Ensemble Model With BERT,RoBERTa And XLNet For Molecular Property Prediction

Junling Hu

Abstract

Utilizing natural language processing methods for predicting molecular properties has proven to be a viable approach. However, the use of Transformer family models often necessitates extensive time for pre-training. This paper introduces an ensemble learning approach, employing supervised fine-tuning of BERT, RoBERTa, and XLNet as an alternative to pre-training for molecular property prediction. Through comparison with existing advanced models, the proposed method demonstrates significant effectiveness. This offers a novel approach for experimental groups with limited computational resources, enabling them to predict molecular properties without the need for extensive pre-training.

Keywords: Ensemble Learning, BERT, RoBERTa, XLNet, SMILES, Molecular Property Prediction

1.Introduction

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 generation, and molecular property prediction[2-4]. Molecular properties play a pivotal role in various fields including chemistry, drug discovery, and healthcare. They connect disciplines such as quantum mechanics, physical chemistry, biophysics, and physiology. Computer-assisted methods facilitate the rapid prediction of molecular properties[5].

Traditional machine learning methods often require manual selection and construction of features for training models, such as Extended-Connectivity Fingerprints (ECFP)[6]. As data volume increases, this process becomes time-consuming and may lead to performance saturation, affecting prediction accuracy[7]. Currently, various deep learning models have been proposed for predicting molecular properties. Common methods for molecular property prediction are based on one-dimensional sequences[8-10], such as SMILES, and two-dimensional representations[11-13], such as molecular graphs.

With the introduction of the Transformer model[14] in 2017 and Google's subsequent release of the BERT model in 2018, pre-training strategies have achieved great success in the field of natural language processing (NLP). Models based on SMILES sequences can be pre-trained using specially designed tasks such as the masked language model[15,16], focusing more on the context of molecular sequences. Graph models typically capture structural features by constructing a graphical

representation of molecules[17,18], 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, pre-training the input model is crucial[19].

However, pre-training models usually require substantial computational resources and large datasets, which is very time-consuming[20]. This presents significant challenges for environments with limited computational resources[21]. Additionally, pre-training may lead to excessive optimization for specific tasks, limiting its applicability in processing small-scale or specific types of datasets. Therefore, this study explores a strategy of fully fine-tuning these pre-trained architectural models from a state of random initialization, abandoning the traditional pre-training steps. To overcome the limitations of individual models and compensate for the lack of pre-training, this paper adopts an ensemble learning method[22,23] using BERT[24], RoBERTa[25], and XLNet[26] models for predicting molecular properties. Ensemble learning combines multiple models, utilizing their different perspectives and strengths to improve prediction accuracy. By stacking the outputs of multiple models, the ensemble method can compensate for the weaknesses of individual models, providing more robust and accurate property predictions. BiLSTM[27] is used as a component of the base predictor, and BaggingRegressor is used as a meta-predictor for the final prediction.

In light of this, the study aims to develop a molecular property prediction scheme that reduces reliance computational resources without significantly sacrificing prediction accuracy. This approach is intended to provide a feasible solution for environments with limited computational resources. The study hopes to demonstrate that effective molecular property prediction can still be achieved without large-scale pre-training, potentially rivaling the performance of current advanced models.

2.Relevant Work

Tokenization is a crucial preprocessing step in natural language processing, 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 series, 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 (Simplified Molecular Input Line Entry System)[28] 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 syntax. 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, prompting researchers to develop new SMILES representations to overcome the deficiencies of traditional representations [29]. DeepSMILES [30] increases the probability of generating valid molecules by introducing closing brackets or single symbols at cyclic positions. SELFIES [31] 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. introduced the Atom in SMILES (AIS) method [32], 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 pairing them with the right models iscrucial in the field of molecular prediction. In terms of models, the rapid development of the Transformer[33] family has facilitated their swift application in molecular language modeling[34,35]. The Transformer outperforms traditional RNN models in terms of performance, becoming the most versatile model to date. Innovations and improvements upon the Transformer framework in models like BERT, RoBERTa, and XLNet have led to superior performance in handling complex language modeling tasks.

In the evolving field of molecular property prediction, the adoption of NLP techniques signifies a leap forward. Ross et al. [36] pioneered with MolFormer, harnessing over 1.1 billion molecules to forecast chemical behaviors, showcasing the transformer model's prowess in capturing intricate molecular details. This approach underlines the potential of large-scale molecular language models in scientific

discovery.

Parallel to Born et al. [41] introduced the Regress Transformer (RT), a novel concoction blending regression analysis with conditional generation tasks. Utilizing the XLNet architecture, RT has surpassed existing models in both chemical and protein language modeling, illustrating the vast potential of combining numeric and text tokens for molecular science.

Ross et al., Wang et al. [37] introduced SMILES-BERT, a model predicated on unsupervised pretraining that has demonstrated remarkable predictive accuracy across several benchmarks. Its success on datasets like QM9 and ESOL highlights the model's ability to decipher complex chemical information, positioning it as a cornerstone for future explorations in drug discovery and material sciences.

Furthering this trajectory, Yu et al. [38] presented SolvBERT, a model specifically fine-tuned for solvation properties, marking a significant advancement in understanding molecular interactions through NLP models. Similarly, Li et al. [39] developed Mol-BERT, leveraging a vast corpus of SMILES strings to achieve unprecedented accuracy in molecular property predictions, illustrating the model's superiority in tasks that span across diverse molecular datasets.

Completing this panorama of innovation, Liu et al. [40] expanded on these foundations with MolRoPE-BERT, integrating innovative position encoding methods to refine predictions further. This model's performance, validated on multiple benchmark datasets, exemplifies the continuous enhancement of molecular property prediction models.

3.Methodology

3.1 Data Set

This project utilized two datasets: zinc250k and zinc350k：

Zinc250k(Fig. 1) is a subset of the zinc12 dataset [40], 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 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.

