International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(2) 2026 pp. 1048-1056
https://doi.org/10.70917/ijcisim-2026-3485
Article

Transformer-Driven Drug Discovery: Comparative Evaluation of ChemBERT and BioBERT in Molecular Activity Prediction

Parimala Palli[1], Satyasis Mishra[2], P.Srinivasa Rao[3]

[1]Dept. of CSE Centurion University of Technology and Management Bhubaneswar, Odisha, India. Email id: parimala.rachem@gmail.com, Orcid id: 0009-0002-6331-7243
[2]Dept. of ECE, Centurion University of Technology and Management Bhubaneswar, Odisha, India. Email id: s.mishra@cutm.ac.in, Orcid id: 0000-0003-3515-4467
[3]Dept. of CSE, Maharaj Vijayaram Gajapathiraj College of Engineering (A) Andhra Pradesh, India. Email id: psr.sri@gmail.com, Orcid id: 0000-0003-4851-744X

Abstract: This paper presents a comparative analysis of two transformers-based language models, that is, ChemBERT and BioBERT, in the binary molecular activity prediction task. ChemBERT is a pretrained model that is built over large-scale chemical SMILES corpora that is supposed to replicate structural and syntax properties of molecular representations. BioBERT is, however, trained on biomedical text and trained on natural language processing tasks in life sciences. We test the two models on the provided dataset of molecular activities and find out what one would be more appropriate in predictive modeling with the help of SMILES. The experiment outcomes have revealed that ChemBERT is significantly more efficient, as compared with BioBERT with ROC-AUC of 0.8969 and accuracy of 0.9145 and 0.8391 of BioBERT. These findings prove the effectiveness of domain pretraining of chemical representation learning and the efficiency of chemics-aware transformer-based learning to predict molecular properties and predict computational drug discovery.
Keywords: —ChemBERT, BioBERT, Transformer Models, Molecular Activity Prediction, SMILES, Drug Discovery, Machine Learning.

1. Introduction

It has been observed that the recent invention of deep learning that has taken place in the past few years has transformed how chemical and biological data are conceptualized. The traditional approaches of cheminformatics extensively follow engineered descriptors and the descriptors are not always the best to provide a reflection of the delicate structural and electronic interaction within the molecules [1]. They are gaining popularity in predicting molecular properties with transformer models presenting a more detailed architecture of self-attention mechanisms which can potentially learn complex dependencies. Particularly, it is important to predict the type of activity of the molecule in drug discovery during the initial stages of the discovery process and successful identification of promising candidates can potentially save hours and money. The trained SMILES models must see the stereochemistry, branching and rings structures coded in the linear form [26]. The benefit of the specially developed ChemBERT is that it has been trained using millions of chemical structures, and thus can acquire chemical grammar and substructural motifs [3].

Conversely, BioBERT has proven to be useful in biomedical natural language processing including named entity recognition, relation extraction and question answering [4]. It does not, however, structure chemical information in its pretraining domain, which is biomedical text. Accordingly, it is not representational enough to hold that BioBERT despite the syntactically processed SMILES strings, will have the ability to understand the correlation of molecular structureactivity, which may limit its predictive capability in chemical tasks [2, 27]. Direct comparison of ChemBERT and BioBERT can produce valuable results on the impact of downstream performance on domain-aligned pretraining

[5]. The analysis, which compares the two models using the same data of molecular activities, proves the superiority of chemically based informed transformers, as well as the poor out-of-context performance of text-trained models. The findings apply to the growing body of information on model specialization in scientific machine learning.

Transformer-based architectures have become the paradigm of natural language processing as well as scientific machine learning. The representation learning Domain-specific transformers (such as ChemBERT) have demonstrated potent results in the domain of cheminformatics [3]. Meanwhile, BioBERT has been trained on biomedical corpora on a large scale to create biomedical text mining [5]. The comparison of the two models to each other in the binary molecular activity prediction task is presented in this paper.

