Supplementary Information for **Explicit relation between thin film chromatography and column chromatography conditions from statistics and machine learning**

Hao Xu[a,b,c], Wenchao Wu[a,d], Yuntian Chen[c,e], Dongxiao Zhang[c,*], Fanyang Mo[a,d,f,g,*]

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

[b] *BIC-ESAT, ERE, and SKLTCS, College of Engineering, Peking University, Beijing, 100871, P. R. China.*

[c] *Zhejiang Key Laboratory of Industrial Intelligence and Digital Twin, Eastern Institute of Technology, Ningbo, Zhejiang 315200, China*

[d] *School of Materials Science and Engineering, Peking University, Beijing, 100871, P. R. China.*

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

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

[g] *Guangdong Provincial Key Laboratory of Nano-Micro Materials Research, Peking University Shenzhen Graduate School, Shenzhen, 518055, China.*

Supplementary text

1. Influence of the sample mass

In this work, explicit equations are discovered that describe the underlying relationship between the R_F values from thin-layer chromatography (TLC) and column chromatography (CC). In these formulas, the influence of the solvent ratio, which is crucial for chromatographic separation, is considered. In addition to the mobile phase ratio, the loading sample mass also influences the retention volume in the CC, although its impact is less dominant. In this section, the influence of loading sample mass is investigated. The model predictions are generated from the surrogate model of CC. For each of the 585 distinct experimental conditions, seven different loading sample mass

conditions are considered, including 50 mg, 100 mg, 150 mg, 200 mg, 250 mg, 300 mg, and 350 mg. To better demonstrate the influence of loading sample mass, for each model prediction, a retention volume of 50 mg was used as the baseline, and the ratio of the retention time under other sample masses to the retention volume of 50 mg was calculated. The mean values of the ratios across different sample masses are shown in Fig. 2e. From the figure, the trend of the mean ratio for V_S and V_E is divergent. The mean ratio of V_1 decreases and remains below 1 as the sample mass increases. This means that the start time will be earlier. In contrast, the mean ratio of V_E increases and remains higher than 1. This discovery aligns with the experience that a large loading sample amount may result in column overloading, leading to increased peak width and decreased resolution.

2. The definition and calculation of the separation index

In this work, the separation index matrix is defined and utilized to measure the possibility of separation in the CC from the aspect of statistics. Here, the definition and calculation are introduced in detail.

For the sake of simplicity, we use the example of two compounds in the mixture for demonstration. The two compounds are termed Compound A and Compound B. Empirically, TLC is conducted under a selected mobile phase proportion, and the R_F values of A and B are obtained. Guided by chemical experience, the mobile phase, in which the R_F value of A is 0.2~0.3, is suitable for separation in CC. This implies that the separation probability is related to the mobile phase and the corresponding R_F values. From Fig. 3a, the range of retention volumes can be estimated from the given mobile phase and R_F values, and the variation principle is similar in divergent mobile phase proportions. Therefore, the separation probability can be measured by comparing the ranges of retention volumes V_S and V_E of both components, which are termed V_{SA}, V_{EA}, V_{SB}, and V_{EB}. Since V_S and V_E represent the solvent volumes corresponding to when the compound just starts to elute and when it has completely eluted, respectively, there are only two scenarios for achieving complete separation, including $V_{SA} > V_{EB}$ and $V_{SB} > V_{EA}$. These two scenarios correspond to Compound A eluting only after Compound B

has completely eluted, and Compound B still not eluting even after Compound A has completely eluted.

