Enhanced Structure-Based Prediction of Chiral Stationary Phases for Chromatographic Enantioseparation from 3D Molecular Conformations

Yuhui Hong,[†] Christopher J. Welch,[‡] Patrick Piras,[¶] and Haixu Tang[*,†]

†*Luddy School of Informatics, Computing, and Engineering, Indiana University Bloomington, Bloomington, IN 47408, USA*

‡*Indiana Consortium for Analytical Science & Engineering (ICASE), Indianapolis, IN 46202, USA*

¶*Aix Marseille Université, CNRS, Centrale Marseille, FSCM, Chiropole, Marseille, France*

E-mail: hatang@indiana.edu

Abstract

The accurate prediction of suitable chiral stationary phases (CSPs) for resolving the enantiomers of a given compound poses a significant challenge in chiral chromatography. Previous attempts at developing machine learning models for structure-based CSP prediction have primarily relied on 1D SMILES strings[1] or 2D graphical representations of molecular structures, and have met with only limited success. In this study, we apply the recently developed 3D molecular conformation representation learning algorithm, which uses rapid conformational analysis and point clouds of atom positions in

[1]The simplified molecular-input line-entry system (SMILES) is a specification in the form of a line notation for describing the structure of chemical species using short ASCII strings.

3D space, enabling efficient chemical structure-based machine learning. By harnessing the power of the rapid 3D molecular representation learning and a dataset comprising over 300,000 chromatographic enantioseparation records sourced from the literature, our models afford notable improvements for the chemical structure-based choice of appropriate CSP for enantioseparation, paving the way for more efficient and informed decision-making in the field of chiral chromatography.

Introduction

Chiral Chromatography has been the most widely used technique for measuring enantiopurity for nearly four decades[1], but the issue of which of the dozens of available chiral stationary phases (CSPs) will be best suited for resolving the enantiomers of a particular compound remains difficult to predict. Presently, the preferred technique for developing chromatographic methods involves the rote screening of multiple chiral columns using high-performance liquid chromatography (HPLC) or supercritical fluid chromatography (SFC) instruments, often equipped with column switching devices. This empirical approach, though widely adopted, can be time-consuming and resource-intensive, particularly when dealing with newly synthesized compounds[2,3] with limited prior characterization data. Machine learning (ML) techniques offer new possibilities for advancing research in chiral chemistry and improving our understanding of chirality in various fields. Many examples demonstrate a variety of applications including chiral chromatography[4–6], asymmetric catalysis[7–9], molecule detection[10], or optical rotation prediction[11].

The availability of Chirbase, a massive database comprised of more than 300,000 individual chiral chromatographic records extracted from the literature by Roussel and coworkers[12,13], has encouraged preliminary attempts at structure-based ML for CSP predictions. Progress to date has been somewhat modest[12], with the scarcity of negative data in literature reports being identified as an impediment to further advancements[14]. Rebalancing the data set to compensate for missing negative data has been somewhat helpful[13], but model

performance still lacks the level of certainty needed for routine use in chiral chromatographic method development.

Chemical structure-based machine learning models typically use traditional one-dimensional (SMILES)[15] or two-dimensional chemical graph[16,17] molecular representations, from which a collection of molecular 'features' are extracted that together describe certain characteristics of the molecule of interest (for example, Molecule A possesses a total of 11 carbon atoms, 12 hydrogen atoms, 2 nitrogen atoms, and 2 oxygen atoms, 5 double bond equivalents, a carbonyl oxygen that is 3 bonds removed from a nitrogen atom, an indole ring system, etc.). Models are then trained on the association of such features with some measurable properties such as solubility, NMR peak shift, etc. While progress in this area has been remarkable in recent years, the use of molecular descriptors derived from 1D or 2D molecular representations may not fully capture the subtle differences in shape and conformation that frequently play a crucial role in molecular properties associated with catalysis, reactivity, and molecular recognition.

We have recently developed the deep neural network (DNN) model, 3DMolMS[18], which represents the chemical structures of the input compound as 'point clouds' in 3D, where each component atom is represented by x, y, z-coordinates and an atom identifier (e.g. carbon atom, oxygen atom, etc.). With this approach, no specific information about bond connectivity is recorded, although this information can of course be deduced from interatomic distances. 3DMolMS encompasses all the advantages realized by three-dimensional molecular representation learning. Notably, it incorporates SE(3) invariance (pertaining to the Euclidean group involving 3D displacement motions like translations and rotations)[19–22], high efficiency, and geometric completeness[23]. The 3DMolMS algorithm applies a rapid calculation of a single lowest energy conformer for each molecule in the training set, which is then converted to a point cloud. The model was initially developed for structure-based prediction of MS fragmentation in tandem mass spectrometry, where it shows improved performance relative to previously developed ML models. Subsequently, transfer learning

employing the 3DMolMS model has afforded improved ML predictive models for disparate properties of compounds such as their retention time (RT) in liquid chromatography (LC) and collisional cross sections (CCS) in ion mobility spectrometry (IMS)[18].

Based on these results, we reasoned that the 3DMolMS model could be useful for improved structure-based enantioselectivity prediction for specific CSPs. Accordingly, we have revisited previously developed ML models, evaluating the prospects for improving performance using the 3DMolMS approach combined with transfer learning. Specifically, we developed 3DMolCSP, which extends the DNN model of 3DMolMS, for enantioselectivity prediction. We trained the model using the previously prepared CSP data set in ChirBase[12,13], and evaluated the model using cross-validation and on an independent CSP dataset CMRT[5]. The results showed 3DMolCSP outperforms the previous ML models for enantioselectivity prediction, while the transfer learning based on the pretrained model on spectra prediction can further improve the model's performance. These results suggest that 3DMolCSP is ready to be used to assist in the selection of appropriate CSP for enantioseparation in the field of chiral chromatography.

