Grenze International Journal of Engineering and Technology, June Issue

ChemSol AI: A Deep Learning Approach to Predict Molecular Solubility

Anvitha A Pai[1], Adithya M[2], Deeksha[3], Ashwini[4] and G Godavari[5]
[1-5]Department of Computer Science and Engineering, Canara Engineering College Bantwal, Karnataka, India
Email: anvithapai09@gmail.com, adithyam@canaraengineering.in, deekshad758@gmail.com, ashwinikiradi@gmail.com, godhavari06@gmail.com

***Abstract*— The ability to predict molecular solubility is an important factor in drug development. Solubility dictates how readily a drug compound will be absorbed into the body and ultimately, how effective that compound will be as a therapeutic agent. Conventional feature-based approaches to solubility prediction have been limited by their inability to adequately model the complexities of a molecule's structure or capture the nonlinear interactions occurring within those structures. A new framework for solubility prediction utilizing deep learning techniques (specifically transformers and graph neural networks) has been designed to generate more complex and nuanced molecule representations. A new framework was trained on carefully curated dataset consisting of verified solubility values combined with extensive structural information about each molecule. Using common regression and classification criteria, six predictive models Random Forest, XGBoost, Gradient Boosting, ChemBERTa, Graph Transformer and GAT were compared. With an MAE of 1.40 and R2 of 0.69, Graph Transformer demonstrated the best regression performance, whereas ChemBERTa demonstrated a balanced performance with 80.6% accuracy in solubility categorization. This method facilitates quicker and more effective early-stage screening of possible drug candidates by providing more precise and trustworthy solubility estimations. The results demonstrate the increasing influence of sophisticated deep learning methods in computational chemistry and their capacity to greatly speed up data-driven pharmaceutical development.**

***Index Terms*— Solubility Prediction, Machine Learning, Deep Learning, Pharmaceutical Modelling, ChemBERTa, Graph Transformer, GAT.**

I. INTRODUCTION

Molecular solubility is an important factor in the drug discovery process because a drug's ability to be absorbed, distributed and ultimately effective is dependent on the degree to which it can dissolve within different solvents. Solubility prediction is also useful to chemists in reducing experimental cost, increasing speed of compound screening and eliminating potential drug candidate failures prior to later stages of drug development. Therefore, accurate solubility predictions continue to remain one of the top challenges in drug research due to the significant influence of solubility on a drug's efficacy and bioavailability. Many current computational approaches including rule-based models and basic machine learning algorithms have limitations that result in inability to adequately capture the complex structural patterns and non-linear chemical interactions that determine solubility.

Grenze ID: 01.GIJET.12.2.65_23

Although deep learning has improved the quality of predictions for solubility, most current models lack the ability to properly encode molecular topology or longer range relationships, and therefore limit their predictability. These limitations shows that there is a need for more new, better developed and robust models that are capable of creating chemical representations far more complex than current ones. The primary goal of this study is to create a deep-learning based model capable of making accurate predictions about solubility of molecule through the utilization of advanced transforms and graph neural networks. The end goal of the study is to determine which models will most accurately capture the molecular interactions required to achieve the highest level of predictive performance.

This research also makes the following key contributions:

- Comparative analysis of machine learning (ML) models and three deep learning (DL) models for the prediction of the solubility of molecules.
- The implementation of a DL model that uses a transformer architecture for both regression and classification type problems.
- The development of a web-based application for the prediction of solubility for use as part of the initial phases of drug screening.

The remainder of this manuscript will describe the design of models employed within this study, the preprocessing steps that were used to prepare the data for modeling and the methods utilized to train these models. Results from experiments that compare the performance of the various ML/DL models and a discussion of the relative merits and demerits of each model will be presented in Section IV. Finally, the last section of this manuscript will provide a summary of the major findings of this research and a description of the potential impact of the findings upon future pharmaceutical research.

