

## Explicit relation between thin film chromatography and column chromatography conditions from statistics and machine learning

Corresponding Author: Professor Fanyang Mo

**This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.**

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

The work described in this manuscript attempts to build a machine learning model linking thin layer chromatography parameters (specifically, the retardation factor) to retention volumes in preparative column chromatography. To me, the major contribution of this work is the development of an automated system for column chromatography, which enables the generation of data suitable for building the regression model.

I do have many concerns and questions about the machine learning/data analytics applied in the work:

The authors argument against mechanistic models is their reliance on idealized assumptions, which they claim can lead to discrepancies when compared to actual experimental conditions. However, there is ample evidence that (partial) differential equation-based models can describe experimental operation in chromatography very accurately. Machine learning models are invoked typically for computational reasons. I think the authors should present results comparing the accuracy of mechanistic models against the data-driven approach they advocate.

Continuing with the same point, these statements in the manuscript are arguable and potentially incorrect: "Notably, statistical and machine learning techniques are adopted here to generate "artificial intelligence (AI) experience" from data, rather than relying on human summarization for the acquisition of experience. This "AI experience" not only exhibits superior accuracy in experimental data and demonstrates robust applicability, but also obtains a closer alignment with the fundamental nature of chemistry. Furthermore, "AI experience" can be quickly replicated through parameter transfer (Fig. 1a). It elucidates the relationship between TLC and column chromatography explicitly, thereby offering a rationale for chemists to better comprehend the determination of separation conditions."

I refer to the fact that the alternative to AI experience from data is not necessarily human summarization for the acquisition of experience, and also point out that 'experience' is an undefined quantity. The notion that "AI experience" exhibits superior accuracy is a claim that needs evidence to back it up, which is why the authors need to show comparisons against mechanistic models. The notion that "AI experience" is in closer alignment with the fundamental nature of chemistry is an even more imprecise statement, given that this claim, too, is made with no evidence, and furthermore, chromatography systems have an interplay of chemistry with physics, convective flows, etc. The next claim, that "AI experience" can be quickly replicated through parameter transfer, begs the question about what the authors mean by parameter transfer. If it simply means refitting the parameters of an equation while retaining its structure, that is quickly done with any model.

Further, the paragraphs purporting to describe the difference between "AI experience" and black box machine learning models do not provide any specifics on how the authors' "AI experience" is different from any other approach.

I would suggest that the authors reframe the sentence that claims that retention volumes are identified and calculated automatically by a laptop – a more technical description is required.

The authors claim that 'we attempted to generate "AI experience" from the data acquired through 4g preparative columns and extrapolates it to other column specifications'. However, no information is provided on the extrapolation, nor are any

results provided.

Model predictions are labeled as 'meta-data', which can be misleading. It would be prudent to simply call them model predictions.

The eluent solvent is described by an 'averaged molecular description' of 6 dimensions. What does the term 'averaged' mean in this context?

Details of the surrogate model's training (such as the early stopping strategy) must be provided.

The predictive performance of surrogate models (Fig. 2b) is described as 'impressive', which I must question. One assumes that this is on the test dataset (it isn't stated explicitly) and that the impressiveness is because of the  $R^2$  values. Were any other tests of accuracy (or error analysis) performed? What should one make of the relatively low  $R^2$  value for Delta V?

In Fig. 2c, many data points are said to be derived 'from predictions on the meta-data'. Does that mean these are predictions based off other predictions, or do the authors mean 'predictions of the meta-data'?

The authors claim (Fig. 2d) that the variance of the retention volume distribution decreases as  $R_f$  increases and calls this a latent, significant correlation between  $R_f$  and retention volume. Correlation speaks to the covariance between variables, and not between a single value and the variance of a distribution.

Also, in Fig. 2e, it is unclear whether the deviation in the mean value with loading sample amount is actually significant.

In the discussion of Fig. 3, the trend between  $R_f$  and retention volume is said to be exemplified by a ratio of PE:EA of 100:1. However, the figure does not have a case with that PE:EA ratio.

Also, the chemical intuition that accompanies the relationship of the polarity of the eluent solvent and variance of retention volumes needs to be stated explicitly.

In supplementary section 2 (defining the separation index), the scenarios for complete separation are identified as  $V_{1A} > V_{2B}$  and  $V_{2A} > V_{2B}$ . The second scenario should be  $V_{1B} > V_{2A}$ . Also, equation (S.1) needs to be explained more clearly – how are these conditions derived?

A citation should be provided to pySR for the symbolic regression. Also, any hyperparameters used must be specified. The parameter  $r$  is defined as the ratio of PE in the eluent, but it seems to be the fraction of PE in the eluent. Finally, how good was the symbolic regression? No performance metrics are provided.

For the decision trees, hyperparameters (e.g., maximum depth, minimum leafs/nodes) should be reported. Also, it is more conventional to refer to roots and leafs/nodes rather than cap and whiskers.

The authors disregard transfer learning (because they involve fine-tuning, i.e., fitting, model parameters with target domain data), but then 'recalibrate' the coefficients of the equations learnt from symbolic regression with data from the target domain. In other words, the authors also fit model parameters with target domain data, and making an artificial distinction with respect to refitting is simply inaccurate. In particular, the statement "remarkably, our model obviates the need for retraining black-box models since it can adapt to diversified column specifications by simply adjusting the coefficients in eq. 1 and 2" is completely at variance with the facts. The authors model, derived through symbolic regression, is a fitted model that is retrained for new systems. It is also a black-box model, because the (fitted) parameters do not possess physical significance.

With respect to Fig. 5, the low RMSE and  $R^2$  values should be discussed in more detail. Is the extrapolation to different column specifications the only reason for the low accuracy?

The writing needs to be improved significantly. There are far too many qualitative objectives used, and I highlight just a few here: 'an exceptionally concise and accurate equation', 'seasoned experimenters', 'paramount significance', 'overly complex', 'pivotal application of knowledge discovery', 'machine learning methods excel unparalleled'. Many sections, especially where motivation is provided for choices made in the work, have somewhat vague and qualitative descriptions, and more precision in the wording would be appreciated. Grammatical errors also need to be corrected.

(Remarks on code availability)

I was able to install and run the code, and it does have a README file. I did not do a thorough investigation to reproduce the results with the code.

Reviewer #2

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)



### Reviewer #3

#### (Remarks to the Author)

English grammar can be improved through out the manuscript.

Rationale for application of current study should be explained.

Results of validation data set should be explained in manuscript which gives accuracy of test results.

More such paper should be cited.

Ex:

1. <https://doi.org/10.1080/10408347.2023.2254379>

The manuscript can be accepted after the revision.

#### (Remarks on code availability)

I am unable to open or run the code file so kindly provide the material which can be executable.

### Reviewer #4

#### (Remarks to the Author)

The method proposed in this manuscript uses an approach based on machine learning and statistical methods to transfer the  $R_f$  values of the solutes to  $k$  then  $V_s$  (number of volumes of mobile phase necessary to elute a solute, expressed in units of  $V_0$  of the column flash or the open column or even a preparative HPLC column) and generate final processing by artificial intelligence. This results in the use of empirical equations (1), (2) and (5) qualifying the adequacy of the calculated retention volumes or  $R_f$  to the experimental retention volumes or  $R_f$ . I have great difficulty understanding the significance of this result, apart from having the highest possible linear regression coefficient.

#### Details:

From a chromatographic point of view

1- For a TLC-column or TLC-HPLC transfer to be truly operational, it is absolutely necessary that the stationary phase must be strictly identical in the two modes: its humidity level must be identical, which is rarely the case, hence the weighting by a transfer coefficient. No mention is made of this for the choice of stationary phases in the construction of their two databases (TLC model and CC model): Authors should clarify this point.

2- The  $R_f$  measurement must be done at the center of the spot which must be as little dispersed as possible (hence the narrowest possible deposit). Nothing is specified on these two levels. How is  $R_f$  measured: by densitometry or manually?

3- How the 387 used solutes in TLC and the 182 in CC were chosen ? Is it from the data in table S1?

4- Without additional explanation I have difficulty to understand the dispersion of retention volumes for a given  $R_f$  value (Fig 3a and S4)

5- I don't understand Fig S7 b

6- In order to better understand, it would be better to replace the indices 1 and 2 (which characterize the start of peak volume and the end of peak volume and which suggest a retention order), by the indices b (beginning) and e (end): ( $V_bA$  instead of  $V_1A$ ) and ( $V_eA$  instead of  $V_2A$ ). This would make it easier to read.

7- Writing error : 2nd page of supplementary information, 6th line from the bottom of the page:

Replace  $V_2A > V_2A$  with  $V_1B > V_2A$  or as suggested above  $V_2A > V_2A$  with  $V_bB > V_eA$

#### (Remarks on code availability)

### Reviewer #5

#### (Remarks to the Author)

The manuscript describes use of machine learning techniques and statistical analysis for investigating how chemists use thin layer chromatography (TLC) to determine column chromatography conditions. For this purpose, the authors developed an automatic platform and used this platform for generating a big experimental dataset for the column chromatography (CC), which includes 192 compounds. This dataset was used for generating a reliable "artificial intelligence (AI) experience" through a knowledge discovery framework, where the relationship between the retardation factor (RF) values from TLC and retention volumes from column chromatography is represented through explicit equations. The evaluation of these equations confirmed their satisfactory accuracy and generalizability, providing a scientific basis for selecting experimental conditions and enhancing the understanding of chromatography. It is expected that this platform will in the future be expanded to much

higher number of compounds. Experimental datasets for TLC and CC and are available for the readers in "DATA – CODE" files. The manuscript is clearly structured. Some small improvements in Figures and "DATA – CODE" files at [https://github.com/woshixuhao/Discovery\\_of\\_column\\_chromatography/tree/main](https://github.com/woshixuhao/Discovery_of_column_chromatography/tree/main) (e.g. English instead of Chinese titles) can easily improve the manuscript. Other corrections needed are listed below. I recommend this manuscript for the publication in Nature Communications after revisions.

The following changes are needed in the in the whole manuscript and supplementary material (text, Figs. Tables) and "DATA-CODE files":

1. "thin layer" should be replaced with "thin-layer".
2. "molecules" should be replaced with "compounds".
3. "eluent solvent" should be replaced with "mobile phase".
4. All abbreviations should be introduced before being used in the text, in Figs., and "DATA-CODE files".
5. The abbreviation "CC" for column chromatography should be used in the whole manuscript (like TLC is used and not thin-layer chromatography).
6. The font size in Figures should be unified and the size of everything written in Figures should be big enough to be readable, which is not the case in the current version.
7. According to IUPAC "R<sub>f</sub>" should be replaced with "R<sub>F</sub>" in the whole manuscript (text, Figs., figure legends and equations and "DATA-CODE files").
8. "(V/V)" should be added after "PE:EA".
9. I hardly believe that the samples were analyzed in this study, but most probably, only mixtures of standard solutions were applied. Therefore, corrections are needed in the whole manuscript.
10. In most cases there is no need to use "in this study" (used 14 x).

#### LIST OF OTHER CORRECTIONS NEEDED

##### ABSTRACT

1. Correct is "thin-layer". Corrections are needed in the whole manuscript.
2. According to IUPAC it should be "R<sub>F</sub>" and not "R<sub>f</sub>".

##### INTRODUCTION

3. 1st paragraph, 1st sentence: Not only in synthetic chemistry.
4. 1st paragraph, 2nd sentence: There is no parallelism in "minute laboratory experiments" vs. "substantial industrial-scale processes".
5. 1st paragraph, 3rd sentence: Not only this, but also in isolation of natural bioactive compounds from natural products (e.g. plant materials, food waste materials), etc.
6. 3rd paragraph, last sentence: The mobile phase is not the only parameter that influence the R<sub>F</sub> values in TLC.
7. 6th paragraph, 4th sentence: Specify, because references [11-15] include HPLC and HPLC-MS.

##### RESULTS

8. Page 10, 2nd line: Replace "error" with "errors".
9. Page 12, 2nd paragraph, 1st sentence: Avoid repetitions (see Introduction).
10. Page 12, 2nd paragraph, 2nd sentence: Which previous analysis? In the case this is referred to the published work a reference should be added.

##### MATERIALS AND METHODS

This section should be written in such a way that the reader can repeat the experiments. However, based on the current version this is not possible.

11. Add data about the standards used.
12. Add the specifications of the stationary phase for TLC and CC.
13. Add the separation conditions for TLC (a developing chamber (type, saturated/unsaturated), developing distance; detection conditions; in case of derivatization also derivatization conditions, etc.) and for CC.
14. Pages 22 and 23: Move the following sentence from the section Materials and Methods: "While automated preparative chromatographs have long been available, they still possess several drawbacks. Firstly, they do not fully automate the entire column chromatography dataset process. For instance, the sample loading process remains manual. However, given the abundant experiments of column chromatography required in this study, manual sample loading is highly inefficient. Secondly, existing preparative chromatographs are highly integrated, which limits their flexibility and precludes their use for data collection as required in this study."
15. Page 23, line 7 (and several other places in the manuscript): I hardly believe that the samples used. Most probably, only mixtures of standard solutions were applied.
16. Page 23, 3rd paragraph, lines 3-4: "solution sample" is not appropriate term, replace it with either "standard solutions" or "mixtures of standard solutions".
17. Page 23, 3rd paragraph: Correct (samples & with UV detector) the sentence "The samples are then eluted and separated, with the UV detector recording the UV signal of the effluent solution during the elution process."
18. Page 23, 4th paragraph: Corrections are needed for "seven experiments are conducted using different eluent ratios": Specify the ratios and explain why 7 if 8 ratios of PE:EA (1:0, 100:1, 50:1, 10:1, 5:1, 2:1, 1:1, and 0:1) are given in the manuscript.
19. Page 24, line 11: Not "molecules", but "compounds". Explain the abbreviation MACCS.
20. Page 25, line 12: Replace "proportion of the developing solvent" with "proportion of the solvents in the developing solvent" or "proportion of the solvents in the mobile phase".

