Supplementary Information for: High-throughput discovery of chemical structure-polarity relationships combining automation and machine learning techniques

Hao Xu[1,2†], Jinglong Lin[1†], Qianyi Liu[3], Yuntian Chen[4], Jianning Zhang[1], Yang Yang[5], Michael C. Young[6], Yan Xu[7], Dongxiao Zhang[8,9*] and Fanyang Mo[1*]

[1*]School of Materials Science and Engineering, Peking University, Beijing, 100871, P. R. China.
[2]BIC-ESAT, ERE, and SKLTCS, College of Engineering, Peking University, Beijing, 100871, P. R. China.
[3]College of Chemistry and Molecular Engineering, Peking University, Beijing, 100871, P. R. China.
[4]EIT Institute for Advanced Study, Yongriver Institute of Technology, Ningbo, 315200, Zhejiang P. R. China.
[5]Department of Chemistry and Biochemistry, University of California Santa Barbara, Santa Barbara, 93106, CA, U.S.
[6]Department of Chemistry & Biochemistry, School of Green Chemistry & Engineering, The University of Toledo, 2801 W. Bancroft St. Toledo, 43606, Ohio, U.S.
[7]Chemistry Service Unit, WuXi AppTec Headquarters, Shanghai, 200131, , P. R. China.
[8]Department of Mathematics and Theories, Peng Cheng Laboratory, Shenzhen, 518000, P. R. China.
[9]School of Environmental Science and Engineering, Southern University of Science and Technology, Shenzhen, 518055, P. R. China.

*Corresponding author(s). E-mail(s): zhangdx@sustech.edu.cn; fmo@pku.edu.cn;
†These authors contributed equally to this work.

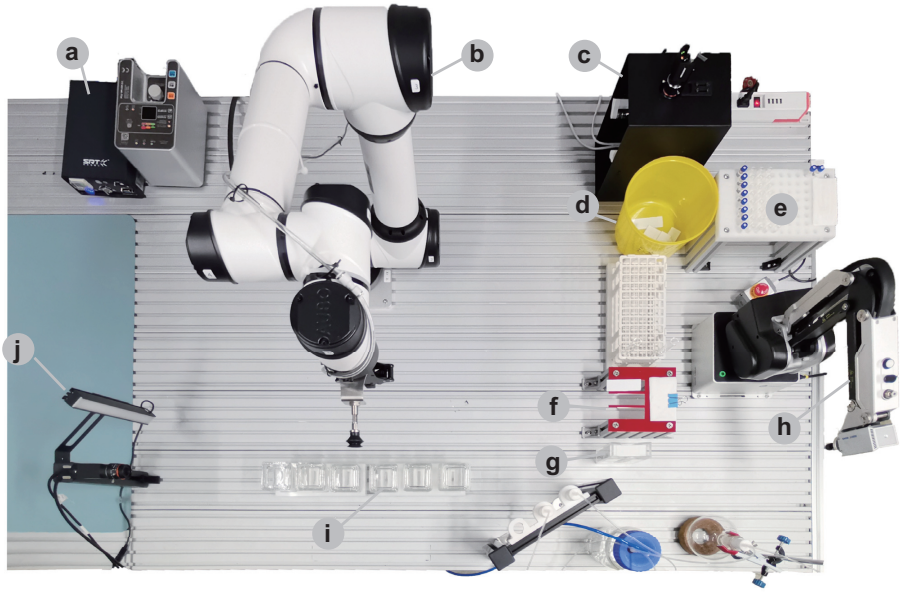
1 The design of automated high-throughput robotic platform for TLC

1.1 Overview of the automated platform

In this work, an automated TLC high-throughput robotic platform for TLC that can automatically complete the TLC analysis process is developed. The overview of the automated platform is provided in SI Figure 1. The core of the designed system is two collaborative robots, namely DOBOT MG400 robot that is equipped with a capillary and AUBO i5 robot that is equipped with a mechanical gripper and a suction cup driven by the air pump. The TLC samples to be analyzed are placed on the TLC sample tray and the TLC plates are placed on the stand, which can load many TLC plates at the same time for high-throughput experiment. In the experiment, the TLC sample is dipped by the capillary on the DOBOT MG400 robot and then spotted onto the TLC plates successively. Then, the AUBO i5 robot uses the gripper to transfer the spotted TLC plate to the corresponding TLC chamber. The sucker is employed to open or close the chamber lid. It is worth noting that there are multiple TLC chambers here, which can be utilized with different developing solvent eluents at the same time, which greatly improves the efficiency of the system. When the developing time reaches the set value (300 seconds in this work), AUBO i5 robot will retrieve the developed TLC plate from the chamber and send it to the ultraviolet and visible light photographic devices respectively to take photos and record the result. The visible light photographic device is utilized to record the position of the frontier of the developing eluent and the ultraviolet photographic device is employed to visualize the spot on the TLC board. The used TLC board is dropped into the garbage can and pristine TLC plates are retrieved from storage and placed on the stand via the suction cup. This cycle will continue until all TLC samples have been tested. We have made a vedio about the above mentioned workflow. Please reach it via the following address: https://www.bilibili.com/video/BV1am4y1o7yE/