Methodology

Data Prepossessing

To train the 3DMolCSP model for enantioselectivity prediction, we exploited the experimental data curated in ChirBase, which contains the measurement of 42,361 unique enantiomer pairs. Similar to the previous studies, we selected 18 different CSPs among 1603 chiral columns, on which a sufficient number of compounds were experimentally tested. In addition to cross-validation on this training dataset, we evaluated the model on an independent testing set, including 6 CSPs collected in CMRT[5]. The number of compounds in the training and testing sets are summarized in Table 1. For the transfer learning approach, we adapted the pretrained model of 3DMolMS for tandem mass (MS/MS) spectra prediction, which

Table 1: Number of compounds with the experimental data of 18 different CSPs available in ChirBase and CMRT. The enantiomers are counted as one compound.

CSP	No. of Compounds				
	All		After preprocessing		
	ChirBase	CMRT	ChirBase	CMRT	Overlap
Chiralcel OD (Lux Cellulose-1)	14395	178	13746	171	8
Chiralpak AD	11194	292	10906	269	14
Chiralcel OJ (Lux Cellulose-3)	4261	111	4170	102	5
Chiralpak AS	3666	156	3605	151	13
Whelk-O	1773	0	1691	0	0
Chiralpak IA	1380	805	1345	727	25
Pirkle (R or S)-DNBPG	1338	0	1334	0	0
Chiralcel OB	1276	0	1257	0	0
Chirobiotic T	1155	0	1155	0	0
Chiralpak IC (Lux i-Cellulose-5)	1035	931	1024	893	22
Chiralpak IB	680	300	679	285	0
Cyclobond I	642	0	639	0	0
Chiral-AGP	574	0	575	0	0
Cyclobond I RN	533	0	553	0	0
Chirobiotic R	462	0	460	0	0
Chirobiotic V	351	0	351	0	0
Chirobiotic TAG	308	0	308	0	0
Ultron-ES-OVM	189	0	189	0	0

was previously developed by us using the training data collected in the spectral libraries of Agilent DCPL and NIST20[24].

We followed the same data preparation procedure as used in the previous studies[13,18]. For each data set record sharing the same chiral chemical structure, we selected the optimal condition recorded for each CSP, ensuring that each enantioselectivity value was unique. We retained the compounds composed solely of the most common atoms (i.e., C, H, O, N, F, S, Cl, P, B, I and Br). The lowest-energy three-dimensional conformations are generated from their respective SMILES strings using the ETKDG[25] algorithm implemented in the RDKit library [2].

Discretization of Enantioselectivity Values

The CSP enantioselectivity values, denoted as α, serve as the critical indicators of the separation efficacy achieved by a CSP in chiral chromatography using high-performance liq-

[2]Open-source cheminformatics: `https://www.rdkit.org/`

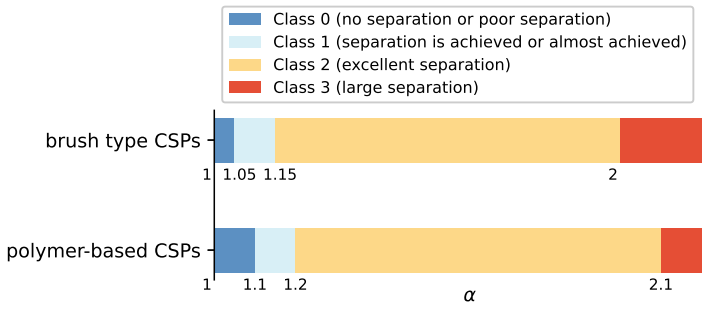


Figure 1: The four classes (denoted as Class 0, 1, 2 and 3, respectively) of compounds are defined based on their α values for the brush type or the polymer-based CSPs. Between these two types of CSPs, the fraction of compounds in the four classes are 10-15%, 20-30%, 45-55%, and 10-15%, respectively.

uid chromatography (HPLC) or gas chromatography (GC). These values are fundamental to assessing the selectivity and discriminatory power of each CSP in resolving enantiomers. Specifically, α, the ratio of retention factors (k_1 and k_2) on the chiral stationary phase (CSP), quantifies the separation between enantiomers, with higher values indicating greater separation. Retention factors, denoted by k, are standardized parameters defined as $(t_R - t_M)/t_M$, where t_R denotes the time the analyte spends in the stationary phase, and t_M denotes the retention time for an unretained analyte. Unlike t_R, which depends on specific instrument conditions such as the column length, column inner diameter, and the flow rate, the retention factor allows for easier comparison and communication of results across different chromatographic systems.

As described in the previous study[13], the CSP's dataset consists of two types of CSPs: the brush type CSPs and the polymer-based (e.g. carbohydrate or protein) CSPs, respectively. For each type, the compounds are grouped into four classes based on their enantioselectivity values, as shown in Figure 1.

Imbalanced Data Handling and Data Augmentation

Two pivotal methodologies are employed on the training dataset: resampling, employed to achieve category balance, and data augmentation, utilized to enhance the model's resilience

in both enantiomeric configurations (S and R). To address the issue of imbalanced training data (i.e., the number of compounds in one class is much greater than those in the other classes), we randomly re-sampled the compounds in classes with fewer compounds within each CSP for training purposes.

It is worth noting that the 3DMolCSP model is geometrically complete[23], which allows it to differentiate between two enantiomeric configurations, as proved in section S1. We also conduct the prediction of the elution order of enantiomers, and showed in the results section that our model is capable of distinguishing enantiomers' configuration. Considering this capability, both enantiomer configurations are given as input into 3DMolCSP, in an attempt to enable the model to learn chiral information from both configurations. The other configuration of enantiomers can be easily calculated by inverting z-coordinate of one configuration conformation, which is also known as flipping data augmentation in 'point clouds'-based methods[26]. When evaluating, we average the predicted results from two configurations of each enantiomer as the final predicted enantioselectivity values.

Neural Network Architecture Optimization

The deep neural network of 3DMolCSP for enantioselectivity prediction in CSPs is based on the architecture of 3DMolMS[18], while we enhanced the elemental convolution and optimized the decoder component to efficiently retrieve the chiral information from compounds 3D conformations.

