JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

# ChiGNN: Interpretable Algorithm Framework of Molecular Chiral Knowledge-Embedding and Stereosensitive Property Prediction

Jiaxin Yan, Haiyuan Wang, Wensheng Yang,* Xiaonan Ma,* Yajing Sun,* and Wenping Hu

Cite This: J. Chem. Inf. Model. 2025, 65, 3239−3247 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

**ABSTRACT:** Molecular chirality-related tasks have remained a notable challenge in materials machine learning (ML) due to the subtle spatial discrepancy between enantiomers. Designing appropriate steric molecular descriptions and embedding chiral knowledge are of great significance for improving the accuracy and interpretability of ML models. In this work, we propose a state-of-the-art deep learning framework, Chiral Graph Neural Network, which can effectively incorporate chiral physicochemical knowledge via Trinity Graph and stereosensitive Message Aggregation encoding. Combined with the quantile regression technique, the accuracy of the chiral chromatographic retention time prediction model outperformed the existing records. Accounting for the inherent merits of this framework, we have customized the Trinity Mask and Contribution Splitting techniques to enable a multilevel interpretation of the model's decision mechanism at atomic, functional group, and molecular hierarchy levels. This interpretation has both scientific and practical implications for the understanding of chiral chromatographic separation and the selection of chromatographic stationary phases. Moreover, the proposed chiral knowledge embedding and interpretable deep learning framework, together with the stereomolecular representation, chiral knowledge embedding method, and multilevel interpretation technique within it, also provide an extensible template and precedent for future chirality-related or stereosensitive ML tasks.


## 1. INTRODUCTION

Chiral enantiomers refer to a pair of molecules that cannot be superimposed in 3D space through rotation or translation. Despite sharing the same chemical composition and bonding order, they indeed exhibit significant differences in physical, chemical, biological, and pharmacology properties, which is crucial in circularly polarized luminescence, drug design, biological processes, and many other scenarios.[1−5] The separation and purification of chiral enantiomers are prerequisites for harnessing their unique properties. Currently, methods like chiral chromatography, membrane separation, and recrystallization have been developed for enantioseparation.[6−10] Among them, chiral chromatography, owing to its high stability and rapid separation speed, has emerged as the most frequently used technique in enantiomeric purification. In this approach, the selection of appropriate separation conditions and the identification of matched chiral chromatographic stationary phases (CSPs) are of paramount importance.[11,12] Due to the unclear interaction between the stationary phase selectors and chiral analytes, the choice of CSPs still necessitates tedious and repetitive wet experiments, significantly impeding the separation and utilization of chiral molecules.

In recent years, data-driven machine learning (ML) methods have played an increasingly important role in many fields such as chemistry, physics, materials, and medicine, gradually becoming a prominent research paradigm in various fields.[13−17] In material science, a multitude of ML algorithms have been devised for the expeditious prediction of an array of material properties and the identification of high-performance materials.[18−22] To illustrate, Cao et al. constructed a ML model to realize the rapid prediction of the band gap and gas adsorption properties of metal−organic framework materials.[23] Similarly, Cheng et al. employed deep learning models to achieve high-throughput screening of luminescent materials.[24] Currently, the development of ML algorithms tailored to chiral materials has been relatively slow. In existing chirality-related models, the integration of spatial arrangement knowledge is limited, and some models even fail to incorporate chiral features using expensive experimental measurements such as Raman spectra instead of direct coding of material structure. Dai et al. used decision tree (DT), random forest (RF),

Received: December 3, 2024
Revised: March 16, 2025
Accepted: March 17, 2025
Published: March 21, 2025

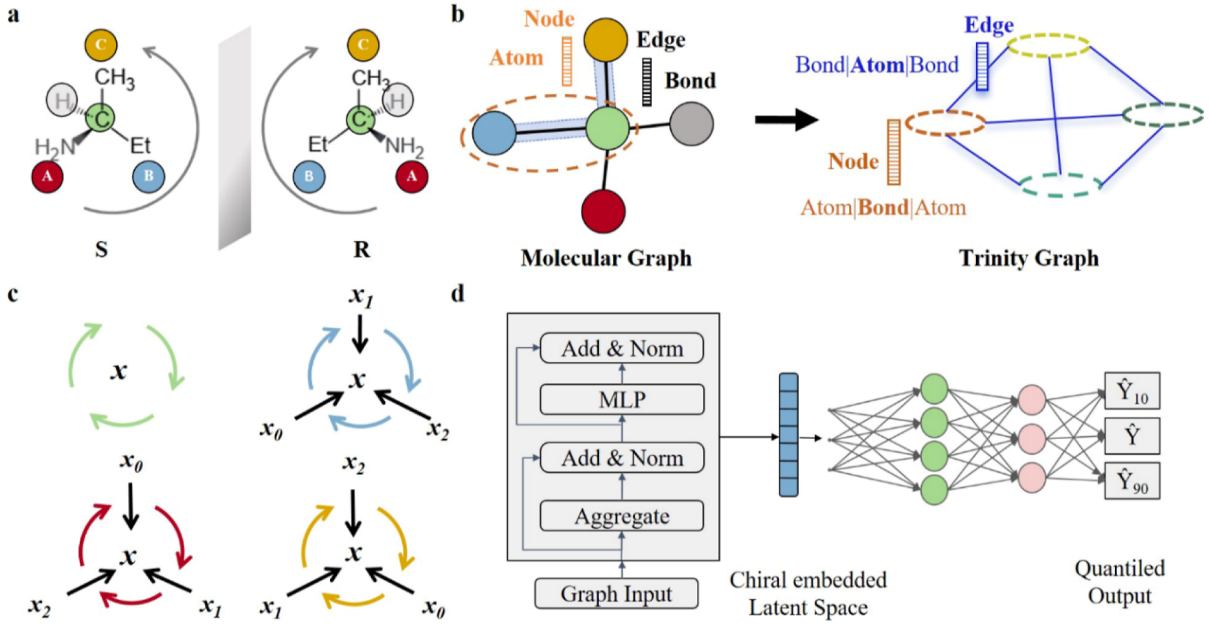


**Figure 1.** (a) Schematic diagram of the spatial structure of chiral enantiomers; (b) construction of the Trinity Graph; (c) stereosensitive message aggregation process; and (d) brief framework of ChiGNN.

support vector machine, linear regression, nearest neighbor algorithm (KNN), and LightGBM algorithms to predict the imbalance factor of circularly polarized luminescent materials, with four macro experimental parameters.[25] Hao et al. used many classical ML models to identify the chirality layered structures from the spectra.[26] Regarding material structures with embedded chiral information, certain models differentiate isomers by directly incorporating chiral labels, such as abstract notations (e.g., "@" or "@@"). Since these labels do not encode the actual physical or chemical basis of chirality, such as molecular geometry, their applicability is significantly limited. Mamede et al. trained a RF model using descriptors with chiral labels, which can be used to predict the sign of optical rotation of chiral compounds.[27] However, this model often fails in enantiomeric molecules with similar chemical structures.[28] In addition, given the crucial role of three-dimensional (3D) structures in identifying spatial isomers, Xu et al. introduced a geometrically enhanced GeoGNN and QGeoGNN algorithm framework. This model demonstrated impressive performance in predicting the retention time (RT) of chiral molecules in chromatography, achieving an $R^2$ of 0.7.[29] Hong et al. developed the 3DMolCSP model to assess the ease of enantiomer separation, achieving commendable results.[30] However, these efforts all hinge on accurate 3D molecular structures, necessitating a substantial amount of training data and computational resources. Gaiński et al. designed a sequence-sensitive message aggregation strategy, which offers a promising algorithm for recognizing spatial heterogeneity.[31] This strategy has proven to exhibit excellent performance on multiple chirality-related tasks, yet its applicability still requires verification.[32] Additionally, regarding model interpretability, given the specificity of chiral tasks, current interpretable models or techniques are incapable of directly explaining atoms with the same chemical environment but differing chiral environments.[33−35] Hence, developing interpretable ML models for chiral knowledge embedding holds significant importance for swiftly predicting chiral sensitivity properties and exploring structure−activity relationships.