1.2 Robots

In the automated platform, two collaborative robots including DOBOT MG400 robot and AUBO i5 robot play a vital role in conducting the experiments automatically, where the specifications are demonstrated in SI Figure 2. The DOBOT MG400 is a four-axis robot, which has the advantage of high efficiency and easy operation since the z-axis is perpendicular to the operating plane, which means that it is particularly suitable for vertical motion. Therefore, it is utilized to complete the task of drawing and spotting samples on the TLC plates, which requires fast and accurate vertical movement. Different from the DOBOT MG400, the AUBO i5 is a six-axis robot that is more flexible and able to handle more complicate tasks since it can act like a human arm. Considering the freedom of its operation, we employ it to complete the transfer task of TLC plate between different components of the platform, including putting a TLC



SI Figure 1. The overview of the designed automated platform for TLC. a, Air pumps; **b**, the AUBO i5 robot; **c**, the UV light photographic device; **d**, the garbage can; **e**, the sample tray; **f**, the TLC plate stand; **g**, the TLC plates storage; **h**, the DOBOT MG400 robot; **i**, TLC chambers; **j**, the visible light photographic device.

plate to TLC chamber, retrieving plates from the chamber, taking the TLC plate to the UV light and visible light photographic device and retrieving plates from storage. The two robots are both controlled by the computer program generally in order to avoid conflict.

1.3 The components of the automated platform

In the automated platform, we designed and utilized some devices to assist the relevant steps in TLC experiment, which are illustrated in the SI Figure 3. Design concept and function of these devices are described below in detail.

TLC chambers (SI Figure 3a). The TLC chambers are employed to store the developing eluent and thus providing the place for the developing process. In each TLC chamber, there is a fixture that is customized to limit the tilt angle of the TLC plate, which facilitates grabbing of the TLC plate. Meanwhile, each TLC chamber has a lid that is opened and closed by the suction cup on the AUBO i5 robot. Herein, multiple TLC chambers are utilized here to store different developing eluents at the same time, which greatly improves the efficiency of the automated experiments.

The sample tray (SI Figure 3b). The sample tray is a customized platform where circular depressions are distributed uniformly in 10×8 grids. The circular

a

Specifications

Robot | Axis Movement | Control Box | Teach Pendant

Degrees of freedom	6
Reach(mm)	886.5
Payload(kg)	5
Weight(kg)	24
Mounting surface diameter(mm)	Ø172
Repeatability (mm)	±0.05
Linear velocity (m/s)	≤4.0
Average power (W)	400
Peak power (W)	2000
Ambient temperature (°C)	0-50
Ambient humidity	25%-90%
Installation orientation	Any ceiling, floor, wall
IP Classification	IP54

