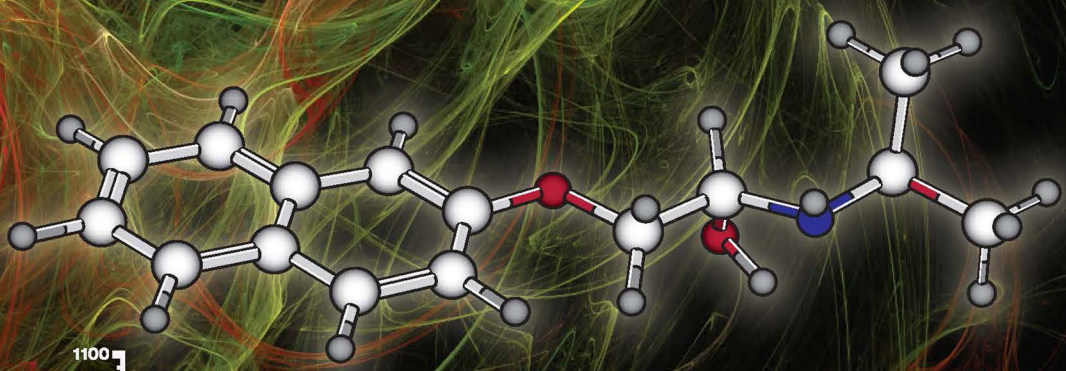
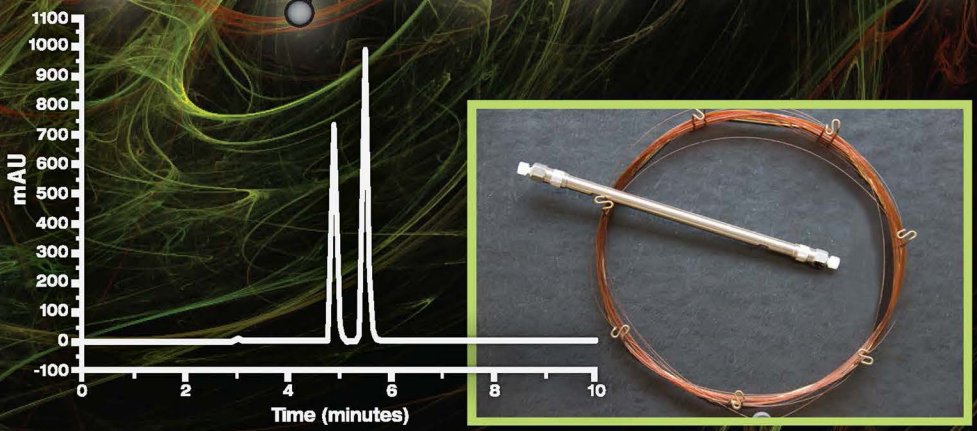
Volume 59

Advances in CHROMATOGRAPHY





EDITORS
Nelu Grinberg • Peter W. Carr

CRC Press
Taylor & Francis Group

Advances in Chromatography

For six decades, scientists and researchers have relied on the Advances in Chromatography series for the most up-to-date information on a wide range of developments in chromatographic methods and applications. The clear presentation of topics and vivid illustrations for which this series has become known make the material accessible and engaging to analytical, biochemical, organic, polymer, and pharmaceutical chemists at all levels of technical skill.

- Describes the thermodynamics and kinetics underlying hydrophobic interaction chromatography of proteins.
- Outlines use of a kinetic model in the predictive modeling of evaporation processes that eliminates the need to know the composition and identity of the chemical constituents in the sample.
- Explores building and employing QSRR models in cyclodextrin modified high-performance liquid chromatography (HPLC).
- Reviews chemometric methods commonly paired with comprehensive 2D separations and key instrumental and preprocessing considerations.

Advances in Chromatography

Volume 59

Edited by Nelu Grinberg and Peter W. Carr

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

First edition published 2023
by CRC Press
6000 Broken Sound Parkway NW, Suite 300, Boca Raton, FL 33487-2742

and by CRC Press
4 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

CRC Press is an imprint of Taylor & Francis Group, LLC

ISBN: 978-1-032-36027-0 (hbk)
ISBN: 978-1-032-36063-8 (pbk)
ISBN: 978-1-003-33008-0 (ebk)

DOI: 10.1201/9781003330080

Typeset in Times
by Apex CoVantage, LLC

Contents

1 Kinetic and Thermodynamic Aspects of Hydrophobic Interaction Chromatography

Dorota Antos, Wojciech Piątkowski

CONTENTS

1.1 INTRODUCTION

Hydrophobic Interaction Chromatography (HIC) is widely used for separation of proteins in both small- and large-scale applications. It is often employed in downstream processing for intermediate purification and in the polishing stage to complement

DOI: 10.1201/9781003330080-1

other chromatography techniques, such as ion-exchange chromatography and affinity chromatography.

The advantage of HIC is that it can be realized under mild conditions with respect to pH, temperature, and solvent environment. HIC is an entropically driven process involving interactions between hydrophobic adsorbent surfaces and hydrophobic fragments of protein molecules. The proteins are separated based on differences in their hydrophobicity. Protein binding is promoted by the presence of a kosmotropic salt, such as ammonium sulfate (AS), which increases the surface tension of the liquid solution and destroys the structure of the water shell hydrating hydrophobic patches on the protein surface [1–5]. The protein release from the adsorbed phase and its elution are imposed by a decrease in the salt concentration, which causes gradual weakening of hydrophobic interactions between the adsorbent surface and the proteins.

Since the salt concentration is a primary process variable used to alter the separation efficiency, HIC is easier to establish compared with ion-exchange chromatography, where protein binding and elution are also affected by pH. This also holds true for comparison with multimodal chromatography, whose efficiency is additionally altered by the type of ligand and its density. This requires additional effort to determine the operating window and optimize the process performance.

HIC is particularly efficient for isolation of targeted therapeutic protein from its aggregated forms; for example, it is applied in the purification of monoclonal antibodies (mAbs) for removal of aggregates [6–8]. It is also efficient for purification of plasmids and removal of bound endotoxin [9–11]. Recently, HIC has gained new areas of interest, such as the manufacture of antibody–drug conjugates (ADCs) and co-formulated mAbs. ADCs, which are destined for targeted therapies, consist of small-molecule drugs conjugated to large-molecule highly hydrophobic mAbs. The components with different degrees of conjugation differ in hydrophobicity; therefore, the HIC step can be employed for their efficient separation as well as adjustment of the drug-to-antibody ratio [12–15]. Co-formulated mAbs are composed of two mAbs destined for synergistic targeting to multiple sites of action. HIC has been reported to be effective in the quantitation of individual mAbs in co-formulated mAbs [16,17].

The development of a robust separation process cannot be accomplished without understanding the mechanism of protein binding and elution (i.e., underlying thermodynamic and kinetic effects) along with their dependencies on the operating variables. The thermodynamic nature of HIC has been investigated in several studies in which mechanistic isotherms were developed [18,19] and the influence of salt on HIC retention, capacity factors, and water release were quantified [3,20–26].

Nevertheless, the complexity of protein behavior and the specificity of their adsorption mean that molecular dynamic models cannot provide general conclusions concerning patterns of protein binding and elution. The adsorption behavior typically assigned to “hydrophobic interactions” arises from a complex combination of long-range non-hydrophobic and short-range hydrophobic interactions [27,28]. The process thermodynamics become even more complex for multicomponent mixtures when they involve either competitive or synergistic adsorption effects or a combination of them [29]. This particularly holds true for adsorption of mixtures of proteins of quite varied sizes, which induces size exclusion effects. In such cases, to quantify the adsorption behavior, more advanced models that account for these phenomena are required.