In this work, departing from the classical graph structure where atoms are represented as nodes and chemical bonds as edges, we constructed the "Trinity Graph" (TG), which uses "atom|bond|atom" as nodes and "bond|atom|bond" as edges, encapsulating the spatial/stereo arrangement of molecules. Building upon this foundation, we developed a chiral knowledge-embedding deep learning algorithm framework, the Chiral Graph Neural Network (ChiGNN), which employs stereosensitive Message Aggregation (SSMA) to distinguish chiral configurations. This approach has achieved the highest accuracy to date in predicting the RTs of enantiomers across various CSP columns. The $R^2$ of the prediction model on the IC, ADH, IA, and ODH chiral CSP column data sets can reach 0.85, 0.78, 0.77, and 0.83, respectively. Furthermore, leveraging the unique Trinity Graph data structure of ChiGNN, we developed a mask-based node splitting interpretation method, Trinity Masking and Contribution Decomposition (TMCD) technique, realizing the multilevel model interpretation. By unravelling the decision-making mechanism of chromatographic RT within the model, particularly for a pair of chiral enantiomers, we discovered that the model's ultrahigh accuracy stems from its thorough learning of the differences in stereo characteristics between the enantiomers. Through the dimensionality reduction analysis of molecular representations, we also derived a stationary phase screening strategy that aligns with the three-point theory of chiral chromatographic separation. This not only reinforces the model's reliability but also suggests that ChiGNN has the potential to become a powerful tool for selecting CSPs in chiral separation tasks. In summary, ChiGNN surpasses chiral label coding that relies on nonchiral physicochemical knowledge and 3D algorithms with expansive parameter dimensionality. By encoding the chiral characteristics of molecules through enantiomeric chemical definitions, ChiGNN can not only guarantee the thorough integration of chiral knowledge but also reduce the model's parameter burden, thereby providing robust support for the


https://doi.org/10.1021/acs.jcim.4c02259
J. Chem. Inf. Model. 2025, 65, 3239−3247

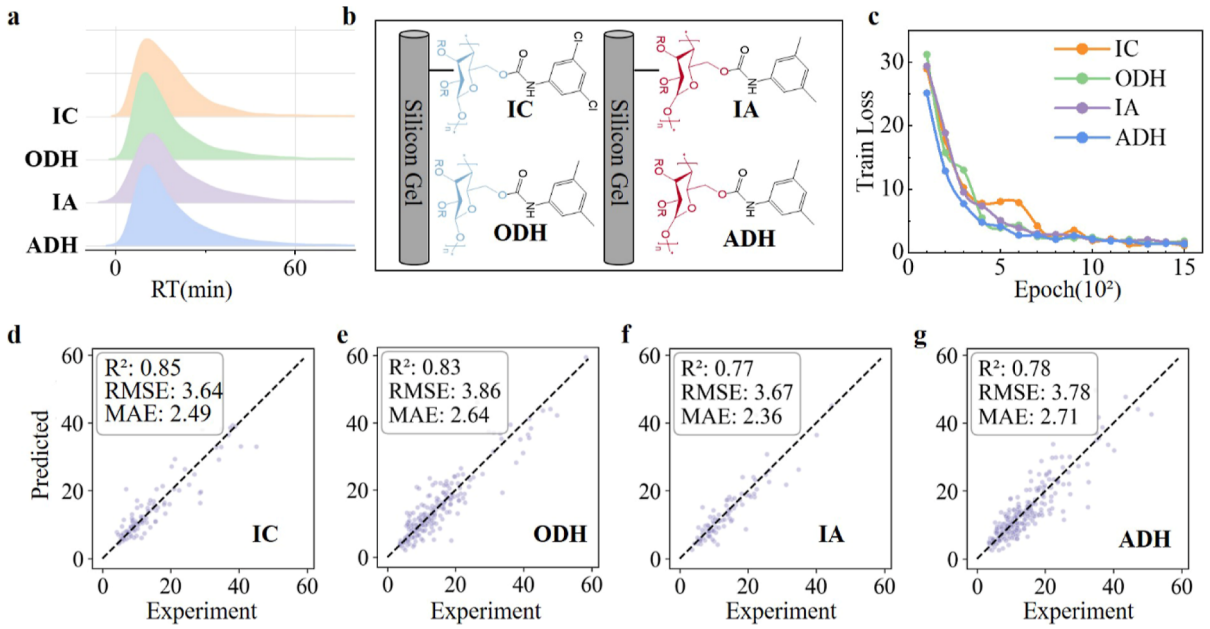


**Figure 2.** (a) Distribution of RT in subdata sets with CSPs of IC, ODH, IA, and ADH; (b) chemical structures of the modifying groups in the four CSPs; (c) learning curve of the model training; and (d−g) performance of the RT predictive model of the four CSPs.

future development and interpretation of chirality-related ML tasks.

## 2. METHODS

**2.1. Construction of ChiGNN.** Given the identical chemical composition and bonding mode of chiral enantiomers, it is of pivotal importance to accurately identify stereo differences in order to enhance the precision of model predictions, as shown in Figure 1a. In molecular graph neural networks, the molecular representation is obtained via an iterative aggregation of the binding environment information (neighbor bonds and atoms) to the nodes (atomic) vector, in a nonsequential manner.[36] Due to the insensitivity of spatial arrangement in message aggregation, classical graph-based algorithms struggle to distinguish between the chiral enantiomers. To address this issue, we have implemented a stereosensitive message aggregation scheme with sequence described by the following formula

$$x' = Wx + \Psi^3(x_0, x_1, x_2) + \Psi^3(x_2, x_0, x_1) + \Psi^3(x_1, x_2, x_0)$$

where $\boldsymbol{x}'$ represents the updated representation of node. $\boldsymbol{x}$ is the original representation and $\boldsymbol{x_0}$, $\boldsymbol{x_1}$, and $\boldsymbol{x_2}$ represent the neighbor nodes. $\boldsymbol{\Psi}$ is the aggregation function. This scheme is based on the physicochemical definition of chirality. The result of message aggregation is unique only if the order of the environmental inputs is consistent. As illustrated in Figure 1b, by sequentially aggregating the message of $x_0{-}x_1{-}x_2{-}x_0$ toward central node $x$, consistent results can be obtained regardless of the starting point. However, should the 3D structure undergo a change, such as the neighbor nodes becoming $x_0{-}x_2{-}x_1{-}x_0$, this will result in a discrepancy. It is thus ensured that the chirality of the encoded molecule is unique. To achieve the aforementioned approach, we chose to eschew the conventional representation of individual elements (atoms and bonds). Instead, we constructed a multielement graph comprising "atoms|bonds|atoms" as nodes and "bonds|atoms|bonds" as edges, as shown in Figure 1c, $\boldsymbol{G}' = (\boldsymbol{X}', \boldsymbol{E}')$, named the Trinity Graph. $\boldsymbol{G}'$ can be represented as

