Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Digital Discovery*, 2025, **4**, 2465

Multimodal learning in synthetic chemistry applications: gas chromatography retention time prediction and isomer separation optimization

Jinglong Lin,†[a] Longyin Song,†[b] Yuntian Chen,[c] Chengchun Liu,[a] Shufeng Chen *[b] and Fanyang Mo *[ade]

Multimodal learning, a key machine learning (ML) approach, has been extensively applied in fields such as medical diagnostics and recommendation systems. The complexity of chemical data offers unique opportunities for multimodal learning, though its application in chemistry remains underexplored. Here, we propose an innovative multimodal framework for gas chromatography (GC) that integrates a geometry-enhanced graph isomorphism network and gated recurrent units. This framework predicts GC retention time across diverse molecular heating profiles with a test set R^2 of 0.995, outperforming traditional ML methods. It effectively recommends optimal chromatographic conditions for separating positional isomers and *cis*/*trans* isomers, minimizing experimental iterations and significantly improving analytical efficiency. Moreover, the model provides insights into the separation challenges of various isomers, enhancing understanding of the relationship between molecular structure and chromatographic behavior. This approach could pave the way for broader applications of multimodal learning in chemistry.

Received 15th November 2024
Accepted 27th July 2025

DOI: 10.1039/d4dd00369a

rsc.li/digitaldiscovery

1 Introduction

Machine Learning (ML) technology is a technique that enables computers to learn from data and make decisions or predictions.[1,2] Multimodal learning, an ML approach that integrates data from different modalities such as text, images, and audio, has been widely applied in fields such as visual question answering,[3] emotion analysis,[4] and recommendation systems.[5] However, its application in the field of chemistry remains relatively limited. In fact, multimodal learning is particularly well-suited for addressing chemical problems due to the inherent diversity and complexity of chemical data. Chemical research involves various types of data, including molecular structures, spectroscopic data (*e.g.*, infrared and nuclear magnetic resonance spectra), textual data (*e.g.*, chemical equations and literature descriptions), image data (*e.g.*, microscopic crystal structures), and numerical data from experimental conditions (*e.g.*, temperature, pressure, concentration). By integrating diverse data modalities, multimodal learning enhances model generalizability and provides deeper insights into chemical mechanisms for more precise interpretations. This study introduces an innovative multimodal learning framework for predicting gas chromatography (GC) retention time (RT) and optimizing chromatographic conditions, serving as a concrete example of multimodal learning applications in chemical research.

GC is a vital technique for separating and analyzing complex mixtures,[6–8] now extensively used in fields such as chemical analysis,[9] pharmaceutical testing,[10] and environmental monitoring.[11] The RT in GC, referring to the duration a specific component stays within the chromatographic column, is one of the key parameters for compound identification. Researchers have proposed various theories,[12] including thermodynamic and kinetic models, to explain analyte behavior in stationary and mobile phases and to account for RT variations under different chromatographic conditions. However, these theories rely on idealized assumptions and simplified conditions, making them challenging to apply in practical research. Therefore, there is an urgent need to develop predictive models for GC-RT to rapidly determine chromatographic conditions and enhance analytical efficiency. Recently, researchers have used ML to predict GC retention indices (GC-RI)[13–16] and have combined ML with techniques like Gas Chromatography-Mass Spectrometry (GC-MS)[17–19] and Gas Chromatography-Ion Mobility Spectrometry (GC-IMS).[20] However, there is currently a lack of predictive models for GC-RT. Our group has primarily

[a]*School of Materials Science and Engineering, Peking University, Beijing, 100871, P. R. China. E-mail: fmo@pku.edu.cn*

[b]*School of Chemistry and Chemical Engineering, Inner Mongolia University, Hohhot, 010021, P. R. China. E-mail: shufengchen@imu.edu.cn*

[c]*Ningbo Institute of Digital Twin, Eastern Institute of Technology, Ningbo, Zhejiang 315200, P. R. China*

[d]*School of Advanced Materials, Peking University Shenzhen Graduate School, Shenzhen, 518055, China*

[e]*AI for Science (AI4S)-Preferred Program, Peking University Shenzhen Graduate School, Shenzhen, 518055, China*

† These authors contributed equally.

focused on predicting RT in thin-layer chromatography (TLC)[21] and high-performance liquid chromatography (HPLC).[22]

The factors influencing GC-RT include molecular properties, temperature conditions, and stationary phases (Fig. 1a). The multimodal learning framework is particularly suited to address these complexities. The multimodal framework developed in this work can be utilized for rapid virtual screening and recommendation under GC temperature programming conditions, identification of internal standard peak positions, and analysis of complex mixtures (Fig. 1b). Our contributions can be summarized as follows:

1. We established a Gas Chromatography Retention Time (GCRT) dataset, covering 250 compounds, 244 types of temperature programs, and 3950 retention time data. We developed a threshold filtering method, GC-PIEA, that can extract target peak information from GC chromatogram data in PDF format in batches automatically.

2. We constructed three conventional ML models and explored model interpretability through correlation and feature importance analysis, thereby validating and uncovering physicochemical principles affecting GC-RT from a statistical perspective.

3. We constructed a multimodal model where a geometry-enhanced graph isomorphism network processes molecular information and a Bidirectional GRU handles temperature curve data. This model can accurately predict the GC-RT of target molecules under different temperature programs, with a test set R^2 reaching up to 0.995, surpassing conventional ML models. The multimodal model demonstrates substantial robustness and generalization capabilities. Even with Gaussian noise at 40%, the R^2 value on the test set remains 0.906. Furthermore, the model achieves an R^2 of 0.900 in predicting entirely novel compounds under nonlinear temperature programs.

4. We introduced the concept of the isomer separation degree, enabling our multimodal model to rapidly perform virtual screening. This approach identifies optimal temperature programs for positional and *cis*/*trans* isomers, ensuring the fastest detection peaks while maintaining separation. Search algorithms can be utilized to explore the separation challenges of various isomers, providing new chemical insights and enhancing the understanding of the relationship between molecular structure and chromatographic behavior. This method advances chromatographic research,

