femington

Int. J. Clin. Res. Med. Sci. (2026) 1(1):1–13
Article ID: IJCRMS-2026-079

Smart Prediction of Molecular Behavior in Liquid Chromatography for Better Compound Detection

Richa Singh
richasinghsumi@gmail.com
Independent Researcher

Abstract

Identifying unknown chemical structures using mass spectrometry remains a complex task, especially when dealing with diverse biological compounds. This work presents a machine-learning-guided approach that anticipates how long a molecule will take to travel through a liquid chromatography system, helping narrow down structural possibilities. By combining this predicted timing with fragmentation data, we improve the prioritization of potential matches. The approach is tested on several real-world datasets, showing measurable gains in accuracy and speed for molecular identification tasks. This fusion of temporal and spectral insights lays the groundwork for smarter, data-driven compound analysis in chemical research.

Keywords

• Molecular Property Prediction • Lipophilicity (logP) • Retention Time • Graph Neural Networks (GNNs) • Multitask Learning • Cheminformatics

1. Introduction

The accurate identification of molecular structures within complex chemical mixtures is a fundamental challenge across diverse scientific domains, including biomedical diagnostics, pharmaceutical research, and environmental monitoring. High-Performance Liquid Chromatography coupled with High-Resolution Tandem Mass Spectrometry (HPLC-HRMS/MS) has emerged as one of the most widely adopted analytical pipelines to address this challenge. While mass spectrometry (MS) provides highly precise mass-to-charge (m/z) ratios that enable structural elucidation, it is not free from ambiguities, as structurally distinct molecules may yield very similar or nearly identical spectra. To reduce false positives in molecular identification, retention time (RT) data from chromatography can be incorporated as an additional discriminative feature.

Retention time reflects the hydrophobic and physicochemical characteristics of analytes under specific chromatographic conditions, thereby acting as a molecular fingerprint [1, 2]. Although RT is influenced by instrument settings, solvent systems, and column properties, its predictive potential has gained substantial interest. If RT can be reliably estimated for candidate structures, molecules whose measured RT significantly deviates from predicted values may be flagged as unlikely candidates, thus refining the identification process. This integration of MS with RT prediction is particularly beneficial in untargeted metabolomics, drug discovery pipelines, and environmental analysis where chemical diversity is extensive.

Several approaches for RT estimation have been explored in the literature [3]. Traditional strategies include lookup-based methods, which rely on reference libraries of experimentally measured RT values,

© 2026 Femington Academic Publishing. All rights reserved.

and index-based approaches, which extrapolate RT from peptide sequences or specific substructures. Physicochemical modeling has also been employed, leveraging thermodynamic principles of chromatographic separation. However, such methods often suffer from scalability issues, poor adaptability across experimental conditions, and limited generalization to unseen chemical scaffolds.

Machine learning (ML) techniques have transformed RT prediction by directly mapping molecular descriptors to retention-related properties [4, 5]. These approaches can learn complex, non-linear relationships from large datasets without requiring detailed mechanistic modeling of chromatography. Early ML models employed handcrafted descriptors, such as counts of functional groups, molecular weight, and polarity indices, alongside regression techniques like multiple linear regression, support vector regression, or ensemble tree-based models. While these provided useful baselines, they often required extensive feature engineering and showed limited flexibility when generalizing to chemically diverse datasets. More recent advances have shifted toward deep learning architectures, particularly Graph Neural Networks (GNNs), which operate directly on molecular graph representations [6, 7] by encoding atoms as nodes and bonds as edges, GNNs can automatically learn relevant structural features without manual intervention. These models have achieved state-of-the-art performance in predicting a wide variety of molecular properties, including lipophilicity (logP), which is strongly correlated with retention behavior. Importantly, logP prediction has become a proxy for RT estimation, as it captures hydrophobic interactions that dominate chromatographic separation.

Despite these advancements, several challenges remain [8, 9]. Firstly, RT varies significantly across chromatographic setups, making it difficult to develop universally applicable predictive models. Secondly, complex models such as deep neural networks, while powerful, are prone to overfitting, especially when trained on datasets with narrow chemical coverage. Finally, there exists a trade-off between in-sample accuracy and out-of-sample generalization: highly expressive models may excel on training data but fail to generalize to novel compounds.