II. Literature Review

The focus of the other research projects is developing better solubility prediction methods through the use of new approaches in artificial intelligence and deep learning. Compared to older models that are based on descriptors (descriptive variable), it appears that newer graph-based neural networks can make a greater number of successful predictions of molecular interactions and structure relationships. Advanced residual and attention-based graph descriptions were used to describe molecules. In this context, the nodes of the graph represent the molecules and edges represent the connection between the molecular. This graph description resulted in improved solubility predictions across of variety of different datasets [1]-[3]. These models have shown to work well for organic and aqueous systems and provided a good understanding of how molecular structures interact. In addition to their potential to make accurate solubility predictions, there has also an increasing interest in developing ensemble and tree-based models. A Machine learning framework that utilizes gradient boosting and random forest has been utilized to predict solubility from molecular fingerprints and descriptors [4]-[7]. Hybrid AI models have been created that utilize domain knowledge, statistical learning and optimization. Hybrid models that include neural network architecture and optimization techniques have been developed to optimize the prediction of drug solubility in supercritical and green solvents [8],[9]. Numerous comparative analyses have been completed to identify which model(s) may be the most suitable to the evaluation of specific data sets comparisons of the predictive abilities of different solubility models and frameworks [10],[11]. Although, these models significantly improve the accuracy of solubility predictions, they are still limited due to data availability and computational resources.

This study provides an expanded solution by developing a single framework combining the benefits of deep neural feature extraction and descriptor-based learning to further improve the consistency of predictions. In contrast to prior studies focusing exclusively on either deep neural or descriptor-based learning, our approach enables the prediction of solubility for new substances while maintaining the usability of the method for practical applications in the pharmaceutical industry. This study aims to find a balance among three criteria: performance, interpretability and applicability in real-world applications.

III. Proposed Method

A. Overview

The main steps in producing a solubility prediction inside the ChemSol AI framework are summarized in the workflow shown in Fig. 1.

B. Dataset and Preprocessing

This study used data from PubChem and DrugBank that contains verified solubility of substances in experimental studies; and, it also contains the names of the solvent/solute pairs and their SMILES representations. The data set after validation/cleaning had 644 entries/rows and nine columns. It contained 139 unique solutes and four unique solvents. Each entry in the table is a unique solvent/solute pair at a specified temperature. The entries include: the solubility, the unit type, and encoded identifiers, as well as the SMILES representation of each molecule. Duplicate entries were removed and incorrect chemical structure(s) were corrected by converting all molecules into canonical SMILES representations using RDKit. All solubility values were converted to mol/L from mass based units (g/L or mg/mL) using

$$\text{C mol/L} = \frac{\textit{mass concentration (g/L)}}{\textit{molecular wei} \quad \textit{(g/mol)}} \qquad (1)$$

and % w/v values were transformed via

$$\text{C g/L} = (\text{value}) \times 10 \qquad (2)$$

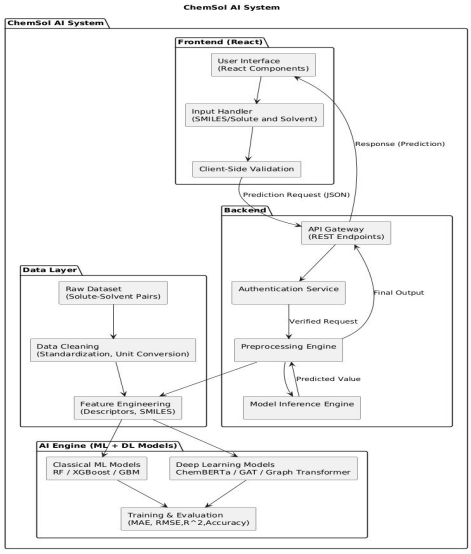

Fig. 1: System Architecture flow diagram of ChemSol AI

This ensures uniform molar representation across all samples. This augmentation was applied independently to both solute and solvent, expanding the training set according to:

$$\text{Augmented Size} = \text{Original Size} \times (1 + \text{num\_aug}) \qquad (3)$$

The number of augmentations (num_aug) was optimized using Optuna within the range:

$$1 \leq \text{num\_aug} \leq 3 \quad (4)$$

Additionally, molecular descriptors such as molecular weight, LogS, hydrogen bond donors/acceptors, TPSA and rotatable bond count were computed to enrich the feature space. The processed dataset was then split 70:20:10 into training, testing and validation sets to enable reliable evaluation of the solubility prediction models.

