Analytica Chimica Acta: X 2 (2019) 100010

Contents lists available at ScienceDirect

Analytica Chimica Acta: X

journal homepage: www.journals.elsevier.com/analytica-chimica-acta-x

Rational design of mixtures for chromatographic peak tracking applications via multivariate selectivity

Daniel W. Cook [a], Kelson G. Oram [b], Sarah C. Rutan [a, *], Dwight R. Stoll [b]

[a] Department of Chemistry, Virginia Commonwealth University, Richmond, VA, 23284, USA
[b] Department of Chemistry, Gustavus Adolphus College, St. Peter, MN, 56082, USA

ARTICLE INFO

Article history:
Received 18 September 2018
Received in revised form
24 January 2019
Accepted 27 February 2019
Available online 5 March 2019

Keywords:
Multivariate curve resolution
Rational design of mixtures
Net analyte signal

ABSTRACT

Chromatographic characterization and parameterization studies targeting many solutes require the judicious choice of operating conditions to minimize analysis time without compromising the accuracy of the results. To minimize analysis time, solutes are often grouped into a small number of mixtures; however, this increases the risk of peak overlap. While multivariate curve resolution methods are often able to resolve analyte signals based on their spectral qualities, these methods require that the chromatographically overlapped compounds have dissimilar spectra. In this work, a strategy for grouping compounds into sample mixtures containing solutes with distinct spectral and, optionally, with distinct chromatographic properties, in order to ensure successful solute resolution either chromatographically or with curve resolution methods is proposed. We name this strategy rational design of mixtures (RDM). RDM utilizes multivariate selectivity as a metric for making decisions regarding group membership (*i.e.*, whether to add a particular solute to a particular sample). A group of 97 solutes was used to demonstrate this strategy. Utilizing both estimated chromatographic properties and measured spectra to group these 97 analytes, only 12 groups were required to avoid a situation where two or more solutes in the same group could not be resolved either chromatographically (*i.e.*, they have significantly different retention times) or spectrally (i.e., spectra are different enough to enable resolution by curve resolution methods). When only spectral properties were utilized (*i.e.*, the chromatographic properties are unknown ahead of time) the number of groups required to avoid unresolvable overlaps increased to 20. The grouping strategy developed here will improve the time and instrument efficiency of studies that aim to obtain retention data for solutes as a function of operating conditions, whether for method development or determination of the chromatographic parameters of solutes of interest (*e.g.*, k_w).

1. Introduction

It is frequently necessary to carry out screening experiments in the course of liquid chromatographic (LC) method development. Typically, a number of target analytes are identified, for example for a metabolite profiling experiment [1–4] or a pharmaceutical degradation or impurity analysis [5–7], and optimal chromatographic conditions for the separation of these analytes are sought. This optimization of chromatographic conditions is arguably the most time-consuming step of any chromatographic analysis in which many parameters, such as stationary phase, temperature, gradient time, mobile phase composition, etc., must be considered. This optimization may be carried out in a number of ways, including trial and error variation of conditions, as well as more systematic approaches, like those used in commercial software such as DryLab [8–10] or our recently developed LC simulation software [11]. In these latter cases, models for the prediction of the retention of the target analytes are used to map the separation space and determine the conditions that optimize the separation to meet a particular goal (e.g., maximize resolution at a particular analysis time). Fits of retention time as a function of mobile phase composition to models such as linear solvent strength (LSS) [12] or Neue-Kuss (NK) [13] are used to obtain model parameters, that then allow for *in-silico* optimization of chromatographic conditions. These parameters (S and k_w in the case of LSS; a, B and k_w in the case of NK) are then used to predict retention using computer calculations or simulations to find the optimal separation condition.

Another situation that requires screening of a large number of

* Corresponding author.
E-mail address: srutan@vcu.edu (S.C. Rutan).

https://doi.org/10.1016/j.acax.2019.100010

compounds at a range of chromatographic conditions is studies aimed at understanding and rationalizing chromatographic selectivity. One example of this type application is the hydrophobic subtraction (HSM) model for chromatographic selectivity [14,15]. In this method for selectivity characterization, a number of model solutes are analyzed on one or more chromatographic columns to determine a set of column selectivity parameters. In the original HSM scheme, 67 compounds were characterized on 10 different columns [16], followed by a more complete study that involved characterization of 16 solutes on more than 300 columns [17]. This is clearly a task where rationally designed mixtures could accelerate the screening process. Other methods based on principal components analysis have also been proposed, which also depend on the screening of a number of different solutes under multiple chromatographic conditions [18–20].

Whether the goal is to parameterize retention or simply screen retention behavior for a large number of solutes, the analyst will want to group the solutes in as few mixtures as possible while retaining the ability to determine which peaks correspond to which compounds that are present in the mixtures. The ability to identify peaks in chromatograms over a range of experimental conditions is generally referred to as peak tracking in the literature [5,7,21–23]. In some cases, this task can be fairly straightforward; for example, if the set of target solutes consists of one or more homologous series of compounds, the members of each homologous series can be prepared in a single mixture, as the retention order of the compounds is known. For more complex sets of solutes, automated peak tracking strategies can be of great help [21,24]. These strategies involve tracking the retention of several solutes over a range of chromatographic conditions in order to optimize the separation of these species of interest. Peak tracking can also be useful when multiple separations are performed on different orthogonal chromatographic columns [25]. While many such methods have been described in the literature, little attention has been paid to the design of sample mixtures themselves in ways that optimize the effectiveness of these peak tracking methods.

While the design of such mixtures would traditionally avoid chromatographic overlap so that interpretation of the results is straightforward, this requirement is unnecessarily stringent, given the capabilities of modern multivariate curve resolution (MCR) strategies [26]. Incorporating spectral information via a diode array ultraviolet-visible or mass spectrometric detector allows for reliable peak detection, even for compounds that have chromatographic resolution (R_S) much less than one [24,27]. These curve resolution strategies, such as multivariate curve resolution-alternating least squares (MCR-ALS) [28–30] and parallel factor analysis (PARAFAC) [31,32] are able to resolve solutes that are significantly overlapped chromatographically based on the differences between their spectra; however, when two (or more) solutes have identical or very similar spectra, curve resolution strategies are unable to resolve these peaks. The implication of this in preparing mixtures of compounds to be characterized for method development and/or parameterization is that solutes with very similar spectra must either be completely resolved chromatographically or placed in separate mixtures.