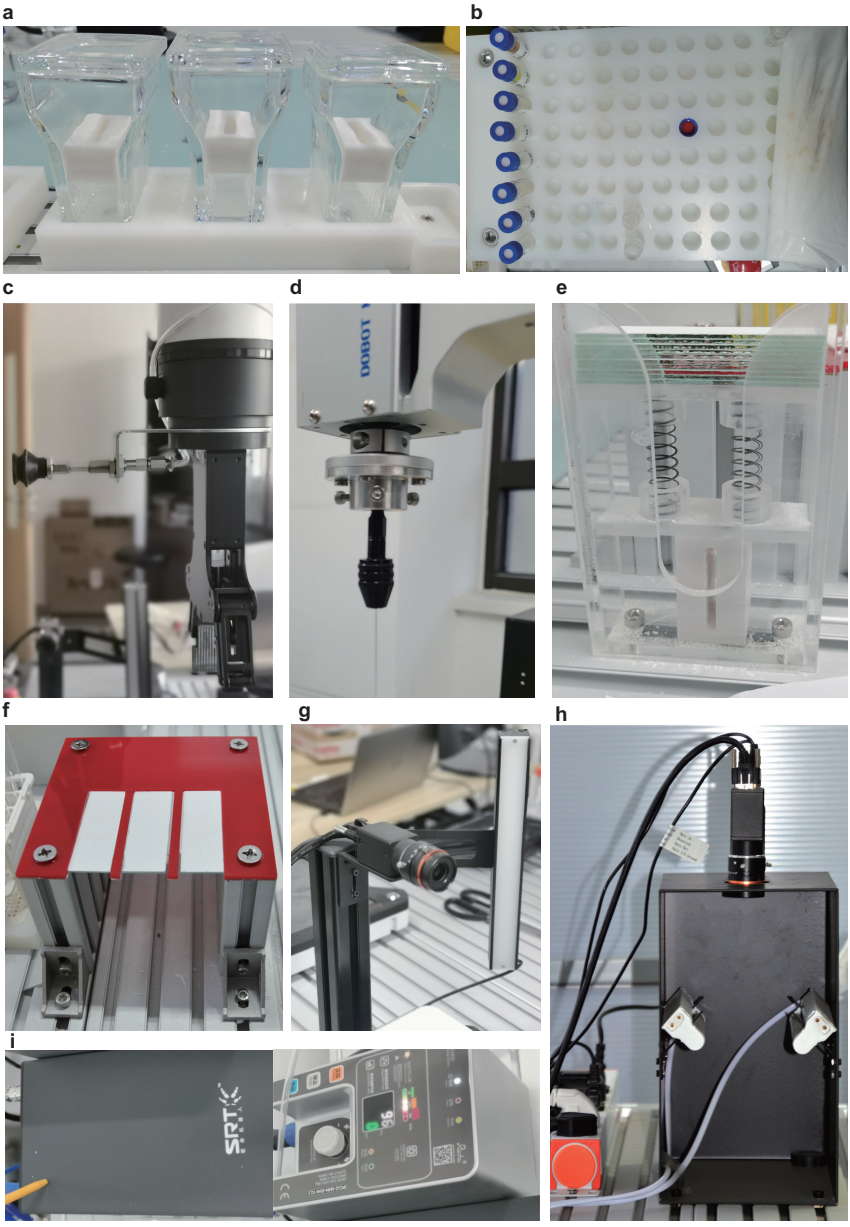
b

Controled Axes	4		
Rated Load	500 g (Max. 750 g)		
Max. Reach	440 mm		
Repeatibility	±0.05 mm		
Motion Parameters	**Joint**	**Motion Range**	**Max. Speed**
	J1	±160°	300 °/s
	J2	-25 ° ~ 85 °	300 °/s
	J3	-25 ° ~ 105 °	300 °/s
	J4	-360 ° ~ 360 °	300 °/s
Power Input	100~240 V AC, 50/60 Hz		
Rated Voltage	DC 48 V		
Rated Power	150 W		
Communication	TCP/IP, Modbus TCP		
Installation	Desktop		
Weight	8 kg		
Base Dimension	190 mm × 190 mm		
Work environment	0 °C~40 °C		
Software	DobotStudio 2020, SCStudio		
Interface	**Base:**		**End:**
	Digital Input	16	Digital Input 2
	Digital Output	16	Digital Output 2
	Ethernet	2	Air Way 1
	USB 2.0	2	
	Encoder Input	1	
	Air Way	1	

SI Figure 2. The specification for robots. a, the AUBO i5 robot; **b**, the DOBOT MG400 robot.

depression is utilized to fix the bottle of TLC sample. A piece of absorbent paper is placed at the end of the sample tray, which is employed to clean the capillary before the next sampling operation. In the experiment, the TLC samples are placed in the circular depressions orderly beforehand and four TLC samples are spotted on the same TLC plate successively in each experimental cycle of the automated platform.

The gripper and suction cup (SI Figure 3c). For AUBO i5 robot, we customized a gripper and a suction cup to facilitate its operation. Both of them are fixed on the robot by a flange. The gripper is controlled by the computer independently and the suction cup works relying on the air pump which is controlled by the IO interface. The gripper can easily pick up and down the



SI Figure 3. Close-up photographs for some components of the platform. a, TLC chambers; **b**, the sample tray; **c**, the gripper and suction cup on the AUBO i5 robot; **d**, the flange to fix the capillary; **e**, the TLC plates storage device; **f**, the TLC plate stand; **g**, the visible light photographic device; **h**, the UV light photographic device; **i**, the air pumps.

TLC plates vertically while the suction cup is able to suck up and down the TLC plates and lips of the TLC chamber from the frontage. The flange to fix the capillary (SI Figure 3d). In the design process, the biggest challenge is to fix the capillary on the DOBOT MG400 robot because the capillary is only a few millimeters thick and is very easy to break. Therefore, we design a customized flange to fix the capillary which can adjust the size of the passable radius by rotating on the thread. With this special flange, it is easy to fix the capillary by rotating it until holding the capillary tube firmly.

The TLC plates storage device (SI Figure 3e). How to store the pristine TLC plates is another challenge. To handle this issue, we design a TLC plates storage device, which is composed of plastic shell, spring and diaphragm. Before the experiment, a number of pristine TLC plates are placed on the diaphragm, which will squeeze the spring. During the experiment, the used TLC plates are dropped into the garbage can and pristine TLC plates are retrieved from storage and placed on the stand via the suction cup in each experimental cycle. Every time a TLC plate is retrieved from the storage device, the spring can push up the next TLC plate in order to be retrieved for the next time.

The TLC plate stand (SI Figure 3f). The TLC plate stand is a platform for placing the TLC plates to be spotted. From the figure, it is obvious that the stand can hold three TLC plates at the same time, which means that we can measure the R_f values of 12 compounds under the same developing eluent or 4 compounds under 3 different developing eluents in an experimental cycle.

The visible light photographic device (SI Figure 3g). In the automated platform, we construct a visible light photographic device which is utilized to record the position of the frontier of the developing eluent. The device consists of an industrial camera that is connect to the computer by router and a light that is controlled by the IO interface of the AUBO i5 robot. When the AUBO i5 robot sends the developed TLC plate to the visible light photographic device, the light is opened and the camera records a photo.

The UV light photographic device (SI Figure 3h). In TLC experiments, a lot of compounds exhibits their color only under the irradiation of ultraviolet light. Therefore, we design the ultraviolet photographic device to visualize and record the spot on the TLC board. The ultraviolet photographic device is composed of a black shading shell, a UV lamp that is controlled by the IO interface of the AUBO i5 robot, and an industrial camera that is connect to the computer by router. The black shading shell provides a black background that facilitates the subsequent computer vision task. When the AUBO i5 robot sends the developed TLC plate to the UV light photographic device, the UV light is opened and the camera records a photo. It is worth mentioning that the height of UV light is adjustable.