C. Feature Extraction

Feature extraction focused on deriving structural and physicochemical descriptors that act as input representation for predicting solubility prediction. For ML models the feature extraction was based on handcrafted features and for DL models the features were selected based on the inbuilt tool RDKit, including molecular weight, LogS, hydrogen bond donors/acceptors, TPSA, rotatable bonds and aromatic ring count were computed and normalized. For DL models molecular representations were learned automatically using SMILES based embeddings and graph based encodings. Molecular descriptors and SMILES embeddings are used as input features, while the experimentally measured solubility value is treated exclusively as the target variable.

D. Model Training and Implementation

Model training was structured to ensure fair evaluation across all approaches. The dataset was split into training, validation and test sets to prevent information leakage. Classical ML models (Random Forest, Gradient Boosting, XGBoost) were trained on RDKit descriptors, with hyperparameters optimized via grid search and cross-validation.

For deep learning, ChemBERTa was fine-tuned on SMILES embeddings using a regression head with Adam optimizer, learning rate schedules and dropout to improve generalization. Graph based models (GAT and Graph Transformer) processed molecular graphs through attention layers, trained with minibatch gradient descent and early stopping to avoid overfitting. All experiments were implemented in the PyTorch deep learning framework with fixed random seeds to ensure reproducibility and consistent evaluation across models. Training was carried out on a system equipped with an NVIDIA GPU (16 GB memory), Intel based CPU and 16 GB RAM. GPU acceleration was used for all deep learning experiments, while classical machine learning models were trained on the CPU.

The hyperparameters shown in Table I reflect how each deep learning architecture was individually optimized using Optuna, with early stopping determining the ideal training duration. The learning rate controls how quickly model weights are updated, with ChemBERTa requiring a much smaller value due the sensitivity of transformer layers, while the Graph Transformer and GAT operate effectively with higher rates. Dropout and L2 regularization act as key regularizers to prevent overfitting, with ChemBERTa using slightly stronger dropout and lower weight decay compared to the graph based models, which benefit from higher regularization due to their tendency to overfit on smaller datasets.

The difference in epoch counts emerged from the tuning process, as each model converges at a different rate and early stopping halted training once validation performance peaked. SMILES augmentations also vary, with graph models requiring more structural variations to learn robust molecular representations, while ChemBERTa needs fewer due to strong contextual embedding capabilities.

Finally, batch size differences arise from the computational complexity of each architecture. On the setup used in this study, ChemBERTa and Graph Transformer required approximately 15-20 minutes per epoch, resulting in a total 2-3 hours depending on early stopping, while inference for all DL models took less than one second per molecule.

TABLE I: MODEL SPECIFIC HYPERPARAMETERS USED FOR TRAINING THE DEEP LEARNING MODELS

Hyperparameter	ChemBERTa	GraphTransformer	GAT
Learning Rate (LR)	$2.78e^{-6}$	0.001905047	0.0034
Dropout	0.206	0.1477	0.1477
L2 Regularization	1×10^{-5}	2.65×10^{-4}	2.6×10^{-4}
Epochs	30	50	17
Augmentations	3	3	3
Batch Size	32	8	32

E. Evaluation Strategy and Model Selection

Solubility prediction is a regression task involving continuous values, while classification assigns solubility to discrete categories using thresholds. Regression performance was evaluated using RMSE, MAE and R^2, while accuracy and a hybrid accuracy metric assessed correct class assignment. By successfully identifying the borders between classes from molecular descriptors, traditional ML models achieved high classification accuracy. Their higher MAE values, however, imply that they have trouble making precise solubility values predictions. This issue likely arises from their dependence on handcrafted features as well as the relatively small size of the dataset. To reduce overfitting, applied hyperparameter tuning, regularization, data augmentation and validation strategies, setting aside 194 samples for validation.

Deep-learning models were analyzed using diagnostic plots. In both regression and classification tasks, ChemBERTa offers balanced performance, while Graph Transformer exhibits reduced numerical error on regression metrics.

ChemBERTa is more appropriate for real world solubility prediction, however, this benefit is accompanied by decreased stability and increased sensitivity to overfitting. These results show that numerical precision and classification accuracy need to be balanced, with deep learning models offering more precise quantitative solubility predictions.

