RESEARCH ARTICLE

Ensemble Model with BERT, RoBERTa and XLNet for Molecular Property Prediction

Junling Hu[1,*]

[1] School of Intelligent Manufacturing, Chongqing Jianzhu College, Chongqing 400072, China

Abstract

Molecular property prediction is a fundamental task in drug discovery and materials science, yet most high-performing approaches depend on large-scale pretraining that demands substantial computational resources. This work proposes a pretraining-free ensemble framework that trains multiple Transformer-based architectures—BERT, RoBERTa, and XLNet—from random initialization using the Atom-in-SMILES (AIS) molecular representation, which provides richer atomic-level semantics than conventional SMILES. The three Transformer encoders are coupled with BiLSTM prediction heads and integrated via a BaggingRegressor to reduce variance and improve generalization. Experiments on the ZINC250k and ZINC310k benchmarks demonstrate that the proposed framework achieves competitive performance against pretrained baselines including GROVER, CHEM-BERT, and D-MPNN, while requiring only task-specific end-to-end training with adaptive early stopping. These results establish that carefully designed molecular representations combined with heterogeneous ensemble learning can serve as a practical and resource-efficient alternative to pretraining-based paradigms in molecular modeling.

Keywords: ensemble learning, BERT, RoBERTa, XLNet, molecular property prediction.

1 Introduction

Molecular properties play a pivotal role in various fields including materials, chemistry, drug discovery, and healthcare. They connect disciplines such as quantum mechanics, physical chemistry, biophysics, and physiology. In recent years, deep learning models have rapidly expanded their application in the field of chemical science [1], including areas such as drug discovery, molecular property prediction, and molecular representation learning [2–4]. Computer-assisted methods facilitate the rapid prediction of molecular properties [5].

Traditional machine learning approaches for molecular property prediction typically rely on hand-made features, such as Extended-Connectivity Fingerprints (ECFP) [6]. While effective to some extent, these methods often require extensive domain expertise for feature engineering and become increasingly inefficient as the volume of molecular data grows. Moreover, manually designed features may fail

Submitted: 03 May 2026
Revised: 13 June 2026
Accepted: 06 July 2026
Published: 06 August 2026

Vol. 3, **No.** 3, 2026.
10.62762/TETAI.2026.604672

***Corresponding author:**
✉ Junling Hu
jlinghu6@gmail.com

Citation
Hu, J. (2026). Ensemble Model with BERT, RoBERTa and XLNet for Molecular Property Prediction. *ICCK Transactions on Emerging Topics in Artificial Intelligence*, 3(3), 170–187.

to capture the complex structural and chemical information inherent in molecules, leading to performance saturation and reduced prediction accuracy. To address these limitations, deep learning has been increasingly adopted in drug discovery to automatically learn informative representations directly from molecular data [7], thereby improving prediction performance without the need for manual feature construction.

Currently, various deep learning models have been proposed for molecular property prediction, leveraging different molecular representations. Common approaches are based on one-dimensional sequences, such as SMILES [8, 9], or two-dimensional molecular graphs [10–13]. Sequence-based models treat SMILES strings similarly to natural language and benefit from advancements in NLP, particularly the introduction of the Transformer model [14] and the BERT architecture [15]. These models often employ pretraining strategies such as masked language modeling, enabling them to capture rich contextual information within molecular sequences. Graph models typically capture structural features by constructing a graphical representation of molecules [16, 17], where nodes represent atoms and edges represent chemical bonds between atoms. Randomly masking parts of a graph and training the model to predict the attributes or relationships of these masked parts is a key aspect of self-supervised learning. Regardless of the molecular representation chosen, pretraining the input model is crucial [18]. Pretraining models usually require substantial computational resources and large datasets, which is very time-consuming [19]. This presents significant challenges not only for environments with limited computational resources, but also for scenarios in which only small amounts of labeled molecular data are available [20]. Additionally, pretraining may lead to excessive optimization for specific tasks, limiting its applicability in processing small-scale or specific types of datasets.

Therefore, Transformer encoder architectures are trained from scratch in this study, with all parameters randomly initialized and optimized end-to-end together with downstream components. Since large-scale pretraining requires substantial computational resources and data, the present work investigates whether competitive molecular property prediction performance can be achieved under more constrained conditions. To overcome the limitations of individual models and compensate for the absence of pretraining, an ensemble learning strategy [21, 22] is adopted, integrating BERT [15], RoBERTa [23], and XLNet [24] for molecular property prediction. Existing studies have demonstrated that encoder-only attention-guided Transformer frameworks can effectively capture complex latent patterns and contextual dependencies in sequential data through bidirectional self-attention mechanisms [25]. In addition, BiLSTM is introduced as a core component of each base predictor, given that bidirectional recurrent encoders have been shown to effectively extract structure–property relationships directly from SMILES syntax by aggregating contextual information in both the forward and reverse directions [26], which is well suited to SMILES representations that lack a strict reading order. Transformer architectures have demonstrated strong capability in molecular property prediction, achieving state-of-the-art performance that surpasses specialized graph neural networks on established molecular benchmarks [27], confirming the general effectiveness of attention-based architectures across both graph-structured and sequence-based molecular representations. Hybrid designs that couple a Transformer encoder with a downstream neural predictor have further achieved strong performance in QSAR-style molecular property regression [28], motivating the use of a BiLSTM head on top of each Transformer backbone. Finally, ensemble learning is applied to integrate the predictions of the Transformer-based models and BiLSTM predictors using a BaggingRegressor [29]. The ensemble helps improve robustness and generalization. Although this introduces additional computational cost compared with a single model, it avoids the much higher expense of large-scale pretraining, offering a practical trade-off between accuracy and efficiency for resource-constrained molecular property prediction tasks.

2 Relevant Work

Tokenization is a crucial preprocessing step in NLP, significantly impacting the quality of predictions. Therefore, the first key issue is how to represent molecules. With the rapid development of NLP models, particularly those in the Transformer family, tokenizers can encode words or sentences. This allows the conversion of one-dimensional SMILES information into a tokenized language understandable by machines. Tokenizers use byte-pair encoding to construct the vocabulary for model inputs. Choosing a one-dimensional approach like SMILES as the input for NLP models has significant advantages

compared to two-dimensional methods. SMILES [30] is a character encoding system used to represent chemical molecules. It transforms complex molecular structures into one-dimensional string representations by sequentially depicting atoms within the molecule. This transformation is achieved by applying a depth-first search algorithm to the molecular graph, generating a linear character sequence that reflects the molecular structure. This approach not only simplifies the model's processing flow but also significantly reduces computational complexity. Due to its structural similarity to sentences in human language, the SMILES format enhances data interpretability, allowing NLP methods to be effectively applied in chemical data analysis. Hence, the SMILES format enables deep learning-based models to more effectively capture fundamental molecular features and generate accurate molecular property predictions.

However, previous studies have indicated limitations in SMILES representation. Different carbon atoms in a molecule may have different relationships with other atoms and occupy different positions, potentially corresponding to different properties. In SMILES, atoms of the same element with different properties are represented in the same way. Therefore, relying solely on SMILES for molecular property prediction is inaccurate. This problem is viewed as a challenge; comprehensive reviews of molecular representation strategies in AI-driven drug discovery have systematically documented these limitations of traditional SMILES, motivating the development of more informative molecular encodings [31]. DeepSMILES [32] increases the probability of generating valid molecules by introducing closing brackets or single symbols at cyclic positions. SELFIES [33] proposes a different molecular representation based on Chomsky Type-2 grammar, introducing a grammar-based molecular representation framework significantly different from traditional SMILES.

Furthermore, Ucak et al. [34] introduced the Atom in SMILES (AIS) method, eliminating ambiguity in property generation from SMILES representations. This formalized AIS description provides comprehensive atomic and environmental details, converting SMILES into nuanced atom-level representations, enhancing understanding of molecular structure and properties. From their research, AIS outperforms other SMILES tokenization methods in prediction accuracy, enabling sequence-based models to effectively utilize high-quality SMILES representations.

Selecting appropriate molecular representations and aligning them with suitable predictive models remain a central challenge in molecular property prediction. A variety of machine learning approaches have been explored for this task, including LSBoost [35], Support Vector Regression (SVR) [36], regression trees [37], Gaussian Process Regression (GPR) [38], and neural networks (NN). LSBoost improves predictive robustness on high dimensional molecular descriptors through ensemble learning. SVR models nonlinear structure property relationships via kernel based mappings. Regression trees provide an interpretable framework for capturing hierarchical molecular patterns. GPR enables probabilistic modeling with reliable uncertainty quantification for noisy experimental data. NN learn complex nonlinear representations directly from molecular descriptors or sequences, offering strong expressive capacity for modeling intricate structure property relationships.

In the evolving field of molecular property prediction, the Transformer architectures have driven their widespread adoption in molecular language modeling and representation learning. Recent work by Ross et al. [39] introduced the MoLFormer model, which utilizes rotary positional embeddings and a linear attention mechanism to process SMILES sequences of over 1.1 billion molecules from the PubChem and ZINC datasets. This approach has demonstrated state-of-the-art (SOTA) performance across several benchmark datasets, surpassing existing supervised and self-supervised models, including graph neural networks. The results suggest that large-scale molecular language models, such as MoLFormer, hold significant potential in capturing intricate molecular details and advancing scientific discovery, particularly in predicting quantum-chemical properties.