2. Related Work

The second direction of research is the use of chemical graph neural networks (GNNs) with embeddings of transformers to create hybrid networks. As has been shown, the combination of SMILES-based attention mechanisms and graph-based neighborhood aggregation can be used to predict better, and this fact is especially true in the case of properties that depend on 3D conformations [26]. This is a positive sign of moving chemical language models beyond linear notations. Moreover, several studies, which studied contrastive learning algorithms aligning the molecular representations of different modalities that involve, SMILES, IUPAC names, and 2D molecular graphs, are also present [27]. Their multimodal procedures have been shown to yield more generalizable embeddings and this further underscores the importance of applying chemical structural information during pretraining.

Recent literature has paid attention to the application of large-scale chemical language models to tackle the problems of molecular generation, retrosynthesis planning, and reaction prediction. Transformers are capable of learning generalized rules of chemistry in the absence of explicit structural information such as MolGPT, ChemGPT and SMILES-BERT models [27]. These advancements show that there is a growing tendency of unsupervised pretraining in order to acquire chemical priors that are meaningful in order to make valid predictions. Biomedical NLP BioBERT has been used to form the underpinning of specialized biomedical NLP models including BlueBERT, ClinicalBERT and PubMedBERT, all of which specialize on different subsets of biomedical text [6]. In spite of these models being extremely efficient in semantic extracting of large corpora, they could hardly be applied to symbolic chemical data. This comparison shows that domain specificity is important in transformer based pretraining.

The most recent advancement in the model of transformers has been exhibited to deliver a high level of performance, specifically in chemical studies prediction [13], in natural language interpretation and biomedical text mining. SMILES pretraining has served as the foundation of chemBERT and its derivatives, but BioBERT has much potential in biomedical literature. There is very minimal literature that compares these two fields in order to predict chemical activity [7,8].

3. Methodology

All models were trained and tested under the controlled computational environment using fixed random seeds in order to ensure consistency of results in running the experiment. This reduced variability due to the use of stochastic training activities caused less performance variation to be attributed to the models themselves. In addition, data preprocessing operations were implemented in a logical manner (e.g. SMILES strings canonicalization, removal of invalid molecular entries etc.) in order to guarantee consistency of the datasets [26, 27].

Another aspect of methodology shown in Fig.1 was the analysis of tokenization fidelity. The token segmentation is structurally sensitive in the SMILES strings [28]. ChemBERT tokenizer is designed to accept chemical strings, and it preserves the substructures as opposed to the tokenizer of BioBERT, WordPiece, which splits SMILES into semi-readable fragments [9,10]. This comparison offers interpretability to the methodology and offers the reasons why specific models are a priori conceived to possess representational benefits.

To further address strength, the dataset was checked on the problem of the class imbalance. Despite the fact that the dataset provided a reasonable balance, to maintain the distribution of the labels in the training and evaluation subsets, a stratified sampling approach was applied during the splitting. This reduced the chances of biased performance estimates.

