

Prediction of Human Liver Microsome Clearance with Chirality-Focused Graph Neural Networks

Chengtao Pu, Lingxi Gu, Yuxuan Hu, Weijie Han, Xiaohe Xu, Haichun Liu, Yadong Chen,* and Yanmin Zhang*



Cite This: *J. Chem. Inf. Model.* 2024, 64, 5427–5438



Read Online

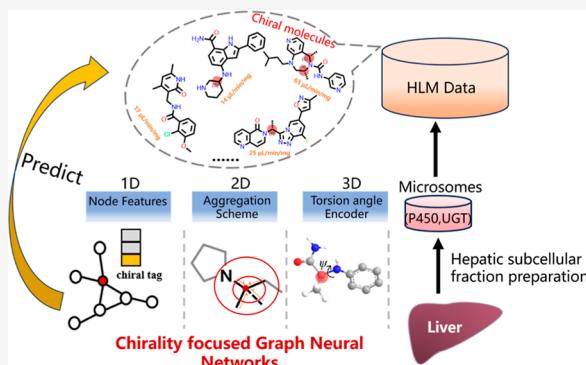
ACCESS |

Metrics & More

Article Recommendations

Supporting Information

ABSTRACT: In drug candidate design, clearance is one of the most crucial pharmacokinetic parameters to consider. Recent advancements in machine learning techniques coupled with the growing accumulation of drug data have paved the way for the construction of computational models to predict drug clearance. However, concerns persist regarding the reliability of data collected from public sources, and a majority of current in silico quantitative structure–property relationship models tend to neglect the influence of molecular chirality. In this study, we meticulously examined human liver microsome (HLM) data from public databases and constructed two distinct data sets with varying HLM data quantity and quality. Two baseline models (RF and DNN) and three chirality-focused GNNs (DMPNN, TetraDMPNN, and ChiRo) were proposed, and their performance on HLM data was evaluated and compared with each other. The TetraDMPNN model, which leverages chirality from 2D structure, exhibited the best performance with a test R^2 of 0.639 and a test root-mean-squared error of 0.429. The applicability domain of the model was also defined by using a molecular similarity-based method. Our research indicates that graph neural networks capable of capturing molecular chirality have significant potential for practical application and can deliver superior performance.



INTRODUCTION

Properly balanced absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties play an important role in the development of new drugs. Many clinical trial failures have been attributed to deficiencies in ADMET properties.^{1,2} Drug clearance, often referred to as total body clearance (CL_{tot}) that comprises both hepatic clearance and renal clearance, is a key pharmacokinetic (PK) parameter that measures the efficiency of the body in eliminating a drug.^{3,4} One primary strategy for assessing hepatic clearance involves measuring the metabolism of the drug in human liver microsomes. Liver microsomes which are products of hepatic subcellular fraction preparation, contain many drug-metabolizing enzymes, such as cytochromes P450 (CYPs) and uridine 5'-diphospho-glucuronosyltransferases (UGTs).⁵ Utilizing the in vitro–in vivo extrapolation (IVIVE) method,⁶ data derived from these in vitro experiments could be employed for the estimation of hepatic clearance. However, conducting large-scale wet experiments in the early stages of drug discovery is often impractical. With the rapid development of artificial intelligence in recent years, researchers seek to use machine learning models to screen out these compounds containing undesirable ADMET properties.^{7–9}

Machine learning methods, which could make reasonable predictions when facing new and unseen data by utilizing

historical data, have been widely applied to human liver microsome clearance prediction. Lee et al.¹⁰ used random forest and Bayesian classification methods with MOE, E-state descriptors, ADME Keys, and ECFP_6 fingerprints to develop highly predictive models on Pfizer private laboratory human liver microsome (HLM) clearance data. Sakiyama et al.¹¹ expanded more machine learning methods for modeling human liver microsomal stability including support vector machine (SVM), logistic regression, and recursive partitioning. Rather than on single species, Hu et al.¹² used the naïve Bayesian classifier based on FCFP_6 fingerprints to build global classification models for rat, mouse, and human liver microsomal stability. Rodríguez-Pérez et al.¹³ also utilized a multitask learning architecture to model the CL_{int} of six species simultaneously.

Since many research studies were carried out earlier, the methods have remained rooted in traditional machine learning approaches, necessitating preprocessing steps such as feature

Received: February 10, 2024

Revised: May 27, 2024

Accepted: June 20, 2024

Published: July 8, 2024



calculation and selection. Also, most of these studies were classification models for the consideration of the complex mechanism behind the stability of liver microsome. However, a unified classification threshold does not yet exist. In recent years, with the rapid development of deep learning and artificial intelligence, one of the most eye-catching advancements is the development of graph neural networks (GNNs) and their application on molecular property prediction task.^{14–16} GNNs are algorithms specially proposed for the data structure of graphs, and the molecule is composed of atoms and bonds that could be viewed as graphs very naturally. As Wu et al.¹⁷ showed, graph-based models outperformed conventional methods on 11/17 data sets ranging from quantum mechanics to physiology, and could achieve comparable performances on the remaining data sets. When constructing graph neural network models, the chirality of molecules, typically represented by a carbon atom with four nonequivalent bonded neighbors, was often overlooked. Two molecules that differ only in tetrahedral chirality are called enantiomers. Enantiomers share many chemical properties such as boiling/melting points and electronic energies but could display strikingly different behavior when it comes to pharmacokinetic or toxicology. A classic example is thalidomide, where the R-enantiomer is sedative and the S-enantiomer is teratogenic, so does the human liver microsome clearance (Figure 1). Recently, several advanced GNNs have been proposed to incorporate the chirality of molecules.^{18–20} However, their applications in real-world tasks are far from explored.

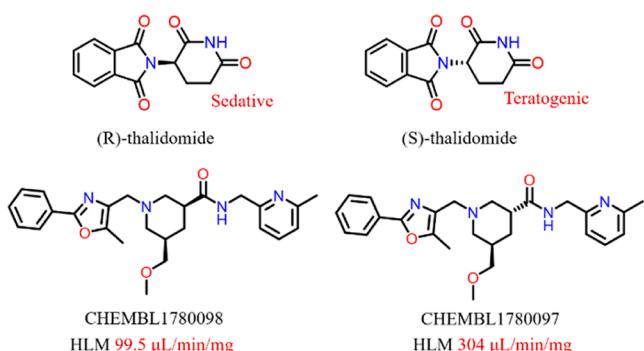


Figure 1. Cases of enantiomers with distinct pharmacokinetic and toxicological behaviors.

Apart from methodological considerations, what impedes the development of artificial intelligence methods on the ADMET prediction task is ascribed to data-related challenges. Most studies on human liver microsome stability used in-house data from large pharmaceutical companies (like Pfizer^{11,21} and NOVARTIS²²). These data sets were not available to the public, although they tended to be more uniform and less noisy and have higher quality. Currently, the ChEMBL database²³ is one of the largest bioactivity databases which cover many drug-like properties and has been widely used as a data source for constructing quantitative structure–property relationship (QSPR) models. We found that there were certain inconsistencies in the human liver microsome clearance data from the ChEMBL data set compared to the original publication as was also noticed by Esaki et al.²⁴ These data were calibrated in our study, and a more detailed description could be found in **Methods and Materials**. Fang et al.²⁵ have also disclosed a portion of the data set regarding the end point of ADMET properties which contain human liver microsome clearance, and these data were also incorporated into our study.

In summary, we manually collected a HLM data set from the article and ChEMBL bioactivity data set meticulously, which, to our knowledge, was the largest publicly available data set for human liver microsome clearance. Then, we trained three graph neural networks which could capture the chirality of the molecule and two baseline models based on these data and evaluated their performance using multiple metrics. By comparing with published works, our regression model achieved better performance on a larger data set. The whole workflow is shown in Figure 2.