Parallel to the work by Born and Manica [40], which introduced the Regression Transformer as a novel blend of regression analysis with conditional generation tasks, recent advances have further explored this integration. The RT, building on the XLNet architecture, abstracts regression as a conditional sequence modeling problem, thus seamlessly bridging sequence regression and conditional sequence generation. This approach has not only matched but often surpassed existing models in property prediction tasks across small molecules, proteins, and chemical reactions. Furthermore, the RT's ability to integrate both numeric and text

tokens highlights its potential in molecular science, particularly in property-driven, local exploration of chemical or protein spaces.

Wang et al. [41] introduced SMILES-BERT, a semi-supervised model leveraging the power of unsupervised pretraining to tackle the challenges of limited labeled data in molecular property prediction. This model, built on a Transformer architecture with an attention mechanism, undergoes pretraining on large-scale unlabeled molecular data using a Masked SMILES Recovery task. Through this process, SMILES-BERT effectively learns to capture intricate chemical structures and properties. The pre-trained model can then be fine-tuned for various molecular property prediction tasks, demonstrating remarkable performance. Specifically, SMILES-BERT has set new benchmarks across datasets such as QM9 and ESOL, outperforming SOTA methods and underscoring its generalization capability. Its ability to decode complex chemical information positions SMILES-BERT as a critical tool for advancing research in drug discovery and materials science.

Furthering this trajectory, Yu et al. [42] presented SolvBERT, a novel model tailored specifically for predicting solvation properties, a key concept in physical and organic chemistry. SolvBERT uniquely processes solute and solvent interactions through their combined SMILES representations. By pretraining on a vast database of computational solvation free energies in an unsupervised manner, the model captures essential features of solvation dynamics. Upon fine-tuning, SolvBERT demonstrates versatility by accurately predicting experimental solvation free energy and solubility, a capability not previously achieved by graph-based models. This work signifies a considerable step forward in leveraging NLP models for complex molecular property predictions.

Similarly, Li and Jiang [43] introduced Mol-BERT, a model that capitalizes on a vast corpus of SMILES strings to push the boundaries of molecular property prediction. By pretraining on millions of unlabeled drug SMILES from datasets like ZINC15 and ChEMBL27, Mol-BERT effectively learns molecular substructure embeddings tailored for property prediction. The model is then fine-tuned on various tasks, consistently outperforming traditional and sequence-based methods. In particular, Mol-BERT demonstrated superior accuracy across multiple benchmarks, achieving at least a 2% improvement in ROC-AUC scores on datasets such as Tox21, SIDER, and ClinTox. This underscores Mol-BERT's capability to generalize across diverse molecular datasets, solidifying its role as a powerful tool in the domain of drug discovery.

Completing this panorama of innovation, Liu et al. [44] expanded on these foundations with MolRoPE-BERT, a sophisticated model that incorporates Rotary Position Embedding (RoPE) to enhance the encoding of position information within SMILES sequences. By integrating this novel position encoding approach with a pretrained BERT model, MolRoPE-BERT improves the extraction of molecular substructure information crucial for property prediction. The model was trained on a substantial dataset of four million unlabeled drug SMILES, including data from ZINC 15 and ChEMBL 27. Rigorous evaluation across four well-established datasets demonstrates that MolRoPE-BERT achieves performance that is either comparable to or exceeds that of conventional and SOTA methods.

Despite the remarkable progress achieved by SMILES-based pretrained language models such as SolvBERT, Mol-BERT, and MolRoPE-BERT, most existing approaches rely on a single backbone architecture and optimize performance through larger-scale pretraining or architectural refinements. While effective, these single-model paradigms remain vulnerable to model-specific biases and instability, and they rarely exploit the complementary strengths of heterogeneous transformer architectures. Ensemble learning has demonstrated strong advantages in general machine learning, it remains relatively underexplored in molecular language modeling, where most studies focus on optimizing single pretrained models. This work addresses this gap by integrating multiple Transformer architectures trained from random initialization, achieving robust performance gains through cross-model complementarity rather than increased pretraining cost or architectural complexity.

3 Methodology

3.1 DataSet

This project maintains two datasets: ZINC250k and ZINC310k. ZINC250k is a subset of the zinc12 dataset [45], which contains 250,000 organic molecules. Each molecule is provided with a SMILES and two properties. The dataset includes real values for the log octanol-water partition coefficient (logP), which is a measure of lipophilicity and indicates how hydrophobic a compound is. Additionally, each

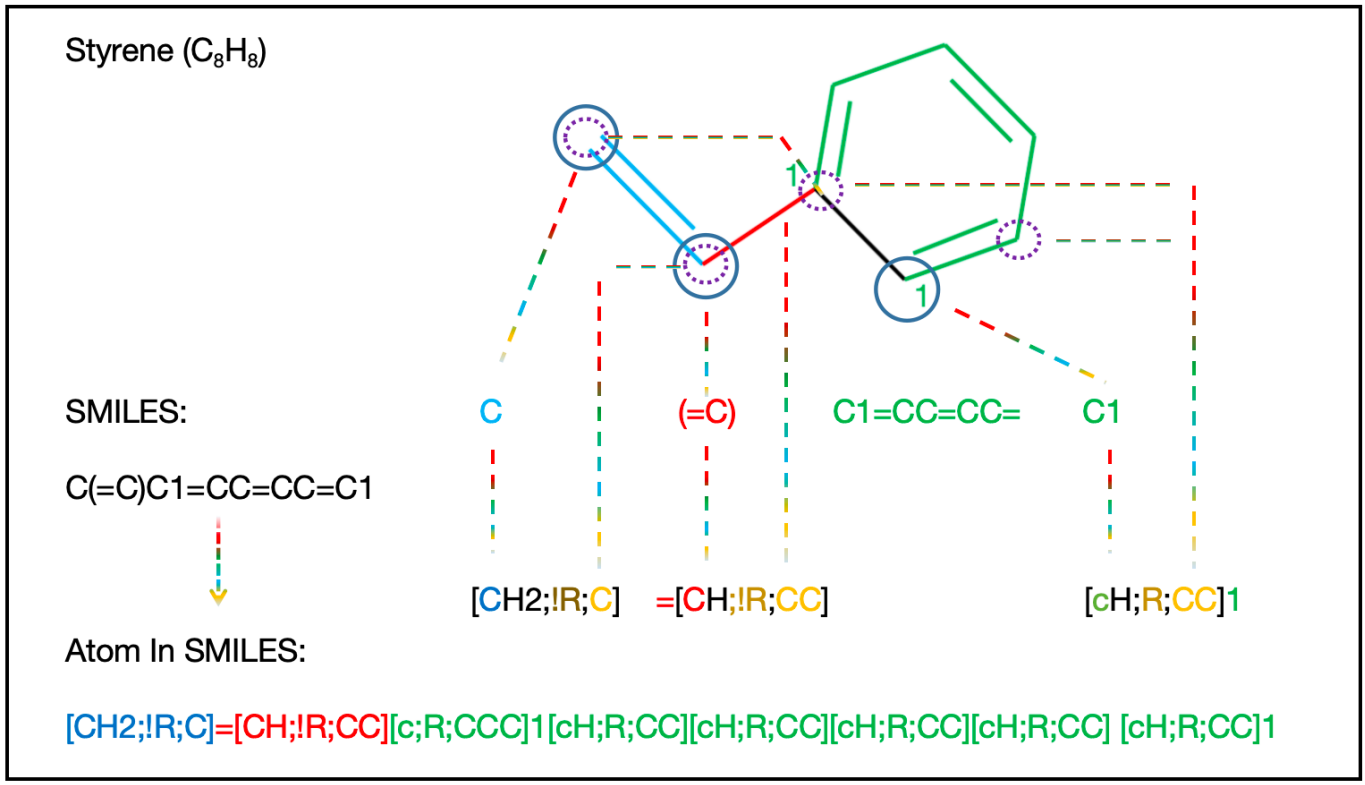

Figure 1. An example illustrates the SMILES expression of the styrene molecule (C(=C)C1=CC=CC=C1) and the step-by-step transformation process into AIS.

molecule is scored with a quantitative estimate of drug-likeness (QED), which reflects the molecule's potential to be a drug based on its physicochemical properties; QED values in this dataset range from 0.11 to 0.95. ZINC310k is a derivative of the complete zinc15 dataset [46], featuring 310,000 molecules. Similar to ZINC250k, each entry in this dataset is associated with a SMILES representation, QED score, and logP value. Furthermore, this dataset includes the molecular weight (MW) of each molecule, which is the mass of a single molecule of a substance and is typically measured in atomic mass units (g/mol). In ZINC310k, QED values span from 0.07 to 0.94, logP values range from -1.99 to 4.99, and MW ranges from 200 to 500 g/mol, providing a diverse set of molecules for the analysis of physicochemical and structural properties relevant to drug discovery.

