Analytica Chimica Acta 1379 (2025) 344733

Contents lists available at ScienceDirect

Analytica Chimica Acta

journal homepage: www.elsevier.com/locate/aca

Machine learning-assisted retention time predictions on a cellulose Tris (3,5)-dimethylphenylcarbamate column in polar organic mode

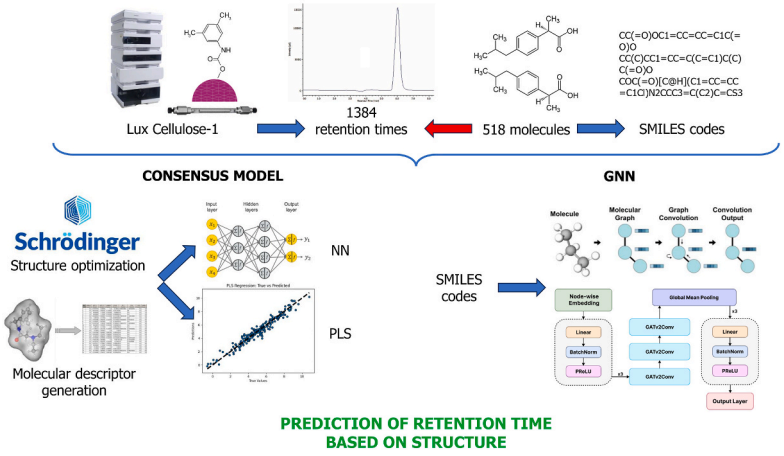
Attila Imre [a,b,1], Gergely Dombi [b,c,1], Máté Dobó [b,c], Ali Mhammad [b,c], Elek Ferencz [d], Balázs Balogh [e,c], Anna Vincze [b,c], Zoltán-István Szabó [f,g], György Tibor Balogh [b,c,h], Anita Rácz [i,2,**], Gergő Tóth [b,c,2,*]

[a] Center for Health Technology Assessment, Semmelweis University, Üllői str. 25, Budapest, 1091, Hungary
[b] Department of Pharmaceutical Chemistry, Semmelweis University, Hőgyes, Endre str. 9, Budapest, 1092, Hungary
[c] Center for Pharmacology and Drug Research & Development, Semmelweis University, Üllői str. 28, Budapest, 1085, Hungary
[d] Emergency County Hospital Miercurea Ciuc, Service of Translational Medicine and Clinical Research, Doctor Dénes László str. 2, Miercurea Ciuc, 530173, Romania
[e] Department of Organic Chemistry, Semmelweis University, Hőgyes Endre str. 7, Budapest, 1092, Hungary
[f] Department of Pharmaceutical Industry and Management, George Emil Palade University of Medicine, Pharmacy, Science and Technology of Targu Mures, Gh. Marinescu 38, Targu Mures, 540142, Romania
[g] Sz-imfidum Ltd., Sat Lunga nr. 504, Covasna, 525401, Romania
[h] Department of Chemical and Environmental Process Engineering, Budapest University of Technology and Economics, Műegyetem rkp. 3, Budapest, 1111, Hungary
[i] Institute of Materials and Environmental Chemistry, HUN-REN Research Centre for Natural Sciences, Magyar Tudósok 2, Budapest, 1117, Hungary

HIGHLIGHTS

- New machine learning based methodology using graph neural network.
- Enantioseparation without a trial and error screening phase.
- Retention time prediction made enabled for chiral stationary phase.
- Web-based tool (chiralscreen.com) for retention time prediction from SMILES codes.

GRAPHICAL ABSTRACT



* Corresponding author. Hőgyes E. str. 9, 1092, Budapest, Hungary.
** Corresponding author.
E-mail addresses: imre.attila@stud.semmelweis.hu (A. Imre), dombi.gergely@stud.semmelweis.hu (G. Dombi), dobo.mate@stud.semmelweis.hu (M. Dobó), ali.mhammad@phd.semmelweis.hu (A. Mhammad), elekferencz@yahoo.com (E. Ferencz), balogh.balazs@semmelweis.hu (B. Balogh), vincze.anna@semmelweis.hu (A. Vincze), zoltan.szabo@umfst.ro (Z.-I. Szabó), balogh.gyorgy.tibor@semmelweis.hu (G.T. Balogh), racz.anita@ttk.hu (A. Rácz), toth.gergo@semmelweis.hu (G. Tóth).
[1] These authors contributed equally and are considered the first authors of this article.
[2] These authors contributed equally and are considered the last authors of this article.

https://doi.org/10.1016/j.aca.2025.344733
Received 5 July 2025; Received in revised form 22 September 2025; Accepted 30 September 2025
Available online 30 September 2025

ARTICLE INFO

Handling Editor: Xiu-Ping Yan

Keywords:
Machine learning
Enantioseparation
QSRR
Neural network
Retention time prediction
Chiral screening

ABSTRACT

Background: Enantioseparation in HPLC is a considerable challenge in analytical chemistry, frequently requiring numerous trials with varying experimental conditions to achieve baseline separations. To address this issue, we propose a solution that utilizes consensus modelling based on partial least squares (PLS) regression method together with neural network (NN) algorithms and a graph neural network (GNN) method to predict the retention times of compounds on Lux Cellulose-1 chiral stationary phase under various polar organic mode mobile phases.

Results: A homogeneous dataset was collected for the developed machine learning methods, consisting of 535 unique molecules and 1,414 retention time measurements under four polar organic mode conditions (acidic and basic methanol, acidic and basic acetonitrile). The PLS + NN consensus model showed outstanding results in condition-specific predictions, achieving R^2 values over 0.70 and RMSE values below 0.40 in most cases. Conversely, the GNN model excelled in combined predictions under all conditions, achieving a R^2 of 0.58 and $RMSE_{CV}$ of 0.49 during cross-validation, as well as a R^2 of 0.85 and $RMSE_{Test}$ of 0.25 on the test set.

Significance: Our research presents a novel approach for predicting chiral separations, offering an easy-to-use, open-access web tool to the scientific community. The robust GNN model was used to create a web server (https://chiralscreen.com) that enables the prediction of retention times, separation capabilities, and elution orders for various compounds. Furthermore, the software helps determine optimal initial chromatographic conditions for separations.

1. Introduction

The number of chiral pharmaceutical compounds is steadily rising, hence propelling the development of chiral pharmaceutical analysis [1]. Various methods exist for chiral impurity analysis, such as capillary electrophoresis, gas chromatography, supercritical fluid chromatography, as well as additional spectroscopic techniques [2]. However, the most significant and the gold standard in this field is high performance liquid chromatography (HPLC) using chiral stationary phases (CSPs) [3,4]. Numerous chiral columns are available commercially, each promising to offer enantiomeric separation for various chiral compounds, however, no single column is universally successful for the enantioseparation of all compounds. Consequently, column manufacturers offer sets including five-six CSPs, capable of effectively separating the enantiomers of a diverse variety of compounds [5]. Nonetheless, success is not guaranteed, and method development often relies on a trial-and-error approach. The greatest challenge with chiral separations is their unpredictability; it is impossible to determine in advance which column and eluent mixture will exhibit chiral recognition towards our analytes. To improve predictive capabilities, researchers have explored the analysis of structurally similar compounds to establish correlations between molecular structure and separation efficiency [2,6]. While this approach has shown some promise, its success has been limited. Mapping ligand-selector interactions has provided a deeper understanding of the processes underlying chiral recognition and has led to the synthesis of more efficient chiral selectors [7–9]. However, the primary goal of improving predictive capabilities has been only partially achieved by this approach. Previously, the quantitative structure activity relationship (QSAR) approach has been utilized; nevertheless, contemporary machine learning (ML) and different artificial intelligence-based approaches currently provide novel prospects for enhancing the predictability of chiral separations [10–13]. However, currently, no fully functional tools are available to scientists in this field. One of the main limitations in developing such tools is the lack of homogeneous datasets. Using literature data or data from various databases can introduce errors during the training phase, consequently leading to less reliable results [14]. Our other insight, that there is no freely available program, which can help for scientists to predict the retention time or resolution in chiral separation.