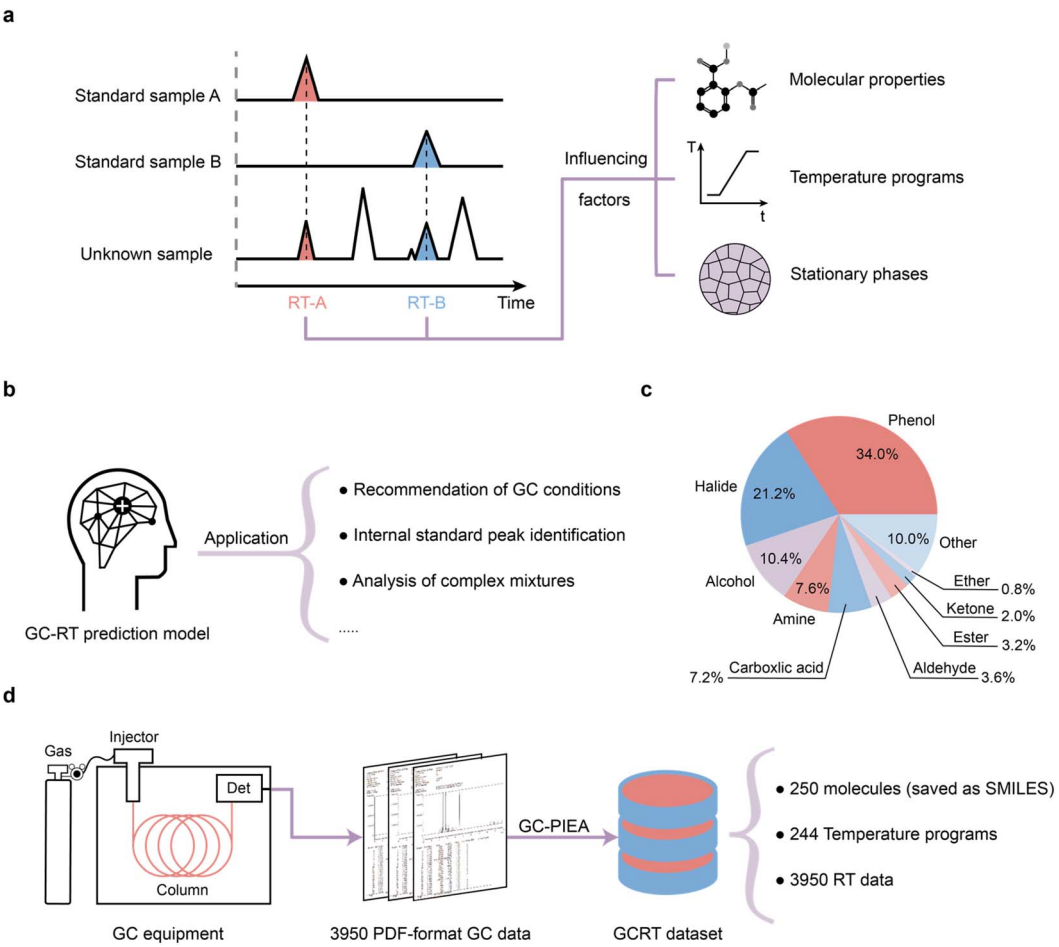

Fig. 1 The scheme for predicting RT in GC. (a) Factors influencing the GC-RT. b, Applications of GC-RT prediction model. (c) Distribution chart of compound types in the GCRT dataset. (d) Workflow for constructing the GC-RT dataset. GC-PIEA, Gas Chromatography Peak Information Extraction Algorithm.

allowing other scientists to apply it to their isomer studies of interest.

5. We proposed a strategy that utilizes GC-RI and temperature programs to predict GC-RT, and provided an ANN pre-trained model. Other researchers can fine-tune this model on their own tasks, thereby achieving universal RT prediction.

2 Results and discussion

2.1 Experiment and GCRT dataset

Gas chromatographic analysis was performed using a Shimadzu GC-2014C (Shimadzu Corporation, Suzhou, Jiangsu, China). Chromatographic separation was carried out on a 30 m × 0.25 mm WondaCAP-5 capillary column, coated internally with 0.25 μm of 5% phenyl–95% dimethylpolysiloxane stationary phase. Nitrogen gas was chosen as the carrier gas, with a constant flow rate set at 1.2 mL min^{-1}. A total of 244 different column temperature programs were used, which are provided in the SI Material 'Temperature_programs.csv'. The dual flame ionization detector (DFID) was operated with hydrogen at 40 mL min^{-1}, air at 400 mL min^{-1}, and a tail gas flow rate of 30 mL min^{-1}, with the detector temperature set at 300 °C to ensure excellent ionization efficiency. All data were processed and analyzed using the GC Solution software provided by Shimadzu Corporation.

The sample preparation process was precise and meticulous. Initially, 0.5 mmol of the sample was dissolved in 1 mL of ethyl acetate. Then, approximately 0.05 mL of this solution was extracted using a disposable pipette and dropped into a dedicated sample vial, which was subsequently diluted to the mark with 1 mL of ethyl acetate. The sample was injected in a volume of 1 μL, employing a 20 : 1 split injection mode, with the injector temperature set at 250 °C.

The GC chromatogram data obtained using specific devices and methods have been saved in PDF format. To efficiently extract and batch-process key information from these PDF-format chromatograms, we have developed a new algorithm: Gas Chromatography Peak Information Extraction Algorithm (GC-PIEA), as detailed in Section 1 of the SI Materials. This algorithm first converts the text in the chromatograms into strings, then filters data by setting a threshold for peak area, thereby accurately locating the position of the target peak and extracting key parameters such as RT, peak height, and peak area.

The temperature program is originally a time-series data set (T–t). We have processed the temperature programs of all data points into series of 32 time steps each, corresponding to the temperatures for 1 to 32 minutes. Furthermore, to enhance the data's usability, molecular information is stored in SMILES format. Based on these methods, we have constructed the Gas Chromatography Retention Time (GCRT) dataset. This dataset comprises 3950 RT data points for 250 different compounds under various temperature programs, with the distribution of compounds illustrated in Fig. 1c and the dataset construction process shown in Fig. 1d.

2.2 Construction of conventional ML models and model interpretability research

To compare with the multimodal model and explore model interpretability, we constructed three conventional ML algorithms as baselines, namely Random Forest (RF), LightGBM (LGB), and Artificial Neural Network (ANN). In the RF and LGB models, we adopted a random split for the training-to-test set ratio of 9 : 1 and utilized 10-fold cross-validation on the training set to determine the optimal hyperparameters. For the ANN model, we designed a structure with two hidden layers and incorporated batch normalization and dropout techniques (see Section 2 of the SI Materials for details), setting a random split for the training, validation, and test set ratio of 8 : 1 : 1. The model's input features total ten dimensions (Fig. 2a), six of which are based on the physicochemical properties of molecules, covering aspects such as polarity and hydrophobicity. These features include Molecular Weight (MW), Topological Polar Surface Area (TPSA), Number of Hydrogen Bond Acceptors (HBA), Number of Hydrogen Bond Donors (HBD), Number of Rotatable Bonds (NROTB), and lipid-water partition coefficient (Log P). These features were selected due to their potential correlation with the compounds' RT in GC. For instance, MW influences boiling points, thereby affecting the migration speed and volatility of compounds within the chromatographic column; Log P is a key descriptor for assessing the hydrophobicity of compounds, which influences their hydrophobic interactions with non-polar stationary phases. All these descriptors were calculated using the RDKit library in Python. The remaining four dimensions are temperature program features, including initial temperature, final temperature, heating rate, and duration of initial temperature, all of which directly impact the analysis outcome in GC. In this study, the efficacy of feature engineering was corroborated from a modeling perspective, as evidenced by the exemplary performance on the test set of three ML models (Fig. 2b): RF ($R^2 = 0.950$), LGB ($R^2 = 0.919$), and ANN ($R^2 = 0.925$).

To further investigate the predictive performance of conventional ML models on unknown compounds, we employed the compound split method (Fig. S1). The compound split method refers to dividing a dataset based on the SMILES of compounds, ensuring that compounds in the test set are never seen in the training set. This ensures that the model's evaluation is based on entirely novel chemical entities, enhancing the robustness and generalizability of the results. For the applications of RF and LGB, we set the ratio of training to test sets at 9 : 1; for the ANN, the data were divided into training, validation, and test set with respective ratios of 8 : 1 : 1. The models' performance on the test sets can be found in Section 4 of the SI Materials, under Fig. S2, where the R^2 values are 0.878 for RF, 0.885 for LGB, and 0.852 for ANN.

In our investigation into model interpretability, we began by computing the feature correlation matrix (Fig. 2c), uncovering significant positive correlations between HBA, HBD, and TPSA, as well as between Log P and MW. Furthermore, the evident negative correlations between TPSA, HBA, and Log P, along with near-zero correlation coefficients among other features, highlighted the independence among features. In the following