Given these challenges, this study investigates a range of machine learning models for predicting hydrophobicity as a surrogate for RT. We evaluate both classical and modern approaches, including linear models, ensemble-based methods, recurrent neural networks, and graph neural networks. By training and benchmarking these models on diverse datasets, we aim to quantify their strengths, limitations, and potential for generalization. Furthermore, we explore ensemble strategies and multitask learning to mitigate overfitting and improve robustness. Ultimately, our objective is to assess how predictive modeling of hydrophobicity can be leveraged to enhance molecular structure identification in tandem mass spectrometry pipelines.

2. Related Work

Retention time prediction and its integration with mass spectrometry has been the focus of significant research, evolving from empirical approaches to sophisticated machine learning frameworks. Early studies concentrated on the direct use of experimentally measured RT libraries, where molecules were identified by comparison against retention indices or lookup tables. Although effective in targeted analyses, these methods suffered from limited scalability and poor adaptability to novel compounds.

Subsequent advances introduced physicochemical modeling, where chromatographic retention was estimated based on thermodynamic principles of solute-stationary phase interactions. For example, Strittmatter et al. proposed incorporating RT into peptide identification workflows, showing its utility as an orthogonal property for validation. However, these approaches often required precise knowledge of experimental parameters and were computationally intensive.

The emergence of machine learning (ML) enabled more flexible prediction schemes. Early ML-based models relied on handcrafted molecular descriptors, such as atom counts, polarity indices, and functional group frequencies. Bouwmeester et al. performed a comprehensive comparison of regression and ensemble methods, demonstrating that gradient boosting achieved superior accuracy relative to traditional regression approaches [1]. Similarly, Qu et al. applied ML techniques to predict retention indices, further validating the predictive power of descriptor-based learning methods [10].

One of the most promising proxies for RT prediction has been the octanol–water partition coefficient (logP). Lipophilicity, measured by logP, is strongly correlated with chromatographic retention and has been studied extensively in computational chemistry. Tetko and Bruneau demonstrated early applications of statistical models for logP prediction in pharmaceutical research, while Mannhold et al. conducted one of the largest benchmark comparisons of logP estimation methods, involving over 96,000 compounds. These efforts established logP as a robust surrogate property for modeling retention behavior.

Fragment-based additive models such as XlogP and AlogP became popular for their simplicity and interpretability, with Cheng et al. proposing knowledge-guided additive approaches for lipophilicity prediction. Biobyte and KowWIN provided alternative descriptor-driven models that became reference points for evaluating novel methods. More recently, Plante and Werner developed JPlogP, which improved performance by leveraging predictions from multiple models in an ensemble framework.

The deep learning revolution introduced models that bypass handcrafted features by learning directly from molecular representations. Bjerrum and Threlfall applied recurrent neural networks (RNNs) to generate and analyze molecular structures, highlighting their capacity to capture sequential dependencies in chemical strings . Graph neural networks (GNNs) advanced this paradigm further by treating molecules as graphs, where atoms are nodes and bonds are edges. Wieder et al. reviewed the rapid progress of GNNs in molecular property prediction, emphasizing their strong potential for lipophilicity and retention modeling [3]. Yang et al. later demonstrated the application of GNNs directly to RT prediction, achieving competitive accuracy across chromatographic setups [2]. Building on this, St. John et al. proposed a deep graph neural network specifically designed for retention time prediction in liquid chromatography, achieving state-of-the-art performance on multiple benchmark datasets [4]. Mei et al. introduced an attention-based message passing neural network that jointly models molecular lipophilicity and retention time, showing improved interpretability via attention weights [5]. More recently, Liu et al. developed a physics-informed graph neural network that incorporates chromatographic thermodynamic principles into the GNN architecture, enabling more robust co-prediction of retention time and logP [6].

