (FAVOR+) [6], rotatory positional embeddings [19], and efficient training strategies (e.g., adaptive bucketing, distributed training) [20], enabling fast processing of large datasets like PubChem [16] and ZINC (>1.1 billion molecules) [21].

Pretraining on massive datasets and finetuning led to state-of-the-art performance, notably on the QM9 benchmark [22]. Despite its strengths, MolFormer lacks geometry-aware inductive biases and is sensitive to optimizer settings. Nonetheless, its scalability and efficiency make it a strong foundation model for molecular property prediction. We used the Molformer-XL-both-10pct [4] in our experiments. It was pretrained on 10% of both datasets with over 50 million parameters and performs similar to Molformer-XL (>1 billion molecules). The model can be accessed from https://huggingface.co/ibm-research/MoLFormer-XL-both-10pct (accessed on 20 January 2026) Huggingface repository.

2.4. Zero-Shot vs. Full vs. Parameter-Efficient Finetuning

Vanilla full finetuning of a CLM updates all parameters of the pretrained language model using task-specific data. On the other hand, zero-shot learning uses prompts on embeddings from the pretrained model to perform new tasks without changing their core parameters. Zero-shot models are cheap and have low latency, but are less accurate for specific tasks [23]. Full finetuning is computationally expensive and resource intensive as it creates models (with revised weights for millions of parameters) that can generalize for a specific task but cannot be easily re-used [24,25]. More details about zero-shot and full finetuning are provided in Supplementary Section S1.

To mitigate issues with pretrained CLMs, we utilize parameter-efficient finetuning (PEFT) techniques. PEFT techniques provide a balance between accuracy and efficiency when using CLMs [26]. The core principle is to achieve comparable or better performance than regular finetuning with significantly less ($\ll$) parameters. This is done by selectively updating the model parameters and making targeted coefficient adjustments. To formally explain PEFT, we can consider the modification of the weight matrices associated with the attention mechanism. In a transformer model [1], the attention function includes the following transformations for an input $\boldsymbol{X}$

$$\boldsymbol{Q} = \boldsymbol{X}\boldsymbol{W}_{\boldsymbol{Q}}, \quad \boldsymbol{K} = \boldsymbol{X}\boldsymbol{W}_{\boldsymbol{K}}, \quad \boldsymbol{V} = \boldsymbol{X}\boldsymbol{W}_{\boldsymbol{V}}$$

where, $\boldsymbol{W}_{\boldsymbol{Q}}$, $\boldsymbol{W}_{\boldsymbol{K}}$, and $\boldsymbol{W}_{\boldsymbol{V}}$ are the weight matrices for the query, key, and value projections respectively. For instance, for vanilla ChemBERTa-2 or MolFormer, all elements of these matrices were updated. However, in PEFT,

the weight matrix $\boldsymbol{W}$ (generic representation for query, key and value weight matrices) is modified as follows:

$$\boldsymbol{W}' = \boldsymbol{W} + \Delta\boldsymbol{W}$$

Here $\Delta\boldsymbol{W}$ represents the parameter-efficient updates, which can be low-rank matrices or other sparse updates. The process of implementing PEFT is often an iterative process with the aim of selecting the appropriate parameters and layers to freeze [26] i.e., these parameters are unchanged during the adapter model training procedure.

There is a growing interest in designing new techniques for PEFT [27], where each method comes with new parameters and adapters represented by $\Delta\boldsymbol{W}$ to finetune pretrained language models. In our work, we apply re-parameterized finetuning techniques [28] to adjust our model parameters. Re-parameterized finetuning introduces adapter weights next to pretrained weights to refine the model as depicted in Figure 2. We selected this method based on comparative analyses of PEFT methods [29], which indicated that Low-Rank Adapters (LoRA) and their derivatives achieve the highest parameter reduction with superior performance for natural language processing (NLP) tasks. In the Supplementary Section S2, we explain how we utilize Low Rank Adapters (LoRA) [30] for finetuning on a classification task.

3. EffiChem Model Architecture

The EffiChem architecture was designed for three classification tasks but can extend to other classification or regression tasks. This model architecture is designed to leverage the strengths of CLMs and chemoinformatic features for molecular property prediction. EffiChem operates in three ways:

- End-to-End: The CLMs are finetuned using LoRA technique in an end-to-end framework for each downstream classification task.
- Representation Learning and Prediction: The CLMs are finetuned for separate molecular classification tasks. These models trained on SMILES representations of molecules, learn rich, task-specific molecular embeddings that capture both syntactic and semantic features of chemical structures. The classification head of the LoRA finetuned model provides appropriate task-specific embedding representation.
- Feature Fusion and Prediction: The learned embeddings from the CLMs are concatenated with molecular descriptors, graphical features and fingerprints. This combined feature vector is then used to train tree-based classifiers for each of the four classification tasks.

The architecture of EffiChem remains consistent across all experiments. This generalization allows EffiChem to serve as a robust and adaptable framework for a wide range of molecule property prediction challenges. The EffiChem architecture is depicted in Figure 2. A detailed description of EffiChem trainer and handling of class imbalance in the datasets is provided in Supplementary Section S3.

3.1. Embeddings + Molecular Feature Integration

EffiChem combines learned molecular embeddings from transformer models with molecular descriptors, graph-based features and fingerprints to create combined feature representation. These representations capture both deep chemical patterns from SMILES and expert-crafted properties, enabling robust molecular prediction using tree-based models like XGBoost [31], LightGBM [32], and CatBoost [33]. We consider the following embedding representations from fine-tuned CLMs in EffiChem model:

- Base Models: The embedding is taken directly from the '[CLS]' token of the last hidden state of the fine-tuned CLM, a standard approach for capturing global molecular features in transformer architectures [34].
- Finetuned MolFormer Model: The '[CLS]' token from the last hidden state is passed through model's classifier head, which includes two dense layers and a non-linear activation (e.g., Gaussian Error Linear Unit (GELU)), yielding a task-adapted representation.
- Finetuned ChemBERTa Models: Similarly, the '[CLS]' token from the last hidden state is processed by a dense layer followed by a 'tanh' activation function, producing a condensed, non-linear embedding for downstream tasks.

For each molecule, the resulting embedding size was 384-dimensional vector (ChemBERTa) and 768-dimensional vector (Molformer). During feature fusion, we combine the learned embeddings with 148 RDKit-based descriptors, 30 graph-based (using 'networkx' v3.5), 128-Morgan and 167-MACCS fingerprints from RDKit v2025.09.1. This creates an enhanced feature space that leverages both deep learning representations and domain knowledge. This hybrid approach has been shown to improve prediction accuracy by capturing complementary aspects of molecular structure and properties [35, 36]. Supplementary Table S1 provides details about the domain-specific features used in EffiChem architecture.