3.2 Data Preprocessing

AIS is selected as the input representation, and SMILES strings are converted into AIS format (Figure 1). This conversion involves three key elements: the central atom, ring information, and neighboring atoms interacting with the central atom, enclosed within square brackets and separated by semicolons. The central atom's representation includes details about the corresponding SMILES atom, along with the count of neighboring hydrogen atoms. !R denotes the atom's exclusion from a ring, whereas R signals its inclusion in a ring. In cases where an atom is part of a benzene ring, its representation employs lowercase letters to indicate aromaticity. The final element of AIS illustrates atoms adjacent to the central atom. AIS ensures a direct mapping of represented atoms to those in the original SMILES, maintaining consistency with non-atomic symbols. Chirality information can be attached to the central atom using @ or @@ suffixes.

This formalized AIS representation offers a simplified and systematic approach for addressing the limitations of traditional SMILES tokenization, thereby enhancing the predictive quality of molecular property prediction. Upon converting SMILES into AIS, each individual atom and special symbol is treated as a separate token to construct the vocabulary (Figure 2). Analysis of the two datasets reveals that conventional SMILES tokenization yields approximately 80 distinct tokens, whereas AIS tokenization expands the vocabulary to approximately 1,000 tokens. This substantial increase reflects the greater richness and granularity of AIS in capturing molecular structural details, rendering the molecular representation more analogous to natural

Index	Token
1	[PAD]
2	[UNK]
3	[CLS]

...

Index	Token
23	[CH3;!R;C]
24	[C;!R;CCCC]
25	[c;R;CCC]

...

Index	Token
1026	[S;R;CCN]
1027	[[N+];!R;CO]
1028	[S;R;NNN]

Index	Token
1	[PAD]
2	[UNK]
3	[CLS]

...

Index	Token
6	C
7	O
8	c

...

Index	Token
83	[SiH2]
84	[C+]
85	[N]

Figure 2. The vocabularies for AIS(Up) and SMILES(Down) were created based on the ZINC250k and ZINC310k datasets.

language. The expanded AIS vocabulary enables deep learning models to construct more precise descriptions of molecular properties, thereby facilitating more informative feature extraction.

3.3 Ensemble Model

An ensemble model is designed (Figure 3) to predict molecular properties by incorporating BERT, RoBERTa, and XLNet as feature extractors to obtain AIS-based token representations. These representations are subsequently processed by BiLSTM prediction heads, which serve as the core components of the base predictors owing to their inherent sensitivity to sequential data and strong capability in capturing temporal dependencies.

The choice of BiLSTM for processing features extracted from the Transformer encoders is motivated by its effectiveness in capturing sequential dependencies. SMILES sequences, while ordered, can be read in both forward and backward directions without altering their meaning. This characteristic makes BiLSTM's bidirectional processing particularly advantageous. It enables the model to capture contextual information from both directions, thereby enhancing its understanding of molecular structures. In contrast to Convolutional Neural Networks (CNNs), which excel in local feature extraction but may not fully utilize the sequential nature of SMILES, or Deep Neural Networks (DNNs), which do not inherently account for sequential order, BiLSTM integrates information from different parts of the sequence effectively. This bidirectional capability ensures that both past and future contexts are considered, leading to a more nuanced understanding of molecular features.

After processing the features with BiLSTM, the results are aggregated using BaggingRegressor, a robust ensemble learning algorithm for regression tasks. BaggingRegressor is based on the Bagging (Bootstrap Aggregating) principle, which improves overall performance by combining predictions from multiple base learners. This ensemble technique reduces variance and enhances the model's robustness and accuracy.

In summary, by integrating BiLSTM's sequential processing with BaggingRegressor's ensemble approach, the proposed model achieves superior performance in predicting molecular properties, leveraging the strengths of both techniques. The pseudocode of the proposed approach is given in Algorithm 1.

To optimize training speed and efficiency, the number of layers in BERT and XLNet is reduced to 6, and a smaller variant of RoBERTa is employed. All models are trained for 100 epochs with a batch size of 16. Mean absolute error (MAE) is adopted as the loss function, and the learning rate is set to 1×10^{-5}. Detailed hyperparameter configurations are summarized in Table 1.

3.4 Baseline Methods

The proposed ensemble model is evaluated against both representative classical machine learning

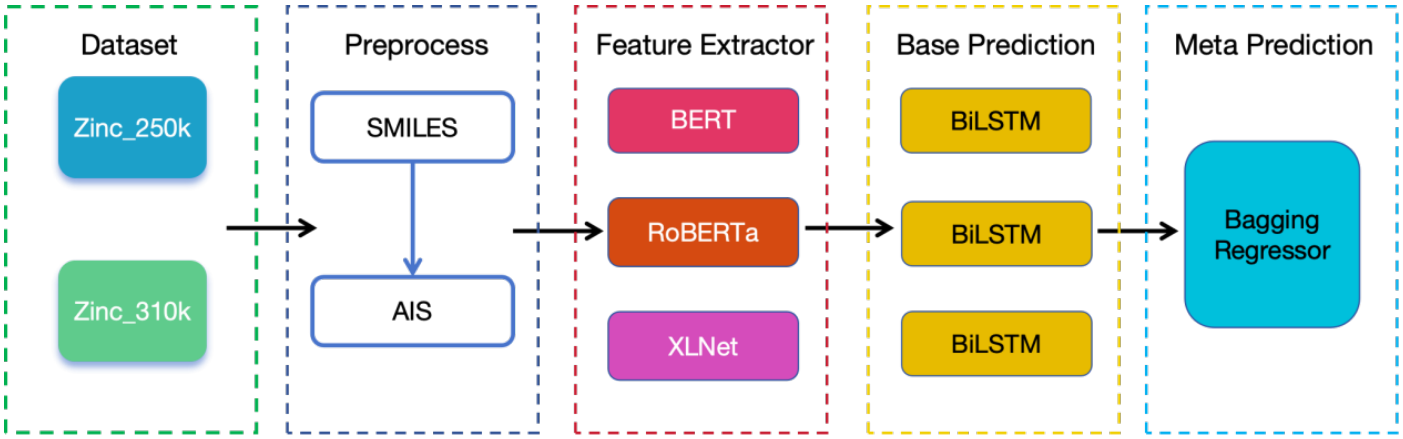

Figure 3. The structure of ensemble model.

Algorithm 1 Ensemble Model for Molecular Property Prediction

Require: SMILES data $\{S_i, Y_i\}_{i=1}^{N}$
Ensure: Predicted properties $\hat{Y}$ and MAE score

for $i = 1$ **to** N **do**
  Tokenize SMILES into structural and atomic symbols;
  Replace atomic tokens with symbolic atom descriptors encoding element type, charge, chirality, ring membership, and local neighborhood context;
  Preserve original structural symbols (e.g., bonds, branches, ring indices) to construct AIS sequence A_i;
  Obtain contextualized embeddings: $F_{\text{BERT}}(A_i)$, $F_{\text{RoBERTa}}(A_i)$, $F_{\text{XLNet}}(A_i)$;
  Predict properties using BiLSTM heads: $P_{\text{BERT}}(A_i)$, $P_{\text{RoBERTa}}(A_i)$, $P_{\text{XLNet}}(A_i)$;
end for
Fuse predictions via BaggingRegressor;
for $i = 1$ **to** N **do**
  $\hat{Y}_i \leftarrow BR\big(P_{\text{BERT}}(A_i),\ P_{\text{RoBERTa}}(A_i),\ P_{\text{XLNet}}(A_i)\big)$;
end for
Compute MAE: $\text{MAE} \leftarrow \frac{1}{N}\sum_{i=1}^{N}|\hat{Y}_i - Y_i|$;
return $\{\hat{Y}_i\}$, MAE;

Table 1. Parameters for model architecture and training.

Parameter Group	Parameter Name	Value
Model Architecture	Activation Function	Gelu
	Num_hidden_layers	6
	Dropout	0.1
	Attention Heads	12
	Hidden Size	768
	BiLSTM Parameters	768×10
	Output Layer	10×1
Training Hyperparameters	Optimizer	Adam
	Epochs	100
	Learning Rate	0.00001
	Batch Size	16
	Loss Function	MAE
Ensemble Parameters	Estimators	200
	Max samples	0.3

methods and advanced deep learning baselines to comprehensively assess its effectiveness.

Random Forest (RF) [47] is a classical ensemble learning method based on bagging of decision trees. It constructs multiple decision trees during training and outputs the mean prediction of individual trees for regression tasks, offering strong robustness and resistance to overfitting on high-dimensional molecular descriptors.

LSBoost is a gradient boosting regression framework [35] based on ensembles of decision trees and has demonstrated strong performance in chemical property prediction and materials engineering applications, particularly under limited data regimes. Its iterative residual correction mechanism enables accurate nonlinear modeling while maintaining computational efficiency.

Regression trees [37] provide an interpretable nonlinear modeling approach and have been widely applied in physical chemistry and materials science for molecular and materials property prediction, benefiting from their ability to capture complex feature interactions.

ASVAE (All SMILES Variational Autoencoder) [48]: ASVAE deploys diversified SMILES representations and employs stacked recursive neural networks to encode individual molecules. By combining

semi-supervised and fully supervised learning, it significantly improves the accuracy of predicting various molecular properties. ASVAE has demonstrated strong performance in predicting attributes such as logP, molecular weight, and drug-likeness (QED), outperforming earlier approaches that learn continuous latent molecular representations from SMILES for property prediction [49, 50]. The same dataset as ASVAE is adopted, with MAE as the primary evaluation metric for comparison, as reported in the original work.