$$X' = \{x_{ij} = x_i|e_{ij}|x_j\colon x_i, x_j \in X, e_{ij} \in E\}$$

$$E' = \{e_{ij,jk} = e_{ij}|x_j|e_{jk}\colon e_{ij}, e_{jk} \in E, x_j \in X\}$$

where $\boldsymbol{X}$ is the set of atoms and $\boldsymbol{E}$ is the set of chemical bonds. The $\boldsymbol{x}_{ij}$ element in the new node set $\boldsymbol{X}'$ is the trinity atom consisting of atom $\boldsymbol{x}_i$, atom $\boldsymbol{x}_j$, and their bonding $e_{ij}$. Similarly, the $\boldsymbol{e}_{ij,jk}$ element in the new edge set $\boldsymbol{E}'$ is the trinity bond consisting of the $\boldsymbol{e}_{ij}$, $\boldsymbol{e}_{jk}$, and their shared atom $\boldsymbol{x}_j$. As shown in Figure 1d, the Trinity Graph representation serves as the input to the model, which sequentially processes it through three stereosensitive message aggregation layers specifically designed for the Trinity Graph, followed by residual connections and normalization operations. To further enhance the accuracy of the predictions, a three-layer quantile regression multilayer perceptron (MLP) is employed. Specifically, the 10th and 90th percentiles are used to capture the lower and upper bounds of the conditional distribution, providing a more robust prediction framework. Unlike traditional regression methods, which focus on the conditional mean, quantile regression estimates the conditional quantiles of the response variable by minimizing a weighted sum of absolute residuals. This enables a more comprehensive analysis of the conditional distribution. By using the 10th and 90th percentiles, the model captures the variability and uncertainty in the predictions, offering a more detailed view of the data distribution. During the training process, the Adam optimizer was used with an initial learning rate of 0.001 and a batch size of 2048, and training was conducted over 1500 epochs. Detailed information on the implementation of the stereosensitive message aggregation scheme within the Trinity Graph, as well as comparisons between the Trinity Graph and conventional graph representations, the quantile regression approach, and model hyperparameters, can be found in Section S1 of Supporting Information.



https://doi.org/10.1021/acs.jcim.4c02259
J. Chem. Inf. Model. 2025, 65, 3239−3247

Table 1. Performances of ChiGNN, Graph-Based Models (Such as QGeoGNN, GCN), and Classic ML Algorithms as Transformer, XGBoost, RF, and MLP[a]

| | IC | | ADH | | IA | | ODH | |
|---|---|---|---|---|---|---|---|---|
| | MAE(mL) | $R^2$ | MAE(mL) | $R^2$ | MAE(mL) | $R^2$ | MAE(mL) | $R^2$ |
| **ChiGNN[a]** | **2.49** | **0.85** | **2.71** | **0.78** | **2.36** | **0.77** | **2.64** | **0.83** |
| ChiGNN[b] | 2.57 | 0.84 | 2.45 | 0.70 | 3.04 | 0.75 | 3.25 | 0.69 |
| QGeoGNN | 2.68 | 0.83 | 2.87 | 0.72 | 2.91 | 0.74 | 2.74 | 0.78 |
| GCN | 5.02 | 0.39 | 4.21 | 0.62 | 4.29 | 0.61 | 3.95 | 0.52 |
| Transformer | 3.70 | 0.67 | 3.94 | 0.54 | 3.68 | 0.67 | 3.44 | 0.71 |
| XGBoost | 3.66 | 0.66 | 3.98 | 0.54 | 3.86 | 0.64 | 3.55 | 0.70 |
| RF | 3.71 | 0.66 | 4.09 | 0.54 | 3.83 | 0.66 | 3.72 | 0.68 |
| MLP | 3.97 | 0.62 | 4.70 | 0.39 | 3.96 | 0.65 | 3.97 | 0.63 |

[a]ChiGNN[a] refers to the ChiGNN model with quantile regression, while ChiGNN[b] refers to the ChiGNN model without quantile regression.

**2.2. Data Set.** The chiral molecular retention time data set comprises a compilation of normal-phase high-performance liquid chromatography experiments from the literature, assembled by Xu et al.[29] The data set encompasses chromatographic RT data for over 25,000 chiral molecules, covering more than 11,720 enantiomers and involving over 20 CSPs. It currently stands as the largest open-source chiral chromatographic data set in terms of the number of enantiomers included. It is important to note that the data set includes duplicate compounds tested under different experimental conditions, resulting in different RT × $v$ values. We describe in Section S1.6 of Supporting Information the results and analysis of model training after data set cleaning and the rationale for ultimately retaining the original data set. When optimizing the model parameters, the data subset of four chiral CSP columns with the largest amount of data in chiral molecular retention time was selected for training and testing. We selected the data-subset with CSPs of ADH, ODH, IC, and IA, which contained the information on RT of 5418, 4972, 2091, and 1849 chiral molecular chromatography data, respectively. The distribution of RTs exhibited a partial Gaussian shape, with the majority concentrated within 60 min, and the chemical structures of the CSPs are shown in Figure 2a,b. Further details can be accessed in Section S2 of Supporting Information. During the model training process, the train, valid, and test sets were allocated in a ratio of 9:0.5:0.5, respectively. Additionally, according to the chromatographic process equation, there is an inverse relationship between RT and flow rate ($v$)

$$\mathrm{RT} = t_0\left(1 + K\frac{V_s}{V_m}\right) \approx \frac{1}{v}(V_m + KV_s)$$

where RT is the retention time, $K$ is the partition coefficient, $v$ is the flow rate, $V_m$ and $V_s$ are the volumes of the mobile phase and stationary phase, respectively, and $t_0$ is the dead time. Therefore, the target value was set as RT × $v$ (min × mL/min), which allows the corresponding RT to be obtained by dividing the predicted value by the set $v$. This approach not only simplifies the relationship between RT and $v$ but also enables the model to generalize across different $v$ conditions without explicitly requiring the flow rate as an input feature during prediction. By training separate models for each column (e.g., ADH, ODH, IC, and IA), we implicitly account for column-specific effects, including the type of CSPs and other column-related parameters such as column length. Given that the source data set includes information like the concentration of iso-propanol, we conducted additional experiments by incorporating it as a descriptor, resulting in the ChiGNN + i-Prop model. The detailed methodology, results, and analysis of these experiments are presented in Section S1.5 of Supporting Information. Although the inclusion of isopropyl alcohol did not lead to significant improvements in model performance with the current data set, this exploration provides valuable insights into the challenges and opportunities of integrating solvent information into predictive models. Future work could focus on optimizing the integration of solvent effects, potentially by utilizing larger data sets or more sophisticated modeling techniques to enhance accuracy and applicability.