IV. RESULTS

This section displays the solubility prediction performance of the Deep Learning and Machine Learning models that were put into practice. To guarantee consistent comparison across approaches, the main metrics used to evaluate all models were Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Root Square and Classification Accuracy.

A. Model Performance

Table II compares the performance of Machine Learning (ML) and Deep Learning (DL) models using Accuracy, Mean Absolute Error (MAE), Root-Mean-Square-Error (RMSE) and R^2 as the comparison metrics to evaluate the performance of solubility prediction models. The results indicate that there are differences in terms of performance among the two model categories. Gradient Boosting and XGBoost, similar to other ML approaches, were also capable of achieving very good accuracies greater than 0.93; however, the best accuracy for the Random Forest algorithm was 0.9397 which is indicative of an excellent ability to classify. Although they exhibited very high levels of class prediction capabilities, these models tended to present relatively elevated values of MAE compared to the DL models. Therefore, it appears that the evaluated models have complementary characteristics in terms of error-based performance. The Graph Transformer demonstrated superior numerical precision in estimating continuous solubility levels due to its lowest value of MAE and best regression performance. ChemBERTa demonstrated the best overall performance in terms of balance between high solubility ion accuracy and high classification accuracy. Traditional machine learning algorithms can be effective at learning coarse decision boundaries between solubility classes, which can result in elevated levels of classification accuracy. However, the elevated levels of MAE demonstrate the reduced numerical accuracy of the models when attempting to determine the exact magnitudes of solubility values. This illustrates the important trade-offs between the accuracy of classification and the accuracy of numerical prediction of solubility values and emphasizes the need to consider both metrics when evaluating solubility prediction algorithms.

TABLE II: PERFORMANCE COMPARISON OF ML AND DL MODELS

Models	Accuracy	MAE	RMSE	R^2
Random Forest	0.9397	3.7241	-	-
Gradient Boosting	0.9336	4.2582	-	-
XGBoost	0.9362	3.9897	-	-
ChemBERTa	0.8062	2.6893	6.343	10.1495
Graph Transformer	0.8895	1.4045	3.8640	0.6967
GAT	0.5969	2.3095	6.4315	0.1923

B. Performance Evaluation

To determine if ChemBERTa performed as expected we have created a series of key diagnostic charts that provide information on how well the model performs with respect to its ability to classify, discrimination between classes, and consistency in its training process. The clear and reliable separation of soluble and insoluble molecules as demonstrated by the diagonal concentration in the confusion matrix (Fig. 4), along with an area under the curve (AUC) of approximately 0.89 for the ROC curve (Fig. 2), indicate the ChemBERTa model has excellent discriminative strength at all thresholds. Additionally, there is evidence of no overfitting and the training-validation accuracy curve (Fig. 3) displays smooth curves that are close together and parallel to each other, both indicators of a steady and consistent learning process. All of this indicates that ChemBERTa has provided a very consistent and reliable performance in its molecular solubility classification task.

C. Web Application Deployment

ChemSol AI Web Tool provides researchers the ability to quickly determine the solubility of potential drug molecules prior to laboratory testing thereby assisting with the initial stages of drug discovery research. In addition to providing an immediate indication of expected solubility value for a molecule, ChemSol AI Web Tool can provide researchers with information regarding solubility class for a particular compound, thus allowing researchers to determine if a compound will be soluble or insoluble in water or another solvent of their choice. By interfacing with a backend prediction model through a low latency API, ChemSol AI Web Tool enables researchers to perform effective inference in real time. Examples of the user interface provided for chemists to enter chemical solutes/solvents and view predicted solubility data are presented in Fig. 5 and Fig. 6.