GROVER (Graph Representation frOm self-superVised mEssage passing tRansformer) [51] is a hybrid graph neural network and Transformer architecture designed to optimize molecular graph representations. It integrates dual GNN Transformers for node-level and edge-level feature learning and employs large-scale self-supervised pretraining to enhance downstream task performance. In the present experiments, the publicly available pretrained GROVER model is fine-tuned on the target dataset for fair comparison.

CHEM-BERT [52] exploits SMILES strings as input and focuses on learning molecular representations through BERT-style pretraining. It incorporates matrix embedding layers to enhance structural learning and has demonstrated strong generalization capability across multiple benchmark datasets, particularly for quantitative estimation of drug-likeness and molecular property prediction tasks. A similar fine-tuning protocol is adopted for performance comparison.

D-MPNN (Directed Message Passing Neural Network) [54] operates directly on molecular graphs and extends the message passing neural network framework [53] by propagating information along directed bonds rather than atoms, enabling more expressive structural representation learning. Unlike most Transformer-based models, D-MPNN does not rely on large-scale pretraining and has consistently outperformed traditional descriptor-based models and other graph neural architectures across a wide range of molecular property benchmarks.

By comparing against both classical machine learning models and modern neural architectures, this study systematically evaluates the advantages of the proposed ensemble Transformer framework. Unlike most existing approaches that rely on a single pretrained backbone or handcrafted molecular descriptors, the proposed method leverages complementary Transformer architectures trained from random initialization, enabling improved performance without the need for costly large-scale pretraining pipelines.

3.5 Evaluation Metrics

This study uses four evaluation metrics to assess model performance: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Symmetric Mean Absolute Percentage Error (sMAPE), and the Coefficient of Determination (R^2).

RMSE is the square root of the average of the squared prediction errors and reflects the volatility in the model's predictions:

MAE measures the average absolute difference between predicted and actual values, providing an indication of the typical prediction error:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

sMAPE quantifies the average relative prediction error in percentage form in a symmetric manner, treating overestimation and underestimation equally and providing greater numerical stability than standard MAPE:

$$\mathrm{sMAPE} = \frac{100}{n}\sum_{i=1}^{n}\frac{2\,|y_i - \hat{y}_i|}{|y_i| + |\hat{y}_i|}$$

Finally, R^2 evaluates the proportion of variance in the observed data explained by the model, with values closer to 1 indicating better predictive performance:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

4 Experiment

In this experiment, each dataset was divided into training, validation, and test sets in an 8:1:1 ratio. Figure 4 illustrates the training and validation MAE of the BERT, RoBERTa, and XLNet models across 100 epochs on the ZINC250k and ZINC310k datasets. To improve robustness and reduce the variance introduced by data partitioning, all experiments were conducted using five-fold cross-validation, and the

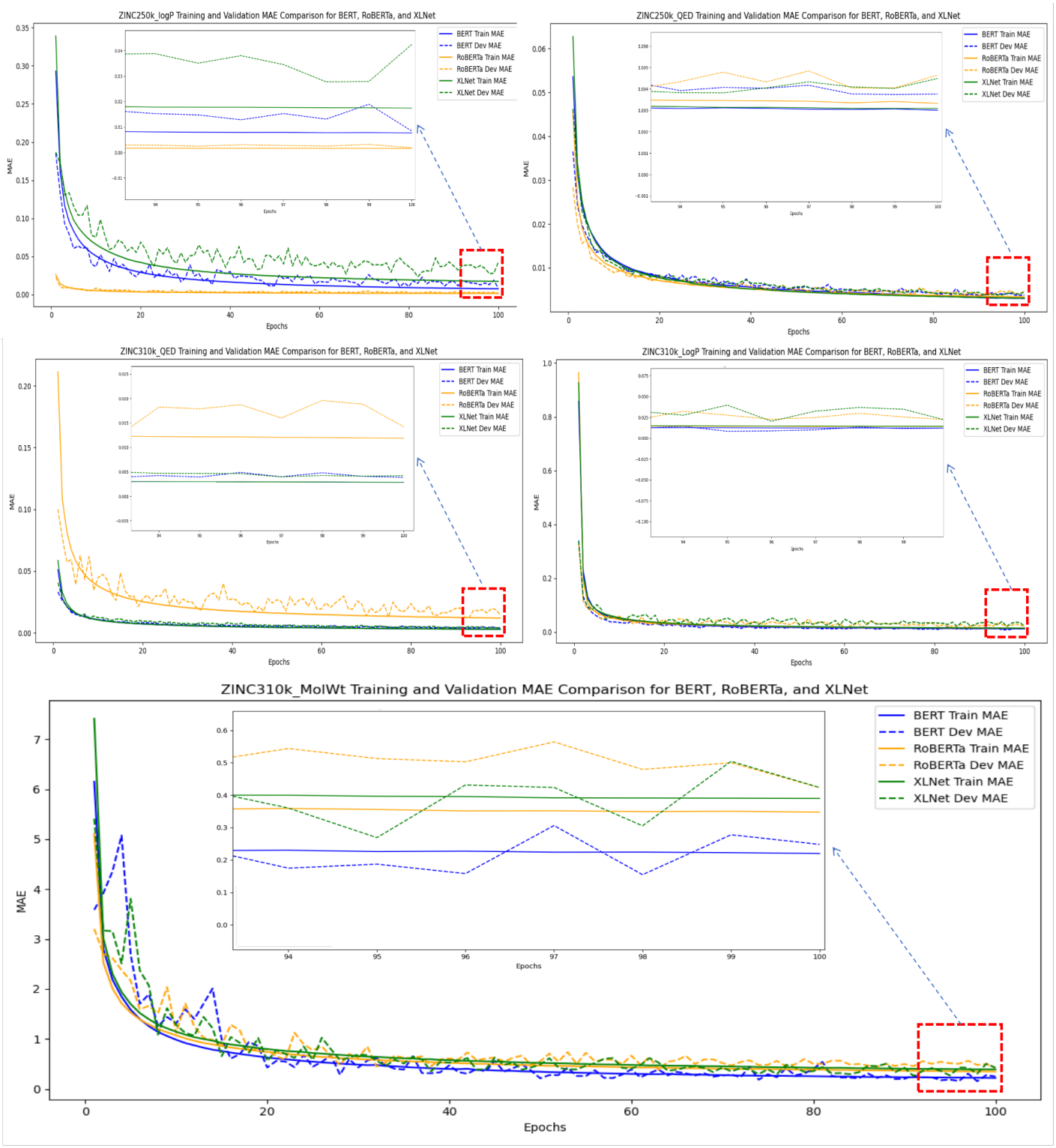

Figure 4. Training and Validation MAE for each model.

reported curves reflect the averaged trends across all folds.

Across both datasets, the training MAE of all models exhibits a clear downward trend as training progresses, indicating effective learning behavior. Among the three architectures, BERT demonstrates comparatively more stable convergence than RoBERTa and XLNet. In the later stages of training, the validation MAE begins to fluctuate without consistent improvement, suggesting the onset of overfitting. Extending the number of training epochs beyond this stage resulted in only marginal reductions in MAE while increasing the risk of overfitting.

To alleviate overfitting and improve training efficiency,

an adaptive early stopping strategy is adopted. During the initial phase of training, early stopping is not activated, and only the validation MAE is recorded at each epoch. Early stopping monitoring is triggered when either the validation MAE fails to improve for three consecutive epochs or the reduction in training MAE falls below 1×10^{-5}. This strategy replaces the conventional use of a fixed epoch threshold.

Once monitoring is activated, the minimum validation MAE achieved is recorded. At each subsequent epoch, two conditions are evaluated: (i) whether the training MAE continues to decrease by more than 1×10^{-6}, indicating ongoing learning, and (ii) whether the validation MAE increases by more than 1×10^{-6} relative to the current minimum, indicating degraded generalization. If both conditions are satisfied for three consecutive epochs, training is terminated and the model corresponding to the minimum validation MAE during the monitoring phase is selected. As a fallback mechanism, if no early stopping criterion is met within 100 epochs, the model achieving the lowest validation MAE across all epochs is retained.

Scatter and histogram plots utilizing predicted and true values offer a direct visual assessment of the model's performance. The final Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) across all attributes in the test set, as presented in Figures 5-6, indicate errors within a tight margin of 0.5%.

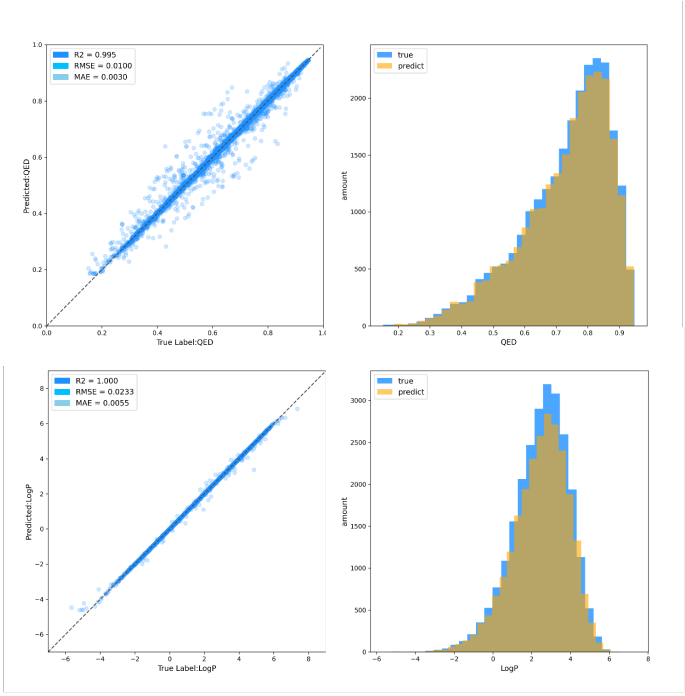

Figure 5. Regressions (left) and distributions (right) of true and predicted values of qed and logP from ZINC250k.