Therefore, we can assess the likelihood of separation on the basis of the ranges of R_F values for both compounds that satisfy these two conditions. Taking V_{SA} and V_{EB} as examples, there can be various scenarios for the relative positions of their ranges, as illustrated in Fig. S8. If Compound A has not yet completely eluted, and if Compound B has already started to elute, that is, if $V_{EB}>V_{SA}$, it is obviously not able to be separated completely. Conversely, if $V_{EB}<V_{SA}$, where Compound B has not started to elute when Compound A has eluted completely, the compounds can be separated. Therefore, when the range of V_{SA} is completely above that of V_{EB}, the separation probability p_1 for V_{SA} and V_{EB} is 1. However, when there is an overlap between the ranges of V_{SA} and V_{EB}, it is difficult to calculate the separation probability. According to the possible scenarios, there are four conditions when the ranges of V_{SA} and V_{EB} overlap, namely overlap with V_{SA} higher (Fig. S8a), overlap with V_{EB} higher (Fig. S8b), V_{SA} contained in V_{EB} (Fig. S8c), and V_{EB} contained in V_{SA} (Fig. S8d). For each condition, the separation probability can be calculated from the conditional probability of separable intervals in the total intervals, which is defined as:

$$p_1 = \begin{cases} \dfrac{V_{SA}^{\max}-V_{EB}^{\max}}{V_{SA}^{\max}-V_{SA}^{\min}} + \dfrac{V_{EB}^{\max}-V_{SA}^{\min}}{V_{SA}^{\max}-V_{SA}^{\min}} \cdot \dfrac{V_{EB}^{\max}-V_{SA}^{\min}}{V_{EB}^{\max}-V_{EB}^{\min}}, V_{SA}^{\max} > V_{EB}^{\max} \,\&\, V_{SA}^{\min} > V_{EB}^{\min} \\ \dfrac{V_{SA}^{\max}-V_{EB}^{\min}}{V_{SA}^{\max}-V_{SA}^{\min}} \cdot \dfrac{V_{SA}^{\max}-V_{EB}^{\min}}{V_{EB}^{\max}-V_{EB}^{\min}}, V_{EB}^{\min} < V_{SA}^{\max} < V_{EB}^{\max} \\ \dfrac{V_{SA}^{\max}-V_{EB}^{\min}}{V_{EB}^{\max}-V_{EB}^{\min}}, V_{SA}^{\max} < V_{EB}^{\max} \,\&\, V_{SA}^{\min} > V_{EB}^{\min} \\ \dfrac{V_{SA}^{\max}-V_{EB}^{\min}}{V_{SA}^{\max}-V_{SA}^{\min}}, V_{SA}^{\max} > V_{EB}^{\max} \,\&\, V_{SA}^{\min} < V_{EB}^{\min} \end{cases} \quad \text{(S.1)}$$

Here, $V_{SA}^{\max}$, $V_{SA}^{\min}$, $V_{EB}^{\max}$, and $V_{EB}^{\min}$ represent the maximum and minimum ranges of V_{SA} and V_{EB}, respectively. To reduce the impact of outliers, the maximum and minimum values within the 95% confidence interval are adopted. If $V_{SA}^{\min}>V_{EB}^{\max}$, the mixture will definitely be separated, and p_1=1. If $V_{SA}^{\max}<V_{EB}^{\min}$, the mixture will never be separated, and p_1=0. Then, for V_{EA} and V_{SB}, the separation probability p_2 can be calculated similarly. From this, the separation index P can be calculated as:

$$P = p_1 + p_2 \tag{S.2}$$

In this work, the R_F value is divided into 10 ranges. Therefore, for Compounds A and B, there are 100 possible combinations of their R_F value ranges, which can be represented by a 10×10 matrix. For each combination, the separation index can be calculated via Eqs. (S.1) and (S.2). All the separation indices form a separation index matrix. Notably, each mobile phase proportion corresponds to a separation index matrix. Considering that the underlying principle for the R_F values is similar in divergent mobile phase proportions, we calculate the average of the separation index matrices from eight commonly used mobile phase proportions to obtain a statistical result, as shown in Fig. 3b.

