SOLVATION META PREDICTOR

Annatu Amadu Somah

A thesis

submitted in partial fulfillment of the

requirements for the degree

Master of Science

University of Washington

2025

Committee:

David Beck

Joseph Hellerstein

Program Authorized to Offer Degree:

Chemical Engineering

University of Washington

Abstract

Solvation Meta Predictor

Annatu Amadu Somah

Chair of the Supervisory Committee:

David Beck

Department of Chemical Engineering

Accurately predicting the aqueous solubility of organic molecules is essential in a wide range of scientific and industrial domains, including drug development, food, and energy storage. This study builds upon prior work by Panapitiya et al. by introducing a multi-stage ensemble learning framework to enhance the predictive performance of solubility models using the SOMAS dataset. The dataset comprises 11,696 molecules with diverse structural and physicochemical properties, including 2D, 3D, and quantum descriptors. Three base models, a Molecular Descriptor Model (MDM), a Graph Neural Network (GNN), and a SMILES model developed by Panapitiya et al. were utilized and evaluated using RMSE, MAE, R^2, and Spearman correlation. Among individual models, MDM achieved the strongest performance, but ensemble methods consistently outperformed standalone models. Simple averaging improved predictive accuracy, while Optuna-based ensemble weight optimization yielded the best overall results. Additionally, a Mixture of Experts (MoE) architecture was implemented to dynamically weight model outputs based on structural input features, demonstrating strong performance and scalability. This work highlights the value of combining diverse molecular representations and advanced ensemble techniques,

providing a robust, adaptive framework for high-accuracy solubility prediction and future data-driven molecular design.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

1.0 INTRODUCTION

Solubility is a fundamental physicochemical property that significantly influences numerous aspects of daily life and scientific research. From the simplicity of dissolving sugar in coffee to complex industrial and environmental applications, solubility plays a critical role. In pharmaceutical development, it directly impacts drug bioavailability, efficacy, and formulation strategies[1][2][3]. In environmental science, solubility governs the mobility, persistence, and remediation of pollutants, shaping effective environmental protection and cleanup efforts[4]. Within the food industry, it affects taste, texture, stability, and the performance of preservatives[5]. Moreover, in the field of energy storage and conversion such as in the design of electrolytes for batteries and the optimization of photovoltaic systems solubility is a key factor in determining material compatibility, efficiency, and system stability[3][6]. Understanding and predicting solubility, therefore, remains essential across diverse scientific and engineering disciplines.

CHAPTER 2

2.0 LITERATURE REVIEW

The determination of solubility, particularly for crystalline substances can be estimated from its absolute solid free energy and excess solvation free energy[7]. The critical importance of determining the solubility of substances is well recognized. However, the experiments, calculations, and predictions involved in assessing solubility, are labor-intensive and prone to human error. The complexity and multidimensional aspects of melting processes, coupled with considerations of chemistry, thermodynamics, kinetics, and morphology, render the prediction of solubility a challenging endeavor. Despite the availability of numerous statistical and machine learning techniques, the complexity of computations remains unmitigated, largely because each relevant feature is either challenging to acquire or necessitates the use of an extensive number of features. Consequently, a straightforward, accurate, and dependable method for predicting solubility based on experimental data would undoubtedly be beneficial across various research domains where precise solubility predictions are crucial[3].

Methods of determining Solubility

Historically, solubility prediction has relied on three main approaches, each presenting distinct challenges. Quantum mechanical methods, such as ADF COSMO-RS (Conductor-like Screening Model for Real Solvents), utilize density functional theory (DFT) to estimate chemical potentials and thermodynamic properties, but they are computationally expensive and often struggle to provide quantitatively accurate comparisons with experimental data. The General Solubility Equation (GSE), particularly the refined version based on Jorgensen and Duffy's work, offers a simplified, less error-prone alternative; however, it demands extensive validation across diverse solvent-solute combinations, which can be resource-intensive. More recently, machine learning and statistical models, including QSAR/QSPR (Quantitative Structure-Activity/Property Relationships) and deep learning-based approaches, have gained traction for their computational efficiency in predicting solubility. While these models can yield rapid predictions, they frequently face limitations in transferability and generalizability, often due to constraints such as fixed solvent environments in training datasets and an incomplete understanding of molecular features that govern solubility[3].

2.0.1 Recent Developments in ML Models to predict Solubility

In addition to QSPR methods, several mechanistic machine learning models have also been developed to predict solubility. Some of these latest developed models are SolTranNet, Delfos, SolPredictor, AquaPed, MLSolvA and SolvBERT. These models utilize cutting-edge methods such as deep learning models, Graph Neural Networks, natural language processing (NLP) techniques, and attention-based mechanisms, establishing new standards in the industry.

SoltranNet

SolTranNet, a molecule attention transformer model which is based on the MAT (Multi Agent Transformer) architecture that predicts aqueous solubility using a molecule's SMILES representation. Contrary to typical expectations, its findings reveal that larger models underperform in this task. SolTranNet's final architecture, which comprises only 3,393 parameters, surpasses traditional linear machine learning methods in accuracy. Specifically, SolTranNet achieved a root-mean-square error (RMSE) of 1.459 on the AqSolDB during a 3-fold scaffold split cross-validation, and a RMSE(Root Mean Square Error) of 1.711 on a separate withheld test set[8].

SolPredictor

The model employs a residual graph neural network convolution (RGNN) architecture designed to effectively capture long-range dependencies within graph-structured data. Utilizing residual connections, the model facilitates the flow of information across various layers, ensuring that essential features and patterns distributed throughout the network are captured and retained. Compiled from the two largest datasets currently available, SolPredictor uses the simplified molecular-input line-entry system (SMILES) for representing molecules. It achieves a Pearson correlation coefficient (R^2) of 0.79±0.02 and a root mean square error (RMSE) of 1.03±0.04 through ten-fold split cross-validation. The effectiveness of SolPredictor was further assessed across five independent datasets. Comprehensive analyses, including error analysis, hyperparameter optimization, and model explainability studies, were conducted to identify the molecular features most critical for accurate prediction[9].

SolvBERT

The SolvBERT is a BERT-based (Bidirectional Encoder Representations from Transformers) regression model which predicts the solvation free-energy as well the solubility. SolvBERT processes the solute and solvent by interpreting the SMILES into a vectorized form. This model underwent unsupervised pre-training using a substantial database filled with computational data on solvation free energies. The model was pre-trained and fine-tuned with the CombiSolv QM (containing approximately 1 million solute-solvent pairs). Transfer learning was done on the Combisolv-Exp-8780 and solubility database from Boobier et Al. SolvBERT model was fine-tuned using the Combisolv-8780 data. SolvBERT demonstrated high accuracy in its predictions, particularly excelling in out-of-sample tests. The model's proficiency in predicting solubility and solvation-free energy was underscored by its successful application of transfer learning, transitioning from pre-training on computational data to fine-tuning with experimental data. Although extrapolating to new compounds not represented in the training set posed a significant challenge, SolvBERT still delivered promising results, showing both accuracy and reliability in these scenarios[10].

Delfos

Delfos(deep learning model for solvation free energies in generic organic solvents) is an innovative machine-learning-based QSPR method combined with a recurrent neural network (RNN) designed to predict solvation free energies across a variety of organic solute and solvent systems. The model is made up of three sub-neural networks: the encoder network for solute and solvent and the predictor network. The two distinct encoder networks for solvents and solutes, which leverage word embeddings and recurrent layers to capture the structural characteristics of the compounds involved. These networks are enhanced with an attention mechanism that identifies critical substructures from the recurrent neural networks' outputs. Subsequently, a predictor network utilizes the encoded features to compute the solvation free energy for specific solvent-solute pairs. Through comprehensive analysis of 2,495 solute-solvent pairs of 418 solutes and 91 solvents, Delfos not only demonstrates accuracy comparable to leading computational chemistry methods but also provides insights into the substructures significantly influencing the solvation process. The model was evaluated using the Minnesota Solvation DataBase[11].