The work here describes a strategy entitled rational design of mixtures (RDM) for rationally grouping solutes into mixtures based on the likelihood that the analytes in a group can be resolved in the chromatographic and/or spectral dimensions. RDM relies on a multivariate selectivity metric [33–35] to group only chromatographically resolved or spectrally distinct analytes together in the same sample, while placing chromatographically overlapped analytes that have similar spectra into separate groups.

2. Theory

2.1. Multivariate curve resolution

The use of chemometric methods is very useful for the analysis of multidimensional data. These additional dimensions provide information that may help to differentiate between different analyte signals and enable the analysis of these analytes even in the presence of interferents, effectively increasing the selectivity of the analysis for each analyte. The use of multivariate detectors, such as a diode array detector or a mass spectrometric detector, for liquid chromatographic analyses provides these additional dimensions of data. This enables the use of multivariate curve resolution (MCR) strategies that can mathematically resolve analyte signals even if they are overlapped in one or more dimensions of the data (*i.e.,* chromatographic and spectral overlap). This results in pure chromatographic and spectral profiles for each of the compounds present in the sample. The spectral profiles assist in the identification of each compound while the chromatographic profiles can be used for quantitation, or in the case of peak tracking, to extract chromatographic parameters for the purpose of modelling retention.

Two of the most popular MCR strategies are multivariate curve resolution by alternating least squares (MCR-ALS) [28–30] and parallel factor analysis (PARAFAC) [31,32]. The principal difference between these two methods is that MCR-ALS is based on a bilinear model, whereas PARAFAC is based on a trilinear or higher (*e.g.,* quadrilinear) model. Because of this PARAFAC is considered to give unique solutions to the curve resolution problem, but doing this successfully requires that no shifts are present in retention time between samples. This is a rather limiting constraint that typically necessitates an additional step of retention time alignment prior to PARAFAC analysis. For applications in which retention changes are deliberately made (e.g., by changing the stationary phase), PARAFAC is not applicable. MCR-ALS on the other hand has no requirement of retention time stability across analyses, but suffers from rotational ambiguity. This ambiguity means that the results from MCR-ALS are not unique and therefore careful application of mathematical constraints during the ALS step is needed to ensure accurate results. It has also been shown that analyzing several samples simultaneously, as would be done for a large peak tracking experiment, greatly decreases rotational ambiguity [36].

2.2. Multivariate selectivity

A figure of merit called multivariate selectivity (SEL) has been previously developed to measure the selectivity of an analysis for a target analyte considering the entirety of the multidimensional data [33,37]. SEL is defined in terms of the net analyte signal (NAS) framework originally developed by Lorber [37–39]. When analyzing first-order data (*i.e.*, a vector of data such as an absorption or emission spectrum, or a chromatogram) the data can be represented in N-dimensional space where N is the number of elements in the vector. When all signals contributing to the data are plotted in this space, the pure analyte signal is represented by a vector, and all other signals constitute a hyperplane. The portion of the analyte vector that is orthogonal to the hyperplane is defined as the NAS. SEL is equal to the sine of the angle between the analyte vector and the hyperplane. As this angle increases, SEL approaches one and the analysis becomes more selective. This definition is easily extended to higher order data such as second-order data (*i.e.*, a matrix) such as that obtained from an LC-DAD analysis. SEL can be calculated using eq. (1) where **A** and **B** are matrices containing the normalized, pure component spectral and chromatographic profiles, respectively, for each signal contributing to the data. The superscript 'T' and '•' refer to a matrix transpose and the Hadamard products,

respectively, and subscript '*i,i*' refers to the *i*-th diagonal element of the SEL matrix [39,40]. For first-order data, such as spectra only, the **B** term drops out of eq. (1). Likewise, for third-order (or higher order) data additional terms can be added.

$$SEL_i = [(\mathbf{A}^T\mathbf{A}\cdot\mathbf{B}^T\mathbf{B})^{-1}]_{i,i}^{-½} \quad (1)$$

Strictly speaking, calculating SEL with both chromatographic and spectral matrices represents the selectivity of an analysis when using a trilinear model, such as PARAFAC, whereas for MCR-ALS, the SEL metric depends solely on the spectral profiles [41]. The use of the MCR-ALS SEL metric is somewhat pessimistic, in that it doesn't account for the fact that chromatographic selectivity does exist, and it is well known that each compound will appear in only one region of the chromatogram. Conversely, the PARAFAC SEL metric is optimistic, in that it assumes that one can rely on an assumption of a trilinear data structure, which is generally not strictly true for LC data. The proposed RDM strategy allows for both metrics to be used to guide the design of samples that can be resolved using the chromatographic and/or spectral dimensions.

3. Strategy

In order to most appropriately assign compounds to samples, the likelihood that the analytes can be differentiated once analyzed must be assessed. This likelihood depends on their chromatographic separation and/or their potential to be separated mathematically via differences in their spectra. As described in the Theory section, multivariate selectivity (SEL) is a measure of this likelihood of resolution from other signals and thus was selected as the metric used to guide assignment of each compound into a group.

RDM enables grouping of compounds based solely on spectra or based on both the spectral and chromatographic properties of the compounds. In liquid chromatography applications, UV absorption spectra are typically obtained experimentally for each compound. Chromatographic profiles may be obtained experimentally as well; however, estimates of chromatographic profiles may be obtained from a variety of sources, including chromatographic simulators [8–10,42–44] or structure activity relationships [45,46]. Typically, a retention factor would be estimated, and a Gaussian peak of a specified width would be generated as the chromatographic profile for each target compound. One means of estimating chromatographic profiles would be to make very rough estimates of LSS parameters. In this case, a single retention factor at one mobile phase composition, and a 'typical' S value for the compound class under investigation can be used to obtain the k_w value using eqn. (2) [12].

$$\ln k = \ln k_w - S\phi \quad (2)$$

Then, the retention time (t_R) and peak width (σ) at an alternative mobile phase composition can be estimated as

$$t_R = t_0(1 + k) \quad (3)$$

$$\sigma = \frac{t_R}{\sqrt{N}} \quad (4)$$

Or, if it is desired to estimate chromatographic profiles for mobile phase gradient conditions, the LSS equations shown in eqs. (5)–(8) can be used to estimate retention time and peak width [12].

$$t_R = \frac{t_0}{b}\ln(k_0 b + 1) + t_0 \quad (5)$$

$$\sigma = \frac{4}{\sqrt{N}}(t_0)\left(1 + \frac{k^*}{2}\right) \quad (6)$$

$$k^* = \frac{2k_0}{bk_0 + 2} \quad (7)$$

$$b = t_0 S\frac{\Delta\phi}{t_G} \quad (8)$$

Here, t_G is the gradient time, k_0 is the retention factor at the start of the gradient, $\Delta\phi$ is the difference between the final and initial solvent composition and k^* is the gradient retention factor (retention factor at the column midpoint). In this case, the gradient compression factor, *G*, is omitted for simplicity. Here, an overly pessimistic value for the efficiency (*N*) is chosen, which will lead to more severe chromatographic overlap. This can help to ameliorate the use of inaccurate estimates for the LSS parameters.

