Predicting Antiviral Potenty and ADMET Using Ensemble Approach of Chemical Language Model and Fingerprint Features

Azmine Toushik Wasi[1]
[1] Shahjalal University of Science and Technology, Sylhet, Bangladesh

Abstract: The **ASAP Discovery x OpenADMET challenge** was organized by ASAP Discovery and OpenADMET with polaris, an initiative designed to replicate real-world challenges in preclinical antiviral drug discovery. ASAP Discovery is in the process of patenting preclinical candidates targeting coronavirus Mpro, and this competition reflects some of the hurdles they have faced over the past three years. By participating in both the ADMET and potency prediction tasks, we aimed to contribute insights into how computational techniques can enhance the drug discovery pipeline. Our study underscores the viability of in silico approaches while emphasizing the importance of robust validation strategies to ensure clinical relevance. Additionally, we discuss the limitations of current methodologies and propose future directions to improve prediction reliability in both pharmacokinetics and antiviral potency modeling.

The drug discovery process is often hindered by late-stage failures due to unfavorable pharmacokinetic and toxicity profiles. Early prediction of ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxicity) properties plays a critical role in mitigating these risks, reducing both the cost and time associated with developing new drugs. In this work, we leverage machine learning and quantitative structure–activity relationship (QSAR) approaches to predict ADMET properties, focusing on key assay readouts such as MLM (microsomal stability), HLM (human liver microsomal stability), KSOL (solubility), LogD (lipophilicity), and MDR1-MDCKII (permeability). The dataset was provided by competition organizers, and we explored feature extraction techniques using molecular descriptors, followed by model development using various machine learning algorithms. Among these, ensemble methods like Random Forests demonstrated high predictive accuracy. The results are analyzed using statistical metrics such as R^2, RMSE, and accuracy, alongside ROC curves and feature importance graphs to illustrate model performance.

In addition to ADMET prediction, we also tackled the Antiviral-Potency-2025 task, which is particularly relevant in computational drug discovery. This task involves predicting the pIC50 values for two coronavirus main protease (Mpro) targets—SARS-CoV-2 Mpro and MERS-CoV Mpro. While potency prediction is one of the most widely studied aspects of computational modeling, it represents only part of the broader drug discovery landscape. We approached this challenge by developing predictive models to estimate the negative log10 of the IC50 values from dose-response curves, mirroring real-world antiviral screening efforts.

1. Introduction

The drug discovery process is inherently complex, with significant challenges arising in late-stage development due to adverse pharmacokinetic and toxicity profiles. These issues often result in costly failures, which can hinder the overall progress of new drug candidates. One of the most critical stages in drug discovery is the prediction of ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxicity) properties, which are essential for assessing the viability of compounds before clinical trials. In modern drug discovery, the in silico prediction of pharmacokinetic and toxicity properties has emerged as a critical component in streamlining the development pipeline. Pharmaceutical companies invest significant resources in experimental assays for ADMET properties, yet many promising compounds fail in later phases because of unexpected ADMET issues. In recent years, the integration of machine learning (ML) and quantitative structure-activity relationship (QSAR) models has shown great promise in providing early-stage predictions of ADMET characteristics, enabling more informed decisions and reducing the high attrition rates in drug development. However, despite the advancements, the field continues to struggle with improving prediction accuracy and ensuring reliable performance across various datasets and drug classes.

The motivation for this work stems from the significant gaps in current methods for ADMET and antiviral potency prediction. Although ML techniques like Random Forests have demonstrated promise in predicting pharmacokinetic properties, challenges remain in developing models that generalize well across diverse chemical spaces. Moreover, existing approaches often lack the robustness required to handle the complex interplay between chemical structures and their pharmacokinetic behavior, particularly in the context of less-explored compounds. The Antiviral-Potency-2025 task further emphasizes the need for more accurate predictive models for drug candidates targeting novel viral strains, such as SARS-CoV-2 and MERS-CoV. These models not only need to predict potency but also must account for the variability observed in real-world biological screening, which often presents significant challenges due to the inherent noise in experimental data and the complexity of biological systems.

In response to these challenges, our work introduces an innovative ensemble approach that combines molecular fingerprints with Chemical Language Models (CLMs) for enhanced feature extraction and model training. By utilizing Gradient Boosting, we build predictive models that integrate diverse molecular descriptors and chemical context, improving the overall performance of ADMET and antiviral potency prediction tasks. Our method leverages both the power of traditional QSAR modeling and the more recent advancements in deep learning for chemical representations, offering a more nuanced understanding of molecular properties. In addition, we propose a simple yet effective weighted ensemble strategy, which further enhances the robustness and accuracy of the predictions. Through this approach, we contribute to bridging the gap between early-stage drug discovery and real-world clinical outcomes, demonstrating the potential of in silico models in addressing the limitations of current methodologies.

2. Background

2.1. Drug Discovery and the Role of ADMET and Computational Approaches

Advancements in computational methods have significantly enhanced the prediction of ADMET properties, supplementing traditional experimental approaches. Machine learning (ML) techniques, including Random Forests, Support Vector Machines (SVM), and Neural Networks, have been effectively employed to model complex relationships between molecular structures and their ADMET outcomes. These methods can automatically learn intricate interactions from molecular descriptors without necessitating explicit mechanistic understanding, thereby streamlining the drug development process [2].

