High-throughput discovery of chemical structure-polarity relationships combining automation and machine learning techniques

Fanyang Mo (fmo@pku.edu.cn)
Peking University https://orcid.org/0000-0002-4140-3020
Hao Xu
Peking University
Jinglong Lin
Peking University
Qianyi Liu
Peking University
Yuntian Chen
Yongriver Institute of Technology
Jianning Zhang
Peking University
Yang Yang
University of California, Santa Barbara https://orcid.org/0000-0002-4956-2034
Michael Young
University of Toledo https://orcid.org/0000-0002-3256-5562
Yan Xu
WuXi AppTec Headquarters
Dongxiao Zhang
Southern University of Science and Technology

Physical Sciences - Article

Keywords: organic compound, thin layer chromatography, machine learning, compound polarity, Rf value

Posted Date: April 14th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1541871/v1

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, OH, 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.

Abstract

As an essential attribute of organic compounds, polarity has a profound influence on many molecular properties such as solubility and phase transition temperature. Thin layer chromatography (TLC) represents a commonly used technique for polarity measurement. However, current TLC analysis presents several problems, including the need for a large number of attempts to obtain suitable conditions, as well as irreproducibility due to non-standardization. Herein, we describe an automated experiment system for TLC analysis. This system is designed to conduct TLC analysis automatically, facilitating high-throughput experimentation by collecting large experimental datasets under standardized conditions. Using these datasets, machine learning (ML) methods are employed to construct surrogate models correlating organic compounds' structures and their polarity using retardation factor (R_f). The trained ML models are able to predict the R_f value curve of organic compounds with high accuracy. Furthermore, the constitutive relationship between the compound and its polarity can also be discovered through these modeling methods, and the underlying mechanism is rationalized through adsorption theories. The trained ML models not only reduce the need for empirical optimization currently required for TLC analysis, but also provide general guidelines for the selection of conditions, making TLC an easily accessible tool for the broader scientific community.

Keywords: organic compound, thin layer chromatography, machine learning, compound polarity, R_f value

Introduction

Thin layer chromatography (TLC) is a commonly used technique in modern chemistry and biology laboratories. As a key chromatography technique, the employment of a solid stationary phase and a liquid mobile phase allows for the separation of individual components of a complex mixture on the basis of their relative affinities for the two phases (Figure 1a)[1]. TLC analysis is currently used routinely for reaction monitoring, product identification, and determination of chromatography conditions for subsequent purification. While highly experienced synthetic practitioners are able to use this tool, TLC techniques often present a significant hurdle for scientists in synthesis-adjacent fields. Furthermore, the identification of TLC conditions for new compound classes requires the judicious selection of several variables, most notably the mobile phases and their ratios, to achieve optimal separation. Traditionally, such goals are accomplished through trial-and-error in an extremely time and labor-intensive manner.

In recent years, cutting-edge techniques in artificial intelligence (AI) have revolutionized the extrapolation of structure-property relationships in the chemical sciences[2]. In particular, machine learning (ML) algorithms are able

to solve complex chemical problems with respect to prediction, surrogate model construction, and constitutive relationship discovery[3–6]. In this vein, we postulated that a trained model might possibly work on predicting the polarity and specifically the R_f value of organic compounds due to the strong structure-property relationship in the physical mechanism underlying TLC.

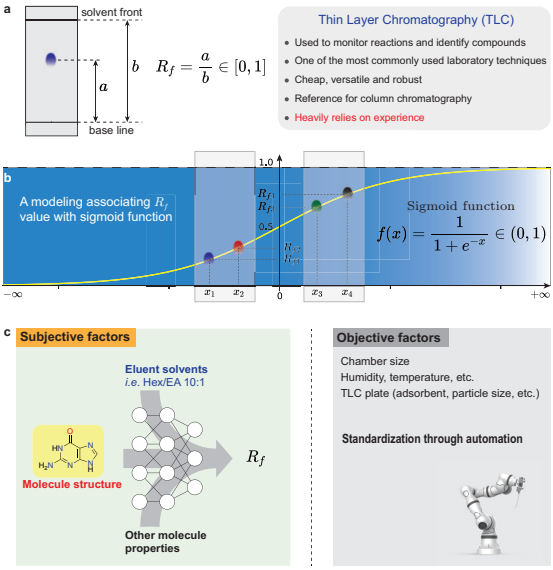


Fig. 1 Context of the work. a, Thin-layer chromatography (TLC) is a chromatography technique used to separate non-volatile mixtures. Synthetic laboratories heavily use TLC techniques to monitor reactions and identify compounds daily. Choosing suitable TLC conditions is usually time-consuming for novices or for new compounds. The retardation factor (R_f) is the fraction of an analyte in the mobile phase of a chromatographic system. It is defined as the ratio of the distance traveled by the center of a spot to the distance traveled by the solvent front. **b**, A sigmoid function is a mathematical function having a characteristic "S"-shaped curve, and has domain of all real numbers with a return value in the range 0 to 1. Considering that the R_f value also has the same value range, we deliberately associate it with sigmoid function. **c**, The subjective and objective factors to compound R_f value measurement. The subjective factors include the compound's structure and other physical properties, as well as elution solvents. The information can be mapped to a vector space via feature engineering, and then fed to ML algorithms. Other factors like chamber size, humidity, etc., can also affect the measurement. The influence of these objective factors should be eliminated as much as possible to avoid their impact on model training.

Essentially, the task of prediction is to establish a certain mathematical model that maps between input and output. We noted that the sigmoid function, a common activation function in neural network algorithms, is tightly bound to the pair of horizontal asymptotes and highly useful in compressing or squashing outputs within a 0 to 1 range. The output of TLC, R_f, is also a value between 0 and 1. As such, we envisioned that R_f values could be reasonably fitted using a sigmoid function (Figure 1b). In this case, all the factors that have an influence on R_f value will be selected and subject to a matrix

operation through certain input forms to determine an x, and the output, R_f value, will be the sigmoid function value of the x. Additionally, in a reverse perspective, through such a transformative operation, the original range of 0 to 1 is extended to the entire real number domain. In other words, we can predict the R_f value by regression in a larger range to improve the accuracy of the model (*vide infra*).