Whether chromatographic profiles and spectra are utilized, or only spectra, the grouping strategy is identical. The sole difference in implementation of the strategy in these two cases is related to how the groups are initialized. Groups are initialized by choosing two similar compounds and placing them into two separate groups. When chromatographic profiles are used, the two compounds with most similar *k* values are selected. When only spectra are used, the two compounds with the most similar spectra as measured by their correlation coefficient are used as the initial two groups. Fig. 1 outlines the entire RDM strategy. After the initialization of the groups with two compounds, a third compound's SEL is calculated against each of the existing groups. The compound is then placed into the group which gives the highest SEL value as long as it exceeds a preset SEL threshold. If the threshold is not met, a new group is created with that compound. Particularly when chromatographic profiles are included in the SEL calculation, SEL values of one (*i.e.*, totally selective) for multiple groups are not uncommon. In this case the compound is placed into the group which has the most different *k* values. This is performed by taking the difference between the *k* of the compound and the most similar *k* value in each group. The compound is placed in the group with the maximum difference between these two *k* values. This encourages a broader range of *k* in each group. This process then continues for a fourth compound and so on until all compounds have been assigned to a sample.

An optional constraint on the algorithm is a limitation on group membership. If used, after each compound is placed into a group, the number of compounds in the group is checked. If the group membership is at the preset limit, no more compounds are allowed to be added to that group. Another optional step is an iterative optimization. As the final step in the grouping, the iterative optimization step removes a single compound at a time and reevaluates the compound's SEL against each group and places the compound in the group with the highest SEL. Once each compound has been reevaluated, the process repeats, analyzing the compounds in a different order. The number of iterations is equal to the number of compounds as each iteration starts by replacing a different compound. This optimization step may allow more than the target number of compounds within a group as it does not take group membership into account; however, it is unlikely to increase group membership drastically in any one group.

4. Experimental

All calculations were performed in MATLAB (R2016a; Mathworks, Inc., Natick, MA) with codes written in-house on a standard

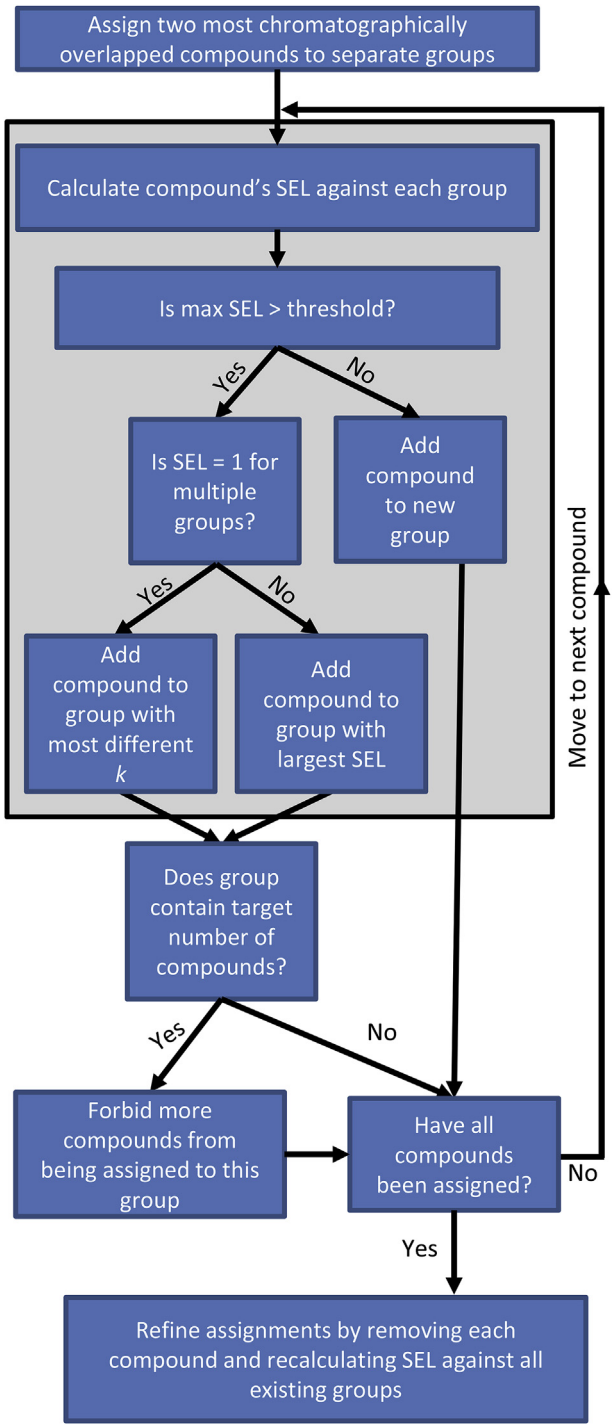

Fig. 1. Strategy for assigning compounds to samples using multivariate selectivity in RDM. The steps enclosed by the gray box represent the core grouping algorithm.

desktop PC. The MATLAB code required for the calculations is provided in the Supplementary Data.