21. Page 25, line 14: Add "(V/V)" after "0.1".

#### FIGURE 1:

22. Replace "Rf" with "RF" in Fig. 1b and Fig. 1 title.

In Fig. 1c:

23. Replace "molecules" with "compounds", and use the term mobile phase;

24. It is not clear what is the source for all the numbers. If this data come from another source a reference should be added.

25. Replace "eluent systems" with "mobile phases".

26. Replace "eluent solvents" with "solvents" or "constituents of mobile phase".

27. A parallel is missing between the column (3 column spec) and plate (no data).

#### FIGURE 2:

28. Explain all abbreviations (e.g. Molecular ACCess System (MACCS) keys) and add the space between "MACCS" and "keys".

29. Replace "Rf" with "RF" (also in the legends), which is correct according to IUPAC.

30. In Fig. 2b add 1.0 to scales with RF.

31. The fonts in some figures (e.g. Fig. 2d; Fig. 2a (the subscripted "e" in "Ve")) are too small.

32. Remove the background in Figs. 2b.

33. Add missing lines for x- and y-axes.

34. Fig. 2e: "Ratio" of what? The "red and black triangles" in blue and red rectangles are not visible are not visible without zooming in on the screen.

35. Add the numbers to the scales everywhere where they are missing – where data presented have higher values than are the values on the scales (Fig. 2c, 2d).

#### FIGURE 3:

36. Replace "Rf" with "RF" (also in the legends and equations), which is correct according to IUPAC.

37. In Fig. 3c the RF scale should start with "0.0" and not "0".

38. The fonts in some figures, figure legends and equations are too small.

39. Fig. 3a: The "red and black triangles" in blue and red rectangles are not visible without zooming in on the screen.

40. Add the numbers to the scales (Fig. 3a, 3d) everywhere where they are missing -data presented have higher values than are the values on the scales.

#### FIGURE 4:

41. The fonts in Fig. 4 are too small to be readable.

42. In the figure and in the figure title add (V/V) after PE:EA (e.g. PE:EA (v/v)).

#### FIGURE 5:

43. Replace "Rf" with "RF".

44. The fonts in some figures and equations are too small.

45. Add the numbers to the scales (Fig. 5a) everywhere where they are missing -data presented have higher values than are the values on the scales.

#### FIGURES S1 and S6:

46. Only one of the figures Fig. S1 or Fig. S6 should be selected, while the other should be deleted.

#### FIGURE S3:

47. Not "molecules" but "compounds".

#### FIGURE S4:

48. See comments for Figs. In the manuscript and prepare corrections accordingly.

#### EQUASIONS 1, 2, 5, 6, 7, 8, 9 and 10:

49. Pages 15, 16, 19, 20, 27, 28: Replace "Rf" with "RF".

#### REFERENCES

50. For the latest papers add paper number and not only volume and year.

51. Ref [4]: The authors and the title of the chapter and relevant pages should be added. Add data about the editor "book title: Preparative Chromatography: of Fine Chemicals and Pharmaceutical Agents; editor: Henner Schmidt-Traub".

52. Ref [7-11, 13-16]: Add names of other authors.

DATA – CODE: ([https://github.com/woshixuhao/Discovery\\_of\\_column\\_chromatography/tree/main](https://github.com/woshixuhao/Discovery_of_column_chromatography/tree/main))

53. File TLC\_dataset.xlsx – corrections are needed, because columns (E-I) with data are not related to the manuscript

54. File dataset\_1013.xlsx - all Chinese column titles should be translated to English.

55. File dataset\_25g\_1218.xlsx - corrections are needed because the column P is without the title, the column O with the Chinese title, but without data.

56. File dataset\_4+4g\_1020.xlsx - corrections are needed because the column P is without a title and there are columns with the Chinese titles.

(Remarks on code availability)

README file is available, but some corrections are needed in the following files (DATA – CODE:

[https://github.com/woshixuhao/Discovery\\_of\\_column\\_chromatography/tree/main](https://github.com/woshixuhao/Discovery_of_column_chromatography/tree/main)):

1. File TLC\_dataset.xlsx – corrections are needed, because columns (E-I) with data are not related to the manuscript
2. File dataset\_1013.xlsx - all Chinese column titles should be translated to English.
3. File dataset\_25g\_1218.xlsx - corrections are needed because the column P is without the title, the column O with the Chinese title, but without data.
4. File dataset\_4+4g\_1020.xlsx - corrections are needed because the column P is without a title and there are columns with the Chinese titles.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

The authors have answered my queries reasonably well. I do not have any more major concerns.

(Remarks on code availability)

I reviewed the code with the original manuscript.

Reviewer #2

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Reviewer #5

(Remarks to the Author)

Although the authors have improved the clarity of the manuscript, the revised manuscript needs some minor revisions listed below.

I recommend this manuscript for the publication in Nature Communications after revisions.

#### LIST OF CORRECTIONS NEEDED:

1. According to IUPAC in “RF” “F” should be in subscript and this should be replaced in the whole manuscript (text, figures (Figs. 1b, 2, 3, 5, S5), figure titles (Figs. 1, 2, 3, S5) and equations (1, 2, 5-10).
2. Page 12, line 301: Replace “From the above analyses of Fig. 2” with “From the results of the analyses presented in Fig. 2”.
3. In the whole manuscript (text, and Figs., figure titles) move “(V/V)” after the numbers of the solvents ratio (e.g., PE:EA (1:0, 100:1, 50:1, 10:1, 5:1, 2:1, 1:1, and 0:1 (V/V)).
4. Figure 1.b: Delete “human”.
5. Figure 2.d: Add missing lines for x- and y-axes and units for frequency.
6. Reference [6] in the revised version should be corrected: M. Michel, A. Epping, A. Jupke are the authors of the chapter 6 “Modeling and Determination of Model Parameters” (<https://onlinelibrary.wiley.com/doi/book/10.1002/3527603484>; <https://onlinelibrary.wiley.com/doi/10.1002/3527603484.ch6>) and Henner Schmidt-Traub is the editor of the book “Preparative Chromatography: of Fine Chemicals and Pharmaceutical Agents”, in which chapter 6 was published. Check also the pages as those on the website differ from those in the manuscript.

(Remarks on code availability)

**Open Access** This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

## Response to the reviewers

We would like to sincerely express our deep gratitude to the editors and referees for their insightful comments and suggestions on the manuscript, which are fair, encouraging, and constructive. After carefully studying the comments, we have made corresponding changes in the revised manuscript. The specific responses to the comments are listed below. In the revised manuscript, text in **red** is the revision for Referee #1 & #2. Text in **blue** is the revision for Referee #3. Text in **violet** is the revision for Referee #4. Text in **orange** is the revision for Referee #5. Changes in the manuscript are also emphasized by underlining within the reply, and the corresponding content is included in a gray box.

---

### Reviewer 1

**Reviewer Point P 1.1** — The work described in this manuscript attempts to build a machine learning model linking thin layer chromatography parameters (specifically, the retardation factor) to retention volumes in preparative column chromatography. To me, the major contribution of this work is the development of an automated system for column chromatography, which enables the generation of data suitable for building the regression model.

**Reply:**

Thanks very much for your interest in our work! We sincerely appreciate your comprehensive reading and understanding of our paper. Your comments are insightful and constructive, which guide us further improve our work. We are grateful for your approval of the automated column chromatography (CC) for constructing the dataset, which constitutes an important part of our work. Apart from the automated platform, the discovery of relationship between TLC and CC through experimental data is also significant contribution of this work. Therefore, according to your comments, we have revised the manuscript to emphasize this aspect.

**Reviewer Point P 1.2** — I do have many concerns and questions about the machine learning/-data analytics applied in the work:

The authors argument against mechanistic models is their reliance on idealized assumptions, which they claim can lead to discrepancies when compared to actual experimental conditions. However, there is ample evidence that (partial) differential equation-based models can describe experimental operation in chromatography very accurately. Machine learning models are invoked typically for computational reasons. I think the authors should present results comparing the accuracy of mechanistic models against the data-driven approach they advocate.

**Reply:** Thanks for your valuable comments! We fully agree that the mechanistic models can describe the chromatographic process very accurately, which has great importance. In the revised manuscript, we have provided a figure to illustrate the mechanism under different scales in chromatography. In fact, we do not want to disparage mechanistic models, but rather to highlight the challenges they present when addressing cross-scale problems, like the relationship between different chromatography processes. Therefore, we thoroughly revised the Introduction section to better explain this (pages 3 and 4).

The revised statements and qualitative comparison in the Introduction section.

As illustrated in Fig. S1, the dynamic of chromatography involves multiple scales from microscopic to macroscopic. For each scale, physicochemical mechanistic models have been established through explicit mathematical equations, which accurately describe the chromatographic process. Nevertheless, the prediction of chromatographic results, such as RF value and retention time, remains a challenge for mechanistic models due to the elusive coupling relationship between mechanisms at different scales. Therefore, when determining separation conditions of chromatography, chemists rely more on expertise than mechanistic models. The empirical insights can reflect the coupling relationships between TLC and CC to some extent. However, the separation results of CC are influenced by a multitude of factors, including choice of mobile phases, column type, and various other experimental settings. Consequently, the chemical expertise is limited, which may lead to the failure of CC separation within sophisticated contexts.

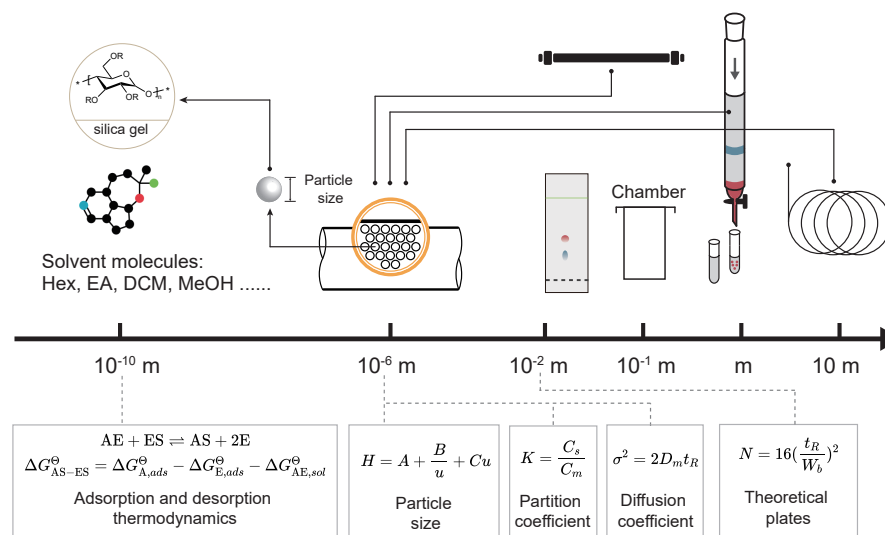


Fig. S1 The mechanism under different scales of the chromatography process

In the revised manuscript, we point out that although the mechanistic models are efficient for the understanding of chromatography, the prediction of chromatographic results remains a challenge for mechanistic models due to the elusive coupling relationship between mechanisms at different scales. That is why we construct the deep learning surrogate model to predict the chromatographic results. In fact, the relationship between TLC and CC has not been derived from mechanic models, since it involves multiple factors. Under this circumstance, we adopt a data-oriented strategy (i.e., knowledge discovery), which discovers equations directly from observation data. This strategy is especially suitable for the situation that theoretical derivation is difficult. Since there are a variety of empirical formulas or chemical expertise in the field of chemistry, where it is difficult to model the potential cross-scale relationships through theoretical deduction, our method is promising.

Finally, we would like to clarify that the proposed knowledge discovery method is not intended to replace the mechanism model, but to complement it. Their common goal is to obtain an explicit equation to describe the phenomenon. (The revised Discussion section, page 22) Honestly, our method is not omnipotent, and it also has some problems, such as the interpretability of the equation terms and

the scope of application. In the revised manuscript, we have revised the statement in the Introduction section and Discussion section to clarify this issue, in which the mechanistic model is analyzed fairly and compared with our method.

The revised statements in the Discussion section.

It is worth noting that the knowledge discovery technique is not intended to replace the mechanism model, but to complement it. Their common goal is to obtain an explicit equation to describe chemical phenomenon.

**Reviewer Point P 1.3** — Continuing with the same point, these statements in the manuscript are arguable and potentially incorrect: “Notably, statistical and machine learning techniques are adopted here to generate “artificial intelligence (AI) experience” from data, rather than relying on human summarization for the acquisition of experience. This “AI experience” not only exhibits superior accuracy in experimental data and demonstrates robust applicability, but also obtains a closer alignment with the fundamental nature of chemistry. Furthermore, “AI experience” can be quickly replicated through parameter transfer (Fig. 1a). It elucidates the relationship between TLC and column chromatography explicitly, thereby offering a rationale for chemists to better comprehend the determination of separation conditions.”

1. I refer to the fact that the alternative to AI experience from data is not necessarily human summarization for the acquisition of experience, and also point out that ‘experience’ is an undefined quantity.
2. The notion that “AI experience” exhibits superior accuracy is a claim that needs evidence to back it up, which is why the authors need to show comparisons against mechanistic models.
3. The notion that “AI experience” is in closer alignment with the fundamental nature of chemistry is an even more imprecise statement, given that this claim, too, is made with no evidence, and furthermore, chromatography systems have an interplay of chemistry with physics, convective flows, etc.
4. The next claim, that “AI experience” can be quickly replicated through parameter transfer, begs the question about what the authors mean by parameter transfer. If it simply means refitting the parameters of an equation while retaining its structure, that is quickly done with any model.

**Reply:** Thanks for point out this. We agree that these statements are inaccurate, so we thoroughly revised the statement about “AI experience” as the Reviewer suggested.

1. In the revised version, we have clarified that the data-driven discovery of “AI experience” is not intended to replace all chemical expertise. We have acknowledged the human summarization from experiments, since it can reflect the coupling relationships between TLC and column chromatography to some extent (page 3). Meanwhile, we point out that the human summarization has its inner limitation, since human is not good at tackling high-dimensional relationship. Considering that AI is apt to this kind of task, we utilize the machine learning and statistic techniques to obtain “AI experience”.