ML methods capture underlying patterns from a large amount of training data in order to make prediction. The availability of large, high-quality datasets is the prerequisite for ML methods. Although the R_f value of newly-prepared organic compounds are available in the chemical literature, the lack of standardization often leads to inconsistent data, which impedes the development of ML models. Generating a sufficient amount of highly standardized data through conventional means is tedious and time-consuming. To address this challenge, we sought to exploit automated instrumentation to accelerate and standardize the measurement of R_f values of organic compounds in TLC analysis. Furthermore, we use these data for ML model training to correlate compound structures and their polarity (Figure 1c).

Data acquisition

The automation of chemical experimentation has been an expanding field in the last decade[7–9]. These automated platforms enable standardization and enhance reproducibility while reducing experimental costs. Most importantly, automation allows for the generation of sufficient data for further statistical analysis. The research paradigm combining automation and ML techniques has already been successfully applied to chemical sciences in various scenarios ranging from reaction condition optimization to mechanism interpretation[10–15].

We realized that the most effective strategy would be to develop an automated robotic platform to collect the necessary TLC data. However, TLC analysis requires several experimental steps, including dissolving the analyte, spotting the analyte, developing the TLC plate, and ultimately measuring and calculating the R_f value for each analyte, thus demanding sophisticated automation.

To address the challenges mentioned above, we built a custom desktop robot system for high-throughput collection of TLC data. Our robotic platform is shown in Figure 2. Two collaborative robots are the core of this system, which are able to accomplish complex actions like human arms with high precision and safety. The system also comprises two cameras, two lights, a router, and a laptop (Figure 2a). The smaller robot, DOBOT MG400, is responsible for drawing TLC samples on the sample tray (Figure 2b), and then spotting samples onto the TLC plates (Figure 2c). Subsequently, the bigger robot, AUBO i5, is responsible for gripping the TLC plate, and then putting it into the chamber for development (Figure 2d & e). Upon completion, the TLC plate is transferred to a box for visualizing and photographing under UV light

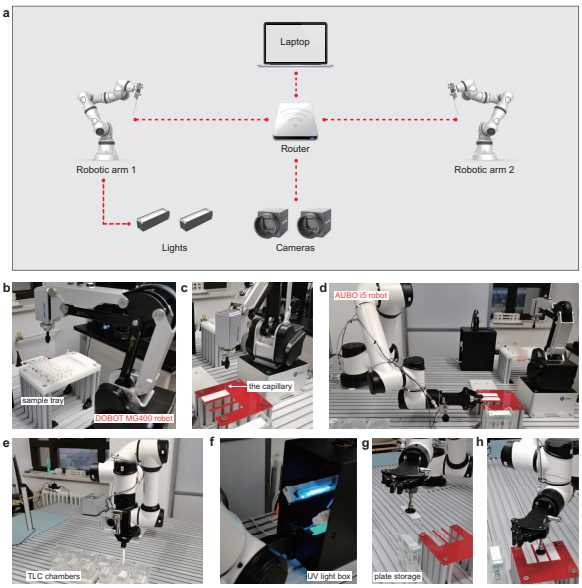


Fig. 2 Automated thin layer chromatography robots and experimental station. **a**, Schematic of the automated robotic platform for TLC data collection. **b**, The DOBOT MG400 robot equipped with a capillary is drawing TLC samples on the sample tray. **c**, The DOBOT MG400 robot is spotting samples onto the TLC plates. **d**, The AUBO i5 robot is gripping a TLC plate prior to moving. **e**, The AUBO i5 robot is putting a TLC plate to TLC chamber (or retrieving a plate from the chamber). **f**, After developing, the AUBO i5 robot is taking the TLC plate to a UV light box to visualize and photograph. **g**, The AUBO i5 robot is retrieving a plate from storage. **h**, The AUBO i5 robot is placing a pristine TLC plate onto the stand.

(Figure 2f). Finally, the AUBO i5 will retrieve a pristine plate from the storage, and place it onto the stand for regenerating the system for the next usage (Figure 2g & h). In order to increase throughput and efficiency, six chambers are placed in the platform with different elution solvents in each. A Python program was built to control the robots, cameras and lights, thus managing the whole workflow. Through this robotic TLC experiment platform, we achieved high-throughput standardized polarity measurements.

After the data collection is completed by the automated TLC system, the R_f values can be automatically calculated by an image analysis computer program, leading to highly standardized TLC runs. For UV-inactive compounds, other visualization methods can be adopted. In a typical task, the R_f values of each compound under different elution solvent ratios are plotted, which helps chemists find outliers based on intuition and experience. Through this method, a high-quality dataset of compound R_f values is established, containing 4944 standardized polarity measurements from 387 organic compounds (Figure 3a) under three elution solvent systems including hexane/ethyl acetate, dichloromethane/methanol and hexane/diethyl ether with a total of 17 different solvent composition (Figure 3b). These typical compounds were collected deliberately based on their types to get better representation. This standard

polarity dataset generated for the first time is very precious since the R_f values are usually sensitive to external factors in the experiments, which means that statistically meaningful data can only be obtained through automated high-throughput standardized experiments.