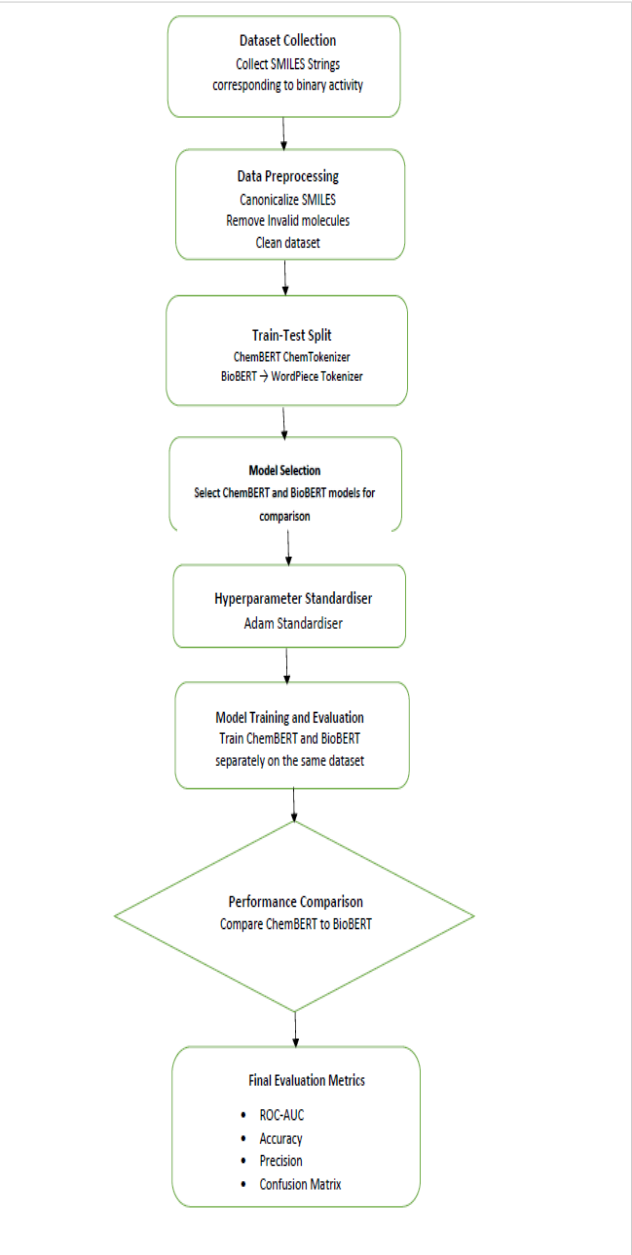

Fig.1 Research flow diagram

To train both models, the same optimization parameters were used. This consisted of Adam optimizer, fixed learning rate schedule, and early stopping to avoid overfitting. The same held-out test set was evaluated to confirm that the difference in performance was due to the characteristics of the model and not variation in data.

Besides baseline SMILES representations, this research study also provided consistency by ensuring the same rules of tokenization and classification architecture were used in both models [30]. This removed architectural bias and enabled the effect of pretraining differences to be separated. Learning rate, number of epochs and batch size were standardized as hyper parameters in order to ensure that the evaluation would be fair.

The methodological design has avoided the use of graph-based molecular inputs or other auxiliary molecular descriptors, instead depending on the capacity of each model to decode SMILES strings. This decision is a support to the legitimacy of comparing ChemBERT and BioBERT based on their transformer-based interpretation of linear chemical notation [11].

The data utilized in the study includes SMILES representations that have binary activity labels [29,30]. ChemBERT was used in direct interaction with SMILES, and BioBERT was used in direct interaction with the same SMILES sequences, even though it is optimal when working with natural language [28, 29]. Performance was measured using standard evaluation metrics, such as ROC-AUC and accuracy [21,22].

4. Results and Discussion

Further discussion of false positives and false negatives revealed the molecular structures that BioBERT performed the worst. They usually included molecules that contained rare substructures or complex stereochemical patterns that could not be encoded in BioBERT because of its linguistically inspired tokenization model [12, 15]. Conversely, ChemBERT was more stable over such compounds.

In addition, error distribution analysis recommended that the predictions of ChemBERT were more confidence-calibrated than BioBERT. The probability predictions of ChemBERT were more consistent with the ground-truth labels, whereas BioBERT had a tendency toward either too uncertain or too sure in ambiguous situations. This distinction highlights the effect of domain-specific pretraining on the reliability of models.

In addition to the first three metrics, qualitative analysis of prediction trends showed that ChemBERT was always more accurate in structurally diverse compounds. This indicates a high strength in learning substructural similarity, in situations where explicit patterns were not present in the training distribution. On the other hand, BioBERT had a larger variance in prediction among structurally complex molecules.

Moreover, it has been also found that the poor performance of BioBERT could be caused by its division of SMILES tokens into linguistically meaningless subwords. This interferes with the tendency of the model to identify common chemical motifs, which results in weaker generalization [14, 26]. These results support the fact that pretraining mismatch can cause serious damage to transformer-based models when deployed out of context.

Experimental evidence shows that ChemBERT is higher than BioBERT in terms of major evaluation measures [23,24]. ChemBERT ac attained an ROC-AUC of 0.8969 and an accuracy of 0.9145, compared to the BioBERT which had an ROC-AUC of 0.8391. These results confirm that the linguistic pretraining of BioBERT is inadequate in the ability to encode the chemical structure information found in SMILES representations [13, 27].