The aim of the present study is to provide suitable solutions to the issues. The main goal was to compile our own database of chiral separation data on a Lux Cellulose-1 column, containing cellulose tris(3,5-dimethylphenylcarbamate) chiral selector under polar organic mode, by analyzing many analytes under various chiral separation conditions. Lux Cellulose-1 column with polar organic mode was chosen as the initial model, however, our findings may potentially extend to include other columns or eluents within this model. Cellulose tris(3,5-dimethylphenylcarbamate)-based CSPs are among the most widely used cellulose-based chiral selectors [11,15–17]. The polar organic mode, using neat alcohol or acetonitrile as eluents, offers significant advantages over traditional normal-phase enantioseparation methods, such as excellent separation capabilities, improved peak shapes, reduced retention times, and are generally more environmentally sustainable while remaining compatible with reversed-phase chromatographic systems [18].

Secondly, our aim was to develop and compare two models - consensus modelling based on PLS and NN algorithms, and a graph neural network – for the prediction of retention times on Lux Cellulose-1 column under four different polar organic mode conditions. Finally, our aim was to develop an open-access web tool capable of predicting the retention time of various analytes and estimating (enantio)separation based on retention time differences.

2. Materials and methods

The chemicals used in the HPLC measurements came from chemical pools available at the Department of Pharmaceutical Chemistry of Semmelweis University and Department of Pharmaceutical Industry and Management of the University George Emil Palade University of Medicine, Pharmacy, Science and Technology of Targu Mures are listed in *Supplementary Document* 1 by simplified molecular input line entry specification (SMILES) codes [19] with corresponding retention times in each measurement condition. Most compounds were purchased from reputable suppliers, including Merck, TCI, and LGC Standards. However, some compounds (e.g. thiourea derivatives) were synthesized at Semmelweis University, and their structures were previously confirmed using NMR, MS, and IR. All substances stored at 4 °C until use. Methanol (MeOH), acetonitrile (ACN), acetic acid (Acac) and diethylamine (DEA) were obtained from Sigma-Aldrich, Hungary (Budapest, Hungary). Chromatographic measurements were carried out on a Phenomenex Lux Cellulose-1 [Cellulose tris(3,5-dimethylphenylcarbamate)] chiral column (150 × 4 mm, particle size 5 μm) which was obtained from Phenomenex (Torrance, CA, USA).

2.1. HPLC analysis

LC-UV analyses of the chemical entities of diverse structure were carried out on an Agilent 1100 HPLC system, consisting of an inline

degasser (G1322A), quaternary pump (G1311A), automatic injector (G1329A) paired with sample thermostat (G1330A), column thermostat (G1316A) and a diode array detector (G1315A) controlled by Agilent Chemstation B04.03-SP2 software (Agilent Technologies, Waldbronn, Germany).

Samples were dissolved in MeOH and diluted to a final concentration of approximately 0.3 mg/mL using pure MeOH then filtered through 0.45 μm pore size PTFE syringe filters. Chromatographic conditions were constant for every measurement, a flow of 1 ml/min, 20 °C column temperature and an injection volume of 1 μL were used. Detection wavelengths were 205, 210, 254, 270, 310 nm. Analysis time was set at 15 min. Measurements were carried out in polar organic mode using neat MeOH and neat ACN with Acac or DEA as a modifier in a concentration of 0.1 %. Each sample was measured in all the four eluents. Each analyte was injected separately, pure enantiomers, where available, were used for retention time determination of the chiral materials, either by spiking the racemic mixture or by injecting separately.

Strict rules were used for analysing the resulting chromatograms. Only peaks with a signal to noise ratio of at least 50 were evaluated and peaks with significant broadening were ruled out as well. During evaluation another important aspect was to be able to identify a peak on at least two different detection wavelengths, this rule was applied to minimise the possibility of evaluating ghost peaks or system peaks as real results. Taking into consideration the above-mentioned principles the retention times of each sample were evaluated and used for further analysis. As a system suitability control after every 50 sample injections, acetone was injected, and the retention time was compared to those previously measured retention time of acetone. Retention times of the standards were subject to a variation of ±0.2 min. It should be noted that in the ACN–DEA mobile phase, a high background absorption was observed at lower wavelengths, limiting detection to compounds with strong absorbance at higher wavelengths. When both a racemic mixture and an enantiomerically pure compound were available, the two samples were injected consecutively, and the individual enantiomers were identified based on their retention times (Supplementary Fig. S1).

2.2. Dataset construction

Measurement data and molecular structures corresponding to those measurements from the HPLC experiments were collected manually and subsequently processed through an automated workflow to ensure consistency and reliability. The processing pipeline included two steps: labelling followed by standardization and was implemented in RDKit 2023.9.4 [20]. During the labelling phase, each measurement was annotated based on its specific experimental conditions, including the solvent composition and modifier used. For standardization, the input SMILES strings were pre-processed to obtain uncharged and unsalted forms, selecting the largest fragment in cases of multiple components. These SMILES were then normalized to isomeric representations and chirality information were preserved.

2.3. Descriptor and Morgan fingerprint generation

We have used the standardized SMILES codes to generate 3D structures at pH = 4.5 (for the acid form) and pH = 9 (for the basic form) with the OPLS4 force field (Schrodinger Suite, Ligprep module) [21]. These structures were used for 2D/3D descriptor calculation with the DRAGON 7.0 software [22]. Morgan fingerprints (with 1024 bits) were calculated with RDKit 2021.9.3 package in KNIME (4.6.4) environment. After the calculation, the descriptors and fingerprints were curated to eliminate the constant and highly correlated descriptors (intercorrelation limit = 0.997) from the dataset [23].

2.4. Consensus modelling based on PLS and NN algorithms

For consensus modelling, we have combined the traditional partial least squares (PLS) regression method [24] with a machine learning-based neural network (NN) [25] to capture both linear and non-linear elements in the relationship between the input molecular descriptors and the output retention times, and therefore to maximize the performance of the final model. PLS is a classical and well-known technique, which explores the linear relationship between the reference (Y) and the input (X) variables with the use of regression coefficients (b). The number of latent variables (PLS components or LVs) is crucial in the modelling, because it affects the performance of the model, and by selecting the optimal number of LVs, we can avoid overfitting as well. The number of LVs was determined based on the global minima of the root mean squared error of cross-validation ($RMSE_{CV}$) values during the validation of the model. For the selection of the most important input variables, we have used genetic algorithm (GA) [26] with a population size of 100, and 100 iteration circles. PLS regression and genetic algorithm were carried out in MATLAB 2019a (MathWorks, Inc.) with PLS Toolbox (v. 8.8.1, Eigenvector Research, Inc.). On the machine learning side of the consensus model, the more complex NN is working with hidden layers and neurons inside of its architecture to adjust the weights of the input variables, to maximize the predictive performance of the model. KERAS neural network (with python deep learning integration in KNIME) was used for the tasks with three densely connected hidden layers. The number of neurons was optimized between 100 and 1000 in a hyperparameter optimization loop [27]. The ReLU activation function was used in each hidden layer. Model validation involved a 5-fold iterative and randomized internal cross-validation, an external test validation with an 80:20 training-to-test split ratio [28], and a permutation test for the detection of possible overfitting. The indexes of the molecules in the train-test splits were subsequently saved to ensure that the same splits are used for both PLS-NN and GNN-based modelling. We have evaluated the models based on the R^2 values for cross-validation (R^2_{CV}) and test validation (R^2_{Test}), and the root-mean-squared errors for both cross-validation ($RMSE_{CV}$) and test validation ($RMSE_{Test}$). The performance parameters can be calculated with the following equations:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} \left(y_i - \underline{y}_i\right)^2} \quad (1)$$

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

where y_i is the measured value (reference), $\hat{y}_i$ is the predicted value, $\underline{y}_i$ is the average of the measured values and n is the number of samples.