3DMolConv 2.0: In the original elemental convolution, we represent each molecular 3D conformation as a point set, denoted by $X = \{x_1, x_2, ..., x_n\}$, where $x_i \in \mathbb{R}^F$ and F is the input dimension. We employ a message-passing method specifically tailored for molecular structures. The features of central points are dynamically updated across layers by aggregating weighted features from both the central point itself and its k-nearest neighbors. The weights are learnable from the atoms' distances (d) and the direction of bonds (ϕ). However, we acknowledge that even though in theory, this operation is sensitive to enantiomers (as

demonstrated in the section S1), in practice, the model encounters the challenge of effectively learning sufficient chiral information to distinguish enantiomers. To address this challenge, we developed an enhanced elemental convolution (denoted as *3DMolConv 2.0*) in an attempt to more explicitly model the bond directions by concatenating them into neighbors' features[3]:

$$x_i^{l+1} = x_i^l + \sum_{j \in \mathcal{N}(x_i^l)} W^l \left[d(x_i, x_j) \right] \circ \left[x_j^l || \phi(x_i^l, x_j^l) \right] \tag{1}$$

where $\circ$ represents the element-wised multiplication, $||$ represents the concatenation, x_j^l represents one (i.e., x_j) of the k-nearest neighbors of the atom x_i in layer l, and the W^l represents the filter on distances. Here, the distance between two atoms x_i and x_j is computed as $d(x_i, x_j) = ||x_i - x_j||$, and the angle between the point vectors x_i and x_j encodes the information related to either the bond angle or the non-bond angles of the edge $\langle x_i, x_j \rangle$: $\phi(x_i, x_j) = \sum_{k \in \mathcal{N}(x_i)} e_{ij}^\top e_{ik}$, where e_{ij} denotes the vector representation related to the edge e_{ij} between x_i and x_j: $e_{ij} = x_i^\top x_j$.

Decoder: Since our model is specifically designed to predict the enantioselectivity for each CSP, which is a single value, we narrowed down the width of the four decoder layers to 512, 256, 128, and 64, respectively. Furthermore, there is no need to concatenate any metadata into the embedded molecules within the latent space. As a result, the total number of parameters in 3DMolCSP is reduced to $11,025,344$.

Feature Enrichment for Elution Orders Prediction

In the CSP enantioselectivity prediction, the two enantiomers with the same chemical structure are expected to demonstrate the same enantioseparation. We use the x, y, z-coordinates and other eight atomic attributes shown in Table 2 as the input atomic features. However, for the prediction of the enantiomers' elution order, the model is expected to learn sufficient

[3]The initial version of *3DMolConv* can be represented as $x_i^{l+1} = x_i^l + \sum_{j \in \mathcal{N}(x_i^l)} W_1^l \left[d(x_i, x_j) \right] \circ W_2^l \left[\phi(x_i^l, x_j^l) \right] \circ x_j$, which contains two filters, filter on distances named W_1^l and filter on directions named W_2^l.

Table 2: The point set encoding of a compound, in which each atom in the compound is encoded as a vector of 22 dimensions, representing the x, y, z-coordinates and other attributes of the atom. The feature marked by an asterisk is only used for the prediction of enantiomers' elution orders.

Index	Description
0-2	x, y, z coordinates
3-14	one-hot encoding of the atom type
15	number of immediate neighbors who are nonhydrogen atoms
16	valence minus the number of hydrogens
17	atomic mass
18	atomic charge
19	number of implicit hydrogens
20	is aromatic
21	is in a ring
22*	is chiral center

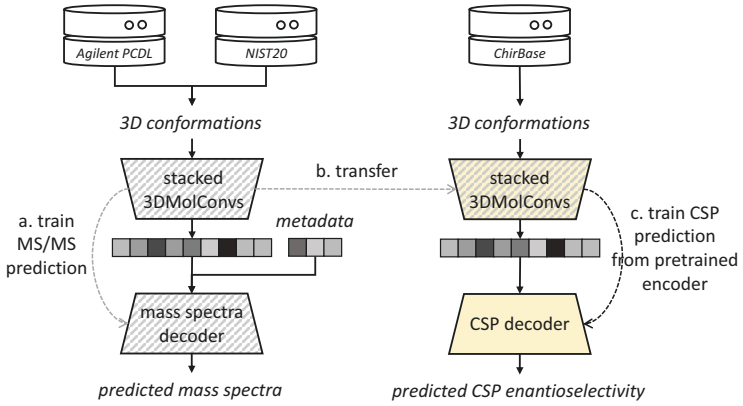


Figure 2: The workflow of building 3DMolCSP-TL model using the transfer learning approach. To build the 3DMolCSP model from scratch (i.e., the independent learning approach), we follow the flow in the right panel only.

distinguishable characteristics of two enantiomers. To guide the model to prioritize chiral atoms in elution order prediction, we enrich the input atomic features by marking whether the atom is a chiral center. It is worth noting that we do not inform the model of the specific chiral tag, i.e. (R) or (S), but let the model learn the chiral types by itself.

Training and Evaluation

The five-fold cross-validation approach is employed to evaluate the DNN model, where each fold retains 20% of the data as the test set, while the remaining 80% is used for training.

Two training approaches are utilized: to initialize the training from scratch (referred to as the *independent learning* approach) and to initialize the training from the weights of the pretrained 3DMolMS model (referred to as the *transfer learning* approach), as illustrated in Figure 2.

Because accuracy could be a biased metric (accuracy paradox)[27] on the extremely imbalanced test dataset, we used three metrics to evaluate the performance of the model predictions: F-measure (F1), Cohen's Kappa (Kappa), and the area under the ROC curve (AUC). As suggested by Piras et al., {Class 2, Class 3} hold the essential information toward modeling the chromatographic enantioseparation. Thus, besides the results of the four-classes classification, we also report the results for the binary classification on two super-classes: {Class 0, Class 1} and {Class 2, Class 3}, respectively.

3DMolCSP is implemented in the PyTorch framework[28]. The model's training process is conducted on a single NVIDIA GeForce RTX 2080 Ti GPU, which takes approximately 175 minutes to train the models for all 18 CSP data sets. The source codes for data preprocessing, training, and validation are made publicly available at `https://github.com/JosieHong/3DMolCSP`.