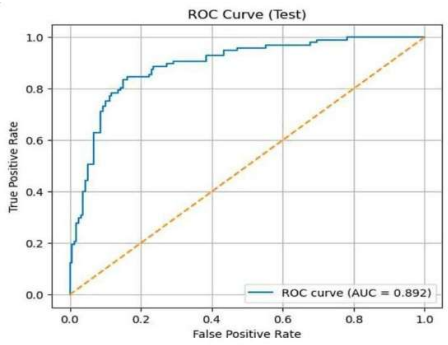

Fig. 2: ROC curve of ChemBERTa on the test set, showing an AUC of 0.892

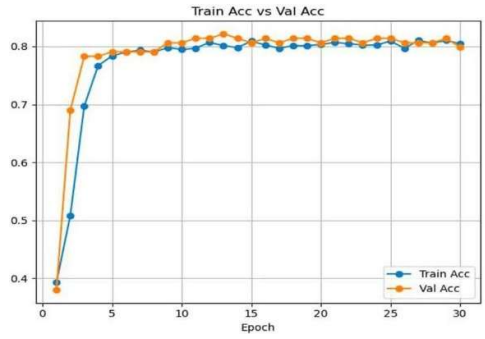

Fig. 3: Training vs. validation accuracy of ChemBERTa stabilizing around 0.80–0.82.

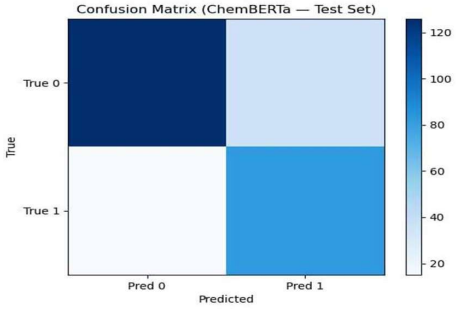

Fig. 4: Confusion matrix of ChemBERTa

Fig. 5: ChemSol AI interface displaying solute-solvent input, predicted solubility value and solubility class

Fig. 6: Summary panel showing molecular properties and final solubility prediction

E. Computational Efficiency

ChemBERTA and Graph Transformer required a total of 15-20 minutes per epoch; therefore, the total training time was approximately 2-3 hours. For molecule predictions, ChemBERTA is fast and takes less than a second. Overall, GPU acceleration increased the performance of both models and reduced the training times. Although the system can run on relatively moderate hardware, graph-based models do need a bit more processing power. Therefore, this framework provides an excellent trade-off in terms of speed and accuracy that is suitable for real-time solubility screening.

V. Discussion

The explanation as to why each model performed differently is due to the efficiency with which each model was able to gather chemical structural information. The Tree-based Models (Gradient Boosting, Random Forest, XGBoost) had been developed with designed descriptors that were able to capture coarse level patterns,

therefore these models yielded the highest amount of accuracy, however their large Mean Absolute Error (MAE), and Root Mean Square Errors (RMSE) showed a lack of ability to capture the fine grain solubility variation in the data set.

On the other hand, the Deep Learning Techniques (ChemBERTa, Graph Transformer), learned more complex structural and contextual representations directly from molecular graph or SMILES, this allowed for better R^2 values and smaller error rates even when there were less accurate predictions. The suggested deep models were especially good at capturing the relationships between the structural elements of a molecule, including how they interacted, without having to manually create features to describe them, however they required large amounts of data, much more computational power and less interpretable outputs. Additionally, failure cases occurred when molecules had unusual functional groups, long SMILES strings, or structural classes that were not well represented in the training data. Ethically, there needs to be consideration given to avoiding over-reliance on model output and ensuring that forecasts are not used to generate harmful chemicals; although, practically speaking, being able to make very accurate solubility predictions would greatly reduce the amount of experimentation needed in drug discovery. Overall, the results demonstrate a trade off between regression accuracy and classification accuracy, with the deep models producing more reliable quantitative predictions.

VI. Conclusion and Future Work

The study demonstrates that while both Graph Transformer and traditional machine learning architectures were able to achieve high levels of accuracy, they often suffer from overfitting and have limited ability to learn exact solubility values. Although ChemBERTa has slightly higher MAE and RMSE than both the Graph Transformer and the traditional machine learning architectures, ChemBERTa along with the Graph Transformer are able to consistently show more predictable learning capabilities than their traditional counterparts through their ability to understand and capture greater amounts of structural and contextual information. Therefore, it appears that as opposed to using artificially created descriptors, deep representations are capable of better representing the chemical space of molecules. Overall, the study findings suggest that deep learning architectures are a more effective way to create useful and reliable predictive systems and that generalizability and the ability to understand structure are more important for predicting solubility than raw accuracy alone.