As reminded by the Reviewer, we have added a clear definition of “AI experience” in order to avoid misunderstanding. “AI experience” is defined as the discovered underlying relationship from experimental data by AI techniques, which is expressed in concise equation form (page 4). Here, “AI experience” is defined to differentiate it from the empirical formulas summarized by humans.

2. In the revised version, we adopt a more cautious description about the performance of “AI experience”, only claim that it is more aligned with experimental data. As replied to Review Point P1.2, we also added the qualitative comparison between mechanistic models and “AI experience”.
3. Regarding to the claim that “AI experience” is in closer alignment with the fundamental nature of chemistry, we admit that it is an imprecise statement as the Reviewer noted. Here, what we want to express is that the “AI experience” is a concise equation which can describe the relationship between TLC and CC. In the revised version, we have revised this sentence.
4. Regarding to the claim that “AI experience” can be quickly replicated through parameter transfer, as the Reviewer suggested, this statement is not clear enough, which may be misleading. What we want to express is that human experience is hard to be replicated since it involving learning and training, while “AI experience”, which is formalized knowledge expressed by concise equations, is easy to replicate, since anyone can utilize this equation to facilitate chromatography easily. Meanwhile, benefited from its interpretability, it is easier to be accepted and understood than conventional machine learning models. Therefore, we revised this sentence in the manuscript (page 4).

The revised statements in page 3 & 4.

Hence, in this work, we attempt to take a step further to discover the coupling relationship between TLC and CC from experimental data, which is expressed in concise equation form. This kind of equation is termed as “artificial intelligence (AI) experience” to differentiate it from the empirical formulas summarized by humans. It elucidates the relationship between TLC and CC explicitly, thereby offering a rationale for chemists to better comprehend the determination of separation conditions. Compared with human expertise, “AI experience” obtains a closer alignment with experimental data and can be quickly replicated without human learning process, which enhances its practicality.

**Reviewer Point P 1.4** — Further, the paragraphs purporting to describe the difference between “AI experience” and black box machine learning models do not provide any specifics on how the authors’ “AI experience” is different from any other approach.

**Reply:** Thank you for the comments. It is very important to demonstrate the difference between “AI experience” and other deep learning approaches as the Reviewer noted. Therefore, in the revised manuscript, we have rewritten the paragraph to emphasize the difference (pages 3 and 4).

The revised statements in pages 3 and 4.

Although various quantitative structure-retention relationship (QSRR) models have been established to predict outcomes for a range of chromatographic techniques, including TLC, CC, and high-performance liquid chromatography (HPLC), they often operate as “black boxes”, in which

the learned relationship cannot be expressed by explicit concise equations. This opacity limits their ability to provide interpretability, a crucial aspect for generating new chemical insights.

In brief, the dominant different between them is the interpretability. In the field of machine learning, especially deep learning, although it can make accurate prediction, we cannot know how it works, so the inner relationship is implicit. That is why they are often called “black-box” models. For example, in our work, we construct two surrogate models for TLC and CC, respectively. However, we do not know the underlying connection between them. Therefore, the knowledge discovery technique is adopted to unveil explicit equation, i.e., AI experience. Through the concise equations, we can explicitly observe the relationship between retention volume, RF values and eluent solvents. This is one of the highlights of our work.

**Reviewer Point P 1.5** — I would suggest that the authors reframe the sentence that claims that retention volumes are identified and calculated automatically by a laptop – a more technical description is required.

**Reply:** Thank you for your suggestion! We have provided a more technical description in the revised version (page 7). In addition, we also uploaded the identification algorithm in the codes of open source (“Auto\_run.py” in the Github).

The revised statements in page 7.

After the completion of an experiment, the absorbance is read and processed by the computer, the peak is identified through the change of absorbance, and the start and end times are obtained from the peak. The identification algorithm is provided in the codes of open source. The retention volumes are calculated based on the retention time and flow rate. All the process are controlled and accomplished by a laptop.

**Reviewer Point P 1.6** — The authors claim that ‘we attempted to generate “AI experience” from the data acquired through 4g preparative columns and extrapolates it to other column specifications’. However, no information is provided on the extrapolation, nor are any results provided.

**Reply:** Thanks for pointing out this issue. The extrapolation and its results are actually explained in the subsequent section **Model generalization across different compounds and column specifications**. The paragraph here is intended to explain why we collect data from 4g column, tandem 4g column and 25g column, and the extrapolation of “AI experience” is conducted in the section below. We have revised this sentence to be understood easier (page 7).

The revised statements in page 7.

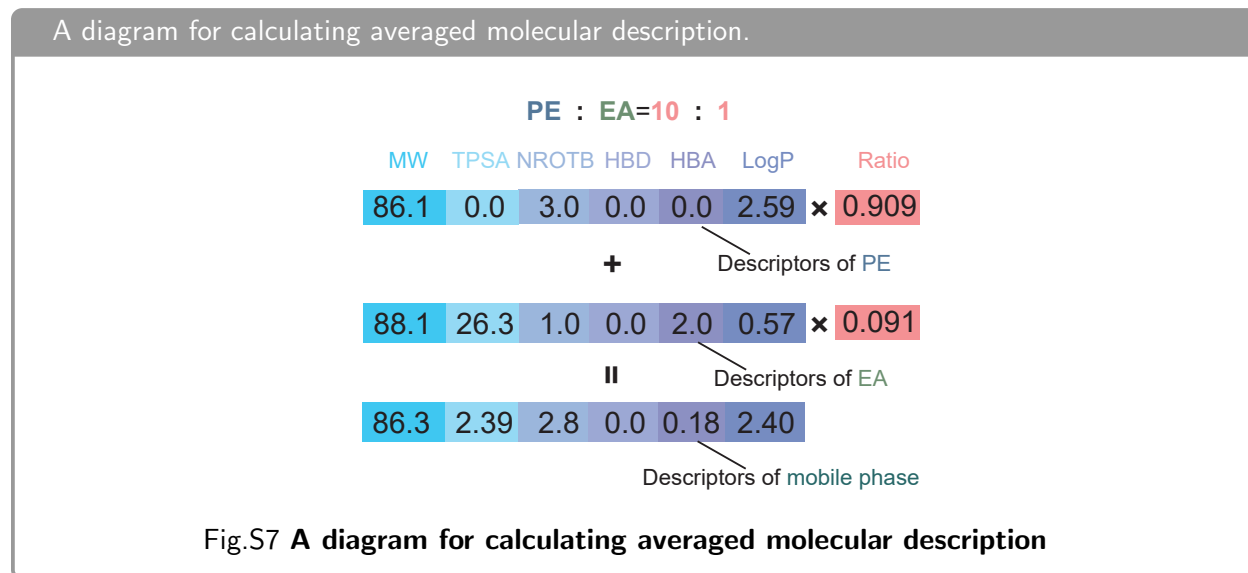
To mitigate this issue, we adopted a strategy that generate “AI experience” from the data acquired through 4g preparative columns and then extrapolates it to other column specifications, including tandem 4g column and 25g column.

**Reviewer Point P 1.7** — Model predictions are labeled as ‘meta-data’, which can be misleading. It would be prudent to simply call them model predictions.

**Reply:** Thanks for your invaluable comments! “meta-data” is a proper noun in the field of knowledge discovery, which means the model predictions on specified grids. We have realized that it may bring misleading, so we have replaced them to “model predictions” as suggested by the Reviewer.

**Reviewer Point P 1.8** — The eluent solvent is described by an ‘averaged molecular description’ of 6 dimensions. What does the term ‘averaged’ mean in this context?

**Reply:** “averaged molecular description” here means that the descriptors of PE and EA are weighted average based on the proportion. Here, we provide a diagram to facilitate the understanding (Fig. S7). We have added the explanation in the Materials and Methods Section (page 25).



**Reviewer Point P 1.9** — Details of the surrogate model’s training (such as the early stopping strategy) must be provided.

**Reply:** Thanks for your suggestion. In the revised manuscript, we have provided details of surrogate model’s training. The training epoch is 10000, and learning rate is  $10^{-3}$ . To avert overfitting, early stopping strategies were employed, where the epoch with the minimum validating error is considered as the best epoch (pages 9 and 10).

**Reviewer Point P 1.10** — The predictive performance of surrogate models (Fig. 2b) is described as ‘impressive’, which I must question. One assumes that this is on the test dataset (it isn’t stated explicitly) and that the impressiveness is because of the  $R^2$  values. Were any other tests of accuracy (or error analysis) performed? What should one make of the relatively low  $R^2$  value for Delta V?

**Reply:** Thanks for your questions. In the revised version, we adopt a more cautious description, “satisfactory”. Meanwhile, we have added the explanation that the figure shows the performance on the testing data (page 10). We also added the reason for the relatively low  $R^2$  for  $\Delta V$ , since it is calculated from the predicted  $V_S$  and  $V_E$ , where the error will accumulate (page 10).

*Note:* the  $V_1$  and  $V_2$  are replaced by the  $V_S$  and  $V_E$  to be clearer as suggested by another Reviewer.

The revised statements in page 10.

The  $R^2$  of  $\Delta V$  is relatively low, since it is obtained from the predicted  $V_S$  and  $V_E$ , which means the error will accumulate.

**Reviewer Point P 1.11** — In Fig. 2c, many data points are said to be derived ‘from predictions on the meta-data’. Does that mean these are predictions based off other predictions, or do the authors mean ‘predictions of the meta-data’?

**Reply:** Thanks for the question. As replied in Reviewer Point P1.6, the meta-data refers to model predictions on grids. Here, we want to express that the datapoints are predicted from the surrogate models instead of experimental data. Therefore, we have added some explanations to avoid misleading (page 10).

The revised statements in page 10.

This facilitated the generation of mobile phase-related curves, where datapoints are derived from model predictions rather than direct experimental data.

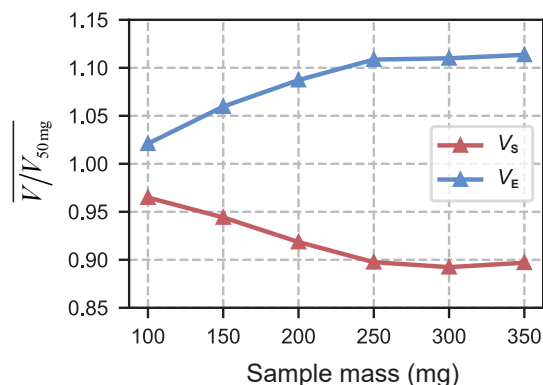
**Reviewer Point P 1.12** — The authors claim (Fig. 2d) that the variance of the retention volume distribution decreases as RF increases and calls this a latent, significant correlation between RF and retention volume. Correlation speaks to the covariance between variables, and not between a single value and the variance of a distribution.

**Reply:** Thanks for pointing out this to us. We admit that the “correlation” here is not rigorous, so we use the “relationship” instead in the revised version (page 10). In fact, Fig. 2d shows that the mean value of retention volume and RF values have an inner relationship, which forms the foundation of subsequent discovery of equations that describe the relationship between TLC and CC.

**Reviewer Point P 1.13** — Also, in Fig. 2e, it is unclear whether the deviation in the mean value with loading sample amount is actually significant.

**Reply:** In order to make the trend in Fig. 2e clearer, we have replaced it with the line chart of mean value v.s. loading sample, which obviously shows the trend.

The revised Fig.2e.



Revised Fig.2e. The influence of loading sample mass. The y-axis is the mean value of the ratio of retention volume under corresponding sample mass to the retention volume under 50 mg with other experimental conditions kept the same.

**Reviewer Point P 1.14** — In the discussion of Fig. 3, the trend between RF and retention volume is said to be exemplified by a ratio of PE:EA of 100:1. However, the figure does not have a case with that PE:EA ratio.

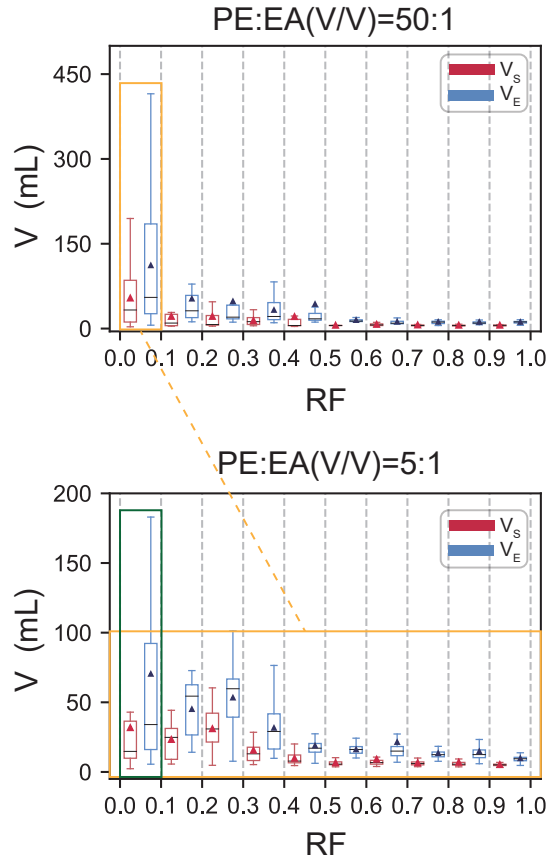
**Reply:** Thanks for reminding us about this issue. We have revised it, the example is PE:EA=50:1 (page 12), and the plots for other ratios are provided in Fig. S5.

**Reviewer Point P 1.15** — Also, the chemical intuition that accompanies the relationship of the polarity of the eluent solvent and variance of retention volumes needs to be stated explicitly.

**Reply:** Thanks for the comments. In the revised version, we have added some explanations about the chemical intuition (pages 12 and 13). In order to make the Reviewer easier to understand the chemical intuition, we also provide a sketch map below. For example, it is clear that, the compounds with RF values between 0 and 0.1 under PE:EA=50:1 (the yellow box) will have a wider RF ranges under the mobile phase with larger polarity (PE:EA=5:1). Therefore, the retention volumes fluctuation in the RF values between 0 and 0.1 under PE:EA=5:1 (green box) is smaller. By comparing within the same image (green box and yellow box in the fig PE:EA=5:1), it can be observed that for identical RF value range, larger polarity of mobile phase will bring lower uncertainty.