AquaPred

This study presents an advanced machine learning framework for predicting molecular solubility, employing an attention-based Graph Neural Network (GNN). Given the central role of solubility in drug discovery and development, the proposed model holds significant value for pharmaceutical research by improving both the accuracy and efficiency of aqueous solubility prediction. The model leverages SMILES (Simplified Molecular Input Line Entry System) strings for molecular representation and incorporates various GNN architectures, including Simple Graph Convolution (SGConv), Graph Isomorphism Network (GIN), Graph Attention Network (GAT), and the Attentive FP network. Among these, the Attentive FP model was selected for its superior performance, combining an attention mechanism for capturing intermolecular interactions with gated recurrent units (GRUs) to model intramolecular properties. Trained on a dataset of 9,943 compounds, the model was further evaluated on 62 anticancer compounds, demonstrating strong generalization and robustness. It achieved a Pearson correlation coefficient (R^2) of 0.52 and a root-mean-square error (RMSE) of 0.61, indicating its effectiveness in accurately predicting molecular solubility[12].

MLSolvA

MLSolvA introduces a novel machine learning approach for predicting solvation free energy by modeling pairwise atomistic interactions. The architecture is framed as a linear regression task, where solvation energy is computed using the inner product of atomic feature vectors extracted via two encoding functions applied to the molecular structure. This design enables a more interpretable representation of intermolecular interactions. Two neural network encoders BiLM (a recurrent neural network-based bidirectional language model) and Graph Convolutional Network (GCN) were evaluated for encoding molecular inputs. The model was trained and validated using 6,239 experimental solvation energy measurements spanning 935 organic solvents and 146 organic solutes, sourced from the FreeSolv and Solv@TUM databases. To improve generalization, pre-training was conducted on a large corpus of organic compounds from the ZINC15 database. MLSolvA achieved strong predictive performance, with mean unsigned errors (MUE) of 0.19 kcal/mol using the BiLM/LSTM encoder and 0.22 kcal/mol using the GCN encoder. However, the model's performance declined in extrapolation tasks particularly when predicting

solvation energies for unseen chemical scaffolds highlighted by experiments using scaffold-based data splits that revealed a notable degradation in accuracy[13].

CHAPTER 3

3.0 OBJECTIVE

This research builds upon the foundational study titled "Evaluation of Deep Learning Architectures for Aqueous Solubility Prediction" by Gihan Panapitiya, Michael Girard, Aaron Hollas, Jonathan Sepulveda, Vijayakumar Murugesan, Wei Wang, and Emily Saldanha[14]. Their work offers a comprehensive assessment of deep learning models in the context of aqueous solubility prediction an area of critical importance in both chemical and pharmaceutical sciences. Expanding on this foundation, the present study seeks to enhance predictive performance by integrating advanced ensemble learning strategies.

Our approach employs a multi-stage methodology. Initially, the K-Nearest Neighbors (KNN) algorithm is utilized to select the most appropriate model for each instance in the test set. Following this, a simple averaging ensemble is applied to aggregate predictions. To further improve predictive accuracy, we leverage Optuna, a hyperparameter optimization framework, to fine-tune model weights in accordance with their individual performance, guided by robust cross-validation techniques for improved generalization.

To further elevate model precision, we incorporate a Mixture of Experts (MoE) framework, which dynamically adjusts the weighting of individual models based on the specific features of each test sample. This adaptive mechanism allows for more nuanced and instance-specific predictions. By extending the deep learning architectures explored in the original study, our work aims to advance the state-of-the-art in solubility prediction and contribute practical, high-accuracy tools for real-world chemical and pharmaceutical applications.

CHAPTER 4

4.0 METHODOLOGY

4.0.1 Data

The dataset used in this study is sourced from the SOMAS (Solubility of Organic Molecules in Aqueous Solution) database, which is based on the extensive compilation originally assembled by Gao et al. to support machine learning-based solubility prediction, particularly for applications in aqueous organic redox flow batteries This dataset integrates aqueous solubility data for 11,696 molecules from multiple reputable sources, including ChEMBL, GDB-13 & GDB-17, NIST Standard Reference Database 106(NIST SRD106), OChem, eChemPortal, and data from Cui et al's work. These molecules exhibit a wide range of structural diversity, with atom counts varying from 1 to 273 and molecular masses ranging from 16 to 972 g/mol. The dataset also encompasses a broad spectrum of aqueous solubility values, Solubility values span from extremely low ($\leq 10^{-10}$ mg/L) to very high ($>10^{6}$ mg/L)[15].

Solubility is represented as Log S, the base-10 logarithm of solubility in mol/L, with distributions of Log S and molecular mass, and in the figure below.

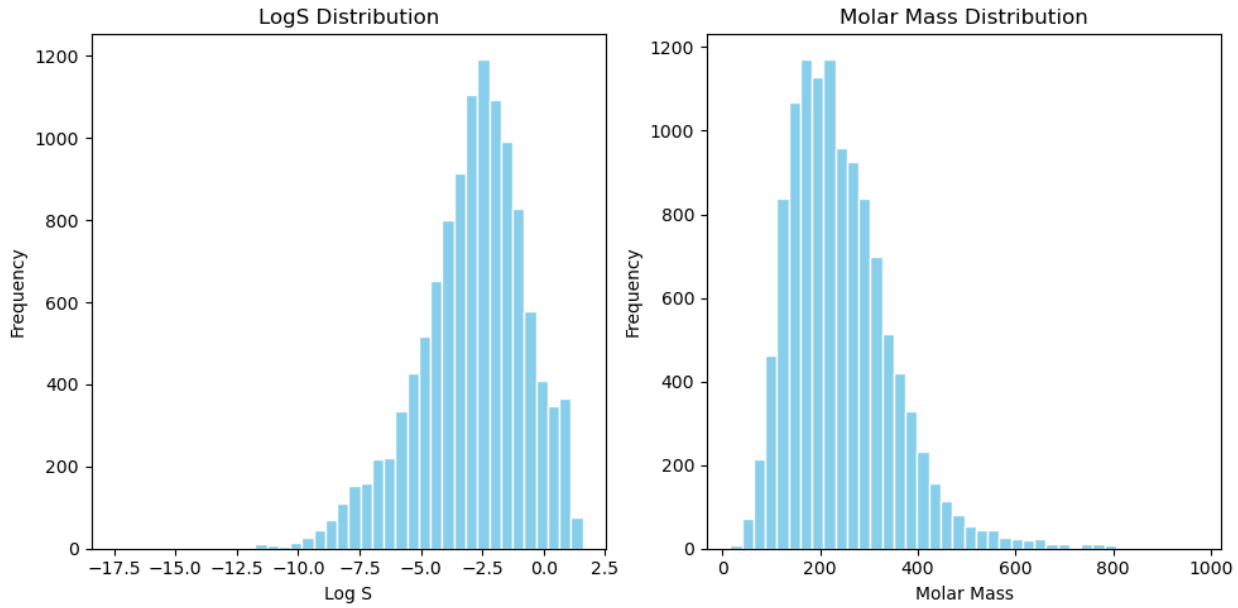

Figure 1:Distribution of Log S and Molar Mass in dataset

To enable predictive modeling, Panapitiya et al.[14] developed a comprehensive molecular feature set capturing structural, topological, geometric, and quantum-level properties. The process began with the generation of 2D molecular descriptors using the Mordred package, which can compute up to 1,613 descriptors from molecular graphs. Due to failures in descriptor generation for some molecules, only 743 descriptors that were successfully computed across the entire dataset were retained. These included features such as atom counts, topological indices, electro topological states, and more, offering a rich representation of molecular connectivity and functionality.

In addition to 2D descriptors, 3D molecular descriptors were calculated to incorporate spatial and conformational information. Atomic coordinates were generated using the Pybel package, followed by structural optimization with the MMFF94 force field. Molecules for which this process failed were removed. Using the optimized coordinates, radial distribution features were computed by counting atoms in six concentric layers around the molecular centroid. Furthermore, geometric shape descriptors were calculated using the method proposed by Ballester and Richards, involving statistical moments (up to order 10) of atomic distances from the centroid and other reference points. Additional spatial metrics, such as the molecular volume derived from the ConvexHull algorithm in SciPy, were included, resulting in a total of 37 distinct 3D descriptors.