A set of 97 probe analytes was utilized to demonstrate the ability of the current strategy to group analytes based on their spectral and chromatographic properties. The names and abbreviations for each analyte are listed in the Supplementary Data Table S1. For each probe analyte, the UV absorption spectra between wavelengths 212–600 nm were obtained using samples prepared by dispersing the pure compound in 50/50 ACN/water at 10 mg/mL, and then diluting to working concentration of either 100 (gradient) or 1,000 (isocratic) μg/mL in 50/50 ACN/water. UV spectra were recorded during elution from a Zorbax SB C18 column (50 mm × 2.1 mm i.d., 3.5 μm particles; Agilent Technologies) under solvent gradient conditions. The A solvent was 10 mM phosphoric acid in water at pH 2.3, and the B solvent was ACN. The solvent gradient was 5-95-95-5-5% B from 0 to 2.0-2.25-2.52-3.5 min., and the column was thermostated at 40 °C. These spectra are provided in the Supplementary Data section. Isocratic retention times were measured on an Eclipse Plus C18 column (50 mm × 2.1 mm i.d., 3.5 μm particles; Agilent Technologies) using a mobile phase of 30/70 (v/v) ACN/60 mM potassium phosphate, pH 2.8. The flow rate was 1.0 mL/min, and the column was thermostated at 35 °C. Isocratic retention factors were then calculated using the elution time of thiourea as a dead time marker, taking care to correct the retention times for the extra-column volume of the instrument. In cases where the retention times were very short or very long in the 30/70 ACN/buffer mobile phase, retention measurements were obtained using lower (*e.g.*, 10, 20 %ACN) ACN percentages or higher (*e.g.*, 40, 50 %ACN) ACN percentages, and then the retention factors were estimated by extrapolation using the LSS retention model. These retention factors are listed in Supplementary Data Table S1. The instrument used for gradient experiments was from Agilent Technologies, comprised of a 1290 binary pump and autosampler, and 1100 column compartment and diode array detector. The instrument used for the isocratic experiments was a HP1090 liquid chromatograph with a diode array detector. Chromatographic profiles for each solute were calculated from k using eqns. (5)–(8). With both the chromatographic and spectral profiles (provided in the Supplementary Data) for each analyte, grouping was performed as outlined in the Strategy section above and as illustrated in Fig. 1.

5. Results and discussion

First, the 97 analytes were grouped based solely on their spectral differences. This would be the case in which retention shifts drastically, such as when different column chemistries are used. In these cases, it is impossible to estimate chromatographic profiles that would be meaningful across all analyses. The compounds were grouped at a SEL threshold 0.2 utilizing the iterative optimization step and not imposing a limit on group membership. Prior to grouping, each spectrum was corrected for a baseline offset by subtracting the intensity at the last wavelength from all other wavelengths. Any negative absorbance values resulting from spectral noise were set to zero to maintain non-negative spectra. Spectra were then normalized to unit length as required for proper calculation of SEL. Application of the proposed grouping strategy resulted in 20 groups with no more than six compounds in each group. Fig. 2 shows the spectra in each group and Table 1 lists the compounds in each group along with each compound's SEL relative to the group.

When LSS parameters can be reasonably estimated, such as in the case of experiments designed to obtain more accurate LSS or NK parameters, chromatographic information can be incorporated via chromatographic profiles. Here, gradient chromatographic profiles were created based on the k value for each compound under isocratic conditions. For the 97 analytes, the k values measured at $\phi = 0.30$ (or extrapolated, for cases where the retention times were too short or too long at $\phi = 0.30$) ranged from 0.05–193 (These values are provided in the Supplementary Data in Table S1. The chromatographic parameters chosen to create the gradient chromatographic profiles are listed in Table 2, with an assumed S value of 10. N was conservatively estimated at 1000 plates to ensure that any overlap that may occur in the experimental data is captured in the simulated profiles. The chromatographic profiles were

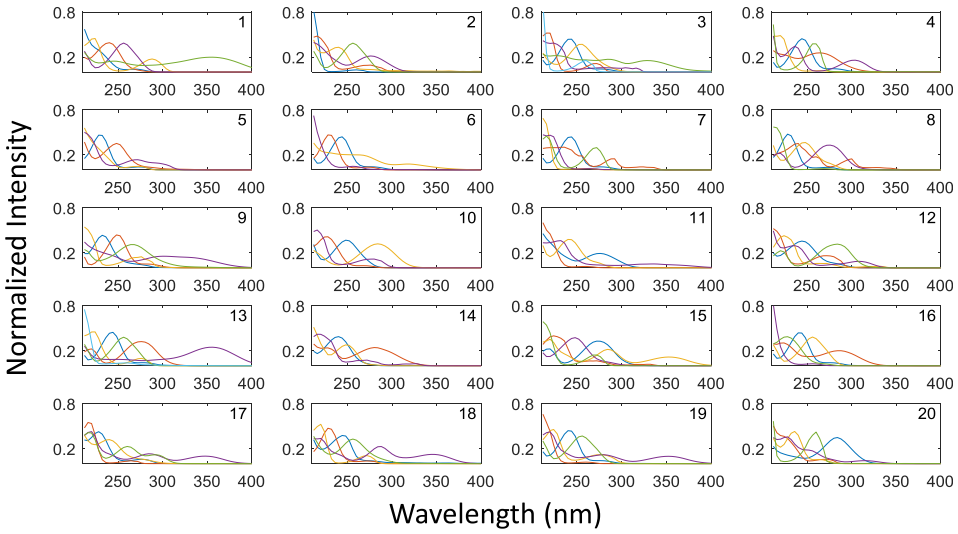

Fig. 2. Spectra of compounds in each of the 20 groups assembled based on spectra alone. Grouping was performed with threshold SEL = 0.2 and iterative optimization was performed. Numbers inside the plot area indicate group number.

normalized to unit length prior to grouping.

In order to initialize the first two groups, two compounds with identical retention factors (or two compounds with the most similar retention factors, if none have identical retention factors) were placed in two separate groups. The remaining 95 compounds were then assigned to groups as described in the Strategy section above. The threshold was selected as 0.95 for this strategy because the addition of chromatographic information greatly increases the selectivity of the analysis [47]. This strategy is dependent on the order in which the compounds are assigned; however, each of the possible results are essentially equivalent, as SEL will always be greater than the threshold and thus will be able to be differentiated by curve resolution methods. To standardize the results, it is suggested to sort compounds based on k before assigning groups. Depending on the final goal of the analysis, the number of analytes per group may be minimized towards a target number. The number of compounds per group can be limited by not allowing analytes to be placed into an existing group that has more than the target number of analytes.