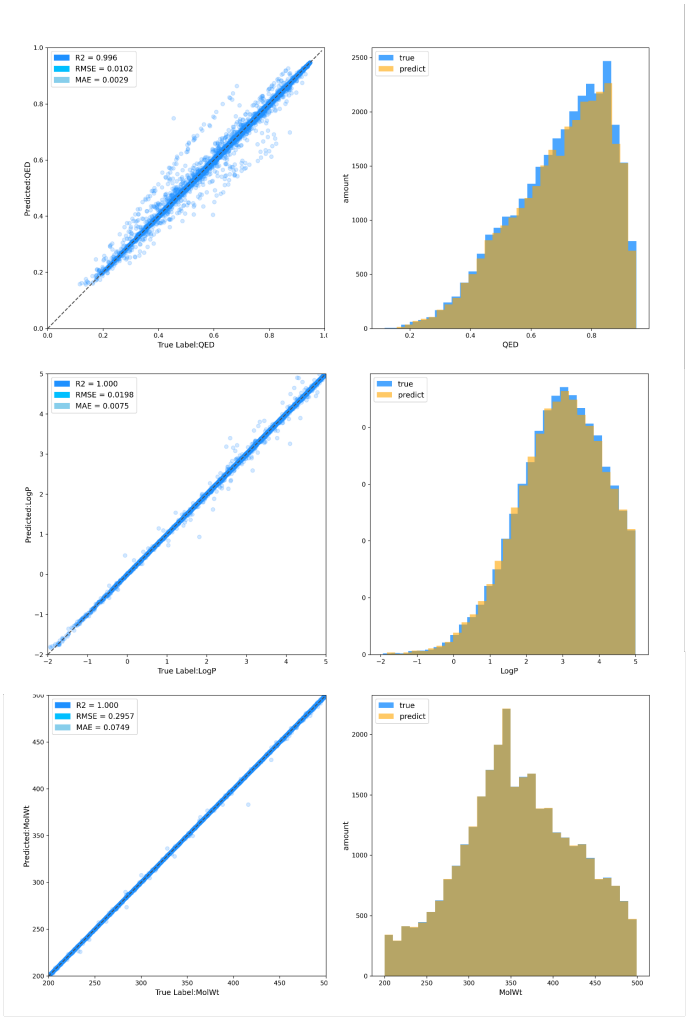

Figure 6. Regressions (left) and distributions (right) of true and predicted values of qed, logP and MolWt from ZINC310k.

Scatter plots exhibit a strong correlation between predictions and actual values, with QED showing an R^2 of 0.996 and logP and MolWt achieving a perfect R^2 of 1.000, suggesting an almost flawless predictive model. The alignment of the true and predicted value distributions in the histograms further confirms the precision of the model.

Tables 2 and 3 report the MAE, RMSE, sMAPE, and R^2 on the ZINC250k dataset. As shown in Table 2, the proposed ensemble model achieves an MAE of 0.003 on the QED property, representing a 42% improvement over the ASVAE baseline and outperforming all other graph-based and sequence-based models. For the logP property, an MAE of 0.006 is achieved, which is comparable to the ASVAE baseline (0.005) but significantly better than all other models.

Tables 4, 5, and 6 report the MAE, RMSE, sMAPE, and R^2 on the ZINC310k dataset. The proposed ensemble

Table 2. The performance of ZINC250k dataset with QED property.

Model	MAE	RMSE	sMAPE(%)	R^2	Molecular representation
RF	0.0791	0.1026	12.33	0.4602	ECFP
LSBoost	0.0639	0.0816	9.89	0.6587	ECFP
ASVAE	0.0052	–	–	–	SMILES
GROVER	0.0056	**0.0083**	0.672	**0.9964**	Graphs
D-MPNN	0.0056	0.0094	0.694	0.9953	Graphs
CHEM-BERT	0.0049	0.0102	0.641	0.9934	SMILES
Ensemble-model	**0.0030±0.0001**	0.0101±0.002	**0.502±0.015**	0.9950±0.0005	AIS

Table 3. The performance of ZINC250k dataset with LogP property.

Model	MAE	RMSE	sMAPE(%)	R^2	Molecular representation
RF	0.7353	0.9348	41.45	0.5737	ECFP
LSBoost	0.4533	0.5856	28.84	0.8327	ECFP
ASVAE	0.005	–	–	–	SMILES
GROVER	0.016	0.036	0.6506	0.9964	Graphs
D-MPNN	0.011	0.029	0.447	**0.9953**	Graphs
CHEM-BERT	**0.005**	0.011	0.2232	0.9934	SMILES
Ensemble-model	**0.006±0.0004**	0.024±0.001	0.2439±0.0163	0.9995±0.0002	AIS

Table 4. The performance of ZINC310k dataset with QED property.

Model	MAE	RMSE	sMAPE(%)	R^2	Molecular representation
RF	0.0851	0.1095	13.11	0.5232	ECFP
LSBoost	0.0688	0.087	10.65	0.6988	ECFP
ASVAE	0.0064	–	–	–	SMILES
GROVER	0.0052	0.0073	0.7478	**0.9977**	Graphs
D-MPNN	0.0059	0.010	0.826	**0.9959**	Graphs
CHEM-BERT	**0.0045**	0.0103	0.6466	0.9948	SMILES
Ensemble-model	0.0029±0.0001	**0.0101±0.0002**	**0.5171±0.012**	0.9958±0.0007	AIS

Table 5. The performance of ZINC310k dataset with LogP property.

Model	MAE	RMSE	sMAPE(%)	R^2	Molecular representation
RF	0.6888	0.8703	29.04	0.4472	ECFP
LSBoost	0.4356	0.5549	19.84	0.7753	ECFP
ASVAE	0.007	–	–	–	SMILES
GROVER	0.009	0.018	0.305	**0.9997**	Graphs
D-MPNN	0.017	0.034	0.5799	**0.9991**	Graphs
CHEM-BERT	**0.008**	0.02	0.5898	0.9997	SMILES
Ensemble-model	**0.007±0.0003**	0.019±0.0003	**0.5653±0.012**	**0.9996±0.0007**	AIS

Table 6. Performance comparison on the ZINC310k dataset for MolWt property prediction.

Model	MAE	RMSE	sMAPE(%)	R^2	Molecular representation
RF	36.5800	46.7800	10.4800	0.5231	ECFP
LSBoost	26.1300	33.5200	7.5500	0.7552	ECFP
ASVAE	0.2100	–	–	–	SMILES
GROVER	0.2600	0.3500	0.0669	**0.9999**	Graphs
D-MPNN	1.5500	3.6300	0.4270	**0.9997**	Graphs
CHEM-BERT	0.4000	0.6300	0.1080	0.9998	SMILES
Ensemble-model	**0.0700±0.016**	**0.2900±0.007**	**0.0216±0.006**	0.9999±0.000	AIS

model achieves an MAE of 0.0029 for QED and 0.07 for MolWt, representing improvements of 54% and 66%, respectively, over the ASVAE baseline. For the logP property, performance remains comparable to the ASVAE baseline. These results further confirm the effectiveness of the proposed ensemble framework in multi-property prediction tasks.

Furthermore, the results indicate that the type of input data significantly influences model performance for the same dataset. Models that use sequences as input perform better overall than models that use graphs as input. Different types of input data, such as SMILES sequences and graph representations, have a notable impact on model performance. Models that use sequence-based inputs (AIS) generally achieve better performance than graph-based models within the proposed ensemble framework. This may be attributed to the richer chemical information provided by the AIS representations, enhancing the learning capacity of the model. The experimental results highlight the excellent performance of the ensemble model in most performance metrics, particularly in terms of MAE. This finding underscores the potential of ensemble methods in improving prediction accuracy.

It is also noted that in certain specific properties, the ensemble model performs comparably to single models, suggesting the need for further exploration of optimal configurations for different model combinations in future research. Since the proposed model is trained from random initialization on the target task, whereas the baseline models undergo large-scale pretraining followed by task-specific adaptation, a direct comparison of computational time is not straightforward. Nevertheless, task-specific training from scratch generally requires substantially less time than large-scale pretraining. Each base model was trained on a single V100 GPU, with the total training time per molecular property not exceeding one day.

5 Ablation Study

The ablation study in this project evaluates the impact of different language models (BERT, RoBERTa, XLNet) in combination with BiLSTM and compares two different input types (AIS and SMILES) on model performance (Tables 7, 8, 9, 10 and 11).