Supplementary tables and figures

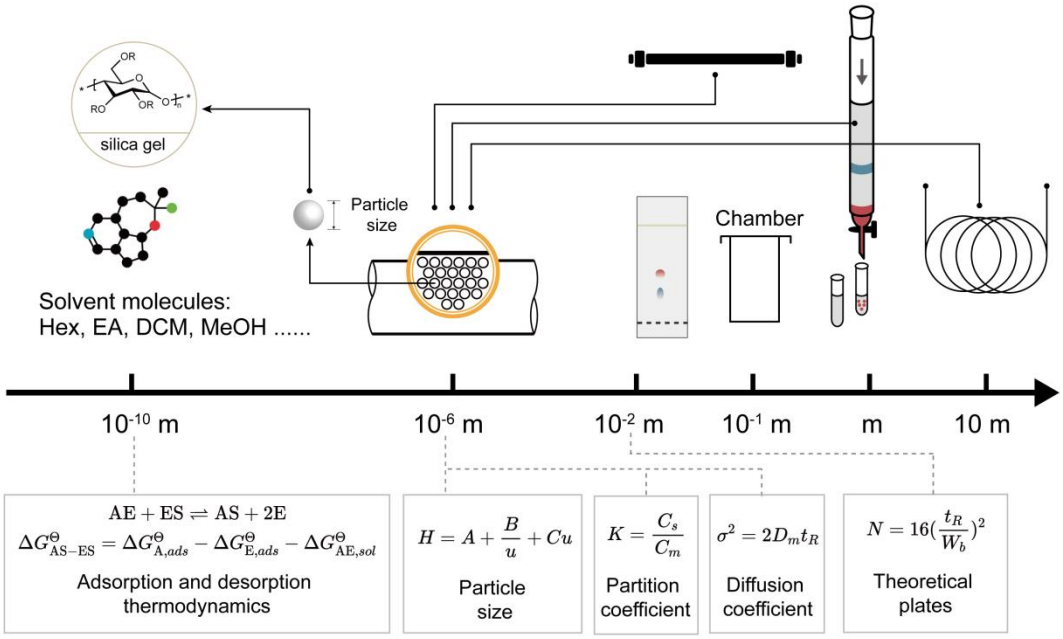

Fig. S1. The mechanism under different scales of the chromatography process.

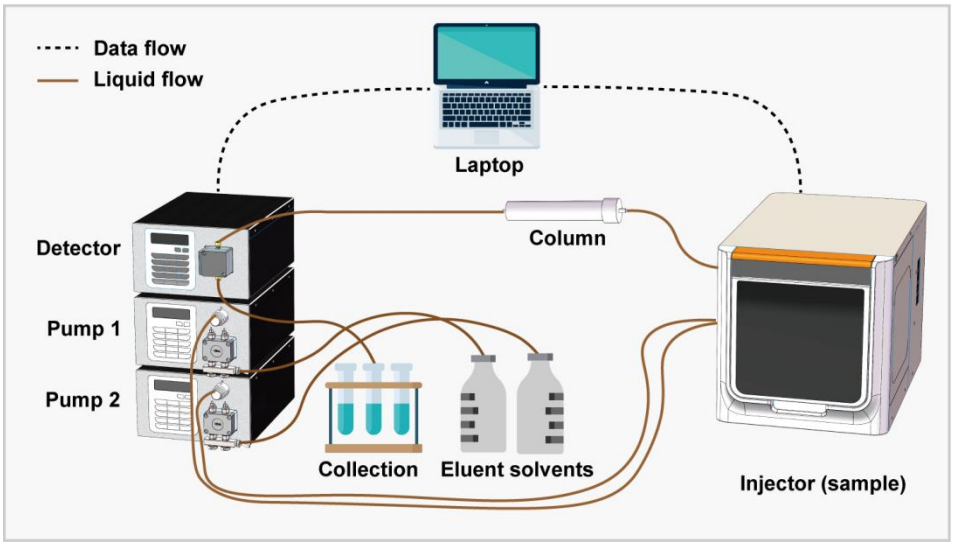

Fig. S2. The sketch map of the automatic platform for conducting column chromatography. The automatic platform is controlled by a laptop.

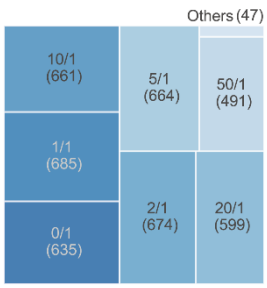

Fig. S3. The distribution of the mobile phase proportions in a 4 g column dataset.