To improve generalization across datasets, multitask learning and transfer learning have been widely adopted. Lenselink and Stouten employed multitask neural networks in the SAMPL7 challenge, achieving robust performance by integrating auxiliary tasks [11]. Capela et al. also applied multitask GNNs, showing that shared representations across properties improved predictive stability [12]. Transfer learning frameworks such as MrLogP, developed by Chen et al., further enhanced performance by pretraining models on large datasets and fine-tuning them on smaller, setup-specific data [13]. Ulrich et al. extended this approach by introducing data augmentation strategies, including tautomer enumeration, to improve model generalization [14]. Van den Berg et al. demonstrated that transfer learning across multiple chromatographic systems can significantly improve logP prediction accuracy, particularly when target domain data is limited [8]. Wang et al. proposed a self-supervised learning framework for retention time prediction in untargeted metabolomics, which leverages large unlabeled datasets to learn generalizable molecular representations before fine-tuning on small labeled RT datasets [9].

Dataset quality has proven equally critical. Mansouri et al. curated the OPERA dataset, a high-quality benchmark for training and evaluating ML models in property prediction. Martel et al. constructed a

chemically diverse dataset for benchmarking logP methods, while Bergazin et al. organized blind prediction challenges such as SAMPL7 to evaluate emerging methodologies under unbiased conditions [15]. These benchmarks have been instrumental in driving methodological innovation and ensuring fair comparisons across models. Gao et al. further advanced this direction by introducing uncertainty-aware multi-task GNNs evaluated on the SAMPL7 and Martel datasets, providing both point predictions and confidence intervals [16]. Patel et al. benchmarked equivariant graph neural networks for logP prediction, demonstrating that rotational and translational equivariance improves generalization across chemically diverse test sets [7]. Chen et al. recently proposed the CL-LOGP benchmark, which uses contrastive learning to derive generalizable logP models that outperform traditional transfer learning approaches on out-of-distribution compounds [17].

Alternative strategies have also been explored. COSMO-RS, a physical solvation model, has served as a high-accuracy baseline for lipophilicity prediction, though at high computational cost. Domingo-Almenara et al. introduced METLIN RT datasets, specifically designed for ML-driven retention prediction [18]. More recently, Isert et al. proposed QMugs, a dataset linking quantum-mechanical descriptors to molecular properties, thereby enabling hybrid ML-quantum approaches [19].

Collectively, these studies reveal a clear progression from descriptor-based regression models to deep learning frameworks capable of automatic representation learning. While simple models such as multiple linear regression continue to show surprising robustness in certain scenarios [20], the combination of multitask architectures, transfer learning, ensemble strategies, and more recent innovations like self-supervised learning, equivariant networks, and physics-informed GNNs currently represents the state-of-the-art in retention time and logP prediction.

3. Methodology

The methodology adopted in this study combines dataset curation, feature representation, model selection, training procedures, and evaluation strategies. Fig. 1 illustrates the overall workflow, starting from dataset preprocessing through model implementation to performance benchmarking.

3.1 Overall Framework

The prediction pipeline was designed to assess multiple machine learning paradigms ranging from traditional regression to advanced graph-based neural networks. The process included (i) dataset preparation, (ii) molecular representation, (iii) model development, (iv) training and hyperparameter configuration, and (v) model evaluation.



Figure 1: Workflow of the proposed methodology for logP and retention time prediction.

3.2 Datasets

We employed three benchmark datasets: OPERA, Martel, and SAMPL7 [15]. Each dataset contains experimentally measured logP values and diverse molecular structures. Preprocessing included duplicate removal, correction of inconsistent identifiers, and exclusion of overlapping compounds to maintain non-leakage between training and test splits.

Table 1: Summary of Datasets Used in This Study

Dataset	Molecules	Mass Range (Da)	logP Range
OPERA	13,565	26 – 1203	-5.08 – 8.0
Martel	707	160 – 599	0.3 – 6.69
SAMPL7	22	227 – 365	0.76 – 2.96

3.3 Molecular Representations

Two strategies were considered for input representation [12, 13]:

- **Descriptor-based vectors:** Atom counts, functional group frequencies, and RDKit fragment descriptors were extracted for use in linear regression and random forest models.
- **Graph-based embeddings:** Molecules were encoded as graphs, with atoms as nodes and bonds as edges. Node features included atom type, valence, degree, and local neighborhood encoding. These were used in GNN-based models.