Recent studies have demonstrated the efficacy of ML models in ADMET prediction. For instance, the development of DeepDelta, a pairwise deep learning approach, has shown significant improvements in predicting differences in ADMET properties between molecular derivatives. This method directly learns property differences from pairs of molecules, outperforming traditional models in various benchmarks [1]. Similarly, platforms like ADMET-AI have been developed to provide rapid and accurate ADMET predictions, utilizing machine learning to process large datasets efficiently [5].

Despite these advancements, challenges persist in developing models that generalize well across diverse chemical spaces. Variability in chemical structures and the complexity of their interactions with biological systems necessitate continuous refinement of computational models. Addressing these challenges is crucial for enhancing the reliability of in silico predictions and their integration into the drug discovery pipeline.

2.2. Computational Approaches in Antiviral Potency Prediction

The accurate prediction of antiviral potency is crucial in the early stages of drug discovery, enabling the identification of effective compounds against viral pathogens. Computational approaches, particularly machine learning (ML) and quantitative structure–activity relationship (QSAR) models, have become instrumental in this domain. These methods analyze large datasets of chemical compounds to predict their efficacy against specific viral targets, thereby reducing the need for extensive experimental screening.

A notable example is the HIV Resistance Response Database Initiative (RDI), which utilized AI to predict patient responses to HIV drugs based on data from over 250,000 patients across 50 countries. Their HIV Treatment Response Prediction System (HIV-TRePS) enabled healthcare professionals to tailor treatments effectively, achieving accuracies around 80%. In the context of the COVID-19 pandemic, projects like COVID Moonshot and Exscalate4Cov have demonstrated the power of computational methods in rapidly identifying potential antiviral agents. COVID Moonshot, for instance, harnessed AI and ML to process thousands of compound submissions, aiming to design inhibitors targeting the SARS-CoV-2 main protease. Similarly, Exscalate4Cov employed high-throughput virtual screening of billions of compounds against SARS-CoV-2 proteins, significantly accelerating the discovery process.

Despite these advancements, challenges persist in developing models that generalize well across diverse chemical spaces and accurately predict antiviral potency. Variability in chemical structures and their interactions with complex viral targets necessitates continuous refinement of computational models. Addressing these challenges is essential for enhancing the reliability of in silico predictions and their integration into the drug discovery pipeline.

3. Problem Statement

3.1. ADMET Challenge

The drug discovery process is inherently complex, and one of the critical stages is the evaluation of the pharmacokinetic and toxicological properties of potential drug candidates. The ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxicology) properties are essential indicators of a compound's likelihood of success in clinical trials. These properties determine how a drug behaves in the human body, its efficacy, and its safety. Unfortunately, the assessment of these properties often occurs late in the discovery pipeline, where failures can lead to costly delays. Therefore, early prediction of ADMET endpoints is a key component in accelerating the drug discovery process and reducing the associated costs and time.

In this challenge, we focus on predicting five crucial ADMET endpoints based on the molecular structure of compounds. The chosen endpoints include Mouse Liver Microsomal stability (MLM), Human Liver Microsomal stability (HLM), solubility (KSOL), lipophilicity (LogD), and cell permeability (MDR1-MDCKII). These properties significantly influence the pharmacokinetics and pharmacodynamics of drug candidates. MLM and HLM assays measure the metabolic stability of compounds in mouse and human liver microsomes, respectively, providing an estimate of how long a compound will remain in the body before being cleared. Solubility (KSOL) and LogD offer insights into a compound's ability to dissolve in aqueous or fatty environments, impacting its absorption and distribution. Finally, cell permeability (MDR1-MDCKII) predicts the ability of drug candidates to permeate cellular barriers, which is particularly important for drug delivery to the brain and other tissues.

The dataset provided for this challenge includes various molecular descriptors, such as the CXSMILES (a textual representation of the 2D molecular structure) of each compound, along with assay readouts for the five endpoints. The goal is to predict the values for these endpoints for a test set of compounds, where only the CXSMILES are provided. The prediction is evaluated based on the Mean Absolute Error (MAE) of the log-transformed endpoints after clipping, which helps reduce the impact of outliers on the evaluation process. A key challenge in this task lies in the accurate prediction of ADMET properties from molecular structures. The molecular descriptors used for training machine learning models must capture the complex interactions between a drug's chemical structure and its pharmacokinetic and toxicological properties. This requires the development of robust models that generalize well to unseen chemical spaces, especially given that the training and test sets are split temporally. The task presents an opportunity to advance computational drug discovery by leveraging state-of-the-art machine learning techniques and incorporating both traditional and novel molecular descriptors to improve prediction accuracy. Thus, the problem involves not only predicting the ADMET properties accurately but also overcoming the complexities associated with molecular interactions, data sparsity, and chemical space diversity. A successful solution to this problem will contribute significantly to the field of computational drug discovery and enable more efficient identification of promising drug candidates in the early stages of development.

3.2. Potency Challenge

The prediction of antiviral potency plays a critical role in the early stages of drug discovery, particularly for diseases caused by rapidly evolving viruses such as the coronaviruses. The potency of a compound refers to its ability to inhibit a specific biological target, which, in this case, are the main proteases (Mpro) of SARS-CoV-2 and MERS-CoV. These proteases are essential for the replication of the viruses, making them prime targets for antiviral drug development. Accurate prediction of potency is crucial to prioritize compounds with the highest likelihood of success, reducing the time and cost associated with experimental screening. However, while potency prediction is an important aspect of antiviral drug discovery, it represents only one part of the broader evaluation of a compound's potential as a drug. The full therapeutic profile requires considering other factors such as ADMET properties and off-target effects, which are not addressed by potency alone.