In the consensus modelling, we have used the average values of the predicted retention times of the compounds, based on the predictions of the PLS and NN models.

2.5. GNN modelling

GNNs represent a class of neural networks designed to operate on graph-structured data, enabling the modelling of complex relationships and interactions between entities. In molecular analysis, GNNs enable the depiction of molecules as graphs, with atoms as nodes and chemical bonds as edges, effectively capturing both local and global molecular features [29–31].

Graph representation for the GNN model were prepared, node embeddings with the deepchem 2.7.1 [32] framework utilizing an implementation of Kearnes et al. work (Supplementary Table S1) [33]. This method constructed graphs, where the molecule is represented by an ν-dimensional node feature vector (where ν is the number of atoms in the molecule), and a list containing the node pairs that related to an edge (bond). The features are listed in Supplementary Table S2, and they include atom type, formal charge, hybridization, hydrogen bonding,

aromaticity, degree, number of hydrogens, and chirality. These data in combination produced a 32-dimensional vector representation for each atom in the molecule.

The GNN architecture consisted of three main components: pre-message passing layers, message passing layers, and post-message passing layers. The pre-message passing layers transform the input node features to prepare them for message passing operations. The message passing layers, which form the core of the GNN, facilitate the exchange of information between connected nodes (atoms). Through multiple iterations, the network aggregates and updates node representations based on their local neighbourhoods, enabling the capture of intricate patterns and interactions within the molecular graph. The post-message passing layers aggregate the refined node information to generate a global representation of the entire molecule. To account for different experimental conditions during retention time prediction, an embedding layer containing a one-hot encoded environment vector was concatenated to the molecular representation after the message passing process.

To incorporate the influence of chromatographic conditions, a one-hot encoded vector representing the solvent and pH environment was embedded through a dedicated transformation block (linear layer, batch normalization, activation, dropout). This solvent embedding was concatenated with the learned molecular representation after pooling and passed through the final prediction layers. This design allows the model to jointly learn the effects of molecular structure and experimental environment on retention time.

The GNN architecture was implemented using PyTorch 1.31.1 [34], PyG 2.3.0 [35]. Parallelization was carried out with the ray 2.8.0 package [36], meanwhile hyperparameter optimization was performed with the Hyperopt 0.2.7 package [37]. Experiment tracking was done using MLflow 2.10.0 [38]. All the experiments were conducted with the Python 3.9.18 programming language. Cosine weight decay [39] was employed during training for 300 epochs, and the model parameters corresponding to the lowest validation loss were used for the final predictions.

2.6. GNN model webserver

The publicly available webserver was implemented in the Dash 2.15.0 package [40] and bundled into a docker 26.1.4 container [41] to be run on a webserver.

3. Results and discussion

A total of 535 compounds were analyzed, yielding a dataset of 1,414 retention times. The mean number of atoms was 19.30 ± 7.35 and the average Tanimoto similarity [42] was 0.10 ± 0.06 as depicted in Fig. 1.

The low average Tanimoto similarity indicates a high degree of diversity within the dataset. Fig. 2 presents a Uniform Manifold Approximation and Projection (UMAP) plot, illustrating the distribution of our dataset within the drug-like molecular space [43].

The projected descriptors of the compounds in our dataset (blue) are compared to FDA-approved molecules in the DrugBank dataset (gray). The dataset is well-dispersed across the chemical space of known drugs, indicating that it sufficiently spans the diversity of drug-like molecules. This broad coverage suggests that the dataset accurately represents a wide range of chemical properties relevant to medicinal chemists. The entire dataset was divided into four subsets based on experimental conditions, and an identical modelling protocol was applied to each: (i) MeOH-acid, (ii) MeOH-base, (iii) ACN-acid, and (iv) ACN-base.

Table 1 depicts summary statistics of retention times and selected molecular properties across the full dataset and each solvent condition. The dataset comprised 1,414 measurements across 535 unique molecules. In the dataset containing all four conditions, the mean retention time was 2.41 min (SD 0.73), the median was 2.32 min with an interquartile range (IQR) from 1.96 to 2.84 min, and values ranged from 1.07 to 5.57 min. In MeOH acidic (n = 474), the mean retention time was 2.08 min (SD 0.71), the median was 2.08 min with an IQR from 1.41 to 2.51 min, and the range was from 1.07 to 4.31 min. In MeOH basic (n = 470), the mean was 2.34 min (SD 0.61), the median was 2.22 min with an IQR from 1.89 to 2.70 min, and the range was from 1.44 to 4.32 min. In ACN acidic (n = 244), the mean was 2.71 min (SD 0.53), the median was 2.55 min with an IQR from 2.26 to 3.01 min, and the range was from 2.10 to 4.25 min. In ACN basic (n = 226), the mean was 2.95 min (SD 0.79), the median was 2.76 min with an IQR from 2.33 to 3.28 min, and the range was from 2.11 to 5.57 min. Molecular weight averaged 274.72 overall (SD 101.93), with solvent specific means of 275.86, 275.01, 263.39, and 283.98 for MeOH acidic, MeOH basic, ACN acidic, and ACN basic, respectively. LogP averaged 2.42 overall (SD 1.78), with solvent specific means of 2.23, 2.27, 2.55, and 3.00, respectively. Histograms of the retention time distributions under the four solvent conditions are provided in Supplementary Fig. S2.

3.1. Influence of molecular properties on retention time

To elucidate the influence of molecular properties on chromatographic behaviour, we examined the correlations between retention time (t_r) and two key molecular descriptors: molecular weight (MW) and lipophilicity (log*P*). The analyses were undertaken in all solvents to assess how solvent composition and pH affect these relationships (Supplementary Figs. S3 and S4).

Pearson correlation was used to quantify these relationships. As shown in Supplementary Fig. S5, the distributions of both MW and logP were approximately normal, with few outliers and no strong skew.

The reported correlations are calculated with Pearson's method [44]. The correlation between t_r and MW was negligible in all environments. The relationship between t_r and log*P* was negligible in all eluents except in basic methanol where it exhibited a slight positive correlation (r = 0.32). These patterns are consistent with differences in hydrogen bonding capacity. In acetonitrile, hydrogen bonds can form between the analyte and the CSP, whereas methanol suppresses this interaction [18,45,46].

3.2. MeOH-acid model

For the MeOH-acid model, we have developed the PLS model on 380

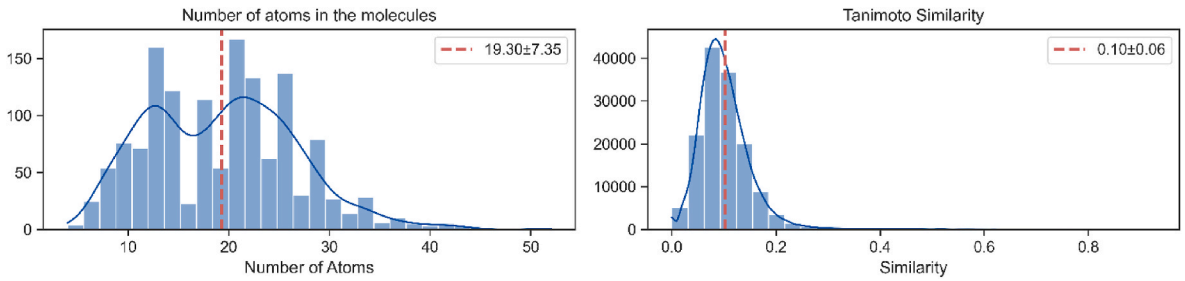


Fig. 1. Distribution of the number of atoms in the molecules (left side) and Tanimoto similarities (right-side).