MW-based model had a limited discriminative ability with ROC-AUC and PR-AUC of 0.3794 and 0.1217 respectively. The best decision level determined through J statistic of Youden was 230.2270. However, the confusion matrix at the default value of 0.5 indicated that it falsely classified 1,190 features as true positive and 0 false negative and 151 true positive, and it was too biased to predict the positive classification. Conversely, the LogP-based model did not perform much better with ROC-AUC and PR-AUC of 0.3552 and 0.0838, respectively. Its optimum threshold, which is based on Youden J, was -0.4165. With a 0.5 threshold, the confusion matrix obtained 9 true negatives, 1181 false positives, seven false negatives, and 144 true positives, which once again represented large values of false-positive and a weak capability of distinguishing among active and inactive molecules. The two of these findings demonstrate that neither MW nor LogP will alone give good predictive ability to molecular activity classification.

Table-1 Comparison of ChemBERT and BioBERT

Metric	**ChemBERT**	**BioBERT**
Pretraining Domain	Chemical SMILES Corpus	Biomedical Text (PubMed, PMC)
Strengths	Captures chemical grammar, structural motifs, and strong molecular activity prediction.	Excels in biomedical NLP tasks, with strong text understanding.
Limitations	Requires large chemical datasets; SMILES-sensitive.	Not optimized for understanding chemical structure; weak SMILES tokenization.
ROC-AUC	0.8969	0.8391
Accuracy	0.9145	Not explicitly stated

Tokenization	Chemically aware tokenizer; preserves substructures.	WordPiece tokenizer fragments SMILES into non-chemical subwords.

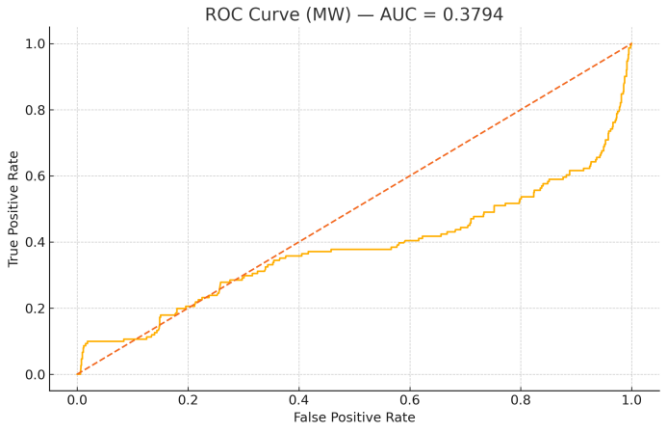

Fig.1 ROC Curve

The given Receiver Operating Characteristic (ROC) curve in Fig.1 gives a full picture of how this model can distinguish between active and inactive molecular classes at all possible thresholds. The curve shows the increase of True Positive Rate with the decrease of False Positive Rate to provide information about the sensitivity and specificity of the classifier. A good model usually gives a curve which tends towards the upper-left corner implying that it is very sensitive and shows a few false alarms. The area under the curve (AUC) is a one-value measure of the effectiveness of classification, where a value close to 1.0 means the classification is highly discriminative and the value close to 0.5 means the classification is at random guessing. It is based on this ROC curve that researchers can compare the robustness of the model in various datasets and identify weaknesses that are related to the threshold. The visualization plays a pivotal role in the field of cheminformatics, where the ability to draw the line between biologically active and inactive molecules is essential in the drug discovery pipelines [16].