**2.3. Comparative Approaches.** QGeoGNN (Quantile Geometry-Enhanced Graph Neural Network) is a graph neural network that incorporates geometric information to predict the RTs of chiral molecules on chromatographic columns.[29] The key feature of this model is the integration of the quantile regression, which outputs predictions for multiple quantiles simultaneously. This enables the model to effectively handle uncertainties and experimental errors within the data. GCN uses a standard graph structure as input, with atoms represented as nodes and bonds as edges. In this approach, there is no explicit encoding of chirality, meaning that the model does not leverage chirality-related information. GCN primarily relies on learning the local connectivity patterns within the molecule. In this study, the Transformer model utilizes SMILES strings with chirality labels, where each atom's chirality is explicitly marked (e.g., "[C@H]" for chiral centers). The model employs an attention mechanism to learn global dependencies between atoms, effectively capturing long-range interactions and chiral information encoded within the SMILES sequence. eXtreme Gradient Boosting (XGBoost) is a gradient-boosted DT algorithm that takes molecular fingerprints with chirality information (such as Morgan fingerprints) as input. The chirality information is encoded in the molecular fingerprint by including features that represent the stereochemistry of the molecule, allowing the model to capture chirality-sensitive properties. RF, similar to XGBoost, uses molecular fingerprints with chirality information as an input. These fingerprints represent a vector of features derived from the molecular structure, which includes chirality-related descriptors. RF builds multiple decision trees and aggregates their predictions, which helps the model handle nonlinear relationships in the data. MLP is a fully connected neural network that also uses molecular fingerprints containing chirality information as an input. This model learns hierarchical representations of molecular features, focusing on capturing complex, nonlinear relationships between molecular structure and target properties. By incorporation


https://doi.org/10.1021/acs.jcim.4c02259
J. Chem. Inf. Model. 2025, 65, 3239−3247

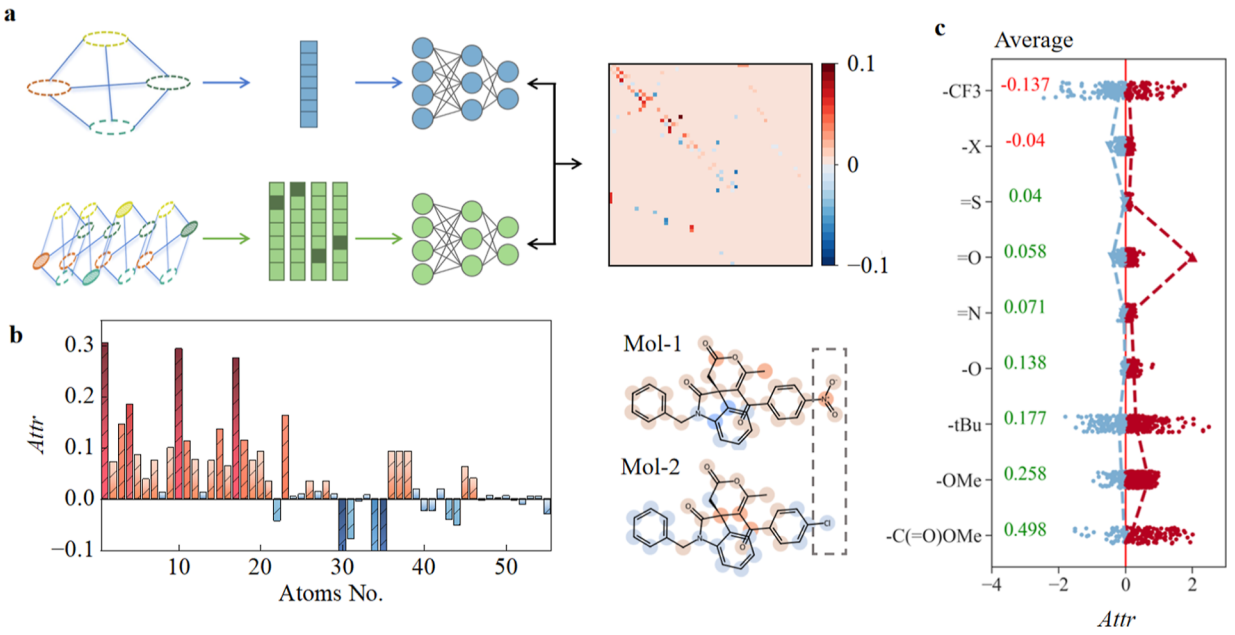


**Figure 3.** (a) Overall of the TMCD process and (b) separated contribution to the chiral RT at the atomic level. The right illustration shows the atomic contribution of Mol-1 and Mol-2. (c) Separated contribution to the RT at the functional group level in IC−CSP data set.

of chirality information in the molecular fingerprints, MLP is better equipped to model chirality-dependent molecular properties.

## 3. RESULTS AND DISCUSSION

**3.1. Performance of ChiGNN.** With extensive training, ChiGNN has achieved the best-ever performance on the four selected CSP data sets. As illustrated in Figure 2c, the model performance stabilizes after 200 epochs of training. On the IC, ADH, IA, and ODH data sets, the mean absolute errors (MAE) for ChiGNN[a] (with quantile regression) are 2.49, 2.71, 2.36, and 2.64 mL, respectively, while the MAE for ChiGNN[b] (without quantile regression) are 2.57, 2.45, 3.04, and 3.25 mL, respectively. These results fully demonstrate ChiGNN's powerful capability in handling chirality-related data and highlight that the introduction of quantile regression significantly improves the model's prediction accuracy and stability. In Table 1, we compare ChiGNN with other common-used models. In contrast, the molecular graph GCN model, lacking any chiral information encoding, performs poorly in all CSP data sets, achieving a maximum $R^2$ of only 0.62 (on the data set of ADH column) and generally exhibiting high MAEs. This strongly suggests that the integration of chiral knowledge is of the utmost significance. In addition, Transformer, Extreme Gradient Boosting (XGBoost), RF, and MLP models, which use SMILES strings with chiral tags or molecular fingerprints containing chiral information (such as Morgan fingerprints) as input, also show suboptimal performance. This indicates that the differentiation of enantiomers by straightforward markers still encounters considerable constraints. Notably, the performance of ChiGNN[a] comprehensively surpasses that of ChiGNN[b] and the previously best-performing QGeoGNN model, which also uses quantile regression and geometric enhancement.[29] Particularly in predicting RT × $\nu$ on IC and ODH chromatographic columns, ChiGNNa achieves $R^2$ values of 0.85 and 0.83, respectively. Quantile regression significantly enhances the model's ability to capture data variability and uncertainty by modeling the lower and upper bounds of the conditional distribution (e.g., the 10th and 90th percentiles), thereby improving prediction robustness. These results not only demonstrate the tremendous potential of ChiGNN in the field of chiral molecular recognition and prediction but also highlight the critical role of quantile regression in enhancing the model performance. By combining a stereosensitive message aggregation scheme with quantile regression, ChiGNN provides a more accurate and reliable tool for chiral molecular research.