3.2. Tree-Based Ensemble Classifiers

The final prediction phase employs three SOTA tree-based ensemble methods: XGBoost, CatBoost and LightGBM. These algorithms are particularly well-suited for handling the combined feature vectors due to their ability to capture complex interactions and handle mixed data types effectively.

- Ensemble Method Advantages: Tree-based models offer several advantages for molecular property prediction including built-in feature selection, robustness to outliers, and interpretability. Each algorithm brings unique strengths: XGBoost excels in gradient boosting optimization, CatBoost handles categorical features efficiently, and LightGBM provides fast training with leaf-wise tree growth.
- Performance Optimization: Bayesian optimization library 'Optuna' v4.5.0 was used for hyperparameter optimization [10], ensuring optimal performance for the specific classification tasks. The ensemble approach allows for comparison and potentially combining predictions from multiple models to achieve superior accuracy. All models were evaluated on the same set of evaluation metrics.

3.3. Evaluation Metrics

EffiChem's performance with other SOTA methods and different finetuning approaches was evaluated using quality metrics like accuracy (ACC), AUC, and MCC. Other metrics used to assess the performance in binary classification problems were precision and recall. These metrics are frequently employed for model comparison [35, 37–40]. Additionally, we utilized F1-micro and F1-macro commonly used for multi-class classification problems.

4. Experimental Results

We conducted a comprehensive ablation study across multiple datasets and task settings. Specifically, we evaluate (a) binary classification tasks on BBBP (Supplementary Tables S2 and S3), ClinTox (Supplementary Tables S6 and S7), and BACE (Supplementary Tables S8 and S9); and (b) a multi-class classification task on the Flavor (FART) dataset (Supplementary Tables S4 and S5). The ablation compares models based on physico-chemical (PC) features, zero-shot embeddings from chemical language models (Base), their combination (PC + Base), end-to-end parameter-efficient fine-tuning using LoRA (E2E LoRA), embeddings extracted from LoRA-fine-tuned models (FT Embed), and FT Embed combined with PC features using tree-based machine learning models. Detailed quantitative results and associated insights from these comparisons are provided in the Supplementary Material.

In Tables 1 and 2, we report the results for the best-performing model configurations. This is defined by input modality, CLM, and loss function (where applicable), combined with the optimal tree-based classifier for the BBBP and Flavor datasets, respectively. For the BACE dataset, Figure 3 presents the AUC and area under the precision-recall curve (AUPR) curves corresponding to the best model configuration. Next, we summarize the task-specific insights derived from these comprehensive comparative evaluations.

4.1. BBBP Dataset Results

The EffiChem framework achieved optimal performance using E2E LoRA finetuning of MolFormer-XL-10pct-both with weighted loss, attaining an MCC of 0.797 and F1 micro of 0.906, representing a 3-5% improvement over baseline methods (Table 1). Among feature-based approaches, the combination of PC features with finetuned MolFormer embeddings (PC + FT Embed) yielded the highest AUC of 0.951 with the second highest MCC of 0.773 using the CatBoost classifier. In particular, the E2E LoRA approach achieved better recall (0.896) compared to all other configurations (minimum improvement of 2.3% over next best recall of 0.873), indicating improved identification of permeable compounds for the blood-brain barrier. For the BBBP dataset, the original MolFormer-XL [4] achieved an AUC of 0.937, MolFormer-XL-10pct-both achieved an AUC of 0.915 and MolFormer-XL-10pct-both fully finetuned attained an AUC of 0.921. Both our E2E LoRA model and PC + FT Embed models (using MolFormer-XL-10pct-both with weighted loss) outperform the state-of-the-art by at least 1.2–3%.

4.2. Flavor Dataset Results

For multi-class flavor prediction, E2E LoRA finetuning of MolFormer-XL-10pct-both with weighted loss achieved the highest performance with MCC of 0.801, F1-macro of 0.809, and AUC of 0.975 (see Table 2). The model demonstrated robust performance across all flavor categories despite severe class imbalance, with strong precision-recall balance (0.802 and 0.817 respectively). Feature fusion approaches showed competitive performance, with PC + FT Embed MolFormer-XL-10pct-both achieving F1-macro of 0.798 and precision of 0.852 using weighted loss. The original FART paper's [14] best reported accuracy, precision, recall and F1-score was 0.884, 0.764, 0.785 and 0.771 respectively. Compared to that, our E2E LoRA on MolFormer-XL-10pct-both with

weighted loss achieves scores of 0.892, 0.802, 0.817, 0.809, respectively achieving $\approx 4\%$ improvement on key quality metrics such as precision, recall, and F1-macro. Furthermore, it indicates that EffiChem can better handle class imbalance including minority class like umami in multi-class prediction problems.

Table 1. Best performing configurations for each modality and CLM combination on the BBBP dataset. For tree-based models, the best performing ML model (XGBoost, LightGBM, or CatBoost) is selected based on average performance across all metrics. For end-to-end (E2E) LoRA models, the best loss function (Focal or Weighted) is selected similarly. Full set of comparisons available in Supplementary Tables S2 and S3. The best metrics are highlighted in bold and *. Here '+' represents the second-best model and '−' represents the third-best model overall.

