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

Yuhui Hong

Luddy School of Informatics, Computing, and Engineering
Indiana University Bloomington

October 11, 2024

About Me

Yuhui Hong

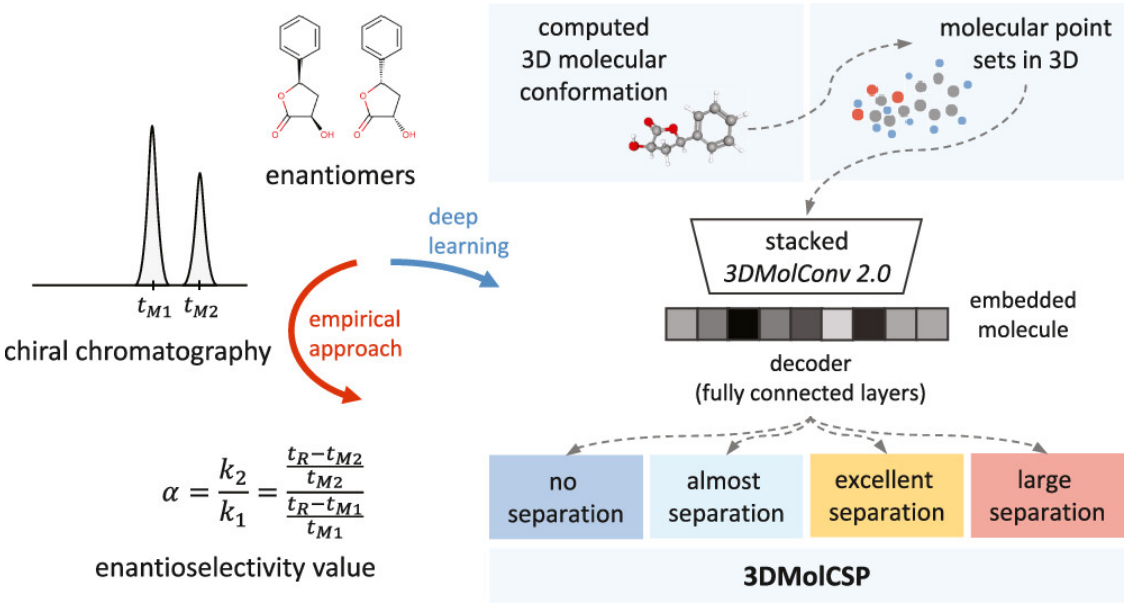
Ph.D. Candidate in Computer Science
Indiana University Bloomington
Advised by Prof. Haixu Tang

My research explores the intersection of **deep learning**, bioinformatics, and cheminformatics, with a focus on advancing the **identification of small molecules** in two key scenarios. The first involves predicting tandem mass spectra and other molecular properties from 3D structures, addressing gaps—often referred to as the "dark matter"—in existing spectral reference libraries. The second approach moves beyond the traditional reliance on database-driven methods by predicting chemical formulas directly from tandem mass spectra. Additionally, I am passionate about developing **reliable and interpretable neural networks** for real-world applications.

Introduction

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*



ChirBase

ChirBase (Chemical Database of Chiral HPLC/SFC Separations) was created in 1988 and has been internationally recognized for over 25 years. The database contains a collection of 306,989 chiral HPLC/SFC separations extracted from literature and patents. All the data are checked by experts.
Chemical structure search (Exact, Substructure, Similarity) can be performed as well as simple or complex searches on other fields such as:

- Compounds: Chemical Structure, IUPAC name, trade name, specific optical rotation, absolute configuration...
- Literature references
- Chromatographic conditions: mobile phase, temperature, flow rate, detection...
- Chiral Stationary phases: Chemical structure, column size, trade name, supplier...
- Chromatographic data: elution order, retention times, enantioselectivity, resolution...

ChirBase (cont.)