The grouping was performed without limiting the number of compounds per group. The iterative optimization step was performed as it was found to more evenly distribute compounds across the groups. Table 3 lists the groups created at a threshold of SEL = 0.95 and Fig. 3 shows the chromatographic profiles for each compound in each group. While most compounds are chromatographically resolved, some compounds do overlap in the chromatographic mode. The benefit of the SEL metric, however, is that it incorporates the additional information contained in the spectral dimension. Fig. 4 shows the chromatographic and spectral profiles of each compound in the 2nd group. It can be seen that while Bz1 and F1Bp are significantly overlapped chromatographically, the spectral correlation between Bz1 and FlBp is 0.627. This dissimilarity allows the chromatographic overlap to be easily overcome

Table 1
Compounds grouped based on spectra only at threshold of SEL = 0.2*.

Group 1	2Hip	4Hip	4MPh	PB4	Pir	
	0.397	0.565	0.456	0.683	0.787	
Group 2	2Bz	2NAA	2PhP	DCFen	PB3	
	0.637	0.313	0.518	0.333	0.607	
Group 3	AP6	Ans	BzP	ClQu	Sul	TFT
	0.420	0.348	0.399	0.291	0.511	0.698
Group 4	AP2	BzTan	Ibp	ScA	dFBz	
	0.561	0.489	0.601	0.508	0.673	
Group 5	BA	BzI	FnP	hInd		
	0.524	0.641	0.276	0.328		
Group 6	AP4	BZ1	IndM	PhAc		
	0.544	0.603	0.467	0.657		
Group 7	AP5	Car	ClBz	Lox	dClPy	
	0.379	0.264	0.498	0.373	0.553	
Group 8	BZ4	CarP	FlBp	HP4	dClBz	
	0.462	0.285	0.398	0.733	0.750	
Group 9	BZPh	BzA	Naph	Nim	NtBz	
	0.547	0.506	0.570	0.543	0.460	
Group 10	BPh	BzAd	FenB	InAA		
	0.721	0.493	0.710	0.464		
Group 11	ACFen	DPol	FBPh	mnQu		
	0.445	0.482	0.469	0.356		
Group 12	AcTP	Ind	Nab	Sal	hBzA	
	0.548	0.467	0.429	0.352	0.678	
Group 13	AP3	HP2	HPhAA	Mel	PB1	i4Bz
	0.540	0.546	0.491	0.857	0.430	0.560
Group 14	AcTan	IndP	Inde	mNap		
	0.537	0.637	0.451	0.493		
Group 15	HP3	NaP	Pan	PhAct	Phenol	
	0.519	0.425	0.428	0.610	0.578	
Group 16	AP7	OxA	PB2	PhAA	nBzAd	
	0.396	0.523	0.544	0.650	0.353	
Group 17	Asp	BzSA	PhBtz	Tof	VA	
	0.332	0.456	0.345	0.358	0.356	
Group 18	BZ2	BZ3	BzO2	FFen	Ket	
	0.368	0.478	0.551	0.602	0.420	
Group 19	AP8	DPM	EtD	MFen	hBA	
	0.613	0.547	0.363	0.332	0.634	
Group 20	DPA	DPE	DPS	DifS	FBz	
	0.772	0.344	0.347	0.223	0.715	

*Values for each compound are the SEL value relative to group.

Table 2
LSS parameters for simulating chromatographic profiles.

Parameter	Value
N	1000
S	10
$\phi_{initial}$	0.20
$\Delta\phi$	0.75
t_M	1 min
t_G	35 min

Table 3
Compounds grouped based on spectra and chromatograms at a threshold of SEL = 0.95.

Group 1	AP3	BPh	PB1	Sal	dClBz	dClPy	mNap					
Group 2	BZ1	DCFen	FBPh	FlBp	Nim	PhAct	i4Bz					
Group 3	AP5	FBz	NaP	Pir	TFT	Tof						
Group 4	Ans	BZ3	DPA	DPS	DPol	PB2						
Group 5	2NAA	BZ4	FenB	Ket	Naph							
Group 6	AP4	AP8	AcTP	AcTan	BzSA	BzTan	InAA	Ind	IndM	Pan	ScA	hBA	hBzA
Group 7	2Bz	ACFen	AP6	BA	BzA	BzI	DPE	MFen	VA	hInd			
Group 8	BZPh	IndP	Mel	Nab	PB4								
Group 9	2Hip	4Hip	AP2	ClBz	FFen	HP3	Ibp	Lox	NtBz	OxA	Phenol	dFBz	
Group 10	Asp	BZ2	FnP	Inde	PB3	PhAc	PhBtz						
Group 11	4MPh	AP7	BzO2	BzP	Car	ClQu	HP4						
Group 12	2PhP	BzAd	CarP	DPM	DifS	EtD	HP2	HPhAA	PhAA	Sul	mnQu	nBzAd	

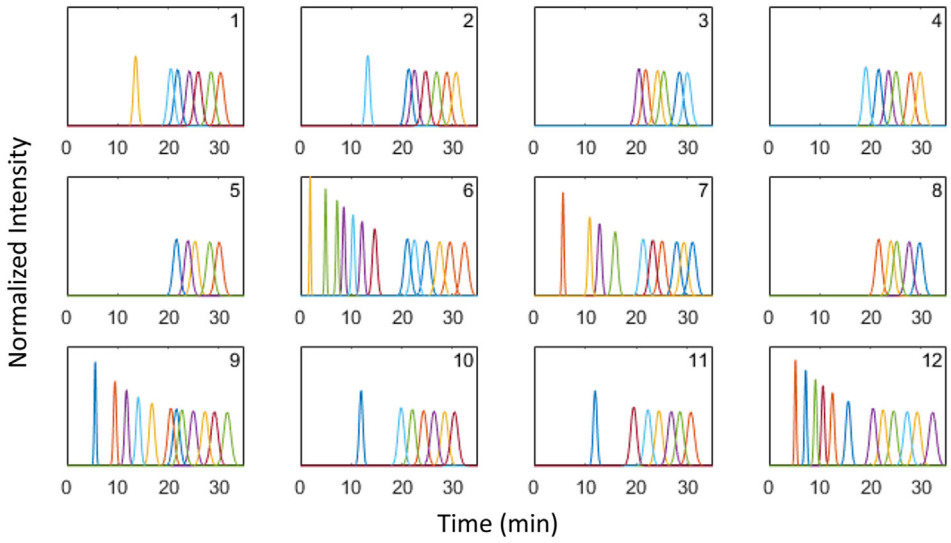

Fig. 3. Chromatographic profiles of compounds in each group designed using both chromatographic and spectral information, and threshold of SEL = 0.95. Numbers indicate group number.