METHODS AND MATERIALS

Data Collection and Preparation. The original HLM data was collected from the ChEMBL bioactivity database (ver.31)²³ by using structured query language (SQL) to query the whole data set. We specified the standard type, standard relation, assay tissue, and assay subcellular fraction of the data, and a snippet of the query can be found in **Figure S1**. All items related to glucuronidation (phase II metabolism), which is mediated by UGT and uridine diphosphate glucuronic acid (UDPGA) cofactor, were excluded first. We also removed the data associated with specific CYP subtypes or including inhibitors targeting CYP enzymes.²⁶ Entries lacking the NADPH cofactor, which is crucial for phase I metabolism,

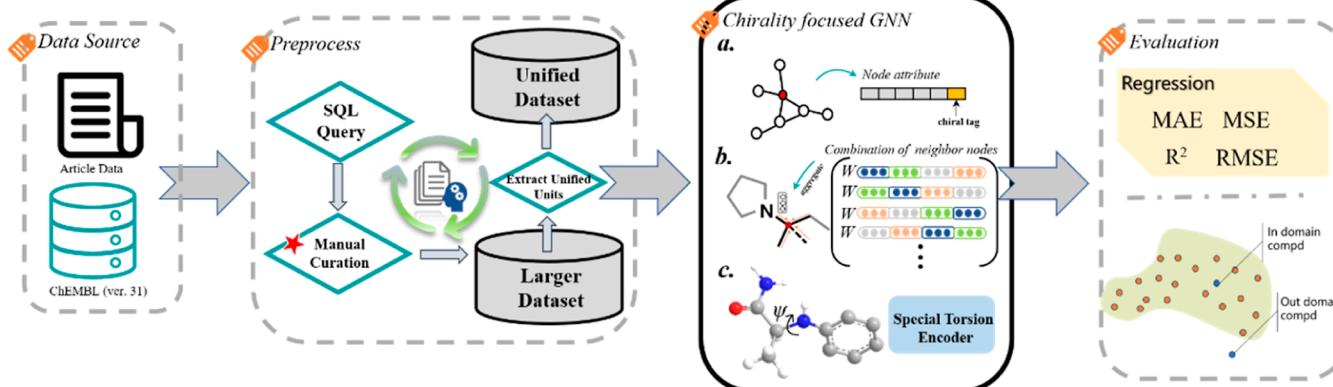


Figure 2. Overview of the workflow. In the preprocess stage, data was achieved by using SQL to query the ChEMBL database.

Datasets Construction

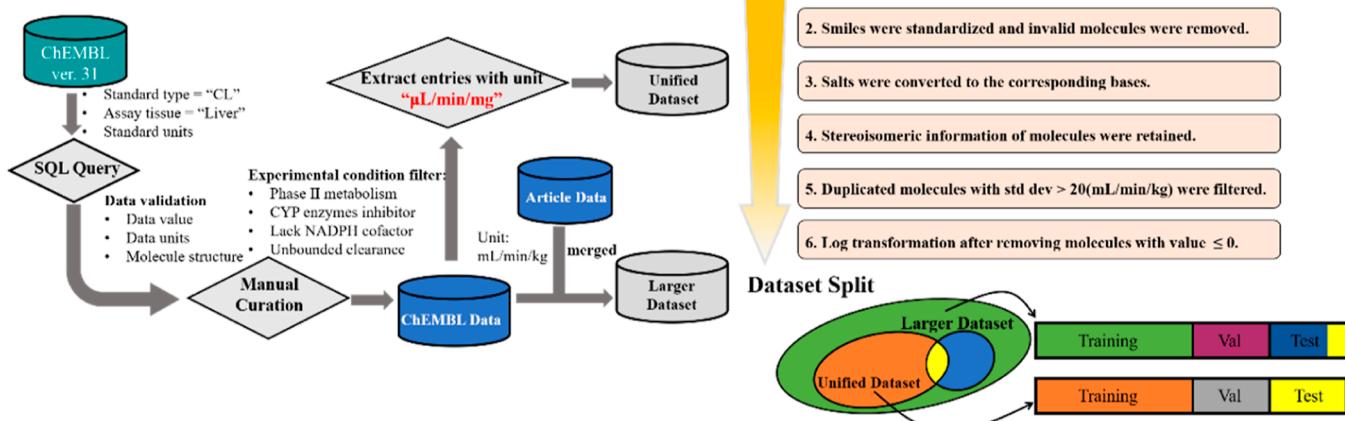


Figure 3. Construction of the “Unified Dataset” and “Larger Dataset”, preprocessing steps, and split of both data sets.

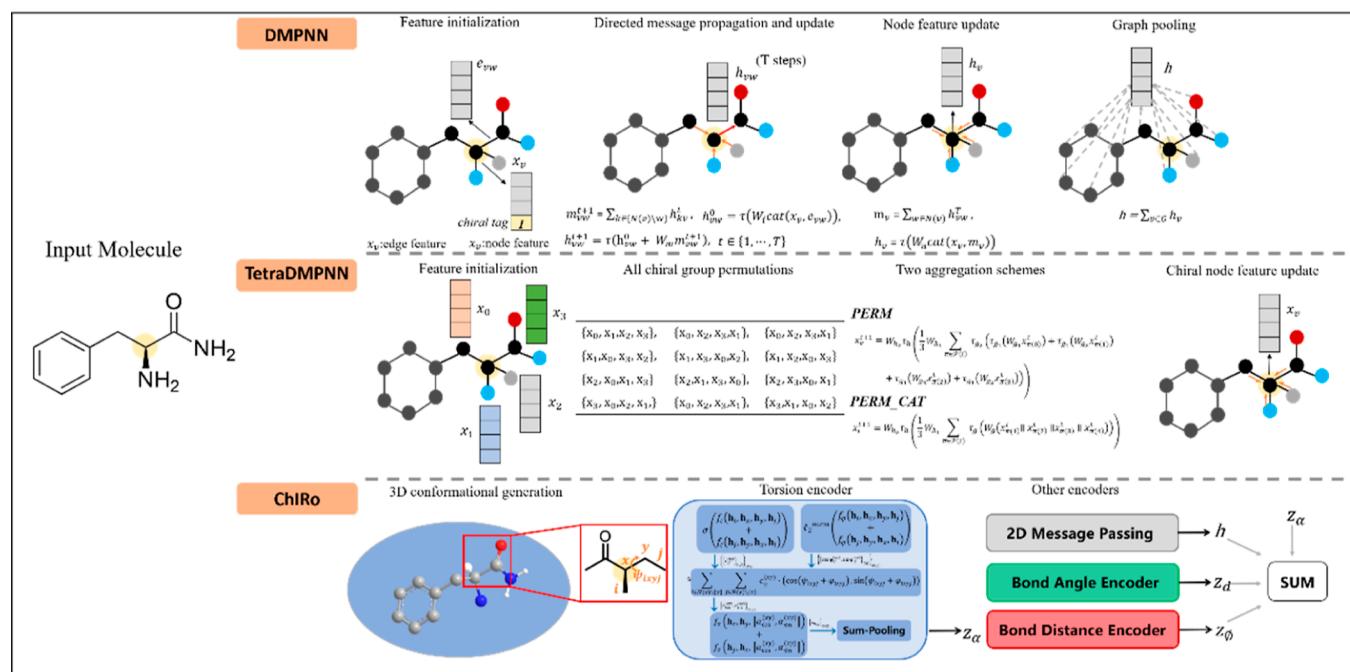


Figure 4. Basic framework of the DMPNN, TetraDMPNN, and ChIRo models as well as their respective module for capturing chirality.

were excluded. Lastly, any data related to unbounded clearance was deleted.

In addition to filtering based on experimental conditions, the HLM data were further examined manually based on roughly 1400 original publications from over a dozen different journals. It is noteworthy that the reported clearance for liver microsomes in scientific literature could exist at various levels, such as the cellular or organ level. These varying levels of microsomal clearance are typically inferred using the IVIVE method, as shown in eq 1. The most intuitive way to recognize these variations is by observing the units used to report the clearance data. Unlike the in-house data set, which adheres to a standard experimental protocol, data collected from public sources may exhibit more inconsistencies. The most palpable of these is when transitioning between different levels of clearance with varying scaling factors, which could introduce additional noise into the data set. Therefore, in this study, we

isolated a subset labeled the “Unified Dataset”, which solely consists of data measured at the cellular level and using the unit of “ $\mu\text{L}/\text{min}/\text{mg}$ ” in the original literature to mitigate this type of noise. Furthermore, we combined the HLM data from Fang et al.²⁷ with all the HLM data from ChEMBL to generate the “Larger Dataset”. Finally, all entries in both data sets were transformed into “ $\text{mL}/\text{min}/\text{kg}$ ” with scaling factor 1 set to 45 and scaling factor 2 set to 20 in eq 1.

$$\text{CL}_{\text{scaled}} = \frac{0.693}{T_{1/2}} \times \frac{\text{incubation volume}}{\text{mg of microsomal protein}} \times \frac{45^a \text{ mg of microsomal protein}}{\text{grams of liver}} \times \frac{20^b \text{ grams of liver}}{\text{kg of body weight}} \quad (1)$$

where “*a*” represents scaling factor 1 and “*b*” represents scaling factor 2.

In addition, crucial preprocessing steps, such as removal of invalid molecules and elimination of duplicate molecules, have been applied to both data sets. For the data set split, given that the “Unified Dataset” constitutes a subset of the “Larger Dataset”, we first divided the latter into training, validation, and test set with a ratio of 8:1:1. Subsequently, the intersection of the test set with the “Unified Dataset” was employed as the test set for the “Unified Dataset”. The validation set was sampled to match the test set’s molecule count, and any remaining molecules were incorporated into the training set. Figure 3 illustrates the assembly of both data sets, the preprocessing steps for each, and the strategy employed for data splitting.