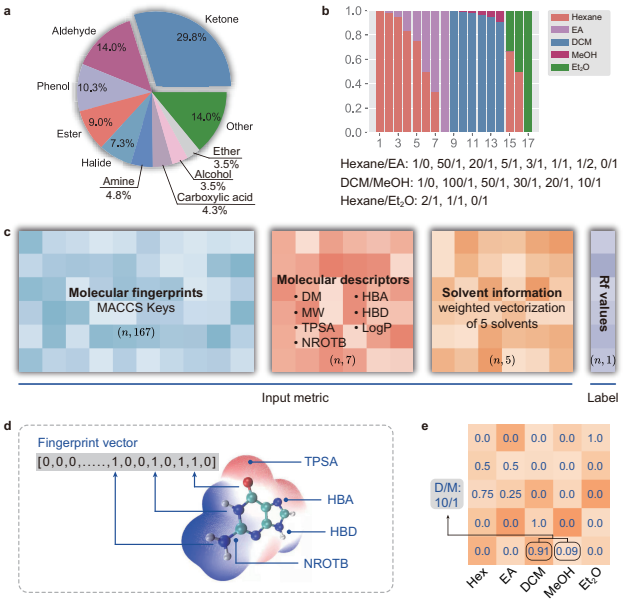


Fig. 3 Description of obtained polarity dataset and data processing. a, The top 10 classes of the compounds in the dataset. **b**, The developing agent and their ratios utilized in this work including EA, ethyl acetate, DCM, and dichloromethane. **c**, The input matrix includes molecular fingerprints, several molecular descriptors and solvent information. The label is the R_f value. DM, dipole moment; MW, molecular weight; TPSA, topological polar surface area; NROTB, number of rotatable bonds; HBA, hydrogen bond acceptor; HBD, hydrogen bond donor; LogP, lipid-water partition coefficient. **d**, The MACCS (Molecular ACCess System) keys are one of the most commonly used structural keys. In these structural keys, the structure of a molecule is encoded into a binary bit string (a sequence of 0's and 1's), each bit of which corresponds to a "pre-defined" structural feature. This binary bit string makes up the fingerprint vector. **e**, Schematic diagram of weighted vectorization of 5 solvents.

Machine learning model for polarity prediction

Before constructing the machine learning model, we first turned to selecting suitable molecular descriptors to represent the compounds in the dataset. In terms of mechanism, molecular polarity is a reflection of a compound's polar bonds and their spatial distribution, thus essentially, its structure. In addition, TLC is based on the principle of separation through adsorption type. For this reason, polarity represented by R_f value is also correlated with the properties that affect the molecular adsorption between stationary phase and mobile phase, for example, hydrogen bonding and topological polar surface area, etc.

While many molecular fingerprints have been developed, we pay more attention to the ones that related to structure. Molecular Access System (MACCS) keys are one of the most commonly used structural keys that are employed to represent molecular structure, fragments, and substructural information[16]. Each compound can be converted into MACCS keys with 167 dimensions. Meanwhile, molecular properties such as molecular weight (MW), topological polar surface area (TPSA), and many others, can also be conveniently extracted by using RDKit software[17]. Moreover, another possible factor, dipole moment, was calculated using Mopac2016 (PM6-D3)[18]. For the elution solvents, vectorization encoding technique was used to express mobile phase information. The values in individual solvent column represent their ratio for a combination of an eluent. To avoid prohibitively time-consuming analysis and logging of computational data, we developed software to automate feature generation. The program requires only the input of SMILES strings[19] of compounds in a Python script. The program then generates the data table that can be used for modeling. In total, 179-dimensional input metrics were extracted by the software to characterize each TLC set of conditions (Figure 3c-e). One of the advantages of ML modeling is that one can define feature engineering relatively freely without recourse to a specific hypothesis. This, to a certain extent, allows the machine to discover the connection, or even knowledge hidden in data without relying on human experience.

With these data in hand, we evaluated the predictive accuracies of a series of ML methods including Bayesian regression, Random Forest (RF), LightGBM (LGB), XGBoost (XGB), and Artificial Neural Network (ANN). Here, the dataset is randomly divided into training, validation, and test set according to the ratio of 80%, 10%, and 10% by TLC data. Considering the physical constraint that the range of R_f values is between 0 and 1, in the training process of the above-mentioned ML algorithms, the output will be mapped from the entire real number domain to [0,1] through the sigmoid function. It is discovered that a deep learning-based algorithm like ANN and decision tree-based algorithms like RF, LGB and XGB all show satisfactory prediction ability with R^2 over 0.93 (Figure 4a). To avoid randomness, 10 independent experiments with different random seeds are conducted and the results are shown in Extended Data Figure 1. On the basis of these techniques, an ensemble method is proposed to form a better model and further improve the accuracy. We are delighted to find that a predictive model with the highest accuracy ($R^2 = 0.961$) can be obtained by a simple weighted average of these methods. This improvement may be attributed to the ablity of the ensemble algorithm to avoid overfitting and decrease the risk of obtaining a local minimum[20].

Next, a more difficult and practical problem is considered, that of predicting the R_f values of out-of-sample compounds under different solvent composition. In this task, the dataset is randomly split into a training dataset with 3922 R_f values from 308 compounds (80% of total compounds), a validation dataset with 484 R_f values from 38 compounds (10% of total compounds), and an out-of-sample test dataset with 473 R_f values from 38 compounds (10% of

total compounds).The proposed ensemble model is trained and the prediction performance on the test set is examined. On average, the MSE of the predicted R_f values of the out-of-sample 38 compounds was 0.117, with an R^2 value of 0.887 (Figure 4b). The effective out-of-sample prediction of the proposed model suggests that the constitutive relationship between the input information and output polarity measurements were captured well by the prediction model, which means that it may be possible to predict the R_f values of compounds under specified developing systems without experiments.

In order to explore the relationship between the predictive power of the model and input data more deeply, different numbers of compounds are utilized to train the model while the test set is fixed. It is surprising to discover that the prediction model is able to maintain satisfactory predictive power with a markedly smaller training data since the R^2 of the model trained from R_f values of 116 compounds (30% of total compounds) achieves 0.870, which is only a decrease of 0.03 compared with the model trained from 70% of the total compounds (Figure 4c). This indicates that there is a tight physical relationship between the structure and properties of the compound and the R_f value that could be learned by the proposed ensemble model through only 116 compounds, so as to predict the polarity of countless kinds of organic compounds, which provide an excellent annotation of "small data, big task"[21].