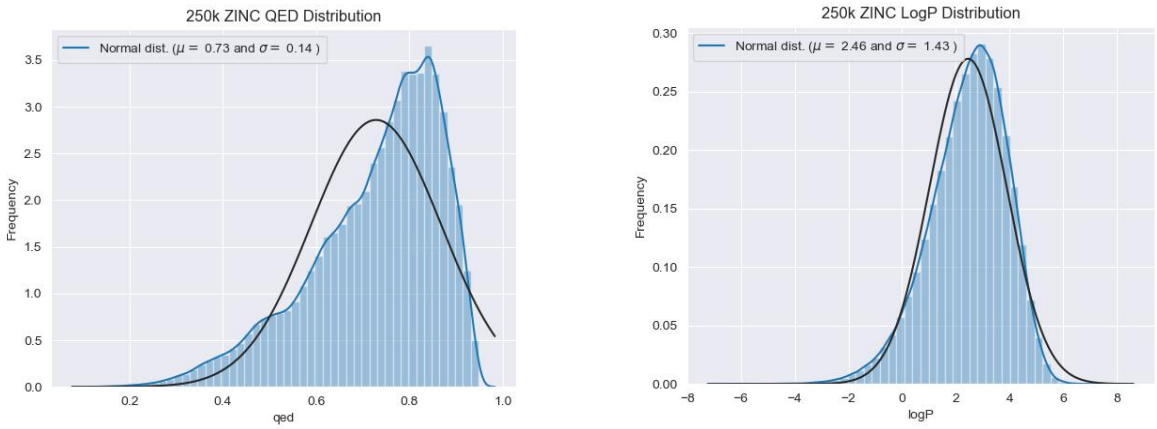

Fig. 1 ZINC250k dataset qed and logP property distribution histogram

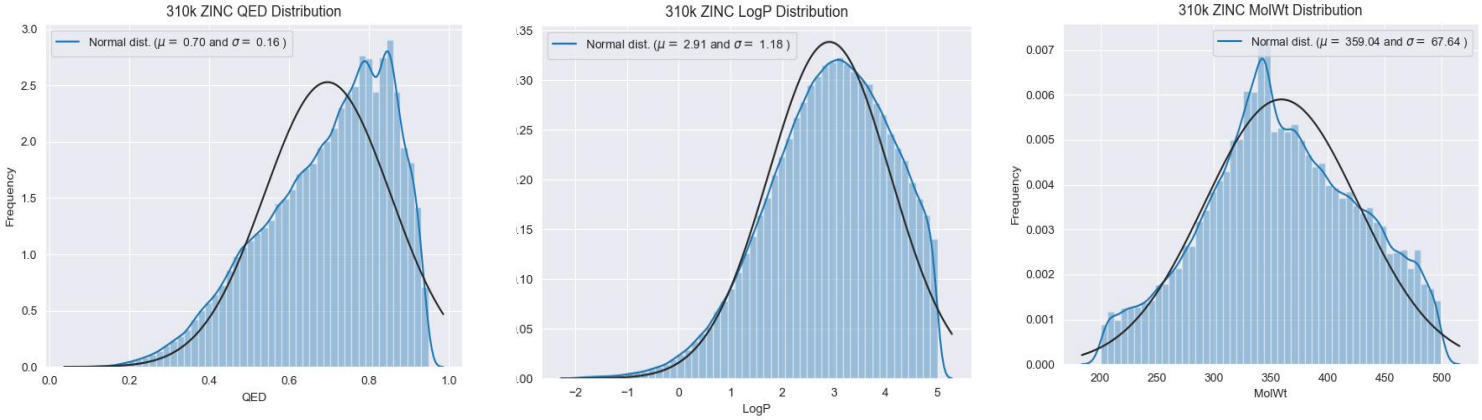

Fig. 2 ZINC310k dataset qed, logP and MolWt property distribution histogram

Zinc310k(Fig. 2) is a derivative of the complete zinc15 dataset [41], 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 are between -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

Our model selects AIS as the input and converts SMILES into AIS representation(Fig. 3).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.

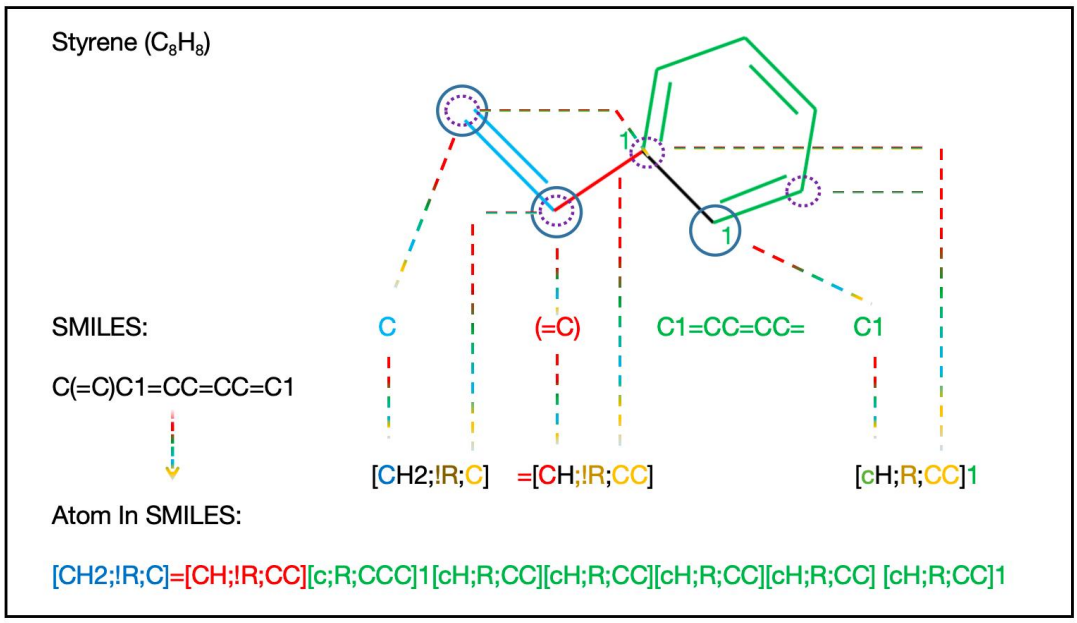

Fig. 3 An example illustrates the SMILES expression of the Styrene molecule (C(=C)C1=CC=CC=C1) and the step-by-step transformation process into AIS.