Modality	CLM	Loss Function	ML Model	AUC	Acc	F1 mac	F1 mic	MCC	Prec	Rec
PC	-	-	XGBoost	0.938	0.875	0.858	0.875	0.728	0.886	0.844
Base	ChemBERTa-77M MLM	-	XGBoost	0.889	0.849	0.818	0.849	0.683	0.894	0.796
Base	ChemBERTa-77M MTR	-	CatBoost	0.914	0.849	0.829	0.849	0.668	0.852	0.817
Base	MolFormer-XL-10pct-both	-	CatBoost	0.915	0.891	0.876	0.891	0.765	0.907	0.859
PC + Base	ChemBERTa-77M MLM	-	LightGBM	0.933	0.880	0.865	0.880	0.739	0.890	0.851
PC + Base	ChemBERTa-77M MTR	-	CatBoost	0.929	0.885	0.869	0.885	0.754	0.903	0.852
PC + Base	MolFormer-XL-10pct-both	-	XGBoost	0.940	0.870	0.851	0.870	0.718	0.887	0.834
E2E LoRA	ChemBERTa-77M MLM	Focal	-	0.923	0.880	0.851	0.865	0.739	0.889	0.851
E2E LoRA	ChemBERTa-77M MTR	Focal	-	0.921	0.870	0.843	0.854	0.715	0.873	0.843
E2E LoRA *	MolFormer-XL-10pct-both	Weighted	-	0.948	**0.906**	**0.898**	**0.906**	**0.797**	0.901	**0.896**
FT Embed	ChemBERTa-77M MLM	Focal	LightGBM	0.905	0.896	0.882	0.896	0.776	0.911	0.866
FT Embed	ChemBERTa-77M MLM	Weighted	CatBoost	0.927	0.870	0.853	0.870	0.716	0.877	0.840
FT Embed	ChemBERTa-77M MTR	Focal	LightGBM	0.902	0.865	0.846	0.865	0.704	0.873	0.833
FT Embed	ChemBERTa-77M MTR	Weighted	CatBoost	0.911	0.865	0.849	0.865	0.703	0.865	0.839
FT Embed −	MolFormer-XL-10pct-both	Focal	XGBoost	0.935	0.901	0.888	0.901	0.787	0.914	0.873
FT Embed	MolFormer-XL-10pct-both	Weighted	CatBoost	0.949	0.880	0.864	0.880	0.741	0.894	0.848
PC + FT Embed	ChemBERTa-77M MLM	Focal	LightGBM	0.888	0.891	0.875	0.891	0.765	0.907	0.859
PC + FT Embed	ChemBERTa-77M MLM	Weighted	CatBoost	0.928	0.870	0.850	0.870	0.718	0.887	0.834
PC + FT Embed	ChemBERTa-77M MTR	Focal	CatBoost	0.928	0.854	0.838	0.854	0.680	0.850	0.830
PC + FT Embed	ChemBERTa-77M MTR	Weighted	CatBoost	0.914	0.859	0.844	0.859	0.692	0.858	0.835
PC + FT Embed +	MolFormer-XL-10pct-both	Focal	LightGBM	0.934	0.901	0.887	0.901	0.789	**0.920**	0.870
PC + FT Embed	MolFormer-XL-10pct-both	Weighted	CatBoost	**0.951**	0.896	0.884	0.896	0.773	0.902	0.872

Table 2. Best performing configurations for each modality and CLM combination on the Flavor (FART) dataset. For tree-based models, the best performing ML model (XGBoost, LightGBM, or CatBoost) is selected based on average performance across all metrics. For E2E LoRA models, the best loss function (Focal or Weighted) is selected similarly. The best results are highlighted in bold.

Modality	CLM	Loss Function	ML Model	AUC	Acc	F1 mac	F1 mic	MCC	Prec	Rec
PC	-	-	XGBoost	0.974	0.888	0.751	0.888	0.803	0.721	0.793
Base	ChemBERTa-77M MLM	-	XGBoost	0.916	0.862	0.634	0.862	0.745	0.626	0.643
Base	ChemBERTa-77M MTR	-	LightGBM	0.973	0.884	0.767	0.884	0.795	0.749	0.790
Base	MolFormer-XL-10pct-both	-	XGBoost	0.960	0.863	0.743	0.863	0.761	0.822	0.737
PC + Base	ChemBERTa-77M MLM	-	XGBoost	0.958	0.882	0.751	0.882	0.793	0.776	0.753
PC + Base	ChemBERTa-77M MTR	-	CatBoost	0.965	0.885	0.801	0.885	0.798	0.782	**0.824**
PC + Base	MolFormer-XL-10pct-both	-	LightGBM	0.966	0.879	0.741	0.879	0.786	0.714	0.776
E2E LoRA	ChemBERTa-77M MLM	Weighted	-	**0.975**	0.882	0.734	0.882	0.785	0.732	0.738
E2E LoRA	ChemBERTa-77M MTR	Weighted	-	0.964	0.890	0.780	0.890	**0.802**	0.780	0.786
E2E LoRA	MolFormer-XL-10pct-both	Weighted	-	**0.975**	**0.892**	**0.809**	**0.892**	0.801	0.802	0.817
FT Embed	ChemBERTa-77M MLM	Focal	XGBoost	0.962	0.872	0.731	0.872	0.773	0.733	0.741
FT Embed	ChemBERTa-77M MLM	Weighted	LightGBM	0.971	0.877	0.739	0.877	0.783	0.709	0.781
FT Embed	ChemBERTa-77M MTR	Focal	XGBoost	0.953	0.881	0.738	0.881	0.793	0.724	0.759
FT Embed	ChemBERTa-77M MTR	Weighted	LightGBM	0.964	0.879	0.742	0.879	0.789	0.711	0.786
FT Embed	MolFormer-XL-10pct-both	Focal	LightGBM	0.950	0.879	0.766	0.879	0.779	0.770	0.766
FT Embed	MolFormer-XL-10pct-both	Weighted	LightGBM	0.970	0.882	0.783	0.882	0.787	0.800	0.778
PC + FT Embed	ChemBERTa-77M MLM	Focal	CatBoost	0.961	0.872	0.723	0.872	0.771	0.715	0.736
PC + FT Embed	ChemBERTa-77M MLM	Weighted	CatBoost	0.971	0.879	0.736	0.879	0.786	0.707	0.780
PC + FT Embed	ChemBERTa-77M MTR	Focal	CatBoost	0.965	0.878	0.727	0.878	0.785	0.712	0.747
PC + FT Embed	ChemBERTa-77M MTR	Weighted	CatBoost	0.964	0.885	0.754	0.885	0.798	0.728	0.788
PC + FT Embed	MolFormer-XL-10pct-both	Focal	LightGBM	0.961	0.881	0.781	0.881	0.783	0.800	0.774
PC + FT Embed	MolFormer-XL-10pct-both	Weighted	LightGBM	0.967	0.883	0.798	0.883	0.789	**0.852**	0.779