Furthermore, the dynamics of protein chromatography are often determined by kinetic effects arising from mass transfer limitations and slow rates of process occurring at the solid–liquid interface, including steric hindrances in porous adsorbents [30–32], slow binding, or conformational changes upon adsorption [33–36]. The latter is a cause of protein unfolding and aggregation, which manifests in elution of multiple peaks of different retentions (e.g., [35,36]). The occurrence of that phenomenon is detrimental for the separation efficiency and often causes strongly hydrophobic HIC resins to be reluctantly used for processing structurally unstable proteins, despite the high selectivity of the separation [37].

Furthermore, a high concentration of kosmotropic salt used for protein binding limits protein solubility in the loading buffers. Therefore, the loading concentration of the protein is often low, which reduces process throughput.

Nevertheless, by proper selection of the operating conditions, the separation can be realized with high yield and purity, while undesirable effects are avoided and the biological activity of the protein is preserved. It cannot be done without recognizing the pitfalls and understanding their origins.

In this chapter, we describe the different thermodynamic and kinetic effects that can accompany HIC separation and cause misinterpretation of the retention data and failure in the process design, including protein conformational changes, multicomponent adsorption, solubility limitations, and thermal heterogeneity of the column in thermally mediated separations. We also provide an elucidation of the mechanisms underlying these effects or the hypothesis of their occurrence.

1.2 PROTEIN BEHAVIOR IN THE ADSORBED PHASE

1.2.1 Conformational Changes—Detection of the Phenomenon

A prerequisite for efficient design of HIC separation is identification and quantification of the process conditions that potentially trigger yield losses due to protein unfolding upon adsorption. To detect this phenomenon, various analytical methods have been employed: circular dichroism, fluorescence, infrared spectroscopy [33,38–41], isothermal titration calorimetry [37,42,43], and hydrogen exchange (e.g., [44–48]). These methods provided insight into the mechanism of protein conformational changes at the absorbent surface as well as a basis for the formulation of mechanistic models describing that phenomenon. However, they involve complex and time-consuming measurement procedures that are not suitable for high-throughput screening of the process conditions in downstream processing. For fast identification of the destabilization of proteins on HIC resins, nano-Differential Scanning Fluorimetry (nanoDSF) can be exploited [49]. In this method, the stability of the protein is determined by the so-called melting temperature, which corresponds to the midpoint of the transition from folded to unfolded forms of the protein or its specific domains. To determine the unfolding transition points, the shifts of intrinsic tryptophan fluorescence at emission wavelengths of 330 nm and 350 nm over the course of temperature gradients are recorded [50]. A protein-specific low critical melting temperature can be identified and used as an indicator of destabilization of protein structure in both liquid and adsorbed phases [49,51]. This approach allows

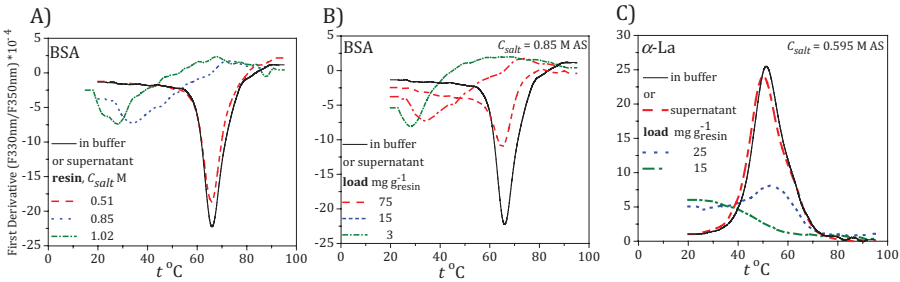


FIGURE 1.1 Changes in the melting temperature of the protein in the liquid solution and adsorbed on a Butyl Sepharose resin recorded using nanoDSF. A) BSA at different salt concentrations in the loading buffer, C_{salt}, at constant protein load on the resin, i.e., 3 mg g^{-1}_{resin}, B) BSA at different protein loads at constant C_{salt}, C) a-La at different protein loads.

From: R. Muca, M. Żurawski, W. Piątkowski, D. Antos, J. Chromatogr. A, 1492 (2017), 79–88 [49].

simultaneous detection of melting temperatures for a large number of small samples; therefore, it can be applied in the development stage of the chromatographic process. Furthermore, this method does not require a dye for protein staining. This is particularly important for HIC resins, which, due to their hydrophobic properties, are prone to interacting with dyes. Figure 1.1 illustrates the changes in the fluorescence ratio at the wavelengths of 330 nm and 350 nm versus temperature for bovine serum albumin (BSA) and α-Lactalbumin (α-La) dissolved in liquid phase as well as adsorbed on an HIC resin under different conditions. Both BSA and α-La were reported to be prone to unfolding upon adsorption on HIC media [3,34,37,43,47,52]. The maximum of the curves corresponds to the melting temperature of the proteins. It can be observed that an increase in salt concentrations, which enhances the strength of hydrophobic interactions, causes a significant reduction in the melting temperature (Figure 1.1A). The opposite effect is induced by increasing loading concentrations, which causes stabilization of the protein in the adsorbed phase (Figures 1.1B and 1.1C). For both proteins, the low critical temperature was about 40°C, and the loading conditions that corresponded to the melting temperature below that value triggered multi-peak elution in the chromatographic process [49].

1.2.2 Conformational Changes—Mechanistic Models

To quantify the phenomenon of protein unfolding on HIC media, several mechanistic models have been developed. Xiao et al. [52] and Deitcher et al. [53] used a four-state model that assumed reversible conformation change and reversible adsorption of proteins. Three-state models were used by Lundström [54] and Heimer et al. [34]. A three-state reversible unfolding model was also used to reproduce the chromatographic band profiles of proteins on HIC media under linear or non-linear isotherm conditions [29,31,49,55–58].

A pictorial representation of these models is shown in Figure 1.2. In the four-state model, the protein can unfold and refold in both the liquid and adsorbed phases

through a sequence of reversible reaction steps (steps 1–4, Figure 1.2A). The ratios of the rate constants of forward and backward reactions determine the preferable paths of the process and the concentration of different protein forms in the liquid and adsorbed phases. However, the unfolded form in the liquid phase is often susceptible to aggregate, which is illustrated by two additional states of the adsorption mechanism assigned to the aggregate formation and its subsequent adsorption (steps 5, 6, Figure 1.2B). It may induce the previously mentioned multiple-peak elution, where different proteins form (i.e., native, refolded, and aggregated), elute with different retention times. When the desorption rate of the unfolded protein, which is usually strongly bound, is infinitely slow, the four-state model converts to the three-state model. Figure 1.2C illustrates a three-state model in which unfolding of the protein is assumed to occur in a sequence of intermediate stages of binding the protein to an active site, followed by its anchoring and spreading due to interactions of the adsorbed molecule with neighboring adsorption sites [36,49,54]. In this model, the unfolded protein can be desorbed only through the backward reaction in the sequence of the unfolding–refolding and adsorption–desorption processes. When the refolding rate is infinitely slow, the unfolding reaction path becomes irreversible.

As mentioned in Section 1.2.1, the unfolding phenomenon is expected to diminish with increasing protein loading on the resin. This stems from the "crowding" effect, where the presence of molecules of proteins on the adsorbent surface reduces their accessibility to adsorption sites and, therefore, their ability to spread and unfold [36,49,59]. A representation of this mechanism is shown in Figure 1.2D.