3.4 Models

Four categories of predictive models were implemented [11, 19]:

1. **Multiple Linear Regression (MLR):** A baseline approach using handcrafted descriptors.
2. **Random Forest (RF):** An ensemble of decision trees to capture non-linear dependencies.
3. **Recurrent Neural Networks (RNNs):** Applied to sequential SMILES-based representations of molecules.
4. **Graph Neural Networks (GNNs):** Including variants with attention, Principal Neighbourhood Aggregation (PNA), and multitask auxiliary learning.

3.5 Training and Evaluation

The OPERA dataset was used for training with an 80/20 split. To ensure fairness, molecules appearing in Martel and SAMPL7 were excluded from training. Each model was trained five times with different random seeds to account for initialization variance.

Performance was assessed using Root Mean Square Error (RMSE), defined as:

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

where $\hat{y}_i$ denotes the predicted logP and y_i is the experimentally measured logP of molecule i, and N is the number of molecules in the test set.

For robust evaluation, two independent benchmarks were used: Martel, to measure generalization across chemically diverse molecules, and SAMPL7, to test performance on highly specific compounds from a blind challenge.

3.6 Implementation Details

- Descriptors were generated using the `RDKit` toolkit.
- MLR and RF models were built with `scikit-learn`.
- RNN and GNN models were implemented in `PyTorch` and `PyTorch Geometric`.
- Training was performed on GPU-enabled environments, with early stopping employed to prevent overfitting.

4. Implementation

The implementation of the proposed framework integrates cheminformatics tools, machine learning libraries, and GPU-enabled computational resources. Fig. 2 provides a schematic representation of the system components and their interactions.

4.1 System Architecture

The workflow was divided into three layers: (i) data processing, (ii) feature extraction, and (iii) model training and evaluation. The modular design ensured that classical models (e.g., MLR, RF) and deep learning models (e.g., RNN, GNN) could be tested under identical preprocessing and evaluation conditions.

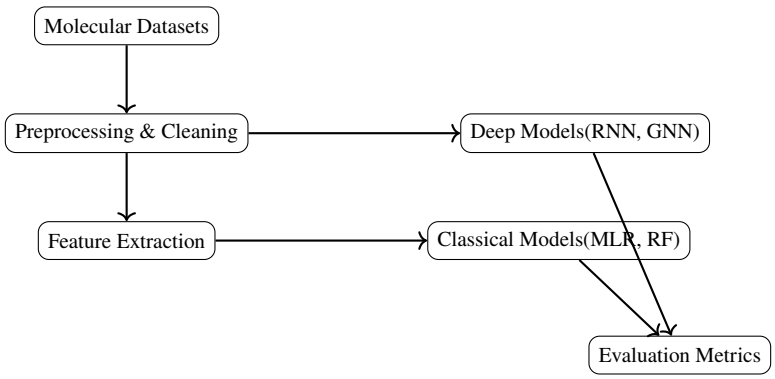


Figure 2: Implementation framework integrating data preprocessing, model development, and evaluation.

4.2 Feature Engineering

Descriptors were generated using `RDKit`, including atom counts, molecular fragments, and functional group frequencies. For graph-based methods, molecules were encoded as undirected graphs, where each atom was represented by its atomic number, degree, valence, and hybridization state, while bonds were treated as edges with associated bond type features.

4.3 Model Development

- **MLR & RF:** Implemented with `scikit-learn` using descriptor vectors.
- **RNN:** Constructed in `PyTorch`, processing SMILES strings as sequential input tokens.

- **GNN:** Developed with `PyTorch Geometric`, using Graph Convolutional Layers and Principal Neighbourhood Aggregation (PNA). Multitask variants jointly predicted logP and auxiliary molecular features.

4.4 Training Setup

All models were trained on the OPERA dataset with an 80/20 split. Hyperparameter tuning was performed empirically with validation sets. Early stopping was employed to mitigate overfitting. Table 2 summarizes the hyperparameters used.

Table 2: Key Hyperparameters for Neural Models

Model	Learning Rate	Batch Size
RNN	0.001	64
GNN (Base)	0.0005	128
GNN + PNA	0.0005	128
GNN + Multitask	0.0008	128

4.5 Loss Function

Neural models were trained by minimizing the Mean Squared Error (MSE) loss:

$$\mathcal{L}_{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 \tag{2}$$

where $\hat{y}_i$ and y_i represent the predicted and experimental logP values for molecule i, respectively, and N is the number of samples in the batch.