The statements in pages 12 and 13 and diagram for the chemical intuition.

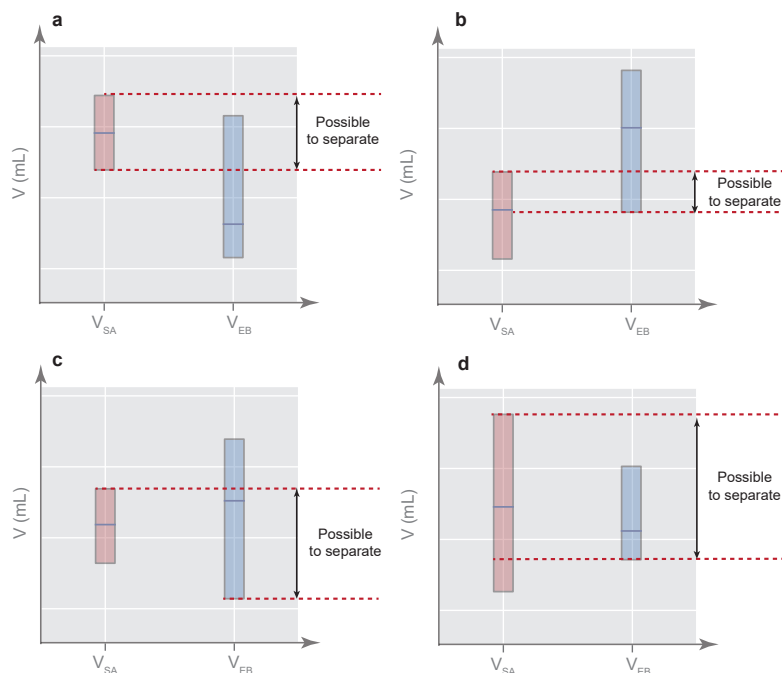
a trend is discernible where higher RF values are associated with a constricted range of retention volume fluctuations and a corresponding reduction in their mean values. This statistical observation is in alignment with chemical intuition, since a higher RF value corresponds to a sample with smaller polarity, where the retention volume is usually small. Meanwhile, for identical RF value range, the retention volumes of mobile phase with larger polarity are discovered to have a smaller variance. It implies that, for identical RF value range, larger polarity of mobile phase will bring lower uncertainty.



**Reviewer Point P 1.16** — In supplementary section 2 (defining the separation index), the scenarios for complete separation are identified as  $V_{1A} > V_{2B}$  and  $V_{2A} > V_{2A}$ . The second scenario should be  $V_{1B} > V_{2A}$ . Also, equation (S.1) needs to be explained more clearly – how are these conditions derived?

**Reply:** Thanks for reminding us of this issue. We have revised it (as another Reviewer suggest, we replace 1 to S (start) and 2 to E (end), so it becomes  $V_{SB} > V_{EA}$ ). In order to facilitate the understanding of Eq. (S.1), we have added some explanations about the derivation for the four conditions, and revised Fig. S8 for better understanding. Here, there are four conditions when the ranges of  $V_{SA}$  and  $V_{EB}$  are overlapped, 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 corresponds to the four conditions in Eq. (S.1).

The revised Fig.S8 and the explanation for the conditions for calculating the separation probability in page 34.



Revised 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.

If compound A has not completely eluted yet and the compound B has already started to elute, that is,  $V_{EB} > V_{SA}$ , they are obviously not able to be separated completely. Conversely, if  $V_{EB} < V_{SA}$ , where compound B has not started elute when the compound A has eluted completely, they 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 possible scenarios, there are four conditions when the ranges of  $V_{SA}$  and  $V_{EB}$  are overlapped, 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.

**Reviewer Point P 1.17** — A citation should be provided to pySR for the symbolic regression. Also, any hyperparameters used must be specified. The parameter  $r$  is defined as the ratio of PE in the eluent, but it seems to be the fraction of PE in the eluent. Finally, how good was the symbolic regression? No performance metrics are provided.

**Reply:** Thanks very much for your comments. We have added a citation for pySR and added the choice of hyperparameters (page 15). Meanwhile we added more details about the pySR (page 15).

Also, we have revised the “ratio” to be “fraction” as reminded by the Reviewer (page 15). In the revised manuscript, we have added the performance of symbolic regression. The  $R^2$  of the symbolic regression for  $\overline{V}_S$  and  $\overline{V}_E$  are 0.882 and 0.900, respectively (page 16).

*Citation: Cranmer, M. Interpretable machine learning for science with PySR and SymbolicRegression. jl. arXiv preprint arXiv:2305.01582 (2023).*

The revised statement in page 15 for the hyperparameters for pySR.

The iteration is 100, the utilized operators are  $+$ ,  $\times$ , and  $/$ , and the criterion is mean squared error. In order to discover a concise equation, only the outcomes with complexity smaller than 10 are considered. Both the structure complexity and regression loss are considered to choose the best equation.

**Reviewer Point P 1.18** — For the decision trees, hyperparameters (e.g., maximum depth, minimum leafs/nodes) should be reported. Also, it is more conventional to refer to roots and leafs/nodes rather than cap and whiskers.

**Reply:** Thanks for your suggestions. We have reported the hyperparameters for decision trees, including maximum depth, minimum leaves, and criterion in the revised manuscript (page 17). As the Reviewer suggested, we have revised the statement of “cap” and “whiskers” to “roots” and “leaves”.

The revised statement in page 17 for the hyperparameters for decision trees.

The maximum depth of the regression tree is 9, the minimum leaf is 1, and the criterion is mean absolute error.

**Reviewer Point P 1.19** — The authors disregard transfer learning (because they involve fine-tuning, i.e., fitting, model parameters with target domain data), but then ‘recalibrate’ the coefficients of the equations learnt from symbolic regression with data from the target domain. In other words, the authors also fit model parameters with target domain data, and making an artificial distinction with respect to refitting is simply inaccurate. In particular, the statement “remarkably, our model obviates the need for retraining black-box models since it can adapt to diversified column specifications by simply adjusting the coefficients in eq. 1 and 2” is completely at variance with the facts. The authors model, derived through symbolic regression, is a fitted model that is retrained for new systems. It is also a black-box model, because the (fitted) parameters do not possess physical significance.

**Reply:** Thanks for your insightful advice! We fully agree that making an artificial distinction with respect to refitting is simply inaccurate. The recalibration of coefficients is essentially a strategy of transfer learning. Therefore, we have revised relevant paragraph to provide a more precise statement (page 19). In fact, here, we want to express that calibration of coefficients in this work has a higher computational efficiency compared with fine-tuning machine learning models, since it only needs a simple regression.



The revised statement in page 19.

The equations can be easily extrapolated to other column specifications with a small amount of data through a transfer learning strategy, which recalibrate the coefficients of the equations learnt from symbolic regression with data from the target domain. This strategy has a higher computational efficiency compared with fine-tuning machine learning models, since it only needs a simple regression.

Meanwhile, we deleted the paragraph mentioned by the Reviewer that is at variance with the facts. However, we point out that the discovered equation is not a black-box model. As explained in Reviewer Point P1.4, in black-box model, the inner relationship cannot be expressed explicitly, and in our work, we manage to extract an explicit equation to describe the relationship between the TLC and CC. Although the parameters are fitted and may not have clear physical meaning, the equation itself can indicate the potential relationship explicitly. We agree that the original statement is not appropriate, so we delete it.

**Reviewer Point P 1.20** — With respect to Fig. 5, the low RMSE and  $R^2$  values should be discussed in more detail. Is the extrapolation to different column specifications the only reason for the low accuracy?

**Reply:** Thanks for the advice. In the revised manuscript, we have added the discussion about the relatively low RMSE and  $R^2$  for the transfer learning (page 20). The error originates from two aspects. On the one hand, the amount of data in tandem 4g and 25g column dataset is much smaller than that in the 4g column, which will affect the model precision. On the other hand, it seems that the difference between the large column and the small column may affect the transfer learning.

The revised statement in page 20.

It is found that the equations can predict the retention volumes well, but the performance is relatively worse than that of 4g column. It may be attributed that the amount of data in larger column is smaller, and the difference between the large column and the small column is notable, affecting the transfer learning.

**Reviewer Point P 1.21** — The writing needs to be improved significantly. There are far too many qualitative objectives used, and I highlight just a few here: ‘an exceptionally concise and accurate equation’, ‘seasoned experimenters’, ‘paramount significance’, ‘overly complex’, ‘pivotal application of knowledge discovery’, ‘machine learning methods excel unparalleled’. Many sections, especially where motivation is provided for choices made in the work, have somewhat vague and qualitative descriptions, and more precision in the wording would be appreciated. Grammatical errors also need to be corrected.

**Reply:** We appreciate a lot for your guidance for the writing of our manuscript, which enables us further polish the article. In the revised manuscript, we have made proof-reading to diminish grammatical errors. Notably, we check all the statements and avoid exaggerated descriptions and too many qualitative objectives. Meanwhile, we have thoroughly revised the Introduction, especially that provides

the motivations, in order to provide a more rigorous and precise descriptions. We sincerely hope that the revised manuscript can bring a better reading experience.

**Reviewer Point P 1.22** — I was able to install and run the code, and it does have a README file. I did not do a thorough investigation to reproduce the results with the code.

**Reply:** Thanks for your effort on testing the code. In the revision, in order to better demonstrate our code and its running results, we also provide a version that can be run online and provides outputs in Bohrium website. (<https://bohrium.dp.tech/notebooks/86427319178>)

---

## Reviewer 2

**Reviewer Point P 2.1** — I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

**Reply:** We sincerely thank you and Reviewer #1 for your effort on reviewing our manuscript. Your suggestions have inspired us a lot and helped us further improve the quality of the article, especially in the writing. In particular, after reading your comments, we made in-depth revisions to make the article more logical and the narrative more objective and accurate. Thank you again!

---

## Reviewer 3

**Reviewer Point P 3.1** — English grammar can be improved throughout the manuscript.

**Reply:** Thanks for your advice, we have made proof-reading to diminish grammatical errors. We hope that after the revision, the manuscript can be more readable.

**Reviewer Point P 3.2** — Rationale for application of current study should be explained.

**Reply:** Thanks for your valuable comments. We have thoroughly revised the Introduction section (pages 3~4), and added the application potential of our work in the Introduction and Discussion (page 22). The application of current study can be summarized in three aspects.

1. In our work, an automated platform for column chromatography is designed and a dataset of retention volumes is established. This inspires the high-throughput experiments in column chromatography.
2. We establish the prediction models for TLC and column chromatography, which shows satisfactory accuracy. These models can facilitate the prediction of R<sub>f</sub> values and retention times, and thus guiding the chromatography process without multiple experiments.

3. We take a step further and discover explicit concise equations to describe the underlying relationship between TLC and column chromatography, which are called “AI experience”. They can promote the chemical understanding and facilitate the chromatography separation.
4. Furthermore, considering that there are a variety of empirical formulas or chemical expertise in the field of chemistry, where it is difficult to model through theoretical deduction. Therefore, our method is promising to discover more “AI experience” in diversified problems.

The added statements in Discussion about the application of the study.

We managed to extracting explicit equations from experimental data to construct an interpretable model. While there are concessions in precision, it can enhance our comprehension of the relationship between TLC and CC, making it more applicable in experimental contexts. Specifically, the constructed TLC and CC surrogate model can provide accurate prediction of RF values and retention times, which can guide the experiments. Meanwhile, the discovered “AI experience” provides an explicit equation to bridge TLC and CC, which can facilitate the chromatographic separation. Furthermore, considering that there are a variety of empirical formulas or chemical expertise in the field of chemistry, where it is difficult to model through theoretical deduction. Therefore, our method is promising to discover more “AI experience” in diversified problems.

**Reviewer Point P 3.3** — Results of validation data set should be explained in manuscript which gives accuracy of test results.

**Reply:** Thanks for your suggestions. In the revised manuscript, we have explained the results of validation data (page 10). The validated  $R^2$  for the TLC and CC model is 0.948 (RF value), 0.838 ( $V_S$ ), and 0.898 ( $V_E$ ).

*Note:* the  $V_1$  and  $V_2$  are replaced by the  $V_S$  and  $V_E$  to be clearer as suggested by another Reviewer.

**Reviewer Point P 3.4** — 4. More such paper should be cited. Ex:

1. <https://doi.org/10.1080/10408347.2023.2254379>

**Reply:** Thanks very much to provide more relevant papers! In the revised manuscript, we have cited this article, and review some other relevant papers (pages 4 and 30).

*Citation:*

Singh, Y. R. et al. *Current trends in chromatographic prediction using artificial intelligence and machine learning. Analytical Methods* 15, 2785–2797 (2023).

Singh, Y. R., Shah, D. B., Maheshwari, D. G., Shah, J. S. & Shah, S. *Advances in AI-Driven retention prediction for different chromatographic techniques: unraveling the complexity. Crit Rev Anal Chem* 1–11 (2023).

**Reviewer Point P 3.5** — I am unable to open or run the code file so kindly provide the material which can be executable

**Reply:** Thanks for reminding us of this issue, in order to better demonstrate our code and its running results, we also provide a version that can be run online and provides outputs in Bohrium website. (<https://bohrium.dp.tech/notebooks/86427319178>)

## Reviewer 4

**Reviewer Point P 4.1** — The method proposed in this manuscript uses an approach based on machine learning and statistical methods to transfer the  $R_f$  values of the solutes to  $k$  then  $V_s$  (number of volumes of mobile phase necessary to elute a solute, expressed in units of  $V_0$  of the column flash or the open column or even a preparative HPLC column) and generate final processing by artificial intelligence. This results in the use of empirical equations (1), (2) and (5) qualifying the adequacy of the calculated retention volumes or  $R_f$  to the experimental retention volumes or  $R_f$ . I have great difficulty understanding the significance of this result, apart from having the highest possible linear regression coefficient.