The reaction scheme for the simplest free-state mechanism (Figure 1.2C) can be expressed by Eqs. 1.1 and 1.2. In this scheme, the protein i, P_i, is reversibly bound to a single adsorption site, S^*, forming a surface complex $\overline{P}_{n,i}\,S^*$ on bare adsorbent surface according to the second-order Langmuir-type reaction kinetics with adsorption and desorption rate constants $k_{a,i}$ and $k_{d,i}$ [29,36,49]:

$$P_i + S^* \underset{k_{d,i}}{\overset{k_{a,i}}{\rightleftharpoons}} \overline{P}_{n,i}S^* \tag{1.1}$$

where: $\overline{P}_{n,i}$ is the adsorbed protein i in the native form.

The second step (2) is induced by the interaction of the protein with a neighboring active site. To describe a two-site interaction (Figure 1.1C), a second-order reaction scheme can be used:

$$\overline{P}_{n,i}S^* + S^* \underset{k_{f,i}}{\overset{k_{u,i}}{\rightleftharpoons}} \overline{P}_{u,i}S_2^* \tag{1.2}$$

where: $k_{u,i}$, $k_{f,i}$ are the kinetic coefficients of protein unfolding and folding, $\overline{P}_{u,i}$, denotes the protein in the adsorbed phase in the unfolded form.

1.2.3 Cluster Formation

At a high protein concentration in the adsorbed phase, another surface reaction is suggested to be active when the adsorbed proteins provide additional interaction sites [29, 60–62]. To describe this phenomenon, Chatelier, and Minton [60,61]

developed kinetic models of positive cooperative protein adsorption. Positive cooperative adsorption, also termed synergistic adsorption, can be induced by attractive intermolecular interactions that lead to the formation of protein clusters on the adsorbent surface [60–62]. Positive cooperative adsorption is assumed to occur due to multilayer and preferred adsorption. In the case of multilayer adsorption, the molecules of proteins present in the adsorbed phase provide additional adsorption sites on their own molecular surface, whereas in the case of preferred adsorption, the protein molecules already present in the adsorbed phase promote or activate adsorption of other molecules on adjacent adsorption sites on the adsorbent surface without direct contact. The positive cooperative adsorption can be described by a simplified reaction scheme with adsorption and desorption rate constants k_{ac} and k_{dc}, as illustrated in Figure 1.2E [29]. When the desorption rate is very slow, the positive cooperative adsorption leads to irreversible formation of aggregates in the adsorbed phase.

A mechanism of the interactions between the *i*-th protein and adsorbed molecules of proteins present in the multicomponent solution can be described as follows:

$$\mathrm{P}_{n,i} + \left(\overline{\mathrm{P}}_k\right)_j \underset{k_{dc,k,ij}}{\overset{k_{ac,k,ij}}{\rightleftarrows}} \overline{\mathrm{P}}_{c,k,ij} \quad i = 1..N, j = 1..N, k = n, u \tag{1.3}$$

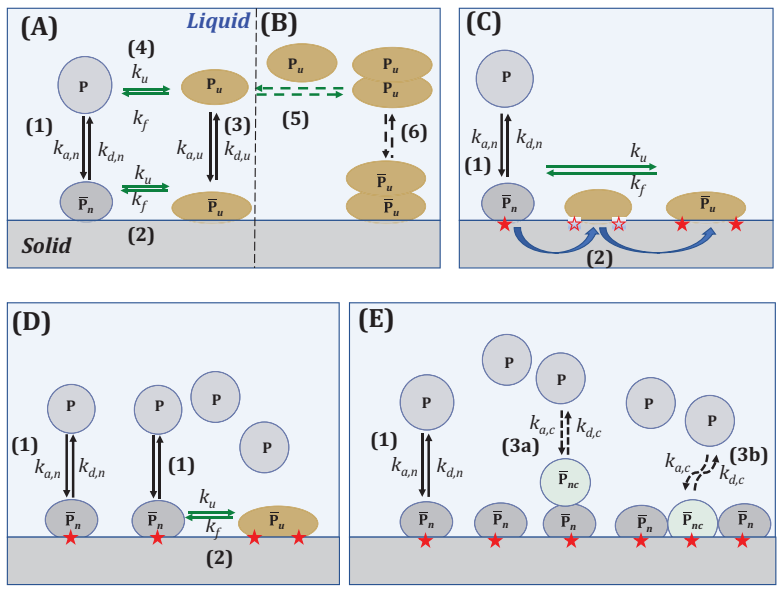


FIGURE 1.2 Cartoon representation of the unfolding models. A) Steps (1),(2),(3),(4) illustrate a four-state unfolding model of reversible reactions of binding of the native form, unfolding in adsorbed phase, binding of the unfolded form, unfolding in liquid phase, B) additional steps (5), (6) correspond to aggregation and binding of aggregates, respectively, C) three-state mechanism: (1) binding of the native form, (2) spreading on the surface by binding to another adsorption site with simultaneous unfolding, D) crowding effect: inhibition of unfolding in the presence of other molecules in the adsorbed phase, E) positive cooperative adsorption: (3a) multilayer adsorption, (3b) preferred adsorption.

where $\overline{P}_{c,k,ij}$ denotes the protein i adsorbed due to the interactions with bound molecules of the same protein when $j = i$ or with bound molecules of a different protein when $j \neq i$, in the native form $k = n$ or the unfolded form $k = u$, $k_{ac,k,ij}$, $k_{dc,k,ij}$ is the corresponding rate constant describing positive cooperative adsorption and desorption, N is the number of the proteins in the solution.

1.2.4 Model of Adsorption Kinetics

The reaction paths presented above can be quantified by kinetic equations that account for the phenomena described above [29].

For the adsorbed phase concentration of the protein $P_{n,i}$ in the native state, $q_{n,i}$, it holds:

$$\frac{\partial q_{n,i}}{\partial t} = k_{d,i}\left[K_{a,i}\ C_i\left(q^{\infty} - \sum_j \sum_l \delta_{l,ij}\, q_{l,j}\right) - q_{n,i}\right] - \frac{\partial q_{u,i}}{\partial t} \tag{1.4}$$

For the adsorbed phase concentration of the protein $P_{u,i}$ in the unfolded state, $q_{u,i}$:

$$\frac{\partial q_{u,i}}{\partial t} = k_{f,i}\left[K_{u,i}\ q_{n,i}\left(q^{\infty} - \sum_j \sum_l \delta_{l,ij}\, q_{l,j}\right) - q_{u,i}\right] \tag{1.5}$$

For positive cooperative adsorption of the protein $P_{c,i}$:

$$\frac{\partial q_{nc,i}}{\partial t} = \sum_j \sum_k k_{dc,k,ij}\left[K_{ac,k,ij}\ C_i q_{k,ij} - q_{c,k,i}\right] \tag{1.6}$$

For the total protein concentration in the adsorbed phase, $q_{tot,i}$:

$$q_{tot,i} = q_{n,i} + q_{u,i} + q_{nc,i} \tag{1.7}$$

$$\frac{\partial q_{tot,i}}{\partial t} = \frac{\partial q_{n,i}}{\partial t} + \frac{\partial q_{u,i}}{\partial t} + \frac{\partial q_{nc,i}}{\partial t} \tag{1.8}$$

$i = 1..N$, $j = 1..N$, $k = n, u$, $l = n, u, c$

In these equations, C_i, q_i are the liquid and adsorbed phase concentrations of the protein i respectively, in the native n, unfolded, u, and in clustered form, c, q^{∞} is the binding capacity, $\delta_{l,ij}$, are the exclusion factors for the native form, $l = n$, the unfolded form, $l = u$, and the clustered form $l = c$ of the protein, respectively. For $j = i$, $\delta_{l,ij}$ account for exclusion effects induced by the presence of all adsorbed molecules of the same protein. For $j \neq i$, $\delta_{l,ij}$ account for the same effects but with respect to different adsorbed proteins. The parameters: $K_{a,i} = \dfrac{k_{a,i}}{k_{d,i}'}$ $K_{u,i} = \dfrac{k_{u,i}}{k_{f,i}}$, $K_{ac,k,ij} = \dfrac{k_{ac,k,ij}}{k_{dc,k,ij}}$ are the equilibrium constants for adsorption on the bare surface, for unfolding, and for positive cooperative adsorption due to the interaction with the native form, $k = n$,

and the unfolded form, $k = u$, respectively. The kinetic model can be simplified by eliminating some of its terms or unifying some of the model coefficients [49].