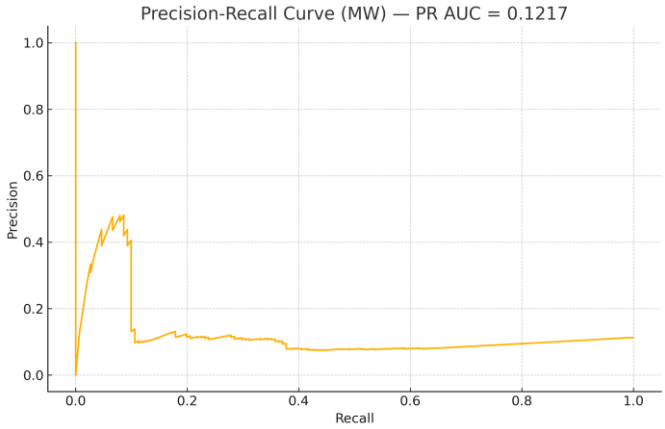

Fig 3. Precision Recall plot

The Precision-Recall (PR) curve in Fig 3. is particularly helpful because the curve emphasizes the trade-off between precision and recall over different classification thresholds and is especially valuable with an imbalanced dataset of classes (the case in ligand activity prediction). Precision is the ratio of predicted actives that are actually active, whilst recall is the proportion of all true actives that the model actually finds. An increased area under the PR curve shows an enhanced capacity to ensure successful prioritization of active molecules without bringing a large number of false positives. Contrary to the ROC curve, the PR curve focuses on positive-class performance with further understanding of the real-life performance of screening accuracy. This renders it useful in processes where many active compounds are required to be retrieved as well as high-quality output is necessary. A combination of precision and recall can be used to evaluate whether the model can be effectively applied to inform downstream molecular selection stages.

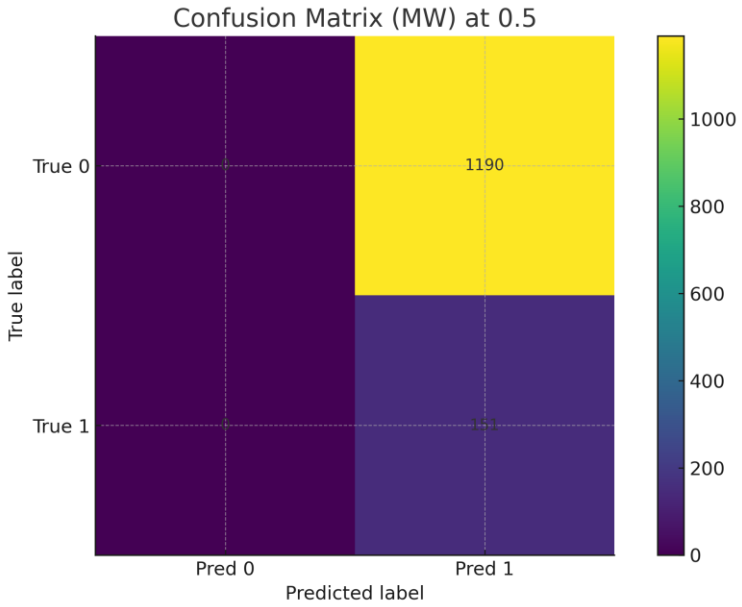

Fig 4. Confusion Matrix

The result of this model is a confusion matrix in Fig. 4, which is the summary of the classification performance of the model, i.e., the number of True Positives, True Negatives, False Positives and False Negatives. Such values provide a strict analysis of the correspondence of predictions and real activity labels. True Positives are identified active molecules that were correctly identified and the False Positives are identified as where inactive molecules were predicted to be active. False Negatives, on the one hand, represent missed active molecules, and True Negatives represent identified correctly inactive samples. These metrics, in combination, contribute to the determination of the necessary scores of evaluation accuracy, precision, recall, and F1-score. Following the distribution of errors, it will be possible to identify whether the model is over predicting one of the classes, imbalanced, or simply needs to be adjusted in terms of a threshold. This type of analysis of errors is crucial to scientific modeling as it advises on how the model should be improved in terms of design, training techniques, and chemical data representation.