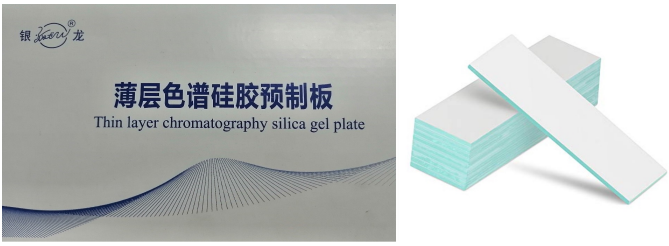
The air pumps (SI Figure 3i). The air pumps are connected to an air compressor and are controlled by the IO interface of the AUBO i5 robot.

1.4 Software

The automated platform is controlled by a computer program written in python to complete the tasks including the robot control, result storage, and R_f value identification. The devices in the automated platform are connected to the computer by a router and interfaces. The computer program can easily run on a personal PC. Once the program is started, the platform will run automatically without any extra effort of humans.

2 Experimental Information

In this work, 387 compounds are configured totally with each compound dissolved in CH_2Cl_2. The concentration is about 1-5 mg/mL. The details of these compounds can be found in the database file. There are three elution solvent systems with a total of 17 different solvent compositions including (1) Hexane/Ethyl acetate system: 1/0, 50/1, 20/1, 5/1, 3/1, 1/1, 1/2, 0/1, (2) Dichloromethane/Methanol system: 1/0, 100/1, 50/1, 30/1, 20/1, 10/1, and (3) Hexane/Diethyl ether system: 2/1, 1/1, 0/1. The brand of TLC plate is Yinlong (SI Figure 4) and the model of TLC is silica gel plate (fluorescent indicator), 75 × 25 mm, GF254, 0.23 mm, 5-20 μm.



SI Figure 4. The brand of TLC plate

In this work, the experiments are carried out automatically via automatic platform for TLC and only some prior preparations are needed manually. First, we power up the robotic platform system, and then prepare the TLC plates and put them in the TLC storage. Next, the samples are arranged on the sample tray and different proportions of elution solvents are poured into the chambers successively. At last, the automatic TLC experiments are started by running the control program. The experiments were all carried by robots automatically at 25 °C and the development time is 300 s. After the experiments, we use an image analysis program based on the computer vision to calculate R_f values automatically.

3 Machine learning methods

3.1 Random forest (RF)

Random forest is an algorithm that integrates multiple trees through the idea of ensemble learning. Its basic unit is a decision tree, and its essence belongs to ensemble learning method which is a major branch of machine learning. RF is insensitive to noise in the training dataset, and is more conducive to obtaining a robust model compared with a single decision tree because it uses a set of unrelated decision trees. Meanwhile, it is able to avoid overfitting to a certain extent. In this work, the RF algorithm is implemented by calling the function *RandomForestRegressor* in the python package *sklearn.ensemble*.

3.2 LightGBM (LGB)

LGB is constructed on the basis of gradient boosting decision tree (GBDT) algorithm, which improves calculation speed and stability by combining gradient-based one-Side sampling (GOSS) technique and exclusive feature bundling method (EFB). LGB algorithm has satisfactory performance in both regression and classification tasks in the case of large data volume and high-dimensional features. The LGB algorithm has been integrated into the *lightgbm* package and can be employed conveniently in Python.

3.3 Extreme gradient boosting (XGBoost)

XGBoost is a widely used GBDT-based machine learning method that is able to complete classification and regression tasks efficiently. The optimization goal of XGBoost is to build K regression trees to perform prediction with high accuracy and generalization ability. Therefore, the target loss function can be written as:

$$L = \sum_i \left(\hat{y}_i - y_i\right)^2 + \sum_j \Omega\left(f_j\right)$$

where

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2$$

Here, $\hat{y}_i$ is the prediction value, y_i is the true value; $\Omega(f_j)$ is the complexity of each regression tree, where T represents the number of leaf nodes and w represents the value of the node. The loss function L can be optimized by greedy algorithm. XGBoost algorithm can be quickly implemented by calling the python package *xgboost*.

3.4 Artificial neural network (ANN)

In this work, a feed forward fully-connected ANN is utilized to perform prediction, which comprises an input layer, an output layer, and one or several layer(s) between the input and output layers that are termed hidden layer(s).

Each hidden layer is composed of multiple neurons. Two adjacent layers are connected as follows:

$$z_l = \sigma\left(W_l \vec{z}_{l-1} + \vec{b}_l\right), l = 1, \ldots, L-1$$

where l denotes the layer index; W denotes the weight matrix; $\vec{b}$ denotes the bias vector; and σ denotes the activation function. Consequently, using a neural network approximation, the relationship between the input vector $\vec{z}_0$ and output prediction $\vec{z}_L$ can be expressed as:

$$\vec{z}_L = NN\left(z_0;\ \theta\right) = W_L \sigma\left(\cdots \sigma\left(W_2 \sigma\left(W_1 \vec{z}_0 + \vec{b}_1\right) + \vec{b}_2\right)\right) + \vec{b}_L$$

where θ denotes the collection of all learnable coefficients, which can be written as:

$$\theta = \{W_1, b_1, W_2, b_2, \ldots, W_L, b_L\}$$

For learning the underlying relationship between the compound properties and polarity, the input vector comprises the information about the structure, properties and eluent solvents, which is represented as $\vec{x}$; and the corresponding R_f value can be termed as $u(\vec{x})$.