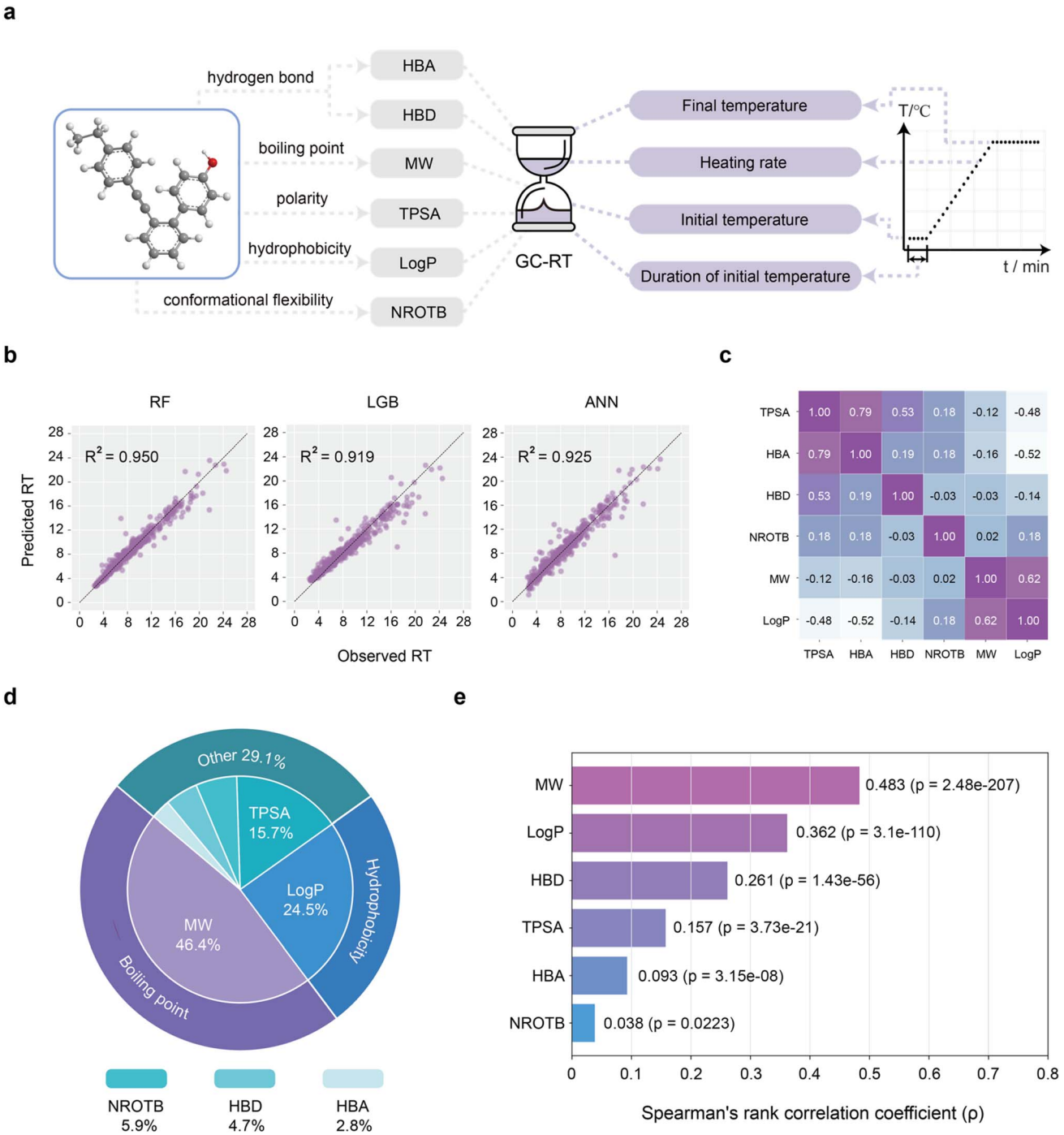

Fig. 2 Conventional ML performance and model interpretability study. (a) Feature engineering schema involving 10 dimensions, comprising 6 molecular descriptors and 4 temperature program features. MW, molecular weight; TPSA, topological polar surface area; HBA, number of hydrogen bond acceptors; HBD, number of hydrogen bond donors; NROTB, number of rotatable bonds; Log *P*, lipid-water partition coefficient. (b) Predictive performance of three ML models on the test set. (c) Correlation matrix for molecular descriptors. (d) Normalized importance of molecular descriptors (based on the GCRT dataset and RF model with SHAP value analysis). (e) Correlation analysis between molecular descriptors and RT, considering a moderate positive correlation when $\rho > 0.3$ and the conclusion significant at p-value < 0.001.

analysis, we focus on the assessment of feature importance. Given the superior performance of the RF model among the three conventional ML approaches, we employed the GCRT dataset and RF model to compute the mean absolute SHapley Additive exPlanations (SHAP) values of various molecular descriptors, thereby quantifying the importance of each feature. The normalized importance ratios of these features are illustrated in Fig. 2d. The concept of SHAP values originates from Shapley values in game theory, which are a cooperative game theory tool designed to fairly allocate the contributions of each

participant to a collective outcome. In this study, each molecular descriptor was considered a "player". The definition of Shapley values is as follows:

$$\phi_i(v) = \sum_{S \subseteq N/\{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} (v(S \cup \{i\}) - v(S)) \quad (1)$$

where N represents the set of all features, S denotes any subset of features that does not include feature i, $v(S)$ is the contribution of the subset S within the model (*i.e.*, the model's predictive output), and ϕ_i is the Shapley value of feature i, which quantifies its average marginal contribution to the model's output. The results indicate that MW (46.4%), Log P (24.5%), and TPSA (15.7%) are the three features with the most significant impact on the RT. This finding underscores the key role of a molecule's molecular weight, hydrophobicity, and polarity in influencing RT. Subsequently, we conducted an analysis of the correlation between molecular descriptors and RT, specifically using the Spearman rank correlation coefficient (ρ) to measure the relationship between each feature and RT (Fig. 2e). The ρ value ranges from [-1, 1], where $\rho > 0$ indicates a positive correlation, and $\rho < 0$ indicates a negative correlation. The analysis reveals that MW and Log P show a moderate positive correlation with RT ($\rho > 0.3$), and their p-values are significantly less than 0.001, lending statistical significance to our findings and further validating the reliability of our discoveries. All these findings are derived entirely from a data-driven analysis.

We provided a chemical rationale for these findings: molecules with higher MW typically exhibit higher boiling points due to stronger van der Waals forces among them, requiring more energy to overcome these forces and transition into the gas phase. This implies that compounds with larger MW tend to linger longer in the stationary phase, as they are more challenging to be transported by the carrier gas, thereby resulting in extended RT. The study utilized a 5% phenyl–95% dimethylpolysiloxane chromatography column, offering an analytical environment ranging from nonpolar to medium polarity. Molecules with higher Log P values, due to their stronger hydrophobicity, are more likely to interact with the stationary phase, leading to an increase in RT. The positive correlation between Log P and MW (correlation coefficient of 0.62) further reveals the synergistic effect of Log P and MW in prolonging the RT of molecules.

2.3 Construction of the multimodal model

Multimodal learning is a method that combines and analyzes data from different modalities, such as text, images, and sound, to enhance learning effectiveness and understanding. In our task, which involves predicting the RT of molecules under various temperature programs, molecular data and temperature program data belong to two distinct modalities. Therefore, we have developed a multimodal model (Fig. 3b) to process and integrate information from these two modalities.

Molecules are naturally represented as graphs, with atoms as nodes and chemical bonds as edges, making them particularly suited for processing with graph neural networks. Previous researchers have conducted extensive work in the field of molecular graph representation learning.[23–26] In the molecular processing branch, we utilize a Geometry-enhanced Graph Isomorphism Network (GeoGIN) framework. Two graphs comprehensively represent molecular information: the atom-bond graph, showing the connections between atoms, and the bond-angle graph, which includes information on the molecule's 3D conformation. The embeddings of nodes and edges for the atom-bond graph and the bond-angle graph are illustrated in Fig. 3a. In the atom-bond graph, node features include formal charge, implicit valence, hybridization, and aromaticity of atoms, among others, while edge features encompass chemical bond types and ring presence, among others. In the bond-angle graph, in addition to calculating bond angles, edge features also embed molecular descriptor information, such as the TPSA and Relative Polar Surface Area (RPSA), merging spatial information with physicochemical properties. Given that chromatographic RT prediction is a structure-sensitive task, we used the Graph Isomorphism Network (GIN) as the molecular encoder to extract features of the graph structure. The key idea of GIN is to distinguish different molecular graph structures more precisely through its message passing and aggregation mechanism, even when these structures appear similar in conventional graph neural networks (like GCN or GAT). In our designed architecture, both atom-bond and bond-angle graphs are processed deeply using five-layer GIN. Through this approach, we have conducted meticulous feature extraction for each graph. Upon completion, sum pooling is employed to aggregate the extracted features, obtaining the embedding representation of the graphs. Subsequently, the embedding representations of the atom-bond and bond-angle graphs are fused to generate a 128-dimensional feature vector. This feature vector accurately captures the structural information and chemical properties of the molecule, laying a solid foundation for subsequent analysis and applications.