In this challenge, the task is to predict the potency of compounds against two important coronavirus targets: SARS-CoV-2 Mpro and MERS-CoV Mpro. The potency is quantified as the negative logarithm (pIC50) of the IC50 values from dose-response curves, which reflect the concentration of a compound required to inhibit 50% of the target enzyme activity. The training dataset includes molecular information, such as the CXSMILES representation of each compound and the corresponding pIC50 values for both SARS-CoV-2 and MERS-CoV Mpro. During the evaluation phase, only the CXSMILES strings are provided, and the goal is to predict the pIC50 values for these test compounds. The predictions are evaluated based on the mean absolute error (MAE), which measures the accuracy of the predicted values against the true values. The challenge introduces several complexities, notably the need to generalize across diverse chemical spaces. Since the test data is temporally split, there may be overlaps between the training and test sets in terms of chemical structures, but the predictive models must still perform well on unseen data.

Moreover, the training data for this task involves compounds with varying degrees of potency, and the relationships between molecular features and antiviral activity are not always straightforward. Therefore, the problem is not only about building a model that predicts potency accurately but also about ensuring that the model generalizes effectively across different molecular scaffolds and behaves robustly when applied to novel compounds. This makes the task highly relevant for real-world drug discovery, where compounds are often tested in different temporal and chemical contexts.

Successfully addressing this challenge will provide valuable insights into how computational techniques can be used to prioritize antiviral candidates early in the drug discovery process. By building models that can predict the potency of compounds against Mpro targets, the scientific community can expedite the identification of promising drug candidates, which is especially critical in the context of emerging viral threats like COVID-19. In addition, advancements made through this challenge may lay the groundwork for future work in antiviral drug discovery and highlight the importance of computational modeling in drug development pipelines.

4. Solution Methodologies

4.1. Molecular Embedding Generation

4.1.1. Transformer Models (ChemLMs)

A key component of this workflow is the extraction of molecular embeddings using pre-trained transformer models. Transformer models have demonstrated remarkable success in various natural language processing tasks and have recently been adapted for representing molecules as well. The SMILES string representation lends itself well to processing by these sequence-based models. This workflow utilizes three distinct pre-trained transformer models to generate molecular embeddings, aiming to capture different aspects of molecular information.

***unikei/bert-base-smiles*: Architecture and Pretraining Details** The first transformer model employed is *unikei/bert-base-smiles*[1]. This model is a bidirectional transformer that has been specifically pre-trained on a large corpus of SMILES strings. The pretraining process involved two primary objectives: masked language modeling and molecular-formula validity prediction. Masked language modeling entails randomly masking certain tokens (characters or sub-sequences) within a SMILES string and training the model to predict these masked tokens based on the surrounding context. This task forces the model to learn the relationships between different atoms and bonds and to understand the chemical grammar encoded in SMILES. The second pretraining objective, molecular-formula validity prediction, trains the model to distinguish between valid and invalid SMILES representations, further enhancing its understanding of chemical structure.

***ibm-research/MoLFormer-XL-both-10pct*: Architecture and Pretraining Details** The second transformer model utilized in this workflow is *ibm-research/MoLFormer-XL-both-10pct*[2] [3, 4]. This model is a state-of-the-art chemical language model specifically designed to process and understand molecular structures through their SMILES string representations. MoLFormer belongs to a class of models that

[1] https://huggingface.co/unikei/bert-base-smiles

[2] https://huggingface.co/ibm-research/MoLFormer-XL-both-10pct

have been pre-trained on vast datasets of SMILES strings, including up to 1.1 billion molecules from the ZINC and PubChem databases. The specific variant used here, MoLFormer-XL-both-10pct, has been pre-trained on 10% of the molecules from both the ZINC and PubChem datasets.

***seyonec/PubChem10M_SMILES_BPE_450k*: Architecture and Pretraining Details** The third pre-trained transformer model utilized in this workflow is *seyonec/PubChem10M_SMILES_BPE_450k*[3]. Unlike the previous two models, the Hugging Face page for this model does not provide a detailed model card with comprehensive information about its architecture and pretraining. However, by examining the config.json file associated with the model, the architecture can be identified as RobertaForMaskedLM. RoBERTa (A Robustly Optimized BERT Pretraining Approach) is a transformer-based model that builds upon the BERT architecture and often achieves improved performance through modifications to the pretraining procedure.

4.1.2. Molecular Fingerprint Generation using RDKit

In addition to using pre-trained transformer models to generate molecular embeddings, this workflow also employs traditional cheminformatics techniques to generate molecular fingerprints from the SMILES strings using the RDKit library. Molecular fingerprints are binary or count-based representations of molecular structure that are widely used in drug discovery for tasks such as similarity searching, virtual screening, and QSAR modeling. This workflow utilizes two different types of fingerprints: ECFP and MACCS.