If $K_{ac,k,ij} = 0$ and $\delta_{k,ii} = \delta_{k,ij,}$ Eqs. 1.4–1.8 are reduced to the multicomponent Langmuir competitive isotherm. Values of $\delta_{k,ij} > \delta_{k,ii}$ indicate negative deviations from the competitive Langmuir isotherm (i.e., negative cooperative adsorption effects), which can be attributed to size exclusion, repulsive interactions, or both [60–62]. The values of $K_{ac,k,ij} > 0$ indicate positive cooperative adsorption effects induced by the presence of proteins in the adsorbed phase. Positive cooperative adsorption corresponds to isotherms with a higher slope relative to a reference Langmuir isotherm, while negative cooperative adsorption corresponds to isotherms with a lower slope.

At low surface loadings, the effects of cooperative adsorption become negligible for both single and multicomponent systems, and Eqs. 1.4–1.8 reduce to the following equations:

$$\frac{\partial q_{n,i}}{\partial t} = k_{d,i}\left[H_{a,i}\ C_i - q_{n,i}\right] - \frac{\partial q_{u,i}}{\partial t} \tag{1.9}$$

$$\frac{\partial q_{u,i}}{\partial t} = k_{f,i}\left[H_{u,i}\ q_{n,i}\ -\ q_{u,i}\right] \tag{1.10}$$

where $H_{a,i} = K_{a,i}\, q^{\infty}$, $H_{u,i} = K_{u,i}\, q^{\infty}$ are the slopes of the isotherm (the Henry constants) for the native and unfolded forms, respectively.

At steady state, an explicit form of the isotherm equation can be derived from Eqs. 1.9 and 1.10:

$$q_{tot,i} = H_{a,i}\left(1 + H_{u,i}\right)C_i = H_{tot,i}C_i \tag{1.11}$$

where $q_{tot,i}$ is the total protein concentration in the adsorbed phase, $H_{tot,i} = H_{a,i}\left(1 + H_{u,i}\right)$ is the total Henry constant that corresponds to the total isotherm slope that is correlated with the retention factor.

1.2.5 Thermodynamic Dependencies of Adsorption on HIC Media

To correlate the Henry constant, H, with the salt concentration, several retention dependencies have been suggested, which typically have a form of logarithmic functions [1–4, 63–70]:

$$\log H = AC_{salt} + B \tag{1.12}$$

where A and B are parameters, which have a physical meaning.

At a sufficiently high salt concentration, the logarithmic dependence of the retention factor on the salt concentration is linear. However, to quantify retention behavior in a wide range of salt concentrations, more sophisticated multi-parameter retention models have to be used [3, 4, 18–26, 63–70]. For practical purposes, an empirical function, $H = f(C_{salt})$, may be employed, whose coefficients are fitted to the retention data acquired experimentally.

The retention properties of proteins are strongly altered by temperature. In general, increasing temperature enhances hydrophobic interactions and protein retention, and lowering temperature promotes protein elution [29,49,67–73]. The temperature dependence of the Henry constant is determined by the van't Hoff plot:

$$\ln H = -\Delta G^0 / RT \tag{1.13}$$

where ΔG^0 is Gibbs three energy $\Delta G^0 = \Delta H^0 + T\Delta S^0$, R is the gas constant, and T is the absolute temperature.

Eq. (1.13) represents a linear relationship $\ln H$ vs $1/T$, provided that ΔH^0 and ΔS^0 do not depend on temperature. If heat capacity changes with temperature, which may accompany conformational changes of proteins in the adsorbed phase, the quadratic dependence can be applied [37,42,43,69–73]:

$$\ln H = a + \frac{b}{T} + \frac{c}{T^2} \tag{1.14}$$

Although the effect of temperature on the adsorption properties of proteins in HIC may be complicated, this parameter can be used to promote elution and separation of proteins under mild conditions without denaturation and to improve the efficiency of the separation process [72,74,75].

1.3 COLUMN DYNAMICS

To predict the course of chromatographic elution, the underlying kinetic equations must be implemented in a dynamic model. For that purpose, a heterogeneous dynamic model (i.e., a general linear rate model) can be used, which directly accounts for extra- and intra-particle mass transport resistances. Since the model contains a number of parameters to be determined and requires an advanced numerical method for solving, it is often replaced with pseudo-homogeneous models, such as the kinetic-dispersive model, which is expressed as follows [43,55–58,76]:

$$\varepsilon_{t,i} \frac{\partial C_i}{\partial t} + u \frac{\partial C_i}{\partial x} + \left(1 - \varepsilon_t\right) \frac{\partial q_{tot,i}}{\partial t} = D_{L,a} \frac{\partial^2 C_i}{\partial x^2} \tag{1.15}$$

where C_i is the concentration of each protein i in the mobile phase, $q_{tot,i}$ is the total adsorbed concentration of each protein i related to the solid matrix volume, u is the superficial velocity in m s^{-1}, t is time in s, x is the axial coordinate, $D_{L,a}$ is the effective axial dispersion coefficient in m^2 s^{-1}, and $\varepsilon_{t,i}$, ε_i are the bed porosity accessible by each protein and the total bed porosity, respectively. For small molecules such as salt ions, which penetrate the whole pore volume of the bed, $\varepsilon_{t,i} = \varepsilon_t$.

The term $\frac{\partial q_{tot,i}}{\partial t}$ is expressed by the underlying kinetic equations (e.g., the model described by Eqs. 1.1–1.8). However, it should be kept in mind that the rate constants $k_{a,i}$ and $k_{d,i}$ lump kinetic limitations arising from the adsorption-desorption process as well as from diffusional mass transfer [49,76,77].

The model must be coupled with standard initial and boundary conditions and solved using a numerical procedure (e.g., [49,75]).

1.4 ELUTION BEHAVIOR UNDER LINEAR ISOTHERM CONDITIONS

1.4.1 Elution of Structurally Stable Proteins

A pattern of isocratic elution on an HIC medium, which is typical for structurally stable proteins, is shown in Figure 1.3A. An increase in kosmotropic salt concentration and temperature enhances the binding of the protein to HIC resins, which results in an increase in the Henry constant and thus in retention time. This is accompanied by band broadening due to slow rates of diffusional mass transport and the adsorption-desorption process. To accelerate the elution progress and mitigate kinetic effects, gradient elution is usually applied to reduce band broadening, where the elution strength of the mobile phase is enhanced by a gradual or stepwise reduction in the salt concentration (Figure 1.3B).

Illustrations of the underlying retention dependencies are shown in Figures 1.4A and 1.4B. They follow typical trends that were described in section 1.2.5: the dependency ln$H$ vs C_{salt} is linear at higher salt concentrations, and the dependency ln$H$ vs 1/T is linear over the range of mild temperature conditions, i.e., 5–25°C.