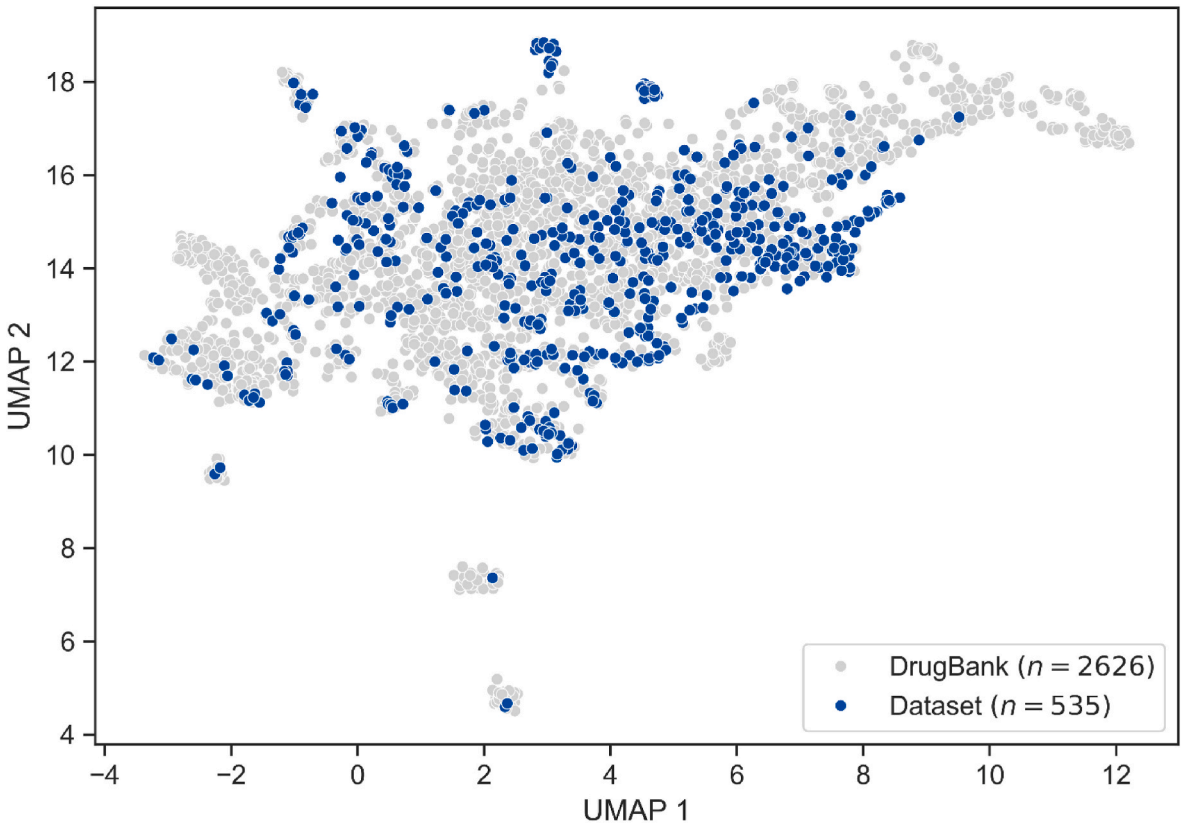


Fig. 2. UMAP visualization of molecular descriptor space: comparison of DrugBank approved drugs and experimental compounds. This UMAP projection demonstrates the chemical space coverage of experimental compounds (blue) relative to the DrugBank-approved drug molecules (gray), indicating that the dataset adequately spans the drug-like chemical space. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Table 1
Summary statistics of retention times and selected molecular properties across datasets.

Dataset	Number of measurements	Number of unique molecules	Train	Test	Retention Time (minutes)				Molecular Weight[a]	LogP[a]
					Mean (SD)	Median (IQR)	Min	Max	Mean (SD)	Mean (SD)
Total	1,414	535	1,140	274	2.41 (0.73)	2.32 (1.96–2.84)	1.07	5.57	274.72 (101.93)	2.42 (1.78)
MeOH-Acid	474	474	380	94	2.08 (0.71)	2.08 (1.41–2.51)	1.07	4.31	275.86 (107.81)	2.23 (1.86)
MeOH-Base	470	470	376	94	2.34 (0.61)	2.22 (1.89–2.70)	1.44	4.32	275.01 (105.33)	2.27 (1.81)
ACN-Acid	244	244	200	44	2.71 (0.53)	2.55 (2.26–3.01)	2.10	4.25	263.39 (94.90)	2.55 (1.60)
ACN-Base	226	226	184	42	2.95 (0.79)	2.76 (2.33–3.28)	2.11	5.57	283.98 (87.97)	3.00 (1.59)

[a] Predicted with RDKit 2023.9.4.

training and 94 test compounds. The covered retention time range of the model was between 1.00 and 4.32 min. Standardization (pre-scaling) was used both on the input variables (X) and target retention times (Y) before regression analysis, and five latent variables were included for the final model according to $RMSE_{CV}$ values of the cross-validation. Genetic algorithm has selected 262 variables out of 2723 with an initial population set of 100 in 100 iteration circles. The performance of the model is discussed in Supplementary Table S3 together with the results of NN and the consensus model.

The selected 262 variables were moved forward to the NN model building, utilizing three hidden nodes with an optimum neuron count for predicting retention time. In the optimization process the final number of neurons in the hidden layers were 1000, 518 and 100, respectively. The same training and test set was used as in the case of PLS, for consistency. Both the training and test set values were standardized, and all the validation protocols were consistent with those utilized during PLS modelling. Finally, the predicted retention time values from both models were used to estimate the average values in the consensus model. Supplementary Table S3 contains the performance parameters for the NN and consensus model as well. The permutation test proved that neither the PLS, nor the NN models are overfitted. As shown in Supplementary Table S3, the consensus model outperformed both the PLS and NN single models based on the R^2 and error metrics (especially in the test validation part). Even though the consensus model provided the most accurate predictions, the two single models are also acceptable. The predicted vs. measured retention time values based on the consensus model are plotted against each other in Fig. 3b.

For comparison we also included the GNN model with an R^2 of 0.51 for the CV set and 0.85 for the test set, considering the fact of the low (n = 478) sample number as presented in Fig. 4b.