**Reply:** We are very grateful for the efforts of the Reviewer for reviewing our manuscript! In order to make it easier to understand the contributions of our work, we have thoroughly revised the Introduction and Discussion section (pages 3~4, and page 22). The Reviewer's understanding is correct, but it is only part of our work's contribution. Here, we want to explain it in detail to the Reviewer. We hope that it can facilitate the understanding of our work. The main contribution of this work can be summarized as four aspects:

1. In our work, an automated platform for column chromatography is designed and a dataset of retention volumes is established. This inspires the high-throughput experiments in column chromatography;
2. We establish the prediction models for TLC and column chromatography, which shows satisfactory accuracy. These models can facilitate the prediction of  $R_f$  values and retention times, and thus guiding the chromatography process without multiple experiments.
3. We take a step further and discover explicit concise equations to describe the underlying relationship between TLC and column chromatography, which are called "AI experience". They can promote the chemical understanding and facilitate the chromatography separation.
4. Furthermore, considering that there are a variety of empirical formulas or chemical expertise in the field of chemistry, where it is difficult to model through theoretical deduction. Therefore, our method is promising to discover more "AI experience" in diversified problems.

**Reviewer Point P 4.2** — From a chromatographic point of view - For a TLC-column or TLC-HPLC transfer to be truly operational, it is absolutely necessary that the stationary phase must be strictly identical in the two modes: its humidity level must be identical, which is rarely the case, hence the weighting by a transfer coefficient. No mention is made of this for the choice of stationary phases in the construction of their two databases (TLC model and CC model): Authors should clarify this point.

**Reply:** Thanks very much for pointing out this issue. We fully agree that the control of experimental settings, like stationary phase and humidity, will affect the discovery of TLC and column chromatography relationships, as the Reviewer suggested. We have noticed this issue, and controlled these influential variables during the design of experiments. In the revised manuscript, we have added some clarifications.

In the revised manuscript, we have added the details about the choice of stationary phases in TLC and column chromatography, which is controlled to be identical (page 24).

The added statements about the choice of stationary phases in TLC and column chromatography.

For TLC, the silica gel plate from Yinlong brand is utilized, and the stationary phase is standard silicone that average particle size is 40  $\mu\text{m}$ . For CC, the column from Agela brand is utilized, and the stationary phase is also standard silicone with average particle size being 40  $\mu\text{m}$ .

Meanwhile, we have added some discussion about this issue in the revised manuscript (pages 4 and 21). The Reviewer's concern is actually the motivation to construct automated platform to collect experimental data in this work. Through automated experiments, the environmental and operational influence can be better controlled, which makes our dataset more consistent and precise. Meanwhile, all the experiments are conducted in the same laboratory with stable humidity and temperature.

The added statements in page 4 and page 21.

*In page 4:*

Notably, in order to investigate the relationship between TLC and CC, the stationary phases in the two modes are maintained to be identical, and the environmental and operational influence are controlled through automated experiments, which is detailed in Materials and Methods.

*In page 21:*

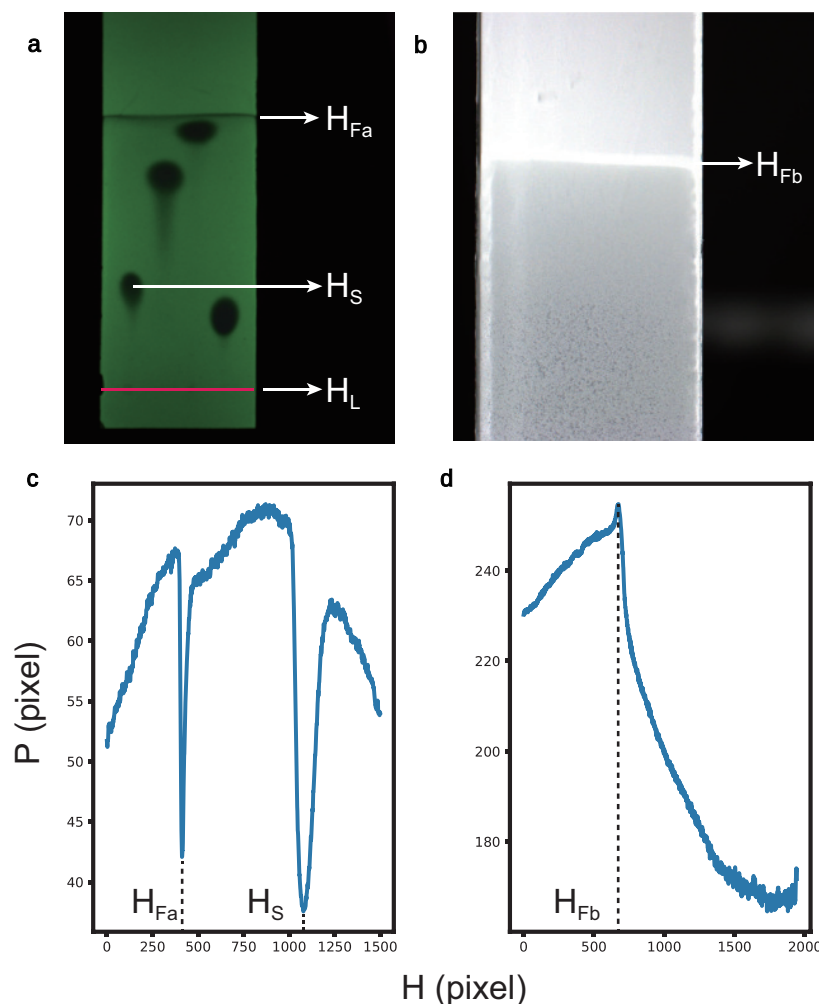
In order to guarantee the effectiveness and generalizability of the discovered relationship between TLC and CC. Automated platforms are established to conduct standardized experiments, where the stationary phase, humidity, temperature, and other environmental and operational influences are controlled.

**Reviewer Point P 4.3** — The Rf measurement must be done at the center of the spot which must be as little dispersed as possible (hence the narrowest possible deposit). Nothing is specified on these two levels. How is Rf measured: by densitometry or manually?

**Reply:** Thanks for your valuable suggestions! We fully agree a precise RF measurement needs cautious operation. Therefore, in the automated TLC platform, the two levels mentioned by the Reviewer are solved.

In the revised version, we have supplemented more details about the Rf measurement (page 24). In the TLC automated platform, the whole process is accomplished by a collaborative robot, in order to guarantee the precision of measured RF values. Specifically, the robot can finely control the spotting procedure, which makes the spot as little dispersive as possible. Meanwhile, the developed TLC plate is photographed, and the computer vision algorithm is adopted to identify the geometry center and calculate RF values. In order to guarantee the quality of data, human verification is adopted to calibrate TLC plates that are incorrectly identified by the computer.

In order to provide the Reviewer a better understanding of the measurement of RF values, we also attach a diagram below.



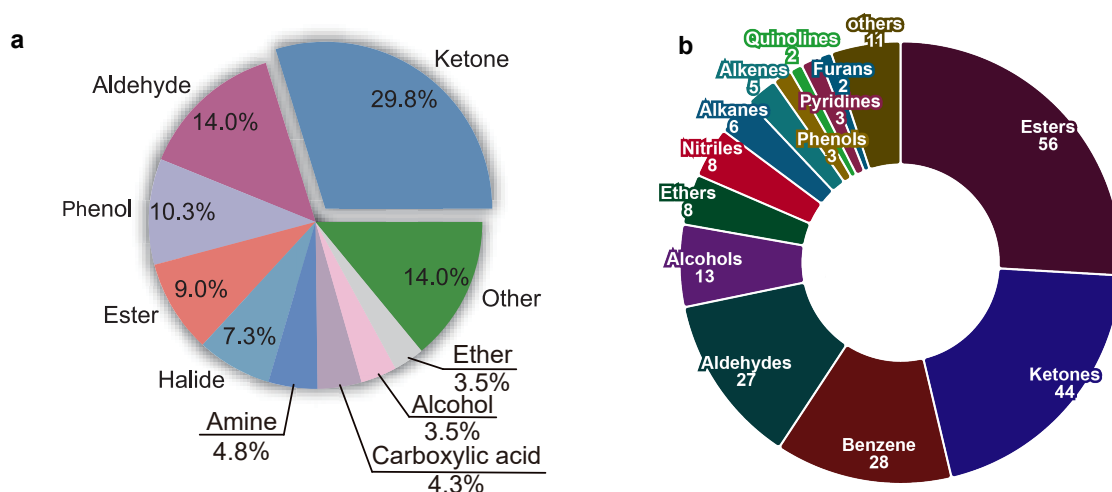
**Calculation of the  $R_f$  values.** (a) and (b) The photos photographed for the same TLC plate under ultraviolet light and visible light, respectively. (a) is photographed from the frontal angle and (b) is photographed from the back angle. (c) The average pixel value after y-axis projection performed 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.

**Reviewer Point P 4.4** — How the 387 used solutes in TLC and the 182 in CC were chosen ? Is it from the data in table S1?

**Reply:** Thanks for your valuable question! In this work, the solutes are chosen according to the diversity and accessibility, which include various types like Ketone, Aldehyde, and Phenol. We have added some explanations in the revised manuscript (page 4). The solutes are not from data in Table S1, and Table

S1 is the utilized molecular descriptors for characterization in the AI model construction. Here, we have provided two fan charts here for better illustration of the types of solutes in the CC and TLC dataset.

Fan charts for the types of solutes in the CC and TLC dataset.



**(a)**, The distribution of solute types for the TLC dataset. **(b)**, The distribution of solute types for the CC dataset.

**Reviewer Point P 4.5** — Without additional explanation I have difficulty to understand the dispersion of retention volumes for a given  $R_f$  value (Fig 3a and S4).

**Reply:** Thanks for reminding of us about this issue. In the revised version, we have added some explanations (page 12).

Fig. 3a and S4 essentially indicates the distribution of retention times of solutes whose  $R_f$  values are in specific ranges under given solvent eluents. We need to emphasize that this is the statistical result of many different solutes, not the result of a single solute. Obviously, even for identical  $R_f$  value, the retention volumes of different solutes in column chromatography are not exactly the same, and will present a distribution. Therefore, in Fig.3a, we counted the  $R_f$  values of all solutes in different eluent solvents, classified them according to their ranges, and displayed the distribution of their retention volumes. From this distribution, we can observe some principles, which constitutes the basis for our subsequent discovery of the explicit equation that describe the relationship between TLC and CC.

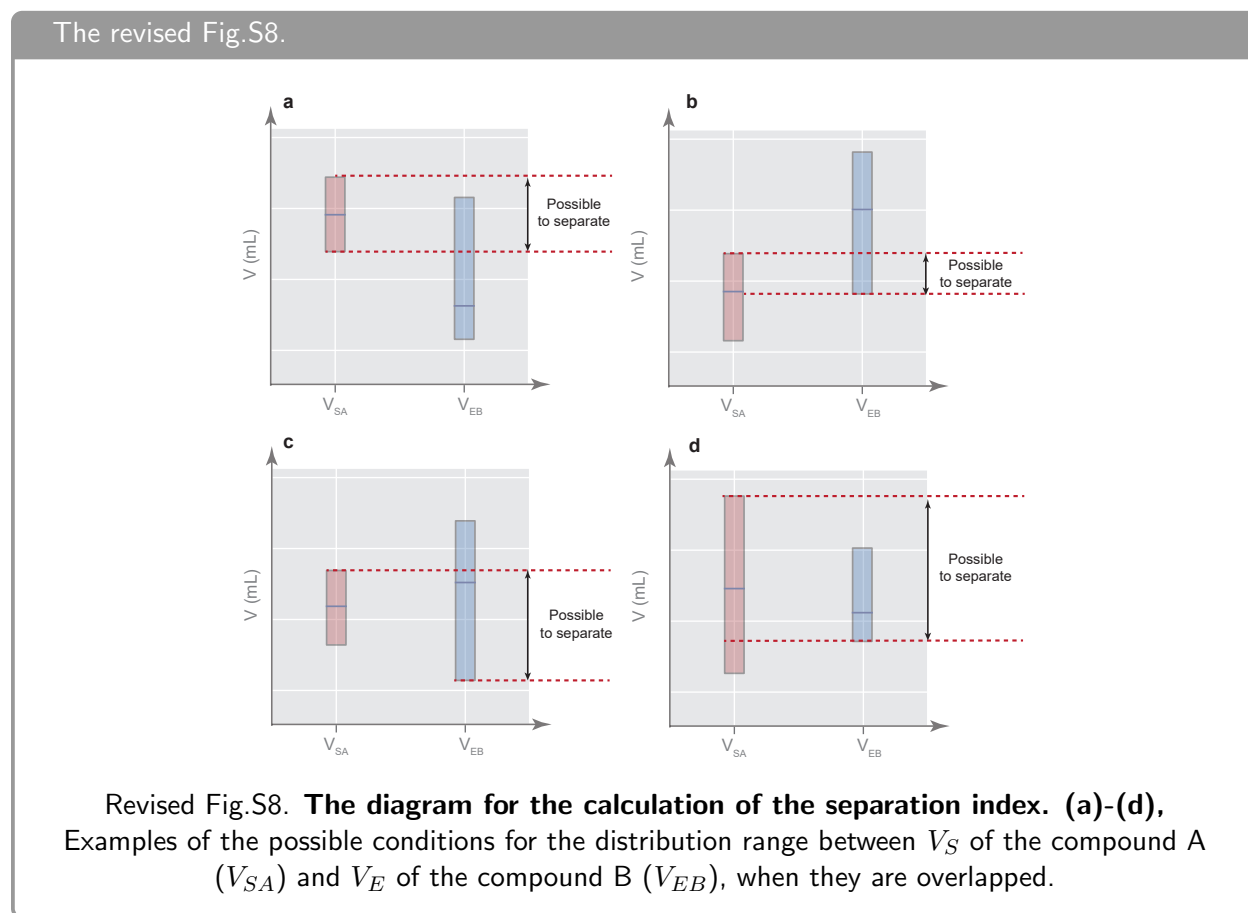
The revised statements in page 12.

Notably, the boxplots refer to the distribution of retention times of compounds whose  $R_f$  values are in specific ranges under given solvent eluents.

**Reviewer Point P 4.6** — I don't understand Fig S7 b

**Reply:** Thanks for your valuable comments. The Fig. S7 aims to facilitate the understanding of Eq. (S.1) which calculates the separation index. For a better illustration and easier understanding, in the revised manuscript, we have revised Fig. S7 (now is Fig. S8) and added some explanation in page 34.