1.4.2 Elution of Structurally Unstable Proteins

The retention of structurally unstable proteins that unfold upon adsorption may involve multipeak elution in the gradient mode and incomplete elution of the protein

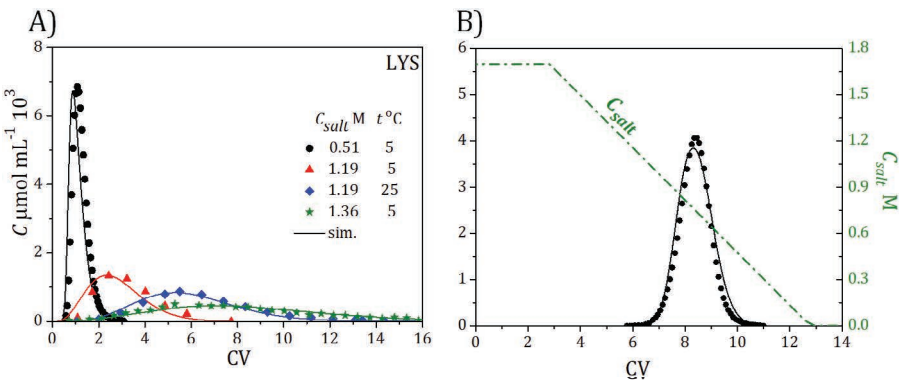


FIGURE 1.3 Illustration of elution pattern of LYS on 4FF Butyl Sepharose, the injection concentration C_{inj} = 0.069 µmol mL^{-1} (1 mg mL^{-1}), the injection volume V_{inj} = 0.1 mL, and the mobile phase flow rate Q = 1 mL min^{-1}. A) Isocratic band profiles of LYS at different salt concentrations and temperatures, B) gradient elution, linear gradient 0–1.7 M AS, C_{salt}, in 10 column volumes (CV) at 25^{0}C.

A from: R. Muca, W. Marek, W. Piątkowski, D. Antos, J. Chromatogr. A, 1217 (2010), 2812–2820 [55].

B from: R. Muca, W. Piątkowski, D. Antos, J. Chromatogr. A, 1216 (2009), 8712–8721 [74]. Symbols—experimental data, lines—model simulations.

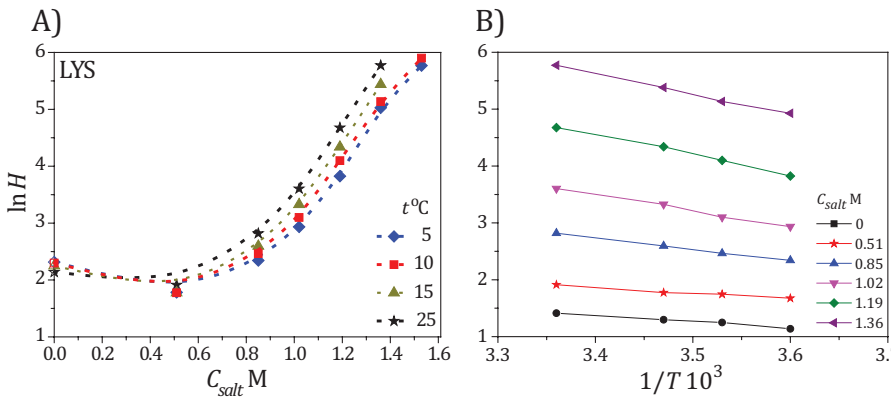


FIGURE 1.4 Variations of the Henry constant H of LYS on an HIC resin vs. the salt concentration and temperature. A) Dependency of ln$H$ on the salt concentration at different temperatures, B) dependency of ln$H$ on 1/T at different salt concentrations. Lines are a guide for the eye.

From: R. Muca, W. Marek, W. Piątkowski, D. Antos, J. Chromatogr. A, 1217 (2010), 2812–2820 [55].

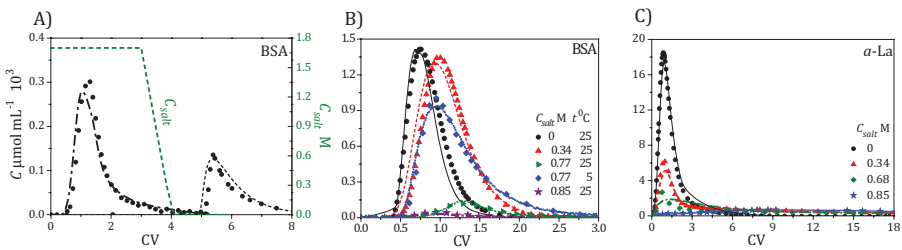


FIGURE 1.5 Elution pattern of BSA and a-La, flow rate Q = 1 mL min^{-1}, V_{inj} = 0.1 mL, $C_{BSA,inj}$ = 0.015 µmol mL^{-1} (1 mg mL^{-1}), $C_{\alpha\text{-La},inj}$ = 0.35 µmol mL^{-1} (5 mg mL^{-1}). A) Illustration of band splitting of the BSA profiles in salt gradient elution, B) effect of the salt concentration, C_{salt}, and temperature on partial elution of BSA in isocratic mode, C) effect of the salt concentration on partial elution of α-La.

From: R. Muca, M. Żurawski, W. Piątkowski, D. Antos, J. Chromatogr. A, 1492 (2017), 79–88 [49] and from: R. Muca, W.K. Marek, W. Piątkowski, D. Antos, J. Chromatogr. A 1217 (2010) 2812–2820 [55].

in isocratic mode. In the gradient mode, the earlier eluting peak is assigned to native or predominantly native proteins, whereas the more retained is assigned to partially unfolded proteins. In the isocratic mode, only an earlier-eluting peak appears at the column outlet; the remaining amount of the protein retains in the column until the salt content in the mobile phase is reduced. For low protein loads corresponding to the linear isotherm range, the extent of incomplete elution or peak-splitting phenomena depends on the salt concentration in the mobile phase, the temperature, and the mobile phase flow rate.

Figure 1.5 demonstrates the influence of the salt concentration and temperature on the elution behavior of BSA and α-La, which as mentioned above, are representatives

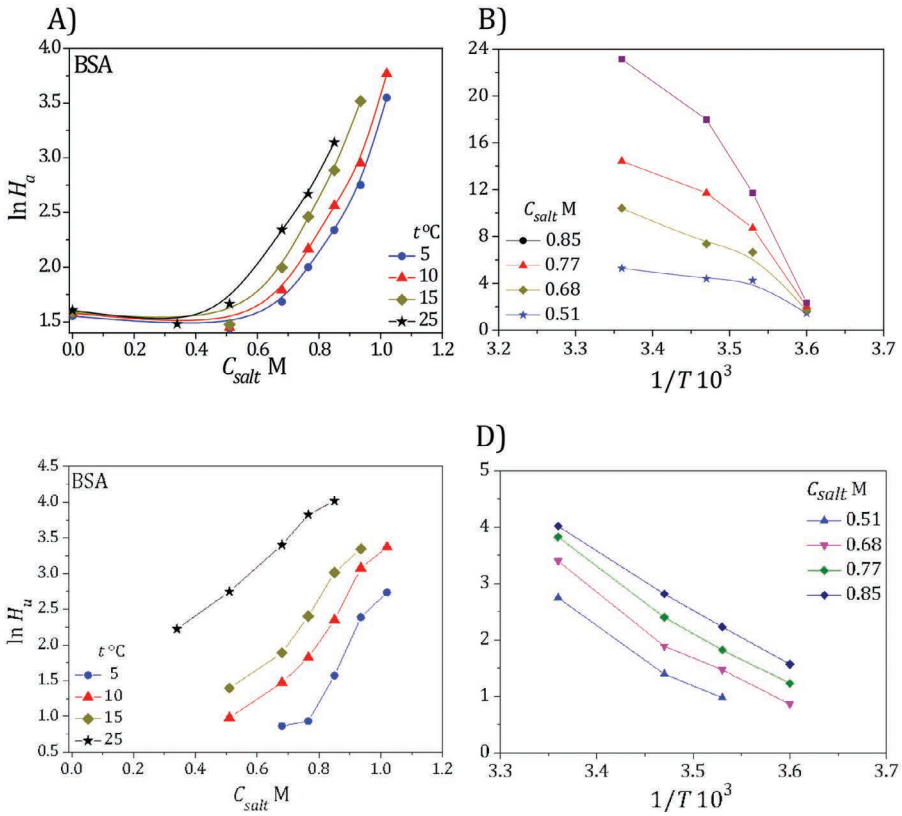


FIGURE 1.6 Variations of the Henry constant H of BSA vs. the salt concentration and temperature. A) Dependency of $\ln H_a$ vs C_{salt} at different temperatures, and B) $\ln H_a$ vs $1/T$ at different salt concentrations, C) dependency $\ln H_u$ vs C_{salt} and D) dependency $\ln H$ vs $1/T$. Lines are a guide for the eye.