This formalized AIS representation offers a simplified and systematic approach to address the limitations of traditional SMILES tokenization, enhancing the predictive quality of molecular property prediction tasks.

After converting SMILES into Atom In SMILES (AIS), we treat each individual atom and special symbol as a separate token to construct the vocabulary(Fig. 4). When analyzing two datasets, we observed that traditional SMILES representation covered approximately 80 different tokens, while AIS representation increased to around 1000 tokens. This significant change highlights the richness and depth of AIS in capturing molecular structural details, making SMILES expression more akin to human language. AIS, with its expanded vocabulary, can provide a more detailed description of molecular properties, thereby offering deeper insights to deep learning models.

Index	Token	...	Index	Token	...	Index	Token
1	[PAD]		6	C		83	[SiH2]
2	[UNK]		7	O		84	[C+]
3	[CLS]		8	c		85	[N]

Index	Token	...	Index	Token	...	Index	Token
1	[PAD]		23	[CH3;!R;C]		1026	[S;R;CCN]
2	[UNK]		24	[C;!R;CCCC]		1027	[[N+];!R;CO]
3	[CLS]		25	[c;R;CCC]		1028	[S;R;NNN]

Fig. 4 The vocabularies for SMILES(Up) and AIS(Down) were created based on the zinc250k and zinc310k datasets

3.3 Ensemble Model

We designed an ensemble model(Fig.5) to predict molecular properties, incorporating BERT, RoBERTa, and XLNet as feature extractors to obtain AIS text features. These features are then passed to the base predictor, BiLSTM, which is an integral part of the base predictor, making full use of its sensitivity to sequential data and strong time-feature capturing capabilities. By integrating information in both forward and backward directions, the results are ultimately aggregated to the meta-learner (BaggingRegressor) to derive the final prediction. BaggingRegressor is a powerful ensemble learning algorithm used for regression tasks. It is based on the Bagging (Bootstrap Aggregating) principle, which improves overall performance by combining predictions from multiple base learners.

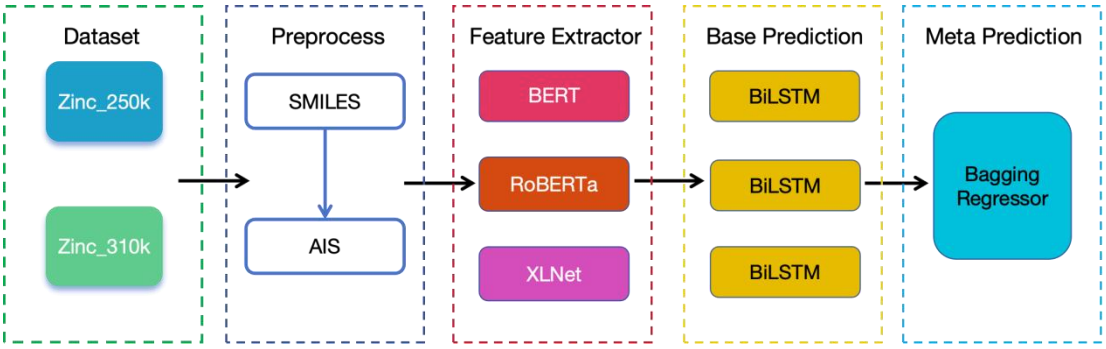

Fig. 5: The structure of ensemble model

To optimize training speed and efficiency, the layers of BERT and XLNet were reduced to six layers, while a small version of RoBERTa was employed. The table 1 below details the specific parameters of the model.

Table 1: Parameters set in model

Hyperparameter Name	Description
Learning Rate	0.00001
Batch Size	16
Dropout	0.1
Hidden Size	768
Attention Heads	12
Epochs	100
Optimizer	Adam
Loss Function	MAE Loss

3.4 Baseline Methods

We compare our ensemble model with the current state-of-the-art baseline models, which include:

ASVAE (All SMILES Variational Autoencoder)[42]: ASVAE utilizes a diversified SMILES representation and employs a stack of 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 its effectiveness in predicting attributes such as logP, molecular weight, and drug likeness (QED) in prior research, surpassing other advanced models [42-46]. We utilize the same dataset as ASVAE and consider it as a benchmark model. Although ASVAE's research reports only the MAE metric, which we consider a critical indicator as it directly reflects the average absolute difference between predicted and actual values, we specifically focus on MAE during the comparison.

GROVER(Graph Representation frOm self-superVised mEssage passing tRansformer)[47]. GROVER is a hybrid graph neural network (GNN) and Transformer architecture that takes molecular graphs as input, aiming to optimize the graph representation of molecules. What makes this model unique is its combination of two GNN Transformers—one tailored for nodes and the other for edges. These components are structurally similar but handle different features. Notably, one GNN component is specifically designed to transform graph information into features required by the Transformer. GROVER employs transfer learning techniques, similar to other sequence-based models, to improve the training efficiency and accuracy of downstream tasks. In our study, we used the pre-trained GROVER model and fine-tuned it on our dataset for performance comparison.

CHEM-BERT[48]. Utilizing SMILES strings as input, CHEM-BERT focuses on learning SMILES features during BERT pre-training. Additionally, this method incorporates matrix embedding layers for structural learning and quantitative estimation of drug similarity (QED) prediction tasks. Their approach has shown outstanding performance on multiple benchmark datasets, demonstrating its effectiveness in generalizing molecular data and improving downstream task prediction accuracy. To compare the performance of different models, we conducted a similar fine-tuning process on CHEM-BERT.