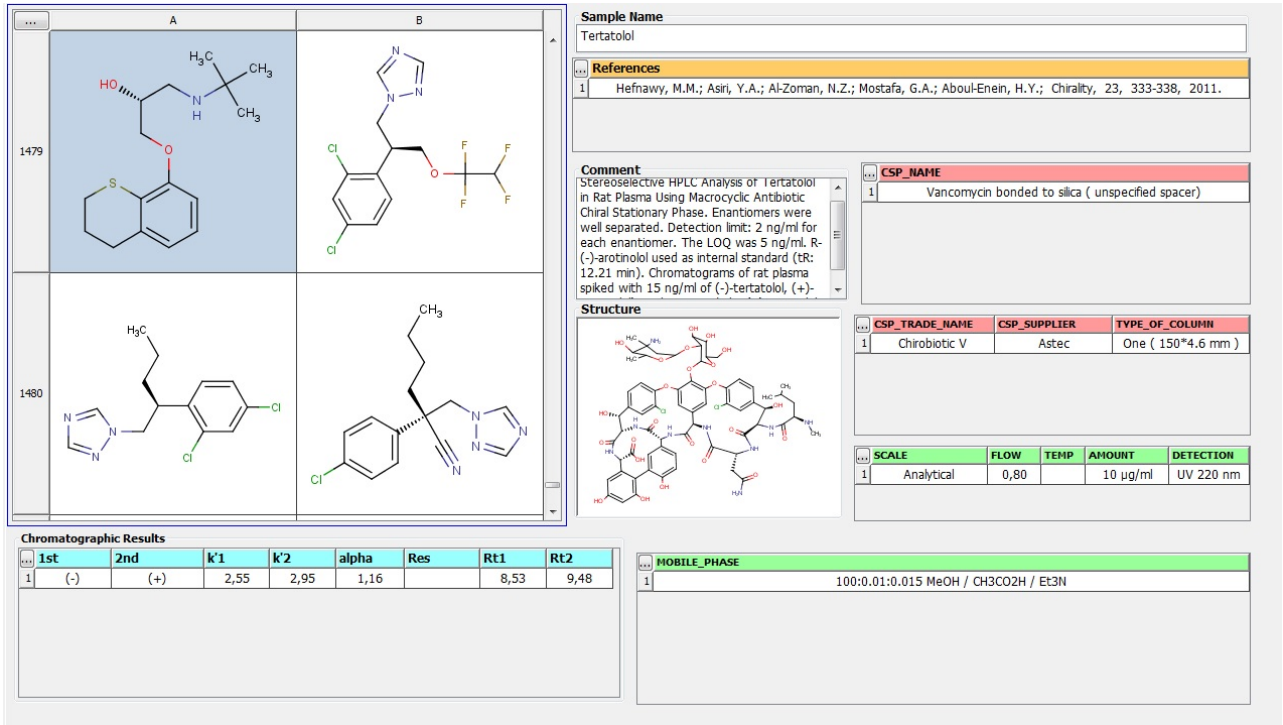


Figure: ChirBase Screen.

Please check the website of ChirBase for more details: `https://chirbase.u-3mrs.fr/`.

Data Prepossessing

Building on previous studies, we selected 18 CSPs from 1603 chiral columns with sufficient experimental data [Sheridan, 2016]. The model was cross-validated on this training set and evaluated on an independent test set of 6 CSPs from CMRT [**?**].

We limited the atom count in each compound to 300 to focus on small molecules, retaining only those composed of common atoms (C, H, O, N, F, S, Cl, P, B, I, Br).

CSP	no. of compounds				
	all		after preprocessing		
	ChirBase	CMRT	ChirBase	CMRT	overlap
Chiralcel OD (Lux Cellulose-1)	14,395	178	13,746	171	8
Chiralpak AD	11,194	292	10,906	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

Figure: Number of compounds with the experimental data of 18 different CSPs available in ChirBase and CMRT.