Results and Discussion

Prediction of CSP Enantioselectivity

We first compared the performances of two types of 3DMolCSP models on four-class CSP enantioselectivity prediction. Specifically, two 3DMolCSP models were built for each CSP using the data sets from ChirBase: one was trained by independent learning (3DMolCSP-SC) and the other was trained by transfer learning (3DMolCSP-TL). We evaluated the models based on five-fold cross-validations: 80% randomly selected compounds were used as the training data while the remaining 20% data were used as the testing data. For each model, we performed five independent cross-validation experiments and reported the average

Table 3: The performance of 3DMolCSP for enantioselectivity prediction. Two types of models are compared, which are trained by independent learning (3DMolCSP-SC) and transfer learning (3DMolCSP-TL), respectively. Specifically, two types of models were built for each CSP and their performances were obtained based on five-fold cross-validations. The performance metrics were then computed on the prediction results of the four desirable classes and averaged on validation samples. The improved performances by the transfer learning are shown in Δ F1 and Δ Kappa.

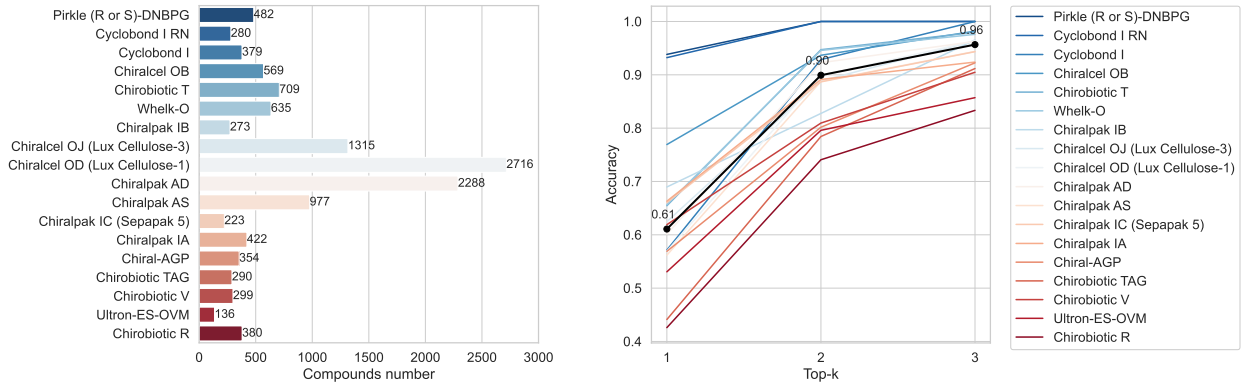
CSP	3DMolCSP-SC		3DMolCSP-TL			
	F1	Kappa	F1	Δ F1	Kappa	Δ Kappa
Chirobiotic R	0.86	0.79	0.90	+0.04	0.85	+0.05
Cyclobond I	0.88	0.67	0.89	+0.01	0.65	−0.02
Cyclobond I RN	0.87	0.79	0.89	+0.02	0.82	+0.03
Chiralpak IB	0.87	0.78	0.88	+0.01	0.78	±0.00
Chiralcel OD (Lux Cellulose-1)	0.90	0.81	0.87	−0.03	0.73	−0.08
Ultron-ES-OVMa	0.78	0.64	0.87	+0.08	0.74	+0.11
Chirobiotic V	0.84	0.77	0.87	+0.03	0.81	+0.04
Chiralpak AS	0.85	0.71	0.87	+0.02	0.70	−0.01
Chiralcel OJ (Lux Cellulose-3)	0.84	0.71	0.87	+0.02	0.73	+0.01
Chirobiotic TAG	0.84	0.78	0.86	+0.02	0.80	+0.03
Pirkle (R or S)-DNBPG	0.81	0.72	0.86	+0.05	0.79	+0.07
Chirobiotic T	0.83	0.75	0.86	+0.03	0.79	+0.04
Chiral-AGP	0.87	0.77	0.85	−0.02	0.72	−0.05
Chiralpak AD	0.88	0.75	0.85	−0.03	0.67	−0.08
Chiralcel OB	0.87	0.75	0.81	−0.06	0.62	−0.12
Chiralpak IC (Sepapak 5)	0.80	0.66	0.80	±0.00	0.62	−0.05
Whelk-O	0.78	0.64	0.79	+0.01	0.64	−0.01
Chiralpak IA	0.80	0.68	0.77	−0.03	0.62	−0.06

Table 4: Comparison of 3DMolCSP-TL and the state-of-the-art ML model for enantioselectivity prediction. The Random Forest (RF) classifier results are extracted from Piras et al.. All of the results are based on five-fold cross-validation, except for the results of the RF classifier on Ultron-ES-OVM where ten-fold cross-validation is used. For comparison purposes, the performance metrics were computed on the binary classification of two superclasses ({Class 0, Class 1} and {Class 2, Class 3}, respectively) and averaged on validation folds. The standard deviations among validation folds are shown in parentheses.