D-MPNN(Directed Message Passing Neural Network)[49]. D-MPNN takes molecular graphs as input and innovatively uses directed edges instead of atoms to convey information within the neural network structure, enhancing the molecular representation learning process. It can directly and efficiently handle molecular structural data without the need for pre-training on large-scale datasets. Extensive evaluations on various public and proprietary datasets have consistently shown that D-MPNN outperforms traditional models using fixed molecular descriptors and other graph neural architectures.

3.5 Evaluation Metrics

This study employs three main evaluation metrics to assess model performance: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination (R-squared or R^2).

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

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

MAE measures the average absolute difference between predicted and actual values, providing an indication of the average error level.

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

R^2 is a statistical measure used to assess the accuracy of model predictions, with values ranging from 0 to 1, where values closer to 1 indicate better model predictive performance.

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

4. Experiment

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 Fig. 6-7, indicate errors within a tight margin of 0.5%. 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 histograms' alignment of true and predicted value distributions further confirms the model's precision.

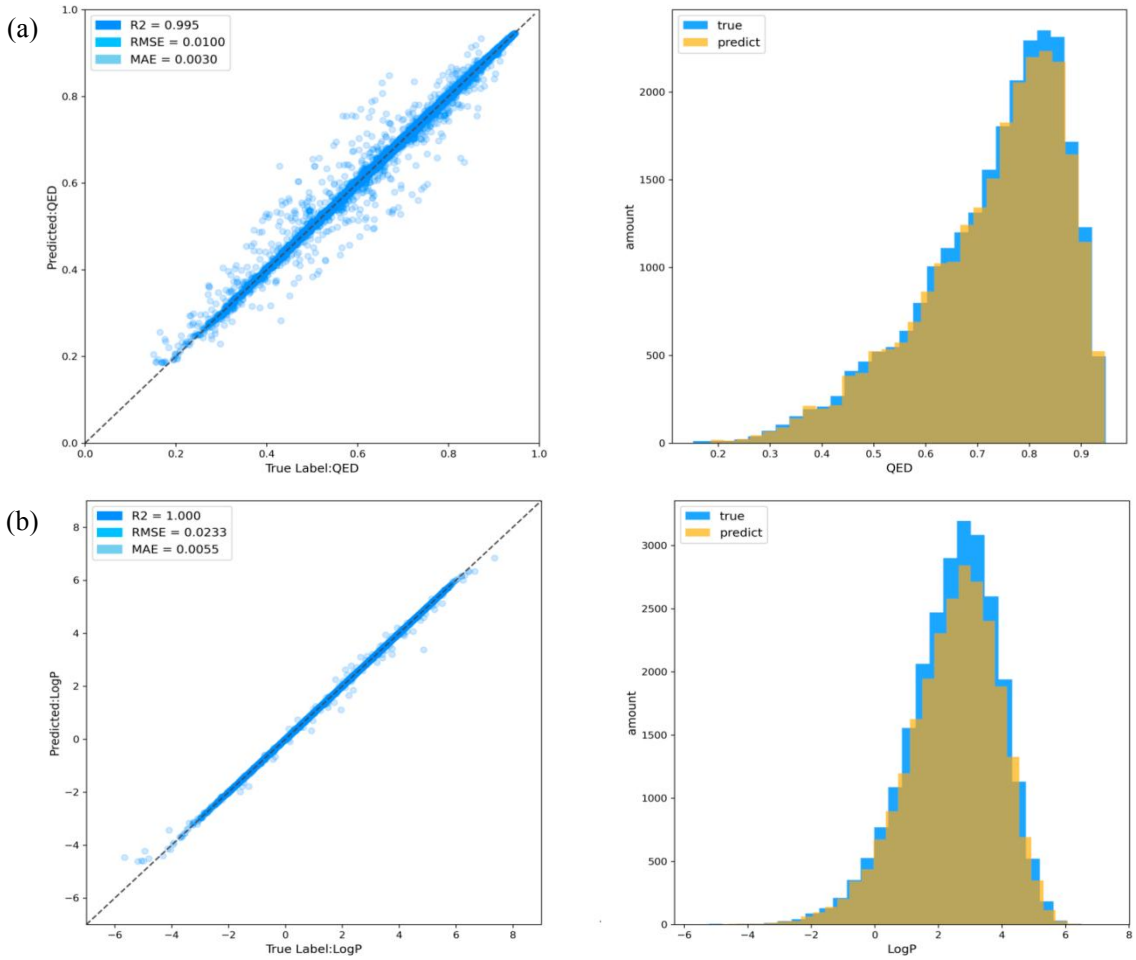

Fig. 6: Regressions (left) and distributions (right) of true and predicted values of qed and logP from ZINC250k