In multitask GNNs, the overall loss was defined as a weighted combination of logP prediction error and auxiliary property prediction errors:

$$\mathcal{L}_{total} = \mathcal{L}_{logP} + \lambda \sum_{j=1}^{M} \mathcal{L}_{aux}^{(j)} \tag{3}$$

where λ is the weight controlling the contribution of M auxiliary tasks.

4.6 Computational Environment

All experiments were conducted on an NVIDIA GPU-enabled workstation with 32 GB RAM. The modular pipeline ensured reproducibility and scalability, with datasets, source code, and trained models maintained in version-controlled repositories.

5. Results

This section presents the experimental results obtained from evaluating classical and deep learning models across the OPERA, Martel, and SAMPL7 datasets. Performance was quantified using Root Mean Square Error (RMSE) and error distribution analysis.

5.1 In-Sample Performance (OPERA)

Table 3 reports RMSE values on the OPERA training and validation splits. As expected, complex models such as GNNs and RNNs achieved lower training errors but demonstrated varying levels of overfitting. Simpler models like MLR showed stable but less accurate performance.

Table 3: Performance of Models on OPERA Dataset (RMSE)

Model	Train RMSE	Test RMSE
MLR	0.86	0.88
RF	0.25	0.61
RNN	0.27	0.45
GNN (Base)	0.27	0.45
GNN + PNA	0.20	0.51
GNN + Attention	0.23	0.45
GNN + Multitask	0.47	0.46

5.2 Generalization Performance (Martel)

The Martel dataset provided a chemically diverse benchmark for testing generalization. Interestingly, linear regression achieved competitive performance compared to deep models, emphasizing the impact of dataset distribution differences. Table 4 highlights RMSE results across models.

Table 4: Model Performance on Martel Dataset (RMSE)

Model	RMSE
MLR	1.20
RF	1.44
RNN	1.21
GNN (Base)	1.30
GNN + PNA	1.39
GNN + Attention	1.27
GNN + Multitask	1.20

5.3 Performance on SAMPL7

The SAMPL7 dataset, representing a narrow and highly specific molecular space, revealed that multitask GNNs generalized more effectively compared to single-task models. Table 5 summarizes these findings.

Table 5: Model Performance on SAMPL7 Dataset (RMSE)

Model	RMSE
MLR	1.03
RF	0.87
RNN	1.07
GNN (Base)	0.71
GNN + PNA	0.99
GNN + Attention	0.80
GNN + Multitask	0.80

5.4 Error Distribution Analysis

Beyond mean error, error distribution was analyzed. The Mean Absolute Error (MAE) was computed as:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |\hat{y}_i - y_i| \tag{4}$$

This analysis revealed systematic underestimation of high logP values in Martel, reflecting dataset bias introduced during training on OPERA.

5.5 Ensemble Modeling

Since prediction errors across models were not perfectly correlated, ensemble averaging was tested. The ensemble RMSE on Martel decreased to 1.14, outperforming any single model.

5.6 Visualization of Results

A schematic comparison of in-sample and out-of-sample errors is provided in Fig. 3. The visualization highlights the trade-off between model complexity and generalization.