In the experimental conditions processing branch, we employ a Bidirectional GRU (Bi-GRU) to handle the temperature programs. Bi-GRU, a lightweight recurrent neural network, is highly effective for extracting series features due to its ability to capture temporal dependencies in data. In our study, by applying average pooling operations along the temporal dimension to the hidden state vectors extracted by Bi-GRU, we successfully obtained a vectorized representation of the temperature program. Subsequently, we employed an attention-weighted fusion strategy that dynamically determines the feature weights of two separate branches: the feature weight for the Bi-GRU branch is denoted as ω_1, while the feature weight for the GeoGIN branch is represented as ω_2. Through this method, we obtained a 128-dimensional fused feature vector that integrates information from both the temperature program and molecular graphs. Considering the differences in feature distributions extracted by Bi-GRU and GeoGIN, we utilized two residual blocks to further enhance the fusion of these two types of features. The use of residual blocks aims to reduce information loss while increasing the model's adaptability to and integration of feature differences. This step ensures that the fused features more comprehensively represent the attributes of the input data. Finally, these meticulously fused features are fed into a fully connected neural network. The output layer of this

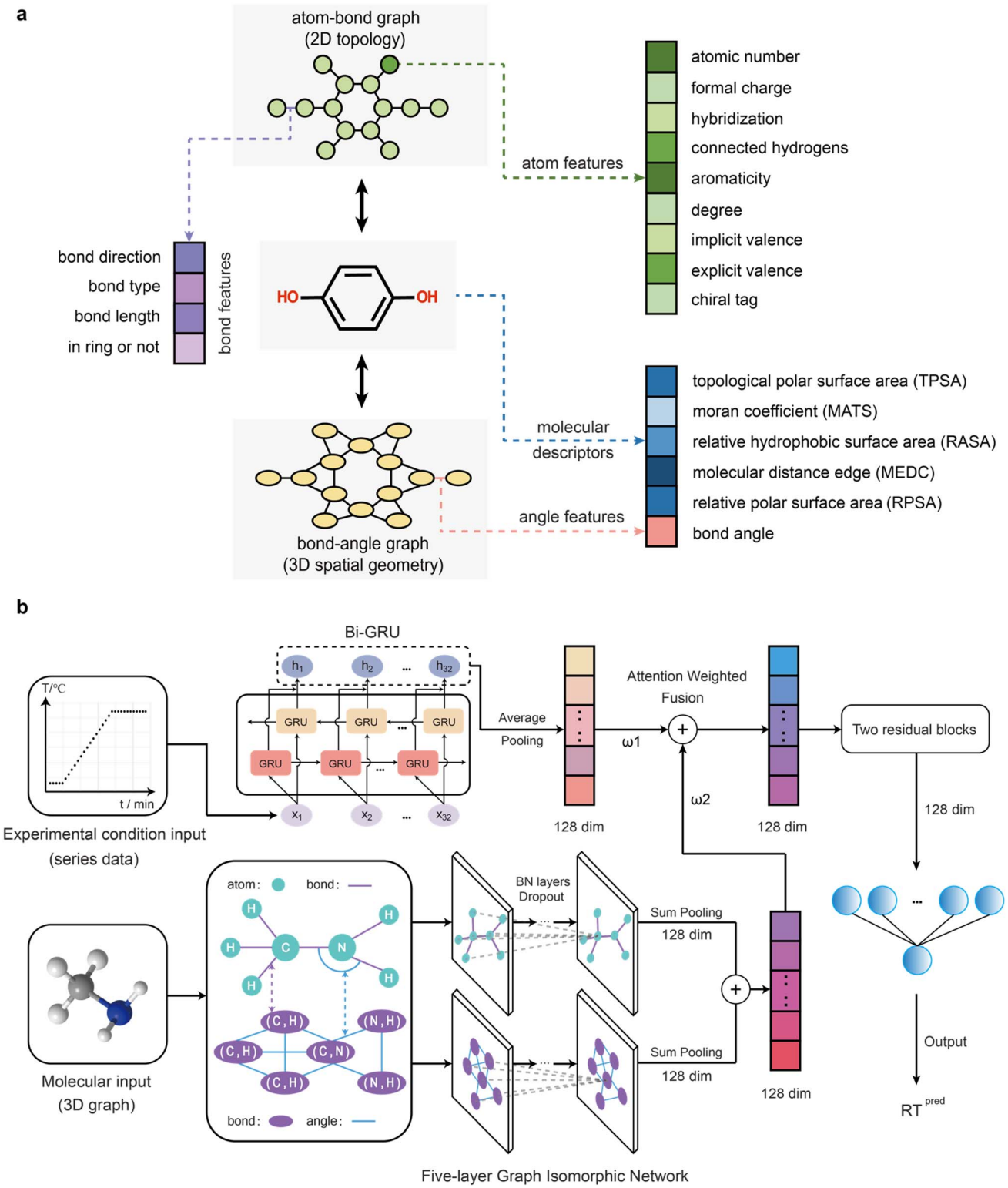

Fig. 3 Multimodal model architecture design. (a) Node and edge embeddings in atom-bond graphs and bond-angle graphs incorporate atom features, bond features, bond angle features, and molecular descriptors, facilitating comprehensive molecular structure characterization. (b) The multimodal model is divided into two branches. In the molecular processing branch, molecules are first represented as atom-bond graphs and bond-angle graphs, which are then processed through GINs. In the experimental conditions processing branch, time series information of temperature over time (T–t) is handled through a Bi-GRU. The multimodal model learns the weights of both branches through attention layers and finally integrates the information through residual blocks to output the prediction results.

network features a single neuron, tasked with delivering the predicted RT. Techniques like Batch Normalization, Layer Normalization, and dropout are applied to prevent overfitting and accelerate convergence. All the aforementioned designs not only improve the accuracy of predictions but also provide richer information for decision-making.

2.4 Performance of multimodal model

In the multimodal model, the dataset was subjected to a random split into training, validation, and test set in an 8 : 1 : 1 ratio. As illustrated in Fig. 4a, the model achieved an R^2 value of 0.995 on the test set, significantly surpassing the performance of conventional ML algorithms. An outlier analysis (see Section 5 of the SI Materials for details) was also performed to investigate the 6 samples with prediction errors greater than 1 minute. The results indicate that the model has limitations in predicting high molecular weight (MW > 200) and highly polar compounds. To further evaluate the model's generalization capabilities towards unknown compounds and to prevent data leakage, we implemented a novel dataset division strategy: 80% of the compound data was designated for training, with 10% allocated for testing and another 10% for validation. Following this approach, the model demonstrated an R^2 value of 0.931 on the test set (Fig. 4b), exhibiting excellent predictive performance and robust generalization to unknown compounds, significantly surpassing the performance of conventional ML algorithms. Above outcomes indicate that architecture of the proposed multimodal model effectively integrates diverse modal information, demonstrating a robust capability to learn the mappings among different variables, significantly augments the model's predictive accuracy for RT.

Additionally, we investigated the robustness of our model by evaluating its performance across different training set ratios and Gaussian noise levels. The method for adding noise to the data is as follows:

$$\hat{X} = X + \varepsilon \cdot \mathrm{std}(X) \cdot N(0,1) \quad (2)$$

where $\hat{X}$ represents the noise features of the temperature programs, X denotes the experimental features of the temperature programs, ε is the Gaussian noise ratio, std(X) is the standard deviation of the features, and N(0,1) refers to the normal distribution. As shown in Fig. 4c, the R^2 value on the test set increases with the rising proportion of the training set. When the proportion of training data reaches 60%, the R^2 value on the test set already exceeds 0.99. Additionally, the multimodal model demonstrates excellent robustness to noise in chromatographic experimental conditions: even with the introduction of 40% Gaussian noise, the R^2 value remains above 0.90, indicating robust performance of the model.

Although our model was initially trained on linear temperature program data, to assess its adaptability to nonlinear heating processes, we conducted further tests. We specifically selected eight nonlinear temperature programs and ten compounds that were distinctly not included in the GCRT dataset. This choice underscores the compounds' novelty, ensuring a rigorous evaluation of the model's performance on entirely new chemical entities. As shown in Fig. 4d, the temperature programs were categorized into rapid nonlinear (reaching final temperature within 8 minutes, represented in red) and slow nonlinear (reaching final temperature after 10 minutes, represented in purple). The orange area represents the coverage of linear temperature programs within the training set. Notably, two of the rapid nonlinear curves fall outside this coverage area. Encouragingly, the test results demonstrated that the multimodal model was able to accurately predict the RT of unknown compounds under various nonlinear conditions, achieving an R^2 of 0.900. This not only demonstrates the model's robust generalization ability but also indicates its effectiveness in capturing the essential temporal information within the heating process.