ECFP Fingerprints: Parameters and Generation ECFP, or Extended Connectivity Fingerprints, are a type of circular fingerprint that captures information about the atom environments and connectivity within a molecule up to a certain radius. They are generated by iteratively assigning unique identifiers to each atom based on its connectivity and the identifiers of its neighbors. The process is repeated for a specified number of iterations, determined by the radius parameter. A larger radius considers a larger neighborhood around each atom, capturing more extensive structural information. The resulting atom identifiers are then hashed and folded into a bit vector of a predefined length, specified by the nBits (or fpSize) parameter.

MACCS Fingerprints: Parameters and Generation The second type of molecular fingerprint generated in this workflow is the MACCS (Molecular ACCess System) fingerprint. Unlike ECFP, which is generated based on a general algorithm, MACCS fingerprints are based on a predefined set of structural keys or rules. The RDKit implementation of MACCS fingerprints uses 167 such keys, each corresponding to the presence or absence of a specific structural feature or pattern within the molecule. The generation of a MACCS fingerprint involves evaluating each of these 167 predefined rules against the molecule. If a particular structural feature is present, the corresponding bit in the fingerprint is set to 1; otherwise, it is set to 0. This results in a fixed-length binary vector of 167 bits, where each bit represents the answer to a specific question about the molecule's structure, such as Is there a ring of size 4? or Is at least one halogen present?. RDKit provides a straightforward function to generate MACCS fingerprints from a molecule object derived from a SMILES string. These fingerprints offer a compact and interpretable representation of molecular structure based on the presence or absence of well-defined structural fragments.

[3] https://huggingface.co/seyonec/PubChem10M_SMILES_BPE_450k

4.2. Handling Missing Target Data

In many real-world datasets, including those pertaining to ADMET/potency properties, the presence of missing values is a common challenge. These missing entries can arise due to various reasons, such as experimental limitations or data collection issues. If left unaddressed, missing data can negatively impact the training and performance of machine learning models. Therefore, a crucial step in this workflow involves handling these missing values in the target ADMET properties of the training data. A common strategy to address this is through imputation, where the missing values are estimated based on the other available information in the dataset. The workflow described utilizes a GradientBoostingRegressor from the scikit-learn library to predict and impute the missing values in the target properties of the training data. Gradient boosting is an ensemble learning technique that builds a strong predictive model by combining the predictions of multiple weaker models, typically decision trees. By training a GradientBoostingRegressor on the samples with complete target property values, the model learns the relationships between the molecular features (likely the embeddings extracted in the previous step or other available descriptors) and the target property. This learned relationship can then be used to predict the missing values for the samples where the target property is not available.

4.3. Generating Predictions

4.3.1. Prediction with Embedding-Based Multi-Layer Perceptron Models

The MLP model is a fully connected neural network designed for regression tasks, consisting of three layers: an input layer, two hidden layers, and an output layer. The first hidden layer has 1024 neurons, while the second has 728, both activated using ReLU functions and regularized with Dropout (0.2) to prevent overfitting. The model is trained using the Adam optimizer with a learning rate of 1e-3, and the Mean Squared Error (MSE) is used as the loss function. The training process includes early stopping with a patience of 10 epochs, where training halts if validation loss does not improve for 10 consecutive epochs. The dataset is split into 90% training and 10% validation, with a batch size of 32, ensuring effective optimization while preventing overfitting.

4.3.2. Prediction with Embedding-Based Gradient Boosting Regressors

The prediction process utilizes ChemLM-derived embeddings as input features to train Gradient Boosting Regressors (GBR) for multiple molecular targets. Separate models are trained for each target using embeddings from three fingerprint representations (model1, model2, model3). The dataset is split into 90% training an 10% validation split from the training data. Each GBR model is trained with 500 trees (n_estimators=500), a learning rate of 0.01, and a maximum depth of 6, optimizing the fit while controlling overfitting. The model is evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), and R^2 score on the validation set.

4.3.3. Prediction with Fingerprint-Based Gradient Boosting Regressors

The prediction process leverages molecular fingerprint embeddings (e.g., ECFP and MACCS) to train Gradient Boosting Regressors (GBR) for multiple molecular targets. Each fingerprint is processed separately, with data split into 90% training and 10% val sets. The GBR is configured with 250 estimators, a learning rate of 0.01, and a maximum depth of 5 for tree structures, optimizing model complexity and prediction accuracy. After training, the model is evaluated using Mean Absolute Error (MAE), Mean

Squared Error (MSE), and R^2 score on the validation set.

4.4. Ensemble and Final Prediction

Lastly, we combined all the test predictions from all the trained models based on weights (mostly calculated from the R^2 scores for each combination) to generate the final prediction. The weights are given in Table 2 and 4 for ADMET and Potency respectively.