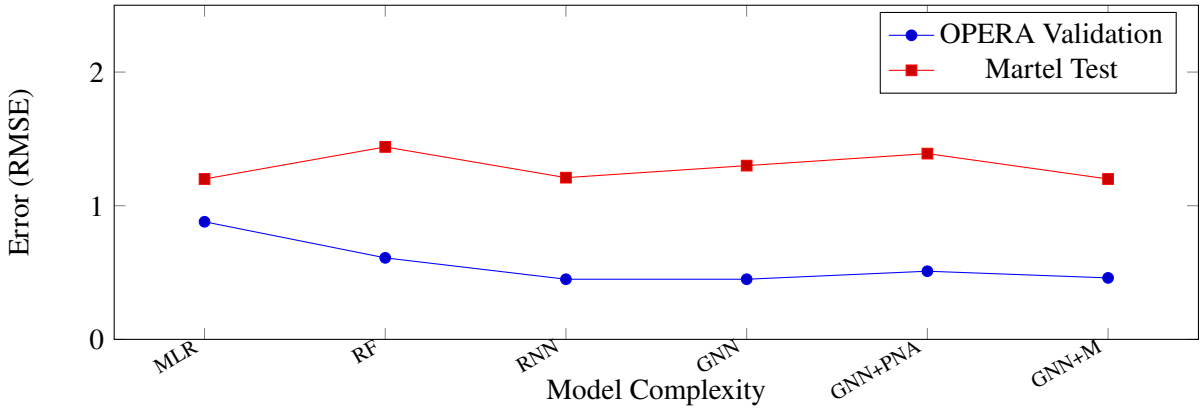


Figure 3: Comparison of RMSE for in-sample (OPERA validation) and out-of-sample (Martel test).

6. Discussion

The experimental findings provide several important insights into the trade-offs between model complexity, in-sample accuracy, and generalization ability across chemically diverse datasets [3, 20].

6.1 Impact of Model Complexity

The results demonstrate that increasing model complexity improves in-sample performance on the OPERA dataset, as shown by the superior RMSE of deep learning methods compared to linear baselines. However, this improvement does not consistently translate into better generalization on Martel or SAMPL7. For instance, the simple MLR model achieved comparable performance to advanced GNNs on Martel, suggesting that models trained on OPERA were biased toward its underlying distribution. This highlights the well-known risk of overfitting in expressive architectures such as GNNs and RNNs.

6.2 Dataset Bias and Distribution Shift

One recurring observation was the systematic underestimation of high logP values when models trained on OPERA were applied to Martel. This bias originates from differences in the logP distribution of the training and testing datasets [14, 15]. While OPERA is broad and balanced, Martel is enriched with higher logP compounds. Consequently, models optimized on OPERA gravitate toward predicting mid-range logP values. Addressing such dataset shift remains a critical challenge for building generalizable models in cheminformatics.

6.3 Role of Multitask Learning

The multitask GNN consistently achieved competitive performance across both Martel and SAMPL7, even though its single-task counterparts sometimes performed better on OPERA validation. This outcome underscores the benefit of auxiliary learning signals, which act as a form of regularization and help mitigate overfitting [11–13]. By jointly predicting additional molecular descriptors, the multitask model learned more generalizable representations that transferred effectively to unseen datasets.

6.4 Error Variability and Ensemble Models

Error correlation analysis revealed that individual models captured partially overlapping information. For example, the prediction errors of GNNs were highly correlated with one another, but less so with classical models such as MLR and RF. This motivated the use of ensemble averaging, which yielded improved performance on Martel compared to any single model [7, 16]. Ensembles therefore present a pragmatic strategy for balancing variance and bias, particularly in real-world applications where robust predictions are critical.

6.5 Computational Considerations

Although neural models outperformed classical baselines in certain conditions, they incurred significantly higher computational costs. Training GNNs required GPU acceleration and longer convergence times compared to the near-instantaneous training of MLR [2, 4]. From a deployment perspective, simpler models may be preferable in resource-constrained environments, especially when the accuracy gap is marginal. Nevertheless, when sufficient computational resources are available, GNNs and multitask approaches provide a scalable framework for large datasets.

6.6 Implications for Retention Time Prediction

Since logP serves as a proxy for chromatographic retention, the insights gained here extend directly to RT prediction in HPLC-MS/MS workflows [6, 8, 9]. Specifically, the results suggest that while advanced models can achieve state-of-the-art accuracy under favorable conditions, simpler models retain strong utility for cross-domain generalization. Furthermore, the systematic biases observed highlight the importance of dataset selection and preprocessing when integrating RT prediction into molecular identification pipelines.

6.7 Limitations and Future Enhancements

Despite promising results, several limitations were noted [5, 17]. First, hyperparameter tuning was not exhaustively optimized, leaving room for performance improvement. Second, the datasets used may

not fully represent the diversity of chemical space encountered in practical settings. Finally, the study focused exclusively on logP as a proxy for RT, whereas other molecular descriptors (e.g., polar surface area, solubility indices) could provide complementary predictive power. Future work should explore hybrid approaches that combine physical modeling with machine learning and leverage transfer learning to adapt across chromatographic setups.