2.5 Applications of multimodal model in isomer separation

Isomers, which have identical molecular formula and similar boiling points, often require repeated adjustments of the temperature programs for effective separation in GC. To address this challenge, we employed the trained multimodal model to perform virtual high-throughput screening of various temperature programs, and recommending the optimal temperature program in seconds. This temperature program not only ensures effective separation of isomers but also significantly shortens the peak elution time, thereby avoiding traditional trial-and-error methods. The multimodal model exhaustively searches through 100 000 temperature programs, balancing search precision with efficiency. On a single Nvidia 4090 GPU, this search is completed in just one minute. The procedure for generating 100 000 random temperature programs starts by establishing boundary conditions for the following variables: initial temperature, final temperature, heating rate, and duration of the initial temperature. Following the setting of these parameters, random sampling is utilized to generate the curves. For further details on this method, refer to Section 6 of the SI Materials. The definition of isomer separation degree is as follows:

$$\Delta_i = \frac{\min\left(\left|y_{\mathrm{b},i} - y_{\mathrm{a},i}\right|, T_{\mathrm{threshold}}\right)}{T_{\mathrm{threshold}}}, \quad \text{for } i = 1, 2, \ldots, 100000 \quad (3)$$

where Δ_i represents the separation degree of isomers under the i-th temperature program, with a range of [0,1]. When $\Delta_i = 1$, we consider the separation of isomers to be successful. $y_{\mathrm{a},i}$ and $y_{\mathrm{b},i}$ respectively denote the peak elution time of the two compounds under the i-th temperature program, while $T_{\mathrm{threshold}}$ represents the minimum retention time difference considered acceptable for effective separation. Given that GC peaks are typically sharp, a small difference in peak elution time between isomer compounds is sufficient for separation. Therefore, we set the default value of $T_{\mathrm{threshold}}$ to 0.5 (30 seconds). $T_{\mathrm{threshold}}$ is a variable parameter; when it is not possible to find a temperature program where $\Delta_i = 1$ for a specific isomer, this value can be reduced to lower the difficulty of the search. This strategy enhances the flexibility and user-friendliness of the search algorithm. We define the separation ratio s to represent the ease

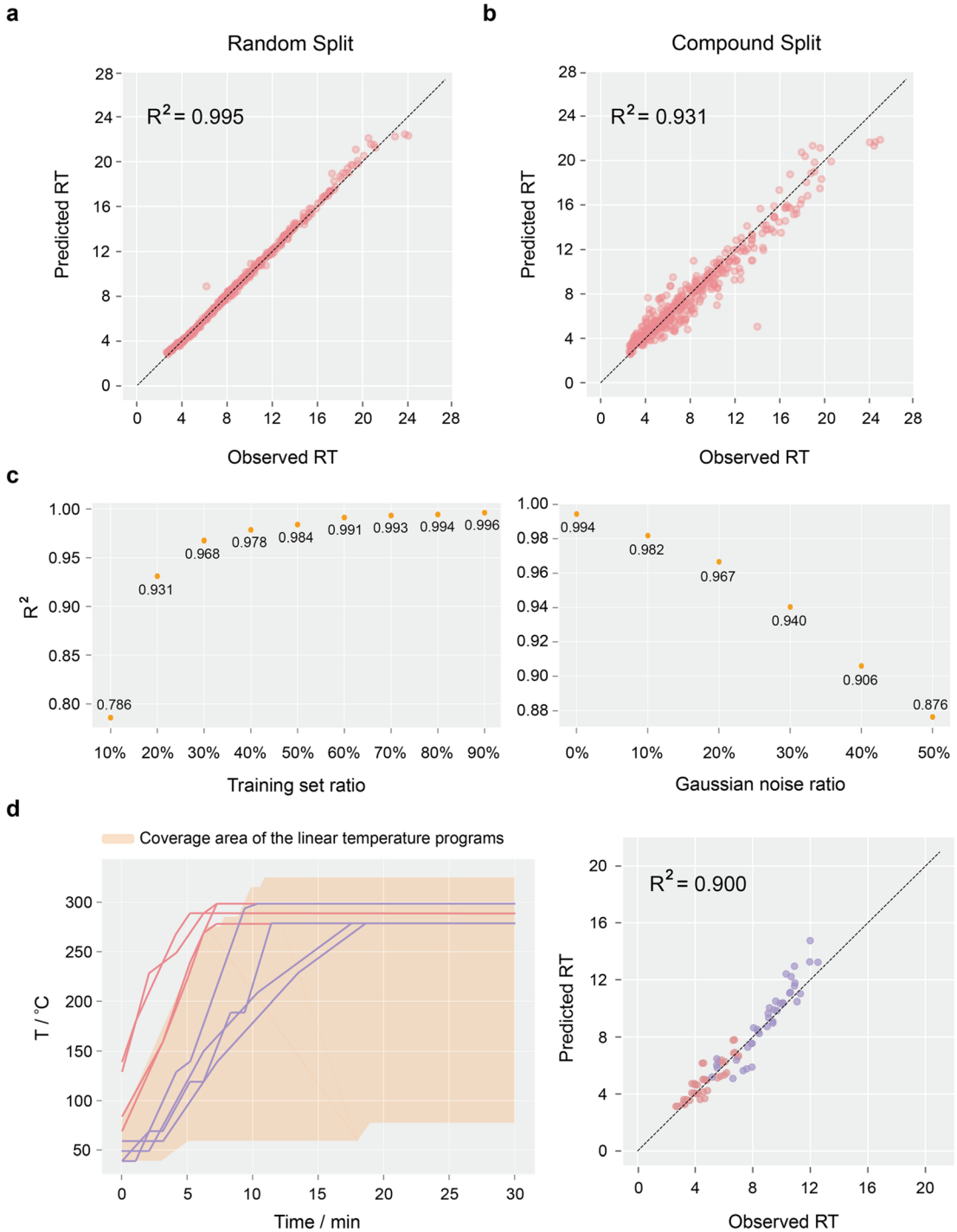

Fig. 4 Predictive performance of the multimodal model. (a) Performance of the multimodal model on the test set, with the dataset divided using the random split method, with training, validation, and test set ratios set at 8 : 1 : 1. (b) Performance of the multimodal model on the test set, with the dataset divided using the compound split method, with training, validation, and test set ratios set at 8 : 1 : 1. (c) Robustness analysis of the multimodal model across various training set ratios and Gaussian noise levels. (d) The multimodal model's performance in predicting RT under nonlinear temperature programs. The orange area represents the coverage of linear temperature programs within the training set. The red curves signify rapid nonlinear heating profiles, reaching the final temperature in less than 8 minutes. Conversely, the purple curves indicate slow nonlinear heating profiles, with the final temperature achieved in more than 10 minutes. All compounds tested were entirely new, outside the scope of the GCRT dataset.

of separating isomers at a specific $T_{\text{threshold}}$ value. A higher separation ratio s indicates that the isomers are more easily separated. The formula for s is as follows:

$$s = \frac{1}{100000} \sum_{i=1}^{100000} 1(\Delta_i = 1) \tag{4}$$

In practical applications, we aim not only for $\Delta_i = 1$ but also for the quickest possible peak elution time to reduce analysis timing cost. Therefore, we have established the following search algorithm:

$$i^* = \arg \min_{i \in \{i\Delta_i = 1\}} \max(y_{a,i},\ y_{b,i}) \tag{5}$$

where i^* represents the index of the optimal temperature program identified by the search algorithm. Through this method, we can find the index of the temperature program that not only achieves $\Delta_i = 1$ but also provides the quickest peak elution time.

To validate the reliability of the temperature program search algorithm, we selected two pairs of positional isomers (2-bromo-4-methylbenzaldehyde and 2-bromo-5-methylbenzaldehyde; 2-iodo-4-methylbenzeneamine and 3-iodo-4-methylbenzeneamine) and two pairs of *cis*/*trans* isomers (*cis*-1,2-diphenylethylene and *trans*-1,2-diphenylethylene; *cis*-1,4-dichloro-2-butene and *trans*-1,4-dichloro-2-butene) as proof-of-concept subjects. More importantly, these four sets of compounds were not included in the GCRT dataset, hence the model had no prior exposure to these compounds during training.