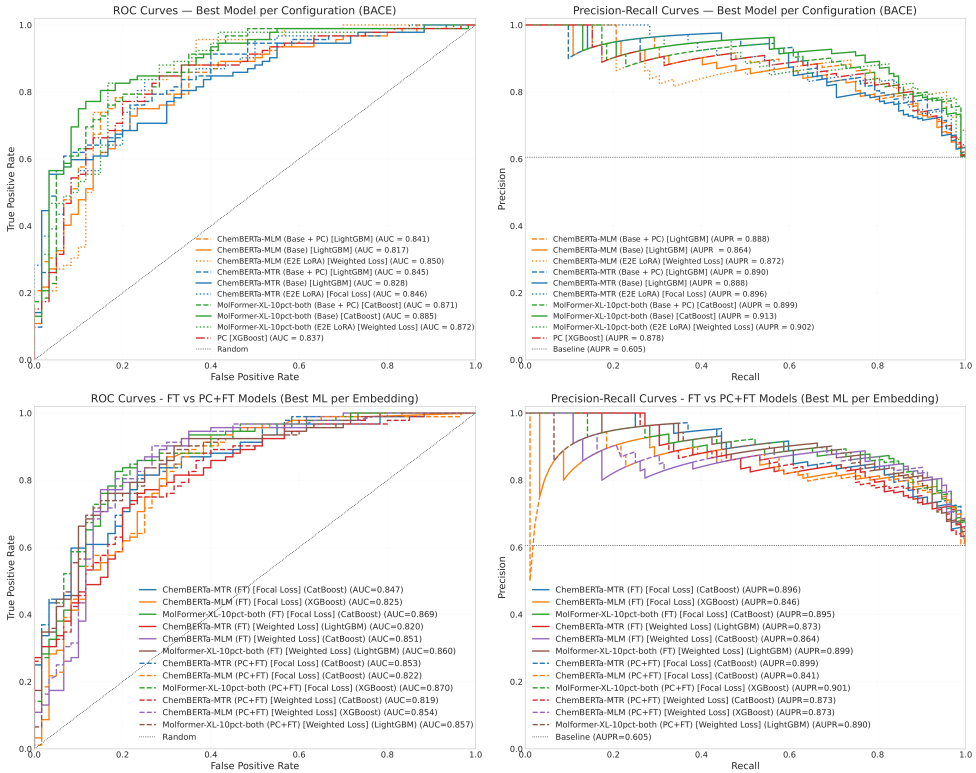

Figure 3. ROC and Precision-Recall curves for the BACE dataset. Panels A-B show the best model per configuration, with MolFormer-XL-10pct-both (Base) achieving the highest AUC (0.885) and AUPR (0.913). Panels C-D compare finetuned (FT) vs. combined physicochemical and finetuned (PC + FT) embeddings, where MolFormer-XL-10pct-both consistently outperforms baseline models across both metrics.

4.3. ClinTox Dataset Results

ClinTox presents two severely imbalanced subtasks: clinical toxicity prediction (CT_TOX, $>92\%$ non-toxic in training) and FDA approval prediction (FDA_APPROVED, $>93\%$ approved). Table 3 reports the best-performing configuration per modality for CT_TOX. Unlike BBBP, Flavor, and BACE, the best results here come from *zero-shot* MolFormer embeddings combined with physicochemical features (PC + Base, LightGBM), achieving AUC $= 0.999$, MCC $= 0.875$, and F1-macro $= 0.934$—outperforming all LoRA-finetuned approaches. This indicates that pretrained MolFormer already captures sufficient toxicity-relevant molecular signals, and LoRA adaptation provides limited gain under extreme class imbalance ($>92\%$ majority class).

A comprehensive ablation for ClinTox is provided in Supplementary Tables S6 and S7; key results are summarised in Table 3 below.

Table 3. Best-performing configurations per modality on the ClinTox CT_TOX task. Best CLM selected per modality by average across all metrics. Full ablation and FDA_APPROVED results in Supplementary Tables S6 and S7. The best results are highlighted in bold and ⋆.

Modality	CLM	Loss Function	ML Model	AUC	Acc	F1 mac	F1 mic	MCC	Prec	Rec
PC	–	–	LightGBM	0.975	0.972	0.850	0.972	0.700	0.850	0.850
Base	MolFormer-XL-10pct-both	–	LightGBM	0.997	**0.986**	0.925	**0.986**	0.850	0.925	0.925
PC + Base ⋆	MolFormer-XL-10pct-both	–	LightGBM	**0.999**	**0.986**	**0.934**	**0.986**	**0.875**	0.889	**0.993**
E2E LoRA	MolFormer-XL-10pct-both	Focal	–	0.993	0.972	0.881	0.972	0.786	0.818	0.985
FT Embed	MolFormer-XL-10pct-both	Focal	XGBoost	0.987	0.979	0.906	0.979	0.827	0.850	0.989
PC + FT Embed	ChemBERTa-77M MTR	Focal	CatBoost	0.995	0.972	0.881	0.972	0.786	0.818	0.985

4.4. BACE Dataset Results

On the BACE inhibitor prediction task, end-to-end approaches and feature-based methods showed comparable performance across all metrics (see Figure 3). MolFormer-based models consistently achieved AUC values above 0.86, with the best configuration reaching 0.885 using base embeddings with CatBoost. The precision-recall curves (Figures 3) demonstrate strong discriminative ability with AUPR curves exceeding 0.89 for most configurations.

Notably, adding physicochemical features to finetuned embeddings (PC + FT) showed minimal improvement over finetuned embeddings alone, suggesting that LoRA adaptation effectively captures task-relevant molecular properties. Even in the presence of a relatively balanced class distribution (54.3% inactive, 45.7% active), the scaffold-based split results in only 90 inactive and 62 active compounds in the test set.

From Figure 3 and Supplementary Tables S8 and S9, we observed that MolFormer-based models substantially outperform ChemBERTa-based models across all configurations, with performance gaps more pronounced than in other tasks. Additionally, the focal loss outperforms weighted loss for MolFormer-based methods, suggesting that focusing on hard-to-classify examples is particularly beneficial for this challenging task (optimal MCC $= 0.577$ for BACE is much lower than that attained for BBBP, Flavor and Clintox datasets as illustrated via Supplementary Tables S2–S9).

The substantially lower MCC on BACE reflects several compounding factors. First, the small test set (152 compounds, 90 inactive and 62 active) means individual misclassifications carry disproportionate weight in MCC. Second, the scaffold split shifts the class balance between training (54.3% inactive) and test (59.2% inactive), introducing a distribution shift on top of the limited test size. Third, BACE-1 inhibition is intrinsically a structure-based recognition problem: binding affinity depends on 3D pocket complementarity, a signal that SMILES-only CLMs such as MolFormer and ChemBERTa cannot fully encode without geometry-aware inductive biases. Fourth, the active/inactive boundary is defined by a continuous potency threshold, making the class boundary inherently less sharp than the binary endpoint tasks in BBBP and ClinTox. Together, these factors make BACE a fundamentally harder generalisation target for SMILES-based models regardless of the adaptation strategy used.

4.5. Parameter Efficiency

We highlight the parameter efficiency of the E2E LoRA models for each CLM across the four diverse tasks in Table 4. The trainable parameter percentage is computed as Trainable% $=$ (Trainable Params/Total Params) $\times$ 100, and %Reduction $= 100 -$ Trainable%. Here, Total Params is the full pretrained backbone; Trainable Params comprises the LoRA adapter matrices (low-rank $\mathbf{A}$, $\mathbf{B}$ injected into query/value projections) plus the task-specific classification head, which collectively contribute $<25\%$ across all configurations. The frozen backbone parameters constitute the remainder and are never updated during finetuning. We observe that both the ChemBERTa (MTR and MLM) models achieve a parameter reduction in the range of 76–94%. The larger MolFormer-XL-10pct-both ($\approx$ 46 million base parameters) attained a reduction of 84–96% across four tasks while obtaining optimal performance. Crucially, LoRA adapters are just a few MB in size (vs. tens of GB for a fully finetuned model) and add only milliseconds to inference time, translating the parameter savings directly into wall-clock and memory efficiency.