Future studies will focus on the development of ensemble models, which combine predictions from multiple architectures (graph neural networks, transformer-based models, and machine learning models) to leverage each architecture's strengths. Using an ensemble approach, the solubility estimates can be made more reliable, reduce overfitting and increase robustness to different classes of chemicals. Some of the additional research areas include expanding the dataset to include more structurally diverse molecules; incorporating uncertainty-aware predictive methods, and exploring tools for enhancing model interpretability. The ultimate goal of all of these efforts is to create a more reliable, more accurate, and more interpretable solubility predictive system that can be applied in practical chemical and pharmaceutical contexts.

Acknowledgement

Dr. Lavanya D Kateel of the Chemistry Department at Canara Engineering College Benjanapadavu, Mangalore, is greatly appreciated by the authors for her wise counsel, which greatly advanced this study on AI based solubility prediction. Additionally, the authors thank the Institution for its unwavering encouragement and support during the completion of this work.

References

[1] W. Ahmad, H. Tayara, H. Shim and K. T. Chong, “SolPredictor: Predicting solubility with residual gated graph neural network,” International Journal of Molecular Sciences, vol. 25, no. 2, p. 715, 2024.

[2] W. Ahmad, H. Tayara and K. T. Chong, “Attention-based graph neural network for molecular solubility prediction,” ACS Omega, vol. 8, no. 3, pp. 3236–3244, 2023.

[3] Z. Fan, L. Chen, X. Wu, Z. Huang and L. Deng, “Enhancing predictions of drug solubility through multidimensional structural characterization exploitation,” IEEE Journal of Biomedical and Health Informatics, 2024.

[4] M. Li, H. Chen, H. Zhang, M. Zeng, B. Chen and L. Guan, “Prediction of the aqueous solubility of compounds based on light gradient boosting machines with molecular fingerprints and the cuckoo search algorithm,” ACS Omega, vol. 7, no. 46, pp. 42027–42035, 2022.

[5] A. D. Vassileiou, M. N. Robertson, B. G. Wareham, M. Soundaranathan, S. Ottoboni, A. J. Florence, T. Hartwig and B. F. Johnston, “A unified MLframework for solubility prediction across organic solvents,” Digital Discovery, vol. 2, no. 2, pp. 356–367, 2023.

[6] M. A. Ghanavati, S. Ahmadi and S. Rohani, “A machine learning approach for the prediction of aqueous solubility of pharmaceuticals: a comparative model and dataset analysis,” Digital Discovery, vol. 3, no. 10, pp. 2085–2104, 2024.
[7] F. Cenci, S. Diab, P. Ferrini, C. Harabajiu, M. Barolo, F. Bezzo and P. Facco, “Predicting drug solubility in organic solvent mixtures: a machine-learning approach supported by high-throughput experimentation,” International Journal of Pharmaceutics, vol. 660, p. 124233, 2024.
[8] S. M. Alshahrani, A. S. Alshetaili, M. M. Alfadhel, A. Belal, M. A. S. Abourehab, A. Al Saqr, B. K. Almutairy, K. Venkatesan, A. M. Alsubaiyel and M. Pishnamazi, “Optimization of tamoxifen solubility in carbon dioxide supercritical fluid and investigating other molecular targets using advanced artificial intelligence models,” Scientific Reports, vol. 13, no. 1, p. 1313, 2023.
[9] B. Huwaimel and A. Alobaida, “Anti-cancer drug solubility development within a green solvent: design of novel and robust mathematical models based on artificial intelligence,” Molecules, vol. 27, no. 16, p. 5140, 2022.
[10] N. Xue, Y. Zhang and S. Liu, “Evaluation of machine learning models for aqueous solubility prediction in drug discovery,” in Proc. 7th Int. Conf. Artificial Intelligence and Big Data (ICAIBD), pp. 26–33, 2024.
[11] I. Aitouhanni and A. Berqia, “SolvPredict: A comprehensive exploration of predictive models for molecule solubility,” in Proc. Int. Conf. Intelligent Systems and Computer Vision (ISCV), pp. 1–8, 2024.