If compound A has not completely eluted yet and the compound B has already started to elute, that is,  $V_{EB} > V_{SA}$ , they are obviously not able to be separated completely. Conversely, if  $V_{EB} < V_{SA}$ , where compound B has not started elute when the compound A has eluted completely, they 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 possible scenarios, there are four conditions when the ranges of  $V_{SA}$  and  $V_{EB}$  are overlapped, 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 (i.e., where the  $V_{SA}$  has the probability to be larger than  $V_{EB}$ ), which corresponds to the four conditions in Eq. (S.1).



**Reviewer Point P4.7** — In order to better understand, it would be better to replace the indices 1 and 2 (which characterize the start of peak volume and the end of peak volume and which suggest a retention order), by the indices b (beginning) and e (end): (VbA instead of V1A) and (VeA instead of V2A). This would make it easier to read.



**Reply:** Thank you for providing us this good advice! We have replaced the indices as suggested by the Reviewer. In addition, as the b (beginning) has the possibility to be confused to the solute B. Therefore, we use s (start), e.g.,  $V_{SA}$ .

**Reviewer Point P 4.8** — Writing error : 2nd page of supplementary information, 6th line from the bottom of the page: Replace  $V_{2A} > V_{2A}$  with  $V_{1B} > V_{2A}$  or as suggested above  $V_{2A} > V_{2A}$  with  $V_{bB} > V_{eA}$

**Reply:** Thanks for reminding us of this issue. We have revised it (page 33).

---

## Reviewer 5

**Reviewer Point P 5.1** — The manuscript describes use of machine learning techniques and statistical analysis for investigating how chemists use thin layer chromatography (TLC) to determine column chromatography conditions. For this purpose, the authors developed an automatic platform and used this platform for generating a big experimental dataset for the column chromatography (CC), which includes 192 compounds. This dataset was used for generating a reliable "artificial intelligence (AI) experience" through a knowledge discovery framework, where the relationship between the retardation factor (RF) values from TLC and retention volumes from column chromatography is represented through explicit equations. The evaluation of these equations confirmed their satisfactory accuracy and generalizability, providing a scientific basis for selecting experimental conditions and enhancing the understanding of chromatography. It is expected that this platform will in the future be expanded to much higher number of compounds. Experimental datasets for TLC and CC and are available for the readers in "DATA – CODE" files. The manuscript is clearly structured. Some small improvements in Figures and "DATA – CODE" files at <https://github.com/woshixuhao/Discovery-of-column-chromatography/tree/main> (e.g. English instead of Chinese titles) can easily improve the manuscript. Other corrections needed are listed below. I recommend this manuscript for the publication in Nature Communications after revisions.

**Reply:** We are particularly grateful for your acknowledgement of our work! In particular, the Reviewer gave us many detailed comments, which makes our manuscript more complete. We sincerely thank you for your efforts! We have revised all the aspects mentioned by the Reviewer carefully.

**Reviewer Point P 5.2** — The following changes are needed in the in the whole manuscript and supplementary material (text, Figs. Tables) and "DATA-CODE files":

**Reply:** Thanks for your valuable advice! We have changed in the whole manuscript as suggested.

**Reviewer Point P 5.3** — "thin layer" should be replaced with "thin-layer".

**Reply:** We have replaced all "thin layer" to be "thin-layer".

**Reviewer Point P 5.4** — "molecules" should be replaced with "compounds".

**Reply:** We have replaced all "molecules" to be "compounds".

**Reviewer Point P 5.5** — “eluent solvent” should be replaced with “mobile phase”.

**Reply:** We have replaced all “eluent solvent” to be “mobile phase”.

**Reviewer Point P 5.6** — All abbreviations should be introduced before being used in the text, in Figs., and “DATA-CODE files”.

**Reply:** We have introduced all abbreviations in the text, Fig, and codes in the revised version.

**Reviewer Point P 5.7** — The abbreviation “CC” for column chromatography should be used in the whole manuscript (like TLC is used and not thin-layer chromatography).

**Reply:** We have utilized the abbreviation “CC” for column chromatography in the revised manuscript.

**Reviewer Point P 5.8** — The font size in Figures should be unified and the size of everything written in Figures should be big enough to be readable, which is not the case in the current version.

**Reply:** Thanks for reminding us of this issue. We have revised the figures to make the font size unified and readable.

**Reviewer Point P 5.9** — According to IUPAC “R<sub>f</sub>” should be replaced with “R<sub>F</sub>” in the whole manuscript (text, Figs., figure legends and equations and “DATA-CODE files”).

**Reply:** We have replaced “R<sub>f</sub>” to be “R<sub>F</sub>” in the whole manuscript.

**Reviewer Point P 5.10** — “(V/V)” should be added after “PE:EA”.

**Reply:** We have added (V/V) after PE:EA.

**Reviewer Point P 5.11** — I hardly believe that the samples were analyzed in this study, but most probably, only mixtures of standard solutions were applied. Therefore, corrections are needed in the whole manuscript.

**Reply:** Thanks for pointing out this for us, we have replaced “samples” to the “mixtures of standard solutions” in the revised manuscript.

**Reviewer Point P 5.12** — In most cases there is no need to use “in this study” (used 14 x).

**Reply:** Thanks, we have deleted redundant “in this study” (only reserve a few that is necessary) in the revised manuscript.

**Reviewer Point P 5.13** — ABSTRACT Correct is “thin-layer”. Corrections are needed in the whole manuscript.

**Reply:** Thanks, we have replaced “thin layer” by “thin-layer” in the whole manuscript as suggested.

**Reviewer Point P 5.14** — According to IUPAC it should be “R<sub>F</sub>” and not “R<sub>f</sub>”.

**Reply:** Thanks, we have replaced “Rf” by “RF” in the whole manuscript as suggested.

**Reviewer Point P 5.15** — 1st paragraph, 1st sentence: Not only in synthetic chemistry.

**Reply:** We have revised it to be “As a pivotal preparative chromatography technique in chemistry” (page 2). Thank you.

**Reviewer Point P 5.16** — 1st paragraph, 2nd sentence: There is no parallelism in “minute laboratory experiments” vs. “substantial industrial-scale processes”.

**Reply:** Thanks. We have revised this sentence to be “Its application spans a wide spectrum, including medicine, chemical industry, and biochemistry” (page 2), in order to better demonstrate the wide application of column chromatography.

**Reviewer Point P 5.17** — 1st paragraph, 3rd sentence: Not only this, but also in isolation of natural bioactive compounds from natural products (e.g. plant materials, food waste materials), etc.

**Reply:** Thanks for your supplement, we have revised this sentence to be “Numerous synthesis laboratories worldwide conduct a staggering volume of CC separations daily to purify synthesized compounds and isolate bioactive compounds from natural products” (page 2), and added some citations.

*Citation:*

Sasidharan, S., Chen, Y., Saravanan, D., Sundram, K. M. & Latha, L. Y. *Extraction, isolation and characterization of bioactive compounds from plants' extracts. African journal of traditional, complementary and alternative medicines* 8, (2011).

Zhang, Q.W., Lin, L.G. & Ye, W.C. *Techniques for extraction and isolation of natural products: A comprehensive review. Chin Med* 13, 1–26 (2018).

**Reviewer Point P 5.18** — 3rd paragraph, last sentence: The mobile phase is not the only parameter that influence the RF values in TLC.

**Reply:** Thanks for pointing out this to us. We have revised this sentence to be “how the experimental conditions and results in TLC experiments (i.e., RF values) influence the outcomes of CC” (page 3).

**Reviewer Point P 5.19** — 6th paragraph, 4th sentence: Specify, because references [11-15] include HPLC and HPLC-MS.

**Reply:** Thanks for pointing out this to us. We have specified the chromatographic techniques, and revised the sentence to be: “Although various quantitative structure-retention relationship (QSRR) models have been established to predict outcomes for a range of chromatographic techniques, including TLC, CC, and high-performance liquid chromatography (HPLC)”. (page 3)

**Reviewer Point P 5.20** — Page 10, 2nd line: Replace “error” with “errors”.

**Reply:** We have revised it by replacing “error” with “errors”.

**Reviewer Point P 5.21** — Page 12, 2nd paragraph, 1st sentence: Avoid repetitions (see Introduction).

**Reply:** Thanks, we have deleted the repetitions.

**Reviewer Point P 5.22** — Page 12, 2nd paragraph, 2nd sentence: Which previous analysis? In the case this is referred to the published work a reference should be added. Page 12, 2nd paragraph, 2nd sentence: Which previous analysis? In the case this is referred to the published work a reference should be added.

**Reply:** Thanks for the valuable question. The previous analysis is the analysis of Fig. 2 in the above section, which shows a potential principle between TLC and CC. For a more clear statement, we have revised it to be “From the above analysis of Fig. 2” (page 11).

**Reviewer Point P 5.23** — MATERIALS AND METHODS This section should be written in such a way that the reader can repeat the experiments. However, based on the current version this is not possible.

**Reply:** Thanks for your comments! In the revised version, we have added a lot of details as suggested, in order to facilitate the readers to better understand and repeat the experiments.

**Reviewer Point P 5.24** — Add data about the standards used.

**Reply:** Thanks for your suggestions! I cannot understand what standards means by the Reviewer. If possible, can you give a more specific description?

**Reviewer Point P 5.25** — Add the specifications of the stationary phase for TLC and CC.

**Reply:** We have added the specifications of the stationary phase for TLC and CC in the revised manuscript (page 24).

The added statements about the choice of stationary phases in TLC and column chromatography.

For TLC, the silica gel plate from Yinlong brand is utilized, and the stationary phase is standard silicone that average particle size is 40  $\mu\text{m}$ . For CC, the column from Agela brand is utilized, and the stationary phase is also standard silicone with average particle size being 40  $\mu\text{m}$ .

**Reviewer Point P 5.26** — Add the separation conditions for TLC (a developing chamber (type, saturated/unsaturated), developing distance; detection conditions; in case of derivatization also derivatization conditions, etc.) and for CC.

**Reply:** We have added the separation conditions for TLC and CC in the revised manuscript (page 24).

The added statements about separation conditions.

In TLC automated platform, the whole process is accomplished by a collaborative robot, in order to guarantee the precision of measured RF values. The developing is conducted in a square flat-bottom developing chamber and the developing time is 300s. Specifically, the robot can finely control the spotting procedure, which makes the spot as little dispersive as possible. Meanwhile, the developed TLC plate is photographed, and the computer vision algorithm is adopted to identify

the geometry center and calculate RF values. In order to guarantee the quality of data, human verification is adopted to calibrate TLC plates that are incorrectly identified by the computer. The detailed conditions can be referred to the protocol<sup>1</sup>.

<sup>1</sup> Xu, H., Zhang, D. & Mo, F. *High-throughput automated platform for thin layer chromatography analysis*. *STAR Protoc* 3, 101893 (2022).

**Reviewer Point P 5.27** — Pages 22 and 23: Move the following sentence from the section Materials and Methods: "While automated preparative chromatographs have long been available, they still possess several drawbacks. Firstly, they do not fully automate the entire column chromatography dataset process. For instance, the sample loading process remains manual. However, given the abundant experiments of column chromatography required in this study, manual sample loading is highly inefficient. Secondly, existing preparative chromatographs are highly integrated, which limits their flexibility and precludes their use for data collection as required in this study."

**Reply:** We have moved these sentences from the section Materials and Methods.

**Reviewer Point P 5.28** — Page 23, line 7 (and several other places in the manuscript): I hardly believe that the samples used. Most probably, only mixtures of standard solutions were applied.

**Reply:** Thanks for pointing out this for us, we have replaced "samples" with "mixtures of standard solutions".

**Reviewer Point P 5.29** — Page 23, 3rd paragraph, lines 3-4: "solution sample" is not appropriate term, replace it with either "standard solutions" or "mixtures of standard solutions".

**Reply:** Thanks for pointing out this for us, we have replaced it with "mixtures of standard solutions".

**Reviewer Point P 5.30** — Page 23, 3rd paragraph: Correct (samples & with UV detector) the sentence "The samples are then eluted and separated, with the UV detector recording the UV signal of the effluent solution during the elution process."

**Reply:** We have corrected the sentence to be "The standard solutions are then eluted and separated, and the UV signal of the effluent solution during the elution process is recorded by the UV detector." (page 23)

**Reviewer Point P 5.31** — Page 23, 4th paragraph: Corrections are needed for "seven experiments are conducted using different eluent ratios": Specify the ratios and explain why 7 if 8 ratios of PE:EA (1:0, 100:1, 50:1, 10:1, 5:1, 2:1, 1:1, and 0:1) are given in the manuscript.

**Reply:** Thanks very much for reminding us of this issue, we find that this is a typo due to counting, the correct number of ratios is 8, and we have corrected it, and specified the ratios (page 23).

**Reviewer Point P 5.32** — Page 24, line 11: Not "molecules", but "compounds". Explain the abbreviation MACCS.

**Reply:** We have replaced the “molecules” to “compounds” and explain the abbreviation MACCS in the revised manuscript, which means “Molecular Access System” (page 24).

**Reviewer Point P 5.33** — Page 25, line 12: Replace “proportion of the developing solvent” with “proportion of the solvents in the developing solvent” or “proportion of the solvents in the mobile phase”.