Molecular Representation. In this study, we utilized two distinct types of molecular representation. The first type was based on molecular fingerprints, while the second was based on graphs. 1024 bits of extended-connectivity fingerprints²⁸ with a radius of 2 (ECFP4) were employed to develop models for random forest (RF) and deep neural networks (DNNs). The graph representation, rooted in atoms and bonds present within the molecule, was constructed for the GNNs.²⁹ The atom and bond features are described in detail in the Supporting Information (Table S1). For the ChIRO method, RDKit³⁰ was utilized to acquire the three-dimensional coordinates of the molecules, and the molecules were optimized using the universal force field. One important note is that, to ensure a fair comparison, the chiral tags of atomic features for TetraDMPNN and ChIRO were masked.

Model Construction. Totally, three graph neural networks capable of capturing molecular chirality were employed in our study: (1) directed message-passing neural networks (DMPNN), (2) TetraDMPNN, and (3) chiral interrotor-invariant neural network (ChIRO). Based on the GNN’s model architecture, more focus would be put on how each method senses chirality instead, as illustrated in Figure 4 and described as follows. Besides, RF and DNN were incorporated as baseline models, considering the popularity of these two methods in ADMET property prediction.

DMPNN. The DMPNN³¹ based on MPNN (message passing neural network) framework contains mainly four stages, feature initialization, directed message propagation, node feature update, and graph pooling. The process of message propagation and update would be repeated in T steps. After the final graph representation is obtained, a simple MLP (Multilayer Perceptron) is added to make predictions. Unlike the other two methods, DMPNN does not include an explicit module to capture molecular chirality. Instead, we added an additional bit that represents whether the atom is a chiral center when constructing node features to consider chirality implicitly.

TetraDMPNN. In the source article,¹⁸ the author put forward two novel node aggregation schemes (PERM and PERM_CAT) for updating the representation of a tetrahedral center x_v . In scheme PERM, W_{hi} ($i \in \{1,2\}$) and W_{gi} ($j \in \{1,2,3,4\}$) represent separate weight matrices where τ_{gm} ($m \in \{1,2\}$) represents nonlinear operators. $x_{\pi(n)}^{(t-1)}$ ($n \in \{1,2,3,4\}$) means different permutations (total 12 permutations) of ordered neighbor node vectors of the chiral center at time step $(t - 1)$. In scheme PERM_CAT, four neighbor nodes were concatenated first, and an individual weight matrix W_g was added before it. Because enantiomers have different

ordered local chiral groups, the schemes of aggregation allow the model to distinguish them while invariant to arbitrary bookkeeping conventions. The aggregation function could be integrated with common graph neural networks, like GCN, GIN, and MPNN. To reduce trainable parameters of the model, we applied scheme PERM_CAT to the DMPNN model and named it as TetraDMPNN. This updated node features of chiral centers with new scheme, while for atoms without chirality, it remained the same process as the DMPNN.

ChIRO. The ChIRO model proposed by Adams³² consists of total four components: the 2D Message Passing part, Bond Angle Encoder, Bond Distance Encoder, and Torsion Encoder. In the 2D Message Passing phase, the edged-conditioned convolution (EConv) and graph attention layers (GAT) were utilized to derive the molecular representation based on the 2D topology. The Bond Distance Encoder and the Bond Angle encoder embed bond distance and angle into learned latent vectors through two MLPs. The most important section of the ChIRO model is the Torsion Encoder, which can capture the chirality of a molecule and be invariant to the rotation of internal molecular bonds at the same time. In the Torsion Encoder, MLP f_c combined with an activation function was applied to node features (h_i, h_x, h_y, h_j) across the torsion angle to get the weight coefficient $\{\{\alpha_{ij}^{(xy)}\}_{(i,j)}\}_{(x,y)}$ and the weighted sum of $\cos(\psi_{ixyj})$ and $\sin(\psi_{ixyj})$ of each coupled torsion, which proved to form a circle, with the radius of the circle remaining unchanged regardless of bond rotation. For chirality invariance, the author added a learned phase shift (ϕ_{ixyj}) which was obtained by using the MLP f_ϕ combined with L2-normalization to the calculation of sine and cosine values of the torsion angle. The weighted sum would result in different radii ($\|\alpha_{\cos}^{(xy)}, \alpha_{\sin}^{(xy)}\|$) for different enantiomers with inverted chiral centers. The Torsion Encoder put the node feature of the internal molecular bond (h_x, h_y) with the learned radius described above together and got the final latent vector z_α by sum-pooling. The 2D graph vector h , learned bond distance vector z_ϕ , learned bond angle vector z_θ , and torsion vector z_α were concatenated to form a new representation. This vector could be fed into an output layer for different regression or classification tasks.

Random Forest. Random Forest is a powerful ensemble learning method that has been widely used in constructing QSPR models.^{33–35} For the regression task, it operates by constructing a multitude of decision trees during the training phase and outputs the average prediction of the individual trees. Each tree is built using a random subset of the training data and a random subset of the features. RF has demonstrated good model performance due to its high accuracy, robustness to outliers, and capability to handle large data sets with high dimensionality.

Deep Neural Networks. DNN³⁶ also known as a multilayer perceptron, typically consists of an input layer, multiple hidden layers, and an output layer. Each layer performs a transformation on the input with a weight matrix. In this study, the number of neurons in the hidden layers was set in a pyramid fashion.³⁷ This approach has the advantage of enabling the network to learn to extract simpler features in the early layers and then combine them to form more complex representations in the deeper layers. Additionally, this network architecture also helps in reducing the number of parameters to be trained.

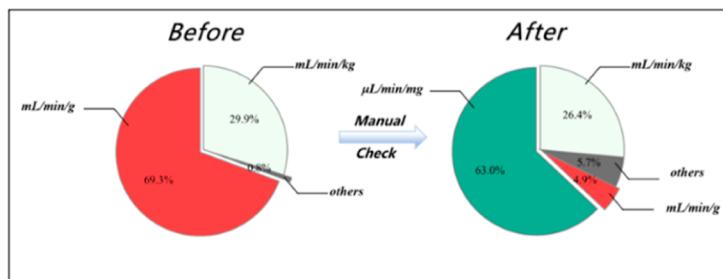
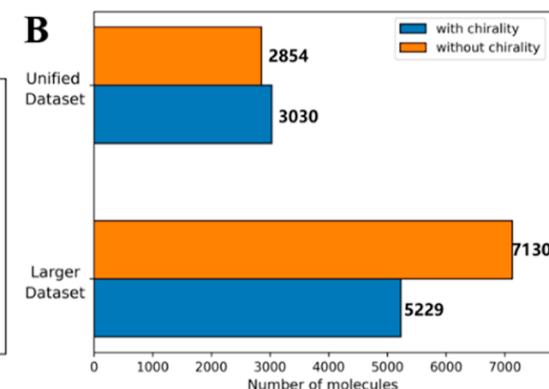
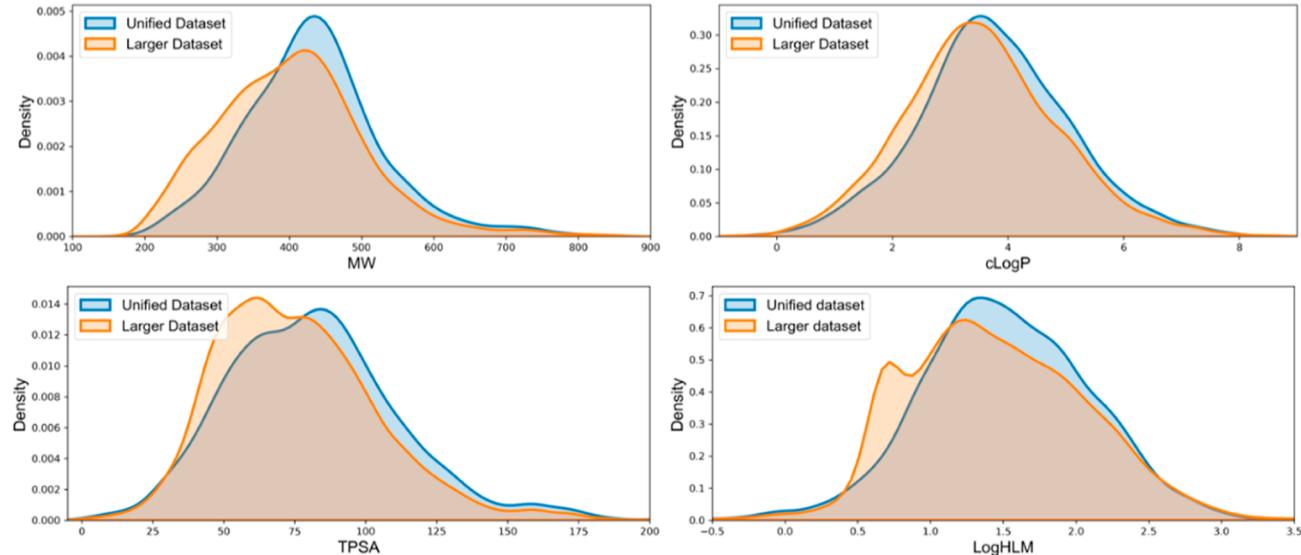
A**B****C**

Figure 5. (A) Units distribution of ChEMBL's HLM data before and after manual check. (B) Count of chiral and nonchiral molecules in the “Unified Dataset” and “Larger Dataset”. (C) Molecular property (MW, cLog P, TPSA, and logHLM) distribution calculated from both data sets. (Tips: the racemic compounds are categorized as “without chirality” in our study.)