Statistical analysis for descriptors

After having obtained a satisfactory predictive model, we sought to explore the physical and chemical knowledge underlying the trained model. The molecular descriptors that characterize molecular properties are contained in the input information. As a consequence, the relative importance of utilized molecular descriptors are evaluated by the percent increase in the predictive model's mean absolute error (MAE) when values of certain descriptors in the test set are randomly reassigned according to data distribution. It was found that, among these descriptors, TPSA shows significant importance over all others (Figure 4d). The correlation coefficients between molecular descriptors was also explored (Figure 4e). It was found that TPSA and HBA have a very strong positive correlation. The TPSA of a molecule is defined as the surface sum over all polar atoms, primarily oxygen and nitrogen, which is often used to evaluate the transport properties of drugs in cells, and it is proven to have an inseparable relationship with polarity in this work. We rationalize that a larger TPSA leads to a stronger interaction between the adsorbate and the adsorbent (e. g. silica gel), leading to smaller R_f values. This is proved in Figure 4f where the spearman coef is -0.654, which indicates a strong negative correlation. As for HBD and HBA, they indicate the number of hydrogen bond donors and acceptors, respectively. Our results show that HBD is more relevant to R_f value over HBA and the R_f values present a clear downward trend as HBD rises which is illustrated in Figure 4g. The explanation is that the solid phase used in this study is silica gel, which contains many bridge oxygens and hydroxyl

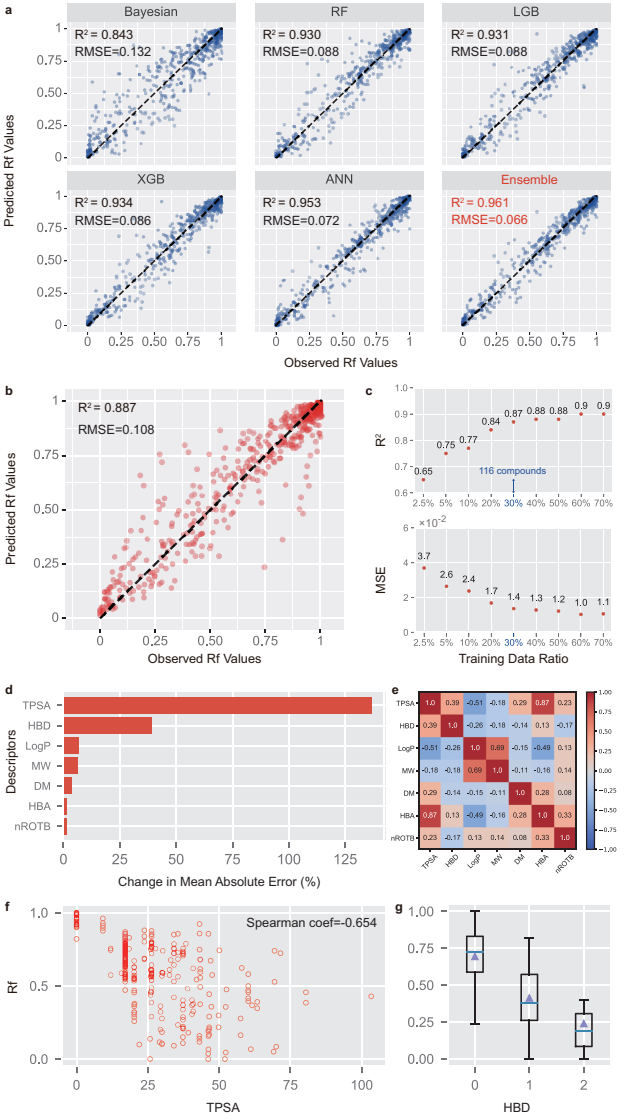


Fig. 4 Polarity prediction with machine learning techniques and the analysis of the descriptors. a, Observed versus predicted R_f values for different ML methods. For all the models, an 80/10/10 random split of training, validation, and test data by TLC data was performed to measure the prediction ability of each model. Only test set data are shown in plots. The dashed line is the $y = x$ line. **b**, Observed versus predicted R_f values for the proposed ensemble method to predict out-of-sample compounds. The dataset is randomly split into 80/10/10 by compounds and only test set data are shown in plots. The dashed line is the $y = x$ line. **c**, Test set performance of the ensemble prediction model with sparse data. The smaller training sets were selected randomly from the entire compounds and the test data are kept the same. A gradual erosion in predictive accuracy occurred from 70% of the entire compounds down to 2.5%. **d**, The relative importance of the molecular descriptors utilized in this work. **e**, The heatmap of correlation coefficients between molecular descriptors. **f**, The scatterplot of R_f values on TPSA of all compounds when PE/EA=5/1. **g**, The boxplot of R_f values of the compounds with different HBD when PE/EA=5/1.

groups on the surface acting as hydrogen bond accepters. As such, the analyte will have more binding interactions with silica gel when it contains hydrogen bond donors rather than hydrogen bond accepters. The relationship between polarity and descriptors is indistinct before and is revealed statistically in this work for the first time.

Application and expansibility

The predictive model has a wide range of applications that will facilitate the acquisition of polarity data and deepen respective understanding. Its primary utilization is to predict the R_f curve of a given compound. Three out-of-sample compounds where the R_f curves present different patterns are examined. The predicted and experimental R_f curves under three elution solvent systems are illustrated in Figure 5. It is discovered that the predicted curves have achieved satisfactory accuracy in all systems although the patterns of curves are different. On this basis, the predictive model is able to provide prior information of the R_f values in different elution solvent systems before chemical experiments, which averts repreated trails for selecting an appropriate system in common practice. Furthermore, from the predicted R_f curves, it is easy to discover the optimal elution solvent system for seperating two compounds, which is a common and significant requirement in organic chemistry and an example is provided in the supplementary information.

Considering the wide variety of organic compounds, it is difficult to complete a sweep of all classes of compounds, therefore, we select some representative classes such as Ketone, Aldehyde, Phenol and others to form the polarity dataset. Although the predicted model has been proven to possess a satisfactory accuracy when predicting the R_f curves of the classes involved in the dataset, the expansibility for other classes that have never appeared in the dataset is also a challenge. As a consequence, the predictive model is employed to predict the polarity of saccharides and the results are displayed in Extended data Figure 2. Properties of saccharides are totally different from the classes considered in the dataset, which brings a challenge to the predictive model. However, it is surprising to find that the accuracy of prediction is slightly affected faced with compounds of new classes, which means that it has great expansibility and can be adapted to predict other classes in some degree. The result implies that the ML model actually learns the strong relationship between the compound properties and polarity, and thus achieving the ability for making reasonable extrapolations.