The $T_{\text{threshold}}$ for two groups of positional isomers was set at 0.33 (20 seconds), while the default value of 0.5 was used for two groups of *cis*/*trans* isomers. This adjustment was necessary because the positional isomers failed to achieve a temperature program satisfying $\Delta_i = 1$ at $T_{\text{threshold}}$ of 0.5. Fig. 5a presents the kernel density estimation (KDE) plots of $\max(RT_A^{pred}, RT_B^{pred})$ values for the four groups of isomers during the search process where $\Delta_i = 1$, illustrating the probability distribution of different isomers' later RT. The KDE was performed using a Gaussian kernel, as described by the following equation:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x - x_i)^2}{2h^2} \right) \tag{6}$$

where n is the number of data points, and h is the bandwidth parameter. Fig. 5a shows the mean and mode for $\max(RT_A^{pred}, RT_B^{pred})$, where the mode refers to the position of the maximum value on the KDE curve $\hat{f}(x)$. The recommended temperature program corresponds to the minimum $\max(RT_A^{pred}, RT_B^{pred})$. For the four groups of isomers from top to bottom in Fig. 5a, the temperature programs recommended by the multimodal model are accelerated by 42.85%, 27.77%, 19.48%, and 21.41%, respectively, relative to the temperature programs corresponding to the mode of $\max(RT_A^{pred}, RT_B^{pred})$. Experimental validation demonstrated that under the recommended temperature programs, all four groups of isomers were effectively separated, and the multimodal model successfully predicted the elution order with absolute errors of 0.395 minutes, 0.613 minutes, 0.693 minutes, and 0.395 minutes, respectively, showcasing excellent predictive accuracy. The experimental chromatograms are shown in Section 7 of the SI Materials. Notably, the s values for these four groups of isomers were 0.562, 0.094, 0.217, and 0.870, respectively, indicating a certain level of separation difficulty, especially for the pair of isomers 2-bromo-4-methylbenzaldehyde and 2-bromo-5-methylbenzaldehyde. The results suggest that this multimodal model, in conjunction with the search algorithm, has the potential to ensure effective isomer separation and enhance analytical efficiency. Given its high precision in prediction, this method can also be employed for internal standard peak identification and rapid analysis of complex mixtures.

To further investigate the separation difficulty of different isomers on a 5% phenyl–95% dimethylpolysiloxane stationary phase, we calculated the s values for various positional and *cis*/*trans* isomers at different $T_{\text{threshold}}$ values as a proof of concept: Fig. 5b illustrates three groups of *ortho*/*meta* positional isomers, differing only by their substituents. The results show that bromine-substituted isomers are easier to separate than chlorine-substituted ones, while methoxy-substituted isomers are the most difficult to separate. The chemical explanation we provide is that methoxy is a strong electron-donating group with significant polarity, producing a relatively consistent polar effect regardless of its position on the benzene ring, leading to minimal polarity differences between positional isomers, making them hard to separate. Chlorine is less polar than methoxy, and bromine is the least polar but has a higher polarizability, resulting in significant differences in substituent effects at different positions on the benzene ring, making positional isomers easier to separate. Fig. 5c shows three groups of halogenated ethylene *cis*/*trans* isomers, with iodine isomers being the easiest to separate, followed by bromine, and chlorine isomers being the hardest. The chemical explanation is that iodine atoms, being larger with high polarizability and weaker polarity, cause significant van der Waals force differences between *cis* and *trans* isomers. This substantial difference in atomic size and polarizability makes it easier to separate *cis*/*trans* isomers in the stationary phase. Bromine has moderate polarizability and size, whereas chlorine atoms are the smallest with the lowest polarizability, resulting in minimal polarity differences between *cis* and *trans* isomers, making separation the most difficult. In summary, different substituents lead to varying polarity differences between isomers, influencing separation difficulty. The data-driven conclusions align with chemical intuition. Other researchers can apply our model to their isomers of interest to explore the relationship between molecular structure and chromatographic behavior, aiding in experimental decision-making without performing experiments.

2.6 Predicting RT using RI

In the field of GC prediction research, the Retention Index (RI) is a fundamental parameter that reflects the retention capability of a specific stationary phase (SP) for molecules. Unique in its

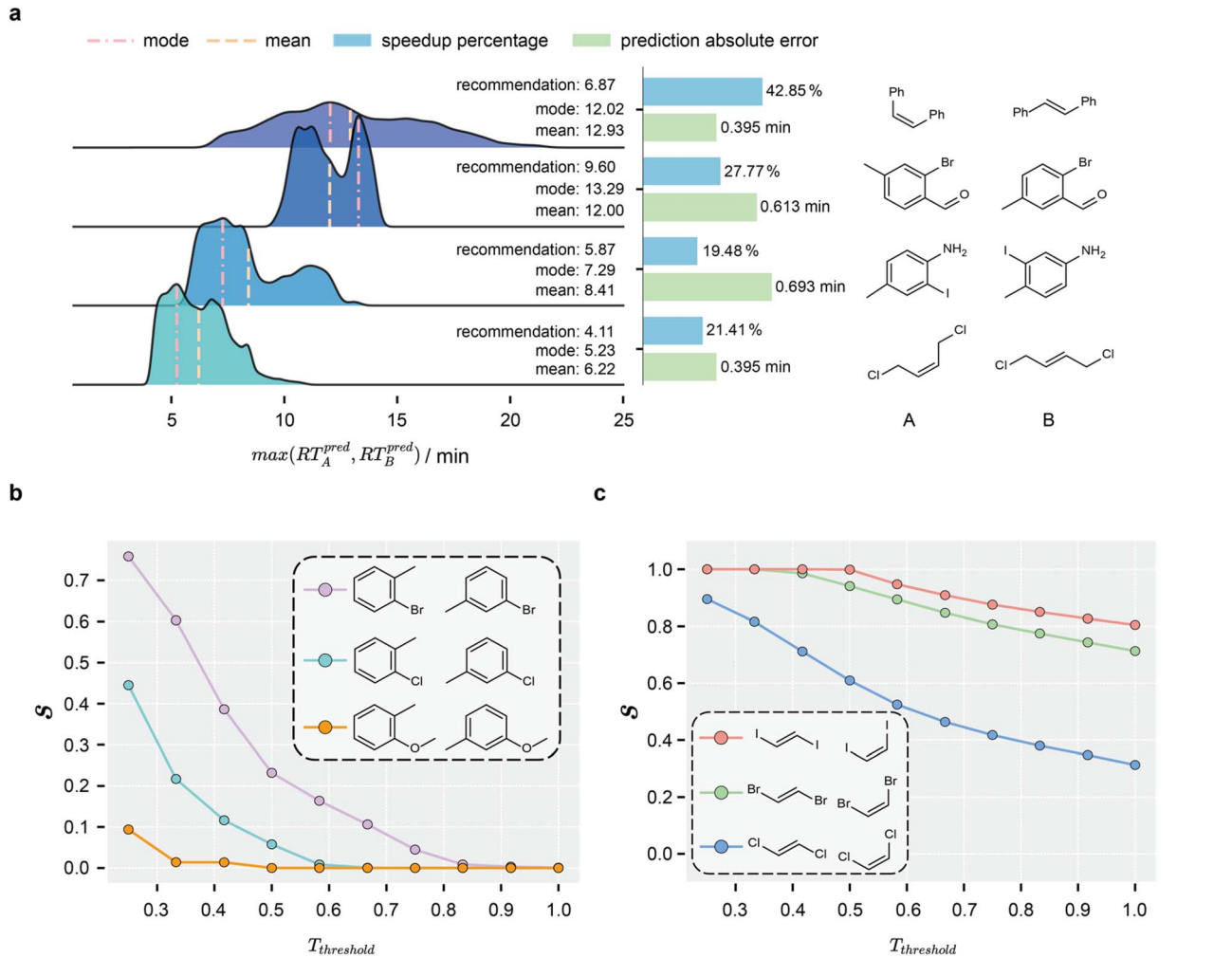

Fig. 5 Experimental validation of the search algorithm and exploration of isomer separation difficulty. (a) KDE curves of the max(RT_A^{pred}, RT_B^{pred}) values for four groups of isomers, showing the acceleration ratios of the recommended temperature programs and the absolute errors between predicted and experimental values. (b) *s* value variation curves at different $T_{threshold}$ for three groups of positional isomers. (c) *s* value variation curves at different $T_{threshold}$ for three groups of halogenated ethylene *cis*/*trans* isomers.

independence from chromatographic conditions, RI is predominantly influenced by the SP and molecular structure, represented as follows:

$$\mathrm{RI} = f(\mathrm{SP},\ \mathrm{Molecule}) \tag{7}$$

This characteristic renders it a critical parameter for the identification of compounds. Researchers like Dmitriy and Aleksandar have advanced ML models to predict the RI of compounds under various SPs.[13–16] Given the applicability of the RI across various GC conditions and instrumentation, integrating RI to enhance the scalability of RT prediction models represents a promising direction.