Suppose that there are N TLC data, $\{u\left(\vec{x}_i\right)\}_{i=1}^{N}$. In order to train the neural network, a loss function is then defined as follows:

$$\text{Loss}(\theta) = \frac{\sum_{i=1}^{N}\left[u\left(\vec{x}_i\right) - NN\left(\vec{x}_i;\ \theta\right)\right]^2}{N}$$

In this work, the Adam optimizer is utilized to minimize the loss function for training the neural network.

4 Supplemantary experiments

4.1 Comparison between different machine learning techniques

In the paper, the coefficient of determination R^2 is utilized to measure the prediction accuracy. In this section, several other metrics including mean square error (MSE), root mean square error (RMSE) and mean absolute error (MAE) are used to show the differences between different machine learning methods more comprehensively. The calculation formulas are written as:

$$\text{MSE} = \frac{\sum_i^N \left(y_i - \hat{y}_i\right)^2}{N}$$

$$\text{RMSE} = \sqrt{\frac{\sum_i^N \left(y_i - \hat{y}_i\right)^2}{N}}$$

$$\mathrm{MAE} = \frac{\sum_i^N |y_i - \hat{y}_i|}{N}$$

$$R^2 = \frac{\sum_i^N (y_i - \hat{y}_i)^2}{\sum_i^N (y_i - \bar{y}_i)^2}$$

where N is the number of test samples, y_i is the true R_f value, $\bar{y}_i$ is the mean of true values and $\hat{y}_i$ is the predicted R_f value. Here, an 80/10/10 random split of training, validation, and test data by TLC data is utilized and 10 independent trails with different random seeds are conducted to obtain a statistical outcome, which reduces the influence of the randomness. The performance of machine learning techniques employed in this work is measured by the metrics mentioned above and the result is shown in SI Table 1. From the table, the difference between different machine learning techniques is obvious and it is discovered that the ensemble method is more accurate and stable compared with the others. Therefore, we adopt the ensemble method as the optimal model in this work.

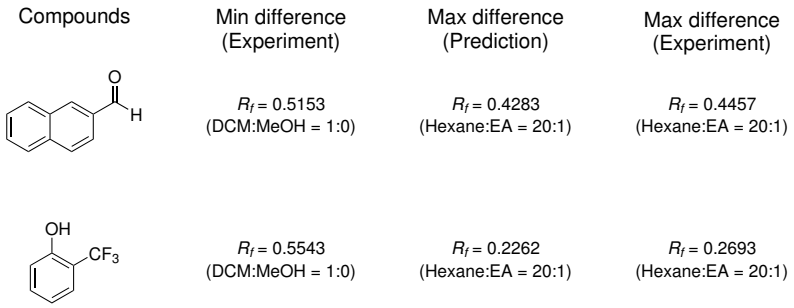
SI Table 1. The MSE, RMSE, MAE and R^2 of the prediction made by XGB, LGB, ANN, Bayesian ridge regression, Random Forest and Ensemble method, respectively. The results are displayed in the form of mean ± standard deviation, which are calculated from 10 independent trails with different random seeds.

	XGB	LGB	ANN	Bayesian	RF	Ensemble
MSE	0.0091 ±0.0013	0.0100 ±0.0015	0.0077 ±0.0016	0.0190 ±0.0016	0.0101 ±0.0021	**0.0063 ±0.0012**
RMSE	0.0951 ±0.0067	0.0997 ±0.0073	0.0872 ±0.0088	0.1378 ±0.0058	0.1002 ±0.0105	**0.0794 ±0.0072**
MAE	0.0619 ±0.0039	0.0641 ±0.0037	0.0519 ±0.0036	0.0947 ±0.0040	0.0624 ±0.0044	**0.0499 ±0.0034**
R^2	0.9147 ±0.0139	0.9091 ±0.0157	0.9300 ±0.0163	0.8277 ±0.0166	0.9077 ±0.0205	**0.9422 ±0.0127**

4.2 Application in selecting elution solvents

In TLC analysis, selecting a suitable elution solvents is the most difficult task and relies heavily on researchers' experience. In some cases, the two compounds may have little difference in R_f values under the same solvent system, resulting inseparability. Therefore, it is necessary to find a suitable solvent system to maximize the difference between the R_f values of the two compounds at this time. Under normal circumstances, this process requires repeated experimentation of different solvent systems, which is time-consuming and labor-intensive. With the predictive model, it is easy to find the optimal elution solvents by comparing the predicted values under different these conditions, which greatly improves the work efficiency. An example is provided in SI Figure 5. It is

found that the difference between the two compound is very small under DCM/MeOH system, while the difference is obvious in Hexane/EA system. Meanwhile, the R_f value under the optimal elution solvents found by prediction model is parallel with the result of repeated experiments, which shows the accuracy and power of prediction model.