To enrich the chemical information, a fragment-based representation was also constructed. Using RDKit, 59 fragments were selected including the most common functional groups and structural motifs in the dataset, such as aromatic rings and halogenated compounds. This representation allowed the model to explicitly learn from substructure presence and frequency.

Finally, quantum chemical descriptors derived from density functional theory (DFT) were incorporated to capture electronic properties. These included solvation energy (kcal/mol), molecular volume ($Å^3$), surface area ($Å^2$), dipole moment (Debye), dipole moment per volume, and quadrupole moments, all calculated using the NWChem package. Due to the high computational cost of DFT calculations, this quantum descriptor set was limited to a subset of 7,764 molecules containing no more than 83 atoms [14].

The combination of 2D, 3D, fragment-based, and DFT derived descriptors provided a multi-dimensional, chemically informed feature space, laying a robust foundation for developing and evaluating deep learning models for aqueous solubility prediction. This elaborate feature engineering pipeline represents one of the most comprehensive solubility modeling datasets

available and serves as the cornerstone for subsequent model development and ensemble strategies in this study.

4.0.2 MODELS.

In Panapitiya et al.'s[14] study, four deep learning models were developed to predict molecular solubility by learning patterns between structural and physicochemical properties and experimentally measured solubility values. These models leveraged different molecular representations, including structural/electrochemical feature vectors, SMILES strings, molecular graphs, and 3D atomic coordinates each paired with a deep learning architecture suited for that representation. Specifically, the four models developed were the SchNet model (3D coordinates), the MDM model (molecular descriptors), the SMILES model, and the GNN model. Their work evaluated which representation and corresponding model architecture were best suited for high-accuracy solubility prediction. In the present study, we employed three of these models MDM, SMILES, and GNN to support our own objectives in solubility prediction.

4.0.2.1 Molecular Descriptor Model

The Molecular Descriptor Model (MDM), developed by Panapitiya et al.[14] is based on a fully connected feedforward neural network designed to predict aqueous solubility from a rich vector of structural and physicochemical features. This feature set includes 2D molecular descriptors, 3D spatial descriptors, and fragment-based features, capturing diverse chemical characteristics relevant to solubility. The features in the training, validation, and test sets were standardized to zero mean and unit variance using parameters derived from the training data, thereby preventing data leakage. The model architecture is implemented as a Sequential neural network beginning with a dense input layer of 128 neurons activated by a sigmoid function, introducing non-linearity. This is followed by a dropout layer with a rate of approximately 10.7% to mitigate overfitting. The second hidden layer comprises 576 neurons with a ReLU activation function, followed by a second dropout layer with a rate of approximately 60.3% to further support generalization.

The network concludes with a single-node dense output layer using a linear activation function, suitable for continuous regression outputs. The model is trained using the RMSprop optimizer, with a learning rate specified externally through a configuration file and optimized using the mean squared error (MSE) loss function, which aligns with the regression objective of solubility prediction.

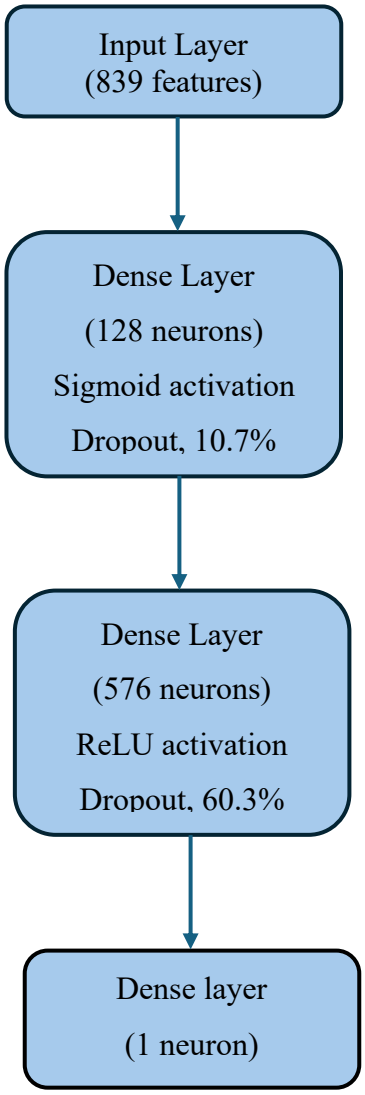

Figure 2: Model Architecture of MDM model

4.0.2.2 SMILES MODEL

The SMILES (Simplified Molecular Input Line Entry System) model, also developed by Panapitiya et al.[14] employed in this study is the simplest, utilizing a character-level sequence modeling approach to predict molecular solubility. SMILES strings, which encode molecular structures as linear text sequences, are first tokenized into individual characters and padded to a uniform length. These character indices are then passed through an embedding layer that projects them into a 960-dimensional continuous vector space, allowing the model to learn meaningful representations of molecular substructures.The embedded sequence is processed through a unidirectional LSTM(Long-Short Term Memory) layer with 256 units, capable of capturing long-range dependencies within the molecular sequence. A dropout layer with a rate of approximately 8.6% is applied to reduce overfitting. This is followed by a bidirectional LSTM

layer, also with 256 units, which enhances contextual learning by processing the sequence in both forward and backward directions. A second dropout layer, with a higher rate of approximately 47.3%, is applied to further improve generalization.

The output from the recurrent layers is passed through a series of dense layers, starting with 704 neurons and employing various non-linear activation functions such as ReLU, SELU, or sigmoid, depending on the configuration. The network concludes with a final dense output layer with linear activation, suitable for generating continuous solubility predictions. The model is trained using the Adam optimizer, with hyperparameters such as learning rate and layer dimensions supplied dynamically from an external configuration. The mean squared error (MSE) loss function is used to guide learning, aligning with the regression nature of the task.

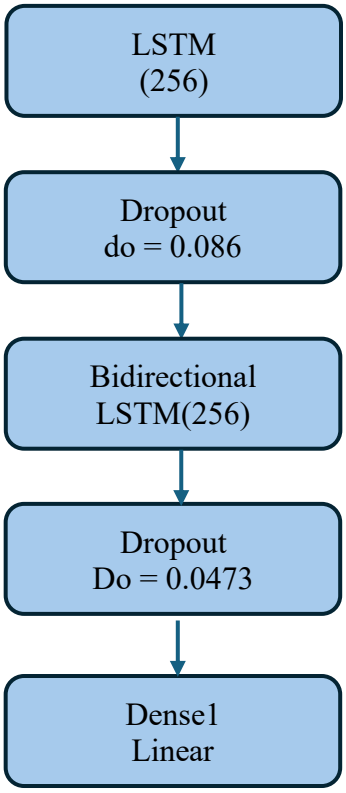

Figure 3: Model Architecture of SMILES model.

4.0.2.3 GNN

The Graph Neural Network (GNN) model implemented in the study is designed to predict molecular solubility by learning from graph-structured molecular representations, where atoms are modeled as nodes and chemical bonds as edges. Each node is initialized using atomic features

defined by the atom features function in the DeepChem library, which includes attributes such as atomic symbol, degree, hybridization, formal charge, hydrogen count, and aromaticity. These features are one-hot encoded or boolean-valued, providing a rich representation of each atom's chemical context.

The architecture employs a stack of graph convolutional layers (GCNConv) using the PyTorch Geometric library, beginning with a GCNConv layer that transforms the 65-dimensional input node features into a 256-dimensional space. This is followed by successive graph convolutional layers with increasing feature sizes (320, 448, and 512 units), each performing message passing to aggregate information from neighboring nodes. Additionally, EdgeConv layers are used to capture edge-based interactions, allowing the model to incorporate bond-level information during node updates.Non-linear activation functions, including ReLU, SELU, and Sigmoid, are applied at various stages to introduce non-linearity and improve model expressiveness. Dropout is strategically applied after each graph and dense layer, with rates reaching up to 94%, to mitigate overfitting and enhance generalization.