Table 1: ADMET Challenge: Evaluation Metrics for Different Models Across Targets (10% Val Set)

Setup	Model/Fingerprint	Target	MSE	MAE	R^2 Score
ChemLM+MLP	*unikei/bert-base-smiles*	MLM	31038.4512	124.2559	-0.4056
	ibm-research/MoLFormer-XL-both-10pct	MLM	27333.5703	103.0010	-0.1408
	seyonec/PubChem10M_SMILES_BPE_450k	MLM	23615.9512	103.5613	0.3609
	unikei/bert-base-smiles	HLM	55577.3789	134.1852	0.3516
	ibm-research/MoLFormer-XL-both-10pct	HLM	18745.2812	90.7356	0.0825
	seyonec/PubChem10M_SMILES_BPE_450k	HLM	47109.2773	100.4175	0.0976
	unikei/bert-base-smiles	KSOL	12874.5293	85.6253	0.4242
	ibm-research/MoLFormer-XL-both-10pct	KSOL	12843.4336	85.2019	0.4770
	seyonec/PubChem10M_SMILES_BPE_450k	KSOL	9979.1807	70.6956	0.4633
	unikei/bert-base-smiles	LogD	1.0951	0.8626	0.2467
	ibm-research/MoLFormer-XL-both-10pct	LogD	0.5554	0.5383	0.5479
	seyonec/PubChem10M_SMILES_BPE_450k	LogD	0.5468	0.5447	0.5331
	unikei/bert-base-smiles	MDR1-MDCKII	41.8468	3.6954	0.2160
	ibm-research/MoLFormer-XL-both-10pct	MDR1-MDCKII	30.2398	3.5705	0.4915
	seyonec/PubChem10M_SMILES_BPE_450k	MDR1-MDCKII	29.6319	3.1879	0.5102
ChemLM+GB	*unikei/bert-base-smiles*	MLM	33995.0650	109.4686	0.6696
	unikei/bert-base-smiles	HLM	8237.8764	61.6570	0.1217
	unikei/bert-base-smiles	KSOL	7335.3774	64.9156	0.6643
	unikei/bert-base-smiles	LogD	0.5214	0.5302	0.5642
	unikei/bert-base-smiles	MDR1-MDCKII	21.7398	3.1533	0.5029
	ibm-research/MoLFormer-XL-both-10pct	MLM	27199.2508	116.8974	0.7357
	ibm-research/MoLFormer-XL-both-10pct	HLM	7824.4482	57.9228	0.1657
	ibm-research/MoLFormer-XL-both-10pct	KSOL	10738.5183	74.2544	0.5085
	ibm-research/MoLFormer-XL-both-10pct	LogD	0.4070	0.4616	0.6598
	ibm-research/MoLFormer-XL-both-10pct	MDR1-MDCKII	44.8666	4.3502	-0.0259
	seyonec/PubChem10M_SMILES_BPE_450k	MLM	100384.7293	142.9475	0.0244
	seyonec/PubChem10M_SMILES_BPE_450k	HLM	13534.8449	63.5901	-0.4431
	seyonec/PubChem10M_SMILES_BPE_450k	KSOL	8519.1801	68.5093	0.6101
	seyonec/PubChem10M_SMILES_BPE_450k	LogD	0.3601	0.4415	0.6990
	seyonec/PubChem10M_SMILES_BPE_450k	MDR1-MDCKII	24.0709	3.6017	0.4496
FP+GB	ECFP	MLM	17206.2617	92.9497	0.8328
	ECFP	HLM	4879.7816	57.5628	0.4797
	ECFP	KSOL	7845.2460	69.0913	0.6409
	ECFP	LogD	0.5691	0.5177	0.5243
	ECFP	MDR1-MDCKII	17.7733	2.8569	0.5936
	MACCS	MLM	18007.8987	103.8011	0.8250
	MACCS	HLM	8039.8588	72.7008	0.1428
	MACCS	KSOL	8257.2350	73.5085	0.6221
	MACCS	LogD	0.6613	0.5570	0.4473
	MACCS	MDR1-MDCKII	16.6995	2.5457	0.6182

Table 2: ADMET Challenge: Model Weights for Different Targets

Target	Model	Weight
MLM	*unikei/bert-base-smiles*-GB	0.03
	ibm-research/MoLFormer-XL-both-10pct-GB	0.07
	seyonec/PubChem10M_SMILES_BPE_450k-GB	0.00
	ECFP	0.45
	MACCS	0.45
HLM	*unikei/bert-base-smiles*-GB	0.10
	ibm-research/MoLFormer-XL-both-10pct-GB	0.10
	seyonec/PubChem10M_SMILES_BPE_450k-GB	0.00
	ECFP	0.60
	MACCS	0.10
KSOL	*unikei/bert-base-smiles*-GB	0.25
	ibm-research/MoLFormer-XL-both-10pct-GB	0.10
	seyonec/PubChem10M_SMILES_BPE_450k-GB	0.20
	ECFP	0.25
	MACCS	0.20
LogD	*unikei/bert-base-smiles*	0.20
	ibm-research/MoLFormer-XL-both-10pct	0.30
	seyonec/PubChem10M_SMILES_BPE_450k	0.40
	ECFP	0.05
	MACCS	0.05
MDR1-MDCKII	*unikei/bert-base-smiles*-GB	0.20
	ibm-research/MoLFormer-XL-both-10pct-GB	0.00
	seyonec/PubChem10M_SMILES_BPE_450k-GB	0.10
	ECFP	0.30
	MACCS	0.40

5. Results and Discussion

5.1. ADMET Challenge

Table 1 presents evaluation metrics for the ADMET challenge across various models and fingerprints on a 10% validation set. It comprises three distinct setups: ChemLM+MLP, ChemLM+GB, and FP+GB. For each setup, multiple targets (MLM, HLM, KSOL, LogD, MDR1-MDCKII) are evaluated using Mean Squared Error (MSE), Mean Absolute Error (MAE), and R^2 score. These metrics provide insights into the models' ability to capture complex relationships between molecular embeddings or fingerprints and the measured endpoints. Overall, the table highlights substantial variability in performance depending on the underlying representation and regression technique. It underscores the critical role of model architecture and feature representation in achieving robust predictions in ADMET studies. This detailed comparison sets the stage for a nuanced discussion on strengths and limitations across different approaches.