Table 4. E2E LoRA parameter efficiency per model and task. Trainable Params = LoRA adapter matrices + classification head; Trainable% = Trainable / Total $\times$ 100; % Reduction = 100 $-$ Trainable%. The frozen backbone constitutes the remaining parameters.

Dataset	Loss Type	Model	r	LoRA α	Dropout	Total Params (Base)	Trainable Params	Trainable %	% Reduction
BACE	Focal Loss	ChemBERTa-77M-MLM	4	32	0	3,428,210	205,826	6.0039	93.9961
BACE	Focal Loss	ChemBERTa-77M-MTR	32	128	0.2	3,428,210	606,338	17.6867	82.3133
BACE	Focal Loss	MoLFormer-XL	64	128	0.2	45,557,762	8,260,610	18.1322	81.8678
BACE	Weighted Loss	ChemBERTa-77M-MLM	16	128	0.1	3,428,210	377,474	11.0108	88.9892
BACE	Weighted Loss	ChemBERTa-77M-MTR	4	128	0	3,428,210	205,826	6.0039	93.9961
BACE	Weighted Loss	MoLFormer-XL	32	8	0	45,557,762	4,721,666	10.3641	89.6359
BBBP	Focal Loss	ChemBERTa-77M-MLM	64	32	0	3,428,210	1,064,066	31.0385	68.9615
BBBP	Focal Loss	ChemBERTa-77M-MTR	16	128	0.2	3,428,210	377,474	11.0108	88.9892
BBBP	Focal Loss	MoLFormer-XL	32	8	0	45,557,762	4,721,666	10.3641	89.6359
BBBP	Weighted Loss	ChemBERTa-77M-MLM	64	32	0	3,428,210	1,064,066	31.0385	68.9615
BBBP	Weighted Loss	ChemBERTa-77M-MTR	16	32	0.2	3,428,210	377,474	11.0108	88.9892
BBBP	Weighted Loss	MoLFormer-XL	16	16	0	45,557,762	2,952,194	6.4801	93.5199
CLINTOX	Focal Loss	ChemBERTa-77M-MLM	4	32	0.1	3,428,210	205,826	6.0039	93.9961
CLINTOX	Focal Loss	ChemBERTa-77M-MTR	4	32	0.2	3,428,210	205,826	6.0039	93.9961
CLINTOX	Focal Loss	MoLFormer-XL	4	32	0.2	45,557,762	1,625,090	3.5671	96.4329
CLINTOX	Weighted Loss	ChemBERTa-77M-MLM	4	128	0.2	3,428,210	205,826	6.0039	93.9961
CLINTOX	Weighted Loss	ChemBERTa-77M-MTR	64	16	0.2	3,428,210	1,064,066	31.0385	68.9615
CLINTOX	Weighted Loss	MoLFormer-XL	64	128	0	45,557,762	8,260,610	18.1322	81.8678
FLAVOR	Focal Loss	ChemBERTa-77M-MLM	64	128	0	3,429,365	1,064,066	31.0281	68.9719
FLAVOR	Focal Loss	ChemBERTa-77M-MTR	16	128	0.2	3,429,365	377,474	11.0071	88.9929
FLAVOR	Focal Loss	MoLFormer-XL	8	32	0	45,560,069	2,067,458	4.5379	95.4621
FLAVOR	Weighted Loss	ChemBERTa-77M-MLM	64	64	0.2	3,429,365	1,064,066	31.0281	68.9719
FLAVOR	Weighted Loss	ChemBERTa-77M-MTR	32	128	0.2	3,429,365	606,338	17.6808	82.3192
FLAVOR	Weighted Loss	MoLFormer-XL	4	64	0.1	45,560,069	1,625,090	3.5669	96.4331

5. Discussion & Conclusions

The development of accurate and efficient molecular property prediction models remains a challenge in computational drug discovery and chemoinformatics. While transformer-based CLMs such as MolFormer and ChemBERTa

have demonstrated remarkable capabilities in learning molecular representations from SMILES sequences, their widespread adoption is hindered by substantial computational requirements for task-specific finetuning. This computational barrier is particularly pronounced in resource-constrained settings, where full finetuning of models with hundreds of millions of parameters for each downstream task becomes prohibitively expensive. Our work addresses this fundamental limitation by introducing EffiChem, a parameter-efficient framework that democratizes access to SOTA molecular property prediction through Low-Rank Adaptation (LoRA) of pretrained CLMs.

Current approaches for molecular property prediction fall into two categories: zero-shot inference using pretrained embeddings or full finetuning of all model parameters. Zero-shot methods, while computationally inexpensive, often underperform on specialized tasks as they cannot adapt to task-specific nuances. Conversely, full finetuning achieves superior performance but requires updating millions of parameters, leading to substantial computational costs, increased risk of overfitting on small datasets, and potential catastrophic forgetting of pretrained knowledge [37,41]. EffiChem innovates by integrating LoRA adapters with CLMs, reducing trainable parameters by 75–95% while exceeding the performance over fully finetuned models. Our comprehensive ablation studies across multiple feature modalities, from hand-crafted physicochemical (PC) features to learned embeddings (FT Embed), systematically evaluate the trade-offs between computational efficiency and predictive accuracy. Furthermore, we employ focal and weighted loss functions to effectively handle the severe class imbalance prevalent in molecular datasets, a critical consideration often overlooked in prior work.

Unlike adapter-only CLM methods that couple adaptation directly to the model's classification head, EffiChem decouples these stages: the LoRA-adapted encoder acts as a feature extractor whose embeddings are fused with molecular descriptors and fingerprints for tree-based ensemble prediction. This SMILES-only design also sets EffiChem apart from graph-based models (GROVER [7], GraphMVP [8]) and 3D-aware models (Uni-Mol [9]) that require graph construction or conformer generation, and from cross-modal models (MoleculeSTM [42], MolT5 [43]) designed for retrieval and generation rather than supervised property classification.

Our experimental evaluation across four diverse molecular property prediction tasks demonstrates the effectiveness of the EffiChem framework. For blood-brain barrier permeability (BBBP), E2E LoRA finetuning of MolFormer-XL-10pct-both with weighted loss achieved optimal performance with MCC of 0.797 representing 3–5% improvement over baseline results (see Table 1 and Supplementary Tables S2 and S3). On multi-class flavor prediction task (FART), our best model attained MCC of 0.801 and F1-macro of 0.809, achieving $\approx 4\%$ improvement in key metrics compared to optimal FART model [14]. For the BACE inhibitor classification task, E2E LoRA MolFormer with focal loss reached MCC of 0.577, demonstrating robust performance on this challenging dataset. Interestingly, the Clintox toxicity prediction task revealed that combining physicochemical features with base MolFormer embeddings (AUC: 0.999, MCC: 0.875, see Supplementary Table S6) outperformed finetuned approaches, suggesting that pretrained CLMs already capture toxicity-related molecular features effectively, and task-specific adaptation may introduce overfitting given the severe class imbalance ($>90\%$ majority class).