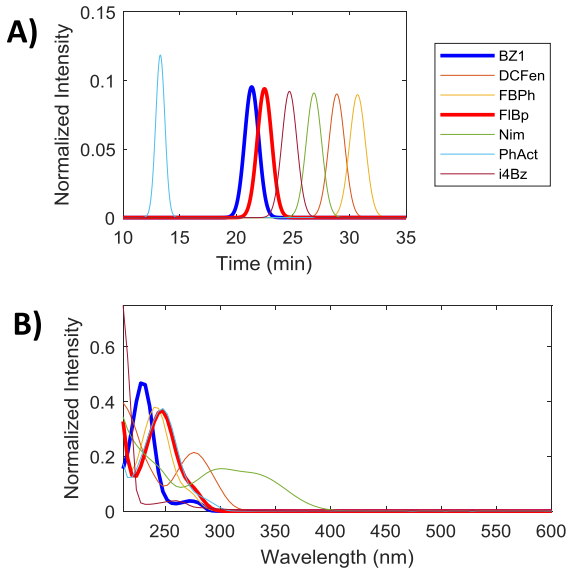

Fig. 4. Chromatographic (A) profiles of each compound in the 2nd group at a threshold SEL = 0.95 and the corresponding spectra (B) of the chromatographically overlapped compounds.

with methods such as MCR-ALS.

In cases where many compounds have very different spectral and chromatographic profiles, it may be desired to limit the number of compounds in each group. This is accomplished during the grouping algorithm as shown in Fig. 1 by not allowing the addition of more compounds to a group past the preset limit. For the 97 compounds analyzed previously, a limit of eight compounds per group led to 15 groups. Because the iterative optimization step chooses the optimal group in which to add each compound, this step cannot be included when a specific number of compounds per group is desired, as iterative optimization will often result in greater than the preset limit of compounds for one or more groups.

We also examined the dependence of the average number of compounds/group on the selection of the threshold SEL value. Fig. 5A shows this dependence for the case where only spectral information is considered. For a low SEL threshold value of 0.1, an average of 8.2 ± 0.4 compounds per group is needed, whereas for a SEL threshold of 0.5, an average of 3.3 ± 0.1 compounds per group is selected. The demands of the particular screening study should guide the selection of an appropriate SEL threshold. For example, if the study requires only approximate retention times, a much lower SEL can be used, than if exact chromatographic peak shape information is required. Fig. 5B shows the average number of compounds/group for the case where approximate chromatographic information is available, as well as the spectral information. For a

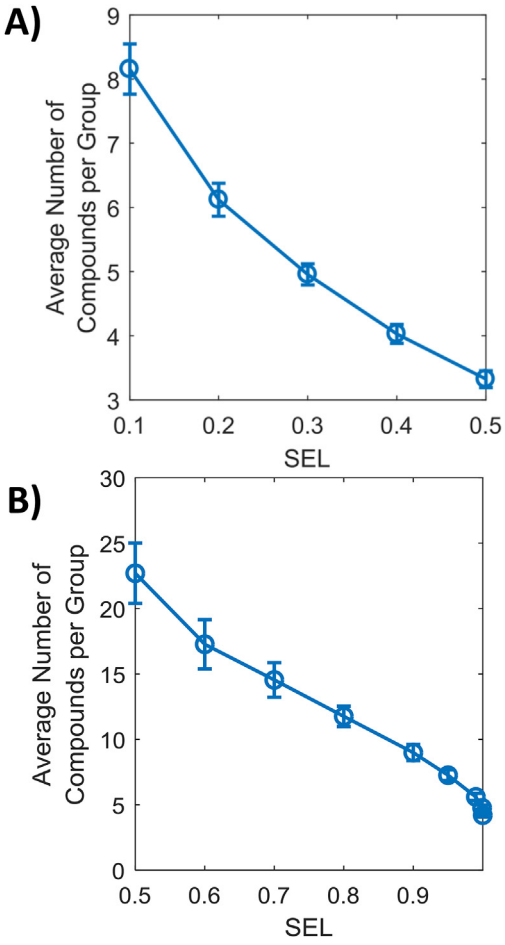

Fig. 5. The average number of compounds per group as a function of the selected SEL threshold with (A) spectra only and (B) chromatographic information included. These data represent the average number of compounds per group across 25 runs with the compound starting orders randomly assigned. Error bars represent the standard deviation of the 25 runs. The last three points in panel (B) are SEL = 0.99, 0.999, and 0.9999, respectively.

threshold SEL value of 0.5, the average number of compounds/group is 22.7 ± 2.3, which clearly offers a huge improvement in screening efficiency as compared to a one compound/injection strategy. Even when a SEL threshold of 0.9999 is used, an average of 4.2 ± 0.1 compounds per group is suggested by this analysis. These values will of course be strongly influenced by the overall spectral similarity within the compound set, as well as the selected column efficiency (peak width).

Finally, in some cases, the retention factor estimates may not be accurate, such that some mixtures may show peaks where unambiguous identification of peaks is not possible. In this case, simple spiking experiments can be done to confirm the identity of each peak.

6. Conclusions

The RDM strategy described in this paper enables rational design of mixtures containing a distinct number of target analytes; this approach may be useful for a number of potential applications. One possible application is when parameterizing retention for *in-silico* optimization studies. Rationally designed mixtures support reliable peak tracking when used in conjunction with curve resolution algorithms such as MCR-ALS. By using a multivariate selectivity metric, we were successfully able to group 97 compounds into only 12 groups, greatly diminishing the time needed to carry out the experiments needed to obtain retention parameters for a set of probe solutes of interest. Even when assigning solutes based on UV absorption spectra alone, just 20 groups were required to create mixtures that would be able to be separated and the peaks identified via curve resolution techniques. At the present time, some of us are using this strategy to characterize solute retention for the development of new column selectivity metrics.

Another application where RDM would be useful is in the design of calibration standards for analysis of complex mixtures. When using curve resolution strategies for these analyses, it is desired to carry out multi-set experiments to improve the precision of the calibration parameters [36,48]. Designing different calibration mixtures where pure variables are present (*i.e.,* low chromatographic and spectral overlap) allows for more robust calibration in these cases, and the RDM strategy should be directly applicable in these cases.