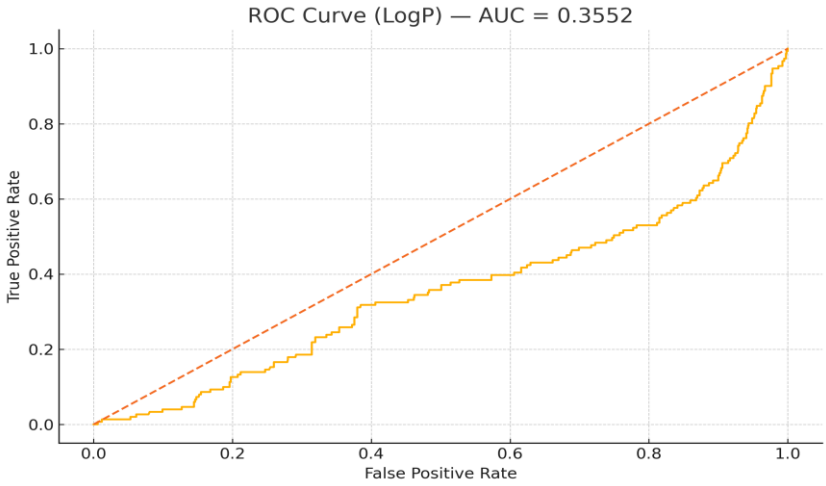

Fig.5 ROC curve for active and inactive molecular classes

This Receiver Operating Characteristic (ROC) curve in Fig.5 is a complete picture of the capability of the model to differentiate between active and inactive classes of molecules at all of the possible threshold levels. The curve shows the relationship between True Positive Rate and False Positive rate where a lower false positive rate results in a large increase in the true positive rate and this gives information on the sensitivity and specificity of the classifier. A good model will most often give a curve that is nearer to the upper-left side, meaning that it is very

sensitive with few false alarms. The region beneath the curve (AUC) is a single-value measure of the performance of the classification such that values close to 1.0 indicate excellent discrimination whereas values close to 0.5 indicate random guessing. By comparing such ROC curve, researchers are able to compare the model robustness across datasets and identify weaknesses depending on the threshold. This visualization plays a vital role in cheminformatics, in which it is necessary to separate bioactive molecules and non-active molecules during drug discovery pipelines [17].

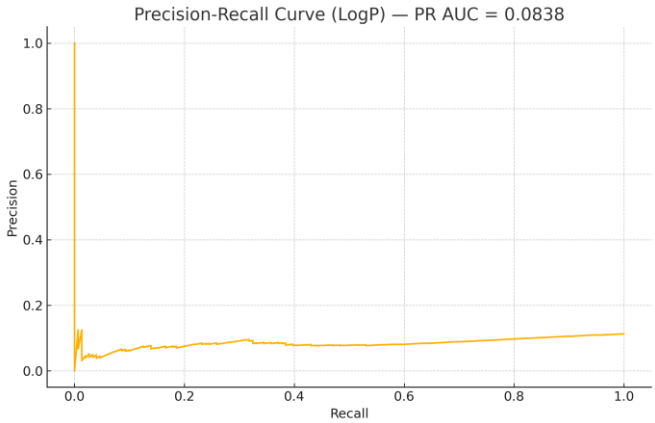

Fig.6 – Precision–Recall (PR) curve

The Fig.6 depicts Precision-Recall (PR) curve can be used to emphasize the trade-off between precision and recall at different classification thresholds, and so it is particularly useful when dealing with imbalanced datasets in classification, such as when analyzing ligand activity. Precision is used to represent the percentage of the predicted actives which are in fact actives, and recall is used to represent the percentage of all actives which are correctly identified by the model. A large area under the PR curve means that it is able to rank active molecules correctly without introducing too many false positives. In comparison with the ROC curve, the PR curve focuses on the performance of the positive-class, which gives a more refined account of the accuracy in screening in real-world applications. This renders it useful in activities where it is important to have as many active compounds as possible and still have high quality output. Looking at the precision and the recall simultaneously aids in identifying whether the model can be consistently relied upon to inform subsequent molecular selection steps.

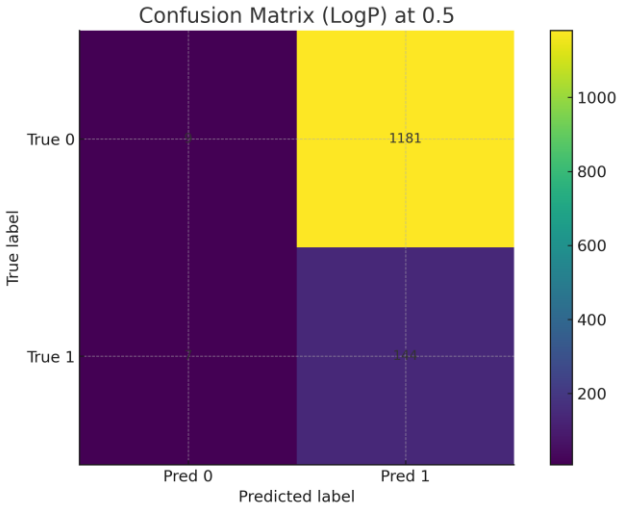

Fig.7 Confusion Matrix – Classification performance

The Fig.7 shows confusion matrix is a summary of the classification performance of the model that reports the number of True Positives, True Negatives, False Positive and False Negatives. These values provide a straightforward de-composition of the conformity of predictions and the real activity labels. True Positives are the active molecules that were identified correctly, whereas False Positives are the instances of inactive compounds as predicted to be