From: R. Muca, W. Marek, W. Piątkowski, D. Antos, J. Chromatogr. A, 1217 (2010), 2812–2820 [55].

of proteins with unstable structures. In Figure 1.5A, a two-peak elution of BSA in gradient mode is shown, where the first peak was assigned to the protein that was both adsorbed and desorbed in the native form, whereas the second one was assigned to the protein that unfolded upon adsorption and desorbed in the native form, according to the three-point mechanism illustrated in Figure 1.2C. This effect is enhanced with increasing salt concentration and thus the binding strength of the protein, which manifests itself by reducing the amount of the protein eluted in the first peak. This is demonstrated in Figures 1.5B and 1.5C, in which the first-eluted peaks recorded at different salt concentrations are superimposed. Similarly, an increase in temperature induces a reduction in the size of the fist-eluting peak (Figure 1.5B).

Such a retention pattern reflects the salt and temperature retention dependencies whose courses depart from the typical pattern reported previously. Figure 1.6 presents

hypothetical salt and temperature dependences of the Henry constants (H_a, H_u) of BSA for adsorption-desorption and unfolding-folding processes under linear isotherm conditions. The dependences were assessed using the peak fitting method, in which the dynamic model parameters were adjusted to reproduce partial elution profiles of the proteins. Both the salt and temperature dependences of $\ln H_a$ and $\ln H_u$ deviate from linearity.

1.5 ELUTION BEHAVIOR UNDER NON-LINEAR ISOTHERM CONDITIONS

1.5.1 Elution of Structurally Stable Proteins

An increase in the protein load up to non-linear isotherm conditions usually results in enhancement of peak asymmetry, that is, formation of a fast-moving sharp concentration front followed by peak tailing, which is characteristic of the Langmuirian type of adsorption mechanism. A typical pattern of changes in peak shape with increasing protein load in isocratic elution mode is illustrated for a model protein, LYS, in Figure 1.7A. The presence of other adsorbing compounds induces the displacement effect, which causes the protein to accelerate its migration along the column. Figure 1.7B illustrates changes in the retention of LYS in a binary mixture with polyethylene glycol (PEG 3.3 kDa). The molar concentration of PEG is much higher than that of LYS, which stems from the differences in their molecular weights and sizes. This causes PEG, which itself is weakly adsorbed on HIC media, to displace LYS and reduce its retention [78]. This implies that the size exclusion effect strongly contributes to the mechanism of multicomponent adsorption.

The shape of the band profiles of structurally unstable proteins differs from the common pattern. An increase in the protein load induces the crowding effect

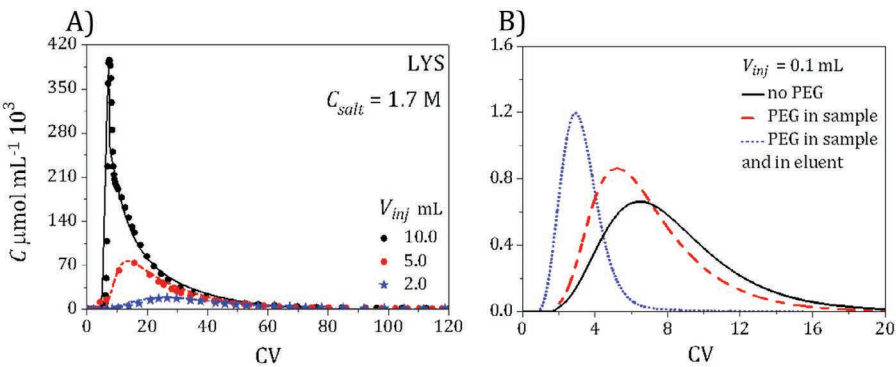


FIGURE 1.7 Illustration of changes in shape of band profiles of LYS on an HIC resin, Q = 1mL min^{-1}, A) at different protein loading volumes, C_{inj} = 0.382 µmol mL^{-1} (5.5 mg/mL^{-1}), B) in the presence of PEG 3.35 kDa, C_{salt} = 1.19 M, $C_{inj,LYS}$ = 0.069 µmol mL^{-1} (1 mg mL^{-1}), C_{PEG} in the sample or in both of the samples in the eluent C_{PEG} = 0.59 µmol mL^{-1} (2 mg mL^{-1}).

A from: I. Poplewska, W. Piątkowski, D. Antos, J. Chromatogr. A. 1386 (2015)] 1–12 [79].

B from: W.K. Marek, W. Piątkowski, D. Antos, Chromatographia 81 (2018)1641–1648 [78].

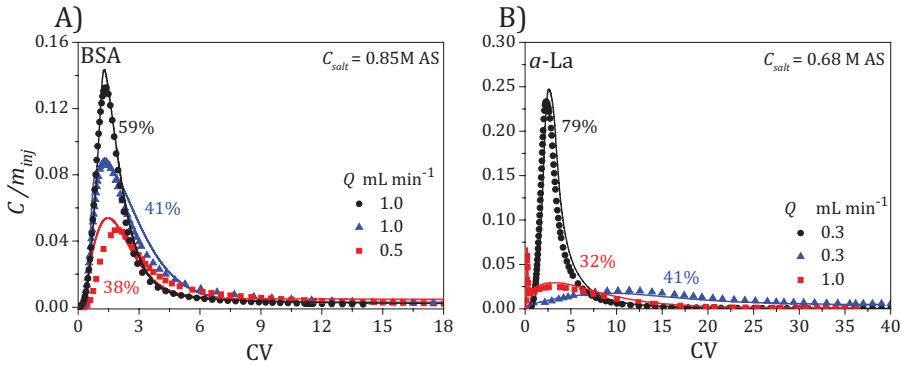


FIGURE 1.8 Effect of protein load and flow rate on the retention behavior of BSA and α-La. The percentage amount denotes the ratio of the mass of the protein eluted to the mass loaded into the column. A) BSA, 59% corresponds to the load 2.9 mg mL^{-1}_{resin}, 41% and 38% correspond to the same protein load of 0.36 mg mL^{-1}_{resin} but at different flow rates, B) α-La, 79% for the protein load of 7.3 mg mL^{-1}_{resin}, 41% and 32% for the protein load of 0.36 mg mL^{-1}_{resin} but at different flow rates.