Acknowledgements

The authors acknowledge funding from National Science Foundation grants (CHE-1507332 and CHE-1508159). We also want to thank Professor Stephen Weber of the University of Pittsburgh for providing the probe compounds used in this study.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.acax.2019.100010.

References

[1] J.-L. Wolfender, G. Marti, A. Thomas, S. Bertrand, Current approaches and challenges for the metabolite profiling of complex natural extracts, J. Chromatogr. A 1382 (2015) 136–164, https://doi.org/10.1016/j.chroma.2014.10.091.
[2] F. Pellati, F. Epifano, N. Contaldo, G. Orlandini, L. Cavicchi, S. Genovese, D. Bertelli, S. Benvenuti, M. Curini, A. Bertaccini, M.G. Bellardi, Chromatographic methods for metabolite profiling of virus- and phytoplasma-infected plants of Echinacea purpurea, J. Agric. Food Chem. 59 (2011) 10425–10434, https://doi.org/10.1021/jf2025677.
[3] M. Gómez-Romero, A. Segura-Carretero, A. Fernández-Gutiérrez, Metabolite profiling and quantification of phenolic compounds in methanol extracts of tomato fruit, Phytochemistry 71 (2010) 1848–1864, https://doi.org/10.1016/j.phytochem.2010.08.002.
[4] R. Wehrens, E. Carvalho, P.D. Fraser, Metabolite Profiling in LC–DAD Using Multivariate Curve Resolution: the Alsace Package for R, Metabolomics, 2014, https://doi.org/10.1007/s11306-014-0683-5.
[5] M.J. Fredriksson, P. Petersson, B.-O. Axelsson, D. Bylund, Combined use of algorithms for peak picking, peak tracking and retention modelling to optimize the chromatographic conditions for liquid chromatography-mass spectrometry analysis of fluocinolone acetonide and its degradation products, Anal. Chim. Acta 704 (2011) 180–188, https://doi.org/10.1016/j.aca.2011.07.047.
[6] B.A. Olsen, B.C. Castle, D.P. Myers, Advances in HPLC technology for the determination of drug impurities, TrAC Trends Anal. Chem. 25 (2006) 796–805, https://doi.org/10.1016/j.trac.2006.06.005.
[7] W. Li, Substitute technology for reference substances in the analysis of impurities in cefonicid for injection with HPLC using a diode array detector, J. AOAC Int. 94 (2011) 531–536. http://www.ncbi.nlm.nih.gov/pubmed/21563687.
[8] L.R. Snyder, J.W. Dolan, D.C. Lommen, Drylab computer simulation for high-performance liquid chromatographic method development I. Isocratic elution, J. Chromatogr. A 485 (1989) 65–89, https://doi.org/10.1016/S0021-9673(01)89133-0.
[9] J.W. Dolan, D.C. Lommen, L.R. Snyder, DryLab computer simulation for high-performance liquid chromatographic method development. II. Gradient elution, J. Chromatogr. 485 (1989) 91–112, https://doi.org/10.1016/S0021-9673(01)89134-2.
[10] I. Molnar, Computerized design of separation strategies by reversed-phase liquid chromatography: development of DryLab software, J. Chromatogr. A 965 (2002) 175–194, https://doi.org/10.1016/S0021-9673(02)00731-8.
[11] L.N. Jeong, R. Sajulga, S.G. Forte, D.R. Stoll, S.C. Rutan, Simulation of elution profiles in liquid chromatography—I: gradient elution conditions, and with mismatched injection and mobile phase solvents, J. Chromatogr. A 1457