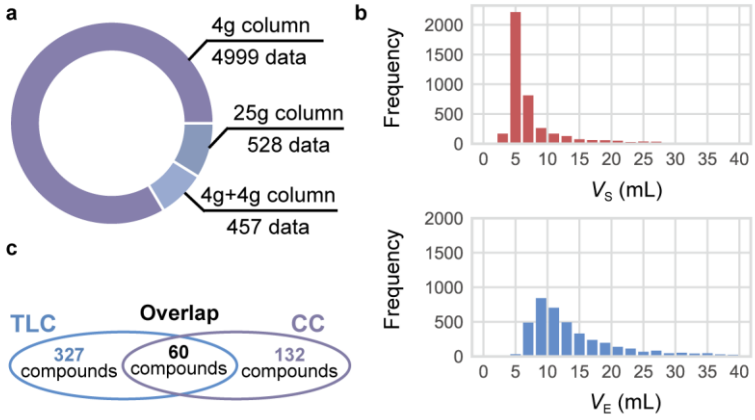

Fig. S4. The automatic platform for column chromatography and the introduction to the column chromatography dataset. (a), The number of datapoints from different column specifications. **(b)**, The distribution of the retention volume in the collected column chromatography (CC) dataset. **(c)**, The intersections between the thin layer chromatography (TLC) dataset and CC dataset. Here, V_S and V_E refer to the retention volume at starting point and retention volume at ending point.

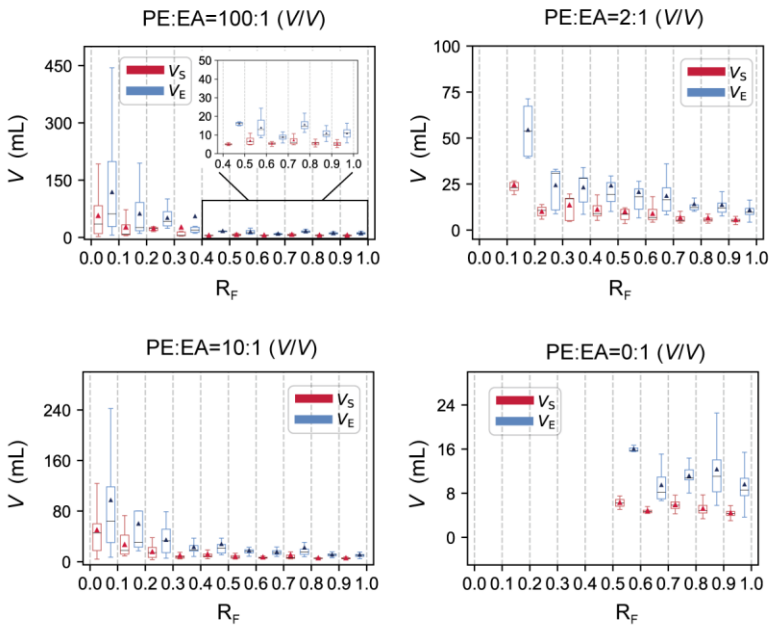

Fig. S5. The distribution of retention volumes in diversified ranges of R_F values under different mobile phase proportions. Here are the situations for PE:EA =100:1, 10:1, 2:1, and 0:1 (V/V). The sample size n=585. The boxplot is defined by the maximum, minimum, and quartiles. The mean values are represented by triangles and the median values are denoted by lines. Here, V_S and V_E refer to the retention volume at starting point and retention volume at ending point. PE refers to petroleum ether and EA refers to ethyl acetate.