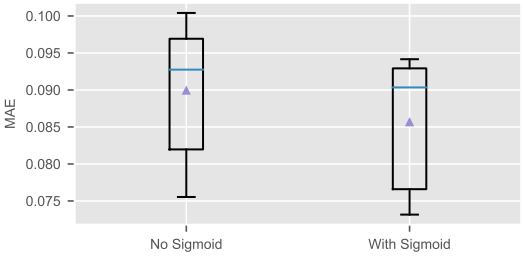
SI Figure 5. An example of selecting elution solvents from prediction model. Min difference (Experiment) and Max difference (Experiment) refer to the elution solvents and corresponding R_f values with the smallest and largest R_f values difference obtained from multiple experiments, respectively. Max difference (prediction) represents the elution solvents and corresponding R_f values with the largest R_f values difference through the prediction model.

4.3 The influence of Sigmoid function constraint

Considering the physical constraint that the R_f values are between 0 and 1, which is associated with the domain of the Sigmoid function, the formula of which is noted as:

$$f(x) = \frac{1}{1 + e^{-x}}$$

As a consequence, in this work, the machine learning techniques are employed to learn a relationship between the input information and the output, which is then mapped to the value domain of (0,1) through the Sigmoid function. In order to better demonstrate the influence of the Sigmoid function constraint, the performance of the proposed ensemble method with or without associating to the Sigmoid function are compared. Here, an 80/10/10 random split of training, validation, and test data by compounds is utilized and 10 independent trails with different random seeds are conducted to obtain a statistical outcome. The result is shown in SI Figure 6. It seems that employing the Sigmoid function as the physical constraint is able to improve the stability and accuracy of prediction. This indicates that the association of the Sigmoid function is meaningful and effective for imporving the generalization ability of the proposed prediction model.



SI Figure 6. The box pot of the performance of the ensemble method with or without associating to the Sigmoid function.

4.4 Analysis for the descriptors

In this work, we analyze the influence of the descriptors on the polarity statistically and we will provide more information here. In the article, the descriptors are analyzed in a representative system in which PE:EA=5:1. Here, the scatterplot of R_f values on TPSA of all compounds and the boxplot of R_f values of the compounds with different HBD in PE/EA system with different proportions including 1:0, 20:1, 5:1, 1:1 and 0:1 are provided in SI Figure 7. From the figure, it is discovered that the TPSA and HBD both show a decrease trend in all propotions which provides a powful evidence for the negative correlation between TPSA, HBD and R_f values.

5 Supplemantary discussion

5.1 Mathematical derivation of TLC adsorption model

We put forward a mathematic model of adsorption equilibria on the surface of silica gel. As the eluent flows on the TLC plate, both the solute molecule **A** and eluent molecule **E** generate a fast adsorption-desorption equilibrium on the surface of silica gel (SI Figure 8a). The adsorbed solute molecule, **AS**, will keep stationary on the TLC plate, while the desorbed solute molecule dissolved in the eluent, **AE**, will move together with the eluent.

Accordingly, a TLC process was imitated as shown in SI Figure 8b. For the sake of presentation, we built a 5×11 matrix of adsorption sites of silica gel on the TLC plate. At the beginning, 5 solute molecules **A** stand on the starting line. In this case, we suppose that the adsorption equilibrium is $[\mathbf{AE}]/[\mathbf{AS}] = 2/3$. At each time point, 2 solute molecules were picked randomly as **AE** and the other 3 as **AS**. Due to the fast equilibration of adsorption and desorption, the solute molecules are rapidly transformed between **AE** and **AS**. The solvent front moved upward for one frame and the desorbed **AE** moved together, while the adsorbed **AS** remained stationary. The ratio in height between the center of the **AS** distributed region and the solvent front, i.e. the R_f value, will be

the ratio of solvated **A**:

$$R_f = \frac{[\mathrm{AE}]}{[\mathrm{AE}] + [\mathrm{AS}]}$$

$$\mathrm{A}_i + \mathrm{E}_j \rightleftharpoons \mathrm{A}_i\mathrm{E}_j$$

$$K_{\mathrm{AE},ij} = \frac{[\mathrm{A}_i\mathrm{E}_j]}{[\mathrm{A}_i]\,[\mathrm{E}_j]}$$

$$\mathrm{A}_i + \mathrm{S} \rightleftharpoons \mathrm{A}_i\mathrm{S}$$

$$K_{\mathrm{AS},i} = \frac{[\mathrm{A}_i\mathrm{S}]}{[\mathrm{A}_i]\,[\mathrm{S}]}$$

$$\mathrm{E}_j + \mathrm{S} \rightleftharpoons \mathrm{E}_j\mathrm{S}$$

$$K_{\mathrm{ES},j} = \frac{[\mathrm{E}_j\mathrm{S}]}{[\mathrm{E}_j]\,[\mathrm{S}]}$$

$$[\mathrm{A}_{i,\mathrm{sol}}] = [\mathrm{A}_i] + \sum_{j=1}^{n} [\mathrm{A}_i\mathrm{E}_j]$$

$$[\mathrm{A}_i\mathrm{E}_j] = K_{\mathrm{AE},ij}\,[\mathrm{E}_j]\,[\mathrm{A}_i]$$

$$[\mathrm{A}_{i,\mathrm{sol}}\,] = [\mathrm{A}_i]\left(1 + \sum_{j=1}^{n} K_{\mathrm{AE},ij}\,[\mathrm{E}_j]\right)$$

$$\mathrm{S}_0 = [\mathrm{S}] + \sum_{i=1}^{m} [\mathrm{A}_i\mathrm{S}] + \sum_{j=1}^{n} [\mathrm{E}_j\mathrm{S}]$$