3.3. MeOH-base model

In the case of the MeOH-base model, the dataset contained 376

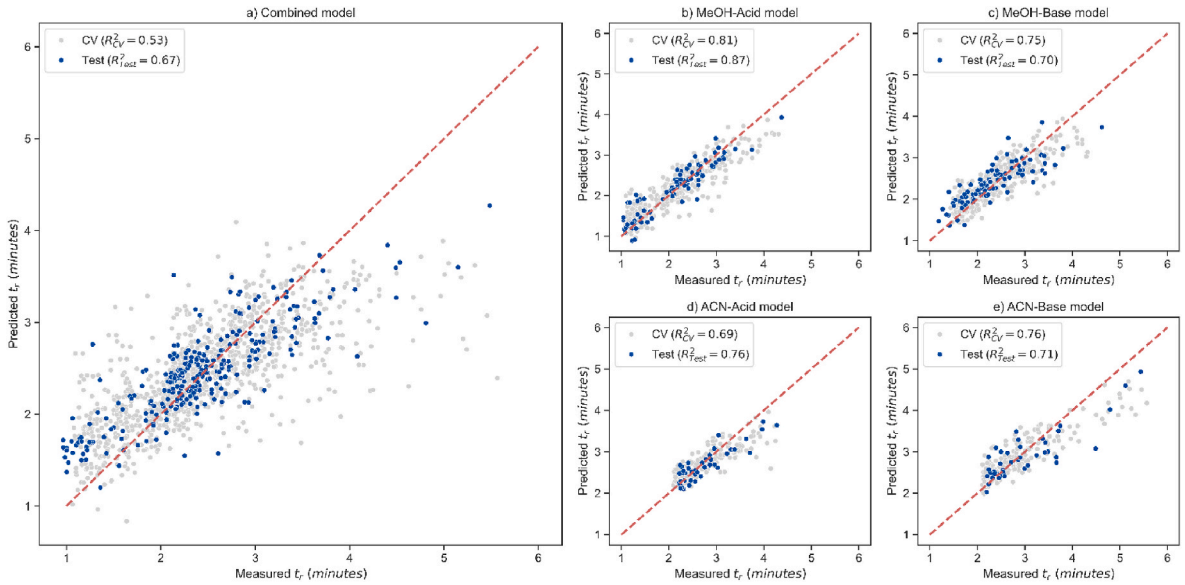


Fig. 3. Comparison of retention time predictions across datasets for the Consensus model. Scatter plots showing the predicted vs. measured retention times (tr) for the combined model and individual sub-datasets. Each subplot compares the cross-validation (CV) and test set performance for different solvent conditions. The red dashed line represents the ideal line where predicted tr equals measured tr. a) The combined model, which uses data from all solvent conditions, with R^2 values of 0.53 for the CV set and 0.67 for the test set. b) MeOH-Acid model exhibits an R^2 of 0.81 for the CV set and 0.87 for the test set c): MeOH-Base model shows slightly lower predictive performance with R^2 values of 0.75 for the CV set and 0.70 for the test set d) ACN-Acid model has similar predictive accuracy, with R^2 values of 0.69 for the CV set and 0.76 for the test set e) ACN-Base model also shows good predictive accuracy, with R^2 values of 0.76 for the CV set and 0.71 for the test set. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

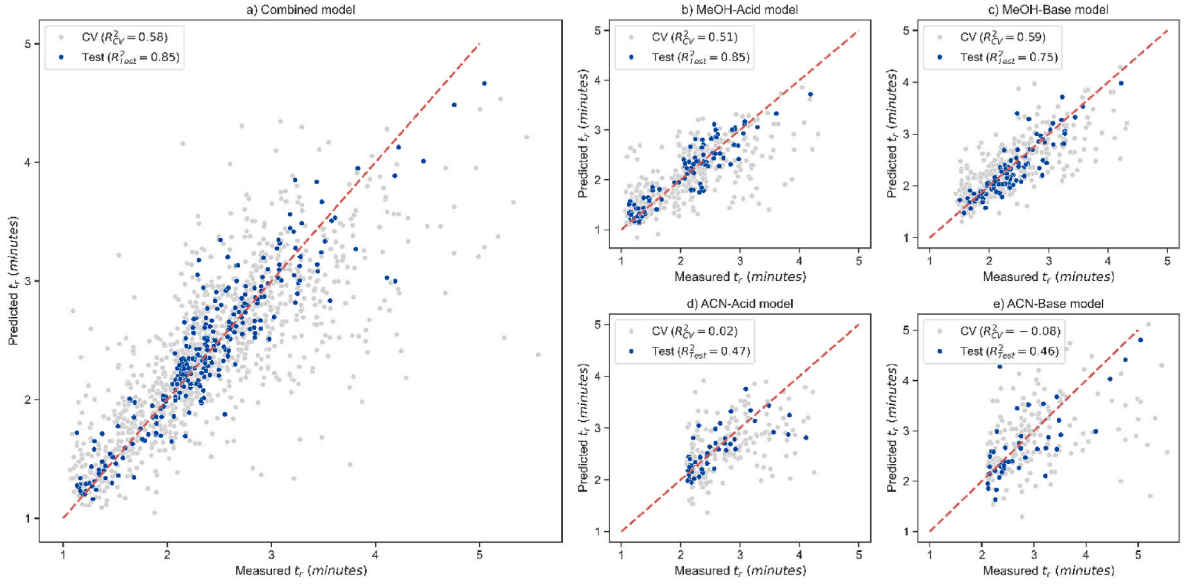


Fig. 4. Comparison of retention time predictions across datasets for the GNN model. Scatter plots showing the predicted vs. measured retention times (t_r) for the combined model and individual sub-datasets. Each subplot compares the cross-validation (CV) and test set performance for different solvent conditions. The red dashed line represents the ideal line where predicted t_r equals measured t_r. a) The combined model, which uses data from all solvent conditions, shows a strong correlation between predicted and measured t_r, with R^2 values of 0.58 for the CV set and 0.85 for the test set. b) MeOH-Acid model exhibits an R^2 of 0.51 for the CV set and 0.85 for the test set c): MeOH-Base model shows slightly lower predictive performance with R^2 values of 0.59 for the CV set and 0.75 for the test set d) ACN-Acid model has lower predictive accuracy, with R^2 values of 0.02 for the CV set and 0.47 for the test set e) ACN-Base model also shows an R^2 of −0.08 for the CV set and 0.46 for the test set. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

training compounds and 94 test compounds. The covered retention time range of the model was between 1.40 and 4.40 min. The same standardization (pre-scaling) was used as for PLS regression. For variable selection, we have selected 298 variables out of the 3833 based on the genetic algorithm. Five latent variables were selected in the final PLS model. The performance of the model is shown together with NN and the consensus model in Supplementary Table S3.

The selected 298 variables were used in the NN model as input data. The optimized number of neurons for the three hidden layers were 1000, 1000 and 100 respectively. The same standardization protocol and training-test split were used for modelling as in the case of PLS. Finally, the predicted retention time values were merged with the same data from PLS, and the average values were calculated for the consensus model. Both the PLS and NN models were additionally tested with a

permutation test, which proved that the models are not overfitted (Supplementary Figs. S6 and S7).

Here, the most accurate model is the consensus one, however these models are slightly worse than those based on the acid variant. On the other hand, the consensus model could still provide an R^2 value above 0.70 even for test validation. Finally, the predicted vs. measured retention time values based on the consensus model are plotted against each other in Fig. 3c.

The GNN model performed acceptably, with an R^2 of 0.59 for the CV set and 0.75 for the test set as presented in Fig. 4c.

3.4. ACN-acid model

The third consensus model was developed for the ACN-acid dataset. In this case 200 compounds were used as training set and 44 as test compounds. The range of the retention time was between 2.10 and 4.30 min. During variable selection, we have reduced the number of variables to 267 out of 2485. The final model contained 9 latent variables. The NN model was developed based on the selected 267 variables. In the hyperparameter optimization phase, the most optimal number of neurons were 769, 1000 and 100 for the three hidden layers respectively. After the same validation process, the predicted retention time values were used together with the PLS results for consensus modelling. The predicted values are plotted against the measured retention time data based on the consensus model in Fig. 3d. The performance of the single PLS, NN and the consensus model is summarized in Supplementary Table S3. The permutation test verified that neither the PLS, nor the NN models are overfitted.

While cross-validation related performances were better in the case of PLS, the consensus model could provide a much better performance in the test validation phase, thus we have selected that one as the most accurate. The NN model alone would not have been accurate enough for the prediction, but it was still useful to improve the consensus model especially in the case of test validation (Supplementary Figs. S6 and S7). The performance of the GNN model diminished notably, with an R^2 of 0.02 for the CV set and 0.47 for the test set as presented in Fig. 4d. This reiterates previous findings, as the GNN model's performance highly depends on the sample size (see MeOH-acid and -base models).