**3.2. Interpretability of ChiGNN.** *3.2.1. TMCD Technique.* In the field of interpretability research on graph models, methods based on node masking technology are frequently employed to elucidate the decision-making processes of deep learning models, rendering black-box models transparent. The conventional approach to node masking entails the concealment or alteration of specific nodes within the graph.[33] Subsequently, the state of the masked nodes is updated by the mass-passing process of the neighboring nodes or edges. Ultimately, the impact of each node on the decision can be inferred based on the changes of the refreshed output. Furthermore, shielding a node involves not only masking the atom itself but also obscuring the chemical environment centered on that atom. By traversing all nodes, it is possible to gain insight into the decision-making mechanism at the level of nodes. Nevertheless, the interpretation of chirality-related models presents a more significant challenge. First, it is not possible for these classical molecular graph structures and deep models to distinguish between stereoisomers directly. Second, the assessment of the contribution of chiral nodes typically requires a comprehensive analysis of a pair of enantiomers rather than focusing on a single chiral molecule. To tackle this issue, a novel node masking method, in conjunction with a contribution decomposition technique, was devised for the Trinity Graph. This approach would offer a clearer understanding of the decision-making mechanism underlying the chirality-related prediction model.


https://doi.org/10.1021/acs.jcim.4c02259
J. Chem. Inf. Model. 2025, 65, 3239−3247

TMCD calculates the difference between the outputs of the masked and unmasked ChiGNN models to determine the contribution of each trinity nodes to the RT, expressed as $Attr_{<Trinity>} = (Y_{mol} - Y_{marked})/Y_{mol}$. $Attr_{<Trinity>}$ represents the contribution of the node in Trinity Graph while $Y_{mol}$ and $Y_{marked}$ denote the prediction results with the inputs of the original and the masked graphs, respectively. Taking the (*S*)-1-benzyl-6′-methyl-5′-(4-nitrobenzoyl)spiro[indoline-3,4′-pyran]-2,2′(3′H)-dione (Mol-1) molecule as an example, the explanation process is illustrated in Figure 3a. As trinity nodes are capable of establishing a correspondence between nodes and atoms via the adjacency matrix, the contribution of nodes can be directly analyzed through utilization of a heat map of the aforementioned matrix. To conduct further analysis, the heat map can be further subdivided into the contribution of each atom to the result based on the vector of Trinity nodes. Considering that the information pertaining to hydrogen atoms is already stored on their directly connected atoms during the process of message aggregation and splitting, we can visualize the contribution of the non-hydrogen atoms of Mol-1 in the molecular structure diagram, realizing the atom-level interpretation, in Figure 3b. What is more, it is easily found in the upper right figure, the nitro-substituent part of the molecule made a positive contribution to the chromatographic RT (extending RT). In contrast, if the substituent in the same position is replaced by a chlorine substituent in Mol-2, this part tends to reduce the RT. This is consistent with the general law of chromatography; that is, in normal phase chromatography, if the molecule contains more hydrophilic functional groups, the RT tends to be longer. The effective utilization of chromatographic principles in the decision-making process indicates that the RT prediction ChiGNN model has acquired pertinent chromatographic knowledge through the training process.

TMCD can be further used to quantify the contribution of different functional groups in molecules to realize an elevated level of model interpretation in the functional groups. To explore the contribution of various functional groups to RT, we utilized the data set of IC−CSP as a case study. We conducted a statistical analysis on the contribution of 9 functional groups with the frequency exceeding 100, as shown in Figure 3c. It is evidently apparent that functional groups of greater hydrophilicity typically result in a more pronounced positive contribution to the RT. For instance, the percentage of positive contributions of ═O (side-chain aldehyde or ketone with Log *P* of −0.12), −O (side-chain hydroxyl with Log *P* of −0.56), and −OMe (methoxy with Log *P* of 0.10) are 84%, 94%, and 86%, aligning with the general principles of normal phase chromatography. Hydrophilic functional groups with a smaller Log *P* tend to prolong the RT. Similarly, a positive correlation can be inferred between the average contribution of functional groups and the frequency of hydrophilic functional groups, with an increase in the former corresponding to an increase in the latter. This demonstrates that ChiGNN can acquire fundamental chemical knowledge of chromatography on the one hand. Furthermore, it provides an efficacious tool for the interpretation of RT predictions of complex molecules. By employing this technique, we can gain a more profound comprehension of the model's decision-making process, thereby establishing a foundation for subsequent optimization and application.

*3.2.2. Identification of Chiral Enantiomers with ChiGNN.* By utilizing the TMCD technique, we not only comprehend the decision-making mechanism of the model in predicting the RT of chiral molecules but also elucidate the reasons behind the high prediction accuracy of ChiGNN. Figure 4a presents

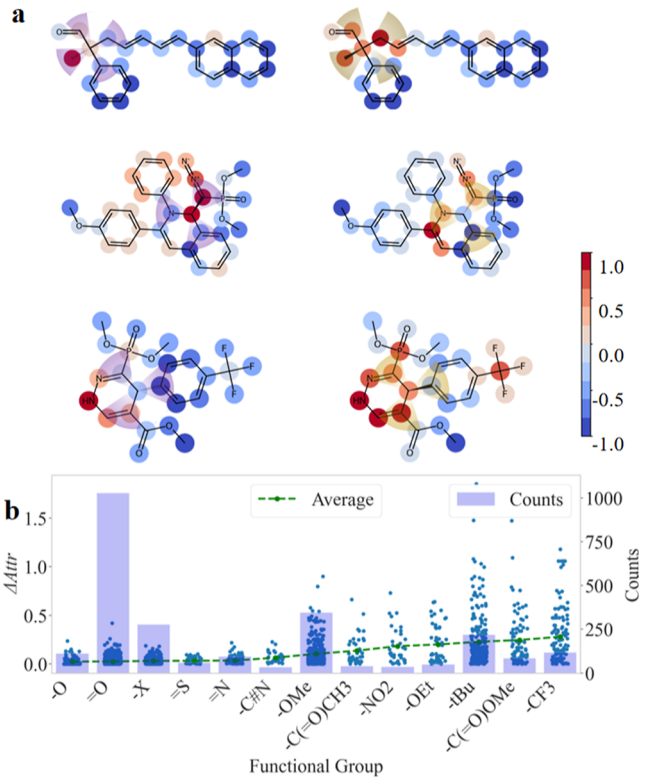


**Figure 4.** (a) Visualization of atomic attribution of three pairs of enantiomers in the test set of IC−CSP and the decision-making disparities primarily raised by the groups linked to chiral atoms and chiral centers. (b) Contribution differences.

three pairs of chiral enantiomers in the test set of IC−CSP. The prediction errors for their R/S stereoisomers are 0.0027/0.7373, 0.0158/0.00217, and 0.0277/0.0662, respectively. Through visual analysis, it can be observed that the decision-making disparities between different stereoisomers primarily manifest in the groups linked to chiral atoms and chiral centers. This further confirms that ChiGNN can accurately capture the chirality differences between well-separated enantiomers when predicting the properties of chiral molecules, demonstrating the model's potential and reliability in chiral molecule analysis.