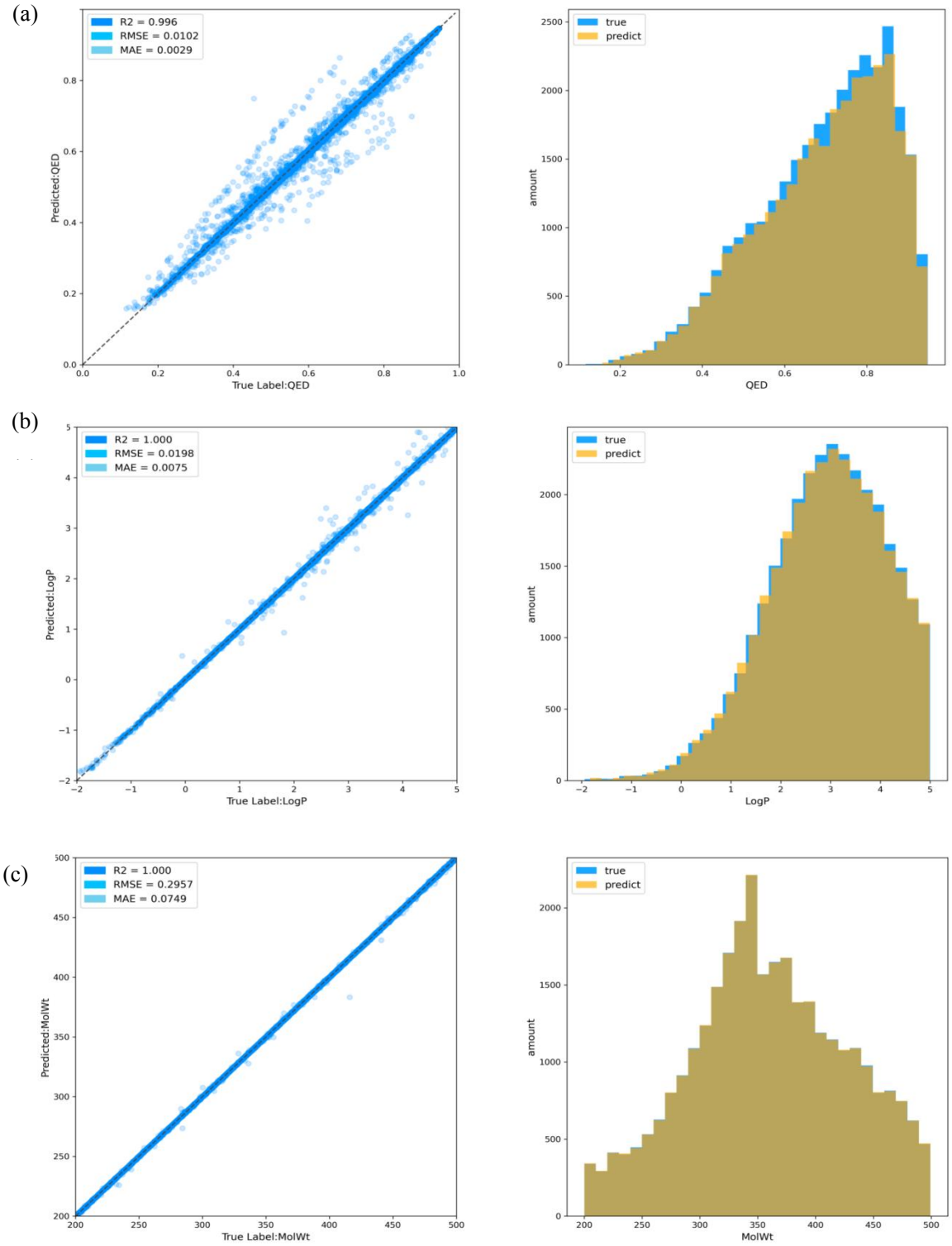

Figure 7: Regressions (left) and distributions (right) of true and predicted values of qed, logP and MolWt from ZINC310k

Table 2 reports the MAE, RMSE, and R^2 on the zinc250k dataset. From this table, we observe that our ensemble model achieves an MAE of 0.003 on the QED property, which is 42% better than the baseline model ASVAE and outperforms other graph-based and sequence-based models. For the logP property, the ensemble model achieves an MAE of 0.005, which is on par with the baseline ASVAE but

significantly better than other models.

Table 2: The performance of ZINC250k dataset

	QED			LogP			
Model name	MAE	RMSE	R2	MAE	RMSE	R2	Molecular representation
ASVAE	0.0052	\	\	0.005	\	\	SMILES
GROVER	0.0056	0.0083	0.9964	0.019	0.0365	0.9993	Graphs
D-MPNN	0.0056	0.0094	0.9953	0.016	0.029	0.9995	Graphs
CHEM-BERT	0.0049	0.0102	0.9934	0.011	0.023	0.9996	SMILES
Ensemble model	**0.0030**	0.0100	0.9948	**0.005**	0.024	0.9997	Atom In SMILES

Table 3 reports the MAE, RMSE, and R2 on the zinc310k dataset. Here, we can observe that our ensemble model achieves an MAE of 0.0029 for QED and 0.07 for MolWt, improving by 54% and 66%, respectively, compared to the baseline ASVAE. For the logP property, it remains comparable to the baseline ASVAE. These results further confirm the effectiveness of the ensemble model in multi-property prediction.

Table 3: The performance of ZINC310k dataset

	QED			LogP			MolWt			
Model name	MAE	RMSE	R2	MAE	RMSE	R2	MAE	RMSE	R2	Molecular representation
ASVAE	0.0064	\	\	0.007	\	\	0.21	\	\	SMILES
GROVER	0.0052	0.0073	0.9977	0.009	0.018	0.9997	0.26	0.35	0.9999	Graphs
D-MPNN	0.0059	0.0100	0.9959	0.017	0.034	0.9991	1.55	3.63	0.9997	Graphs
CHEM-BERT	0.0045	0.0103	0.9948	0.008	0.020	0.9997	0.40	0.63	0.9998	SMILES
Ensemble model	**0.0029**	0.0101	0.9958	**0.007**	0.019	0.9997	**0.07**	0.29	0.9999	Atom In SMILES

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. In our experiments, models that use SMILES sequences as input outperform those that use graphs as input. This may be attributed to the richer chemical information provided by SMILES sequences, enhancing the model's learning capacity.

Our experimental results highlight the excellent performance of the ensemble model on most performance metrics, particularly in terms of MAE. This finding underscores the potential of ensemble methods in improving prediction accuracy. However, we also note that on certain specific properties, the ensemble model performs comparably to single models. This suggests the need for further exploration of optimal configurations for different model combinations in future research.

5. Ablation Study

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(Table 4-8).