3.5. ACN-base model

Finally, a consensus model for the ACN-base dataset was also developed. Here, the dataset contained 184 compounds in the training set and 42 in the test set. The range of the retention time was 2.10 and 5.60 min. As previously, the dataset was standardized before modelling. We have selected 285 variables out of 3482 with the genetic algorithm. The PLS model contained nine LVs. The developed NN model contained 891, 988 and 525 neurons in the three hidden layers, which were optimized in a separate hyperparameter optimization loop. The permutation tests have verified that neither the PLS, not the NN models were overfitted. The predicted retention time values based on the PLS and NN models have been averaged in the consensus model. The performance of the models is summarized in Supplementary Table S3. In the case of test validation, the consensus model outperformed the single ones, while in the CV phase, the PLS model was slightly better. However, the two single models strengthen each other in the consensus variant. The final consensus model can predict the retention times with acceptable accuracy. The predicted and measured retention time values are plotted against each other for the consensus model in Fig. 3e.

The GNN model's performance continues to drastically underperform, with an R^2 of −0.08 for the CV set and 0.46 for the test set as presented in Fig. 4e. In summary, all the four consensus models are acceptable for the prediction of the retention times, however the MeOH-acid,MeOH-base and CAN-base models were slightly more accurate than the ACN-acid model.

3.6. Combined model

The performance of the combined model was thoroughly evaluated using PLS and NN based consensus and a GNN approach. In this case, the GNN model achieved higher R^2 and RMSE values compared to the consensus model: R^2 of 0.58 and $RMSE_{CV}$ of 0.49 during cross-validation and an impressive R^2 of 0.85 and $RMSE_{Test}$ of 0.25 on the hold-out test set. These results, depicted in Fig. 4a and Supplementary Table S3, encompass both the aggregated dataset and the individual sub-datasets, highlighting the model's robust performance across different data segments. It is well-established that the performance of deep learning models is highly dependent on the size and quality of the training dataset [47]. In this context, the GNN model's strong performance suggests that it effectively captures the complex, non-linear relationships among solvents, molecular structures, and retention times inherent in chromatographic analyses. This capability is particularly advantageous in modelling retention behaviour, where interactions can be highly intricate.

3.7. Elution order analysis

After evaluating the combined model, we examined whether the GNN could also predict the elution order of stereoisomers. This analysis provides a direct measure of the model's practical value in enantioseparation. In MeOH-acid, 58 stereoisomer pairs were identified, of which 32 were experimentally separated, corresponding to 55.17 %. The average measured difference in retention time for separated pairs was 0.13 ± 0.23 min, while the average predicted difference was 0.09 ± 0.06 min. The elution order was correctly predicted in 68.75 % of separated pairs.

In MeOH-base, 56 pairs were identified and 31 were separated, corresponding to 55.36 %. The average measured difference was 0.12 ± 0.18 min, the average predicted difference was 0.08 ± 0.06 min, and the correct elution order was achieved in 77.42 % of separated pairs.

In ACN-acid, 27 pairs were identified and 20 were separated, corresponding to 74.07 %. The average measured difference was 0.09 ± 0.11 min, the average predicted difference was 0.04 ± 0.03 min, and the correct elution order was obtained in 80.00 % of separated pairs.

In ACN-base, 31 pairs were identified and 22 were separated, corresponding to 70.97 %. The average measured difference was 0.09 ± 0.11 min, the average predicted difference was 0.06 ± 0.10 min, and the correct elution order was achieved in 81.82 % of separated pairs.

3.8. Web server implementation and usage

To facilitate the practical application of our GNN model, we developed a publicly accessible web server that allows users to predict retention times of molecules on a cellulose tris(3,5-dimethylphenylcarbamate) column under polar organic mode conditions. The web server, designed for user-friendliness and requiring minimal input, is intended to support researchers and practitioners in the field of chiral chromatography. The web server is freely available for research purposes at https://chiralscreen.com. Users can submit their compounds of interest by uploading a comma-separated values (CSV) file containing two columns: one for the compound name and one for the SMILES. The SMILES strings must be valid and accurately represent the correct stereochemistry of the chiral centres to ensure precise predictions. This straightforward input method allows users to efficiently process multiple compounds simultaneously. Upon submission, the server processes the input compounds using the trained GNN model and predicts the retention times for each compound under all four solvent conditions studied (MeOH-acid, MeOH-base, ACN-acid, ACN-base).

The predicted retention times are presented in a tabulated format, allowing users to easily interpret and compare the results across different solvent conditions. Additionally, the webserver provides the option to download the results as a CSV file, facilitating further analysis

or integration into the user's workflow. This web server is a valuable tool for various applications. In method development, it assists analytical chemists in predicting retention times of questionable compounds, potentially reducing the time and resources required for experimental trials. When differences exist between the predicted values of the molecules to be separated, chiral recognition ability is likely to be observed. An example of web server application is presented in Supplementary Fig. S8, illustrating the separation of racecadotril enantiomers. The Lux Cellulose-1 column exhibited enantioselectivity for racecadotril, with the highest enantioselective resolution observed in acidic methanol. Furthermore, the web server allows for easy prediction of the enantiomer elution order. It should be noted that the model is not limited to enantiomers; it is applicable to any type and number of compounds. It can be used for the separation of chiral and achiral compounds alongside each other, that is useful for impurity analysis. The enantiomeric elution order can also be easily deduced from the model, which can be especially useful when analysing racemic mixtures. The database was created using the Lux Cellulose-1 column, but our results can also be applied to columns from other manufacturers containing the same selector and can also be extrapolated for different column lengths. For compound screening, it could enable medicinal chemists to rapidly assess the chromatographic behaviour of large libraries of compounds, thereby facilitating high-throughput screening processes. Moreover, the web server can be utilized for educational purposes, demonstrating the practical application of machine learning in chromatographic prediction and enhancing the understanding of both chromatography and computational modelling.

3.8.1. Limitations

While the web server provides a practical and accessible platform for retention time prediction, users should be aware of its limitations. The accuracy of predictions is generally higher for compounds that are structurally like those included in the training dataset. As a result, compounds that fall outside the chemical space of the training data may yield less accurate predictions (Supplementary Fig. S9). It is important to emphasize that axial chirality cannot be specified using the SMILES code; hence, our model is inapplicable for predicting the separation of axially chiral compounds. Additionally, the predictions are currently limited to the specified CSP, the four solvent conditions and the chromatographic parameters mentioned in this study. Therefore, retention times under different chromatographic conditions are not available currently. It should also be noted that the constructed model predicts retention times and not retention factors. It is well-known, that retention time is highly dependent on the applied HPLC system, however the determination of dead time (t_0) on cellulose column in ACN-based or MeOH-based mobile phase was problematic and unnecessary for our work. We focused on predicting the difference between peaks rather than their absolute values. Future developments aim to enhance the utility of the web server. Plans include expanding the dataset by incorporating additional compounds to broaden the chemical space and improve the model's generalization capabilities. We also intend to extend the range of solvent systems and chromatographic parameters available for prediction, thereby increasing the applicability of the model to a wider array of chromatographic conditions. Enhancements to the user interface are also considered to improve usability and incorporate features based on user feedback.

3.9. Discussion

The comparative evaluation of the two modelling strategies—the condition-specific PLS-NN consensus and the global GNN approach—highlights a trade-off between predictive sharpness under narrowly defined circumstances and operational versatility in real-world settings.