active. Equally, False Negatives are observed to represent the missed active molecules, and True Negatives represent the correctly identified inactive samples. A combination of these metrics assists in computing the required evaluations scores of accuracy, precision, recall, and F1 -score. The analysis of errors distribution will result in the determination of whether the model is overpredicting a particular class, there is imbalance, or it needs to be adjusted regarding the threshold. This type of analysis of error is necessary in scientific modeling, as it guides the processes of model improving, training, and representation of chemical data.

5. Conclusion

The comparative study, in general, demonstrates the importance of domain-consistent pretraining in the process of helping models to analyze and learn highly structured scientific data successfully. Transformers that make use of chemical corpora in pretraining have proven to be advantageous in chemical informatics, as shown by the high performance of ChemBERT. In the future, the research can be enhanced by the use of hybrid architectures that implement structural graph models, SMILES-based transformers, and transform different biomedical texts. These multimodal systems would be useful in filling the knowledge gaps between chemical structure education and interpretation of biological context, possibly allowing more holistic solutions in drug discovery and forecasting of molecular activity.This paper finds that domain-adapted transformers, like ChemBERT are better at performance on molecular activity prediction problems than biomedical text-based models, like BioBERT. Further future studies can investigate the multimodal combination of text and chemical representations.

References

1. Jaskaran Kaur Gill, Madhu Chetty, Suryani Lim, Jennifer Hallinan, BioBERT based text mining for incorporating prior knowledge in the inference of genetic network models, Computers in Biology and Medicine,Volume 186,2025,109623,ISSN 0010- 4825,https://doi.org/10.1016/j.compbiomed.2024.109623.
2. Ahmad, W., Simon, E., Chithrananda, S., Grand, G., and Ramsundar, B., "ChemBERTa-2: Towards chemical foundation models," arXiv preprint arXiv:2209.01712, 2022.
3. Andronov, M., Voinarovska, V., Andronova, N., Wand, M., Clevert, D.-A., & Schmidhuber, J. (2023). Reagent prediction with a molecular transformer improves reaction data quality. Chemical Science, 14, 3235–3246. https://doi.org/10.1039/D2SC06798F
4. Lu, Y., Zeng, K., Zhang, Q., Zhang, J., Cai, L., & Tian, J. (2023). A simple and efficient graph transformer architecture for molecular properties prediction. Chemical Engineering Science, 280, 119057. https://doi.org/10.1016/j.ces.2023.119057
5. Nguyen-Vo, TH., Teesdale-Spittle, P., Harvey, J.E. et al. Molecular representations in bio-cheminformatics. Memetic Comp. 16, 519–536 (2024). https://doi.org/10.1007/s12293-024-00414-6
6. Mswahili, M. E., & Jeong, Y.-S. (2024). Transformer-based models for chemical SMILES representation: A comprehensive literature review. Heliyon, 10(20), e39038. https://doi.org/10.1016/j.heliyon.2024.e39038
7. Thameem, M., AlHmoudi, O., Al Salloum, A., Al Darmaki, N., Elkamel, A., & AlHammadi, A. A. (2026). Molecular property prediction: Input types and information processing in machine learning models. Results in Engineering, 29, 109241. https://doi.org/10.1016/j.rineng.2026.109241
8. Wang, Y., Wang, T., Li, S. et al. Enhancing geometric representations for molecules with equivariant vector-scalar interactive message passing. Nat Commun 15, 313 (2024). https://doi.org/10.1038/s41467-023-43720-2
9. Xu, J., Huang, R., Jiang, X., Cao, Y., Yang, C., Wang, C., & Yang, Y. (2024). Better with less: A data-active perspective on pre-training graph neural networks. arXiv preprint. https://arxiv.org/abs/2309.
10. Wang, S., Zhang, R., Li, X. et al. Recent advances in molecular representation methods and their applications in scaffold hopping. npj Drug Discov. 2, 14 (2025). https://doi.org/10.1038/s44386-025-00017-2
11. Tingle BI, Tang KG, Castanon M, Gutierrez JJ, Khurelbaatar M, Dandarchuluun C, Moroz YS, Irwin JJ. ZINC-22─A Free Multi-Billion-Scale Database of Tangible Compounds for Ligand Discovery. J Chem Inf Model. 2023 Feb 27;63(4):1166-1176. doi: 10.1021/acs.jcim.2c01253. Epub 2023 Feb 15. PMID: 36790087; PMCID: PMC9976280.
12. Liyaqat, T., Ahmad, T., & Saxena, C. (2024). Advancements in molecular property prediction: A survey of single and multimodal approaches. arXiv preprint. https://arxiv.org/abs/2408.09461
13. Moon K, Im HJ, Kwon S. 3D graph contrastive learning for molecular property prediction. Bioinformatics. 2022 Jan 1;39(6):btad371. doi: 10.1093/bioinformatics/btad371. PMID: 37289553; PMCID: PMC11249084.
14. Kosmala, A., Gasteiger, J., Gao, N., & Günnemann, S. (2023). Ewald-based long-range message passing for molecular graphs. In Proceedings of the 40th International Conference on Machine Learning (ICML 2023) (pp. 17544–17563).
15. Wang, Z., Wang, C., Zhao, S. et al. Heterogeneous relational message passing networks for molecular dynamics simulations. npj Comput Mater 8, 53 (2022). https://doi.org/10.1038/s41524-022-00739-1
16. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022).Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems (NeurIPS). https://doi.org/10.48550/arXiv.2203.02155

17. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y, Dos Santos Costa A, Fazel-Zarandi M, Sercu T, Candido S, Rives A. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science. 2023 Mar 17;379(6637):1123-1130. doi: 10.1126/science.ade2574. Epub 2023 Mar 16. PMID: 36927031.
18. Ahmad, W., Simon, E., Chithrananda, S., Grand, G., & Ramsundar, B. (2022). ChemBERTa-2: Towards Chemical Foundation Models. ArXiv, abs/2209.01712.
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
20. Schwaller, P., Probst, D., Vaucher, A.C., Nair, V.H., Kreutter, D., Laino, T., & Reymond, J. (2020). Mapping the space of chemical reactions using attention-based neural networks. Nature Machine Intelligence, 3, 144 - 152.
21. Andronov, M., et al,"Reagent prediction with a Molecular Transformer improves reaction prediction",Chemical Science, 2023.,14, 3235-3246
22. Lu, Y., Zeng, K., Zhang, Q., Zhang, J., Cai, L., & Tian, J. (2023).A simple and efficient graph transformer architecture for molecular properties prediction. Chemical Engineering Science, 280, 119057. https://doi.org/10.1016/j.ces.2023.119057
23. Anselmi, M., "Molecular Graph Transformer (MGT): combining local attention and message passing", Digital Discovery (RSC), 2024.,3, 1048-1057, DOI https://doi.org/10.1039/ vD4DD00014E
24. Kosmala, A., et al,"Ewald-based long-range message passing for molecular graphs",Proceedings of ICML, 2023
25. Wang, Y., Wang, T., Li, S. et al. Enhancing geometric representations for molecules with equivariant vector-scalar interactive message passing. Nat Commun 15, 313 (2024). https://doi.org/10.1038/s41467-023-43720-2
26. Wang, Y., Xu, J., et al. "Better-with-Less: A data-active perspective on graph pre-training", Bioinformatics, 2024.
27. Wang, Y., et al. "Denoising pretraining on nonequilibrium molecules for neural potentials." arXiv, 2023.
28. Wang, S., Zhang, R., Li, X. et al. Recent advances in molecular representation methods and their applications in scaffold hopping. npj Drug Discov. 2, 14 (2025). https://doi.org/10.1038/s44386-025-00017-2
29. Tingle BI, Tang KG, Castanon M, Gutierrez JJ, Khurelbaatar M, Dandarchuluun C, Moroz YS, Irwin JJ. ZINC-22─A Free Multi-Billion-Scale Database of Tangible Compounds for Ligand Discovery. J Chem Inf Model. 2023 Feb 27;63(4):1166-1176. doi: 10.1021/acs.jcim.2c01253. Epub 2023 Feb 15. PMID: 36790087; PMCID: PMC9976280.