Discretization of Enantioselectivity Values

Enantioselectivity value, also known as the ratio of retention factor, is defined as:

$$\alpha = \frac{k_2}{k_1} = \frac{\frac{t_R - t_M 2}{t_M 2}}{\frac{t_R - t_M 1}{t_M 1}} \tag{1}$$

where t_R denotes the time the analyte spends in the stationary phase and t_M denotes the retention time for an unretained analyte.

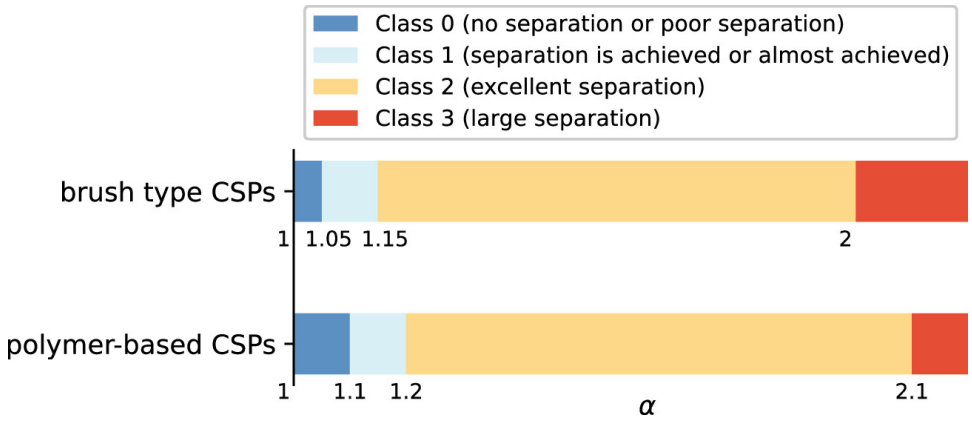


Figure: Four classes (denoted as Class 0, 1, 2, and 3, respectively) of compounds are defined based on their α values for the brush-type or the polymer-based CSPs. Between these two types of CSPs, the fraction of compounds in the four classes is 10–15, 20–30, 45–55, and 10–15%, respectively.