To obtain a graph-level representation, the model applies a global pooling operation either global add pool or global mean pool that aggregates node features across the entire molecular graph. This pooled representation is processed through a series of fully connected layers to refine the learned features. The final prediction is generated via a linear output layer, suitable for continuous solubility regression. The model is trained to minimize mean squared error (MSE), using a learning rate defined via external configuration.

This GNN architecture, inspired by the framework described in Panapitiya et al.[14], effectively leverages both local atomic environments and global molecular structure through message passing, making it well-suited for molecular property prediction tasks such as solubility estimation.

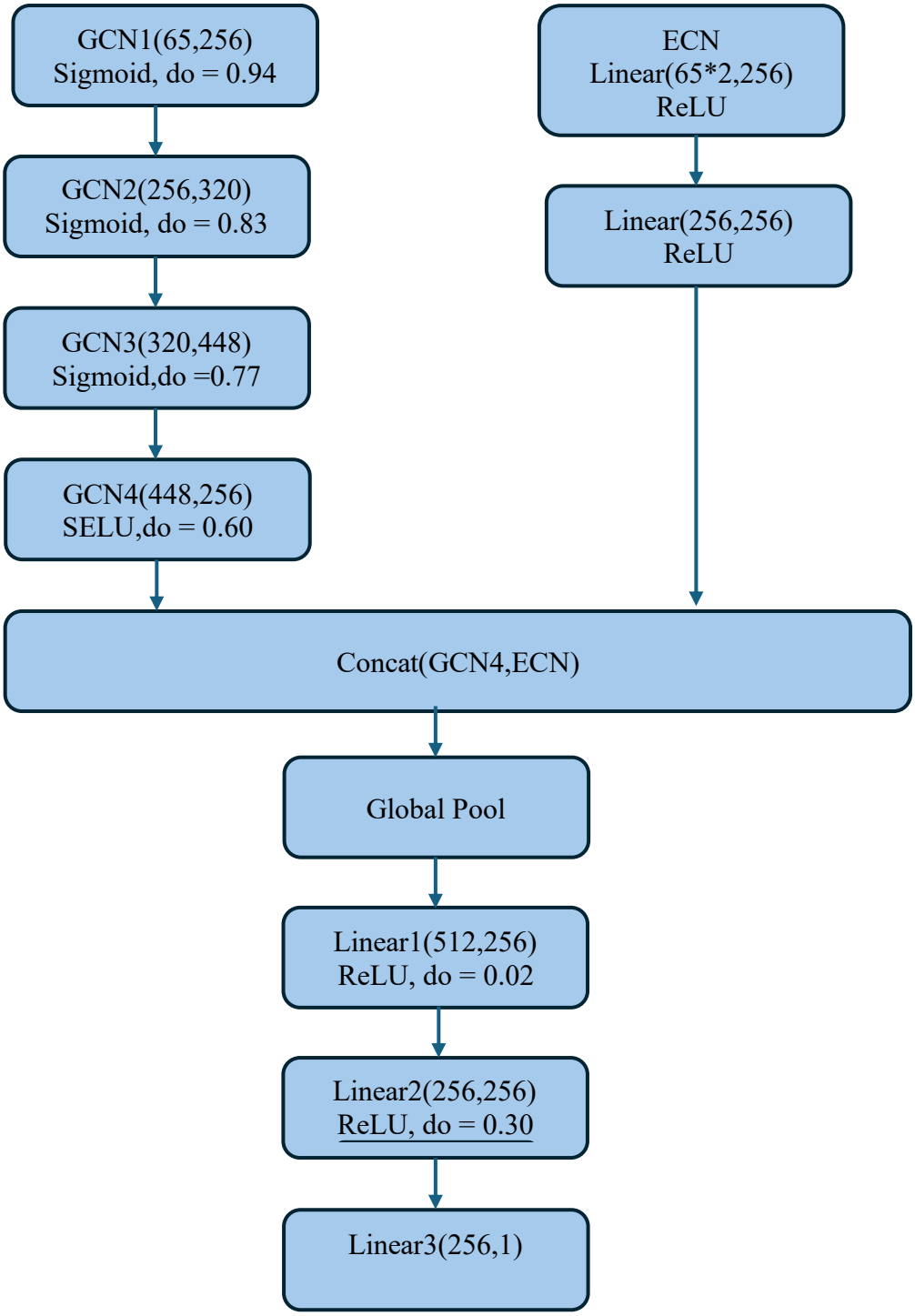

Figure 4: Model Architecture of GNN model[14].

4.0.3 KNN

The k-nearest neighbors (KNN) algorithm is a non-parametric, supervised learning method that predicts outcomes based on the proximity of a given data point to others in the feature space [16][17]. In this study, KNN is employed not as a direct predictor of solubility, but as a dynamic

model selection mechanism to determine which among the three solubility prediction models MDM, GNN, and SMILES is best suited for each test molecule based on molecular similarity. These features are essential for capturing structural and chemical similarities between molecules. For every molecule in the test dataset, the 'K' most similar molecules in the validation dataset are identified using Euclidean distance in the defined feature space. This distance is calculated as:

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$ Equation 1 [17].

where x_i and y_i represent the coordinates input which in this case represent molecules.

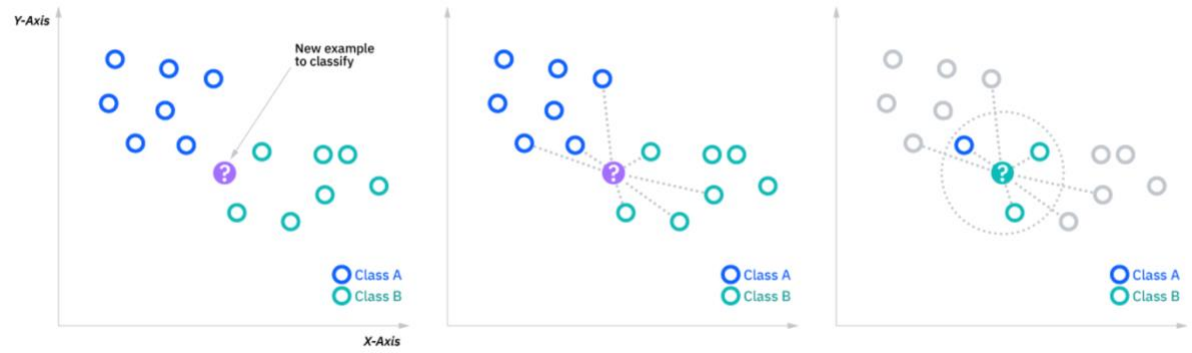

Figure 5: Illustration of how KNN works [17].

Once the nearest neighbors are identified, the predicted solubility values (LogS) for those molecules are retrieved from each of the three models (MDM, GNN, and SMILES). The absolute error between each model's prediction and the experimentally measured solubility is computed for the K neighbors. The model that yields the lowest average absolute error across the K neighbors is selected as the best-performing model for the corresponding test molecule.

This KNN-based approach allows for instance-specific model selection, adapting the prediction strategy to the unique characteristics of each molecule. By leveraging local structural similarity within the dataset, this method enhances both the accuracy and robustness of solubility predictions. Ultimately, it enables the ensemble to dynamically harness the strengths of each individual model in context-specific scenarios, optimizing the overall performance of machine learning applications in molecular property prediction[18].

4.0.4 ENSEMBLE METHODS

Ensemble methods are powerful predictive modeling techniques that combine outputs from multiple models to produce more accurate and robust predictions. By aggregating predictions through simple approaches like averaging or voting, or more advanced strategies that weight model contributions ensembles can leverage the strengths of individual models while mitigating their weaknesses. The primary benefits of ensemble learning are improved predictive performance and increased robustness, as they reduce the variability in predictions and smooth out extreme outcomes. This makes ensembles particularly valuable in applications where high accuracy and consistent performance across diverse datasets are critical [19].