Conclusion

TLC experiments are performed extensively every day in synthetic laboratories around the world. Determination of suitable TLC conditions usually requires ample experience of chemists, and a large number of attempts would be inevitable. Here we have built a robotic experimentation platform and its

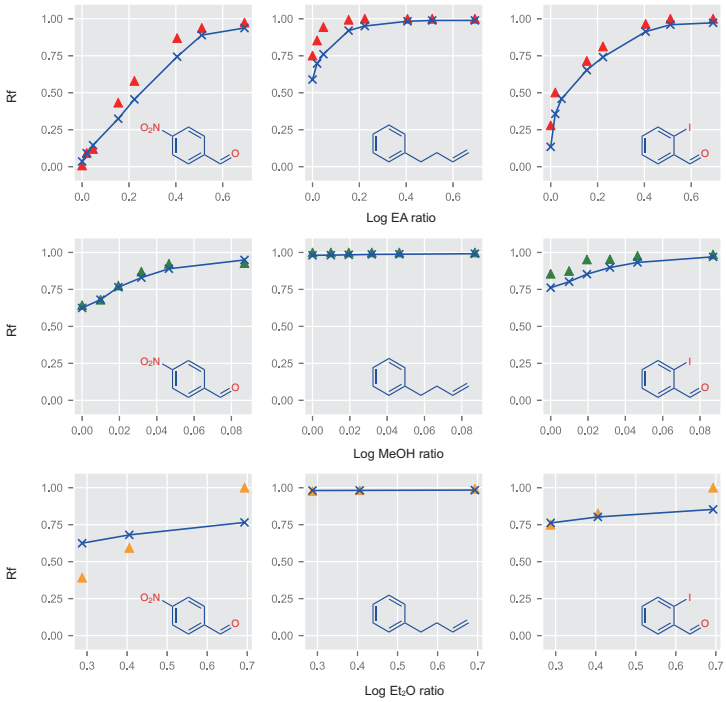


Fig. 5 Predicted R_f curves. Predicted R_f curves and experimental R_f curves for three out-of-sample compounds under three elution solvent systems including hexane/ethyl acetate (upper), dichloromethane/methanol (middle) and hexane/diethyl ether (lower). The log ratio is calculated as $log(r + 1)$ where r is the ratio of eluent.

software for TLC data collection, which generates a large amount of standardized compound's structure-polarity data automatically. Using the dataset, we further developed ML models that can predict compound's R_f values and experiments have proved that the proposed predictive model shows satisfactory accuracy and expansibility, which has a hopeful prospect for facilitating the acquirements and application of R_f values. From the analysis of descriptors, the influencing factors of polarity are revealed statistically. Meanwhile, the success of the predictive model also indicates that the strong relationship between the compound properties and polarity has been learned well. We expect that this TLC prediction approach will prove to be useful to the synthetic community in facilitating laboratory efficiency.

Supplementary information. Supplementary files will be available along with the publication of this article.

Acknowledgments. We thank Prof. Suwei Dong at Peking University Health Science Center, Prof. Shufeng Chen at Inner Mongolia University and Prof. Qing Xiao at Third Military Medical University for generously providing some organic compounds.

Declarations

Funding. This work is supported by the Natural Science Foundation of China (Grant Nos. 22071004, 21933001, 22150013).

Competing interests. F.M., H.X. and D.Z. are inventors on two patent applications (CN 202111638511.2 and 202122346010.9) submitted by Peking University that cover an organic chemistry laboratory automation system and a machine learning method for TLC conditions prediction, respectively.

Availability of data and materials. The TLC dataset is available on the website https://github.com/woshixuhao/ML-Prediction-for-Rf-values. A video associated with this project can be found via the following links:

English version: https://www.bilibili.com/video/BV1am4y1o7yE/

Chinese version: https://www.bilibili.com/video/BV17R4y1j7jz/

Code availability. The code is available on the website https://github.com/woshixuhao/ML-Prediction-for-Rf-values.

Authors' contributions. H.X. and F.M. built the robotics platform for high-throughput experimentation. J.L. performed TLC experiments. H.X. performed chemoinformatic and machine learning studies. Q.L. derived the theoretical formula for TLC mechanisms. J.Z. and Q.L. performed DFT calculation. All authors analysed the data. H.X. and F.M. wrote the manuscript. F.M. conceived the idea and designed the overall research. F.M. and D.Z. supervised the whole project.

References

[1] Sherma, J. & Fried, B. *Handbook of thin-layer chromatography* (CRC press, 2003).

[2] Muratov, E. N. *et al.* Qsar without borders. *Chemical Society Reviews* **49** (11), 3525–3564 (2020) .

[3] Yang, Q. *et al.* Holistic prediction of the pka in diverse solvents based on a machine-learning approach. *Angewandte Chemie International Edition* **59** (43), 19282–19291 (2020) .

[4] Howarth, A., Ermanis, K. & Goodman, J. M. Dp4-ai automated nmr data analysis: straight from spectrometer to structure. *Chemical Science* **11** (17), 4351–4359 (2020) .

[5] Reid, J. P. & Sigman, M. S. Holistic prediction of enantioselectivity in asymmetric catalysis. *Nature* **571** (7765), 343–348 (2019) .

[6] Kirkpatrick, J. *et al.* Pushing the frontiers of density functionals by solving the fractional electron problem. *Science* **374** (6573), 1385–1389 (2021) .

[7] Ley, S. V., Fitzpatrick, D. E., Ingham, R. J. & Myers, R. M. Organic synthesis: March of the machines. *Angewandte Chemie International Edition* **11** (54), 3449–3464 (2015) .

[8] Häse, F., Roch, L. M. & Aspuru-Guzik, A. Next-generation experimentation with self-driving laboratories. *Trends in Chemistry* **1** (3), 282–291 (2019) .