Architecture of Neural Network

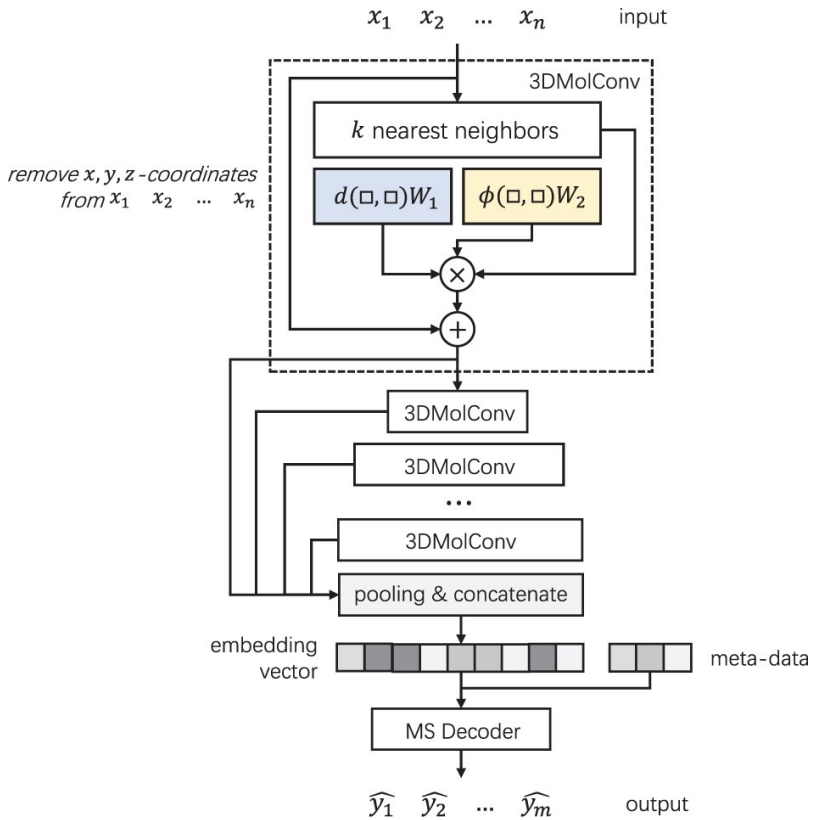


Figure: Architecture of neural network for tandem mass spectra prediction from 3D molecular conformations.

General Idea of Neural Network on Point Sets

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

[a]The feature marked by an asterisk is only used for the prediction of enantiomers' elution orders.

Figure: Point set encoding of a compound, in which each atom in the compound is encoded as a vector of 22 dimensions, representing the $x-$, $y-$, and $z-$ coordinates and other attributes of the atom.

$$f(\{x_1, x_2, ..., x_n\}) \approx g(b(x_1), b(x_2), ..., b(x_n)) \tag{2}$$

where f is the representation function on the input point set, which is from the elemental operation b on each element x_i in the point set through an aggregated function g.

Elemental Convolution on 3D Conformation

In 3DMolConv 1.0, the elemental operation is:

$$x_i^{l+1} = x_i^l + \sum_{j \in \mathcal{N}(x_i^l)} d(x_i^l, x_j^l) W_1^l \circ \phi(x_i^l, x_j^l) W_2^l \tag{3}$$

where $\circ$ represents the element-wised multiplication, W_1^l represents the learnable filter on distance, W_2^l represents the filter on direction, and x_j^l represents one of the k-nearest neighbors of point x_i^l. The distance between two points x_i and x_j is computed as $d(x_i, x_j) = ||x_i - x_j||$, and the angle between the point vector x_i and x_j is computed as $\phi(x_i, x_j) = \sum_{k \in \mathcal{N}(x_i)} e_{ij}^T e_{ik}$, where $e_{ij} = x_i^T x_j$.

In 3DMolConv 2.0, we improve the filter on the direction. Then the elemental operation is:

$$x_i^{l+1} = x_i^l + \sum_{j \in \mathcal{N}(x_i^l)} d(x_i^l, x_j^l) W_1^l \circ [x_j^l || \phi(x_i^l, x_j^l)] \tag{4}$$

where $||$ represents concatenation of two vector.