Focusing on the ChemLM+MLP setup, we observe that the performance varies considerably across different embedding models. For instance, for the MLM target, the *unikei/bert-base-smiles* embedding yields an R^2 score of -0.4056, while the *seyonec/PubChem10M_SMILES_BPE_450k* model improves this to 0.3609. Similar trends are evident in other endpoints, where some embeddings even result in negative R^2 values, indicating poor predictive power or potential overfitting. The variability in MAE and MSE across targets suggests that the MLP architecture may not be robust enough to handle all aspects of the ADMET prediction tasks consistently. This setup, while innovative in leveraging ChemLM embeddings, demonstrates that the choice of embedding critically influences model performance. Moreover, the differences in performance highlight the challenges of mapping high-dimensional molecular representations to accurate quantitative predictions.

Table 3: Potency Challenge: Evaluation Metrics for Different Models Across Targets (10% Val Set)

Setup	Model/Fingerprint	Target	MSE	MAE	R^2 Score
ChemLM+MLP	*unikei/bert-base-smiles*	pIC50 (MERS-CoV Mpro)	0.7632	0.5399	-0.1889
	ibm-research/MoLFormer-XL-both-10pct	pIC50 (MERS-CoV Mpro)	0.6887	0.5967	-0.1787
	seyonec/PubChem10M_SMILES_BPE_450k	pIC50 (MERS-CoV Mpro)	0.5013	0.4509	0.2507
	unikei/bert-base-smiles	pIC50 (SARS-CoV-2 Mpro)	0.4879	0.5826	0.5125
	ibm-research/MoLFormer-XL-both-10pct	pIC50 (SARS-CoV-2 Mpro)	0.5822	0.5959	0.4260
	seyonec/PubChem10M_SMILES_BPE_450k	pIC50 (SARS-CoV-2 Mpro)	0.4309	0.4992	0.4571
ChemLM+GB	*unikei/bert-base-smiles*	pIC50 (MERS-CoV Mpro)	0.6859	0.5033	0.1179
	ibm-research/MoLFormer-XL-both-10pct	pIC50 (MERS-CoV Mpro)	0.7695	0.5035	0.0104
	seyonec/PubChem10M_SMILES_BPE_450k	pIC50 (MERS-CoV Mpro)	0.7699	0.4992	0.0098
	unikei/bert-base-smiles	pIC50 (SARS-CoV-2 Mpro)	0.3471	0.4523	0.5560
	ibm-research/MoLFormer-XL-both-10pct	pIC50 (SARS-CoV-2 Mpro)	0.3077	0.4251	0.6064
	seyonec/PubChem10M_SMILES_BPE_450k	pIC50 (SARS-CoV-2 Mpro)	0.2929	0.4087	0.6254
FP+GB	ECFP	pIC50 (MERS-CoV Mpro)	0.7338	0.5172	0.0562
	ECFP	pIC50 (SARS-CoV-2 Mpro)	0.2718	0.4293	0.6524
	MACCS	pIC50 (MERS-CoV Mpro)	0.7227	0.4864	0.0706
	MACCS	pIC50 (SARS-CoV-2 Mpro)	0.2831	0.4198	0.6379

Table 4: Potency Challenge: Model Weights for pIC50 Prediction

Target	Model	Weight
pIC50 (MERS-CoV Mpro)	*unikei/bert-base-smiles*-GB	0.2
	ibm-research/MoLFormer-XL-both-10pct-GB	0.05
	seyonec/PubChem10M_SMILES_BPE_450k-GB	0.05
	seyonec/PubChem10M_SMILES_BPE_450k-MLP	0.6
	ECFP	0.05
	MACCS	0.05
pIC50 (SARS-CoV-2 Mpro)	*unikei/bert-base-smiles*-GB	0.0
	ibm-research/MoLFormer-XL-both-10pct-GB	0.1
	seyonec/PubChem10M_SMILES_BPE_450k-GB	0.15
	seyonec/PubChem10M_SMILES_BPE_450k-MLP	0.0
	ECFP	0.4
	MACCS	0.35

In contrast, the ChemLM+GB setup appears to offer more stable performance for most targets. With Gradient Boosting as the regressor, the *unikei/bert-base-smiles* embedding, for instance, achieves an R^2 of 0.6696 for MLM, a significant improvement over the MLP counterpart. Although the performance for other targets like HLM and KSOL remains moderate, the overall trend suggests that ensemble methods are better suited to extract value from ChemLM embeddings. The GBR approach, with its inherent capacity to handle non-linear relationships and complex feature interactions, tends to deliver lower MSE and MAE values and improved R^2 scores, particularly evident in endpoints such as LogD where the *seyonec/PubChem10M_SMILES_BPE_450k* embedding attains an R^2 of 0.6990. This comparative analysis underlines the advantage of ensemble techniques over deep MLP architectures when applied to similar ChemLM-derived inputs.