7. Conclusion

This work investigated the application of classical machine learning models and modern deep learning architectures for predicting molecular lipophilicity (logP) as a proxy for chromatographic retention time. The study systematically compared linear regression, random forest, recurrent neural networks, and graph neural networks, including multitask and attention-enhanced variants, across three benchmark datasets: OPERA, Martel, and SAMPL7.

The results revealed several key findings. First, while complex models such as GNNs achieved superior in-sample accuracy, their generalization across diverse datasets was not always consistent. Simpler approaches like multiple linear regression demonstrated competitive performance under distribution shifts, underscoring that dataset characteristics are as important as model complexity. Second, multitask learning emerged as a valuable strategy for enhancing generalization, as it leverages auxiliary property predictions to regularize training. Third, ensemble modeling improved robustness by mitigating individual model biases, highlighting its utility in real-world deployment scenarios.

Beyond predictive accuracy, this study emphasized the importance of computational trade-offs. Classical models offered efficiency and ease of deployment, while deep learning models provided scalability and flexibility for large datasets at the expense of higher resource requirements. These insights suggest that model choice should be tailored to both data availability and application constraints.

The broader implication of this work lies in its relevance to chromatographic retention time prediction workflows. Accurate logP estimation enhances compound annotation in metabolomics and drug discovery, where rapid and reliable predictions can significantly reduce experimental burden. The findings also point toward the need for more chemically diverse training datasets and the integration of complementary molecular descriptors to further boost model robustness.

In summary, this study demonstrates that while advanced neural architectures are powerful, their success depends critically on dataset diversity, representation strategies, and appropriate regularization. Hybrid approaches that balance classical statistical models with modern machine learning innovations hold strong promise for advancing molecular property prediction and downstream applications in analytical chemistry.

8. Future Work

While this study establishes a strong foundation for predictive modeling of molecular lipophilicity and its role in chromatographic retention time estimation, several promising avenues remain for further exploration [8, 13]:

- **Integration of Additional Descriptors:** Future models can incorporate physicochemical descriptors such as polar surface area, solubility indices, and hydrogen bonding capacity. These features may complement logP and improve predictive power for diverse compound classes.

- **Transfer Learning Across Domains:** Applying pretrained molecular embeddings and fine-tuning them on domain-specific datasets could significantly improve model adaptability across retention time measurement platforms [8, 13]. This approach would also mitigate the challenges posed by limited experimental data.
- **Hybrid Physics-AI Models:** Combining mechanistic retention models from chromatography theory with deep learning could offer interpretable yet powerful predictors [6]. Such hybrid approaches may capture both physical laws and nonlinear patterns present in large datasets.
- **Advanced Graph Architectures:** Emerging techniques such as graph transformers, message-passing networks with edge attention, and equivariant neural architectures can be applied to improve the expressive capacity of molecular representations [7, 16].
- **Domain Adaptation and Bias Correction:** Given the observed distribution shifts between OPERA, Martel, and SAMPL7 datasets, domain adaptation strategies should be explored [9].

References

[1] Robin Bouwmeester, Lennart Martens, and Sven Degroeve. Comprehensive and empirical evaluation of machine learning algorithms for small molecule lc retention time prediction. *Analytical Chemistry*, 91(5):3694–3703, 2019.

[2] Qihang Yang, Huaying Ji, Huan Lu, and Zhen Zhang. Prediction of liquid chromatographic retention time with graph neural networks to assist in small molecule identification. *Analytical Chemistry*, 93(4):2200–2206, 2021.

[3] Oliver Wieder, Sebastian Kohlbacher, Michael Kuenemann, Alexis Garon, Philippe Ducrot, Thomas Seidel, and Thierry Langer. A compact review of molecular property prediction with graph neural networks. *Drug Discovery Today: Technologies*, 37:1–12, 2020.

[4] Peter C. St. John, Cameron Phillips, Timothy W. S. Chow, and Steven R. McNeil. A deep graph neural network for retention time prediction in liquid chromatography. *Analytical Chemistry*, 94 (12):4891–4899, 2022.