Strategic model selection depends critically on analytical workflow requirements. The PLS–NN consensus approach exhibits superior performance precision under single mobile phase conditions, consistently achieving test-validation metrics of RMSEP <0.40 and $R^2_{test} > 0.70$. This performance profile establishes its utility for standardized analytical protocols requiring stringent validation criteria, particularly in method validation, impurity quantification, and regulatory compliance applications where analytical precision supersedes operational flexibility.

On the other hand, our GNN architecture is highly adaptable in multi-condition screening scenarios. In this case, the web-based implementation facilitates SMILES-to-retention conversion with parallel enantiomeric elution prediction, optimising high-throughput screening workflows where operational efficiency compensates for minor precision trade-offs.

Thus, the PLS-NN consensus model offers superior precision for single-condition optimization, whereas the GNN's holistic representation, coupled with web-based deployment, delivers broader applicability and immediate translational impact. Selecting between them should be guided by the analytical context: local accuracy versus multi-condition adaptability.

The elution order analysis further illustrates this distinction. The GNN achieved between 68 % and 82 % accuracy across the four solvent conditions, showing that it not only predicts retention times but also reproduces the relative order of stereoisomers with meaningful accuracy. This strengthens its role in exploratory and screening contexts, where knowledge of likely elution order can guide experimental design even if absolute retention predictions are less precise than those of the consensus models.

4. Conclusions

This study presents a large, homogeneous dataset containing retention of 535 structurally diverse molecules on the Lux Cellulose-1 CSP under four polar organic mode conditions (a total of 1,414 retention time measurements) and developed two machine learning strategies for retention time prediction. When modelling each condition separately, the PLS–NN consensus method generally provide the most accurate results, routinely achieving R^2 values over 0.70 and RMSEP values below 0.40 during test-validation. By contrast, in a unified model including all four conditions, the GNN method excelled with an R^2 of 0.58 and an $RMSE_{CV}$ of 0.49 in cross-validation, and an R^2 of 0.85 with an $RMSE_{Test}$ of 0.25 on the external test set. The resulting GNN model was implemented as an open-access web tool. Users can input SMILES strings with correct stereochemical annotations to predict retention times of new compounds under each polar organic mode and to approximate enantiomeric separation potential. This resource can optimize method development by minimizing the necessity for extensive experimental trial-and-error and assist in compound screening and impurity analysis. At the same time, several limitations warrant attention. Predictions are expected to be more reliable for molecules that lie within the broad, but still finite chemical space of our training set. Additionally, the predictions extend only to the four tested mobile-phase conditions and currently focus on retention time. Utilizing over a thousand experimentally obtained data points, we have developed an ML-generated web server, accessible free at www.chiralscreen.com. Our approach enables the accurate prediction of retention times on the Lux Cellulose-1 column, providing a reliable solution to the challenges associated with the unpredictability of chiral separations.

CRediT authorship contribution statement

Attila Imre: Writing – review & editing, Writing – original draft, Project administration, Methodology, Data curation, Conceptualization. **Gergely Dombi:** Writing – review & editing, Writing – original draft, Software, Resources, Investigation, Funding acquisition, Conceptualization. **Máté Dobó:** Writing – review & editing, Project administration, Methodology, Investigation, Conceptualization. **Ali Mhammad:** Validation, Software, Methodology, Investigation. **Elek Ferencz:** Validation, Project administration, Methodology, Investigation, Formal

analysis, Data curation. **Balázs Balogh:** Supervision, Software, Resources, Data curation. **Anna Vincze:** Project administration, Methodology, Formal analysis. **Zoltán-István Szabó:** Writing – original draft, Project administration, Funding acquisition, Conceptualization. **György Tibor Balogh:** Writing – review & editing, Writing – original draft, Validation, Methodology, Funding acquisition, Formal analysis, Conceptualization. **Anita Rácz:** Writing – original draft, Validation, Software, Methodology, Funding acquisition, Formal analysis, Conceptualization. **Gergő Tóth:** Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Conceptualization.

Funding

Supported by the 2024–2.1.1-EKÖP-2024-00004 and 2025–2.1.1-EKÖP-2025-00014 University Research Scholarship Programme of the Ministry for Culture and Innovation from the Source of the National Research, Development and Innovation Fund (G.D.).

This work was funded by the National Research, Development, and Innovation Office, Hungary (grant: NKFIH FK 146930).

This work was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences (G.T. and A.R.).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.aca.2025.344733.

Data availability

Data will be made available on request.