[9] Wilbraham, L., Mehr, S. H. M. & Cronin, L. Digitizing chemistry using the chemical processing unit: From synthesis to discovery. *Accounts of Chemical Research* **54** (2), 253–262 (2020) .

[10] Ahneman, D. T., Estrada, J. G., Lin, S., Dreher, S. D. & Doyle, A. G. Predicting reaction performance in c–n cross-coupling using machine learning. *Science* **360** (6385), 186–190 (2018) .

[11] Granda, J. M., Donina, L., Dragone, V., Long, D.-L. & Cronin, L. Controlling an organic synthesis robot with machine learning to search for new reactivity. *Nature* **559** (7714), 377–381 (2018) .

[12] Steiner, S. *et al.* Organic synthesis in a modular robotic system driven by a chemical programming language. *Science* **363** (6423) (2019) .

[13] Burger, B. *et al.* A mobile robotic chemist. *Nature* **583** (7815), 237–241 (2020) .

[14] Newman-Stonebraker, S. H. *et al.* Univariate classification of phosphine ligation state and reactivity in cross-coupling catalysis. *Science* **374** (6565), 301–308 (2021) .

[15] Shields, B. J. *et al.* Bayesian reaction optimization as a tool for chemical synthesis. *Nature* **590** (7844), 89–96 (2021) .

[16] Durant, J. L., Leland, B. A., Henry, D. R. & Nourse, J. G. Reoptimization of mdl keys for use in drug discovery. *Journal of Chemical Information and Computer Sciences* **42** (6), 1273–1280 (2002) .

[17] Landrum, G. Rdkit documentation. *Release* **1** (1-79), 4 (2013) .

[18] Stewart, J. J. Stewart computational chemistry. *http://openmopac. net/* (2007) .

[19] Weininger, D. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences* **28** (1), 31–36 (1988) .

[20] Sagi, O. & Rokach, L. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **8** (4), e1249

(2018) .

[21] Qi, G.-J. & Luo, J. Small data challenges in big data era: A survey of recent progress on unsupervised and semi-supervised methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020) .

Methods

The automation platform for TLC data collection

In this work, an automated TLC analysis system (Auto-TLC system) that can automatically complete the entire TLC analysis process is developed. The core of Auto-TLC system is two collaborative robots DOBOT MG400 robot equipped with a capillary and AUBO i5 robot equipped with a mechanical gripper and a sucker driven by air pump. The solution sample stored on the sample tray is dipped by the capillary and then spotted onto the TLC plates placed on the rack. It is worth noting that the special rack is customized and can place several TLC plates at the same time for high-throughput experimentation. Then, the AUBO i5 robot uses the gripper to transfer the spotted TLC plates to a 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 elution solvents at the same time, which greatly improves the efficiency of the system. When the developing time reaches the designated value (300 seconds in this work), AUBO i5 robot will retrieve the corresponding TLC plate from the chamber and send it to the ultraviolet and visible light camera devices respectively to take photos and record the result. The visible light camera device is utilized to record the position of the frontier of the developing eluent and the ultraviolet camera device is employed to visualize and photograph the spot on the TLC plate. The used TLC plate is dropped into the waste container and pristine TLC plates are retrieved from storage via the sucker. This cycle will continue until all samples have been tested.

The image recognition algorithm for calculating the R_f values

In the Auto-TLC system, the R_f values are calculated automatically from an image recognition algorithm based on the recorded photos, which is shown in Extended data Figure 3. For each experiment, four samples are spotted on one TLC plate and two photos are taken from ultraviolet and visible light camera devices, respectively. Benefitting from the Auto-TLC analysis platform, the TLC plate is spotted standardized so that the initial position of TLC spot (x_s, $\mathrm{H_L}$) ($s = 1, 2, 3, 4$) fixed beforehand. In order to calculate the R_f value, the height of each TLC spot $\mathrm{H_s}$, and the height of solvent front $\mathrm{H_{Fa}}$ and $\mathrm{H_{Fb}}$ are identified by the projection method. Prior to the projection, the main part of the TLC board can be easily distinguished from the black background by the threshold method. In order to prevent the four TLC spots from interfering with each other during recognition, the projection is made merely in the

neighborhood of each TLC spot $[x_s - \Delta x, x_s + \Delta x]$, and the size of the neighborhood Δx is selected according to the distance between the points which is fixed in advance. The projection curve of the first TLC spot is taken as an example in Extended Data Figure 3. It is obvious that the curve has two minimum values and $\mathrm{H_{Fa}}$ and $\mathrm{H_s}$ can be identified from them, since only these two positions can appear black under ultraviolet light in the neighborhood, thereby significantly reducing the average pixel value under the y-axis projection. In the same way, under visible light irradiation, the solvent front is the brightest, so its y-axis projection curve has a maximum value at this position, which corresponds to the value of $\mathrm{H_{Fb}}$. Considering that the solvent front is not apparent in some cases and cannot be accurately identified under ultraviolet light, the $\mathrm{H_{Fb}}$ identified from the photo taken under visible light can play an important complementary role.

Since the focal lengths of the cameras are predetermined, the mapping relationship between the pixel distance on the photo and the actual distance can be determined in advance. Here, the mapping functions for ultraviolet and visible light camera device are referred as f_a and f_b, respectively. Therefore, the actual height of the solvent front $\mathrm{H_F}$ is expressed as:

$$\mathrm{H_F} = \begin{cases} \frac{1}{2}\left(f_a\left(\mathrm{H_{Fa}}\right) + f_b\left(\mathrm{H_{Fb}}\right)\right), \text{ if } \frac{\left|f_a\left(\mathrm{H_{Fa}}\right) - f_b\left(\mathrm{H_{Fb}}\right)\right|}{f_b\left(\mathrm{H_{Fb}}\right)} < 5\% \\ f_b\left(\mathrm{H_{Fb}}\right), \text{ other condition} \end{cases}$$