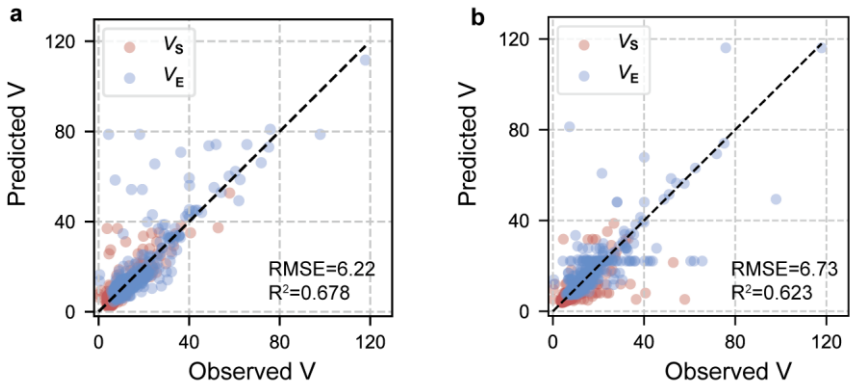

Fig. S6. The comparison between the interpretable model based on the discovered equations and other methods. **(a),** The predicted and observed retention volume with our interpretable model. **(b),** The predicted and observed retention volume with direct tree model training. Here, V_S and V_E refer to the retention volume at starting point and retention volume at ending point. RMSE and R^2 refer to the root mean squared error and coefficient of determination, respectively.

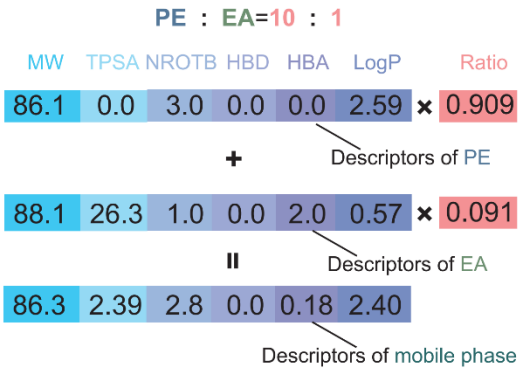

Fig. S7. The diagram for calculating an averaged molecular description. PE refers to petroleum ether and EA refers to ethyl acetate.