4.0.4.1 Simple Averaging Ensemble

Simple averaging is a foundational ensemble learning technique where predictions from multiple machine learning models are combined to produce a more accurate and robust output. This method is particularly effective when the individual models offer complementary perspectives on the data, as it reduces variance and mitigates the risk of overfitting associated with any single model [20].

In this study, simple averaging is applied to the predictions of three independently trained models: the Molecular Descriptor Model (MDM), the Graph Neural Network (GNN), and the SMILES based model. Each model predicts the logarithm of the solubility (LogS) for every molecule in the test dataset. The ensemble prediction for each molecule is computed by taking the arithmetic mean of these individual predictions, as expressed in Equation 2.

$$Predicted\ Log\ S\ = \frac{Pred\ MDM + Pred\ SMILES\ Model + Pred\ GNN\ model}{Number\ of\ models} \quad \text{Equation 2}$$

This results in a single consensus prediction for each test instance, based on the combined output of all participating models. The performance of the ensemble is then evaluated by comparing its predictions against experimental solubility values, allowing for an assessment of its effectiveness in reducing prediction error relative to individual model outputs.

Simple averaging offers several advantages. It is straightforward to implement, requiring no additional training once the individual models are complete. It helps to cancel out individual model biases or noise, leading to more stable and generalizable predictions[21][22]. By aggregating outputs from diverse model architectures, the ensemble benefits from model diversity, which often improves predictive performance over any single model[21].

However, this method also has limitations. It assumes all models contribute equally, regardless of their individual performance. In scenarios where certain models are significantly more accurate, simple averaging does not weight their outputs more heavily, potentially reducing ensemble effectiveness[23]. Furthermore, it lacks adaptability, as it does not account for context-specific performance variations across different molecules or regions of the input space [24].

Despite these limitations, simple averaging remains a widely used and effective ensemble strategy, particularly when model independence and diversity are present. In the context of this work, it serves as a valuable baseline for combining model predictions in chemical property estimation tasks such as aqueous solubility prediction.

4.0.4.2 Ensemble Weight Optimization Using Optuna

Optuna is an open-source software framework for automatic hyperparameter optimization, designed to efficiently search high-dimensional parameter spaces and identify optimal configurations [25]. In this study, we utilize Optuna v3.6.1 to fine-tune the weights assigned to each model GNN, SMILES, and MDM within an ensemble framework for solubility prediction. The goal is to determine the optimal combination of model weights that minimizes the overall prediction error.

The optimization process is structured around an objective function, where Optuna is tasked with exploring the space of possible weight combinations. During each trial, Optuna proposes a set of weights for the individual models within the range [0, 1]. These weights are then normalized to ensure that their sum equals one, preserving proportional influence across the models and preventing skewed ensemble outputs.

In every trial, the ensemble prediction is computed as a weighted average of the predictions from the three base models, using the weights suggested by Optuna. The resulting ensemble predictions are then evaluated against the true solubility values using the mean squared error (MSE) as the objective metric. Optuna systematically searches for the combination of weights that yields the lowest MSE, iterating through a predefined number of trials to thoroughly explore the parameter space.

Upon completion of the optimization process, the best-performing weight configuration is selected. These optimized weights reflect the most effective way to balance contributions from the GNN, SMILES, and MDM models to minimize prediction error. This approach enhances the

ensemble's ability to generalize by leveraging the unique strengths of each model while compensating for their individual weaknesses.

By automating the ensemble tuning process, Optuna provides a principled and reproducible method for improving model performance. This optimization not only leads to more accurate and robust solubility predictions but also ensures that the ensemble model is well-calibrated to the specific structure and distribution of the validation dataset, making it highly applicable in real-world chemical informatics tasks.

4.0.4.3 Ensemble Weight Optimization using Cross Validation with Optuna

Cross-validation (CV) is a robust statistical technique that systematically partitions the dataset into multiple subsets, or folds, where each fold serves as a validation set while the remaining data are used for training [26]. This rotation ensures that every data point is utilized for both training and validation, thereby reducing the bias that may arise from relying on a single train-test split. In the context of ensemble learning, CV plays a critical role in enhancing model reliability and generalizability. By exposing the model to diverse partitions of the data, cross-validation helps mitigate overfitting. It provides a more realistic estimate of model performance by requiring the ensemble to generalize across multiple independent scenarios. If the model exhibits consistent performance across all folds, it indicates robustness and suggests that the model is less likely to underperform on unseen data[27].

In this study, cross-validation is used in conjunction with Optuna to optimize the weights assigned to individual base models (GNN, SMILES, and MDM) within the ensemble. For each trial in the Optuna search space, weights are proposed and then evaluated using k-fold cross-validation. In each fold, the ensemble's weighted predictions are compared against the true solubility values, and the mean squared error (MSE) is calculated. The average MSE across all folds serves as the objective function for Optuna to minimize.

This integrated approach enables the optimization process to identify weight combinations that generalize well across different subsets of the data, rather than being overfitted to a single partition. As a result, the ensemble becomes more robust and capable of delivering reliable predictions in practical applications. Using cross-validation within Optuna ensures that the final ensemble weights are not only optimal with respect to a particular training configuration but are also statistically validated for broader use.

4.0.4.4 Mixture Of Experts

The Mixture of Experts (MoE) architecture, originally introduced by Jacobs et al. (1991) in their seminal paper Adaptive Mixture of Local Experts [28], is a modular neural network design where multiple specialized sub-networks called experts are dynamically selected to handle different regions of the input space. A central component of this architecture is the gating network, which determines how inputs are routed to experts. For a given input, the gating network computes a vector of probabilities indicating the contribution of each expert to the final output. This is commonly achieved using a softmax activation function, which ensures that the gating weights are positive and sum to one. The output of the MoE model is then computed as a weighted sum of the experts' outputs, with the weights derived from the softmax distribution [28][29].

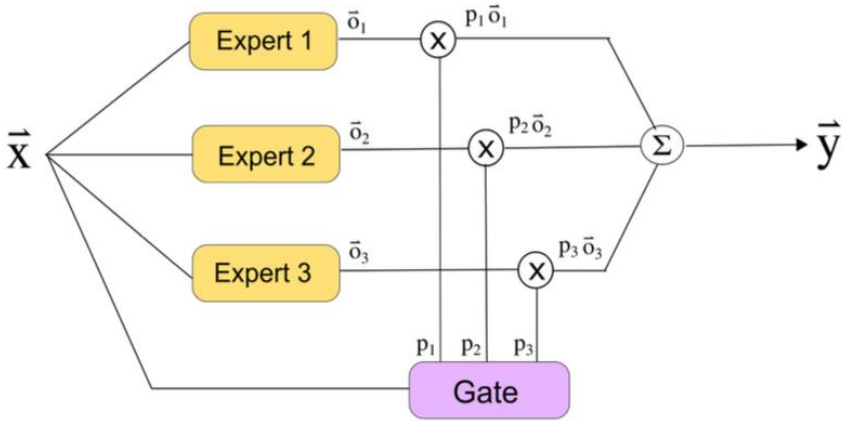

Figure 6: Model Architecture of Mixture of Experts with 3 experts and a gate [29].

The output of the model y=p1*o1 + p2*o2 +p3*o3 where p1, p2,p3 are gate outputs and o1,o2,o3 are expert outputs[29]

Simply put the output from the MOE model, y is summarized in equation 3 below

$$y = \sum_{i=1}^{n} G(x)_i E_i(x)$$ Equation 3[30].

Where G(x) is the gate output which is computed using equation 4 below

$$G_\sigma(x) = Softmax(x * W_g)\ Equation\ 4$$[30].

$$Softmax\ Activation\ Function = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}\ Equation\ 5[31].$$

'x' represents the input to the gate and expert, W_g is the weight matrix, often referred to as the "SoftMax weight matrix", z_i represents the input to the softmax function, and the denominator in equation 5 is the sum of all the raw scores in the output layer.