Workflow of building 3DMolCSP-TL

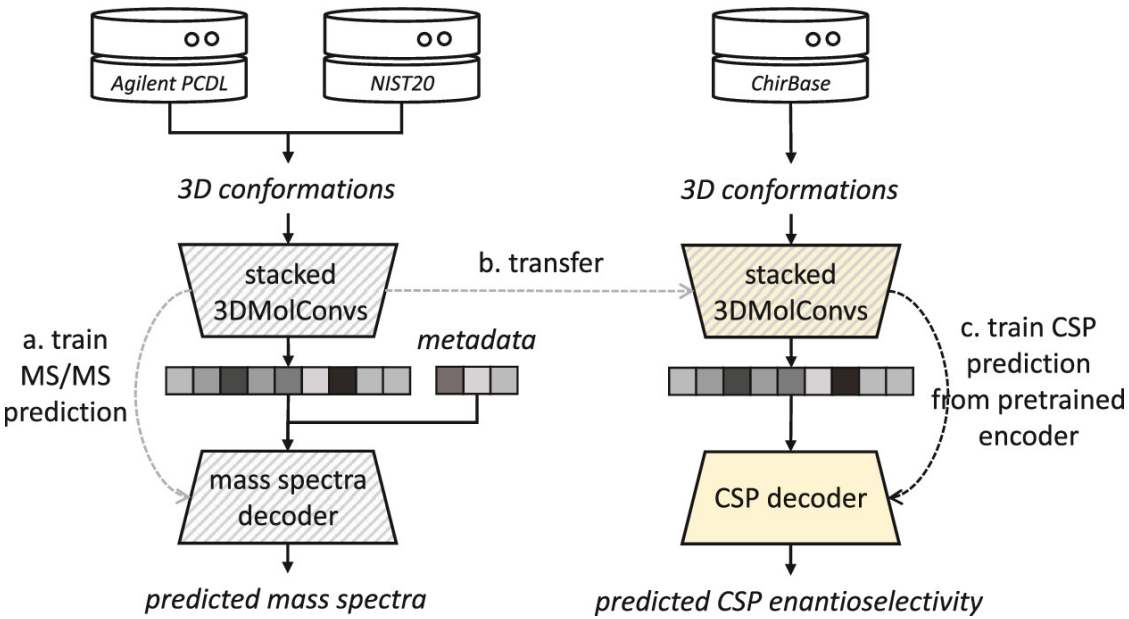


Figure: Workflow of building the 3DMolCSP-TL model using the transfer learning approach. To build the 3DMolCSP model from scratch (i.e., the independent learning approach), we follow the flow in the right panel only.

Prediction of CSP Enantioselectivity

CSP	3DMolCSP-SC		3DMolCSP-TL			
	F1	kappa	F1	ΔF1	kappa	Δkappa
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-OVM	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

Figure: Performance of 3DMolCSP for Enantioselectivity Prediction.

Prediction of CSP Enantioselectivity (cont.)

CSP	RF classifier			3DMolCSP-TL		
	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)

Figure: Comparison of 3DMolCSP-TL and the State-of-the-Art ML Model for Enantioselectivity Prediction.

Assistance in CSP Selections

Assume we have some available compounds, and for each compound, there are several column options. Can our model help us choose the optimal column?