Several critical insights emerge from our comprehensive analysis. First, **task-dependent optimal strategies**: while E2E LoRA finetuning consistently achieves top performance for BBBP, Flavor, and BACE tasks, the Clintox task benefits most from combining physicochemical features with base embeddings, highlighting the importance of task-specific model selection. Second, **MolFormer's consistent superiority**: across all experimental conditions and datasets, MolFormer-XL-10pct-both substantially outperformed both ChemBERTa-77M variants (MLM and MTR), with performance gaps particularly pronounced on the challenging BACE task. This underscores the value of MolFormer's linear attention mechanism and pretraining on 1.1 billion molecules. Third, **task-dependent value of feature fusion**: the benefit of adding hand-crafted PC features to FT embeddings varies by task and metric. For BBBP, PC fusion yields a marginal AUC gain for MolFormer (0.951 vs. 0.949, Table 1), yet the best MCC is still delivered by E2E LoRA without any fusion—suggesting LoRA already captures the dominant permeability signals and PC features contributeredundant information on this dataset. For Flavor, PC+FT Embed achieves the highest precision (0.852) but a lower MCC than E2E LoRA, indicating selective complementarity for precision-sensitive metrics only. For BACE, fusion provides negligible improvement over FT embeddings alone (Figure 3), likely because the scaffold split has already removed the scaffold families where hand-crafted descriptors would add the most diversity signal. ClinTox is the exception: PC + Base embeddings outperform all LoRA-finetuned approaches, which we attribute to the extreme class imbalance ($>92\%$ majority class) causing LoRA adaptation to overfit, while pretrained MolFormer embeddings combined with physicochemical features suffice to capture toxicity-relevant structure-activity patterns. Fourth, **loss function selection matters**: focal loss proved superior for challenging, relatively balanced tasks (BACE), while weighted loss excelled for imbalanced tasks (BBBP, Flavor), emphasizing the importance of loss function design for molecular property prediction. These findings provide actionable guidance for practitioners developing molecular AI tools under computational constraints.

While EffiChem demonstrates substantial improvements in efficiency and performance, several avenues for future work remain. First, integration with multi-modal chemical language models: recent advances in multi-modal molecular representations, such as Mole-BERT [44], which combines SMILES with molecular graphs, and MolT5 [43], which leverages text-molecule pairs, suggest promising directions for further enhancement. Evaluating EffiChem's LoRA adaptation strategy on emerging multi-modal models like nach0 (which unifies molecular structures, properties, and textual descriptions) [45] and Omni-Mol (integrating 2D, 3D structural information with bioactivity data) [46] could yield additional performance gains while maintaining parameter efficiency. Second, extension to regression tasks and uncertainty quantification: our current work focuses on classification; extending EffiChem to molecular property regression with calibrated uncertainty estimates would broaden its applicability. Third, model interpretability: applying gradient-based attribution (e.g., integrated gradients), attention visualization, tree-stage feature importance, and UMAP/t-SNE embedding plots would expose which structural fragments and physicochemical properties drive each prediction, providing chemically actionable insights beyond aggregate metrics. Finally, as molecular foundation models continue to grow (e.g., models with billions of parameters), the parameter efficiency of LoRA becomes increasingly critical; evaluating EffiChem on large-scale models as and when they are designed would demonstrate its scalability.

In conclusion, EffiChem establishes a practical and efficient framework for adapting CLMs to specialized molecular property prediction tasks while reducing computational requirements by 75-95%. Through comprehensive evaluation across BBBP, toxicity, flavor classification, and BACE inhibition tasks, we demonstrate that PEFT via LoRA achieves performance exceeding fully finetuned models and state-of-the-art including MolFormer-XL. Our systematic ablation studies reveal task-dependent optimal strategies, with E2E LoRA finetuning proving most effective for BBBP, Flavor, and BACE tasks, while feature-enhanced base embeddings excel for toxicity prediction. By making our complete code-base publicly available, we provide the research community with reproducible tools to leverage parameter-efficient adaptation for their specific molecular property prediction challenges. Thus, EffiChem represents a significant step toward democratizing AI-driven drug discovery, enabling resource-constrained laboratories and organizations to deploy high-performance molecular AI tools without prohibitive computational infrastructure.

6. Limitations

A limitation of the current evaluation is its reliance on scaffold-based splitting, which, while the standard MoleculeNet protocol for enforcing chemical dissimilarity, carries inherent caveats. The Bemis–Murcko scaffold definition does not capture all axes of chemical diversity (e.g., stereochemistry or side-chain variation), and the resulting splits can introduce pronounced distribution shifts—as observed in BBBP, where the permeable-compound ratio differs by over 15% between training (78.8%) and test (63.4%) sets. Such shifts may simultaneously underestimate performance on structurally interpolated compounds and overestimate extrapolation difficulty for certain scaffold families. Complementary strategies such as temporal, cluster-based, or property-based splits would provide a more complete picture of model generalizability, and we identify this as an important direction for future benchmarking.

A second limitation is the absence of model interpretability analysis. While the ablation results indicate that LoRA-adapted CLMs internalize task-relevant molecular properties (evidenced by the minimal gain from adding hand-crafted features), direct mechanistic evidence is not provided. Techniques such as gradient-based attribution (e.g., integrated gradients), attention-weight visualization, feature importance from the tree-based stage, and embedding-space visualization (UMAP/t-SNE) would strengthen these claims and reveal which structural fragments drive predictions. We regard this as an important direction for future work.

7. Key Points

- EffiChem uses LoRA to efficiently adapt chemical language models, reducing trainable parameters by 75–95% compared to full finetuning while exceeding performance.
- End-to-end LoRA finetuning of MolFormer consistently delivers the best results, achieving 3–5% AUC improvement over state-of-the-art baselines across BBBP, flavor prediction, and BACE tasks.
- The benefit of adding hand-crafted physicochemical features to finetuned embeddings is task-dependent: negligible for BACE and mostly marginal for BBBP/Flavor (except precision on Flavor), but decisive for ClinTox where extreme class imbalance makes LoRA adaptation prone to overfitting.
- MolFormer outperforms ChemBERTa across all tasks and configurations, highlighting the advantage of its linear attention mechanism and larger-scale pretraining on 1.1 billion molecules.
- Loss function choice is task-dependent: weighted loss works best for imbalanced datasets (BBBP, Flavor), while focal loss excels on more balanced, challenging tasks (BACE).