RT is primarily influenced by the SP, the molecule, and the temperature program, which can be represented as follows:

$$\mathrm{RT} = g(\mathrm{SP},\ \mathrm{Molecule},\ \mathrm{Temperature\ Program}) \tag{8}$$

Therefore, we used RI and the temperature program to predict RT, formulated as follows:

$$\mathrm{RT} = z(\mathrm{RI},\ \mathrm{Temperature\ Program}) \tag{9}$$

Linearly temperature-programmed retention index (LTPRI) is the most commonly used type of RI that depends on the RT of adjacent standard alkanes. Its calculation formula is as follows:

$$I = 100 \times \left(n + \frac{t_{\mathrm{R}} - t_{\mathrm{R}n}}{t_{\mathrm{R}n+1} - t_{\mathrm{R}n}}\right) \tag{10}$$

In this context, I represents the LTPRI of the target compound; n is the number of carbon atoms in the nearest smaller n-alkane relative to the target compound; t_{R} denotes the RT of the target compound; $t_{\mathrm{R}n}$ is the RT of the n-alkane with n carbon atoms; $t_{\mathrm{R}n+1}$ is the RT of the n-alkane with $n + 1$ carbon atoms. Under the stationary phase of 5% phenyl–95% dimethylpolysiloxane, we measured the RT of n-alkanes from C5 to C25 across various temperature programs. Utilizing these data, we calculated the LTPRI RI values for compounds in the GCRT dataset according to formula (10). Ultimately, we established a database comprising RI–RT values for 219 compounds under different temperature programs, totaling 3316 data points.

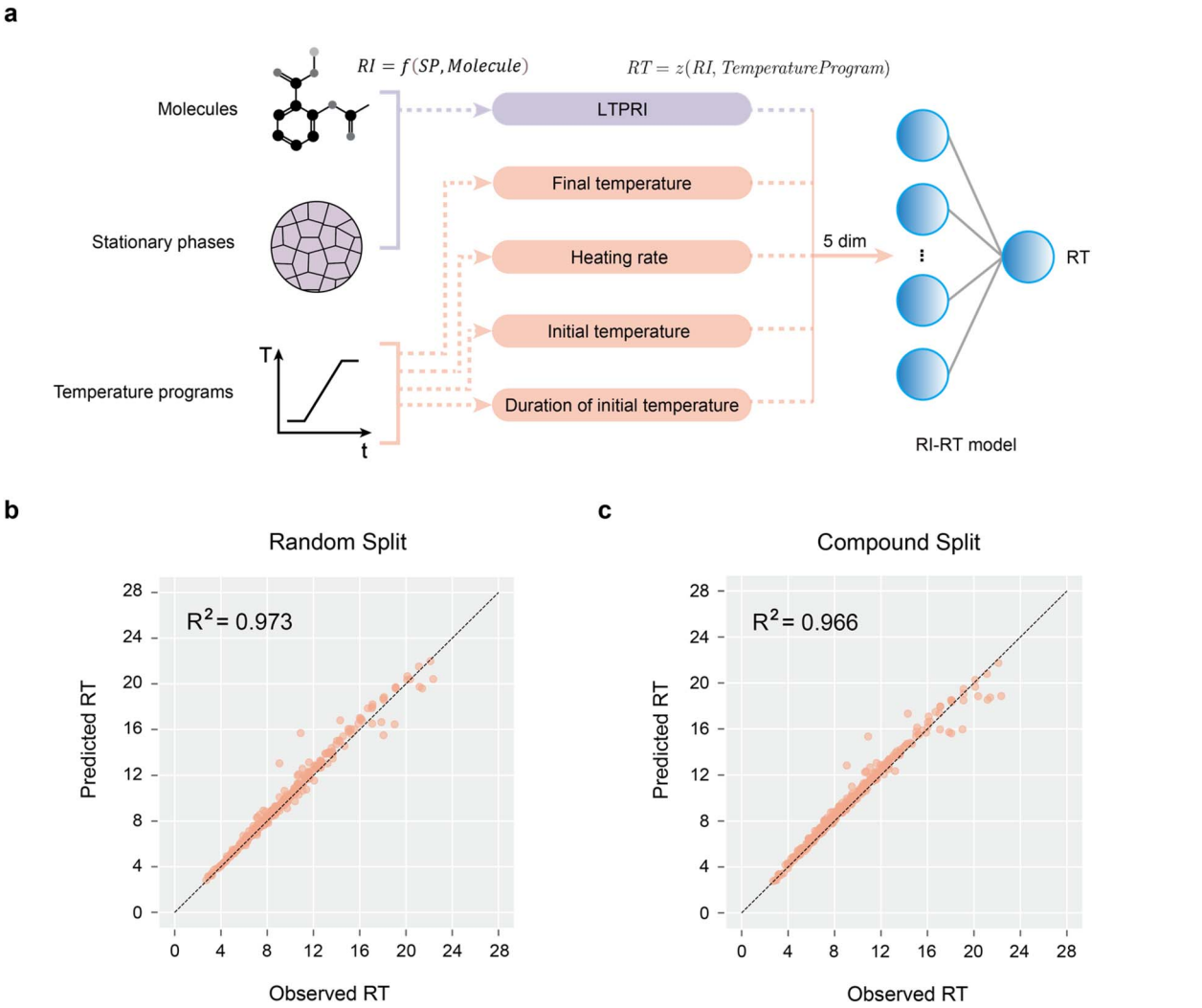

Fig. 6 Predictive performance of the RI-RT model. (a) The feature engineering approach of the RI-RT model. (b) Performance of the RI-RT model on the test set, with the dataset divided using the random split method, with training, validation, and test set ratios set at 8 : 1 : 1. (c) Performance of the RI-RT model on the test set, with the dataset divided using the compound split method, with training, validation, and test set ratios set at 8 : 1 : 1.

We developed an ANN model with two hidden layers, referred to as the RI-RT model (see Section 8 of the SI Materials for details), to predict RT. The model incorporates five input dimensions: LTPRI, final temperature, heating rate, initial temperature, and duration of initial temperature (as illustrated in Fig. 6a). The LTPRI encapsulates information about the molecule and stationary phase, while the remaining four features describe the dynamics of the temperature program. The dataset was randomly divided in an 8 : 1 : 1 ratio for training, validation, and test, achieving a predictive R^2 of 0.973 on the test set (shown in Fig. 6b). To assess the model's generalization capability and prevent data leakage, we re-partitioned the dataset by compound type, maintaining the same distribution ratio. Under these conditions, the model's predictive R^2 on the test set was 0.966 (as shown in Fig. 6c), surpassing the performance of the multimodal learning model.

The exceptional performance of the RI-RT model validates the feasibility of using RI to predict RT, demonstrating that a single parameter, RI, can encompass information about both the molecule and the stationary phase, significantly broadening the applicability of RT prediction models. We offer a pre-trained RI-RT model. Researchers can first employ existing RI prediction models[13–16] to calculate the RI of different molecules on their specific stationary phases. Subsequently, with minimal temperature program testing, they can fine-tune the RI-RT pre-trained model to achieve universal RT prediction.

3 Conclusions

In this study, we firstly developed a GC peak extraction algorithm, termed GC-PIEA, and utilized it to construct the GCRT database. Subsequently, we explored the application of various conventional ML algorithms for predicting RT, accompanied by a thorough analysis of model interpretability. Furthermore, we investigated the impact of different molecular descriptors on RT from a chemical perspective. Following this, we constructed

a multimodal prediction model integrating geometry-enhanced graph isomorphism network for extracting molecular information and Bi-GRU for analyzing the time-series information of temperature programs. This enabled us to achieve highly accurate predictions of GC-RT, outperforming conventional ML algorithms. Notably, despite being trained on linear temperature programs, the multimodal model demonstrated accurate predictions even under non-linear temperature conditions. As a highlight of our research, we developed a temperature program search algorithm that empowered the multimodal model to rapidly identify optimal chromatographic separation conditions. This facilitated effective separation of positional isomeric and *cis*/*trans* isomeric compounds while ensuring rapid peak elution, thereby avoiding traditional trial-and-error method and significantly enhancing analytical efficiency. Moreover, this search algorithm can be employed to investigate the separation difficulty of various isomers, thereby elucidating the relationship between molecular structure and chromatographic behavior, and providing an in-depth understanding of separation mechanisms.