From: R. Muca, M. Żurawski, W. Piątkowski, D. Antos, J. Chromatogr. A, 1492 (2017) [55].

mentioned earlier, in which the presence of the adsorbed molecules prevents the protein from spreading and unfolding. Therefore, at high protein loadings, protein unfolding is inhibited and the native form of the protein prevails in the adsorbed phase. The influence of the protein load on the extent of incomplete elution is demonstrated in Figure 1.8. Yet, when unfolding is accompanied by aggregation in the liquid or adsorbed phase (Figure 1.2A stages 5, 6 or Figure 1.2E), an increase in the protein load may cause contradictory effects: on the one hand, it can diminish formation of the unfolded form and its subsequent aggregation, but on the other hand, it can enhance aggregation, which is a higher-order reaction, thus accelerating with increasing protein concentration.

The extent of the incomplete elution effect also depends on the mobile phase flow rate, which is also demonstrated in Figures 1.8A and 1.8B. As the rate of the unfolding process is low [36,41], reduction in the mobile phase flow rate, thus increasing the contact time of the protein with the adsorbed surface, favors formation of the unfolded form.

1.5.2 Elution of Structurally Unstable Proteins

The presence of other proteins in the adsorbed phase also induces the crowding effect, which is illustrated in Figure 1.9, where the isocratic elution profiles of binary mixtures of BSA and LYS, and BSA and α-La are shown. The incomplete elution of BSA, which results from the protein unfolding, is diminished in the presence of smaller proteins (i.e., LYS) and α-La. In both cases, the elution of the smaller protein was almost unaffected by the presence of BSA. This confirms that the size

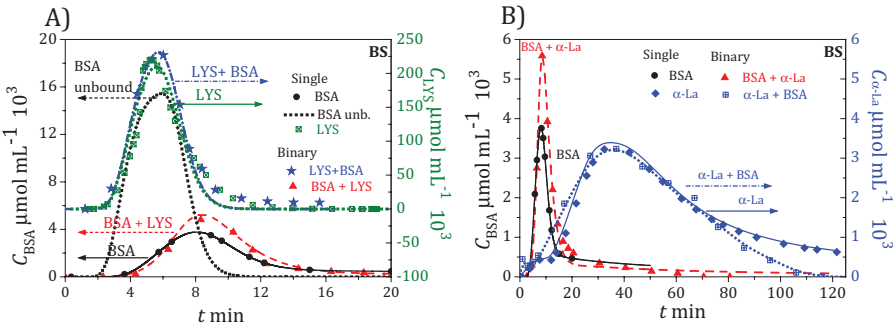


FIGURE 1.9 Elution patterns of BSA in binary mixtures with LYS and α-La on the BS resin at C_{salt} = 0.765 M AS, on the BS resin. A) Upper plot—single LYS and LYS in a binary mixture with BSA, lower plot—single BSA, and BSA in a binary mixture BSA-LYZ, $C_{inj,BSA}$ = 16 x 10^{-3} µmol mL^{-1} (1.1 mg mL^{-1}), $C_{inj,LYZ}$ = 217 × 10^{-3} µmol mL^{-1} (3.1 mg mL^{-1}), V_{inj} = 2 mL for both single and binary component samples, B) BSA and LAC in single and binary component samples, $C_{inj,BSA}$ = 15 × 10^{-3} mmol mL^{-1} (1.0 mg mL^{-1}), $C_{inj,LAC}$ = 70.5 × 10^{-3} µmol mL^{-1} (1.0 mg mL^{-1}), V_{inj} = 2 mL.

From R. Muca, M. Kołodziej, W. Piątkowski, G. Carta, D. Antos, J. Chromatogr. A, 1625 (2020) 461309 [29].

exclusion effect exerts a dominant influence on the mechanism of multicomponent adsorption.

1.5.3 Adsorption Equilibrium: Isotherm Courses

The isotherm course of proteins adsorbed on HIC media reflects the underlying adsorption mechanism; therefore, it can be affected by unfolding and positive and negative cooperative adsorption. Examples of the isotherm shapes measured for different proteins on HIC resins using static and dynamic methods are shown in Figures 1.10 and 1.11.

In the static method, the resin slurry stayed connected with the protein solutions until adsorption equilibrium was established. The equilibrium data were exploited to determine the thermodynamic coefficients of the adsorption model. In the dynamic method, the thermodynamic coefficients were adjusted by fitting predictions by the dynamic model to the experimental peaks recorded for different loading conditions. The resins used for acquiring the data differed in the pore size and the matrix structure, that is, 4FF Butyl Sepharose (BS) with a pore size of 30 nm based on agarose matrix and TOYOPEARL Butyl-650C (TP) with a pore size of 100 nm based on methacrylic polymer. The ranges of molar concentrations selected for the isotherms are relevant to the chromatographic band profiles presented above.

For all static measurements presented in Figure 1.10, a continuous increase in the adsorbed protein concentration with increasing liquid phase concentration can be observed. The isotherm follows the course of favorable isotherms with a decreasing

slope as the protein concentration increases. Adsorption capacity is enhanced with increasing salt concentration, as exemplified in Figures 1.10C and 1.10D. Similar adsorption behavior of proteins on HIC media has been indicated in several studies [25, 29, 49, 80–82].

Furthermore, all curves exhibited convex and nonlinear curvature in a wide concentration range. In the case of proteins whose structure underwent conformational changes upon adsorption (e.g., BSA and α-La), the isotherm curvature bended at low protein concentrations, which indicated that, in the chromatographic process, nonlinear isotherm conditions can already be encountered at low protein loads. Furthermore, the molar concentration of the protein in the adsorbed phase at equilibrium q^* was much lower for BSA than for a-La, which again confirmed the significant contribution of exclusion effects in the adsorption mechanism.

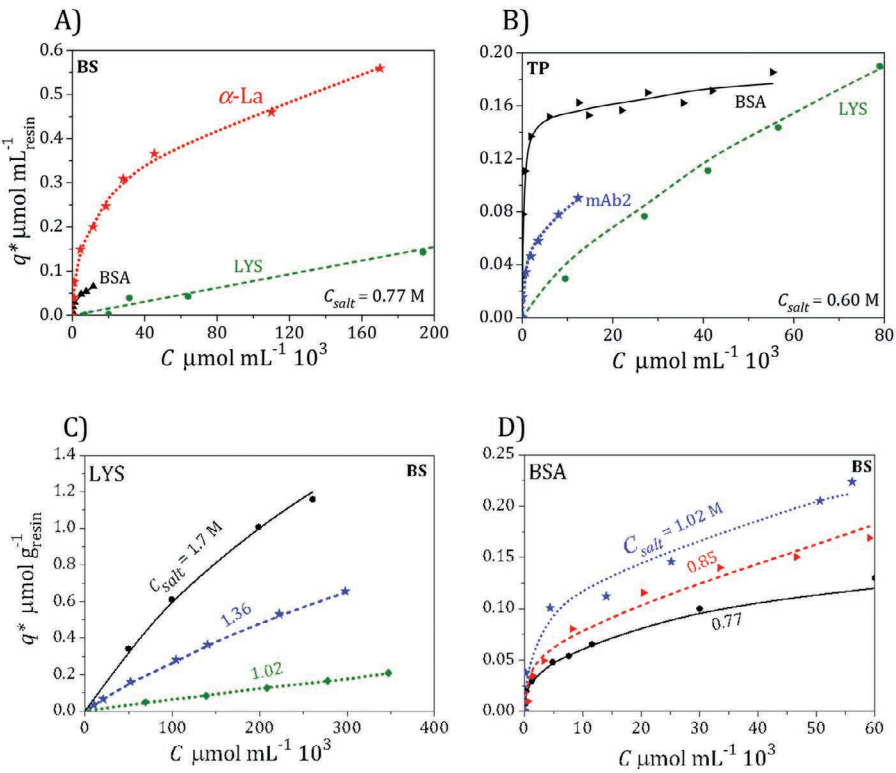


FIGURE 1.10 Isotherm courses for single proteins. A) BSA, LYS, α-La on the BS resin, B) BSA, LYS, and a monoclonal antibody (mAb2) on the TP resin, C) influence of the salt concentration on the isotherm course for LYS, and D) for BSA.