Model Evaluation. The performance of models was evaluated on the test data set. In this study, we employed four commonly used metrics in regression tasks:³⁸ mean absolute error (MAE), mean-squared error (MSE), root-mean-squared error (RMSE), and R-squared (R^2). MAE (eq 2) measures the average absolute differences between predicted and actual values. MSE (eq 3) measures the average of the squares of the errors, penalizing larger errors more heavily than the MAE. RMSE (eq 4) is the square root of the MSE and provides an easy-to-interpret measure of the average error in the same unit as the target variable. R^2 (eq 5) represents the proportion of variance of y that has been explained by the independent variables in the model, and a value closer to 1 indicates better model performance.

$$\text{MAE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} |y_i - \hat{y}_i| \quad (2)$$

$$\text{MSE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} (y_i - \hat{y}_i)^2 \quad (3)$$

$$\text{RMSE}(y, \hat{y}) = \sqrt{\frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} (y_i - \hat{y}_i)^2} \quad (4)$$

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (5)$$

Hyperparameter Optimization and Model Training Protocols.

We optimized the hyperparameters for the DNN and three GNNs using the Optuna hyper-optimization framework.³⁹ A total of 50 trials were conducted using TPESampler to sample new trials for model training. The HyperbandPruner algorithm was employed to prune unpromising trials on the basis of the RMSD on the validation set. For the RF model, we randomly sampled 50 groups of hyperparameters and found the best hyperparameter with the lowest RMSD on the validation set. The whole searching space could be found in the Supporting Information (Table S2).

After the optimal hyperparameters were determined, both the GNNs and the DNN model were trained on the training set for a total of 200 epochs. The mean-squared error between the predicted values and the labels was calculated, and the Adam optimizer⁴⁰ was used to update the parameters of the neural networks. The best model was saved based on the RMSE value on the validation set. The final model performance was assessed on a test set. The RF model was also saved based on the RMSE on the validation set and evaluated on the test set.

Definition of Applicability Domain. Defining the applicability domain (AD) of the model which was emphasized by OECD (Organization for Economic Co-operation and Development) in its five standards, acts as a key component in building QSAR models.⁴¹ In this study, we adopted a kNN (*k*-nearest neighbors) AD method⁴² which is based on the structural similarity between train sets and test sets. Each compound was represented as 1024 ECFP4 fingerprints. This method defines a distance cutoff value D_c to determine whether a compound is in the domain or out domain. The detailed formula is as follows

$$D_c = \bar{y} + Z\sigma \quad (6)$$

where \bar{y} represents the average Euclidean distance of the *k* nearest neighbors for each compound in the training set, σ is the corresponding standard deviation, and Z is an empirical parameter utilized to regulate the significance level. The structural similarity between each compound in the test set and training set was calculated, *k* nearest compounds were kept, and the average value was calculated as D_i . The compound was judged as out of domain (OD)⁴³ if D_i is larger than D_c ; otherwise, it had fallen into the domain (ID).

RESULTS AND DISCUSSION

Data Set Analysis. The HLM data used in this research was primarily obtained from Fang et al.²⁷ and the public database ChEMBL. It should be emphasized that we carefully inspected these entries from ChEMBL based on the original literature studies, discarding data lacking source references. Notably, we corrected various inconsistencies, such as mislabeled units, inaccurate data records, and flawed unit conversion. Figure 5A presents the unit distribution of the ChEMBL's HLM data pre and post manual verification. One of the most notable issues was that many data points originally labeled “ $\mu\text{L}/\text{min}/\text{mg}$ ” were inaccurately marked as “ $\text{mL}/\text{min}/\text{g}$ ”. Subsequently, after filtering out some data that did not meet the experimental conditions, 3087 entries from the article²⁵ were merged into the remaining data. We conducted the preprocessing on the entire data set, removing 88 duplicate molecules with a standard dev $>20 \text{ mL}/\text{min}/\text{kg}$ and 55 entries whose label <0 , as well as 3 molecules that failed to embed into three-dimensional space. Ultimately, we obtained the “Larger Dataset” containing 12 359 molecules. The “Unified Dataset”, which was extracted from the “Larger Dataset” and retaining only entries with original unit “ $\mu\text{L}/\text{min}/\text{mg}$ ”, comprised a total of 5884 entries. The number of molecules in the “Larger Dataset” was approximately twice that of the Unified Dataset.

Based on the two established data sets, we conducted a statistical analysis of the number of chiral molecules contained in each data set first (Figure 5B). The “Larger Dataset” contains 42.3% chiral molecules, while the proportion in the “Unified Dataset” increases to 51.5%, signifying that over half the molecules are chiral. Analyzing the HLM data gathered from public sources, it was evident that chirality was a noteworthy aspect. Kernel density curves of fundamental molecular characteristics, including molecular weight (MW), cLog P, TPSA, and Log HLM, were also plotted for each data set. As can be seen in Figure 5C, the property distributions of the two data sets were generally similar, with MW concentrated in the range of 250–600 Da, cLog P primarily between 1 and 6, TPSA between 25 and 125, and Log HLM mainly between 0.5 and 2.5. Subsequently, we divided both data sets into training, validation, and test sets, in accordance

with the methodology detailed in the “Materials and Methods” section. A comprehensive count of molecules within each set is provided in Table 1.

Table 1. Size of the Training, Validation, and Test Sets for Two HLM Datasets

data set	training	validation	test	total
Unified Dataset	4726	579	579	5884
Larger Dataset	9888	1235	1236	12 359

To better showcase the data sets we constructed, as shown in Figure 6, t-SNE⁴⁴ and PCA,⁴⁵ which have been widely used for visualizing high-dimensional data into a lower-dimensional space, were used to visualize the chemical space^{46,47} of the two data sets. Each data set covers a certain range of chemical space. At the same time, it could be seen that the partitioning of the training, validation, and test sets was reasonable with no individual set clustering in a specific area. This allowed us to better validate, evaluate, and compare the models used in this study.

Model Evaluation. In this study, we utilized two conventional methods (RF and DNN) as well as three GNNs to construct regression models for predicting HLM. The first two models primarily serve as comparative baseline models, and the three subsequent GNNs each contain specific modules designed to capture molecular chirality information. The performance of the built models was evaluated on the test set of both data sets.

In the “Unified Dataset”, DMPNN performed well on this data set, achieving highest test R^2 (0.546) and lowest test RMSE (0.419). TetraDMPNN’s performance with test R^2 (0.533) and test RMSE (0.419) was inferior to DMPNN but performed slightly better than two baseline models. The last method, ChiRo, showed the poorest performance on the test set, not even reaching the metrics of the baseline models. The more detailed statistical results are presented in Table 2.

When it comes to the “Larger Dataset”, the test R^2 values for RF and DNN were 0.516 and 0.591, respectively. The TetraDMPNN model which extracts chiral information explicitly from molecular topology achieved the highest test set R^2 (0.639) and lowest test RMSE (0.429). The test R^2 and RMSE of the DMPNN method on the “Larger Dataset” were worse than those of TetraDMPNN. For the ChiRo model, it achieved comparable performance with two baseline models. The more detailed statistical results are presented in Table 3.

Accessing the Impact of Data Quality and Quantity on the Model. In this study, we constructed two data sets, where the “Unified Dataset” is a subset of the “Larger Dataset” and contains about half the number of molecules compared to the latter. We aimed to explore whether these models could benefit from a larger amount of data, although the increased data might contain some noise.