The FP+GB setup, which employs traditional fingerprint representations like ECFP and MACCS with Gradient Boosting, consistently demonstrates robust performance. For example, ECFP achieves an R^2 of 0.8328 for MLM, and similar strong results are observed for KSOL and MDR1-MDCKII. The MACCS fingerprints also show competitive metrics, although with slight variations; for instance, MACCS yields an R^2 of 0.8250 for MLM and 0.6221 for KSOL. These findings indicate that traditional chemical fingerprints, when combined with powerful ensemble methods, can sometimes outperform or rival the more modern ChemLM embeddings. The FP+GB results emphasize the value of well-established molecular representations, as they provide reliable predictions across multiple ADMET endpoints. Such insights are crucial for guiding future methodological choices in computational drug discovery.

Overall, the comparative analysis reveals that model performance is highly contingent upon both the chosen representation and the regression method. While the ChemLM+MLP setup exhibits significant variability and occasionally subpar performance, the ChemLM+GB approach generally offers improved stability and accuracy. The FP+GB setup, leveraging traditional fingerprints, stands out for its robustness and superior performance in several key endpoints. These results suggest that, in the context of ADMET prediction, ensemble methods like Gradient Boosting can better exploit the information contained in both modern and classical molecular representations. Future work could explore hybrid approaches that combine the strengths of ChemLM embeddings and traditional fingerprints to further enhance prediction reliability and clinical relevance.

Choosing Ensemble Model Weights for ADMET Challenge The model weights in the table 2 reflect the importance or contribution of each model to the prediction for different targets in the ADMET challenge, based on two primary metrics: MSE (Mean Squared Error) and R^2 (R-squared) score. These weights indicate the relative influence each model or fingerprint has in the overall performance for each target. For example, in the MLM (multi-layer model) target, the ECFP and MACCS fingerprints have high weights of 0.45 each, indicating that they are significantly contributing to the predictions. Conversely, models like *seyonec/PubChem10M_SMILES_BPE_450k-GB* have a weight of 0.00, suggesting minimal or no impact on the MLM target. This weight distribution helps highlight which models and fingerprints are most effective for each specific task, guiding further optimization efforts in ADMET predictions.

5.2. Potency Challenge

Table 3 presents a detailed comparison of different models and fingerprints used for predicting pIC50 values for MERS-CoV Mpro and SARS-CoV-2 Mpro in the Potency Challenge. Three primary setups—ChemLM+MLP, ChemLM+GB, and FP+GB—are evaluated for their performance across multiple models and fingerprints. The evaluation metrics include MSE (Mean Squared Error), MAE (Mean Absolute Error), and R^2 Score. These metrics are critical in assessing the accuracy and reliability of the models in predicting the potency of compounds against these viral targets.

In the ChemLM+MLP setup, the models based on the *seyonec/PubChem10M_SMILES_BPE_450k* fingerprint perform the best for both MERS-CoV and SARS-CoV-2 Mpro. For MERS-CoV Mpro, it achieves an R^2 score of 0.2507, which is notably higher than the other models. Similarly, for SARS-CoV-2 Mpro, it achieves an R^2 of 0.4571, indicating better predictive power compared to the other models. In contrast, the *unikei/bert-base-smiles* model performs relatively poorly for MERS-CoV Mpro (with a negative R^2 score of -0.1889) but shows better performance for SARS-CoV-2 Mpro, with an R^2 of 0.5125. This suggests that the *seyonec/PubChem10M_SMILES_BPE_450k* model is more suited for the task than other models, particularly for SARS-CoV-2.

The ChemLM+GB setup reveals a shift in performance, especially with the *unikei/bert-base-smiles* model, which demonstrates significantly improved performance compared to the MLP setup. For instance, its R^2 score for SARS-CoV-2 Mpro rises to 0.5560, a notable improvement. In comparison, the ibm-research/MoLFormer-XL-both-10pct and *seyonec/PubChem10M_SMILES_BPE_450k* models exhibit consistently lower R^2 scores across both targets. This suggests that the addition of gradient boosting in the ChemLM+GB setup enhances the model's generalization ability, particularly for SARS-CoV-2 Mpro. However, the other models like MoLFormer and PubChem still underperform relative to unikei/bert-base-smiles.

The FP+GB setup focuses on specific molecular fingerprints (ECFP and MACCS) and their performance when paired with gradient boosting. Interestingly, the ECFP fingerprint yields a strong R^2 score of 0.6524 for SARS-CoV-2 Mpro, outshining other fingerprints. This is a stark contrast to its performance for MERS-CoV Mpro, where its R^2 score is much lower (0.0562). MACCS fingerprints, while also showing moderate performance for both targets, do not match the predictive capability of ECFP for SARS-CoV-2. These results highlight the importance of choosing the appropriate molecular fingerprint, with ECFP proving to be more effective for SARS-CoV-2.

From a model comparison perspective, it is evident that no single model outperforms across all setups and targets. *seyonec/PubChem10M_SMILES_BPE_450k* stands out for MERS-CoV Mpro in the MLP setup, while *unikei/bert-base-smiles* with gradient boosting excels for SARS-CoV-2 Mpro. Additionally, traditional molecular fingerprints like ECFP and MACCS, when paired with gradient boosting, show competitive performance, especially in the FP+GB setup for SARS-CoV-2. The variance in performance underscores the complexity of predicting potency across different viral targets, as well as the need for tuning models and fingerprints to the specifics of the task at hand.