	RF Classifier			3DMolCSP-TL		
CSP	F1	Kappa	AUC	F1	Kappa	AUC
Chirobiotic R	0.80	0.61	0.90	0.95 (±0.02)	0.88 (±0.04)	0.97 (±0.01)
Chirobiotic T	0.85	0.74	0.94	0.93 (±0.01)	0.81 (±0.04)	0.93 (±0.03)
Chirobiotic TAG	0.77	0.52	0.83	0.93 (±0.03)	0.83 (±0.07)	0.96 (±0.01)
Ultron-ES-OVM	0.58	0.34	0.63	0.92 (±0.04)	0.74 (±0.14)	0.92 (±0.06)
Cyclobond I RN	0.82	0.62	0.88	0.92 (±0.04)	0.84 (±0.08)	0.96 (±0.01)
Chiralpak IB	0.72	0.46	0.81	0.92 (±0.01)	0.82 (±0.04)	0.95 (±0.02)
Cyclobond I	0.69	0.38	0.75	0.92 (±0.01)	0.60 (±0.08)	0.75 (±0.03)
Chiral-AGP	0.76	0.42	0.80	0.92 (±0.03)	0.73 (±0.10)	0.88 (±0.02)
Chirobiotic V	0.78	0.51	0.85	0.92 (±0.04)	0.82 (±0.09)	0.98 (±0.02)
Chiralcel OD (Lux Cellulose-1)	0.74	0.48	0.81	0.91 (±0.01)	0.74 (±0.05)	0.85 (±0.04)
Chiralpak AS	0.72	0.43	0.80	0.91 (±0.01)	0.73 (±0.03)	0.84 (±0.02)
Chiralcel OJ (Lux Cellulose-3)	0.73	0.47	0.81	0.91 (±0.02)	0.75 (±0.04)	0.86 (±0.04)
Pirkle (R or S)-DNBPG	0.82	0.68	0.90	0.91 (±0.01)	0.81 (±0.03)	0.95 (±0.01)
Chiralpak AD	0.75	0.50	0.82	0.90 (±0.02)	0.71 (±0.05)	0.84 (±0.03)
Whelk-O	0.82	0.63	0.90	0.89 (±0.03)	0.70 (±0.08)	0.84 (±0.06)
Chiralcel OB	0.74	0.47	0.80	0.87 (±0.02)	0.68 (±0.06)	0.85 (±0.02)
Chiralpak IC (Sepapak 5)	0.74	0.48	0.83	0.86 (±0.01)	0.67 (±0.04)	0.84 (±0.03)
Chiralpak IA	0.78	0.56	0.86	0.85 (±0.02)	0.67 (±0.04)	0.86 (±0.02)

performance in Table 3. It is worth noting that we did not show the results of AUC because AUC is not suitable for evaluating the four-class prediction given that the four classes are not independent (e.g., it is more valuable to predict a compound from Class 0 to be from Class 1 than to predict it to be from Class 3). As shown in Table 3, transfer learning improved the performances of 3DMolCSP on 12 out of 18 CSPs, indicating the pretrained model on spectra prediction can indeed enhance the prediction of enantioselectivity on CSPs.

Next, we compared the 3DMolCSP-TL with the state-of-the-art Random Forest (RF)-based ML model for enantioselectivity prediction as reported previously[13]. To ensure a fair comparison, we evaluated the performance of binary classification in order to be consistent with the approach employed in the previous model[13]. As shown in Table 4, 3DMolCSP-TL outperforms the RF Classifier in terms of F1 and Kappa on all 18 CSPs within the test set.



(a) (b)

Figure 3: Prediction of potentially optimal CSP. The accuracy was evaluated on the compounds resolved by (i.e., falling into Class 2 or Class 3) more than one CSP in ChirBase. (a) The number of compounds whose best enantioseparation is achieved by each CSP. (b) The top-k (k=1, 2, and 3, respectively) accuracy for the potentially optimal CSP prediction. The colored lines represent the accuracy of compounds in each CSP, while the black solid line and dashed line represent the average predicted accuracy and average random guess accuracy across all compounds, respectively.

For AUC, 3DMolCSP-TL performs better than the RF Classifier on 16 CSPs except for the CSP of Whelk-O and Chirobiotic T, on which 3DMolCSP-TL performs only slightly worse than the RF Classifier on Whelk-O and performs equally as RF Classifier on Chirobiotic T.

Assistance in CSP Selections

Given a compound with multiple available CSPs, researchers are interested in two critical tasks: (1) determining whether the compound can be enantioseparated (i.e., classified as Class 2 or Class 3, as defined in Figure 1) by any available chiral column; and (2) identifying the most effective CSP for enantioseparation. Our model could be extended to assist in these two selections. It is worth noting that the definition of CSP enantioselectivity is based on ideal experimental conditions, wherein the effects of the mobile phase are not considered. Consequently, while our model can suggest potentially optimal CSPs, the final selection of a suitable CSP and mobile phase must be made by researchers based on the specific requirements of their experiments.

Addressing the first task, compounds that can be enantioseparated (i.e., classified in Class 2 or Class 3) by at least one CSP are termed *resolvable* compounds. Conversely, compounds that cannot be separated by any CSP are labeled as *unresolvable* compounds. Our model enhances its functionality by predicting whether compounds are resolvable or unresolvable across all available columns, achieving accuracies of 0.95 and 0.72 for resolvable and unresolvable compounds, respectively. The detailed results are presented in Section S2.

Moving to the second task, when a compound has multiple CSP options for enantioseparation, determining the most effective CSP becomes critical. This selection should be based on the CSPs' enantioselectivity, prioritizing the chiral column that achieves enantioseparation and demonstrates the highest enantioselectivity for the specific compound. To aid in this decision-making process, 3DMolCSP provides valuable insights by suggesting the potentially optimal CSP. This is achieved through the analysis of predicted probabilities from Class 2 and Class 3 validations across all available CSPs.

To evaluate our approach in selecting the potentially optimal CSP, we conducted an experiment. The results, depicted in Figure 3b, showcase the top-k (for k=1, 2, and 3, respectively) accuracy of 3DMolCSP in potentially optimal chiral column selection. The top-k accuracy is calculated as the proportion of compounds for which the potentially optimal CSP is predicted within the top-k selected CSPs by 3DMolCSP, as defined by the following equation:

$$\text{top-k accuracy} = \frac{|\text{top-k selected CSPs} \cap \text{potentially optimal CSPs}|}{|\text{potentially optimal CSPs}|} \tag{2}$$

Here, 'top-k selected CSPs' refers to the CSPs predicted to achieve enantioseparation and ranked within the top k positions for enantioselectivity.

For all compounds, the top-1, top-2 and top-3 accuracies are 0.61, 0.90, and 0.96, respectively. As shown in Figure 3b, 3DMolCSP achieved satisfactory accuracies (e.g., top-1 accuracy > 0.5[4]) on the compounds that are optimally enantioseparable by most CSPs.

[4]Note that the top-1 accuracy of the random guess of the potentially optimal CSP would be 0.41, given

However, for two CSPs, namely Chirobiotic R and Chirobiotic TAG, the top-1 accuracy falls below 0.5. These relatively low accuracies can be attributed to the smaller size of the data set available for these particular CSPs as corroborated by Table 1 and Figure 3a. On the other hand, for the CSP sets containing over 2000 compounds, the accuracies of potentially optimal CSP predicted by 3DMolCSP are significantly better: the minimum top-1 accuracy is 0.56.