Here, other conditions includes the difference between $\mathrm{H_{Fa}}$ and $\mathrm{H_{Fb}}$ is large, or $\mathrm{H_{Fa}}$ is not identified successfully. With the calculated $\mathrm{H_s}$ and $\mathrm{H_F}$, the R_f value can be finally calculated as:

$$R_{fs} = \frac{f_a\left(\mathrm{H_s}\right) - f_a\left(\mathrm{H_L}\right)}{\mathrm{H_F} - f_a\left(\mathrm{H_L}\right)}$$

With the image recognition algorithm, R_f values can be identified quickly and automatically. During the experiments, there may emerge some extreme situations that the algorithm is difficult to judge, such as tailing and fusion of two TLC points. As a consequence, after the high-throughput experiment is completed, the entire result will be manually verified by expert's experience to deal with the extreme situations mentioned above to guarantee the accuracy and reliability of the obtained datasets. A program of visual interface is developed that makes manual verification easy by simply clicking on the correct TLC spot, starting point and solvent front on the image and the image recognition algorithm will correct the mistakes automatically. It is worth mentioning that the extreme situations are rare since the correct R_f value can be obtained through image recognition in most experiments.

Description of the polarity dataset

In this work, a valuable dataset of compound polarity is obtained from Auto-TLC analysis platform. In order to ensure the diversity of compounds, the R_f curves of 387 organic compounds, including ketone, aldehyde, ether, halide, alcohol and other categories, under 17 different solvent composition

are measured, thereby collecting 4944 standardized polarity measurements after removing outliers. Some of the compounds are shown in Extended data Figure 4. Three elution solvent systems including hexane/ethyl acetate (EA), dichloromethane (DCM)/methanol (MeOH) and hexane/diethyl ether (Et_2O) are utilized in this work. For hexane/EA system, eight proportions are employed, namely 1:0, 50:1, 20:1, 5:1, 3:1, 1:1, 1:2 and 0:1. For DCM/MeOH system, six proportions are employed including 1:0, 100:1, 50:1, 30:1, 20:1 and 10:1. For hexane/Et_2O system, three proportions are employed including 2:1, 1:1, 0:1.

Dataset preprocessing

Before machine learning, the dataset acquired from Auto-TLC analysis platform needs preprocessing. Considering that chemical formulas cannot be directly used as input for machine learning, the molecular fingerprint and descriptors are extracted in advance. In this work, the molecular fingerprint Molecular Access System (MACCS) keys are employed to represent molecular structure, fragments or substructure information. MACCSkeys is a one-hot vector with a length of 167 bits, each bit represents a specific molecular structure, 1 represents the structure exists, and 0 represents the structure does not exist. The meaning of each bit can be found in open resource. It is worth mentioning that although each compound corresponds to a unique MACCkeys, each MACCkeys may correspond to several compounds with a similar structure. For example, 134^{th} bit represents the existence of halogen (Cl, Br, I). Therefore, bromobenzene and chlorobenzene have the same MACCkeys since the 134^{th} bit of both is 1 while other substructure is the same. This feature of MACCkeys allows it to extract common features (such as the presence or absence of certain substructure) from countless compounds to predict polarity, thereby achieving the goal of "small data, big tasks". This is why the polarities of new compounds can be well predicted by only relying on hundreds of different kinds of compounds.

In addition to MACCkeys, several other molecular descriptors that may affect polarity are employed in this work, including molecular weight (MW), the number of hydrogen bond acceptors (HBA), the number of hydrogen bond donors (HBD), lipid-water partition coefficient (LogP), rotatable key (NROTB) and topological polar surface area (TPSA). These molecular descriptors contain some physical and chemical properties of the molecule, which may contribute to the polarity of the compound. MACCkeys and molecular descriptors utilized in this work can be easily accessed from the python package RDKit, which is a widely utilized tool combining computational chemistry and machine learning. Moreover, another property closely related to polarity, the dipole moment (DM), has also been derived through computational chemistry in this work.

In addition to compound-related information, eluent solvents-related information will also be imported into the machine learning model. Considering that there are multiple developing agent systems in the dataset, a five-length

vector [Hexane, EA, DCM, MeOH, Et_2O] that represents the proportion of each eluent solvent, is employed to describe the solvent composition. For example, [0.75, 0.25, 0, 0, 0] represents the eluent is Hexane/EA = 3/1. With this method, the influence of solvent composition on the R_f values can also be extrapolated through machine learning.

The ensemble method

In this work, an ensemble method is proposed to improve the stability and accuracy of the prediction performance. The formula of our proposed ensemble method is written as:

$$u_{\text{pred}} = w_{\text{RF}}u_{\text{RF}} + w_{\text{LGB}}u_{\text{LGB}} + w_{\text{XGB}}u_{\text{XGB}} + w_{\text{ANN}}u_{\text{ANN}}$$

$$w_{\text{RF}} + w_{\text{LGB}} + w_{\text{XGB}} + w_{\text{ANN}} = 1$$

where u_{pred} is the prediction of the ensemble method; u_{RF}, u_{LGB}, u_{XGB} and u_{ANN} are the prediction of RF, LGB, XGBoost and ANN, respectively. $w_{\text{RF}}, w_{\text{LGB}}, w_{\text{XGB}}, w_{\text{ANN}}$ are the respective weights.

The hyperparamters utilized in this work

There are many hyperparameters used in machine learning methods, which are often selected based on the experience of scientists and trial-and-errors to search for a better model. Therefore, the selection of hyperparameters is of great importance to the repeatability of research. Here, the utilized hyperparamters in this work that are selected by grid-search and repeated trails, are presented below.

For XGBoost, the number of estimators is 200, maximum depth is 3 and the learning rate is chosen to be 0.1.

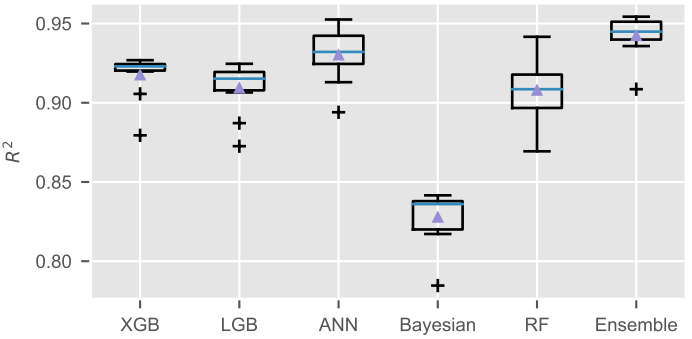
For LGB, the number of estimators is 1000, random state is 1 and the number of jobs is chosen to be 1.