A systematic evaluation of input format reveals that AIS-based models—BERT + BiLSTM, RoBERTa + BiLSTM, and XLNet + BiLSTM—consistently achieve lower MAE and RMSE as well as higher R^2 on both training and test sets, indicating that the richer informational content of AIS enables more effective learning and prediction. This advantage is further reflected in reduced sMAPE values, suggesting that AIS not only lowers absolute prediction errors but also improves scale-normalized accuracy across different molecular property ranges. In contrast, all three architectures experience a noticeable performance drop when fed raw SMILES strings—BERT + BiLSTM(SMILES), RoBERTa + BiLSTM(SMILES), and XLNet + BiLSTM(SMILES)— likely because SMILES encodes comparatively less structural information, which can hinder comprehensive model learning and leads to increased relative errors as captured by sMAPE.

Across the three molecular properties studied (Figure 7), the combined analyses consistently reveal differentiated roles of BERT, RoBERTa, and XLNet within the ensemble model. The correlation heatmaps indicate that predictions from all three Transformer backbones are highly correlated with the ground truth, confirming that each model captures core structure–property relationships. However, the strong inter-model correlations also suggest a certain degree of redundancy, highlighting the importance of analyzing their marginal contributions beyond raw

Table 7. The ablation study of ZINC250k dataset with QED property.

Model	Training				Validation				Test			
	MAE	RMSE	sMAPE (%)	R^2	MAE	RMSE	sMAPE (%)	R^2	MAE	RMSE	sMAPE (%)	R^2
BERT+BiLSTM	0.0030	0.0080	0.4013	0.9953	0.0037	0.0137	0.6215	0.9889	0.0038	0.0144	0.6473	0.9870
RoBERTa+BiLSTM	0.0035	0.0081	0.4577	0.9947	0.0040	0.0130	0.6420	0.9899	0.0039	0.0127	0.6347	0.9899
XLNet+BiLSTM	0.0030	0.0063	0.4654	0.9974	0.0038	0.0108	0.6055	0.9899	0.0039	0.0114	0.6286	0.9918
BiLSTM	0.0054	0.0152	0.8354	0.9860	0.0068	0.0186	1.1019	0.9795	0.0070	0.0194	1.1228	0.9765
BERT+BiLSTM(SMILES)	0.0035	0.0090	0.5401	0.9949	0.0042	0.0141	0.7258	0.9882	0.0041	0.0142	0.7198	0.9874
RoBERTa+BiLSTM(SMILES)	0.0035	0.0083	0.5380	0.9957	0.0040	0.0121	0.6918	0.9913	0.0041	0.0126	0.7038	0.9904
XLNet+BiLSTM(SMILES)	0.0038	0.0092	0.5840	0.9947	0.0045	0.0127	0.7769	0.9904	0.0044	0.0127	0.7725	0.9903
BiLSTM(SMILES)	0.0107	0.0221	1.6431	0.9701	0.0120	0.0250	1.9615	0.9620	0.0118	0.0246	1.9169	0.9627
Ensemble(SMILES)	–	–	–	–	–	–	–	–	0.0035	0.0101	0.6503	0.9958

Table 8. The ablation study of ZINC250k dataset with LogP property.

Model	Training				Validation				Test			
	MAE	RMSE	sMAPE (%)	R^2	MAE	RMSE	sMAPE (%)	R^2	MAE	RMSE	sMAPE (%)	R^2
BERT+BiLSTM	0.0077	0.0203	0.8233	0.9999	0.0085	0.0244	0.8536	0.9996	0.0087	0.0268	0.8455	0.9996
RoBERTa+BiLSTM	0.0174	0.0255	1.9637	0.9996	0.0186	0.0366	0.2010	0.9993	0.0186	0.0346	1.990	0.9994
XLNet+BiLSTM	0.0190	0.0248	3.0448	0.9996	0.0233	0.0412	3.0868	0.9991	0.0233	0.0396	3.1312	0.9992
BiLSTM	0.0183	0.0337	2.2661	0.9992	0.0205	0.0430	2.3488	0.9989	0.0209	0.0460	2.4386	0.9990
BERT+BiLSTM(SMILES)	0.0207	0.0356	2.5594	0.9994	0.0241	0.0425	2.7602	0.9991	0.0240	0.0442	2.7403	0.9996
RoBERTa+BiLSTM(SMILES)	0.0206	0.0372	2.5466	0.9994	0.0203	0.0371	2.3276	0.9994	0.0207	0.0460	2.3908	0.9991
XLNet+BiLSTM(SMILES)	0.0220	0.0367	2.7209	0.9993	0.0263	0.0465	3.0173	0.9989	0.0262	0.0473	3.0080	0.9990
BiLSTM(SMILES)	0.0349	0.0574	4.3168	0.9981	0.0353	0.0624	4.0495	0.9976	0.0340	0.0621	3.8935	0.9978
Ensemble(SMILES)	–	–	–	–	–	–	–	–	0.0148	0.0286	0.1885	0.9995

Table 9. The ablation study of ZINC310k dataset with QED property.

Model	Training				Validation				Test			
	MAE	RMSE	sMAPE (%)	R^2	MAE	RMSE	sMAPE (%)	R^2	MAE	RMSE	sMAPE (%)	R^2
BERT+BiLSTM	0.0028	0.0074	0.4243	0.9963	0.0038	0.0139	0.6643	0.9904	0.0037	0.0132	0.6389	0.9917
RoBERTa+BiLSTM	0.0030	0.0102	0.5044	0.9949	0.0038	0.0142	0.6635	0.9906	0.0037	0.0133	0.6480	0.9916
XLNet+BiLSTM	0.0033	0.0075	0.5573	0.9973	0.0040	0.0115	0.6878	0.9942	0.0040	0.0114	0.6947	0.9938
BiLSTM	0.0046	0.0118	0.3930	0.9926	0.0063	0.0171	0.5161	0.9863	0.0063	0.0168	0.5117	0.9871
BERT+BiLSTM(SMILES)	0.0030	0.0082	0.2563	0.9967	0.0039	0.0119	0.3194	0.9931	0.0039	0.0122	0.3333	0.9930
RoBERTa+BiLSTM(SMILES)	0.0033	0.0089	0.2820	0.9962	0.0040	0.0129	0.3515	0.9920	0.0039	0.0128	0.3462	0.9923
XLNet+BiLSTM(SMILES)	0.0036	0.0093	0.3077	0.9958	0.0044	0.0143	0.3864	0.9902	0.0044	0.0141	0.3906	0.9905
BiLSTM(SMILES)	0.0098	0.0204	0.8376	0.9801	0.0118	0.0237	0.9664	0.9735	0.0107	0.0233	0.8891	0.9741
Ensemble(SMILES)	–	–	–	–	–	–	–	–	0.0033	0.0100	0.6445	0.9959

Table 10. The ablation study of ZINC310k dataset with LogP property.

Model	Training				Validation				Test			
	MAE	RMSE	sMAPE (%)	R^2	MAE	RMSE	sMAPE (%)	R^2	MAE	RMSE	sMAPE (%)	R^2
BERT+BiLSTM	0.0125	0.0181	0.5479	0.9997	0.0083	0.0204	0.5715	0.9996	0.0084	0.0215	0.6095	0.9995
RoBERTa+BiLSTM	0.0172	0.0246	0.9997	0.9994	0.0159	0.0270	1.0403	0.9993	0.0160	0.0277	1.0516	0.9993
XLNet+BiLSTM	0.0147	0.0201	1.267	0.9996	0.0204	0.0300	1.3847	0.9992	0.0204	0.0315	1.3403	0.9991
BiLSTM	0.0170	0.0320	0.4985	0.9990	0.0180	0.0380	0.6186	0.9987	0.0180	0.0370	0.6179	0.9987
BERT+BiLSTM(SMILES)	0.0156	0.0216	0.4575	0.9996	0.0214	0.0314	0.7354	0.9991	0.0162	0.0288	0.5561	0.9992
RoBERTa+BiLSTM(SMILES)	0.0192	0.0269	0.5630	0.9993	0.0265	0.0376	0.9107	0.9987	0.0267	0.0391	0.9166	0.9998
XLNet+BiLSTM(SMILES)	0.0181	0.0249	0.5308	0.9994	0.0276	0.0366	0.9485	0.9988	0.0278	0.0368	0.9543	0.9987
BiLSTM(SMILES)	0.0336	0.0565	0.9853	0.9972	0.0421	0.0675	1.4467	0.9959	0.0343	0.0612	1.1775	0.9967
Ensemble(SMILES)	–	–	–	–	–	–	–	–	0.0122	0.0222	0.8832	0.9996

Table 11. The ablation study of ZINC310k dataset with MolWt property.