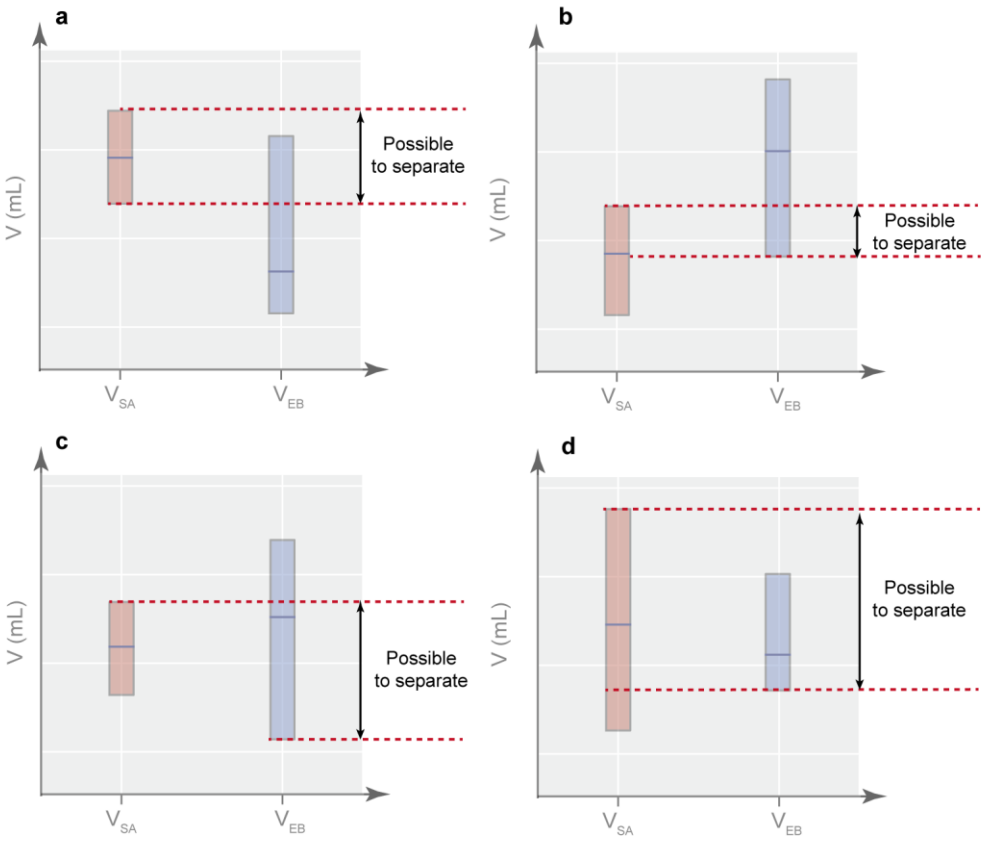

Fig. S8. The diagram for the calculation of the separation index. (a)-(d), Examples of the possible conditions for the distribution range between V_S of the Compound A (V_{SA}) and V_E of the Compound B (V_{EB}), when they are overlapped.

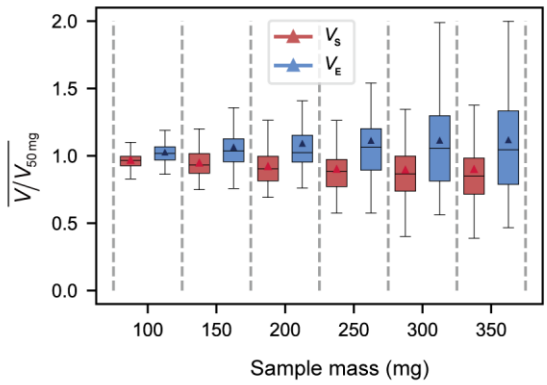

Fig. S9. The influence of loading sample mass. The y-axis represents the mean value of the ratio of the retention volume under the corresponding sample mass (V) to the retention volume under 50 mg (V_{50mg}) with other experimental conditions kept the same. The distribution is illustrated in the box-plot, the number of samples is 192. The black line is median and the triangle is the mean value.

Table S1. The abbreviations of the experimental conditions and descriptors of compounds and eluents, and their corresponding meaning.

Abbreviation	Meaning
MolWt	molecular weight
nRotB	number of rotatable bonds
HBD	number of hydrogen bond donor
HBA	number of hydrogen bond acceptor
LogP	lipid-water partition coefficient
AATS1s	averaged Moreau-Broto autocorrelation of lag 1 weighted by intrinsic state
ATSC0m	centred Moreau-Broto autocorrelation of lag 0 weighted by mass
AXp-2dv	2-ordered averaged Chi path weighted by valence electrons
AXp-3dv	3-ordered averaged Chi path weighted by valence electrons
AETA_eta_FL	averaged local ETA functionality index
AMID_N	averaged molecular ID on N atoms
AATSC0p	averaged and centred moreau-broto autocorrelation of lag 0 weighted by polarizability
TPSA(NO)	topological polar surface area (use only nitrogen and oxygen)
TPSA	topological polar surface area
ETA_dEpsilon_D	ETA delta epsilon (type: D)
PEOE_VSA6	MOE Charge VSA Descriptor 6 (-0.10 <= x < -0.05)
MolWt_e	molecular weight of eluent
TPSA_e	topological polar surface area of eluent
nRotB_e	number of rotatable bonds of eluent
HBD_e	number of hydrogen bond donor of eluent
HBA_e	number of hydrogen bond acceptor of eluent
LogP_e	lipid-water partition coefficient of eluent
e	sample mass
m	sample solvent type
V_e	solvent volume

Table S2. The coefficient of determination for the regression tree to fit the column chromatography factor under different solvent ratios.

Mobile phase (V/V)	R^2
PE:EA =1:0	0.8083
PE:EA=100:1	0.9026
PE:EA=50:1	0.9166
PE:EA=10:1	0.8160
PE:EA=5:1	0.8977
PE:EA=2:1	0.8779
PE:EA=1:1	0.9213
PE:EA=0:1	0.7678