The concept of conditional computation underpins MoE models only a subset of experts is activated per input, which significantly reduces computational cost while preserving model expressiveness. In modern implementations, such as Switch Transformers and GShard [30][31]. this principle has been scaled to enable large language models (LLMs) with trillions of parameters. For instance, DeepSeek-R1 [32]utilizes 671 billion parameters, yet activates only 37 billion per forward pass, thanks to a learned gating mechanism that routes tokens to a small number of relevant experts. This selective activation supports faster inference, lower energy consumption, and modular scalability critical traits for large-scale AI systems. MoEs have been successfully applied in natural language processing (NLP), machine translation, image captioning, continual learning, and multi-task learning [29][30][31].

Recent advancements have introduced alternative gating strategies to improve load balancing and expert utilization. Approaches like Noisy Top-k Gating [30], Hash-based routing, and Expert Choice Routing [31] aim to mitigate the common challenge of expert collapse, where certain experts dominate the routing distribution. Expert Choice, in particular, flips the conventional token-to-expert routing paradigm by allowing each expert to select a fixed number of top-scoring tokens, improving load balance without relying on auxiliary loss terms. Other variations such as Deep Mixture of Experts (DMoE) [33]use layer-specific gating to enable exponentially many expert paths, dynamically constructing computation graphs that reflect the input's structure.

Despite their advantages, MoEs are not without challenges. These include training instability, inefficient memory usage due to inactive but resident experts, and overfitting during fine-tuning [30]. Techniques such as auxiliary losses, expert capacity thresholds, and load balancing regularization (e.g., Softplus and Gaussian noise injection in the gating network) have been explored to address these limitations. Nevertheless, MoEs remain among the most scalable and flexible architectures for managing model complexity in both pretraining and fine-tuning contexts.

In the context of this study, we adopt a simple Mixture of Experts (MoE) framework to construct a gated ensemble model composed of the three pretrained solubility prediction models: a Graph Neural Network (GNN), a Molecular Descriptor Model (MDM), and a SMILES-based LSTM model. Each model serves as an expert, and a separate gating network is trained to assign dynamic weights to each expert's prediction for a given input molecule. The gating system uses structural features such as the number of atoms, number of bonds, and average node degree extracted from the molecular graph representation to assess similarity and relevance among inputs. These features are standardized and embedded into a uniform vector space to improve learning effectiveness[34][35]. The gating network, composed of a fully connected layer followed by a softmax activation, outputs a normalized weight distribution over the experts. Using this softmax-based scoring system, the model computes a weighted sum of predictions, where each expert's contribution is scaled by its relevance to the input. Importantly, during MoE training, the expert models remain "frozen" to preserve their individual strengths while only the gating network parameters are updated. The model is trained using Mean Squared Error (MSE) loss and optimized via the Adam optimizer with a learning rate of 0.001. This approach embraces the principles of expert specialization and conditional computation, enabling the ensemble to adaptively select the most suitable model based on input characteristics, thereby improving prediction accuracy, robustness, and generalizability across diverse molecular structures.

The complete dataset, comprising 11,696 molecules, was divided into 85% for training, 7.5% for validation, and 7.5% for testing. All models were trained with continuous monitoring of the mean squared error (MSE) on the validation set, and the model parameters corresponding to the lowest validation error were retained. Training was halted using an early stopping criterion, where optimization ceased if the validation error failed to improve over 25 consecutive steps. This strategy was employed to prevent overfitting and to ensure that model generalization was prioritized.

CHAPTER 5

5.0 RESULTS

To assess model performance, each predictive model was trained using a comprehensive set of chemical and physical descriptors known to influence aqueous solubility (LogS), as described in prior sections. Once trained, the optimized models were evaluated on a held-out test set comprising 878 molecular samples. Performance was measured using Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) to quantify the magnitude of prediction errors, along with Coefficient of Determination (R^2) and Spearman's rank correlation coefficient to assess how well the models captured the underlying ranking and variance in solubility. These metrics were chosen to provide a well-rounded evaluation of both the predictive accuracy and the consistency of the rankings produced by each model. This framework ensures that model efficacy is evaluated not only in terms of exact value prediction but also in terms of relative ordering of solubility across molecules. Table 1 below summarizes the metrics obtained.

Table 1: Summary of metrics obtained from individual models, KNN and the different ensembling techiniques

Model/Metric	R^2	Spearman	RMSE	MAE
MDM	0.8414	0.9128	0.8640	0.5836
GNN	0.8243	0.9019	0.9095	0.6338
SMILES	0.8157	0.8997	0.9314	0.6459
K-NN	0.8190	0.9004	0.9230	0.625
Simple Averaging (Ensemble)	0.8531	0.9195	0.8314	0.5611
Ensemble Using Optuna	0.8541	0.9199	0.8286	0.5578
Ensemble Using Optuna and Cross-Validation	0.8542	0.9197	0.8285	0.5565
Mixture of experts	0.8504	0.9181	0.8390	0.5667

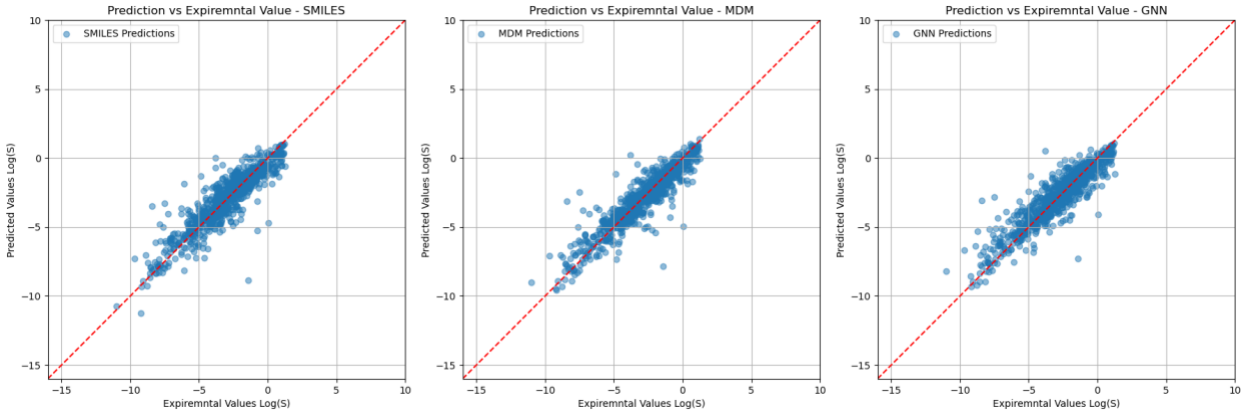

Figure 7: Predictions from individual models

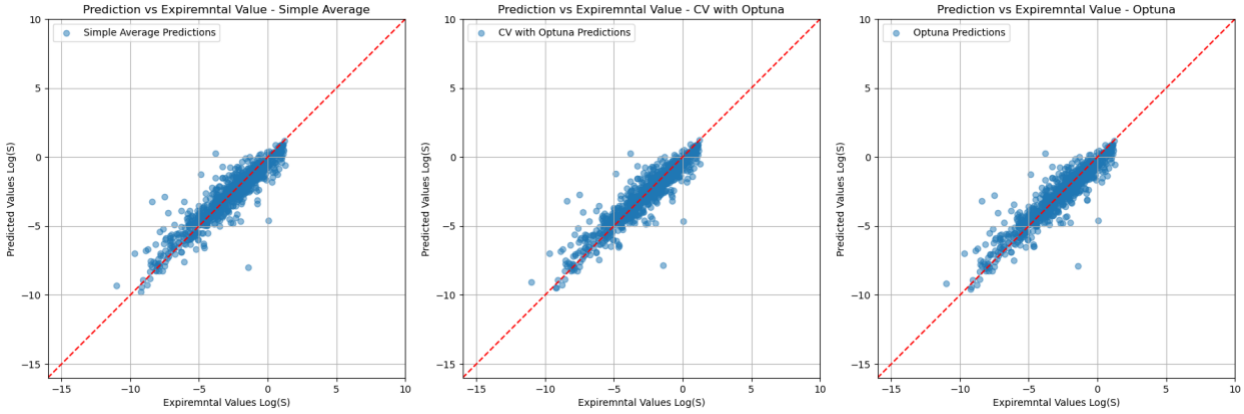

Figure 8: Predictions from Simple averaging ensemble, Weighted average ensemble using optuna , Weighted average using optuna with CV