As shown in Figure 7 (blue and orange bars), the MAE, MSE, and RMSE values of the test set slightly increased for the “Larger Dataset”. Even the top-performing model, TetraDMPNN, demonstrated a performance comparable to that on the “Unified Dataset”. In terms of R^2 , except for the traditional machine learning algorithm RF, all deep learning methods showed improvement in the “Larger Dataset”. It is worth noting that metrics MAE, MSE, and RMSE share a common drawback that their values could range between zero and $+\infty$. Therefore, they are affected by outliers in the

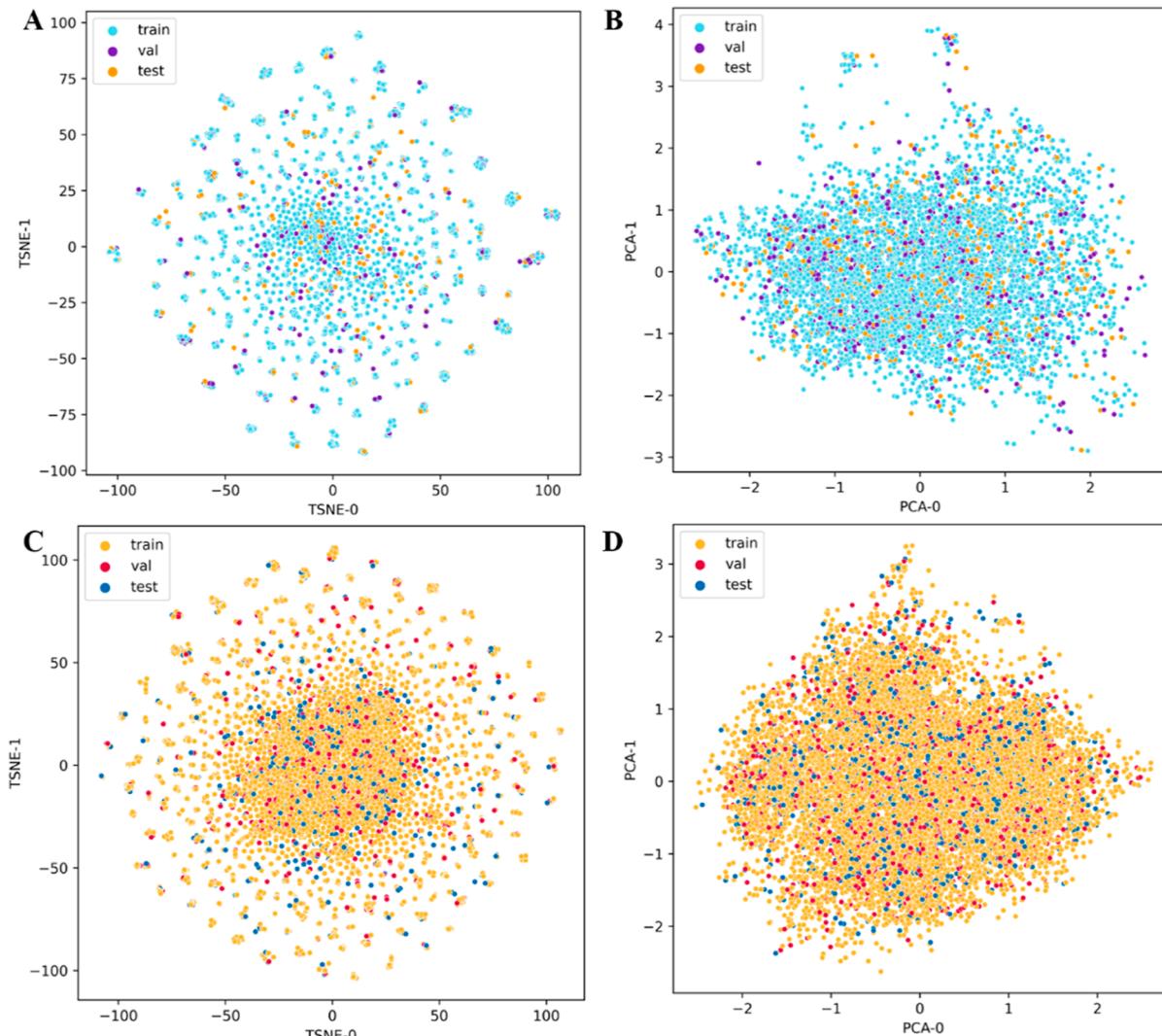


Figure 6. Chemical space distribution visualization. (A,B) t-SNE and PCA of the training set (blue), validation set (purple), and test set (orange) for the Unified Dataset. (C,D) t-SNE and PCA of the training set (orange), validation set (red), and test set (blue) for the Larger Dataset.

Table 2. Regression Metrics for RF, DNN, DMPNN, TetraDMPNN, and ChiRo on the Test Set of “Unified Dataset” (Best in Bold, Second in Italics)

model	Unified Dataset			
	test MAE	test MSE	test RMSE	test R^2
RF	0.322	0.182	0.426	0.530
DNN	0.383	0.276	0.526	0.516
DMPNN	0.312	0.176	0.419	0.546
TetraDMPNN	0.318	0.180	0.425	0.533
ChiRo	0.341	0.199	0.446	0.485

data set. It was suggested by Chicco et al.⁴⁸ using R^2 as a standard metric in regression tasks. However, it remains inconclusive whether deep learning methods truly benefit from larger data volumes, given that the “Larger Dataset” test set may contain a higher proportion of simpler molecules.

To delve deeper into this matter, we have ensured that all entries in the test set of “Unified Dataset” could be found in the “Larger Dataset” test set during the data set splitting process. The performance of the model was then evaluated within this subset (as shown in Table 4), and we also depicted this performance graphically in Figure 7 (indicated by the

Table 3. Regression Metrics for RF, DNN, DMPNN, TetraDMPNN, and ChiRo on the Test Set of “Larger Dataset” (Best in Bold, Second in Italics)

model	Larger Dataset			
	test MAE	test MSE	test RMSE	test R^2
RF	0.383	0.276	0.526	0.516
DNN	0.354	0.233	0.483	0.591
DMPNN	0.324	0.196	0.442	0.616
TetraDMPNN	0.318	0.184	0.429	0.639
ChiRo	0.354	0.233	0.482	0.592

yellow bar). We observed that MAE, MSE, and RMSE values for this subset were essentially identical with those of the “Unified Dataset” test set. For the TetraDMPNN model, these values were even lower, further reinforcing the hypothesis that the increase in these indicators was due to outliers in the data set. In terms of R^2 , only the DNN and TetraDMPNN methods outperformed models trained exclusively on the “Unified Dataset”, with the TetraDMPNN model exhibiting the best performance. This suggests that models trained on larger, noisier data sets do not necessarily outperform those trained on a smaller, less noisy data set. However, the findings also

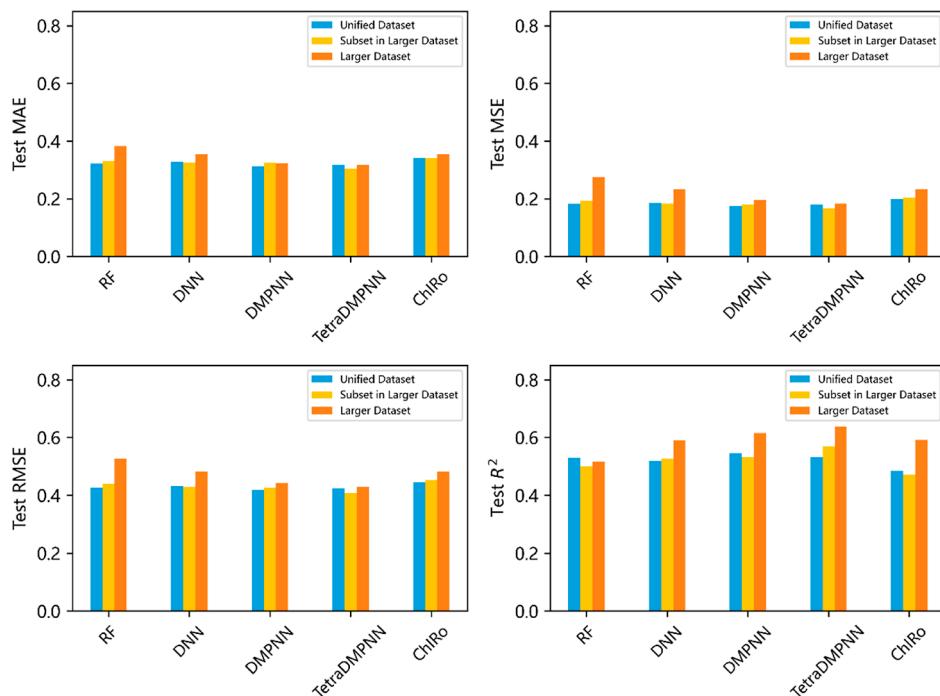


Figure 7. Comparisons of the model test set performance on the “Unified Dataset”, “Subset in Larger Dataset”, and “Larger Dataset”.

Table 4. Regression Metrics for RF, DNN, DMPNN, TetraDMPNN, and ChiRo on the Subset of “Larger Dataset” (Best in Bold, Second in Italics)

model	subset in Larger Dataset			
	test MAE	test MSE	test RMSE	test R^2
RF	0.332	0.193	0.440	0.501
DNN	0.326	0.184	0.429	0.526
DMPNN	0.325	0.181	0.426	0.532
TetraDMPNN	0.305	0.167	0.408	0.570
ChiRo	0.341	0.205	0.453	0.472

highlighted that deep learning methods have certain advantages over traditional machine learning methods RF when trained on a larger data set, except for the ChiRo method

(which will be discussed in the following part). Notably, the TetraDMPNN method showed considerable promise, achieving the best performance on our compiled data sets.