$$[\mathrm{A}_i\mathrm{S}] = K_{\mathrm{AS},i}\,[\mathrm{A}_i]\,[\mathrm{S}]$$

$$[\mathrm{E}_j\mathrm{S}] = K_{\mathrm{ES},j}\,[\mathrm{E}_j]\,[\mathrm{S}]$$

$$\mathrm{S}_0 = [\mathrm{S}]\left(1 + \sum_{i=1}^{m} K_{\mathrm{AS},i}\,[\mathrm{A}_i] + \sum_{j=1}^{n} K_{\mathrm{ES},j}\,[\mathrm{E}_j]\right)$$

$$R_{f,i} = \frac{[\mathrm{A}_{i,sol}]}{[\mathrm{A}_{i,sol}] + [\mathrm{A}_i\mathrm{S}]} = \frac{1}{1 + [\mathrm{A}_i\mathrm{S}]\,/\,[\mathrm{A}_{i,sol}]}$$

$$\frac{[\mathrm{A}_i\mathrm{S}]}{[\mathrm{A}_{i,sol}]} = \frac{[\mathrm{A}_i\mathrm{S}]}{[\mathrm{A}_i]}\frac{[\mathrm{A}_i]}{[\mathrm{A}_{i,sol}]} = K_{\mathrm{AS},i}[\mathrm{S}]\frac{1}{1 + \sum_{j=1}^{n} K_{\mathrm{AE},ij}\,[\mathrm{E}_j]}$$

$$= \frac{K_{\mathrm{AS},i}\mathrm{S}_0}{\left(1 + \sum_{i=1}^{m} K_{\mathrm{AS},i}\,[\mathrm{A}_i] + \sum_{j=1}^{n} K_{\mathrm{ES},j}\,[\mathrm{E}_j]\right)\left(1 + \sum_{j=1}^{n} K_{\mathrm{AE},ij}\,[\mathrm{E}_j]\right)}$$

$$R_{f,i} = \frac{1}{1 + K_{\mathrm{AS},i}\mathrm{S}_0 / \left(1 + \sum_{i=1}^{m} K_{\mathrm{AS},i}\left[\mathrm{A}_i\right] + \sum_{j=1}^{n} K_{\mathrm{ES},j}\left[\mathrm{E}_j\right]\right)\left(1 + \sum_{j=1}^{n} K_{\mathrm{AE},ij}\left[\mathrm{E}_j\right]\right)}$$

$$1 + \sum_{i=1}^{m} K_{\mathrm{AS},i}\left[\mathrm{A}_i\right] + \sum_{j=1}^{n} K_{\mathrm{ES},j}\left[\mathrm{E}_j\right] \approx \sum_{j=1}^{n} K_{\mathrm{ES},j}\left[\mathrm{E}_j\right]$$

$$1 + \sum_{j=1}^{n} K_{\mathrm{AE},ij}\left[\mathrm{E}_j\right] \approx \sum_{j=1}^{n} K_{\mathrm{AE},ij}\left[\mathrm{E}_j\right]$$

$$R_{f,i} = \frac{1}{1 + K_{\mathrm{AS},i}\mathrm{S}_0 / \left(\sum_{j=1}^{n} K_{\mathrm{ES},j}\left[\mathrm{E}_j\right]\right)\left(\sum_{j=1}^{n} K_{\mathrm{AE},ij}\left[\mathrm{E}_j\right]\right)}$$

Through mathematical model derivation, we can prove that the R_f value is determined by the following equation, where S_0 is the total amount of adsorption sites of silica gel.

$$R_f = \frac{1}{1 + \left(\frac{K_{\mathrm{AS}}}{K_{\mathrm{ES}}K_{\mathrm{AE}}}\right)\left(\frac{\mathrm{S}_0}{[\mathrm{E}]^2}\right)}$$

By comparing it with Sigmoid function:

$$f(x) = \frac{1}{1 + e^{-x}}$$

We find that the independent variable of Sigmoid, x, represents the change of energy between the competitive adsorption of **A** and **E**:

$$x = -\ln\left(\frac{K_{\mathrm{AS}}}{K_{\mathrm{ES}}K_{\mathrm{AE}}}\frac{\mathrm{S}_0}{[\mathrm{E}]^2}\right) = c + \frac{\Delta G^{\Theta}_{\mathrm{A},ads} - \Delta G^{\Theta}_{\mathrm{E},ads} - \Delta G^{\Theta}_{\mathrm{AE},sol}}{RT}$$