Table 4: The Ablation Study of ZINC_250k dataset with QED property

	Training			Validation			Test		
Model	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2
BERT + BiLSTM	0.0030	0.0080	0.9953	0.0037	0.0137	0.9889	0.0038	0.0144	0.9870
RoBERTa + BiLSTM	0.0035	0.0081	0.9947	0.0040	0.0130	0.9899	0.0039	0.0127	0.9899
XLNet + BiLSTM	0.0030	0.0063	0.9974	0.0038	0.0108	0.9899	0.0039	0.0114	0.9918
BiLSTM	0.0054	0.0152	0.9860	0.0068	0.0186	0.9795	0.0070	0.0194	0.9765
BERT + BiLSTM(SMILES)	0.0035	0.0090	0.9949	0.0042	0.0141	0.9882	0.0041	0.0142	0.9874
RoBERTa + BiLSTM(SMILES)	0.0035	0.0083	0.9957	0.0040	0.0121	0.9913	0.0041	0.0126	0.9904
XLNet + BiLSTM(SMILES)	0.0038	0.0092	0.9947	0.0045	0.0127	0.9904	0.0044	0.0127	0.9903
BiLSTM(SMILES)	0.0107	0.0221	0.9701	0.0120	0.0250	0.9620	0.0118	0.0246	0.9627
Ensemble(SMILES)							0.0035	0.0101	0.9958

Table 5: The Ablation Study of ZINC_250k dataset with LogP property

	Training			Validation			Test		
Model	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2
BERT + BiLSTM	0.0077	0.0103	0.9999	0.0085	0.0244	0.9996	0.0087	0.0268	0.9996
RoBERTa + BiLSTM	0.0174	0.0255	0.9996	0.0186	0.0366	0.9993	0.0186	0.0346	0.9994
XLNet + BiLSTM	0.0190	0.0248	0.9996	0.0233	0.0412	0.0991	0.0233	0.0396	0.9992
BiLSTM	0.0183	0.0337	0.9992	0.0205	0.0430	0.9989	0.0209	0.0460	0.9990
BERT + BiLSTM(SMILES)	0.0207	0.0356	0.9994	0.0241	0.0425	0.9991	0.0240	0.0442	0.9916
RoBERTa + BiLSTM(SMILES)	0.0206	0.0372	0.9994	0.0203	0.0371	0.9994	0.0207	0.0460	0.9991
XLNet + BiLSTM(SMILES)	0.0220	0.0367	0.9993	0.0263	0.0465	0.9989	0.0262	0.0473	0.9990
BiLSTM(SMILES)	0.0349	0.0574	0.9981	0.0353	0.0624	0.9976	0.0340	0.0621	0.9978
Ensemble(SMILES)							0.0148	0.0286	0.9995

Table 6: The Ablation Study of ZINC_310k dataset with QED property

	Training			Validation			Test		
Model	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2
BERT + BiLSTM	0.0028	0.0074	0.9963	0.0038	0.0139	0.9904	0.0037	0.0132	0.9917
RoBERTa + BiLSTM	0.0030	0.0102	0.9949	0.0038	0.0142	0.9906	0.0037	0.0133	0.9916
XLNet + BiLSTM	0.0033	0.0075	0.9973	0.0040	0.0115	0.9942	0.0040	0.0114	0.9938
BiLSTM	0.0046	0.0118	0.9926	0.0063	0.0171	0.9863	0.0063	0.0168	0.9871
BERT + BiLSTM(SMILES)	0.0030	0.0082	0.9967	0.0039	0.0119	0.9931	0.0039	0.0122	0.9930
RoBERTa + BiLSTM(SMILES)	0.0033	0.0089	0.9962	0.0040	0.0129	0.9920	0.0039	0.0128	0.9923
XLNet + BiLSTM(SMILES)	0.0036	0.0093	0.9958	0.0044	0.0143	0.9902	0.0044	0.0141	0.9905
BiLSTM(SMILES)	0.0098	0.0204	0.9801	0.0118	0.0237	0.9735	0.0107	0.0233	0.9741
Ensemble(SMILES)							0.0033	0.0100	0.9959

Table 7: The Ablation Study of ZINC_310k dataset with LogP property

	Training			Validation			Test		
Model	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2
BERT + BiLSTM	0.0125	0.0181	0.9997	0.0083	0.0204	0.9996	0.0084	0.0215	0.9995
RoBERTa + BiLSTM	0.0172	0.0246	0.9994	0.0159	0.0270	0.9993	0.0160	0.0277	0.9993
XLNet + BiLSTM	0.0147	0.0201	0.9996	0.0204	0.0300	0.9992	0.0204	0.0315	0.9991
BiLSTM	0.0170	0.0320	0.9990	0.0180	0.0380	0.9987	0.0180	0.0370	0.9987
BERT + BiLSTM(SMILES)	0.0156	0.0216	0.9996	0.0214	0.0314	0.9991	0.0162	0.0288	0.9992
RoBERTa + BiLSTM(SMILES)	0.0192	0.0269	0.9993	0.0265	0.0376	0.9987	0.0267	0.0391	0.9998
XLNet + BiLSTM(SMILES)	0.0181	0.0249	0.9994	0.0276	0.0366	0.9988	0.0278	0.0368	0.9987
BiLSTM(SMILES)	0.0336	0.0565	0.9972	0.0421	0.0675	0.9959	0.0343	0.0612	0.9967
Ensemble(SMILES)							0.0122	0.0222	0.9996

Table 8: The Ablation Study of ZINC_310k dataset with MolWt property

	Training			Validation			Test		
Model	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2
BERT + BiLSTM	0.22	0.34	0.99	0.15	0.82	0.99	0.15	0.34	0.99
RoBERTa + BiLSTM	0.34	0.61	0.99	0.42	0.95	0.99	0.42	0.67	0.99
XLNet + BiLSTM	0.40	0.51	0.99	0.25	0.84	0.99	0.26	0.46	0.99
BiLSTM	0.22	0.41	0.99	0.22	0.88	0.99	0.23	0.47	0.99
BERT + BiLSTM(SMILES)	0.23	0.37	0.99	0.18	0.30	0.99	0.18	0.29	0.99
RoBERTa + BiLSTM(SMILES)	0.39	0.55	0.99	0.52	0.73	0.99	0.52	0.74	0.99
XLNet + BiLSTM(SMILES)	0.42	0.56	0.99	0.45	0.63	0.99	0.45	0.62	0.99
BiLSTM(SMILES)	0.34	0.58	0.99	0.24	0.53	0.99	0.23	0.59	0.99
Ensemble(SMILES)							0.12	0.38	0.99

First, we examine the impact of AIS and SMILES inputs on model performance:

AIS Input: When employing AIS input, models such as BERT + BiLSTM, RoBERTa + BiLSTM, and XLNet + BiLSTM generally exhibit lower MAE and RMSE and higher R^2 on both the training and test sets. This suggests that the rich informational content of AIS input facilitates more effective learning and prediction by the models.