**Reply:** Thanks for pointing out this for us, we have revised it.

**Reviewer Point P 5.34** — Page 25, line 14: Add “(V/V)” after “0.1”.

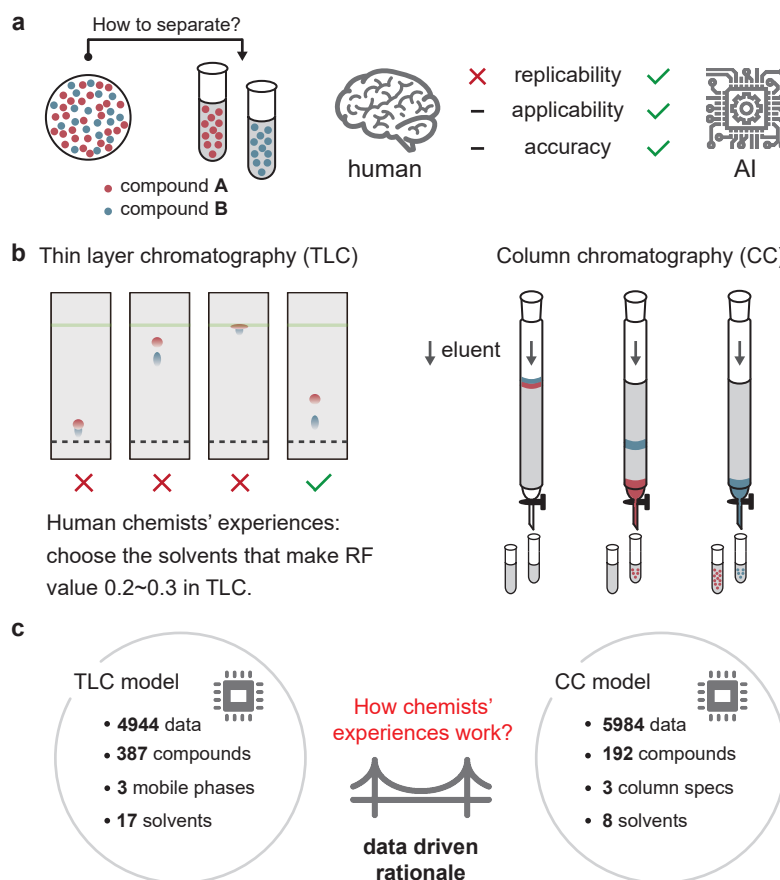
**Reply:** Thanks for pointing out this for us, we have added (V/V) for PE/EA.

**Reviewer Point P 5.35** — FIGURE 1:

1. Replace “Rf” with “RF” in Fig. 1b and Fig. 1 title.
2. In Fig. 1c: Replace “molecules” with “compounds”, and use the term mobile phase;
3. It is not clear what is the source for all the numbers. If this data come from another source a reference should be added.
4. Replace “eluent systems” with “mobile phases”.
5. Replace “eluent solvents” with “solvents” or “constituents of mobile phase”.
6. A parallel is missing between the column (3 column spec) and plate (no data).

**Reply:** Thanks for your suggestions for improving the figure! We have revised the figure as suggested. In the revised version, we have pointed out that the numbers in the figure come from statistics of the dataset. In TLC, we only use the same type of plate, so we do not make a parallel here. Instead, we want to convey more information, so we noted there are 3 mobile phases in TLC experiments.

The revised Fig.1.



Revised Fig.1. **The overview of the framework.** (a), The comparison of human experience and AI experience, exemplified by the task of separating mixtures. (b), The chemical expertise about the determination of separation conditions in CC through RF values of thin-layer chromatography when separating mixtures. (c), The data-driven rationale for the chemists' experiences by statistics and interpretable machine learning through the experimental dataset acquired from the automated platform. The numbers in the figure come from statistics of the dataset.

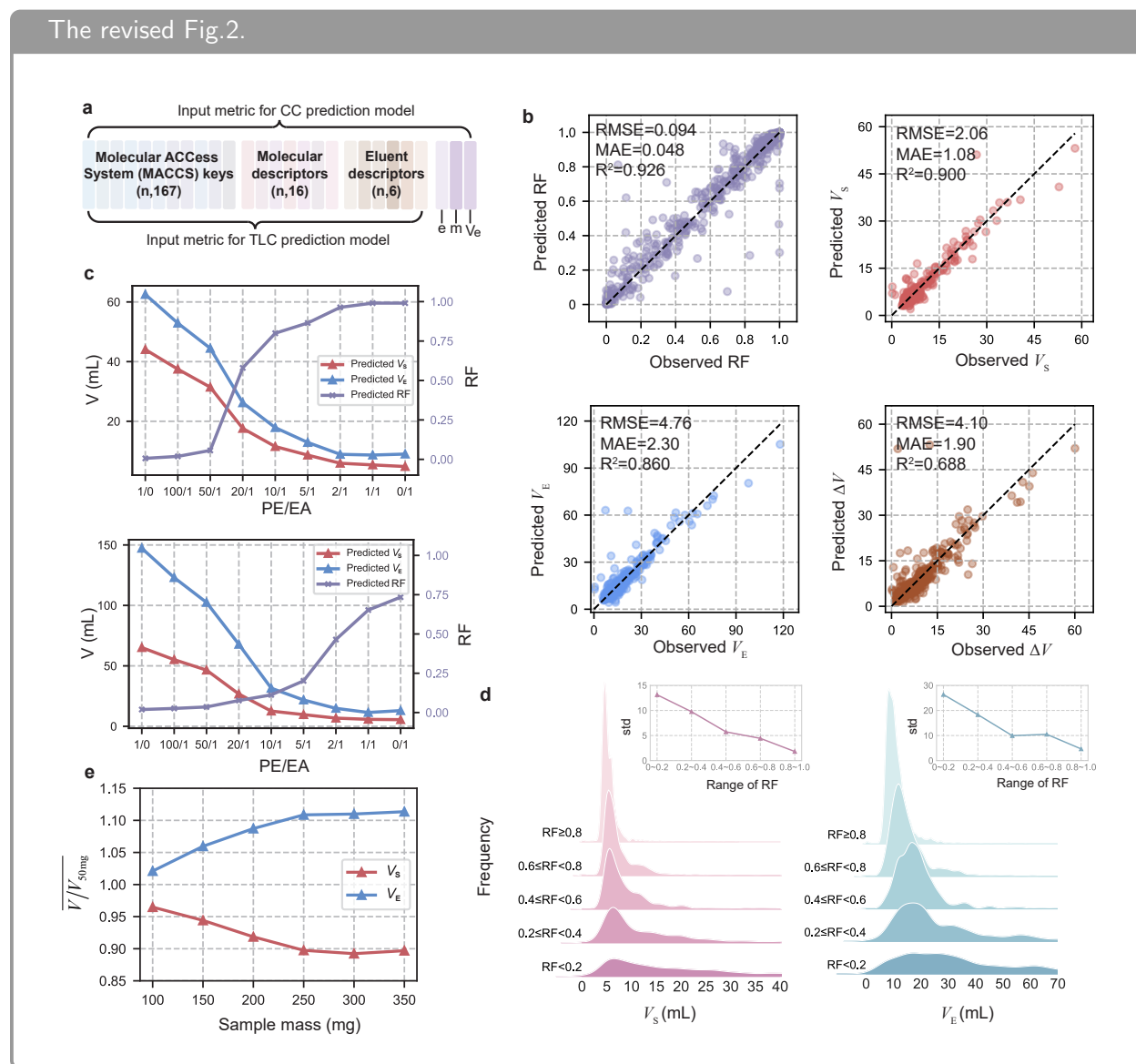
**Reviewer Point P 5.36 — FIGURE 2:**

1. Explain all abbreviations (e.g. Molecular ACCess System (MACCS) keys) and add the space between "MACCS" and "keys".
2. Replace "Rf" with "RF" (also in the legends), which is correct according to IUPAC.
3. In Fig. 2b add 1.0 to scales with RF.
4. The fonts in some figures (e.g. Fig. 2d; Fig. 2a (the subscripted "e" in "Ve")) are too small.
5. Remove the background in Figs. 2b.

6. Add missing lines for x- and y-axes.
7. Fig. 2e: "Ratio" of what? The "red and black triangles" in blue and red rectangles are not visible are not visible without zooming in on the screen.
8. Add the numbers to the scales everywhere where they are missing – where data presented have higher values than are the values on the scales (Fig. 2c, 2d).

**Reply:** Thanks for your suggestions for improving the figure! We have revised the figure as suggested. In the revised version, we have explained all abbreviations and added the space, and replaced all "Rf" with "RF". Meanwhile, we have revised the Fig. 2e to better illustrate the line chart of mean value v.s. loading sample, which obviously shows the trend. We have also added the numbers to the scales where data presented have higher values than the values on the scales. Furthermore, we have enlarged the fonts in these figures.

The revised Fig.2.





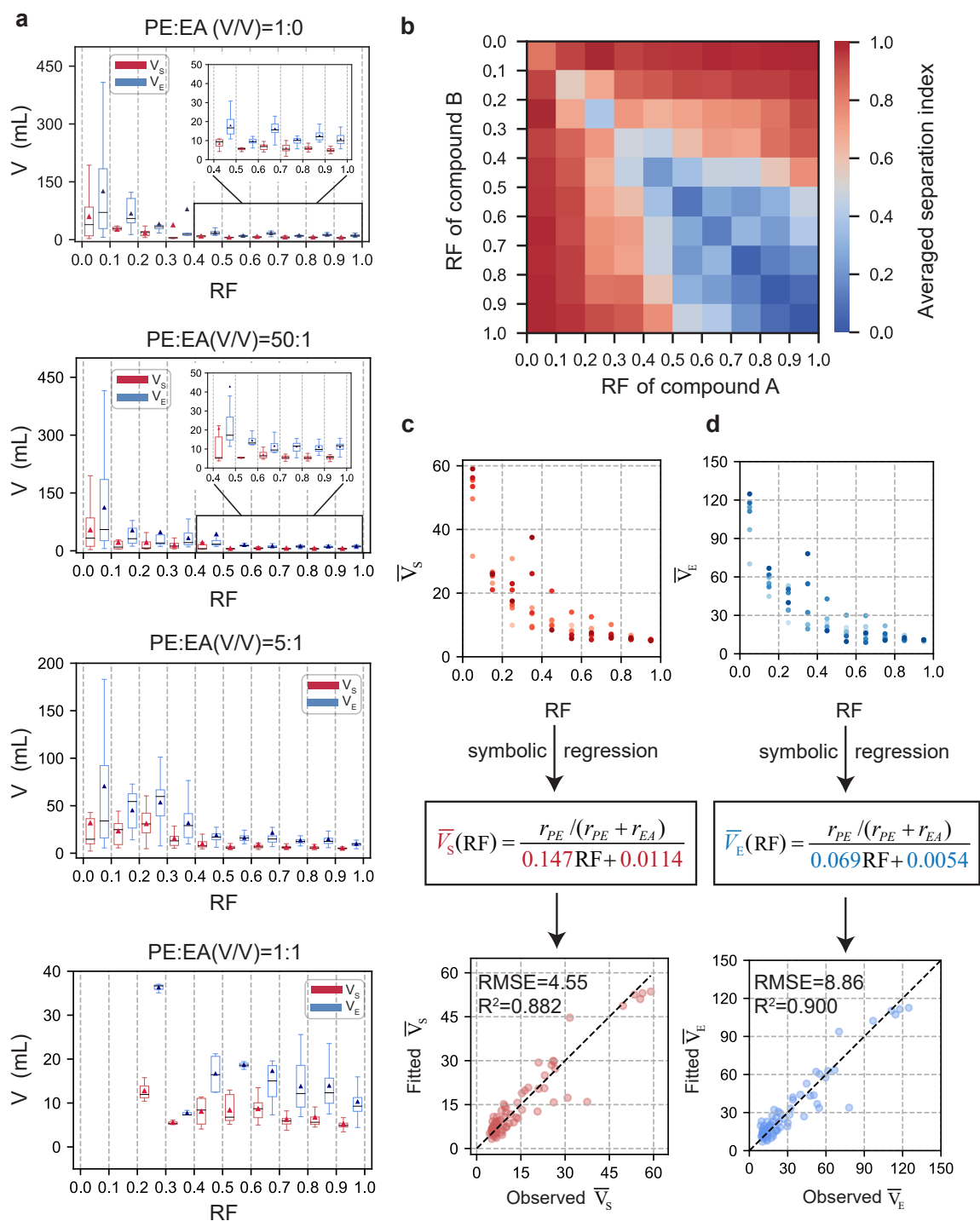
Revised Fig.2. **The results of surrogate models.** **(a)**, The featurization of compound information and experimental conditions for thin-layer chromatography (TLC) and CC dataset. **(b)**, the predictive performance of the surrogate models. **(c)**, The predicted curve of mobile phase and predicted retention volumes and RF value. **(d)**, The distribution of retention volume under the experimental conditions with different RF value ranges, and the standard deviation (std) for different RF value ranges. **(e)**, The influence of loading sample mass. The y-axis is the mean value of the ratio of retention volume under corresponding sample mass to the retention volume under 50 mg with other experimental conditions kept the same.

**Reviewer Point P 5.37 — FIGURE 3:**

1. Replace “Rf” with ”RF” (also in the legends and equations), which is correct according to IUPAC.
2. In Fig. 3c the RF scale should start with ”0.0” and not ”0”.
3. The fonts in some figures, figure legends and equations are too small.
4. Fig. 3a: The ”red and black triangles” in blue and red rectangles are not visible without zooming in on the screen.
5. Remove the background in Figs. 2b.
6. Add the numbers to the scales (Fig. 3a, 3d) everywhere where they are missing -data presented have higher values than are the values on the scales.

**Reply:** Thanks for your suggestions for improving the figure! We have revised the figure as suggested. In the revised version, we have revised the figure to make the mean value more visible, and replaced all “Rf” with “RF”. Furthermore, we have enlarged the fonts in these figures.

The revised Fig.3.