To further investigate the contributions of functional groups to enantiomer separation, we analyzed the ΔAttr of functional groups in enantiomers using the IC column as an example (Figure 4b), where ΔAttr = |Attr(R) − Attr(S)|. We found a strong correlation between the size of functional groups and their ΔAttr. Larger functional groups, such as *tert*-butyl (-*t*Bu), showed higher contributions to enantiomer separation compared with smaller groups, such as carbonyl (═O). Correlation analysis, as shown in Figure S7, indicates a high Pearson correlation coefficient of 0.921 ($p$-value = $8.13 \times 10^{-6}$), suggesting a strong positive correlation between functional group size and ΔAttr. This trend may be attributed to the enhanced steric interactions of larger groups with the chiral selector in the stationary phase. In addition, the structure of the stationary IC column may still play a role. The cellulose backbone substituted with 3,5-dichlorophenylcarbamate introduces electron-withdrawing chlorine atoms, enhancing the activity of the −NH− group and potentially strengthening


https://doi.org/10.1021/acs.jcim.4c02259
J. Chem. Inf. Model. 2025, 65, 3239−3247

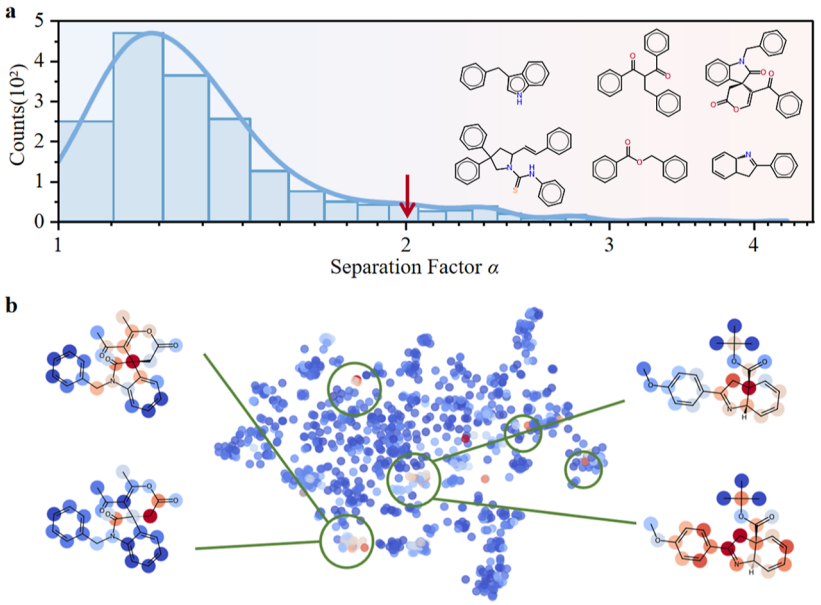


**Figure 5.** (a) Distribution of the chiral separation factor $\alpha$ in the data set of IC−CSP and the high frequency skeleton structure with large chiral separation factor. (b) Dimension reduction analysis of the chiral molecules in the data set of IC−CSP.

specific interactions, such as hydrogen bonding. However, correlation analysis suggests that size-related steric effects are more critical than specific interactions (e.g., hydrogen bonding) in explaining the observed ΔAttr trends. For instance, functional groups such as −OEt, −C(═O)OMe, and −C(═O)CH3 exhibit higher ΔAttr values than −O or ═O due to their larger size. Interpretable ChiGNN can reveal the specific contribution of functional groups in chiral separation processes, providing an insight into the relationship between molecular structure and chromatographic behavior. This finding contributes to the development of more selective and efficacious stationary phases for chiral chromatography.

*3.2.3. Guidance on the Selection of CSPs.* In chromatographic separation, isolating the RT through the CSP column is necessary to achieve efficient separation or purification of a specific analyte. In practice, the chiral separation factor $\boldsymbol{\alpha}$ is used to determine the ability of the CSP column to separate chiral enantiomers. $\boldsymbol{\alpha}$ can be calculated as follows

$$= \frac{t_R - t_0}{t_S - t_0}$$

$t_R$ and $t_S$ are the RTs of the enantiomers and $t_0$ is the dead time of the solvent flow. Likewise, the chiral separation factor is a pivotal metric for gauging the efficacy of enantiomer separation in chromatographic processes. It is preferable to have a larger value. In this study, analysis of the distribution of $\boldsymbol{\alpha}$ can be used to evaluate the separation performance of the relevant CSP column. Taking an IC chromatographic column as an example, Figure 5a shows the distribution of separation factor in the data set. Generally, chiral enantiomers are considered to be of significant separation if their separation factor is greater than 2. The factor between 1.2 and 2 is deemed to represent an excellent degree of separation. Conversely, the factor between 1 and 1.2 would be considered indicative of no separation or poor separation. Subsequently, we conducted t-SNE dimensionality reduction analysis on the latent representation of ChiGNN of chiral enantiomers in the data set. In principle, molecules with similar latent representation tend to be consistent in the chemical space after dimensionality reduction. The average distance between a pair of enantiomers in the data set is only 0.0475, which is far lower than the average distance of 7.67 in the whole data set. With dimension reduction analysis, it is feasible to categorize the molecular structures with high resolution on an IC column. As shown in the red area in Figure 5b, we circled out the regions with high separation factors and selected two representative enantiomers for visual analysis. Through the visualization of contribution, it can be observed that in these enantiomers with high separation factor, the chiral center or the four groups directly connected to it have the extremely opposite effect on the RT. This phenomenon is consistent with the classical three-point interaction theory of chiral chromatography; that is, in chiral recognition, there must be three simultaneous interactions between the CSP and at least one of the enantiomers, with at least one of these being chiral-dependent.

In addition, we counted the molecular skeleton structure of the enantiomer molecules in the higher separation factor region ($\alpha > 2$). Figure 5a shows the most common skeleton structure. The results show that the IC column has higher chiral selectivity for the molecules with these above skeleton structures, which is related to the skeleton structure and the interaction sites on the chiral column. This discovery provides an important clue for the selection of proper CSPs in chiral enantiomer chromatographic separation and helps to improve the efficiency of chromatographic separation. Specifically, different CSPs exhibit selectivity toward particular molecular skeleton, while the latent space of ChiGNN demonstrates the capability to cluster structurally similar molecules together. As depicted in Figure S9, by selecting a pair of enantiomers not included in the IC data set and utilizing ChiGNN to predict their RTs across various chromatographic columns, it is possible to swiftly identify columns that are more conducive to achieving effective separation. The prediction results distinctly recommend the use of the IC column, likely due to the frequent occurrence of the enantiomers' framework within the high-separation-factor frameworks of the IC data set. This predictive capability enables researchers to optimize the



https://doi.org/10.1021/acs.jcim.4c02259
J. Chem. Inf. Model. 2025, 65, 3239−3247

selection of CSPs, thereby enhancing the overall efficiency and success rate of chiral chromatographic separations.

## 4. CONCLUSION

In the prediction of chirality-related properties, embedding chiral knowledge to improve model accuracy and interpretability is of great importance in the aspects of scientific research and application. In this work, we first proposed the Trinity Graph to alternate the classical atom-bond representation of molecules. Upon this fundamental, the ChiGNN algorithmic framework is designed by embedding the chiral physicochemical knowledge with the stereosensitive message aggregation approach. ChiGNN can be used in training high-precision chiral chromatographic RT predictive models. ChiGNN can effectively identify the differences in the spatial arrangement of chiral centers, successfully overcoming the limitations of traditional graphical models in distinguishing chiral enantiomers from stereosensitive ML tasks, which can efficiently improve the prediction accuracy of chiral molecular properties.