SMILES Input: A noticeable performance drop is observed in BERT + BiLSTM(SMILES), RoBERTa + BiLSTM(SMILES), and XLNet + BiLSTM(SMILES) when using SMILES as input. This could be attributed to the comparatively less rich information encoded in the SMILES format, potentially leading to less comprehensive learning by the models.

Among all the models, BERT + BiLSTM consistently shows higher prediction accuracy, regardless of whether AIS or SMILES is used as input, with R^2 values nearing or achieving 0.99 on the test set. This highlights the effectiveness of BERT's pre-trained representations in accurately predicting a variety of chemical properties.

In contrast, both RoBERTa + BiLSTM and XLNet + BiLSTM demonstrate superior performance with AIS input compared to SMILES, particularly in terms of

R^2. This indicates a more effective learning capability from the AIS format in RoBERTa and XLNet models.

The standalone BiLSTM model shows minimal variance in performance between AIS and SMILES inputs, suggesting that the type of input data has a limited impact on its effectiveness.

Lastly, the Ensemble(SMILES) model demonstrates commendable predictive accuracy across most cases, especially for the MolWt property in the second dataset. This finding reinforces the potency of ensemble strategies in amalgamating predictions from diverse models to enhance overall performance.

By comparing the impact of different input types on model performance, we conclude that while AIS input typically leads to better predictive results, similar performance can be achieved even with less structured SMILES input through appropriate model selection and ensemble methods. This emphasizes the importance of choosing suitable model structures and input types for specific chemical property prediction tasks. Future work may involve optimizing these models to better handle different types of chemical representations and exploring other potential ensemble strategies to enhance predictive performance.

6. Conclusion

This study's results highlight the significant capability of our ensemble model in handling chemical data, achieving state-of-the-art (SOTA) performance in predicting a variety of chemical properties. This accomplishment underscores the importance of integrating different models and techniques to enhance predictive accuracy, particularly when dealing with complex chemical structures and properties.

Additionally, we discovered that utilizing the AIS representation as input for deep learning models is highly effective. The structured and information-rich nature of the AIS representation allows deep learning models to more accurately capture and learn key features of molecules. This finding emphasizes the importance of choosing appropriate data representations to boost model performance.

Importantly, our research also validates the feasibility of effective training without the necessity for extensive pretraining. This suggests that models tailored for specific tasks can achieve efficient learning and accurate prediction, even in the absence of vast, generalized pretraining data. This aspect is particularly crucial in situations with limited resources or restricted access to data, paving new paths for future research.

Data Availability

The pre-training data and molecular prooerty prediction results in this work are available in the: https://github.com/jlinghu/AIS-Ensemble-model

Reference

1. Goh G B, Hodas N O, Vishnu A. Deep learning for computational chemistry[J]. Journal of computational chemistry, 2017, 38(16): 1291-1307.
2. Lavecchia A. Deep learning in drug discovery: opportunities, challenges and future prospects[J]. Drug discovery today, 2019, 24(10): 2017-2032.
3. Xu Y, Lin K, Wang S, Wang L, Cai C, Song C, Lai L, Pei J. Deep learning for molecular generation. Future Med Chem. 2019 Mar;11(6):567-597. doi: 10.4155/fmc-2018-0358. Epub 2019 Jan 30. PMID: 30698019.
4. Li Z, Jiang M, Wang S, Zhang S. Deep learning methods for molecular representation and property prediction. Drug Discov Today. 2022 Dec;27(12):103373. doi: 10.1016/j.drudis.2022.103373. Epub 2022 Sep 24. PMID: 36167282.
5. Walters W P, Barzilay R. Applications of deep learning in molecule generation and molecular property prediction[J]. Accounts of chemical research, 2020, 54(2): 263-270.
6. Rogers D, Hahn M. Extended-connectivity fingerprints[J]. Journal of chemical information and modeling, 2010, 50(5): 742-754.
7. LeCun Y, Bengio Y, Hinton G. Deep learning[J]. nature, 2015, 521(7553): 436-444.
8. Goh G B, Hodas N O, Siegel C, et al. Smiles2vec: An interpretable general-purpose deep neural network for predicting chemical properties[J]. arXiv preprint arXiv:1712.02034, 2017.
9. Pinheiro G A, Mucelini J, Soares M D, et al. Machine learning prediction of nine molecular properties based on the SMILES representation of the QM9 quantum-chemistry dataset[J]. The Journal of Physical Chemistry A, 2020, 124(47): 9854-9866.
10. Jo J, Kwak B, Choi H S, et al. The message passing neural networks for chemical property prediction on SMILES[J]. Methods, 2020, 179: 65-72.
11. Wieder O, Kohlbacher S, Kuenemann M, et al. A compact review of molecular property prediction with graph neural networks[J]. Drug Discovery Today: Technologies, 2020, 37: 1-12.
12. Zhang Z, Liu Q, Wang H, et al. Motif-based graph self-supervised learning for molecular property prediction[J]. Advances in Neural Information Processing Systems, 2021, 34: 15870-15882.
13. Jiang D, Wu Z, Hsieh C Y, et al. Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph- based models[J]. Journal of cheminformatics, 2021, 13(1): 1-23.

14. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.
15. Nozza, Debora, Federico Bianchi, and Dirk Hovy. "What the [mask]? making sense of language-specific BERT models." arXiv preprint arXiv:2003.02912 (2020).
16. Nozza, Debora, Federico Bianchi, and Dirk Hovy. "What the [mask]? making sense of language-specific BERT models." arXiv preprint arXiv:2003.02912 (2020).
17. M. Sun, S. Zhao, C. Gilvary, O. Elemento, J. Zhou, F. Wang, Graph convolutional networks for computational drug development and discovery, Brief Bioinform 21 (2020) 919–935.
18. J. Xiong, Z. Xiong, K. Chen, H. Jiang, M. Zheng, Graph neural networks for
19. automated de novo drug design, Drug Discov Today 26 (2021) 1382–1393.
20. Bianchi F, Terragni S, Hovy D. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence[J]. arXiv preprint arXiv:2004.03974, 2020.
21. Abnar S, Dehghani M, Neyshabur B, et al. Exploring the limits of large scale pre training[J]. arXiv preprint arXiv:2110.02095, 2021.
22. Dietterich T G. Ensemble methods in machine learning[C]//International workshop on multiple classifier systems. Berlin, Heidelberg: Springer Berlin Heidelberg, 2000: 1-15.
23. Dong X, Yu Z, Cao W, et al. A survey on ensemble learning[J]. Frontiers of Computer Science, 2020, 14: 241-258. 2021
24. Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.
25. Liu Y, Ott M, Goyal N, et al. Roberta: A robustly optimized bert pretraining approach[J]. arXiv preprint arXiv:1907.11692, 2019.
26. Yang Z, Dai Z, Yang Y, et al. Xlnet: Generalized autoregressive pretraining for language understanding[J]. Advances in neural information processing systems, 2019, 32.
27. Siami-Namini, Sima, Neda Tavakoli, and Akbar Siami Namin. "The performance of LSTM and BiLSTM in forecasting time series." 2019 IEEE International c onference on big data (Big Data). IEEE, 2019.
28. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules[J]. Journal of chemical information and computer sciences, 1988, 28(1): 31-36.
29. Skinnider M A, Stacey R G, Wishart D S, et al. Chemical language models enable navigation in sparsely populated chemical space[J]. Nature Machine Intelligence, 2021, 3(9): 759-770.
30. O'Boyle N, Dalke A. DeepSMILES: an adaptation of SMILES for use in machine learning of chemical structures[J]. 2018.

31. Krenn M, Häse F, Nigam A K, et al. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation[J]. Machine Learning: Science and Technology, 2020, 1(4): 045024.
32. Ucak U V, Ashyrmamatov I, Lee J. Improving the quality of chemical language model outcomes with atom-in-SMILES tokenization[J]. Journal of Cheminformatics, 2023, 15(1): 55.
33. Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).
34. Honda S, Shi S, Ueda H R. Smiles transformer: Pre-trained molecular fingerprint for low data drug discovery[J]. arXiv preprint arXiv:1911.04738, 2019.
35. Tetko I V, Karpov P, Van Deursen R, et al. State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis[J]. Nature communications, 2020, 11(1): 5575.
36. Ross J, Belgodere B, Chenthamarakshan V, et al. Large-Scale Chemical Language Representations Capture Molecular Structure and Properties[J]. arXiv preprint arXiv:2106.09553, 2021.
37. Born, Jannis, and Matteo Manica. "Regression Transformer enables concurrent sequence regression and generation for molecular language modelling." Nature Machine Intelligence 5.4 (2023): 432-444.
38. Wang S, Guo Y, Wang Y, et al. Smiles-bert: large scale unsupervised pre-training for molecular property prediction[C]//Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics. 2019: 429-436.
39. Yu, Jiahui, et al. "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 (2023): 409-421.
40. Li, Juncai, and Xiaofei Jiang. "Mol-BERT: an effective molecular representation with BERT for molecular property prediction." Wireless Communications and Mobile Computing 2021 (2021): 1-7.
41. Liu Y, Zhang R, Li T, Jiang J, Ma J, Wang P. MolRoPE-BERT: An enhanced molecular representation with Rotary Position Embedding for molecular property prediction. J Mol Graph Model. 2023 Jan;118:108344. doi: 10.1016/j.jmgm.2022.108344. Epub 2022 Sep 29. PMID: 36242862.
42. Irwin, John J., et al. "ZINC: a free tool to discover chemistry for biology." Journal of chemical information and modeling 52.7 (2012): 1757-1768.
43. Sterling, Teague, and John J. Irwin. "ZINC 15–ligand discovery for everyone." Journal of chemical information and modeling 55.11 (2015): 2324-2337.
44. Alperstein, Zaccary, Artem Cherkasov, and Jason Tyler Rolfe. "All SMILES Variational Autoencoder for Molecular Property Prediction and Optimization." QSPR/QSARAnalysis Using SMILES and Quasi-SMILES. Cham: Springer International Publishing, 2023. 85-115.
45. Rogers D, Hahn M. Extended-connectivity fingerprints. J Chem Inf Model. 2010 May 24;50(5):742-54. doi: 10.1021/ci100050t. PMID: 20426451.
46. Gómez-Bombarelli R, Wei JN, Duvenaud D, Hernández-Lobato JM, Sánchez-

Lengeling B, Sheberla D, Aguilera-Iparraguirre J, Hirzel TD, Adams RP, Aspuru-Guzik A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. ACS Cent Sci. 2018 Feb 28;4(2):268-276. doi: 10.1021/acscentsci.7b00572. Epub 2018 Jan 12. PMID: 29532027; PMCID: PMC5833007.

47. Duvenaud, David K., et al. "Convolutional networks on graphs for learning molecular fingerprints." Advances in neural information processing systems 28 (2015).
48. Kang, Seokho, and Kyunghyun Cho. "Conditional molecular design with deep generative models." Journal of chemical information and modeling 59.1 (2018): 43-52.
49. Rong, Yu, et al. "Self-supervised graph transformer on large-scale molecular data." Advances in Neural Information Processing Systems 33 (2020): 12559-12571.
50. Kim, Hyunseob, et al. "A merged molecular representation learning for molecular properties prediction with a web-based service." Scientific Reports 11.1 (2021): 11028.
51. K. Yang, K. Swanson, W. Jin, C. Coley, P. Eiden, H. Gao, A. Guzman-Perez, T. Hopper, B. Kelley, M. Mathea, A. Palmer, V. Settels, T. Jaakkola, K. Jensen and R. Barzilay, Analyzing learned molecular representations for property prediction.