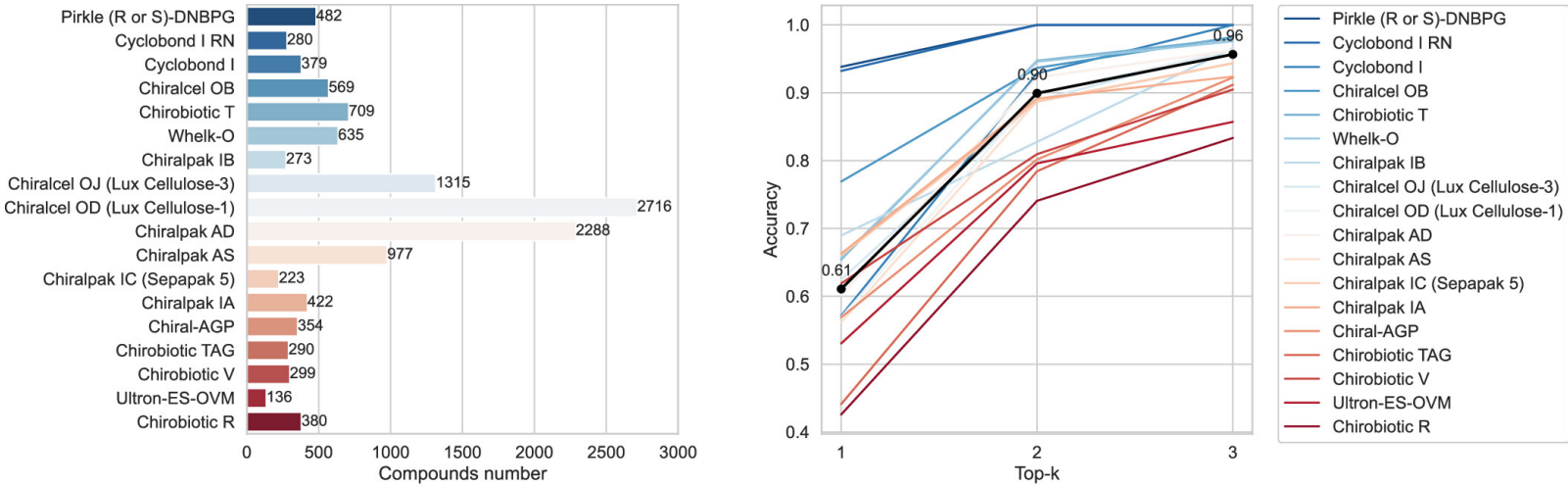


Figure: Prediction of potentially optimal CSP. The accuracy was evaluated on the compounds resolved by (i.e., falling into Class 2 or Class 3) more than one CSP in ChirBase. (a) Number of compounds whose best enantioseparation is achieved by each CSP. (b) Top-k (k = 1, 2, and 3) accuracy for the potentially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

Prediction of Enantiomers' Elution Orders

We extracted the elution orders of enantiomeric pairs from ChirBase, retaining only those with high enantioselectivity (Class 2, Class 3). In total, we collected 5094 pairs for Chiralpak AD, 7173 for Chiralcel OD (Lux Cellulose-1), 662 for Chiralpak IA, and 513 for Chiralpak IC (Lux i-Cellulose-5).

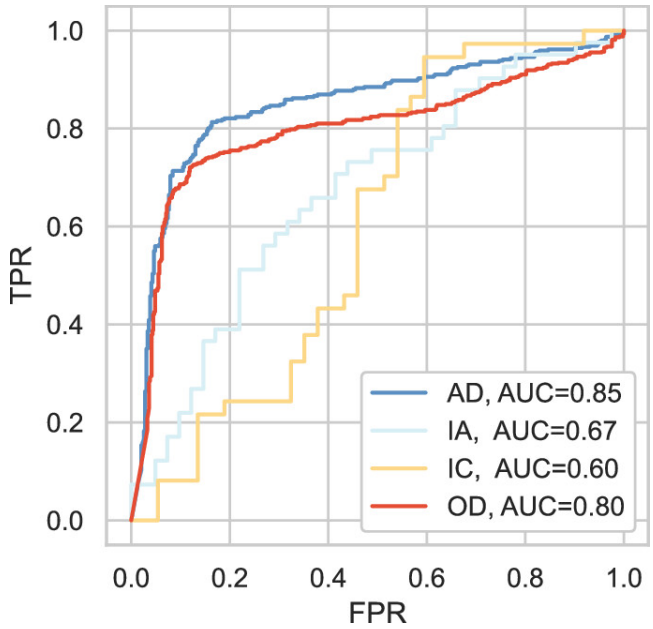


Figure: AUC-ROC curve of elution order prediction. The CSP names are shortened, AD: Chiralpak AD, IA: Chiralpak IA, IC: Chiralpak IC (Lux i-Cellulose-5), and OD: Chiralcel OD (Lux Cellulose-1).

Take Away

- In this work, a neural network based on molecular 3D conformations is proposed, capable of predicting enantioselectivity on CSP columns.
- The prediction of enantioselectivity can assist in the selection of CSP columns.
- Since the neural network is geometrically complete, it can also predict the elution order of enantiomers.

Acknowledgement

We are grateful to the NSF IUCRC Center for Bioanalytic Metrology (CBM) for the financial support provided under the 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).

References

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

Hong, Y., Welch, C. J., Piras, P., & Tang, H. (2024).
Enhanced structure-based prediction of chiral stationary phases for chromatographic enantioseparation from 3D molecular conformations.
Analytical Chemistry, 96(6), 2351-2359.

Hong, Y., Li, S., Welch, C. J., Tichy, S., Ye, Y., & Tang, H. (2023).
3DMolMS: prediction of tandem mass spectra from 3D molecular conformations.
Bioinformatics, 39(6), btad354.

Thanks!