The ChiGNN-based deep learning model has achieved hitherto superior results in the prediction of chiral chromatographic RTs for a wide range of stationary chromatographic phase columns. The predictive model has reached $R^2$ of 0.85, 0.78, 0.77, and 0.83 on four data sets of CSP with IC, ADH, IA, and ODH, respectively. This direct embedding of chiral stereo knowledge not only significantly improves the accuracy of the stereosensitive task but also makes the model quite interpretable. Combined with our proposed TMCD technique, the model can achieve a multilevel interpretation approach at the atomic level, functional group level, and molecular feature level. At the atomic level, the differences in RTs in the well-separated chiral enantiomers are predominantly found in the chiral center or atoms directly connected to the chiral center, confirming that chiral chromatographic splitting relies mainly on the differences in stereoisomerism therein, consistent with the classical three-point rules. At the functional group level, the model learns the basics of chromatography, e.g., the model correctly identifies that hydrophilic functional groups in normal-phase chromatography can generally improve chromatographic RTs. In addition, this hierarchy allows us to derive the contribution of functional groups to chiral chromatography splitting. At the molecular feature level, via unsupervised clustering of chiral molecular latent vectors, we can explore molecular backbones that different CSPs excel at, which provides important clues for chiral CSP column selection and contributes to the improvement of chromatographic separation efficiency. Owing to its multilevel interpretability, the ChiGNN model shows potential for future applied research. This model could offer support for a deeper understanding of chiral mechanisms and provide insights into the design and development of novel chiral molecules and functional materials. Its applications may extend to areas such as chiral drug development, circularly polarized luminescent materials design, and other research involving chiral-related chemical and physical properties.

## ■ ASSOCIATED CONTENT

### Data Availability Statement

The code is open source at https://github.com/YajingSun-Group/ChiGNN. The relevant data sets, model architecture, and generated data can be found in the links.

### Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.4c02259.

Detailed model architecture and training procedures, data set descriptions, functional groups in group-level interpretation, and molecular skeleton in molecular-level interpretation (PDF)

## ■ AUTHOR INFORMATION


### Corresponding Authors

**Wensheng Yang** − *Institute of Molecular Plus, Tianjin University, Tianjin 300072, P. R. China*; Email: wsyang@tju.edu.cn

**Xiaonan Ma** − *Institute of Molecular Plus, Tianjin University, Tianjin 300072, P. R. China*; orcid.org/0000-0002-3591-2451; Email: xiaonanma@tju.edu.cn

**Yajing Sun** − *Key Laboratory of Organic Integrated Circuits, Ministry of Education and Tianjin Key Laboratory of Molecular Optoelectronic Sciences, Department of Chemistry, School of Science, Tianjin University, Tianjin 300072, China; Haihe Lab of ITAI, Tianjin 300051, P. R. China*; orcid.org/0000-0003-2807-9382; Email: syj19@tju.edu.cn

### Authors

**Jiaxin Yan** − *Key Laboratory of Organic Integrated Circuits, Ministry of Education and Tianjin Key Laboratory of Molecular Optoelectronic Sciences, Department of Chemistry, School of Science, Tianjin University, Tianjin 300072, China; Institute of Molecular Plus, Tianjin University, Tianjin 300072, P. R. China; Haihe Lab of ITAI, Tianjin 300051, P. R. China*

**Haiyuan Wang** − *Key Laboratory of Organic Integrated Circuits, Ministry of Education and Tianjin Key Laboratory of Molecular Optoelectronic Sciences, Department of Chemistry, School of Science, Tianjin University, Tianjin 300072, China*

**Wenping Hu** − *Key Laboratory of Organic Integrated Circuits, Ministry of Education and Tianjin Key Laboratory of Molecular Optoelectronic Sciences, Department of Chemistry, School of Science, Tianjin University, Tianjin 300072, China*; orcid.org/0000-0001-5686-2740

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.4c02259


### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

This work was financially supported by the National Natural Science Foundation of China (22473085, 22003046, and 52121002), "A Multi-Scale and High-Efficiency Computing Platform for Advanced Functional Materials" program, funded by Haihe Laboratory in Tianjin (grants no. 22HHXCJC00007), the National Key R&D Program (grant no. 2022YFA1204401), and Xiaomi Young Talents Program. The authors gratefully acknowledge the National Supercomputer Center in Tianjin (Tianhe 3F) and the Scientific Computing Center of CIC, Tianjin University for providing computation facilities.

## REFERENCES


(1) Furlan, F.; Moreno-Naranjo, J. M.; Gasparini, N.; Feldmann, S.; Wade, J.; Fuchter, M. J. Chiral Materials and Mechanisms for Circularly Polarized Light-Emitting Diodes. *Nat. Photonics* **2024**, *18* (7), 658−668.

(2) Yang, H.; Yu, H.; Stolarzewicz, I. A.; Tang, W. Enantioselective Transformations in the Synthesis of Therapeutic Agents. *Chem. Rev.* **2023**, *123* (15), 9397−9446.

(3) Wang, L.; Urbas, A. M.; Li, Q. Nature-Inspired Emerging Chiral Liquid Crystal Nanostructures: From Molecular Self-Assembly to DNA Mesophase and Nanocolloids. *Adv. Mater.* **2020**, *32* (41), 1801335.

(4) Ma, Y.; Oleynikov, P.; Terasaki, O. Electron Crystallography for Determining the Handedness of a Chiral Zeolite Nanocrystal. *Nat. Mater.* **2017**, *16* (7), 755−759.

(5) Sang, Y.; Han, J.; Zhao, T.; Duan, P.; Liu, M. Circularly Polarized Luminescence in Nanoassemblies: Generation, Amplification, and Application. *Adv. Mater.* **2020**, *32* (41), 1900110.

(6) Peluso, P.; Chankvetadze, B. Recognition in the Domain of Molecular Chirality: From Noncovalent Interactions to Separation of Enantiomers. *Chem. Rev.* **2022**, *122* (16), 13235−13400.

(7) De Gauquier, P.; Vanommeslaeghe, K.; Heyden, Y. V.; Mangelings, D. Modelling Approaches for Chiral Chromatography on Polysaccharide-Based and Macrocyclic Antibiotic Chiral Selectors: A Review. *Anal. Chim. Acta* **2022**, *1198*, 338861.

(8) Jiang, H.; Yang, K.; Zhao, X.; Zhang, W.; Liu, Y.; Jiang, J.; Cui, Y. Highly Stable Zr(IV)-Based Metal−Organic Frameworks for Chiral Separation in Reversed-Phase Liquid Chromatography. *J. Am. Chem. Soc.* **2021**, *143* (1), 390−398.

(9) Lorenz, H.; Seidel-Morgenstern, A. Processes To Separate Enantiomers. *Angew. Chem. Int. Ed* **2014**, *53* (5), 1218−1250.

(10) Yu, C.; Yin, B. H.; Wang, Y.; Luo, S.; Wang, X. Advances in Membrane-Based Chiral Separation. *Coord. Chem. Rev.* **2023**, *495*, 215392.

(11) Scriba, G. K. E. Chiral Recognition in Separation Science − an Update. *Journal of Chromatography A* **2016**, *1467*, 56−78.

(12) Kim, B.-H.; Lee, S. U.; Moon, D. C. Chiral Recognition of N-Phthaloyl, N-Tetrachlorophthaloyl, and N-Naphthaloyl $\alpha$-Amino Acids and Their Esters on Polysaccharide-Derived Chiral Stationary Phases. *Chirality* **2012**, *24* (12), 1037−1046.