Comparison of Different GNNs Focusing on Molecular Chirality. Considering the abundance of chiral molecules in the two HLM data sets obtained from public sources, we believed that this crucial information should be leveraged in constructing property prediction tasks. In this study, we employed three GNNs known for effectively capturing the molecular chirality. The DMPNN incorporates the chirality information by using chiral bits on atomic features, while the TetraDMPNN utilizes different permutations of neighboring atoms around chiral centers. The ChiRo, on the other hand, employs a specially designed torsion encoder to capture molecular chirality. Interestingly, these three methods consider

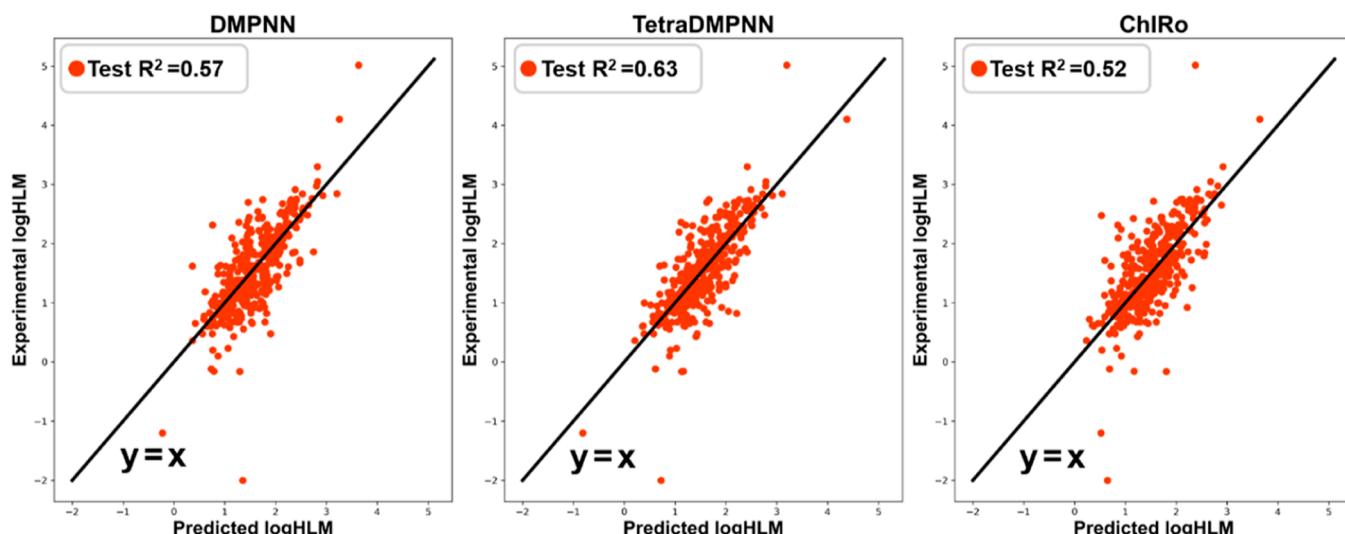
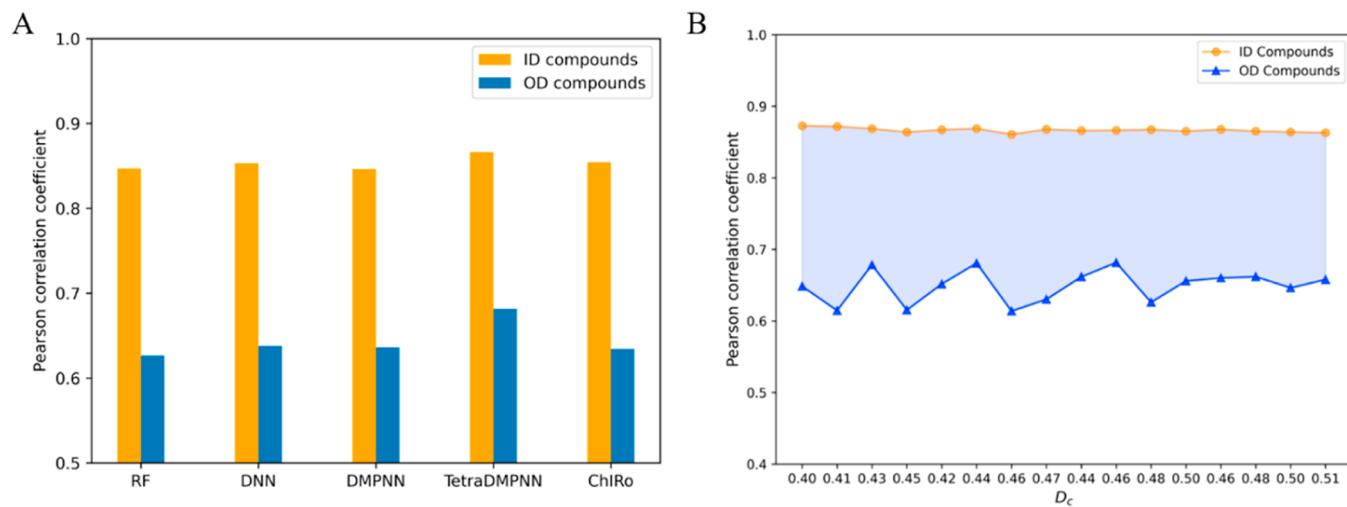


Figure 8. Plots of experimental human Log HLM versus in silico prediction on the chiral subset by the three GNN methods.

Table 5. Summary of In Silico Models for HLM Prediction

year	author	data set size	data availability	molecular representation	method	task type	model performance
2007	Lee et al. ¹⁰	14 557	×	MOE, E-state descriptors, ADME keys, ECFP_6 fingerprints	RF, Bayesian classification	class	test acc = 0.75
2008	Sakiyama et al. ¹¹	1952	×	MOE descriptors	SVM, logistic regression, recursive partitioning	class	test acc > 0.8
2010	Hu et al. ¹²	6426	×	FCFP_6 fingerprints	Naïve Bayesian classifier	class	test acc = 0.77
2019	Esaki et al. ²⁴	5278	✓	Mordred descriptors, ECFP_4	RF, AdaBoost, SVM	class	test acc = 0.77
2019	Wenzel et al. ³⁷	5348	✓	AP-DP descriptors	DNN	reg	test R^2 = 0.586 (single task)
2021	He et al. ⁴⁹	144 300	×	graph representation	DMPNN	reg	test R^2 = 0.557
2024	ours	12 359 (larger data set)	✓	graph representation	DPMNN, TetraDMPNN, ChIRo	reg	test R^2 = 0.639

**Figure 9.** (A) Pearson correlation coefficient of ID and OD compounds with different models. (B) Variation of the Pearson correlation coefficient of ID and OD compounds against D_c for the TetraDMPNN model.

chirality from 1D, 2D, and 3D perspectives of the molecule, respectively.

To provide a more comprehensive comparison of the three GNN models' performances, we gathered all chiral molecules, a total of 399 instances, from the test set of the "Larger Dataset". We then evaluated the performance of the models for this chiral subset. Figure 8 showcases the scatter of the predicted Log HLM values and experimental values on the chiral subset. As demonstrated, TetraDMPNN outperformed other models, while DMPNN secured second place. Disappointingly, the ChIRo model evidenced the lowest performance. The underperformance of ChIRo might be attributed to the limited precision of molecular conformation, which was generated by a low-precision force field, and its bond distance encoder, bond angle encoder, and torsion encoder might fail to effectively learn the molecular information. Additionally, the 3D information was obtained under vacuum conditions, which could significantly differ from the biological conditions in the binding site of metabolic enzymes. Even when the model is adept at processing 3D information, the accuracy of its learning is contingent upon the initial data. If the initial 3D information is flawed or incorrect, it inevitably leads to inaccuracies in the knowledge the model acquires. DMPNN captures chirality implicitly through chiral tag; it is possible that chiral information was submerged during the message propagation process. TetraDMPNN could be viewed as a modification of the DMPNN method, which

adopts an explicit module to sense the chirality of molecules. It achieved the best performance in our chiral subset with $R^2 = 0.63$. In our study, we have not employed the model to predict the properties of enantiomers, primarily because our data set consists more of chiral individual molecules as opposed to pairs of enantiomers. Given the relatively low proportion of enantiomers, their distinctions may become subdued during the training process. Notwithstanding, we place a higher emphasis on the model's utilization of the property of molecular chirality. All things considered, the TetraDMPNN outperforms the other two methods in capturing molecules at the 2D topological level.

Comparison with Previous Studies. Table 5 provides a summary of prior studies of HLM prediction models. Direct comparisons with these in silico models are challenging due to differences in research objectives, data sources, and data sizes for each study. Regarding the availability and volume of data, the "Larger Dataset" we assembled in this study was currently the most extensive among these accessible data sets. Even though some data sets also originated from the ChEMBL database, they were from older versions. We also incorporated entries from literature sources. It is significant to mention that there were certain issues (mislabeled HLM data units) with the data collected by Wenzel et al.³⁷ from ChEMBL, but they have been rectified in our study. When considering the type of task, many previous models were classification models, while the models we constructed were all regression models. Among

them, TetraDMPNN demonstrated exceptional performance with a test R^2 value of 0.639. Furthermore, these GNN methods utilized in this study eliminated the necessity for complex processes such as feature calculation and selection.