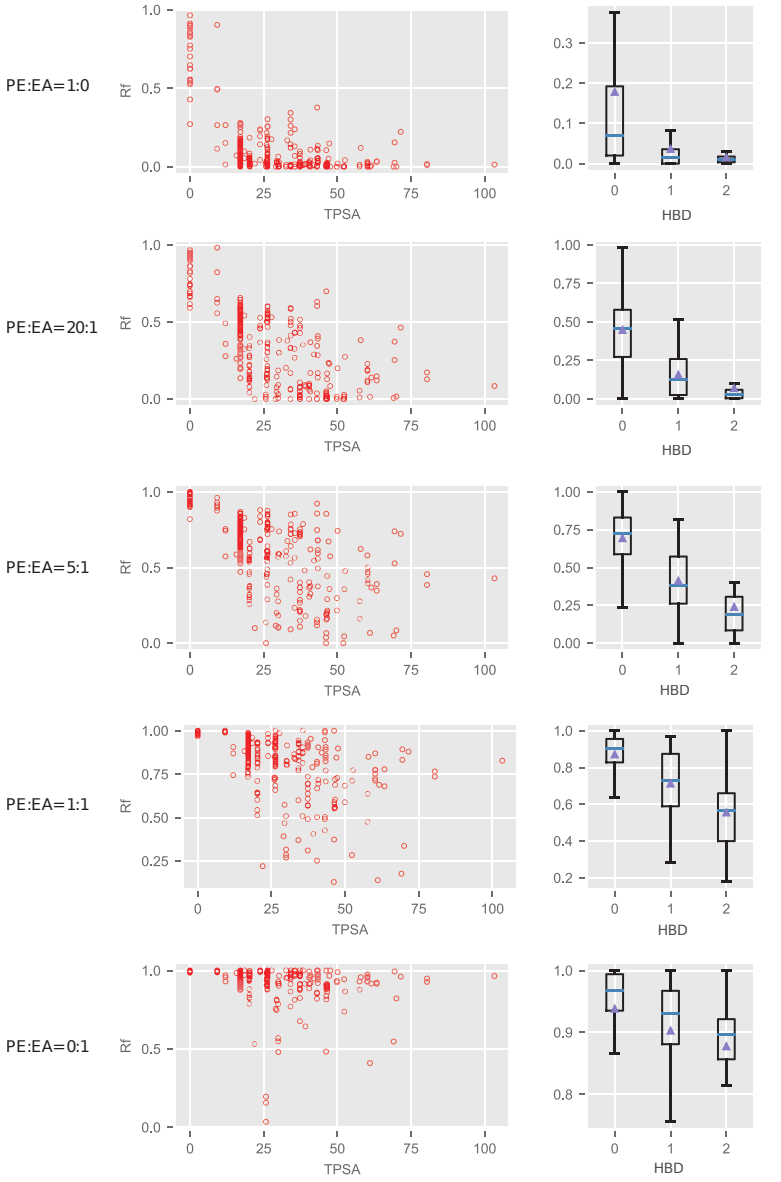
In other words, the R_f value is a Sigmoid function of the change of Gibbs free energy for the competitive adsorption reaction:

$$\mathrm{AE} + \mathrm{ES} \rightleftharpoons \mathrm{AS} + 2\mathrm{E}$$

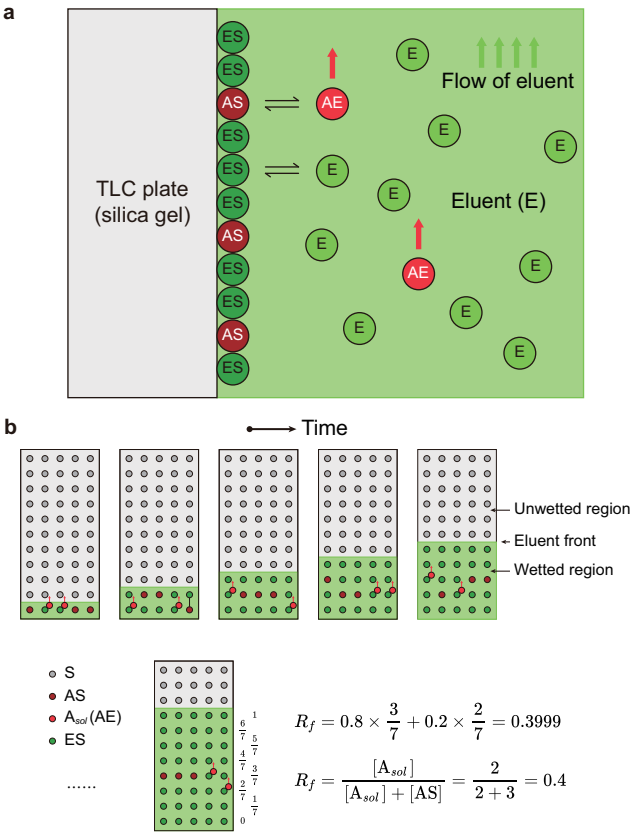
$$\Delta G^{\Theta}_{\mathrm{AS-ES}} = \Delta G^{\Theta}_{\mathrm{A},ads} - \Delta G^{\Theta}_{\mathrm{E},ads} - \Delta G^{\Theta}_{\mathrm{AE},sol}$$

$$R_f = \frac{1}{1 + e^{-\left(c + \frac{\Delta G^{\Theta}_{\mathrm{AS-ES}}}{RT}\right)}}$$

Where the constant c is based on the concentration difference between the eluent and the adsorption sites on the silica gel.



SI Figure 7. The scatterplot of R_f values on TPSA of all compounds (left) and the boxplot of R_f values of the compounds with different HBD in PE/EA system with different proportions (right). The blue triangle is the mean of R_f values for each HBD.



SI Figure 8. Adsorption model of a TLC plate. a, Adsorption equilibrium of solute (**A**) and eluent (**E**) on the surface of silica gel. **b**, Relation of R_f value and adsorption equilibrium.