(13) Meuwly, M. Machine Learning for Chemical Reactions. *Chem. Rev.* **2021**, *121* (16), 10218−10239.

(14) Sajjan, M.; Li, J.; Selvarajan, R.; Sureshbabu, S. H.; Kale, S. S.; Gupta, R.; Singh, V.; Kais, S. Quantum Machine Learning for Chemistry and Physics. *Chem. Soc. Rev.* **2022**, *51* (15), 6475−6573.

(15) Kang, Y.; Park, H.; Smit, B.; Kim, J. A Multi-Modal Pre-Training Transformer for Universal Transfer Learning in Metal−Organic Frameworks. *Nat. Mach Intell* **2023**, *5* (3), 309−318.

(16) Zhou, H.-Y.; Yu, Y.; Wang, C.; Zhang, S.; Gao, Y.; Pan, J.; Shao, J.; Lu, G.; Zhang, K.; Li, W. A Transformer-Based Representation-Learning Model with Unified Processing of Multimodal Input for Clinical Diagnostics. *Nat. Biomed. Eng.* **2023**, *7* (6), 743−755.

(17) Sanchez-Fernandez, A.; Rumetshofer, E.; Hochreiter, S.; Klambauer, G. CLOOME: Contrastive Learning Unlocks Bioimaging Databases for Queries with Chemical Structures. *Nat. Commun.* **2023**, *14* (1), 7339.

(18) Zhang, Q.; Hu, Y.; Yan, J.; Zhang, H.; Xie, X.; Zhu, J.; Li, H.; Niu, X.; Li, L.; Sun, Y.; Hu, W. Large-Language-Model-Based AI Agent for Organic Semiconductor Device Research. *Adv. Mater.* **2024**, *36* (32), 2405163.

(19) Kuenneth, C.; Ramprasad, R. polyBERT: A Chemical Language Model to Enable Fully Machine-Driven Ultrafast Polymer Informatics. *Nat. Commun.* **2023**, *14* (1), 4099.

(20) Xie, T.; France-Lanord, A.; Wang, Y.; Lopez, J.; Stolberg, M. A.; Hill, M.; Leverick, G. M.; Gomez-Bombarelli, R.; Johnson, J. A.; Shao-Horn, Y.; Grossman, J. C. Accelerating Amorphous Polymer Electrolyte Screening by Learning to Reduce Errors in Molecular Dynamics Simulated Properties. *Nat. Commun.* **2022**, *13* (1), 3415.

(21) Lin, T.-S.; Coley, C. W.; Mochigase, H.; Beech, H. K.; Wang, W.; Wang, Z.; Woods, E.; Craig, S. L.; Johnson, J. A.; Kalow, J. A.; Jensen, K. F.; Olsen, B. D. BigSMILES: A Structurally-Based Line Notation for Describing Macromolecules. *ACS Cent. Sci.* **2019**, *5* (9), 1523−1531.

(22) Miyake, Y.; Saeki, A. Machine Learning-Assisted Development of Organic Solar Cell Materials: Issues, Analyses, and Outlooks. *J. Phys. Chem. Lett.* **2021**, *12* (51), 12391−12401.

(23) Cao, Z.; Magar, R.; Wang, Y.; Barati Farimani, A. MOFormer: Self-Supervised Transformer Model for Metal−Organic Framework Property Prediction. *J. Am. Chem. Soc.* **2023**, *145* (5), 2958−2967.

(24) Cheng, Z.; Liu, J.; Jiang, T.; Chen, M.; Dai, F.; Gao, Z.; Ke, G.; Zhao, Z.; Ou, Q. Automatic Screen-out of Ir(III) Complex Emitters by Combined Machine Learning and Computational Analysis. *Adv. Opt. Mater.* **2023**, *11*, 2301093.

(25) Dai, Y.; Zhang, Z.; Wang, D.; Li, T.; Ren, Y.; Chen, J.; Feng, L. Machine-Learning-Driven G-Quartet-Based Circularly Polarized Luminescence Materials. *Adv. Mater.* **2024**, *36* (4), 2310455.

(26) Hao, H.; Li, K.; Ji, X.; Zhao, X.; Tong, L.; Zhang, J. Chiral Stacking Identification of Two-Dimensional Triclinic Crystals Enabled by Machine Learning. *ACS Nano* **2024**, *18* (21), 13858−13865.

(27) Mamede, R.; de-Almeida, B. S.; Chen, M.; Zhang, Q.; Aires-de-Sousa, J. Machine Learning Classification of One-Chiral-Center Organic Molecules According to Optical Rotation. *J. Chem. Inf. Model.* **2021**, *61* (1), 67−75.

(28) Adams, K.; Pattanaik, L.; Coley, C. W. Learning 3D Representations of Molecular Chirality with Invariance to Bond Rotations. *arXiv* **2021**, arXiv:2110.04383.

(29) Xu, H.; Lin, J.; Zhang, D.; Mo, F. Retention Time Prediction for Chromatographic Enantioseparation by Quantile Geometry-Enhanced Graph Neural Network. *Nat. Commun.* **2023**, *14* (1), 3095.

(30) Hong, Y.; Welch, C. J.; Piras, P.; Tang, H. Enhanced Structure-Based Prediction of Chiral Stationary Phases for Chromatographic Enantioseparation from 3D Molecular Conformations. *Anal. Chem.* **2024**, *96* (6), 2351−2359.

(31) Gaiński, P.; Koziarski, M.; Tabor, J.; Śmieja, M. ChiENN: Embracing Molecular Chirality with Graph Neural Networks. In *Machine Learning and Knowledge Discovery in Databases: Research Track*; Koutra, D., Plant, C., Gomez Rodriguez, M., Baralis, E., Bonchi, F., Eds.; Springer Nature Switzerland: Cham, 2023; pp 36−52.

(32) Pu, C.; Gu, L.; Hu, Y.; Han, W.; Xu, X.; Liu, H.; Chen, Y.; Zhang, Y. Prediction of Human Liver Microsome Clearance with Chirality-Focused Graph Neural Networks. *J. Chem. Inf. Model.* **2024**, *64* (14), 5427−5438.

(33) Wu, Z.; Wang, J.; Du, H.; Jiang, D.; Kang, Y.; Li, D.; Pan, P.; Deng, Y.; Cao, D.; Hsieh, C.-Y.; Hou, T. Chemistry-Intuitive Explanation of Graph Neural Networks for Molecular Property Prediction with Substructure Masking. *Nat. Commun.* **2023**, *14* (1), 2585.

(34) Xiang, Y.; Tang, Y.-H.; Lin, G.; Reker, D. Interpretable Molecular Property Predictions Using Marginalized Graph Kernels. *J. Chem. Inf. Model.* **2023**, *63* (15), 4633−4640.

(35) Yoshikai, Y.; Mizuno, T.; Nemoto, S.; Kusuhara, H. Difficulty in Chirality Recognition for Transformer Architectures Learning Chemical Structures from String Representations. *Nat. Commun.* **2024**, *15* (1), 1197.

(36) Xu, K.; Hu, W.; Leskovec, J.; Jegelka, S. How Powerful Are Graph Neural Networks? *arXiv* **2019**, arXiv:1810.00826.




https://doi.org/10.1021/acs.jcim.4c02259
J. Chem. Inf. Model. 2025, 65, 3239−3247