AD Analysis. When constructing QSPR models, analyzing the model's AD is often an essential step. In this study, we further examined the AD of five models trained on the Larger Dataset. The threshold D_c on the distance matrix of the training set was calculated based on eq 6, requiring the manual setting of two parameters, k and Z . To find the optimal parameters for each method, we explored different combinations of k (2, 3, 4, 5) and Z (0.2, 0.3, 0.4, 0.5). The number of OD molecules in the test set and the corresponding D_c value are recorded in Table S3.

We evaluated the model's performance on ID or OD compounds using the Pearson correlation coefficient (r),²⁷ a standard measure for gauging the correlation between predicted and actual values in regression tasks. Models are generally effective for molecules that resemble those in the training set, but their predictive accuracy for OD compounds is often lacking. Therefore, we retained the optimal k and Z values (Table S4) for each model that yielded the best predictive performance for the OD compounds.

As depicted in Figure 9A, there was a marginal difference in performance among the models for ID compounds when they were evaluated using the Pearson correlation coefficient. However, the TetraDMPNN model surpassed the others with a correlation coefficient of 0.682 for OD compounds. We further plotted the r value of ID and OD compounds against the threshold D_c (Figure 9B). The performance of the TetraDMPNN model remained stable for the ID compounds, exceeding 0.85. However, at lower threshold levels of D_c , the Pearson correlation coefficient of OD compounds experienced significant fluctuations, suggesting potential struggles of the model in accurately predicting OD compounds that notably differ from those within the training set. As the value of D_c increased, the number of OD compounds tended to stabilize, thus reducing the fluctuations in the correlation coefficient.

CONCLUSIONS

The human liver microsome clearance is an important indicator of drug clearance. In this study, we tackled the quality issues of data from public sources and focused on molecular chirality, which is often overlooked in other molecular property prediction research. Regarding the data, we corrected some errors in the HLM data from the ChEMBL database and combined remaining entries with data published in the literature to construct our data set. In addition, we retained a subset with original unit consistency, which means it contained less noise. Our results indicate that the performance of models trained on a larger data set but contains some noise does not necessarily achieve higher performance than that on a small but cleaner data set. Furthermore, deep learning models tended to benefit from more data compared to traditional machine learning methods like RF. In terms of molecular chirality, we employed three graph neural networks which could capture molecular chirality and compared their performances. With respect to the data we collected, the method which utilizes chirality information from 2D perspective (TetraDMPNN) achieved the highest performance on test sets of the "Larger Dataset", and its corresponding R^2 on the test set of the "Unified Dataset" was also top. The analysis of the AD

for all models further demonstrated the better predictive performance of the TetraDMPNN model for OD compounds.

As the saying goes, "Garbage in, Garbage out". We recommend conducting additional checks on the data obtained from public databases. Also, these models that could capture chirality should be applied to a wider range of practical tasks to test their effectiveness.

ASSOCIATED CONTENT

Data Availability Statement

All scripts used were written based on Python 3.7. The source code of chiral GNNs and HLM data are available on https://github.com/pctpct/HLM_chiral_gnn. The ChEMBL (ver. 31) database used in this study could be accessed from its official Web site (<https://www.ebi.ac.uk/chembl/>). All molecule fingerprints and physicochemical properties of molecules were calculated using the RDKit package.

Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.jcim.4c00243>.

Code snippet of SQL queries; atom and bond features of molecular graph; hyperparameter search space for five methods; number of OD compounds and corresponding threshold D_c at different Z and k values; and best combination of k and Z values for the models trained on the Larger Dataset ([PDF](#))

AUTHOR INFORMATION

Corresponding Authors

Yadong Chen – Laboratory of Molecular Design and Drug Discovery, School of Science, China Pharmaceutical University, Nanjing 211198, China;  orcid.org/0000-0002-2189-9356; Email: ydchen@cpu.edu.cn

Yanmin Zhang – Laboratory of Molecular Design and Drug Discovery, School of Science, China Pharmaceutical University, Nanjing 211198, China;  orcid.org/0000-4075-7556; Email: yanminzhang@cpu.edu.cn

Authors

Chengtao Pu – Laboratory of Molecular Design and Drug Discovery, School of Science, China Pharmaceutical University, Nanjing 211198, China

Lingxi Gu – Laboratory of Molecular Design and Drug Discovery, School of Science, China Pharmaceutical University, Nanjing 211198, China

Yuxuan Hu – State Key Laboratory of Natural Medicines, China Pharmaceutical University, Nanjing 211198, China;  orcid.org/0000-0003-2931-2991

Weijie Han – Laboratory of Molecular Design and Drug Discovery, School of Science, China Pharmaceutical University, Nanjing 211198, China

Xiaohu Xu – Laboratory of Molecular Design and Drug Discovery, School of Science, China Pharmaceutical University, Nanjing 211198, China;  orcid.org/0009-0008-8493-0235

Haichun Liu – Laboratory of Molecular Design and Drug Discovery, School of Science, China Pharmaceutical University, Nanjing 211198, China

Complete contact information is available at:
<https://pubs.acs.org/doi/10.1021/acs.jcim.4c00243>

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was financially supported by the National Natural Science Foundation of China (nos. 82073704 and 81973182) and the “Double World-classes” Construction Program of China Pharmaceutical University.

ABBREVIATIONS

HLM, human liver microsome; ADMET, absorption, distribution, metabolism, excretion and toxicity; PK, pharmacokinetics; CYP, cytochrome; UGTs, uridine 5'-diphosphoglucuronosyltransferases; QSPR, quantitative structure–property relationship; RF, random forest; DNN, deep neural network; GNN, graph neural network; PCA, principal component analysis; t-SNE, t-distributed stochastic neighbor embedding; AD, applicability domain; SQL, structured query language