Model	Training				Validation				Test			
	MAE	RMSE	sMAPE (%)	R^2	MAE	RMSE	sMAPE (%)	R^2	MAE	RMSE	sMAPE (%)	R^2
BERT+BiLSTM	0.22	0.34	0.0418	0.99	0.15	0.82	0.0438	0.99	0.15	0.34	0.0425	0.99
RoBERTa+BiLSTM	0.34	0.61	0.1186	0.99	0.42	0.95	0.1203	0.99	0.42	0.67	0.1202	0.99
BiLSTM	0.22	0.41	0.0561	0.99	0.22	0.88	0.0602	0.99	0.23	0.47	0.0650	0.99
BERT+BiLSTM(SMILES)	0.23	0.37	0.0586	0.99	0.18	0.30	0.0493	0.99	0.18	0.29	0.0509	0.99
RoBERTa+BiLSTM(SMILES)	0.39	0.55	0.0994	0.99	0.52	0.73	0.1424	0.99	0.52	0.74	0.1470	0.99
XLNet+BiLSTM(SMILES)	0.42	0.56	0.1071	0.99	0.45	0.63	0.1232	0.99	0.45	0.62	0.1272	0.99
BiLSTM(SMILES)	0.34	0.58	0.0867	0.99	0.24	0.53	0.0657	0.99	0.23	0.59	0.1650	0.99
Ensemble(SMILES)	–	–	–	–	–	–	–	–	0.12	0.38	0.0434	0.99

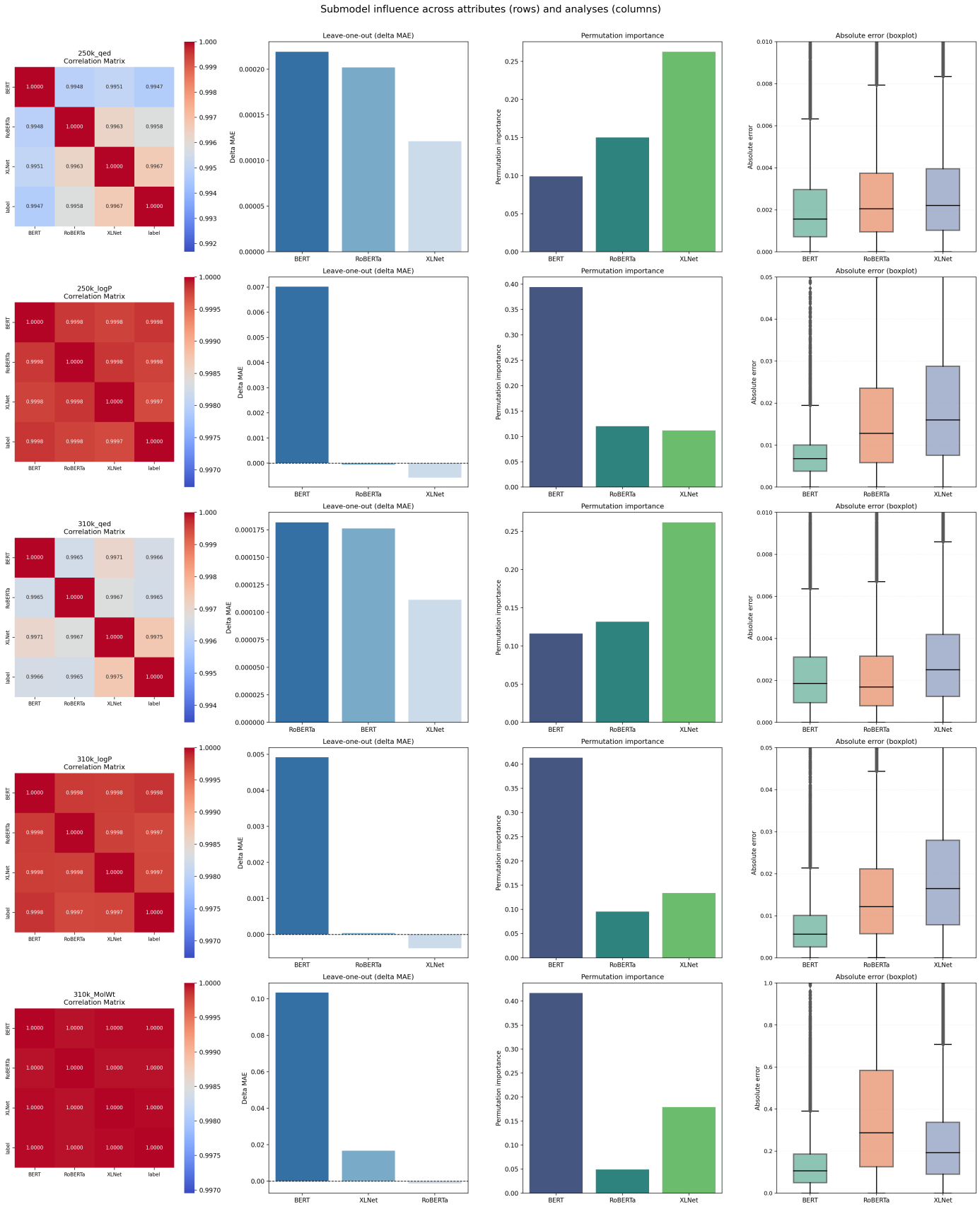

Figure 7. Contribution analysis of base models in the ensemble model.

predictive accuracy.

The leave-one-out experiments provide clearer evidence of each model's contribution to the ensemble. Removing BERT generally leads to the largest increase in MAE, indicating that BERT serves as the most influential backbone in stabilizing ensemble performance. RoBERTa exhibits a moderate but consistent contribution, while excluding XLNet typically results in the smallest degradation and, in some cases, negligible changes, implying a comparatively limited marginal benefit from XLNet.

Permutation importance analysis further corroborates these observations. BERT consistently receives the highest importance scores across properties, suggesting that the meta-learner relies most heavily on BERT's predictions. RoBERTa contributes additional complementary information, whereas XLNet shows lower and less stable importance, reinforcing the conclusion that its role in the ensemble is secondary.

Finally, the absolute error distributions illustrate that BERT-based predictions tend to exhibit tighter error spreads, contributing to improved robustness of the ensemble. RoBERTa shows slightly higher variance, while XLNet often presents broader error distributions, which limits its effectiveness in reducing overall prediction uncertainty.

Overall, these results indicate that the ensemble performance is primarily driven by BERT, with RoBERTa providing supplementary gains, while XLNet contributes relatively little additional information. This finding suggests that future ensemble designs could replace XLNet with lighter or more specialized models to achieve a better trade-off between computational efficiency and predictive performance.

6 Limitations

Despite the competitive performance achieved by the proposed approach, several limitations merit further investigation. First, the present study does not conduct a fine-grained analysis of AIS token usage. Although AIS enriches atomic-level semantics, potential redundancy among tokens and their individual contributions to model performance remain unexplored, which may introduce unnecessary computational overhead. Second, while an adaptive early stopping strategy is employed to improve training efficiency, training three Transformer models from scratch still incurs a non-negligible computational cost. Nevertheless, this cost remains substantially lower than that required by large-scale pretraining-based approaches. Third, the selection of BERT, RoBERTa, and XLNet as base models is primarily motivated by their widespread adoption and strong performance in various downstream tasks, rather than by an exhaustive architecture search tailored to molecular property prediction. Ablation results indicate that XLNet contributes relatively less to the ensemble, suggesting that lighter architectures could potentially replace it to further reduce computational complexity. Finally, model hyperparameters are not systematically optimized in this work, as all experiments rely on general-purpose default settings. Consequently, additional performance gains may be achievable through dedicated hyperparameter tuning in future studies.

7 Discussion and Conclusion

This study demonstrates that effective molecular property prediction can be achieved without relying on extensive pretraining, provided that appropriate molecular representations and model architectures are carefully designed. A systematic comparison of input formats reveals that the AIS representation enables deep learning models to capture molecular features more effectively than raw SMILES, leading to consistently improved predictive performance. These results highlight the critical role of representation learning in molecular modeling tasks.

Furthermore, the ensemble strategy integrating multiple Transformer backbones enhances robustness and accuracy by leveraging complementary predictive patterns. Although the use of BERT, RoBERTa, and XLNet introduces non-negligible computational overhead during end-to-end training from scratch, the overall training cost remains substantially lower than that of large-scale pretraining-based approaches. This balance between performance and efficiency makes the proposed framework particularly suitable for academic environments and small research groups with limited computational resources.

Overall, the present findings confirm that task-specific model design combined with structured molecular representations and ensemble learning offers a practical and scalable alternative to heavily pretrained models. Future work will extend this framework to more complex material systems, such as alloy property prediction, by adapting the AIS methodology to heterogeneous data modalities and integrating lightweight base learners within an ensemble learning paradigm.

Data Availability Statement

The code and data supporting the findings of this study are openly available on GitHub at https://github.com/jlinghu/AIS-Ensemble-model.

Funding

This work was supported without any funding.

Conflicts of Interest

The author declares no conflicts of interest.

AI Use Statement

The author declares that no generative AI was used in the preparation of this manuscript.

Ethical Approval and Consent to Participate

Not applicable. This study is purely computational and does not involve human participants, animal subjects, or identifiable personal data. No ethical approval or informed consent was required.