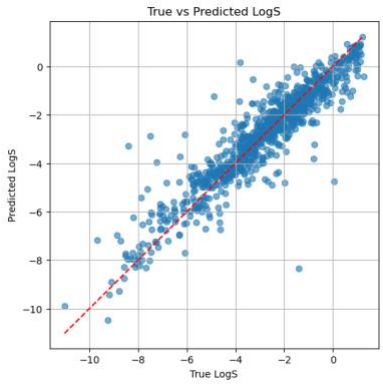

Figure 9: Predictions from Mixture of Experts

Based on the performance metrics presented in the table, the Molecular Descriptor Model (MDM) achieved the strongest overall results among the individual models, with the highest R^2 value of 0.8414, Spearman correlation of 0.9128, and the lowest RMSE (0.8640) and MAE (0.5836),

indicating both strong explanatory power and predictive accuracy. The Graph Neural Network (GNN) and SMILES-based LSTM models also performed reasonably well but lagged slightly behind MDM across all metrics. The K-Nearest Neighbors (KNN) model yielded competitive performance, surpassing SMILES in terms of R^2 and Spearman correlation but still falling short of the MDM model. Among ensemble approaches, simple averaging significantly improved performance across all metrics compared to individual models, with an R^2 of 0.8531, Spearman of 0.9195, RMSE of 0.8314, and MAE of 0.5611. Further optimization using Optuna boosted performance slightly, achieving an R^2 of 0.8541 and the lowest MAE of 0.5578 among all methods. Incorporating cross-validation into the Optuna-based ensemble yielded marginal additional gains, with a slightly higher R^2 (0.8542) and comparable error metrics. The Mixture of Experts (MoE) model also demonstrated strong performance, with an R^2 of 0.8504 and Spearman correlation of 0.9181, while maintaining competitive RMSE (0.8390) and MAE (0.5667) values. Although the MoE did not outperform the Optuna-based ensembles, its performance highlights the value of dynamic expert weighting and conditional computation in ensemble learning. Overall, ensemble methods especially those tuned with Optuna proved most effective in enhancing predictive performance and reducing errors.

5.1 DISCUSSION

The findings of this study demonstrate clear advancements in the predictive modeling of aqueous solubility by leveraging ensemble learning strategies, particularly in comparison to individual model performances reported in the foundational work by Panapitiya et al[14]. The Molecular Descriptor Model (MDM) emerged as the most effective among individual models, showing the highest coefficient of determination ($R^2 = 0.8414$) and Spearman correlation ($\rho = 0.9128$), as well as the lowest prediction errors (RMSE = 0.8640, MAE = 0.5836). These results validate the importance of well-engineered molecular descriptors in capturing solubility-relevant features and suggest that handcrafted features, when derived from domain-specific knowledge, remain highly effective despite the growing popularity of deep learning.

Nevertheless, ensemble approaches substantially outperformed individual models across all metrics, underscoring the strength of model aggregation in capturing diverse aspects of molecular behavior. Simple averaging, though computationally inexpensive, improved both accuracy and rank consistency ($R^2 = 0.8531$, $\rho = 0.9195$), suggesting that diverse modeling architectures offer complementary insights into solubility prediction. Optuna-optimized ensemble models yielded

even better performance, achieving the lowest MAE (0.5578) and highest R^2 (0.8542), thus confirming the value of hyperparameter tuning and weight optimization in ensemble design. These results highlight the utility of treating ensemble weighting as a search problem, where performance-based tuning yields tangible improvements in predictive precision.

Importantly, the Mixture of Experts (MoE) model demonstrated strong, competitive performance ($R^2 = 0.8504$, $\rho = 0.9181$), nearly matching the optimized ensembles. While it did not surpass the Optuna-based models in raw performance, its dynamic weighting mechanism reflects a promising direction for future ensemble learning frameworks. The ability of MoE to assign context-specific weights to expert models offers a more flexible and interpretable alternative to static ensembling, particularly in settings where model reliability varies with input characteristics.

From a computational standpoint, simple averaging ensembles offer the most efficient approach, making them ideal for large-scale screening tasks. Optuna-optimized ensembles, although highly accurate, incur a high computational cost due to repeated evaluations during hyperparameter tuning. The MoE model strikes a balance, offering dynamic adaptability with moderate computational overhead during inference, especially when the number of active experts is limited. This makes MoE a practical and scalable solution for intelligent, high-throughput chemical screening.

Overall, this study's multi-stage methodology from KNN-based model selection to static and dynamic ensembling presents a robust framework that balances predictive accuracy, adaptability, and scalability. By extending the deep learning architectures explored in the original study, this work advances the state-of-the-art in solubility prediction and contributes practical, high-accuracy tools for real-world chemical and pharmaceutical applications.

CHAPTER 6

6.0 CONCLUSION

This study presents a comprehensive approach to improving aqueous solubility prediction of organic molecules through the integration of ensemble learning techniques and model interpretability strategies. Building upon the foundational work by Panapitiya et al.[14], we evaluated and extended multiple machine learning models using a robust dataset derived from the SOMAS database, which includes experimentally measured solubilities along with quantum and molecular descriptors. Among individual models, the Molecular Descriptor Model (MDM) showed the highest standalone performance, demonstrating the continued relevance of carefully engineered descriptors in solubility prediction.

However, ensemble methods consistently outperformed individual models in both predictive accuracy and robustness. Simple averaging offered significant gains with minimal computational cost, while Optuna-based ensembles achieved the best overall metrics through performance-driven weight optimization. The Mixture of Experts (MoE) model also proved to be a strong contender, leveraging instance-specific weighting to dynamically adapt to input characteristics making it a scalable and efficient option for high-throughput screening scenarios.

By combining predictive power with computational feasibility, this research provides a flexible and extensible framework for solubility prediction that is well-suited to both academic exploration and real-world application in fields such as drug discovery, catalysis, and energy storage. Ultimately, the methodologies developed here not only advance the state of the art but also lay the groundwork for data-driven molecular design in solubility-critical domains.

6.2 FUTURE WORK

Future efforts will focus on expanding the solubility dataset to include a broader range of organic molecules, such as charged species, organometallic compounds, and temperature-dependent measurements, to enhance the model's applicability across diverse chemical domains. Additionally, incorporating a wider variety of predictive models, including graph transformers, attention-based networks, and hybrid physics-informed architectures, may further improve performance and generalization. To enhance the flexibility and efficiency of the ensemble framework, future work will also explore advanced gating mechanisms, particularly by leveraging the Top-K expert selection strategy proposed by Shazeer et al.[36], which activates only the most relevant expert models per sample. This approach can significantly reduce computational cost during inference while maintaining high predictive accuracy, enabling more scalable deployment of Mixture of Experts models in large-scale molecular screening and design pipelines.

REFERENCES

[1] K. T. Savjani, A. K. Gajjar, and J. K. Savjani, "Drug solubility: importance and enhancement techniques," *ISRN Pharm.*, vol. 2012, p. 195727, 2012, doi: 10.5402/2012/195727.

[2] A. Kurotani, T. Kakiuchi, and J. Kikuchi, "Solubility Prediction from Molecular Properties and Analytical Data Using an In-phase Deep Neural Network (Ip-DNN)," *ACS Omega*, vol. 6, no. 22, pp. 14278–14287, Jun. 2021, doi: 10.1021/acsomega.1c01035.