Supplementary Materials

The following supporting information can be downloaded at: https://media.sciltp.com/articles/others/26071017 21004013/TI-26050030-SM.pdf, Table S1: RDKit-derived molecular descriptors, 'networkx' based graphical features, Morgan and MACCS fingerprints obtained from RDKit, together represent the set of engineered features. Table S2: Comparison of tree-based models built on physico-chemical (PC) features, base CLM embeddings, PC + base CLM embeddings. Best results highlighted in bold. Table S3: Comparison of models built through end-to-end (E2E) LoRA finetuning with tree-based models built on top of finetuned (FT) CLM embeddings and PC + FT Embeddings. Best results are highlighted in bold. Table S4: Comparison of models built through end-to-end (E2E) LoRA finetuning with tree-based models built on top of finetuned (FT) CLM embeddings and PC + FT Embeddings. Best results are highlighted in bold. Table S5: Comparison of tree-based models built on physico-chemical (PC) features, base CLM embeddings, PC + base CLM embeddings. Best results are highlighted in bold. Table S6: Comparison of tree-based models built on physico-chemical (PC) features, base CLM embeddings, PC + base CLM embeddings. Best results are highlighted in bold. Table S7: Comparison of models built through end-to-end (E2E) LoRA finetuning with tree-based models built on top of finetuned (FT) CLM embeddings and PC + FT Embeddings. Best results are highlighted in bold. Table S8: Comparison of tree-based models built on physico-chemical (PC) features, base CLM embeddings, PC + base CLM embeddings. Best results are highlighted in bold. Table S9: Comparison of models built through end-to-end (E2E) LoRA finetuning with tree-based models built on top of finetuned (FT) CLM embeddings and PC + FT Embeddings. Best results are highlighted in bold. References [47,48] are cited in supplementary materials.

Author Contributions

K.A. and H.H.: contributed to methods, experiments, validation, writing—original draft; I.S.: contributed to investigation, writing—review & editing; M.C. and H.B.: participated in investigation, infrastructure, writing—review & editing; S.G.: participated in conceptualization, project supervision, validation, writing—review & editing; R.M.: conceived, conceptualized, supervised the project, contributed in method, writing—review & editing. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

All the datasets and code are available at https://github.com/kavyagl2/EffChem-2.0.

Acknowledgments

The authors thank the Qatar National Library (QNL) for their support in providing resources and funding the APC, which was instrumental for publishing this research.

Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. R.M. is Scientific Advisor at Eternal, U.A.E.

Use of AI and AI-Assisted Technologies

During the preparation of this work the authors used ChatGPT (OpenAI) and Claude (Anthropic) to assist with phrasing, grammar correction, and refinement of English expressions in the manuscript. After using this tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

1. Vaswani, A.; Shazeer, N.; Parmar, N.; et al. Attention Is All You Need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
2. Ahmad, W.; Simon, E.; Chithrananda, S.; et al. ChemBERTa-2: Towards Chemical Foundation Models. *arXiv* **2022**, arXiv:2209.01712.
3. Huang, K.; Fu, T.; Glass, L.M.; et al. DeepPurpose: A Deep Learning Library for Drug–Target Interaction Prediction. *Bioinformatics* **2020**, *36*, 5545–5547. https://doi.org/10.1093/bioinformatics/btaa1005.
4. Ross, J.; Belgodere, B.; Chenthamarakshan, V.; et al. Large-Scale Chemical Language Representations Capture Molecular Structure and Properties. *Nat. Mach. Intell.* **2022**, *4*, 1256–1264. https://doi.org/10.1038/s42256-022-00580-7.
5. Chithrananda, S.; Grand, G.; Ramsundar, B. ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction. *arXiv* **2020**, arXiv:2010.09885.
6. Likhosherstov, V.; Choromanski, K.; Dubey, A.; et al. Favor#: Sharp Attention Kernel Approximations via New Classes of Positive Random Features. *arXiv* **2023**, arXiv:2302.00787.
7. Rong, Y.; Bian, Y.; Xu, T.; et al. Self-Supervised Graph Transformer on Large-Scale Molecular Data. *Adv. Neural Inf. Process. Syst.* 2020; *33*, 12559–12571.
8. Liu, S.; Wang, H.; Liu, W.; et al. Pre-Training Molecular Graph Representation with 3D Geometry. In Proceedings of the International Conference on Learning Representations, online, 25–29 April 2022.
9. Zhou, G.; Gao, Z.; Ding, Q.; et al. Uni-Mol: A Universal 3D Molecular Representation Learning Framework. In Proceedings of the Eleventh International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023.
10. Akiba, T.; Sano, S.; Yanase, T.; et al. Optuna: A Next-Generation Hyperparameter Optimization Framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 2623–2631. https://doi.org/10.1145/3292500.3330701.
11. Martins, I.F.; Teixeira, A.L.; Pinheiro, L.; et al. A Bayesian Approach to *In Silico* Blood-Brain Barrier Penetration Modeling. *J. Chem. Inf. Model.* **2012**, *52*, 1686–1697. https://doi.org/10.1021/ci300124c.
12. Wu, Z.; Ramsundar, B.; Feinberg, E.N.; et al. MoleculeNet: A Benchmark for Molecular Machine Learning. *Chem. Sci.* **2018**, *9*, 513–530. https://doi.org/10.1039/C7SC02664A.
13. Novick, P.A.; Ortiz, O.F.; Poelman, J.; et al. SweetLead: An *In Silico* Database of Approved Drugs, Regulated Chemicals, and Herbal Isolates for Computer-Aided Drug Discovery. *PLoS One* **2013**, *8*, e79568. https://doi.org/10.1371/journal.pone.0079568.
14. Zimmermann, Y.; Sieben, L.; Seng, H.; et al. A Chemical Language Model for Multi-Class Molecular Taste Prediction. *NPJ Sci. Food* **2025**, *9*, 122. https://doi.org/10.1038/s41538-025-00474-z.
15. Ramsundar, B.; Eastman, P.; Walters, P.; et al. *Deep Learning for the Life Sciences*; O'Reilly Media: Sebastopol, CA, USA, 2019.
16. Kim, S.; Chen, J.; Cheng, T.; et al. PubChem 2023 Update. *Nucleic Acids Res.* **2023**, *51*, D1373–D1380. https://doi.org/10.1093/nar/gkac956.
17. Han, X.; Jia, M.; Chang, Y.; et al. Directed Message Passing Neural Network (D-MPNN) with Graph Edge Attention (GEA) for Property Prediction of Biofuel-Relevant Species. *Energy AI* **2022**, *10*, 100201. https://doi.org/10.1016/j.egyai.2022.100201.
18. Pan, Z. Large Language Model for Molecular Chemistry. *Nat. Comput. Sci.* **2023**, *3*, 5.
19. Heo, B.; Park, S.; Han, D.; et al. Rotary Position Embedding for Vision Transformer. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2024; pp. 289–305.
20. Isard, M.; Yu, Y. Distributed Data-Parallel Computing Using a High-Level Programming Language. In Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data, Providence, RI, USA, 29 June–2 July 2009; pp. 987–994.
21. Irwin, J.J.; Tang, K.G.; Young, J.; et al. ZINC20—A Free Ultralarge-Scale Chemical Database for Ligand Discovery. *J. Chem. Inf. Model.* **2020**, *60*, 6065–6073. https://doi.org/10.1021/acs.jcim.0c00675.
22. Ruddigkeit, L.; van Deursen, R.; Blum, L.C.; et al. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2012**, *52*, 2864–2875. https://doi.org/10.1021/ci300415d.
23. Schuh, M.G.; Boldini, D.; Sieber, S.A. Synergizing Chemical Structures and Bioassay Descriptions for Enhanced Molecular Property Prediction in Drug Discovery. *J. Chem. Inf. Model.* **2024**, *64*, 4640–4650.
24. Nowakowska, S. ChemBERTa-2: Fine-Tuning for Molecule's HIV Replication Inhibition Prediction. *ChemRxiv* **2023**. https://doi.org/10.26434/chemrxiv-2023-b57vx.
25. Lee, D.; Cho, Y. Fine-Tuning Pocket-Conditioned 3D Molecule Generation via Reinforcement Learning. In Proceedings of the ICLR 2024 Workshop on Generative and Experimental Perspectives for Biomolecular Design, Vienna, Austria, 7–11 May 2024.
26. Balne, C.C.S.; Bhaduri, S.; Roy, T.; et al. Parameter Efficient Fine Tuning: A Comprehensive Analysis Across Applications. *arXiv* **2024**, arXiv:2404.13506.
27. Zhou, H.; Wan, X.; Vulić, I.; et al. AutoPEFT: Automatic Configuration Search for Parameter-Efficient Fine-Tuning. *Trans. Assoc. Comput. Linguist.* **2024**, *12*, 525–542.
28. Fu, Z.; Yang, H.; So, A.M.C.; et al. On the Effectiveness of Parameter-Efficient Fine-Tuning. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA,7–14 February 2023; pp. 12799–12807.
29. Xu, L.; Xie, H.; Qin, S.Z.J.; et al. Parameter-Efficient Fine-Tuning Methods for Pretrained Language Models: A Critical Review and Assessment. *IEEE Trans. Pattern Anal. Mach. Intell.* **2026**, *48*, 6107–6126.