However, the multimodal model still exhibits limitations. Due to practical constraints, obtaining RT data for compounds across various stationary phases and temperature programs is difficult. Despite being trained on the widely used 5% phenyl–95% dimethylpolysiloxane stationary phase, extending its application to other stationary phases remains challenging. To address this issue, we propose a strategy for predicting RT using RI and provide a corresponding pre-trained model. Since RI inherently contains information about the stationary phase, researchers can apply our pre-trained model to their specific stationary phases, fine-tuning it with minimal data to achieve universal RT prediction. Through this approach, we aim to overcome the limitations posed by different stationary phases and further enhance the model's applicability. Additionally, in future work, the collection of chromatographic data under optimized flow rates and injection conditions would significantly enhance the quality of the dataset.

Author contributions

L. S. conducted experiments and collected the GCRT dataset. J. L analyzed the data. J. L. performed chemoinformatic and machine learning studies. F. M. conceived the idea and designed the overall research. F. M. and S. C. supervised the whole project. All authors wrote the manuscript.

Conflicts of interest

The authors declare no competing interests.

Data availability

The dataset and source code supporting the findings of this study are publicly available. The dataset used for training and evaluation is available on Zenodo at: **https://zenodo.org/records/16220196**, and the complete source code used for gas chromatography retention time prediction and isomer separation optimization is archived on Figshare at: **https://doi.org/10.6084/m9.figshare.29634785**.

Supplementary information includes details of model architectures, outlier analysis, and experimental validation. See DOI: **https://doi.org/10.1039/d4dd00369a**.

Acknowledgements

This work is supported by the Natural Science Foundation of China (Grant No. 22071004, 21933001, 22150013, 22061031). We thank the High-Performance Computing Platform of Peking University for machine learning model training.

References

1 M. I. Jordan and T. M. Mitchell, Machine learning: Trends, perspectives, and prospects, *Science*, 2015, **349**(6245), 255–260.
2 C. Janiesch, P. Zschech and K. Heinrich, Machine learning and deep learning, *Electron. Mark.*, 2021, **31**(3), 685–695.
3 S. Lu, M. Liu, L. Yin, Z. Yin, X. Liu and W. Zheng, The multimodal fusion in visual question answering: a review of attention mechanisms, *PeerJ Comput. Sci.*, 2023, **9**, e1400.
4 P. Nandwani and R. Verma, A review on sentiment analysis and emotion detection from text, *Soc. Netw. Anal. Min.*, 2021, **11**(1), 81.
5 H. Ko, S. Lee, Y. Park and A. Choi, A survey of recommendation systems: recommendation models, techniques, and application fields, *Electronics*, 2022, **11**(1), 141.
6 X. Domingo-Almenara, A. Perera, N. Ramirez, N. Canellas, X. Correig and J. Brezmes, Compound identification in gas chromatography/mass spectrometry-based metabolomics by blind source separation, *J. Chromatogr. A*, 2015, **1409**, 226–233.
7 X. Domingo-Almenara, A. Perera, N. Ramirez and J. Brezmes, Automated resolution of chromatographic signals by independent component analysis—orthogonal signal deconvolution in comprehensive gas chromatography/mass spectrometry-based metabolomics, *Comput. Methods Programs Biomed.*, 2016, **130**, 135–141.
8 X. Domingo-Almenara, A. Perera and J. Brezmes, Avoiding hard chromatographic segmentation: A moving window approach for the automated resolution of gas chromatography–mass spectrometry-based metabolomics signals by multivariate methods, *J. Chromatogr. A*, 2016, **1474**, 145–151.
9 J. Novák. Quantitative analysis by gas chromatography, in *Adv Chromatogr*, Boca Raton (FL), CRC Press, 2021. pp. 1–71.
10 W. Wichitnithad, O. Sudtanon, P. Srisunak, K. Cheewatanakornkool, S. Nantaphol and P. Rojsitthisak, Development of a sensitive headspace gas chromatography–mass spectrometry method for the simultaneous determination of nitrosamines in losartan active pharmaceutical ingredients, *ACS Omega*, 2021, **6**(16), 11048–11058.

11 Y. Picó and D. Barceló, Pyrolysis gas chromatography–mass spectrometry in environmental analysis: Focus on organic matter and microplastics, *TrAC, Trends Anal. Chem.*, 2020, **130**, 115964.
12 R. L. Grob. Theory of gas chromatography, in *Modern Practice of Gas Chromatography*, Hoboken (NJ): Wiley, 2004, pp. 23–63.
13 D. D. Matyushin, A. Y. Sholokhova and A. K. Buryak, A deep convolutional neural network for the estimation of gas chromatographic retention indices, *J. Chromatogr. A*, 2019, **1607**, 460395.
14 D. D. Matyushin and A. K. Buryak, Gas chromatographic retention index prediction using multimodal machine learning, *IEEE Access*, 2020, **8**, 223140–223155.
15 D. D. Matyushin, A. Y. Sholokhova and A. K. Buryak, Deep learning based prediction of gas chromatographic retention indices for a wide variety of polar and mid-polar liquid stationary phases, *Int. J. Mol. Sci.*, 2021, **22**(17), 9194.
16 A. M. Veselinović, D. Velimorović, B. Kaličanin, A. Toropova, A. Toropov and J. Veselinović, Prediction of gas chromatographic retention indices based on Monte Carlo method, *Talanta*, 2017, **168**, 257–262.
17 F. Qiu, Z. Lei and L. W. Sumner, MetExpert: An expert system to enhance gas chromatography–mass spectrometry-based metabolite identifications, *Anal. Chim. Acta*, 2018, **1037**, 316–326.
18 C. Jirayupat, K. Nagashima, T. Hosomi, T. Takahashi, W. Tanaka, B. Samransuksamer, *et al.*, Image processing and machine learning for automated identification of chemo-/biomarkers in chromatography–mass spectrometry, *Anal. Chem.*, 2021, **93**(44), 14708–14715.
19 Y. Fan, C. Yu, H. Lu, Y. Chen, B. Hu, X. Zhang, *et al.*, Deep learning-based method for automatic resolution of gas chromatography–mass spectrometry data from complex samples, *J. Chromatogr. A*, 2023, **1690**, 463768.
20 C. Capitain and P. Weller, Non-targeted screening approaches for profiling of volatile organic compounds based on gas chromatography–ion mobility spectroscopy (GC-IMS) and machine learning, *Molecules*, 2021, **26**(18), 5457.
21 H. Xu, J. Lin, Q. Liu, Y. Chen, J. Zhang, Y. Yang, *et al.*, High-throughput discovery of chemical structure–polarity relationships combining automation and machine-learning techniques, *Chem*, 2022, **8**(12), 3202–3214.
22 H. Xu, J. Lin, D. Zhang and F. Mo, Retention time prediction for chromatographic enantioseparation by quantile geometry-enhanced graph neural network, *Nat. Commun.*, 2023, **14**(1), 3095.
23 S. Kearnes, K. McCloskey, M. Berndl, V. Pande and P. Riley, Molecular graph convolutions: Moving beyond fingerprints, *J. Comput.-Aided Mol. Des.*, 2016, **30**, 595–608.
24 K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko and K.-R. Müller, Schnet—a deep learning architecture for molecules and materials, *J. Chem. Phys.*, 2018, **148**(24), 241722.
25 X. Fang, L. Liu, J. Lei, D. He, S. Zhang, J. Zhou, *et al.*, Geometry-enhanced molecular representation learning for property prediction, *Nat. Mach. Intell.*, 2022, **4**(2), 127–134.
26 S.-W. Li, L.-C. Xu, C. Zhang, S.-Q. Zhang and X. Hong, Reaction performance prediction with an extrapolative and interpretable graph model based on chemical knowledge, *Nat. Commun.*, 2023, **14**(1), 3569.