[3] S. Lee, M. Lee, K.-W. Gyak, S. D. Kim, M.-J. Kim, and K. Min, "Novel Solubility Prediction Models: Molecular Fingerprints and Physicochemical Features vs Graph Convolutional Neural Networks," *ACS Omega*, vol. 7, no. 14, pp. 12268–12277, Apr. 2022, doi: 10.1021/acsomega.2c00697.

[4] D. Varrica and M. G. Alaimo, "Determination of Water-Soluble Trace Elements in the PM10 and PM2.5 of Palermo Town (Italy)," *Int. J. Environ. Res. Public. Health*, vol. 20, no. 1, p. 724, Dec. 2022, doi: 10.3390/ijerph20010724.

[5] N. Yousefi and S. Abbasi, "Food proteins: Solubility & thermal stability improvement techniques," *Food Chem. Adv.*, vol. 1, p. 100090, Oct. 2022, doi: 10.1016/j.focha.2022.100090.

[6] Y. Liang *et al.*, *High-throughput solubility determination for data-driven materials design and discovery in redox flow battery research*. 2023. doi: 10.26434/chemrxiv-2023-985h7.

[7] "Computational methodology for solubility prediction: Application to the sparingly soluble solutes | The Journal of Chemical Physics | AIP Publishing." Accessed: May 17, 2024. [Online]. Available: https://pubs.aip.org/aip/jcp/article/146/21/214110/973269/Computational-methodology-for-solubility

[8] P. G. Francoeur and D. R. Koes, "SolTranNet-A Machine Learning Tool for Fast Aqueous Solubility Prediction," *J. Chem. Inf. Model.*, vol. 61, no. 6, pp. 2530–2536, Jun. 2021, doi: 10.1021/acs.jcim.1c00331.

[9] "SolPredictor: Predicting Solubility with Residual Gated Graph Neural Network - PMC." Accessed: May 17, 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10815788/

[10] J. Yu *et al.*, “SolvBERT for solvation free energy and solubility prediction: a demonstration of an NLP model for predicting the properties of molecular complexes,” Jul. 27, 2022, *ChemRxiv*. doi: 10.26434/chemrxiv-2022-0hl5p.

[11] “Delfos: deep learning model for prediction of solvation free energies in generic organic solvents - Chemical Science (RSC Publishing).” Accessed: May 17, 2024. [Online]. Available: https://pubs.rsc.org/en/content/articlelanding/2019/sc/c9sc02452b

[12] 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, Jan. 2023, doi: 10.1021/acsomega.2c06702.

[13] H. Lim and Y. Jung, “MLSolvA: solvation free energy prediction from pairwise atomistic interactions by machine learning,” *J. Cheminformatics*, vol. 13, no. 1, p. 56, Jul. 2021, doi: 10.1186/s13321-021-00533-z.

[14] “Evaluation of Deep Learning Architectures for Aqueous Solubility Prediction | ACS Omega.” Accessed: May 06, 2025. [Online]. Available: https://pubs.acs.org/doi/full/10.1021/acsomega.2c00642

[15] P. Gao *et al.*, “SOMAS: a platform for data-driven material discovery in redox flow battery development,” *Sci. Data*, vol. 9, no. 1, p. 740, Dec. 2022, doi: 10.1038/s41597-022-01814-4.

[16] J. Sun, W. Du, and N. Shi, “A Survey of kNN Algorithm,” *Inf. Eng. Appl. Comput.*, vol. 1, May 2018, doi: 10.18063/ieac.v1i1.770.

[17] “What is the k-nearest neighbors algorithm? | IBM.” Accessed: May 07, 2025. [Online]. Available: https://www.ibm.com/think/topics/knn

[18] K. Taunk, S. De, S. Verma, and A. Swetapadma, “A Brief Review of Nearest Neighbor Algorithm for Learning and Classification,” May 2019, pp. 1255–1260. doi: 10.1109/ICCS45141.2019.9065747.

[19] J. Brownlee, “Why Use Ensemble Learning?,” MachineLearningMastery.com. Accessed: May 17, 2025. [Online]. Available: https://www.machinelearningmastery.com/why-use-ensemble-learning/

[20] D. Ganta, H. Das Gupta, and V. Sheng, *Knowledge Distillation via Weighted Ensemble of Teaching Assistants*. 2022. doi: 10.48550/arXiv.2206.12005.

[21] Y. Lim, “Ensemble Averaging - Improve machine learning performance by voting,” Towards Data Science. Accessed: May 17, 2025. [Online]. Available: https://towardsdatascience.com/ensemble-averaging-improve-machine-learning-performance-by-voting-246106c753ee/

[22] S. Bhatnagar, “Ensemble Methods in Machine Learning,” Medium. Accessed: May 17, 2025. [Online]. Available: https://medium.com/@shashank25.it/ensemble-methods-in-machine-learning-2d4cc7513c77

[23] M. Mahoney, “Model averaging methods: how and why to build ensemble models,” Mike Mahoney. Accessed: May 17, 2025. [Online]. Available: https://www.mm218.dev/posts/2021/01/model-averaging/

[24] D. Arpit, H. Wang, Y. Zhou, and C. Xiong, “Ensemble of Averages: Improving Model Selection and Boosting Performance in Domain Generalization,” Oct. 10, 2022, *arXiv*: arXiv:2110.10832. doi: 10.48550/arXiv.2110.10832.

[25] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, “Optuna: A Next-generation Hyperparameter Optimization Framework,” in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, in KDD ’19. New York, NY, USA: Association for Computing Machinery, Jul. 2019, pp. 2623–2631. doi: 10.1145/3292500.3330701.

[26] J. Brownlee, “A Gentle Introduction to k-fold Cross-Validation,” MachineLearningMastery.com. Accessed: May 07, 2025. [Online]. Available: https://www.machinelearningmastery.com/k-fold-cross-validation/

[27] J. M. Gorriz, R. M. Clemente, F. Segovia, J. Ramirez, A. Ortiz, and J. Suckling, “Is K-fold cross validation the best model selection method for Machine Learning?,” Nov. 08, 2024, *arXiv*: arXiv:2401.16407. doi: 10.48550/arXiv.2401.16407.

[28] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, “Adaptive Mixtures of Local Experts,” *Neural Comput.*, vol. 3, no. 1, pp. 79–87, Mar. 1991, doi: 10.1162/neco.1991.3.1.79.

[29] “(PDF) Improving Expert Specialization in Mixture of Experts.” Accessed: May 17, 2025. [Online].Available:https://www.researchgate.net/publication/368877500_Improving_Expert_Specialization_in_Mixture_of_Experts

[30] “Mixture of Experts Explained.” Accessed: May 07, 2025. [Online]. Available: https://huggingface.co/blog/moe

[31] Y. Zhou *et al.*, “Mixture-of-Experts with Expert Choice Routing,” Oct. 14, 2022, *arXiv*: arXiv:2202.09368. doi: 10.48550/arXiv.2202.09368.

[32] “Exploring DeepSeek-R1’s Mixture-of-Experts Model Architecture - AI Resources.” Accessed: May 07, 2025. [Online]. Available: https://www.modular.com/ai-resources/exploring-deepseek-r1-s-mixture-of-experts-model-architecture

[33] D. Eigen, M. Ranzato, and I. Sutskever, “Learning Factored Representations in a Deep Mixture of Experts,” Mar. 09, 2014, *arXiv*: arXiv:1312.4314. doi: 10.48550/arXiv.1312.4314.

[34] “What is Embedding? - Embeddings in Machine Learning Explained - AWS,” Amazon Web Services, Inc. Accessed: May 17, 2025. [Online]. Available: https://aws.amazon.com/what-is/embeddings-in-machine-learning/

[35] “What is Embedding? | IBM.” Accessed: May 17, 2025. [Online]. Available: https://www.ibm.com/think/topics/embedding

[36] N. Shazeer *et al.*, “Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer,” Jan. 23, 2017, *arXiv*: arXiv:1701.06538. doi: 10.48550/arXiv.1701.06538.