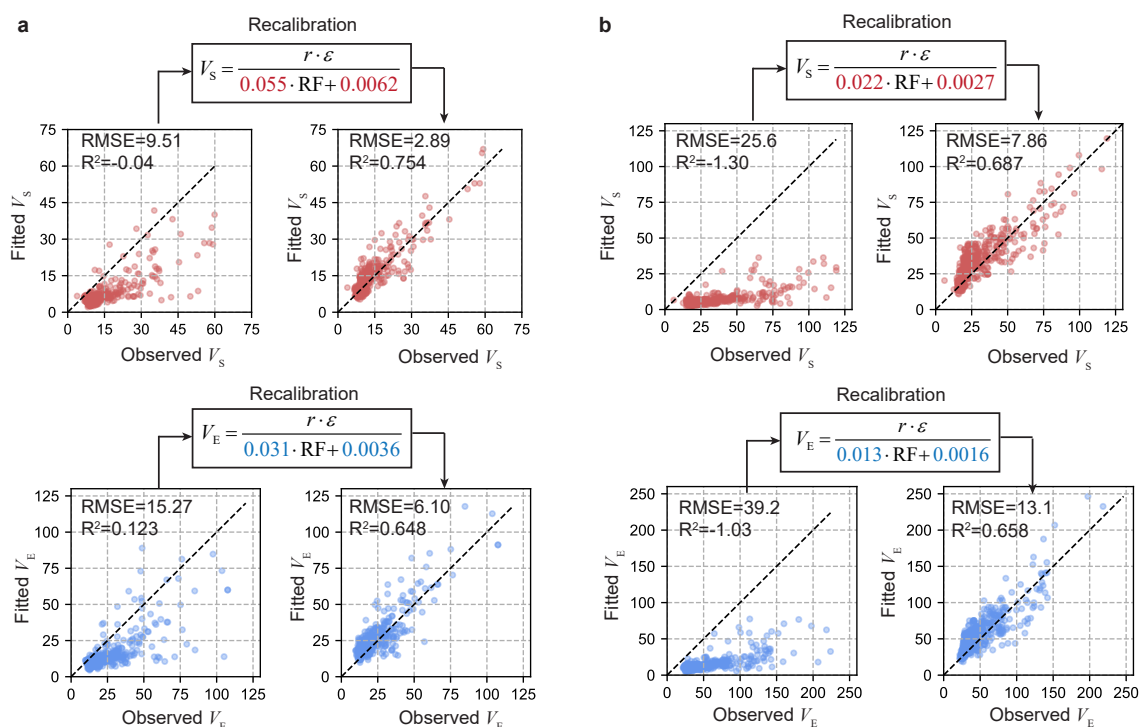


**Reviewer Point P 5.39** — FIGURE 5:

1. Replace "Rf" with "RF".
2. The fonts in some figures and equations are too small.
3. Add the numbers to the scales (Fig. 5a) everywhere where they are missing -data presented have higher values than are the values on the scales.

**Reply:** Thanks for your suggestions for improving the figure! We have revised the figure as suggested. In the revised version, we have enlarged the fonts in Fig.5. Meanwhile, we have added the numbers to the scales where data presented have higher values than the values on the scales.

The revised Fig.5.



Revised Fig.5. **The adaption of the model to diversified compounds. (a)**, The fitted and observed retention volume by directly utilizing discovered equations from 4g column (left) and the fitted equations by nonlinear regression and its predictive performance on the extrapolation for tandem 4g column (right). **(b)**, The fitted and observed retention volume by directly utilizing discovered equations from 4g column (left), and the fitted equations by nonlinear regression and its predictive performance on the extrapolation for tandem 25g column (right).

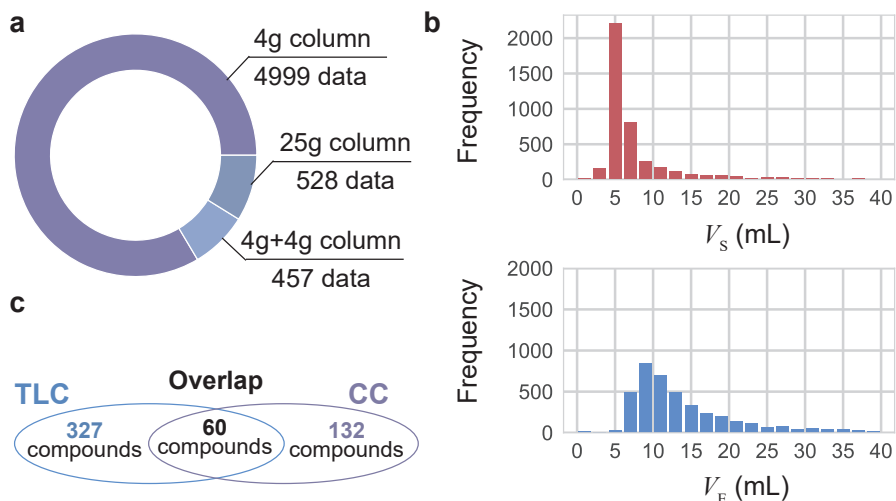
**Reviewer Point P 5.40** — FIGURES S1 and S6: Only one of the figures Fig. S1 or Fig. S6 should be selected, while the other should be deleted.

**Reply:** Thanks for your suggestions, we have deleted Fig. S6 and reserved Fig. S1, which is clearer.

**Reviewer Point P 5.41** — FIGURE S3: Not “molecules” but “compounds”.

**Reply:** Thanks for pointing out this for us, we have revised it.

The revised Fig.S3.

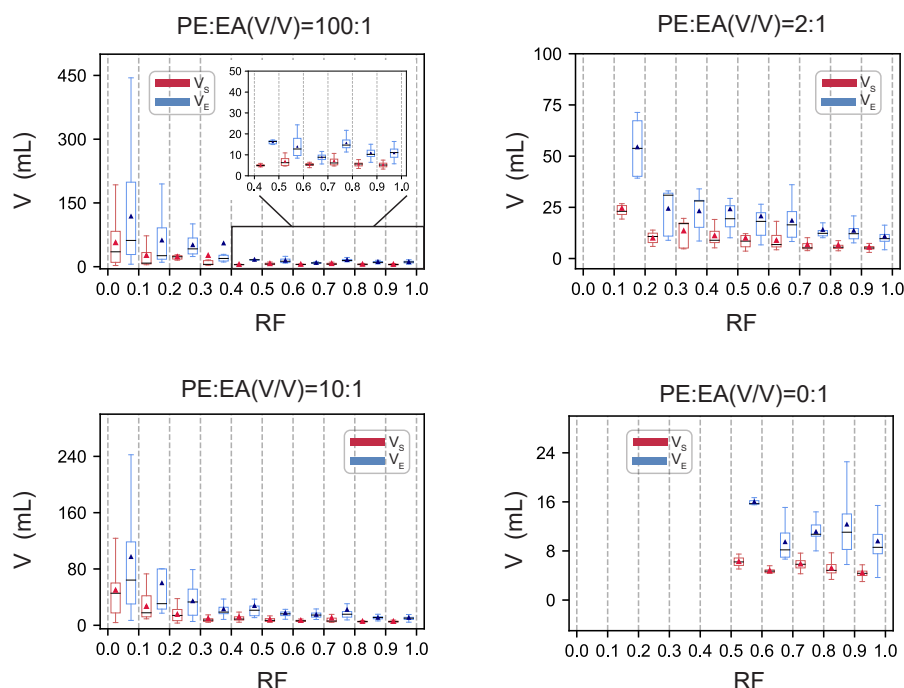


Revised Fig.S3. **The automatic platform for column chromatography and the introduction to the column chromatography dataset.** (a), he number of datapoints from different column specifications. (b), The distribution of the retention volume in the collected column chromatography (CC) dataset. (c), The intersections between the TLC dataset and CC dataset.

**Reviewer Point P 5.42** — FIGURE S4: See comments for Figs. In the manuscript and prepare corrections accordingly.

**Reply:** Thanks, we have revised the figure as suggested.

The revised Fig.S4.



**Revised Fig.S4. The distribution of retention volumes in diversified ranges of RF values under different mobile phase proportions.** Here are the situations for PE:EA (V/V)=100:1, 10:1, 2:1, and 0:1.

**Reviewer Point P 5.43** — EQUASIONS 1, 2, 5, 6, 7, 8, 9 and 10: Pages 15, 16, 19, 20, 27, 28: Replace “Rf” with “RF”.

**Reply:** Thanks for pointing out this for us, we have revised these equations.

**Reviewer Point P 5.44** — REFERENCES

1. For the latest papers add paper number and not only volume and year.
2. Ref [4]: The authors and the title of the chapter and relevant pages should be added. Add data about the editor "book title: Preparative Chromatography: of Fine Chemicals and Pharmaceutical Agents; editor: Henner Schmidt-Traub".
3. Ref [7-11, 13-16]: Add names of other authors.

**Reply:** Thanks for pointing out the problems existed in the reference, we have revised it as suggested.

**Reviewer Point P 5.45** — DATA – CODE: (<https://github.com/woshixuhao/Discovery-of-column-chromatography/tree/main>)

1. File TLC-dataset.xlsx – corrections are needed, because columns (E-I) with data are not related to the manuscript
2. File dataset-1013.xlsx - all Chinese column titles should be translated to English.
3. File dataset-25g-1218.xlsx - corrections are needed because the column P is without the title, the column O with the Chinese title, but without data.
4. File dataset-4+4g-1020.xlsx - corrections are needed because the column P is without a title and there are columns with the Chinese titles.

**Reply:** Thanks for pointing out the problems existed in the data file, we have revised it as suggested and reupload the revised version of dataset in github.

**Reviewer Point P 5.46** — Remarks on code availability: README file is available, but some corrections are needed in the following files (DATA – CODE: <https://github.com/woshixuhao/Discovery-of-column-chromatography/tree/main>):

1. File TLC-dataset.xlsx – corrections are needed, because columns (E-I) with data are not related to the manuscript
2. File dataset-1013.xlsx - all Chinese column titles should be translated to English.
3. File dataset-25g-1218.xlsx - corrections are needed because the column P is without the title, the column O with the Chinese title, but without data.
4. File dataset-4+4g-1020.xlsx - corrections are needed because the column P is without a title and there are columns with the Chinese titles.

**Reply:** Thanks for pointing out the problems existed in the data file, we have revised it as suggested.

## Response to the reviewers

We would like to sincerely express our deep gratitude to the editors and referees for their insightful comments and suggestions on the manuscript, which are fair, encouraging, and constructive. After carefully studying the comments, we have made corresponding changes in the revised manuscript. The specific responses to the comments are listed below. In the revised manuscript, text in **red** is the revision for Referee #5. text in **blue** is the revision for the Editors. Changes in the manuscript are also emphasized by underlining within the reply, and the corresponding content is included in a gray box.

---

### Reviewer 5

**Reviewer Point P 5.1** — Although the authors have improved the clarity of the manuscript, the revised manuscript needs some minor revisions listed below. I recommend this manuscript for the publication in Nature Communications after revisions.

**Reply:**

Thanks very much for your approval of our work! We sincerely appreciate your comprehensive reading and understanding of our paper. According to your comments, we have revised the manuscript by some corrections.

**Reviewer Point P 5.2** — According to IUPAC in “RF” “F” should be in subscript and this should be replaced in the whole manuscript (text, figures (Figs. 1b, 2, 3, 5, S5), figure titles (Figs. 1, 2, 3, S5) and equations (1, 2, 5-10).

**Reply:** We have replaced “RF” to “R<sub>F</sub>” in all texts, figures and legends.

**Reviewer Point P 5.3** — Page 12, line 301: Replace “From the above analyses of Fig. 2” with “From the results of the analyses presented in Fig. 2”.

**Reply:** Thanks for point out this. We have replaced this sentence:

The revised statements in page 12.

From the results of the analyses presented in Fig. 2

**Reviewer Point P 5.4** — In the whole manuscript (text, and Figs., figure titles) move “(V/V)” after the numbers of the solvents ratio (e.g., PE:EA (1:0, 100:1, 50:1, 10:1, 5:1, 2:1, 1:1, and 0:1 (V/V)).

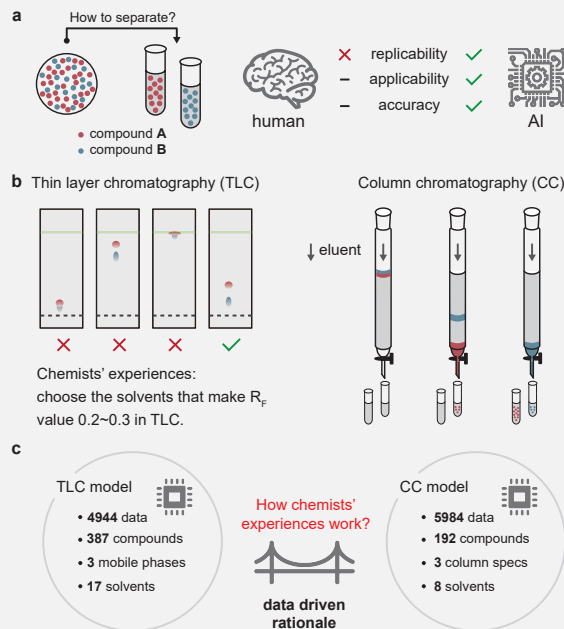
**Reply:** Thank you for the comments. We have moved (V/V) after numbers of the solvents ratio in the revised manuscript.

**Reviewer Point P 5.5** — Figure 1.b: Delete “human”.

**Reply:** Thank you for your suggestion! We have deleted “human” in Fig. 1b.



revised Fig.1.



**Reviewer Point P 5.6** — Figure 2.d: Add missing lines for x- and y-axes and units for frequency.

**Reply:** Thank you for highlighting this matter. In a joint probability density diagram, the x and y axes are often concealed for clearer visualization, as its primary purpose is to depict variations in distribution across diverse conditions. For rigorous, we provided the y-axis of the probability of density. The "frequency" may lead to a misunderstanding, so we replace it as "density", which does not bear any units.

**Reviewer Point P 5.7** — Reference [6] in the revised version should be corrected: M. Michel, A. Epping, A. Jupke are the authors of the chapter 6 "Modeling and Determination of Model Parameters" (<https://onlinelibrary.wiley.com/doi/book/10.1002/3527603484>; <https://onlinelibrary.wiley.com/doi/10.1002/3527603484.ch6>) and Henner Schmidt-Traub is the editor of the book "Preparative Chromatography: of Fine Chemicals and Pharmaceutical Agents", in which chapter 6 was published. Check also the pages as those on the website differ from those in the manuscript.

**Reply:** Thanks for your invaluable comments! We have revised the reference.

The revised reference 6.

Michel, M., Epping, A., Jupke A. Modelling and determination of model parameters In *Preparative Chromatography: Of Fine Chemicals and Pharmaceutical Agents*. Eds. Henner Schmidt-Traub, Michael Schulte, and Andreas Seidel-Morgenstern. Weinheim: Wiley, 2005, 215-312.