Evaluation on An External Dataset CMRT

Finally, we evaluated the performance of 3DMolCSP using an external dataset namely CMRT[5] which contains the chiral HPLC retention times of 11,720 pairs of enantiomers. Because only the retention times (t_R) of enantiomers are collected in CMRT, we computed the enantioselectivity values using a void time $t_M = 2.9$ min estimated according to standard experimental parameters (a 250×4.6 mm chiral HPLC column at 1 mL/min). Out of the compounds that were separated using six different CSPs in the CMRT data set, a total of 87 compounds were also present in our dataset. We compared the α values of these compounds in these two datasets. As shown in Figure S2, the r^2 between the α values of the same compounds is 0.846. After partitioning these compounds into two super-classes based on their α values (see Methodology Section), we observed that 96.6% compounds fall into the same super-class, indicating the external dataset provides a reliable basis for evaluating the performance of 3DMolCSP-TL in a consistent manner. Therefore, we used enantioselectivity computed on the six CSPs in the CMRT dataset to evaluate 3DMolCSP-TL trained on the data in ChirBase.

In order to assess the performance of 3DMolCSP-TL on the CMRT dataset, we trained the model using the complete ChirBase data set and subsequently evaluated its performance on the CMRT data, excluding compounds that are shared between these two datasets. As shown in Figure 4, the performance of 3DMolCSP on the CMRT data is relatively lower

the number of CSPs for each compound.

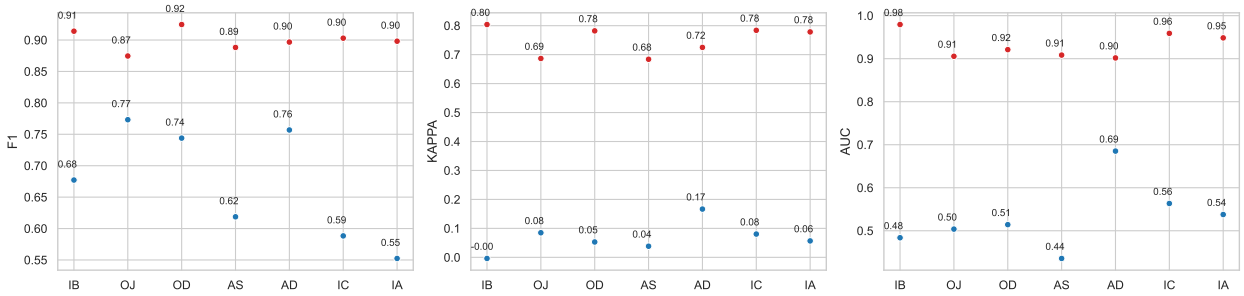


Figure 4: Evaluation of 3DMolCSP-TL (trained using ChirBase data) on the external testing dataset, CMRT. Here, the performances of 3DMolCSP-TL on all compounds in CMRT excluding those also present in ChirBase (blue) are shown in comparison with the cross-validation performances evaluated on ChirBase data (red) on six CSPs. The CSP names are shortened, OD: Chiralcel OD (Lux Cellulose-1), AS: Chiralpak AS, AD: Chiralpak AD, IA: Chiralpak IA, IC: Chiralpak IC (Lux i-Cellulose-5), and OJ: Chiralcel OJ (Lux Cellulose-3).

Table 5: Comparison of qGeoGNN and 3DMolCSP-TL on the testing set (10%) of CMRT. 3DMolCSP-TL$_{0.5}$ denotes 3DMolCSP-TL using the threshold as 0.5, and 3DMolCSP-TL$_{0.9}$ denotes 3DMolCSP-TL using the threshold as 0.9. qGeoGNN predicts α values, which can not be evaluated by AUC.

	One enantiomer used for training			No enantiomer used for training			All		
No. of compounds	248			16			264		
	qGeo-GNN	3DMol CSP-TL$_{0.5}$	3DMol CSP-TL$_{0.9}$	qGeo-GNN	3DMol CSP-TL$_{0.5}$	3DMol CSP-TL$_{0.9}$	qGeo-GNN	3DMol CSP-TL$_{0.5}$	3DMol CSP-TL$_{0.9}$
F1	0.62	0.67	0.68	0.75	0.75	0.88	0.63	0.68	0.69
KAPPA	0.20	0.19	0.25	0.50	0.00	0.71	0.22	0.19	0.26
AUC	-	0.70		-	0.83		-	0.70	

than on the ChirBase data (throughout cross-validation), while the performance metrics (F1, KAPPA, and AUC scores) are still satisfactory, especially on the CSPs with more training data (such as Chiralcel OD, Chiralpak AD, and Chiralcel OJ).

Next, we compared 3DMolCSP-TL and qGeoGNN7 on a randomly chosen 10% testing data from the CMRT dataset in accordance with the qGeoGNN paper's methodology[5]. Because qGeoGNN is designed to predict vt_R, where v is the flow rate, we converted its predictions into corresponding α values using the experimental value of v and the estimated t_M. When computing alpha values during testing, it is imperative to utilize both configurations. A potential information leakage arises from a situation in which the test set from CMRT includes only one configuration, while the other configuration is employed in the training set to determine the predicted alpha values. This occurrence may lead to leakage as a portion of the training set ends up being used for testing purposes. To account for potential information leakage, we partitioned the testing set into two distinct subsets: the first subset consisted of 248 compounds, wherein one enantiomer was utilized for training and the other for testing, while the second subset comprised 16 compounds including both enantiomers within the testing set. Given the notable imbalance between positive and negative samples, we recommend adopting a higher threshold (e.g., 0.9) instead of 0.5 for categorizing predicted probability. In Table 5, we present performance results for both 0.5 and 0.9 thresholds, with the latter demonstrating markedly superior performance. For the first subset, 3DMolCSP-TL exhibits an acceptable performance that is slightly better than qGeoGNN. On the other hand, for the second subset where no information leakage exists, 3DMolCSP-TL outperforms qGeoGNN on almost all measures. These results demonstrated the robustness and effectiveness of 3DMolCSP-TL in predicting the enantioselectivity of the compounds not similar to those in the training set.