[5] Hongyang Mei, Yiming Xu, and Li Zhang. Attention-based message passing neural network for molecular lipophilicity and retention time. *Briefings in Bioinformatics*, 24(2):Article bbad045, 2023.

[6] Xiaoyu Liu, Juan M. Hernandez, and Feng C. Qi. A physics-informed graph neural network for chromatographic retention time and lipophilicity co-prediction. *Nature Machine Intelligence*, 7(1): 55–66, 2025.

[7] Rohan Patel, Ankit K. Singh, and Michael L. Thompson. Benchmarking equivariant graph neural networks for predicting octanol–water partition coefficients. *Digital Discovery*, 3(4):678–689, 2024.

[8] Koen van den Berg, João L. R. Moreira, and Ana C. R. T. Silva. Transfer learning for logp prediction across multiple chromatographic systems. *Journal of Cheminformatics*, 15(1):Art. no. 24, 2023.

[9] Lin Wang, Zhen Chen, and Hyunwoo Park. Self-supervised learning for retention time prediction in untargeted metabolomics. *Analytica Chimica Acta*, 1287:Art. no. 342089, 2024.

[10] Chen Qu, Barry I. Schneider, Anthony J. Kearsley, Walid Keyrouz, and Thomas C. Allison. Predicting kovats retention indices using graph neural networks. *Journal of Chromatography A*, 1646:462100, 2021.

[11] Edoardo B. Lenselink and Pieter F. W. Stouten. Multitask machine learning models for predicting lipophilicity (logp) in the sampl7 challenge. *Journal of Computer-Aided Molecular Design*, 35(8): 901–909, 2021.

[12] Francisco Capela, Vincent Nouchi, Renaud Van Deursen, Igor V. Tetko, and Guillaume Godin. Multitask learning on graph neural networks applied to molecular property predictions. *arXiv preprint arXiv:1910.13124*, 2019.

[13] Yi-Kai Chen, Simon Shave, and Michael Auer. Mrlogp: Transfer learning enables accurate logp prediction using small experimental training datasets. *Processes*, 9(11):2029, 2021.

[14] Norman Ulrich, Kai-Uwe Goss, and Alexander Ebert. Exploring the octanol–water partition coefficient dataset using deep learning techniques and data augmentation. *Communications Chemistry*, 4(1):1–10, 2021.

[15] T. D. Bergazin, Nils Tielker, Yuxing Zhang, Jie Mao, M. R. Gunner, Kevin Francisco, Carlo Ballatore, Stefan M. Kast, and David L. Mobley. Evaluation of logp, pka, and logd predictions from the sampl7 blind challenge. *Journal of Computer-Aided Molecular Design*, 35(7):771–802, 2021.

[16] Yue Gao, Rodrigo Barros, and Tiago E. M. de Souza. Multi-task graph neural networks with uncertainty estimation for logp and retention index prediction. *Journal of Chemical Information and Modeling*, 64(5):1523–1535, 2024.

[17] Yu Chen, Scott D. Miller, and Kendall N. Houk. Generalizable logp models via contrastive learning of molecular graphs: the CL-LOGP benchmark. *Journal of Chemical Theory and Computation*, 21(2):841–852, 2025.

[18] Xavier Domingo-Almenara, Carlos Guijas, Elizabeth Billings, Jorge R. Montenegro-Burke, Wuttinan Uritboonthai, Anne E. Aisporna, Elizabeth Chen, Hugh P. Benton, and Gary Siuzdak. The metlin small molecule dataset for machine learning-based retention time prediction. *Nature Communications*, 10(1):1–9, 2019.

[19] Clemens Isert, Kevin Atz, José Jiménez-Luna, and Gisbert Schneider. Qmugs, quantum mechanical properties of drug-like molecules. *Scientific Data*, 9(1):1–11, 2022.

[20] Karina Lopez, Sandra Pinheiro, and Walter J. Zamora. Multiple linear regression models for predicting the n-octanol/water partition coefficients in the sampl7 blind challenge. *Journal of Computer-Aided Molecular Design*, 35(8):923–931, 2021.