REFERENCES

- (1) Morgan, P.; Van Der Graaf, P. H.; Arrowsmith, J.; Feltner, D. E.; Drummond, K. S.; Wegner, C. D.; Street, S. D. A. Can the Flow of Medicines Be Improved? Fundamental Pharmacokinetic and Pharmacological Principles toward Improving Phase II Survival. *Drug Discovery Today* **2012**, *17*, 419–424.
- (2) Wang, Y.; Xing, J.; Xu, Y.; Zhou, N.; Peng, J.; Xiong, Z.; Liu, X.; Luo, X.; Luo, C.; Chen, K.; Zheng, M.; Jiang, H. *In Silico* ADME/T Modelling for Rational Drug Design. *Q. Rev. Biophys.* **2015**, *48*, 488–515.
- (3) Iwata, H.; Matsuo, T.; Mamada, H.; Motomura, T.; Matsushita, M.; Fujiwara, T.; Kazuya, M.; Handa, K. Prediction of Total Drug Clearance in Humans Using Animal Data: Proposal of a Multimodal Learning Method Based on Deep Learning. *J. Pharm. Sci.* **2021**, *110*, 1834–1841.
- (4) Toutain, P. L.; Bousquet-mélou, A. Plasma Clearance. *Vet. Pharmacol. Ther.* **2004**, *27* (6), 415–425.
- (5) Smith, D. A.; Beaumont, K.; Maurer, T. S.; Di, L. Clearance in Drug Design: Miniperspective. *J. Med. Chem.* **2019**, *62*, 2245–2255.
- (6) Obach, R. S.; Baxter, J. G.; Liston, T. E.; Silber, B. M.; Jones, B. C.; Macintyre, F.; Rance, D. J.; Wastall, P. The Prediction of Human Pharmacokinetic Parameters from Preclinical and In Vitro Metabolism Data. *J. Pharmacol. Exp. Ther.* **1997**, *283*, 46.
- (7) Cumming, J. G.; Davis, A. M.; Muresan, S.; Haeberlein, M.; Chen, H. Chemical Predictive Modelling to Improve Compound Quality. *Nat. Rev. Drug Discov.* **2013**, *12*, 948–962.
- (8) Xiong, G.; Wu, Z.; Yi, J.; Fu, L.; Yang, Z.; Hsieh, C.; Yin, M.; Zeng, X.; Wu, C.; Lu, A.; Chen, X.; Hou, T.; Cao, D. ADMETlab 2.0: An Integrated Online Platform for Accurate and Comprehensive Predictions of ADMET Properties. *Nucleic Acids Res.* **2021**, *49*, W5–W14.
- (9) Daina, A.; Michelin, O.; Zoete, V. SwissADME: A Free Web Tool to Evaluate Pharmacokinetics, Drug-Likeness and Medicinal Chemistry Friendliness of Small Molecules. *Sci. Rep.* **2017**, *7*, 42717.
- (10) Lee, P. H.; Cucurull-Sánchez, L.; Lu, J.; Du, Y. J. Development of in Silico Models for Human Liver Microsomal Stability. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 665–673.
- (11) Sakiyama, Y.; Yuki, H.; Moriya, T.; Hattori, K.; Suzuki, M.; Shimada, K.; Honma, T. Predicting Human Liver Microsomal Stability with Machine Learning Techniques. *J. Mol. Graphics Modell.* **2008**, *26*, 907–915.
- (12) Hu, Y.; Unwalla, R.; Denny, R. A.; Bikker, J.; Di, L.; Humblet, C. Development of QSAR Models for Microsomal Stability: Identification of Good and Bad Structural Features for Rat, Human and Mouse Microsomal Stability. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 23–35.
- (13) Rodríguez-Pérez, R.; Trunzer, M.; Schneider, N.; Faller, B.; Gerebtzoff, G. Multispecies Machine Learning Predictions of In Vitro Intrinsic Clearance with Uncertainty Quantification Analyses. *Mol. Pharmaceutics* **2023**, *20*, 383–394.
- (14) Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; Bengio, Y. Graph Attention Networks. *arXiv* **2018**, arXiv:1710.10903v3.
- (15) Xu, K.; Hu, W.; Leskovec, J.; Jegelka, S. How Powerful Are Graph Neural Networks? *arXiv* **2019**, arXiv:1810.00826v3.
- (16) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural Message Passing for Quantum Chemistry. *arXiv* **2017**, arXiv:1704.01212.
- (17) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: A Benchmark for Molecular Machine Learning. *Chem. Sci.* **2018**, *9*, 513–530.
- (18) Pattanaik, L.; Ganea, O.-E.; Coley, I.; Jensen, K. F.; Green, W. H.; Coley, C. W. Message Passing Networks for Molecules with Tetrahedral Chirality. *arXiv* **2020**, arXiv:2012.00094v2.
- (19) Adams, K.; Pattanaik, L.; Coley, C. W. Learning 3D Representations of Molecular Chirality with Invariance to Bond Rotations. *arXiv* **2021**, arXiv:2110.04383v1.
- (20) Gaiński, P.; Koziarski, M.; Tabor, J.; Śmieja, M. ChiENN: Embracing Molecular Chirality with Graph Neural Networks. *arXiv* **2023**, arXiv:2307.02198.
- (21) Chang, C.; Duignan, D. B.; Johnson, K. D.; Lee, P. H.; Cowan, G. S.; Gifford, E. M.; Stankovic, C. J.; Lepsy, C. S.; Stoner, C. L. The Development and Validation of a Computational Model to Predict Rat Liver Microsomal Clearance. *J. Pharm. Sci.* **2009**, *98*, 2857–2867.
- (22) Di Lascio, E.; Gerebtzoff, G.; Rodríguez-Pérez, R. Systematic Evaluation of Local and Global Machine Learning Models for the Prediction of ADME Properties. *Mol. Pharmaceutics* **2023**, *20*, 1758–1767.
- (23) Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Félix, E.; Magariños, M.; Mosquera, J. F.; Mutowo, P.; Nowotka, M.; Gordillo-Marañón, M.; Hunter, F.; Junco, L.; Mugumbate, G.; Rodriguez-Lopez, M.; Atkinson, F.; Bosc, N.; Radoux, C. J.; Segura-Cabrera, A.; Hersey, A.; Leach, A. R. ChEMBL: Towards Direct Deposition of Bioassay Data. *Nucleic Acids Res.* **2019**, *47*, D930–D940.
- (24) Esaki, T.; Watanabe, R.; Kawashima, H.; Ohashi, R.; Natsume-Kitatani, Y.; Nagaoka, C.; Mizuguchi, K. Data Curation Can Improve the Prediction Accuracy of Metabolic Intrinsic Clearance. *Mol. Inf.* **2019**, *38*, 1800086.
- (25) Fang, C.; Wang, Y.; Grater, R.; Kapadnis, S.; Black, C.; Trapa, P.; Scialoba, S. Prospective Validation of Machine Learning Algorithms for Absorption, Distribution, Metabolism, and Excretion Prediction: An Industrial Perspective. *J. Chem. Inf. Model.* **2023**, *63*, 3263–3274.
- (26) Li, L.; Lu, Z.; Liu, G.; Tang, Y.; Li, W. In Silico Prediction of Human and Rat Liver Microsomal Stability via Machine Learning Methods. *Chem. Res. Toxicol.* **2022**, *35*, 1614–1624.
- (27) Fang, C.; Wang, Y.; Grater, R.; Kapadnis, S.; Black, C.; Trapa, P.; Scialoba, S. Prospective Validation of Machine Learning Algorithms for Absorption, Distribution, Metabolism, and Excretion Prediction: An Industrial Perspective. *J. Chem. Inf. Model.* **2023**, *63*, 3263–3274.
- (28) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.
- (29) Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular Graph Convolutions: Moving beyond Fingerprints. *J. Comput.-Aided Mol. Des.* **2016**, *30*, 595–608.
- (30) RDKit: Open-Source Cheminformatics. <https://www.Rdkit.Org/>, accessed 15 Jul 2023.
- (31) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.; Barzilay, R. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59*, 3370–3388.

- (32) Adams, K.; Pattanaik, L.; Coley, C. W. Learning 3D Representations of Molecular Chirality with Invariance to Bond Rotations. *arXiv* **2021**, arXiv:2110.04383.
- (33) Palmer, D. S.; O'Boyle, N. M.; Glen, R. C.; Mitchell, J. B. O. Random Forest Models To Predict Aqueous Solubility. *J. Chem. Inf. Model.* **2007**, *47*, 150–158.
- (34) Guha, R.; Jurs, P. C. Development of Linear, Ensemble, and Nonlinear Models for the Prediction and Interpretation of the Biological Activity of a Set of PDGFR Inhibitors. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2179–2189.
- (35) Schroeter, T.; Schwaighofer, A.; Mika, S.; Ter Laak, A.; Suelzle, D.; Ganzer, U.; Heinrich, N.; Müller, K. R. Machine Learning Models for Lipophilicity and Their Domain of Applicability. *Mol. Pharmaceutics* **2007**, *4*, 524–538.
- (36) Korkmaz, S. Deep Learning-Based Imbalanced Data Classification for Drug Discovery. *J. Chem. Inf. Model.* **2020**, *60*, 4180–4190.
- (37) Wenzel, J.; Matter, H.; Schmidt, F. Predictive Multitask Deep Neural Network Models for ADME-Tox Properties: Learning from Large Data Sets. *J. Chem. Inf. Model.* **2019**, *59*, 1253–1268.
- (38) Chicco, D.; Warrens, M. J.; Jurman, G. The Coefficient of Determination R-Squared Is More Informative than SMAPE, MAE, MAPE, MSE and RMSE in Regression Analysis Evaluation. *PeerJ Comput. Sci.* **2021**, *7*, No. e623.
- (39) Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A Next-Generation Hyperparameter Optimization Framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*; ACM, 2019, pp 2623–2631.
- (40) Kingma, D. P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980v9.
- (41) Sahigara, F.; Mansouri, K.; Ballabio, D.; Mauri, A.; Consonni, V.; Todeschini, R. Comparison of Different Approaches to Define the Applicability Domain of QSAR Models. *Molecules* **2012**, *17*, 4791–4810.
- (42) Shen, M.; LeTiran, A.; Xiao, Y.; Golbraikh, A.; Kohn, H.; Tropsha, A. Quantitative Structure-Activity Relationship Analysis of Functionalized Amino Acid Anticonvulsant Agents Using k Nearest Neighbor and Simulated Annealing PLS Methods. *J. Med. Chem.* **2002**, *45*, 2811–2823.
- (43) Tropsha, A.; Golbraikh, A. Predictive QSAR Modeling Workflow, Model Applicability Domains, and Virtual Screening. *Curr. Pharm. Des.* **2007**, *13*, 3494–3504.
- (44) van der Maaten, L.; Hinton, G. Visualizing Data Using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579.
- (45) Wold, S.; Esbensen, K.; Geladi, P. Principal Component Analysis. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37–52.
- (46) Cihan Sorkun, M.; Mullaj, D.; Koelman, J. M. V. A.; Er, S. ChemPlot, a Python Library for Chemical Space Visualization**. *Chem. Methods* **2022**, *2*, No. e202200005.
- (47) Dobbelaere, M. R.; Plehiers, P. P.; Van De Vijver, R.; Stevens, C. V.; Van Geem, K. M. Machine Learning in Chemical Engineering: Strengths, Weaknesses, Opportunities, and Threats. *Engineering* **2021**, *7*, 1201–1211.
- (48) Chicco, D.; Warrens, M. J.; Jurman, G. The Coefficient of Determination R-Squared Is More Informative than SMAPE, MAE, MAPE, MSE and RMSE in Regression Analysis Evaluation. *PeerJ Comput. Sci.* **2021**, *7*, No. e623.
- (49) He, J.; You, H.; Sandström, E.; Nittinger, E.; Bjerrum, E. J.; Tyrchan, C.; Czechtizky, W.; Engkvist, O. Molecular Optimization by Capturing Chemist's Intuition Using Deep Neural Networks. *J. Cheminform.* **2021**, *13*, 26.

CAS BIOFINDER DISCOVERY PLATFORM™

ELIMINATE DATA SILOS. FIND WHAT YOU NEED, WHEN YOU NEED IT.

A single platform for relevant, high-quality biological and toxicology research

Streamline your R&D

CAS 
A division of the
American Chemical Society