For RF, the number of estimators is 1000,the random state is 1, the number of jobs is chosen to be 1 and the criterion is MSE loss.

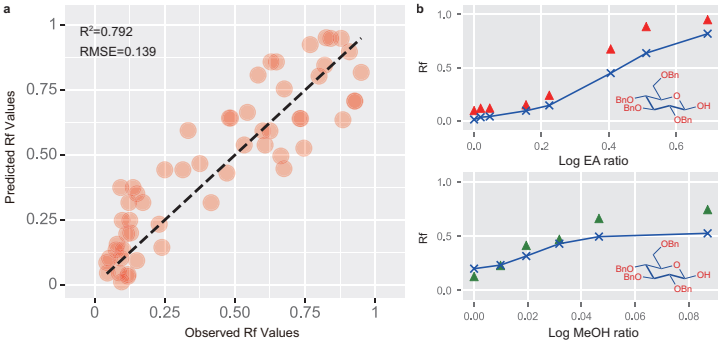
For ANN, the number of hidden layers is 4 with 128 neurons in each hidden layer, the number of input neurons is 179 and the number of output neurons is 1. The optimizer is chosen to be Adam with the learning rate 0.005. The maximum training epoch is 5000 and the early stop technique is adopted to prevent overfitting on the basis of the training loss and validating loss.

For the ensemble method, $w_{\text{RF}} = w_{\text{LGB}} = w_{\text{XGB}} = 0.2$ and $w_{\text{ANN}} = 0.4$.

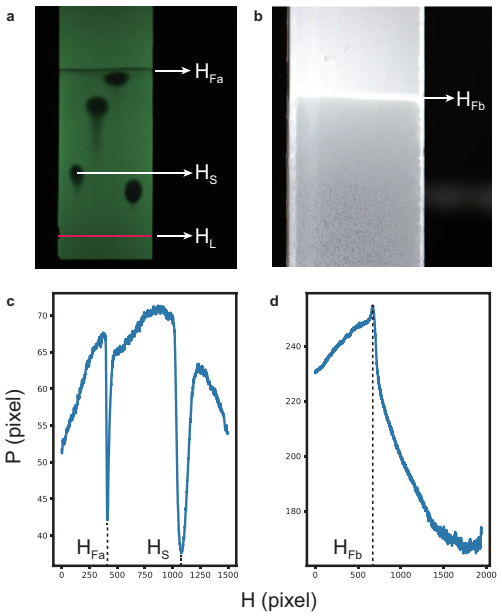
In addition, in order to avoid the influence of the selection of random seeds, different random seeds are used to conduct multiple experiments and take the statistical average.



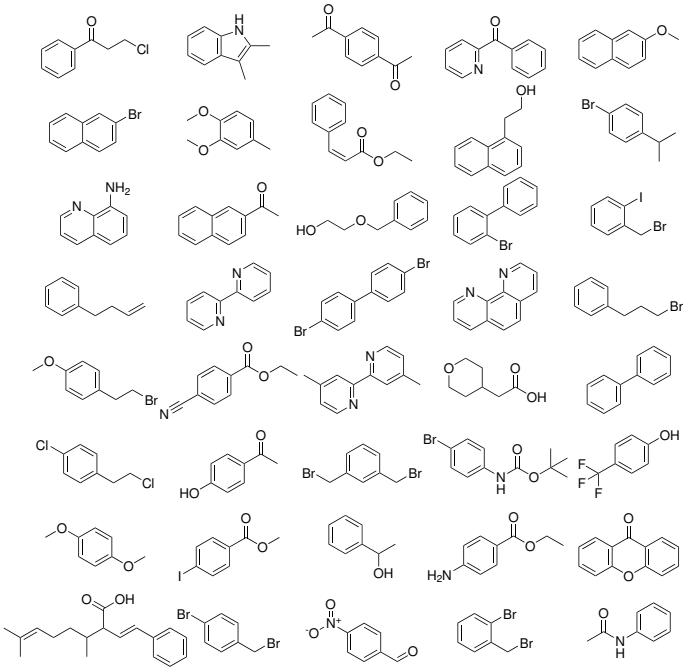
Extended Data Figure 1 The boxplot of the R^2 of different machine learning techniques in multiple trails when the dataset is split by TLC data. The blue triangles are the means of the R^2 for each machine learning technique and the black plus signs are the outliers.



Extended Data Figure 2 The performance of the predictive model for predicting the polarity of saccharides. a, Observed versus predicted R_f values for the predictive model to predict four different saccharides under hexane/ethyl acetate and dichloromethane/methanol systems. The dashed line is the $y = x$ line. **b**, Predicted R_f curves and experimental R_f curves for one of the saccharides under two elution solvent systems including hexane/ethyl acetate (upper) and dichloromethane/methanol (lower). The log ratio is calculated as log(r + 1) where r is the ratio of eluent.



Extended Data Figure 3 Calculation of the R_f values. a, and **b**, The photos photographed for the same TLC plate under ultraviolet light and visible light, respectively. **a** is photographed from the frontal angle and **b** is photographed from the back angle. **c,** The average pixel value after y-axis projection performed on the neighborhood of the first TLC spot. **d,** The average pixel value after y-axis projection performed on the TLC plate photoed under visible light. In this figure, H_{Fa} and H_{Fb} refers to the height of solvent front in both photos respectively, H_L refers to the initial height of the TLC spot, H_s refers to the height of TLC spot and x_s represents the horizontal position of the TLC spot. H is the position of pixel and P is the average pixel value after y-axis projection.



Extended Data Figure 4 Randomly selected compounds in the dataset.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInfomation.pdf