References

[1] R.U. McVicker, N.M. O'Boyle, Chirality of new drug approvals (2013-2022): trends and perspectives, J. Med. Chem. 67 (2024) 2305–2320.
[2] K. Barman, M.M. Islam, K.S. Das, N. Singh, S. Priya, B.D. Kurmi, P. Patel, Recent advances in enantiorecognition and enantioseparation techniques of chiral molecules in the pharmaceutical field, Biomed. Chromatogr. 39 (2025) e6073.
[3] C. De Luca, S. Felletti, F.A. Franchina, D. Bozza, G. Compagnin, C. Nosengo, L. Pasti, A. Cavazzini, M. Catani, Recent developments in the high-throughput separation of biologically active chiral compounds via high performance liquid chromatography, J. Pharm. Biomed. Anal. 238 (2024) 115794.
[4] L.A. Papp, Z.I. Szabo, G. Hancu, L. Farczadi, E. Mircia, Comprehensive review on chiral stationary phases in single-column simultaneous chiral-achiral HPLC separation methods, Molecules 29 (2024).
[5] A. Tarafder, L. Miller, Chiral chromatography method screening strategies: past, present and future, J. Chromatogr. A 1638 (2021) 461878.
[6] G.K.E. Scriba, Update on chiral recognition mechanisms in separation science, J. Separ. Sci. 47 (2024) e2400148.
[7] P. De Gauquier, J. Peeters, K. Vanommeslaeghe, Y. Vander Heyden, D. Mangelings, Modelling the enantiorecognition of structurally diverse pharmaceuticals on O-substituted polysaccharide-based stationary phases, Talanta 259 (2023) 124497.
[8] J. Teixeira, M.E. Tiritan, M.M.M. Pinto, C. Fernandes, Chiral stationary phases for liquid chromatography: recent developments, Molecules 24 (2019).
[9] I. Varfaj, M. Labikova, R. Sardella, H. Hettegger, W. Lindner, M. Kohout, A. Carotti, A journey in unraveling the enantiorecognition mechanism of 3,5-dinitrobenzoyl-amino acids with two cinchona alkaloid-based chiral stationary phases: the power of molecular dynamic simulations, Anal. Chim. Acta 1314 (2024) 342791.
[10] Y. Fan, Y. Deng, Y. Yang, X. Deng, Q. Li, B. Xu, J. Pan, S. Liu, Y. Kong, C.-E. Chen, Modelling and predicting liquid chromatography retention time for PFAS with no-code machine learning, Environ. Sci. Adv. 3 (2024) 198–207.
[11] M. Perez-Baeza, Y. Martin-Biosca, L. Escuder-Gilabert, M.J. Medina-Hernandez, S. Sagrado, Artificial neural networks to model the enantioresolution of structurally unrelated neutral and basic compounds with cellulose tris(3,5-dimethylphenylcarbamate) chiral stationary phase and aqueous-acetonitrile mobile phases, J. Chromatogr. A 1672 (2022) 463048.
[12] Y.R. Singh, D.B. Shah, M. Kulkarni, S.R. Patel, D.G. Maheshwari, J.S. Shah, S. Shah, Current trends in chromatographic prediction using artificial intelligence and machine learning, Anal. Methods 15 (2023) 2785–2797.
[13] L. Xin, H. Yu, S. Liu, G.-G. Ying, C.-E. Chen, POPs identification using simple low-code machine learning, Sci. Total Environ. 921 (2024) 171143.
[14] H. Xu, J. Lin, D. Zhang, F. Mo, Retention time prediction for chromatographic enantioseparation by quantile geometry-enhanced graph neural network, Nat. Commun. 14 (2023) 3095.
[15] A. Aranyi, I. Ilisz, Z. Pataj, I. Szatmari, F. Fulop, A. Peter, High-performance liquid chromatographic enantioseparation of 1-(phenylethylamino)- or 1-(naphthylethylamino)methyl-2-naphthol analogs and a temperature-induced inversion of the elution sequence on polysaccharide-based chiral stationary phases, J. Chromatogr. A 1218 (2011) 4869–4876.
[16] S. Khater, Y. Zhang, C. West, In-depth characterization of six cellulose tris-(3,5-dimethylphenylcarbamate) chiral stationary phases in supercritical fluid chromatography, J. Chromatogr. A 1303 (2013) 83–93.
[17] Z.I. Szabo, A. Bartalis-Fabian, G. Toth, Simultaneous determination of escitalopram impurities including the R-enantiomer on a cellulose tris(3,5-Dimethylphenylcarbamate)-Based chiral column in reversed-phase mode, Molecules 27 (2022).
[18] M. Dobo, M. Foroughbakhshfasaei, P. Horvath, Z.I. Szabo, G. Toth, Chiral separation of oxazolidinone analogues by liquid chromatography on polysaccharide stationary phases using polar organic mode, J. Chromatogr. A 1662 (2022) 462741.
[19] D. Weininger, Smiles, a chemical language and information-system .1. introduction to methodology and encoding rules, J. Chem. Inf. Comput. Sci. 28 (1988) 31–36.
[20] G. Landrum, P. Tosco, B. Kelley, Ric, D. Cosgrove, sriniker, gedeck, R. Vianello, N. Schneider, E. Kawashima, G. Jones, N. Dan, A. Dalke, B. Cole, M. Swain, S. Turk, A. Savelyev, A. Vaucher, M. Wójcikowski, I. Take, V.F. Scalfani, D. Probst, K. Ujihara, G. Godin, A. Pahl, R. Walker, J. Lehtivarjo, F. Berenger, jasondbiggs, strets12,rdkit/rdkit: 2023_09_4 (Q3 2023) Release, Zenodo, 2024.
[21] LigPrep, LigPrep 2019-4, Schrödinger, Schrödinger, LLC, New York, NY, 2019.
[22] A. Mauri, V. Consonni, M. Pavan, R. Todeschini, Dragon software: an easy approach to molecular descriptor calculations, Match-Commun. Math Co. 56 (2006) 237–248.
[23] A. Racz, D. Bajusz, K. Heberger, Intercorrelation limits in molecular descriptor preselection for QSAR/QSPR, Mol.Inform. 38 (2019) e1800154.
[24] P. Geladi, B.R. Kowalski, Partial least-squares regression - a tutorial, Anal. Chim. Acta 185 (1986) 1–17.
[25] R. Hecht-Nielsen, III.3 - theory of the backpropagation neural network**based on "nonindent" by Robert Hecht-Nielsen, which appeared in proceedings of the international joint conference on neural networks 1, 593–611, June 1989. © 1989 IEEE, in: H. Wechsler (Ed.) Neural Networks for Perception, Academic Press1992, pp. 65-93.
[26] R. Leardi, Genetic algorithms in chemistry, J. Chromatogr. A 1158 (2007) 226–233.
[27] M. Belkin, D. Hsu, S. Ma, S. Mandal, Reconciling modern machine-learning practice and the classical bias-variance trade-off, Proc. Natl. Acad. Sci. U. S. A. 116 (2019) 15849–15854.
[28] A. Racz, D. Bajusz, K. Heberger, Effect of dataset size and Train/Test split ratios in QSAR/QSPR multiclass classification, Molecules 26 (2021).
[29] F. Scarselli, M. Gori, A.C. Tsoi, M. Hagenbuchner, G. Monfardini, The graph neural network model, IEEE Trans. Neural Network. 20 (2009) 61–80.
[30] J. Zhou, G.Q. Cui, S.D. Hu, Z.Y. Zhang, C. Yang, Z.Y. Liu, L.F. Wang, C.C. Li, M. S. Sun, Graph neural networks: a review of methods and applications, Ai Open 1 (2020) 57–81.
[31] A. Imre, B. Balogh, I. Mandity, GraphCPP: the new state-of-the-art method for cell-penetrating peptide prediction via graph neural networks, Br. J. Pharmacol. 182 (2025) 495–509.
[32] B. Ramsundar, P. Eastman, P. Walters, V. Pande. Deep Learning for the Life Sciences, O'Reilly Media, Inc, 2019.
[33] S. Kearnes, K. McCloskey, M. Berndl, V. Pande, P. Riley, Molecular graph convolutions: moving beyond fingerprints, J. Comput. Aided Mol. Des. 30 (2016) 595–608.
[34] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, PyTorch: an imperative style, high-performance deep learning library. Proceedings of the 33rd International Conference on Neural Information Processing Systems, Curran Associates Inc., 2019. Article 721.
[35] M. Fey, J.E. Lenssen, Fast graph representation learning with PyTorch geometric, ArXiv, abs/1903 (2019) 02428.
[36] P. Moritz, R. Nishihara, S. Wang, A. Tumanov, R. Liaw, E. Liang, W. Paul, M. I. Jordan, I. Stoica, Ray: a distributed framework for emerging AI applications, ArXiv, abs/1712 (2017) 05889.
[37] J. Bergstra, D. Yamins, D. Cox, Making a science of model search: hyperparameter optimization in hundreds of dimensions for vision architectures. International Conference on Machine Learning, PMLR, 2013, pp. 115–123.
[38] A. Chen, A. Chow, A. Davidson, A. Dcunha, A. Ghodsi, S.A. Hong, A. Konwinski, C. Mewald, S. Murching, T. Nykodym, P. Ogilvie, M. Parkhe, A. Singh, F. Xie, M. Zaharia, R. Zang, J. Zheng, C. Zumar, Developments in MLflow: a system to accelerate the machine learning lifecycle. Proceedings of the Fourth International Workshop on Data Management for end-to-end Machine Learning, Association for Computing Machinery, Portland, OR, USA, 2020. Article 5.
[39] I. Loshchilov, F. Hutter, SGDR: stochastic gradient descent with warm restarts, arXiv: Learning (2016).

[40] Dash, 2024.
[41] D. Merkel, Docker: lightweight linux containers for consistent development and deployment, Linux J. 239 (2014) 2.
[42] D. Bajusz, A. Racz, K. Heberger, Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? J. Cheminf. 7 (2015) 20.
[43] L. McInnes, J. Healy, J. Melville, Umap: Uniform Manifold Approximation and Projection for Dimension Reduction, 2018 arXiv preprint arXiv:1802.03426.
[44] J. Lee Rodgers, W.A. Nicewander, Thirteen ways to look at the correlation coefficient, Am. Statistician 42 (1988) 59–66.
[45] R.B. Kasat, S.Y. Wee, J.X. Loh, N.H. Wang, E.I. Franses, Effect of the solute molecular structure on its enantioresolution on cellulose tris(3,5-dimethylphenylcarbamate), J. Chromatogr., B: Anal. Technol. Biomed. Life Sci. 875 (2008) 81–92.
[46] M. Lammerhofer, Chiral recognition by enantioselective liquid chromatography: mechanisms and modern chiral stationary phases, J. Chromatogr. A 1217 (2010) 814–856.
[47] Y.B. Ian Goodfellow, Aaron Courville, Deep Learning, the MIT Press, 2016.