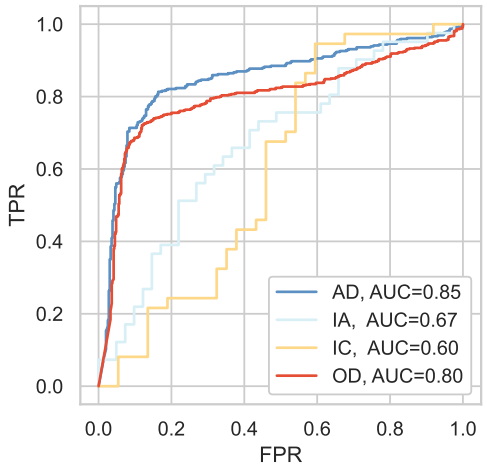


Figure 5: AUC-ROC curve of elution order prediction. The CSP names are shortened, AD: Chiralpak AD, IA: Chiralpak IA, IC: Chiralpak IC (Lux i-Cellulose-5), and OD: Chiralcel OD (Lux Cellulose-1).

Prediction of Enantiomers' Elution Orders

To highlight the model's proficiency in discerning chiral structures, we utilize 3DMolCSP to predict the elution order of two enantiomers with the same chemical structure. Initially, we extracted the elution orders of enantiomeric pairs from ChirBase and retained only those pairs within high enantioselectivity (labeled as {Class 2, Class 3}). A total of 5094, 7173, 662, and 513 enantiomeic pairs were collected in Chiralpak AD, Chiralcel OD (Lux Cellulose-1), Chiralpak IA, and Chiralpak IC (Lux i-Cellulose-5), respectively. Subsequently, we randomly divided this data into the training and testing sets within a 9:1 ratio, ensuring that any specific enantiomer pair does not appear in both the training and testing sets.

As shown in Figure 5, 3DMolCSP effectively predicts the elution orders when provided with ample training data (i.e. for Chiralpak AD and Chiralcel OD [Lux Cellulose-1]). The performances on the columns with smaller training sets (for Chiralpak IA and Chiralpak IC [Lux i-Cellulose-5]) are relatively lower. The successful prediction of the enantiomers' elution order validates our model's capability to distinguish the compounds within the same chemical structure but different 3D configurations in practical scenarios. This characteristic of 3DMolConv 2.0 opens avenues for predicting other configuration-sensitive chemical

properties, such as the retention time of enantiomers, etc.

Conclusion

In this paper, we introduce 3DMolCSP, a novel approach for predicting the enantioselectivity of compounds in chiral chromatography based on their 3D conformations. The enhanced structure-based model is proved to be geometrically complete, thus enabling it to capture chirality-sensitive insights from both configurations (R and S) of compounds. In addition, we show that employing transfer learning with a pretrained model for tandem mass spectra prediction improves the prediction of enantioselectivity.

Based on the cross-validation on the data in ChirBase, our transfer learning-enabled model, 3DMolCSP-TL, outperforms the previous machine learning model based on the Random Forest (RF) Classifier. 3DMolCSP can also be applied to selecting the potentially optimal CSP and predicting unresolvable compounds on 18 CSP columns with sufficient training data. When tested on an external data set, CMRT, 3DMolCSP outperforms the previous deep-learning model, known as qGeoGNN, on enantioselectivity prediction. Furthermore, the experiments on predicting the configuration's elution order on 4 exampled CSPs proved the capability to distinguish configurations in enantiomers in practice. These compelling outcomes demonstrate the capability of 3DMolCSP to serve as a valuable tool for selecting suitable CSPs, thereby facilitating successful enantioseparation within the realm of chiral chromatography.

Our experiments point out the limitations of current deep learning methods in predicting elution order for CSPs with limited training data. Future work could involve incorporating CSP structural information into these models, facilitating knowledge transfer from data-rich to data-poor CSPs, thereby enhancing prediction accuracy for new-generation CSPs on which only limited experimental data are available.

Acknowledgement

We are grateful to the NSF IUCRC Center for Bioanalytic Metrology (CBM) for the financial support provided under National Science Foundation Grant No. IIP-1916645 and for valuable discussions with CBM industry partners and staff. We would also like to extend our gratitude to Prof. Christian Roussel for his exceptional dedication and pioneering efforts in the creation and development of the ChirBase database. This work was also partially supported by the National Science Foundation (Grant No: DBI-2011271).

Supporting Information Available

The following files are available free of charge.

- `AnalChem_3DMolCSP_supp.pdf`: proof of geometric completeness; the details of implement and training settings; classes portion of datasets (Figure S1); consistency between estimated alpha values from CMRT and experimental alpha values from ChirBase (Figure S2).
- The codes for data preprocessing, model training, and validation are available at GitHub: `https://github.com/JosieHong/3DMolCSP`. The data on one chiral stationary phase collected in ChirBase, recognized as Chirobiotic V, is shared as a demo dataset with the source codes.

The entire ChirBase is commercially available at `https://chirbase.u-3mrs.fr/`. To reproduce 3DMolCSP, potential users may retrain the model on the ChirBase data using the train code shared at GitHub.

References

(1) Ward, K. D.; Bravenec, A. D.; Ward, T. J. *Encyclopedia of Analytical Chemistry*; John Wiley & Sons, Ltd, 2019; pp 1–28.

(2) Mattrey, F. T.; Makarov, A. A.; Regalado, E. L.; Bernardoni, F.; Figus, M.; Hicks, M. B.; Zheng, J.; Wang, L.; Schafer, W.; Antonucci, V., et al. Current challenges and future prospects in chromatographic method development for pharmaceutical research. *TrAC Trends in Analytical Chemistry* **2017**, *95*, 36–46.

(3) Barhate, C. L.; Joyce, L. A.; Makarov, A. A.; Zawatzky, K.; Bernardoni, F.; Schafer, W. A.; Armstrong, D. W.; Welch, C. J.; Regalado, E. L. Ultrafast chiral separations for high throughput enantiopurity analysis. *Chemical Communications* **2017**, *53*, 509–512.