30. Hu, E.J.; Shen, Y.; Wallis, P.; et al. LoRA: Low-Rank Adaptation of Large Language Models. In Proceedings of the 2022 International Conference on Learning Representations (ICLR), Online, 25 April 2022.
31. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
32. Ke, G.; Meng, Q.; Finley, T.; et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp.3146–3154.
33. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; et al. CatBoost: Unbiased Boosting with Categorical Features. In Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montreal, QC, Canada, 3–8 December 2018; Volume 31, pp. 6638–6648.
34. Chang, J.; Ye, J.C. Bidirectional Generation of Structure and Properties Through a Single Molecular Foundation Model. *Nat. Commun.* **2024**, *15*, 2323.
35. Mall, R.; Elbasir, A.; Almeer, H.; et al. A Modelling Framework for Embedding-Based Predictions for Compound-Viral Protein Activity. *Bioinformatics* **2021**, *37*, 2544–2555.
36. Alhosani, H.; Mall, R.; Singh, A.; et al. Feature-PHLA: Physico-Chemical Features Efficiently Predict Peptide-HLA Binding Affinity. In Proceedings of the 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Lisbon, Portugal, 3–6 December 2024; pp. 4–11.
37. Mall, R.; Singh, A.; Patel, C.N.; et al. VISH-Pred: An Ensemble of Fine-Tuned ESM Models for Protein Toxicity Prediction. *Brief. Bioinform.* **2024**, *25*, bbae243.
38. Khurana, S.; Rawi, R.; Kunji, K.; et al. DeepSol: A Deep Learning Framework for Sequence-Based Protein Solubility Prediction. *Bioinformatics* **2018**, *34*, 2605–2613.
39. Elbasir, A.; Mall, R.; Kunji, K.; et al. BCrystal: An Interpretable Sequence-Based Protein Crystallization Predictor. *Bioinformatics* **2020**, *36*, 1429–1438.
40. Jethalia, M.; Jani, S.P.; Ceccarelli, M.; et al. PanCancer Network Analysis Reveals Key Master Regulators for Cancer Invasiveness. *J. Transl. Med.* **2023**, *21*, 558.
41. Mall, R.; Kaushik, R.; Martinez, Z.A.; et al. Benchmarking Protein Language Models for Protein Crystallization. *Sci. Rep.* **2025**, *15*, 2381.
42. Liu, S.; Nie, W.; Wang, C.; et al. Multi-Modal Molecule Structure–Text Model for Text-Based Retrieval and Editing. *Nat. Mach. Intell.* **2023**, *5*, 1447–1457.
43. Edwards, C.; Lai, T.; Ros, K.; et al. Translation Between Molecules and Natural Language. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, United Arab Emirates, 7–11 December 2022.
44. Xia, J.; Zhao, C.; Hu, B.; et al. Mole-BERT: Rethinking Pre-Training Graph Neural Networks for Molecules. In Proceedings of the Eleventh International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023.
45. Livne, M.; Miftahutdinov, Z.; Tutubalina, E.; et al. Nach0: Multimodal Natural and Chemical Languages Foundation Model. *Chem. Sci.* **2024**, *15*, 8380–8389.
46. Hu, C.; Li, H.; Yuan, Y.; et al. Omni-Mol: Multitask Molecular Model for Any-to-Any Modalities. *Adv. Neural Inf. Process. Syst.* **2026**, *38*, 66945–66988.
47. Maleki, S.; Huetter, J.C.; Chuang, K.V.; et al. Efficient Fine-Tuning of Single-Cell Foundation Models Enables Zero-Shot Molecular Perturbation Prediction. *arXiv* **2024**, arXiv:2412.13478.
48. Ren, W.; Li, X.; Wang, L.; et al. Analyzing and Reducing Catastrophic Forgetting in Parameter Efficient Tuning. *arXiv* **2024**, arXiv:2402.18865.