(2016) 41–49, https://doi.org/10.1016/j.chroma.2016.06.016.
[12] L.R. Snyder, J.W. Dolan, High-Performance Gradient Elution: the Practical Application of the Linear-Solvent-Strength Model, Wiley, New York, NY, 2006.
[13] U.D. Neue, H.-J. Kuss, Improved reversed-phase gradient retention modeling, J. Chromatogr. A 1217 (2010) 3794–3803, https://doi.org/10.1016/j.chroma.2010.04.023.
[14] L.R. Snyder, J.W. Dolan, D.H. Marchand, P.W. Carr, The hydrophobic-subtraction model of reversed-phase column selectivity, in: Adv. Chromatogr., vol 50, CRC Press, Boca Raton, FL, 2012, pp. 297–376.
[15] J.W. Dolan, L.R. Snyder, The hydrophobic-subtraction model for reversed-phase liquid chromatography: a reprise, LCGC North Am. 34 (2016) 730–741.
[16] N.S. Wilson, M.D. Nelson, J.W. Dolan, L.R. Snyder, R.G. Wolcott, P.W. Carr, Column selectivity in reversed-phase liquid chromatography. I. A general quantitative relationship, J. Chromatogr. A 961 (2002) 171–193, https://doi.org/10.1016/S0021-9673(02)00659-3.
[17] L.R. Snyder, J.W. Dolan, P.W. Carr, The hydrophobic subtraction model of reversed-phase column selectivity, J. Chromatogr. A 1060 (2004) 77–116, https://doi.org/10.1016/j.chroma.2004.08.121.
[18] M.R. Euerby, P. Petersson, W. Campbell, W. Roe, Chromatographic classification and comparison of commercially available reversed-phase liquid chromatographic columns containing phenyl moieties using principal component analysis, J. Chromatogr. 1154 (2007) 138–151, https://doi.org/10.1016/j.chroma.2007.03.119.
[19] L. Lopez, S.C. Rutan, Comparison of methods for characterization of reversed-phase liquid chromatographic selectivity, J. Chromatogr. A 965 (2002) 301–314, https://doi.org/10.1016/S0021-9673(02)00002-X.
[20] T. Németh, E. Haghedooren, B. Noszál, J. Hoogmartens, E. Adams, Three methods to characterize reversed phase liquid chromatographic columns applied to pharmaceutical separations, J. Chemom. 22 (2008) 178–185, https://doi.org/10.1002/cem.1108.
[21] M.J. Fredriksson, P. Petersson, B.-O. Axelsson, D. Bylund, A component tracking algorithm for accelerated and improved liquid chromatography-mass spectrometry method development, J. Chromatogr. A 1217 (2010) 8195–8204, https://doi.org/10.1016/j.chroma.2010.10.083.
[22] J.K. Strasters, H.A.H. Billiet, L. de Galan, B.G.M. Vandeginste, Strategy for peak tracking in liquid chromatography on the basis of a multivariate analysis of spectral data, J. Chromatogr. A 499 (1990) 499–522, https://doi.org/10.1016/S0021-9673(00)96996-6.
[23] I. Molnar, R. Boysen, P. Jekow, Peak tracking in high-performance liquid chromatography based on normalized band areas, J. Chromatogr. A 485 (1989) 569–579, https://doi.org/10.1016/S0021-9673(01)89163-9.
[24] P. V van Zomeren, a Hoogvorst, P.M.J. Coenegracht, G.J. de Jong, Optimisation of high-performance liquid chromatography with diode array detection using an automatic peak tracking procedure based on augmented iterative target transformation factor analysis, Analyst 129 (2004) 241–248, https://doi.org/10.1039/b313165c.
[25] G. Xue, A.D. Bendick, R. Chen, S.S. Sekulic, Automated peak tracking for comprehensive impurity profiling in orthogonal liquid chromatographic separation using mass spectrometric detection, J. Chromatogr. A 1050 (2004) 159–171, https://doi.org/10.1016/j.chroma.2004.08.030.
[26] C. Tistaert, Y. Vander Heyden, Bilinear decomposition based alignment of chromatographic profiles, Anal. Chem. 84 (2012) 5653–5660, https://doi.org/10.1021/ac300735a.
[27] D.W. Cook, Chemometric Curve Resolution for Quantitative Liquid Chromatographic Analysis, Ph.D. Dissertation, Virginia Commonwealth University, 2016.
[28] R. Tauler, Multivariate curve resolution applied to second order data, Chemom. Intell. Lab. Syst. 30 (1995) 133–146, https://doi.org/10.1016/0169-7439(95)00047-X.
[29] S.C. Rutan, A. de Juan, R. Tauler, Introduction to multivariate curve resolution, in: Compr. Chemom., Elsevier, Oxford, 2009, pp. 249–259.
[30] A. de Juan, J. Jaumot, R. Tauler, Multivariate Curve Resolution (MCR). Solving the mixture analysis problem, Anal. Methods 6 (2014) 4964, https://doi.org/10.1039/c4ay00571f.
[31] R. Bro, PARAFAC. Tutorial and applications, Chemom. Intell. Lab. Syst. 38 (1997) 149–171, https://doi.org/10.1016/S0169-7439(97)00032-4.
[32] A. Smilde, R. Bro, P. Geladi, Multi-way Analysis: Applications in the Chemical Sciences, Wiley, New York, 2004.
[33] A.C. Olivieri, Analytical figures of merit: from univariate to multiway calibration, Chem. Rev. 114 (2014) 5358–5378, https://doi.org/10.1021/cr400455s.
[34] M.T. Cantwell, S.E.G. Porter, S.C. Rutan, Evaluation of the multivariate selectivity of multi-way liquid chromatography methods, J. Chemom. 21 (2007) 335–345, https://doi.org/10.1002/cem.1055.
[35] N.J. Messick, J.H. Kalivas, P.M. Lang, Selectivity and related measures for n th-order data, Anal. Chem. 68 (1996) 1572–1579, https://doi.org/10.1021/ac951212v.
[36] A.C. Olivieri, R. Tauler, The effect of data matrix augmentation and constraints in extended multivariate curve resolution-alternating least squares, J. Chemom. 31 (2017) e2875, https://doi.org/10.1002/cem.2875.
[37] A. Lorber, Error propagation and figures of merit for quantification by solving matrix equations, Anal. Chem. 58 (1986) 1167–1172, https://doi.org/10.1021/ac00297a042.
[38] A. Lorber, K. Faber, B.R. Kowalski, Net analyte signal calculation in multivariate calibration, Anal. Chem. 69 (1997) 1620–1626, https://doi.org/10.1021/ac960862b.
[39] A.C. Olivieri, Computing sensitivity and selectivity in parallel factor Analysis and related multiway techniques: the need for further developments in net analyte signal theory, Anal. Chem. 77 (2005) 4936–4946, https://doi.org/10.1021/ac050146m.
[40] K.G. Kraiczek, G.P. Rozing, R. Zengerle, Relation between chromatographic resolution and signal-to-noise ratio in spectrophotometric HPLC detection, Anal. Chem. 85 (2013) 4829–4835, https://doi.org/10.1021/ac4004387.
[41] M.C. Bauza, G.A. Ibañez, R. Tauler, A.C. Olivieri, Sensitivity equation for quantitative analysis with multivariate curve resolution-alternating least-squares: theoretical and experimental approach, Anal. Chem. 84 (2012) 8697–8706, https://doi.org/10.1021/ac3019284.
[42] ACD/LC & GC Simulator—Model and Optimize LC and GC Separation Methods, 2015. http://www.acdlabs.com/products/com_iden/meth_dev/lc_sim/. (Accessed 7 January 2015).
[43] L. Wang, J. Zheng, X. Gong, R. Hartman, V. Antonucci, Efficient HPLC method development using structure-based database search, physico-chemical prediction and chromatographic simulation, J. Pharm. Biomed. Anal. 104 (2015) 49–54, https://doi.org/10.1016/j.jpba.2014.10.032.
[44] L.N. Jeong, R. Sajulga, S.G. Forte, D.R. Stoll, S.C. Rutan, Simulation of elution profiles in liquid chromatography - I: gradient elution conditions, and with mismatched injection and mobile phase solvents, J. Chromatogr. A 1457 (2016) 41–49, https://doi.org/10.1016/j.chroma.2016.06.016.
[45] ChromSword© Offline http://www.chromsword.com/offline/ (accessed March 9, 2019)
[46] K.P. Xiao, Y. Xiong, F.Z. Liu, A.M. Rustum, Efficient method development strategy for challenging separation of pharmaceutical molecules using advanced chromatographic Technologies, J. Chromatogr. A 1163 (2007) 145–156, https://doi.org/10.1016/j.chroma.2007.06.027.
[47] J.M. Davis, S.C. Rutan, P.W. Carr, Relationship between selectivity and average resolution in comprehensive two-dimensional separations with spectroscopic detection, J. Chromatogr. A 1218 (2011) 5819–5828, https://doi.org/10.1016/j.chroma.2011.06.086.
[48] R. Tauler, M. Maeder, A. de Juan, Multiset data analysis: extended multivariate curve resolution, in: S.D. Brown, R. Tauler, B. Walczak (Eds.), Compr. Chemom., Elsevier, Oxford, 2009, pp. 473–501.