Overall, the results from the table suggest that model selection and setup optimization are crucial factors in achieving high predictive accuracy for drug potency. While deep learning-based models such as BERT and MoLFormer show promise, simpler approaches like using ECFP or MACCS fingerprints with gradient boosting can be highly effective, particularly for specific targets like SARS-CoV-2. Future research should focus on further fine-tuning these models, exploring hybrid setups, and potentially incorporating additional data sources to improve the robustness and generalizability of these predictive models.

Choosing Ensemble Model Weights for Potency Challenge Table 4 presents the model weights assigned for predicting pIC50 values for MERS-CoV Mpro and SARS-CoV-2 Mpro based on two performance metrics: Mean Squared Error (MSE) and R^2 score. These weights indicate the relative importance of each model and fingerprint in the final prediction ensemble. For MERS-CoV Mpro, the model *seyonec/PubChem10M_SMILES_BPE_450k*-MLP is given the highest weight of 0.6, suggesting that it contributed most significantly to the final predictions, followed by *unikei/bert-base-smiles*-GB with a weight of 0.2. Other models like ibm-research/MoLFormer-XL-both-10pct-GB, ECFP, and MACCS have much lower weights, around 0.05, indicating they contributed less to the final model.

For SARS-CoV-2 Mpro, the highest weight is assigned to the ECFP fingerprint (0.4), followed closely by MACCS (0.35), indicating these fingerprints had a significant role in the predictions. Interestingly, the *unikei/bert-base-smiles*-GB model has a weight of 0.0, reflecting that it did not contribute meaningfully to the ensemble's prediction. This suggests that for SARS-CoV-2, simpler molecular representations like ECFP and MACCS may perform better than more complex models like BERT-based approaches. The weights reflect how each model's performance, based on MSE and R^2, translates into its contribution to the final prediction ensemble.

6. Concluding Remarks

In summary, the machine learning workflow described herein for predicting ADMET and Potency properties of molecules is a comprehensive, multi-modal approach. It begins with the preparation of molecular data, primarily in the form of SMILES strings. These SMILES representations are then used to generate two distinct types of molecular features: embeddings derived from three different pre-trained transformer models (*unikei/bert-base-smiles*, *ibm-research/MoLFormer-XL-both-10pct*, and *seyonec/PubChem10M-SMILES-BPE-450k*) and molecular fingerprints (ECFP and MACCS) generated using RDKit. Missing values in the target ADMET properties are handled through imputation using a GradientBoostingRegressor. Subsequently, Multi-Layer Perceptron (MLP) models are trained using the molecular embeddings, and GradientBoostingRegressor models are trained using the molecular fingerprints to predict the ADMET properties. Finally, the predictions from these diverse models are combined using a weighted ensemble approach to yield the final ADMET and Potency property predictions.

This methodology leverages the strengths of different molecular representations and machine learning models. The transformer-based embeddings capture contextual information from the SMILES strings, while the molecular fingerprints provide fixed-length representations based on structural features and atom environments. The use of both MLP and Gradient Boosting models allows for capturing potentially different types of relationships between the molecular features and the ADMET properties. The final weighted ensemble aims to improve the prediction accuracy and robustness by combining the outputs of these diverse models. Such a comprehensive workflow has significant potential for efficient and accurate prediction of ADMET properties early in the drug discovery process. By identifying promising drug candidates with favorable ADMET profiles, this approach can contribute to reducing the high failure rates associated with poor drug properties, ultimately accelerating the development of new and effective therapeutics.

Future research could focus on further optimizing this workflow. This could involve exploring other advanced pre-trained transformer architectures and fine-tuning strategies for molecular representation. Additionally, more sophisticated hyperparameter optimization techniques could be applied to fine-tune the MLP and Gradient Boosting models. Investigating alternative ensemble methods, such as using a meta-learner to combine the predictions, could also lead to further improvements in predictive performance. Finally, evaluating the methodology on a broader range of ADMET endpoints and comparing its performance to other state-of-the-art approaches would be valuable for further validating its effectiveness.

References

[1] Zachary Fralish, Ashley Chen, Paul Skaluba, and Daniel Reker. Deepdelta: predicting admet improvements of molecular derivatives with deep learning. *Journal of Cheminformatics*, 15(1), October 2023. ISSN 1758-2946. doi: 10.1186/s13321-023-00769-x. URL http://dx.doi.org/10.1186/s13321-023-00769-x.

[2] Lei Jia and Hua Gao. *Machine Learning for In Silico ADMET Prediction*, page 447–460. Springer US, November 2021. ISBN 9781071617878. doi: 10.1007/978-1-0716-1787-8_20. URL http://dx.doi.org/10.1007/978-1-0716-1787-8_20.

[3] Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-scale chemical language representations capture molecular structure and properties, 2021. URL https://arxiv.org/abs/2106.09553.

[4] Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12):1256–1264, 2022. doi: 10.1038/s42256-022-00580-7.

[5] Kyle Swanson, Parker Walther, Jeremy Leitz, Souhrid Mukherjee, Joseph C Wu, Rabindra V Shivnaraine, and James Zou. Admet-ai: a machine learning admet platform for evaluation of large-scale chemical libraries. *Bioinformatics*, 40(7), June 2024. ISSN 1367-4811. doi: 10.1093/bioinformatics/btae416. URL http://dx.doi.org/10.1093/bioinformatics/btae416.