(4) Bryant, C.; Adam, A.; Taylor, D.; Rowe, R. Towards an expert system for enantioseparations: induction of rules using machine learning. *Chemometrics and Intelligent Laboratory Systems* **1996**, *34*, 21–40.

(5) Xu, H.; Lin, J.; Zhang, D.; Mo, F. Retention time prediction for chromatographic enantioseparation by quantile geometry-enhanced graph neural network. *Nature Communications* **2023**, *14*, 3095.

(6) Ju, R.; Liu, X.; Zheng, F.; Lu, X.; Xu, G.; Lin, X. Deep neural network pretrained by weighted autoencoders and transfer learning for retention time prediction of small molecules. *Analytical Chemistry* **2021**, *93*, 15651–15658.

(7) Gallarati, S.; Fabregat, R.; Laplaza, R.; Bhattacharjee, S.; Wodrich, M. D.; Corminboeuf, C. Reaction-based machine learning representations for predicting the enantioselectivity of organocatalysts. *Chemical Science* **2021**, *12*, 6879–6889.

(8) Hoque, A.; Sunoj, R. B. Deep learning for enantioselectivity predictions in catalytic asymmetric β-C–H bond activation reactions. *Digital Discovery* **2022**, *1*, 926–940.

(9) Betinol, I. O.; Lai, J.; Thakur, S.; Reid, J. P. A Data-Driven Workflow for Assigning and Predicting Generality in Asymmetric Catalysis. *Journal of the American Chemical Society* **2023**,

(10) Chen, Y.; Zhang, F.; Dang, Z.; He, X.; Luo, C.; Liu, Z.; Peng, P.; Dai, Y.; Huang, Y.; Li, Y., et al. Chiral detection of biomolecules based on reinforcement learning. *Opto-Electronic Science* **2023**, *2*, 220019–1.

(11) Chen, M.; Wu, T.; Xiao, K.; Zhao, T.; Zhou, Y.; Zhang, Q.; Aires-de Sousa, J. Machine learning to predict the specific optical rotations of chiral fluorinated molecules. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **2019**, *223*, 117289.

(12) Sheridan, R.; Schafer, W.; Piras, P.; Zawatzky, K.; Sherer, E. C.; Roussel, C.; Welch, C. J. Toward structure-based predictive tools for the selection of chiral stationary phases for the chromatographic separation of enantiomers. *Journal of Chromatography A* **2016**, *1467*, 206–213.

(13) Piras, P.; Sheridan, R.; Sherer, E. C.; Schafer, W.; Welch, C. J.; Roussel, C. Modeling and predicting chiral stationary phase enantioselectivity: An efficient random forest classifier using an optimally balanced training dataset and an aggregation strategy. *Journal of separation science* **2018**, *41*, 1365–1375.

(14) Lin, S.; Dikler, S.; Blincoe, W. D.; Ferguson, R. D.; Sheridan, R. P.; Peng, Z.; Conway, D. V.; Zawatzky, K.; Wang, H.; Cernak, T., et al. Mapping the dark space of chemical reactions with extended nanomole synthesis and MALDI-TOF MS. *Science* **2018**, *361*, eaar6236.

(15) Wang, S.; Guo, Y.; Wang, Y.; Sun, H.; Huang, J. Smiles-bert: large scale unsupervised pre-training for molecular property prediction. Proceedings of the 10th ACM inter-

national conference on bioinformatics, computational biology and health informatics. 2019; pp 429–436.

(16) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural message passing for quantum chemistry. International conference on machine learning. 2017; pp 1263–1272.

(17) Chen, C.; Ye, W.; Zuo, Y.; Zheng, C.; Ong, S. P. Graph networks as a universal machine learning framework for molecules and crystals. *Chemistry of Materials* **2019**, *31*, 3564–3572.

(18) Hong, Y.; Li, S.; Welch, C. J.; Tichy, S.; Ye, Y.; Tang, H. 3DMolMS: prediction of tandem mass spectra from 3D molecular conformations. *Bioinformatics* **2023**, *39*, btad354.

(19) Gasteiger, J.; Giri, S.; Margraf, J. T.; Günnemann, S. Fast and uncertainty-aware directional message passing for non-equilibrium molecules. *arXiv preprint arXiv:2011.14115* **2020**,

(20) Gasteiger, J.; Becker, F.; Günnemann, S. Gemnet: Universal directional graph neural networks for molecules. *Advances in Neural Information Processing Systems* **2021**, *34*, 6790–6802.

(21) Schütt, K.; Unke, O.; Gastegger, M. Equivariant message passing for the prediction of tensorial properties and molecular spectra. International Conference on Machine Learning. 2021; pp 9377–9388.

(22) Fang, X.; Liu, L.; Lei, J.; He, D.; Zhang, S.; Zhou, J.; Wang, F.; Wu, H.; Wang, H. Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence* **2022**, *4*, 127–134.

(23) Wang, L.; Liu, Y.; Lin, Y.; Liu, H.; Ji, S. ComENet: Towards complete and efficient

message passing for 3D molecular graphs. *Advances in Neural Information Processing Systems* **2022**, *35*, 650–664.

(24) Yang, X.; Neta, P.; Stein, S. E. Extending a tandem mass spectral library to include MS2 spectra of fragment ions produced in-source and MSn spectra. *Journal of The American Society for Mass Spectrometry* **2017**, *28*, 2280–2287.

(25) Riniker, S.; Landrum, G. A. Better informed distance geometry: using what we know to improve conformation generation. *Journal of chemical information and modeling* **2015**, *55*, 2562–2574.

(26) Zhang, W.; Wang, Z.; Loy, C. C. Exploring data augmentation for multi-modality 3d object detection. *arXiv preprint arXiv:2012.12741* **2020**,

(27) Valverde-Albacete, F. J.; Peláez-Moreno, C. 100% classification accuracy considered harmful: The normalized information transfer factor explains the accuracy paradox. *PloS one* **2014**, *9*, e84217.

(28) Paszke, A. et al. *Advances in Neural Information Processing Systems 32*; Curran Associates, Inc., 2019; pp 8024–8035.

For Table of Contents Only