References

[1] Goh, G. B., Hodas, N. O., & Vishnu, A. (2017). Deep Learning for Computational Chemistry. *Journal of Computational Chemistry, 38*(16), 1291-1307. [CrossRef]

[2] Lavecchia, A. (2019). Deep learning in drug discovery: opportunities, challenges and future prospects. *Drug discovery today, 24*(10), 2017-2032. [CrossRef]

[3] Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., ... & Pande, V. (2018). MoleculeNet: a benchmark for molecular machine learning. *Chemical science, 9*(2), 513-530. [CrossRef]

[4] Li, Z., Jiang, M., Wang, S., & Zhang, S. (2022). Deep Learning Methods for Molecular Representation and Property Prediction. *Drug Discovery Today, 27*(12), 103373. [CrossRef]

[5] Walters, W. P., & Barzilay, R. (2020). Applications of deep learning in molecule generation and molecular property prediction. *Accounts of chemical research, 54*(2), 263-270. [CrossRef]

[6] Rogers, D., & Hahn, M. (2010). Extended-connectivity fingerprints. *Journal of chemical information and modeling, 50*(5), 742-754. [CrossRef]

[7] Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., & Blaschke, T. (2018). The rise of deep learning in drug discovery. *Drug discovery today, 23*(6), 1241-1250. [CrossRef]

[8] Goh, G. B., Hodas, N. O., Siegel, C., & Vishnu, A. (2017). Smiles2vec: An Interpretable General-Purpose Deep Neural Network for Predicting Chemical Properties. *arXiv preprint arXiv:1712.02034*. [CrossRef]

[9] Pinheiro, G. A., Mucelini, J., Soares, M. D., Prati, R. C., da Silva, J. L. F., & Quiles, M. G. (2020). Machine Learning Prediction of Nine Molecular Properties Based on the SMILES Representation of the QM9 Quantum-Chemistry Dataset. *The Journal of Physical Chemistry A, 124*(47), 9854-9866. [CrossRef]

[10] Jo, J., Kwak, B., Choi, H. S., & Yoon, S. (2020). The message passing neural networks for chemical property prediction on SMILES. *Methods, 179*, 65-72. [CrossRef]

[11] Wieder, O., Kohlbacher, S., Kuenemann, M., Garon, A., Ducrot, P., Seidel, T., & Langer, T. (2020). A compact review of molecular property prediction with graph neural networks. *Drug Discovery Today: Technologies, 37*, 1-12. [CrossRef]

[12] Zhang, Z., Liu, Q., Wang, H., Lu, C., & Lee, C. K. (2021). Motif-based graph self-supervised learning for molecular property prediction. *Advances in Neural Information Processing Systems, 34*, 15870-15882.

[13] Jiang, D., Wu, Z., Hsieh, C. Y., Chen, G., Liao, B., Wang, Z., ... & Hou, T. (2021). Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models. *Journal of cheminformatics, 13*(1), 12. [CrossRef]

[14] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems, 30*.

[15] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171-4186). [CrossRef]

[16] Sun, M., Zhao, S., Gilvary, C., Elemento, O., Zhou, J., & Wang, F. (2020). Graph convolutional networks for computational drug development and discovery. *Briefings in bioinformatics, 21*(3), 919-935. [CrossRef]

[17] Xiong, Z., Wang, D., Liu, X., Zhong, F., Wan, X., Li, X., ... & Zheng, M. (2019). Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of medicinal chemistry, 63*(16), 8749-8760. [CrossRef]

[18] Chithrananda, S., Grand, G., & Ramsundar, B. (2020). ChemBERTa: large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*. [CrossRef]

[19] Irwin, R., Dimitriadis, S., He, J., & Bjerrum, E. J. (2022). Chemformer: a pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology, 3*(1), 015022. [CrossRef]

[20] Altae-Tran, H., Ramsundar, B., Pappu, A. S., & Pande,

V. (2017). Low data drug discovery with one-shot learning. *ACS central science, 3*(4), 283. [CrossRef]

[21] Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *Wiley interdisciplinary reviews: data mining and knowledge discovery, 8*(4), e1249. [CrossRef]

[22] Zhou, Z. H. (2021). Ensemble learning. In *Machine learning* (pp. 181-210). Singapore: Springer Singapore. [CrossRef]

[23] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*. [CrossRef]

[24] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems, 32*.

[25] Fabian, B., Edlich, T., Gaspar, H., Segler, M., Meyers, J., Fiscato, M., & Ahmed, M. (2020). Molecular representation learning with language models and domain-relevant auxiliary tasks. *arXiv preprint arXiv:2011.13230*. [CrossRef]

[26] Zheng, S., Yan, X., Yang, Y., & Xu, J. (2019). Identifying structure–property relationships through SMILES syntax analysis with self-attention mechanism. *Journal of chemical information and modeling, 59*(2), 914-923. [CrossRef]

[27] Ying, C., Cai, T., Luo, S., Zheng, S., Ke, G., He, D., ... & Liu, T. Y. (2021). Do transformers really perform badly for graph representation?. *Advances in neural information processing systems, 34*, 28877-28888.

[28] Karpov, P., Godin, G., & Tetko, I. V. (2020). Transformer-CNN: Swiss knife for QSAR modeling and interpretation. *Journal of cheminformatics, 12*(1), 17. [CrossRef]

[29] Breiman, L. (1996). Bagging predictors. *Machine learning, 24*(2), 123-140. [CrossRef]

[30] Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences, 28*(1), 31-36. [CrossRef]

[31] David, L., Thakkar, A., Mercado, R., & Engkvist, O. (2020). Molecular representations in AI-driven drug discovery: a review and practical guide. *Journal of cheminformatics, 12*(1), 56. [CrossRef]

[32] O'Boyle, N., & Dalke, A. (2018). DeepSMILES: An Adaptation of SMILES for Use in Machine Learning of Chemical Structures. *ChemRxiv preprint*. [CrossRef]

[33] Krenn, M., Häse, F., Nigam, A., Friederich, P., & Aspuru-Guzik, A. (2020). Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Machine Learning: Science and Technology, 1*(4), 045024. [CrossRef]

[34] Ucak, U. V., Ashyrmamatov, I., & Lee, J. (2023). Improving the Quality of Chemical Language Model Outcomes with Atom-in-SMILES Tokenization. *Journal of Cheminformatics, 15*(1), 55. [CrossRef]

[35] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics, 29*(5), 1189-1232. [CrossRef]

[36] Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing, 14*(3), 199-222. [CrossRef]

[37] Loh, W. Y. (2011). Classification and regression trees. *Wiley interdisciplinary reviews: data mining and knowledge discovery, 1*(1), 14-23. [CrossRef]

[38] Seeger, M. (2004). Gaussian processes for machine learning. *International journal of neural systems, 14*(02), 69-106. [CrossRef]

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

[40] Born, J., & Manica, M. (2023). Regression Transformer Enables Concurrent Sequence Regression and Generation for Molecular Language Modelling. *Nature Machine Intelligence, 5*(4), 432-444. [CrossRef]

[41] Wang, S., Guo, Y., Wang, Y., Sun, H., & Huang, J. (2019, September). Smiles-bert: large scale unsupervised pre-training for molecular property prediction. In *Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics* (pp. 429-436). [CrossRef]

[42] Yu, J., Zhang, C., Cheng, Y., Yang, Y.-F., She, Y.-B., Liu, F., Su, W., & Su, A. (2023). SolvBERT for Solvation Free Energy and Solubility Prediction: A Demonstration of an NLP Model for Predicting the Properties of Molecular Complexes. *Digital Discovery, 2*(2), 409-421. [CrossRef]

[43] Li, J., & Jiang, X. (2021). Mol-BERT: an effective molecular representation with BERT for molecular property prediction. *Wireless Communications and Mobile Computing, 2021*(1), 7181815. [CrossRef]

[44] Liu, Y., Zhang, R., Li, T., Jiang, J., Ma, J., & Wang, P. (2023). MolRoPE-BERT: An Enhanced Molecular Representation with Rotary Position Embedding for Molecular Property Prediction. *Journal of Molecular Graphics and Modelling, 118*, 108344. [CrossRef]

[45] Irwin, J. J., Sterling, T., Mysinger, M. M., Bolstad, E. S., & Coleman, R. G. (2012). ZINC: a free tool to discover chemistry for biology. *Journal of chemical information and modeling, 52*(7), 1757-1768. [CrossRef]

[46] Sterling, T., & Irwin, J. J. (2015). ZINC 15–Ligand Discovery for Everyone. *Journal of Chemical Information and Modeling, 55*(11), 2324-2337. [CrossRef]

[47] Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32. [CrossRef]

[48] Alperstein, Z., Cherkasov, A., & Rolfe, J. T. (2019).

All smiles variational autoencoder. *arXiv preprint arXiv:1905.13343*. [CrossRef]

[49] Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., ... & Aspuru-Guzik, A. (2018). Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science, 4*(2), 268. [CrossRef]

[50] Winter, R., Montanari, F., Noé, F., & Clevert, D. A. (2019). Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chemical science, 10*(6), 1692-1701. [CrossRef]

[51] Rong, Y., Bian, Y., Xu, T., Xie, W., Wei, Y., Huang, W., & Huang, J. (2020). Self-Supervised Graph Transformer on Large-Scale Molecular Data. *Advances in Neural Information Processing Systems, 33*, 12559-12571.

[52] Kim, H., Lee, J., Ahn, S., & Lee, J. R. (2021). A merged molecular representation learning for molecular properties prediction with a web-based service. *Scientific Reports, 11*(1), 11028. [CrossRef]

[53] Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., & Dahl, G. E. (2017, July). Neural message passing for quantum chemistry. In *International conference on machine learning* (pp. 1263-1272). Pmlr.

[54] Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., ... & Barzilay, R. (2019). Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling, 59*(8), 3370-3388. [CrossRef]