A) and B) from: R. Muca, M. Żurawski, W. Piątkowski, D. Antos, J. Chromatogr. A, 1492 (2017) 79–88 [49], C) from: I. Poplewska, W. Piątkowski, D. Antos, J. Chromatogr. A. 1386 (2015)] 1–12, [79], D) from: R. Muca, M. Kołodziej, W. Piątkowski, G. Carta, D. Antos, J. Chromatogr. A, 1625, (2020) 461309 [29].

Moreover, the courses of the curves did not follow the Langmuir-type isotherms, as demonstrated on the Scatchard plots in Figures 1.11B and 1.11D. The maximum of the Scatchard plot can be attributed to a change in the adsorption mechanism from negative to positive cooperative adsorption.

In the isotherm courses predicted by the dynamic method, the upward-sloping curve is missing; therefore, they significantly differ from those determined using the static method. In the case of BSA, the isotherm courses obtained by both methods converge only at very low protein concentrations, whereas for a-La they are different over the whole concentration range. This discrepancy can be attributed to kinetic limitations, which hinder protein binding, unfolding, and formation of protein clusters in short-lasting chromatographic elution. Therefore, the determination of all parameters of the dynamic model must be based on both dynamic profiles and adsorption equilibrium data. The partial elution profiles can be used to assess the kinetics of protein binding and unfolding, whereas the isotherm data can be used to quantify the effect of the cluster formation in the adsorbed phase.

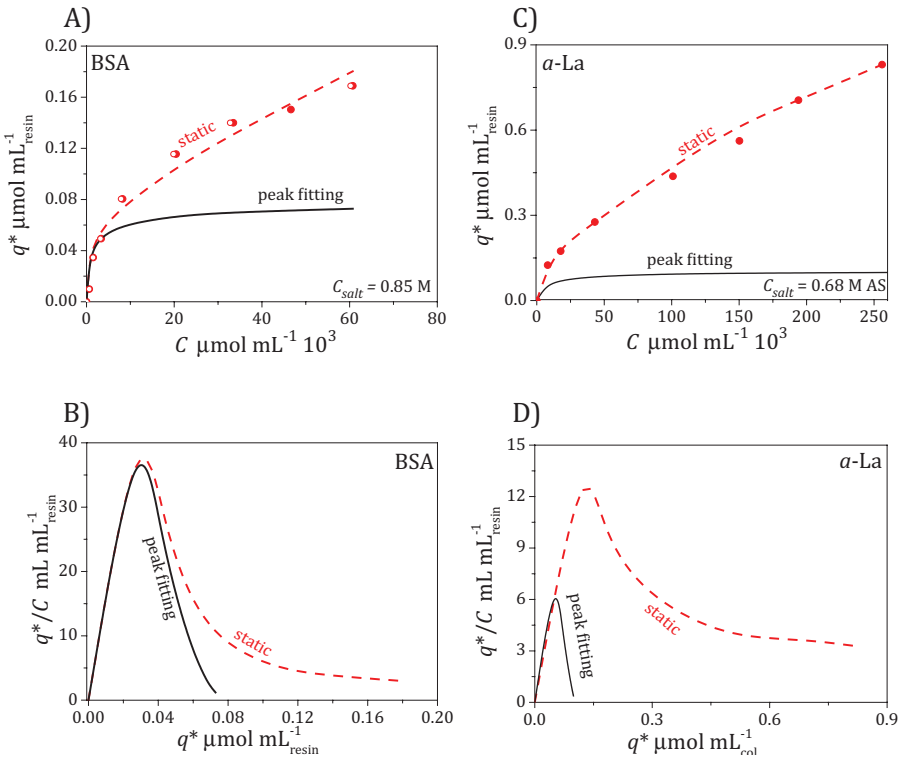


FIGURE 1.11 Adsorption isotherms determined by static and dynamic methods. A) Isotherm for BSA, and B) the corresponding Scatchard plot $q^*/C = f(q^*)$, C) isotherm of α-La, and D) the corresponding Scatchard plot.

From: R. Muca, M. Żurawski, W. Piątkowski, D. Antos, J. Chromatogr. A, 1492 (2017) [49].

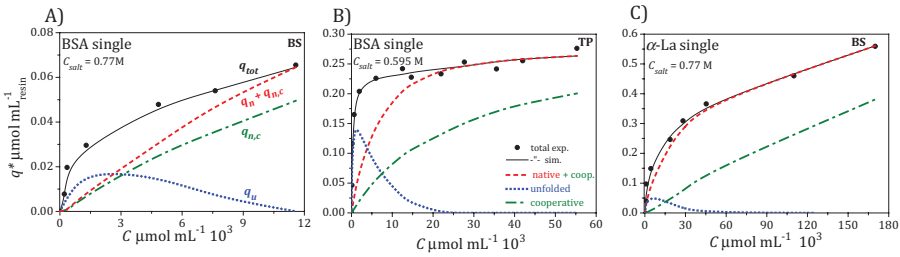


FIGURE 1.12 Predicted contributions of different adsorption mechanisms to total protein adsorption. A), B) BSA on the BS and TP resins, respectively, and C) LAC on the BS resin. q_{tot}, q_n, q_u, q_{nc} are the adsorbed phase concentrations corresponding to adsorption: total, native form, unfolded form, and native in clusters, respectively (Eqs. 1.1–1.8).

From Muca, M. Kołodziej, W. Piątkowski, G. Carta, D. Antos, J. Chromatogr. A, 1625, (2020) 461309 [29].

Figure 1.12 presents the hypothetical contribution of different surface mechanisms to the isotherm course of BSA and α-La predicted based on the three-state mechanistic model (Eqs. 1.1–1.8). The model predicted a maximum formation of the unfolded form over the range of low protein concentrations and its decay at higher concentrations accompanied by an increasing contribution of the cluster formation.

Figures 1.13 and 1.14 illustrate the adsorption isotherms obtained for binary mixtures of different proteins. For comparison, the individual isotherms are overlaid on the same plots. Figures 1.13A–1.13D present the isotherms for LYS and mAb2 in the form of incremental concentrations of the proteins for both resins; Figures 1.13A and 1.13C illustrate the effect of the presence of LYS on the adsorption of mAb2, and Figures 1.13B and 1.13D illustrate the effect of the presence of mAb2 on the adsorption of LYS. In all cases, increasing concentrations of LYS cause significant reduction in the binding strength of mAb2, whereas increasing concentrations of mAb2 almost do not influence the adsorption of LYS. This can be explained by the difference in the molar concentrations of BSA and LYS correlated with the molecular size of the proteins, which triggers the size exclusion effect; small molecules of LYS replace large molecules of mAb2 that are not accommodated well in the 30 nm pores of the BP resin.

The isotherms were more favorable for the TP resin compared with the BP resin; therefore, the effect of competitive binding (negative cooperative adsorption) was more pronounced for the former (Figures 1.13C and 1.13D). The size exclusion effect was still active, that is, the presence of LYS affected the adsorption of mAb2 more than the presence of mAb the adsorption of LYS, but it was weaker than for the BP resin. This can be explained by the difference in the pore size of the resins; mAb2 molecules have better access to 100 nm pores of the TP resin compared with the BS resin.

Figures 1.14A and 1.14B present the isotherms for the pair LYS and BSA on the BS resin; in Figure 1.14A, the effect of the presence of LYS on the adsorption of BSA is presented, while in Figure 1.14B, the effect of the presence of BSA on the adsorption

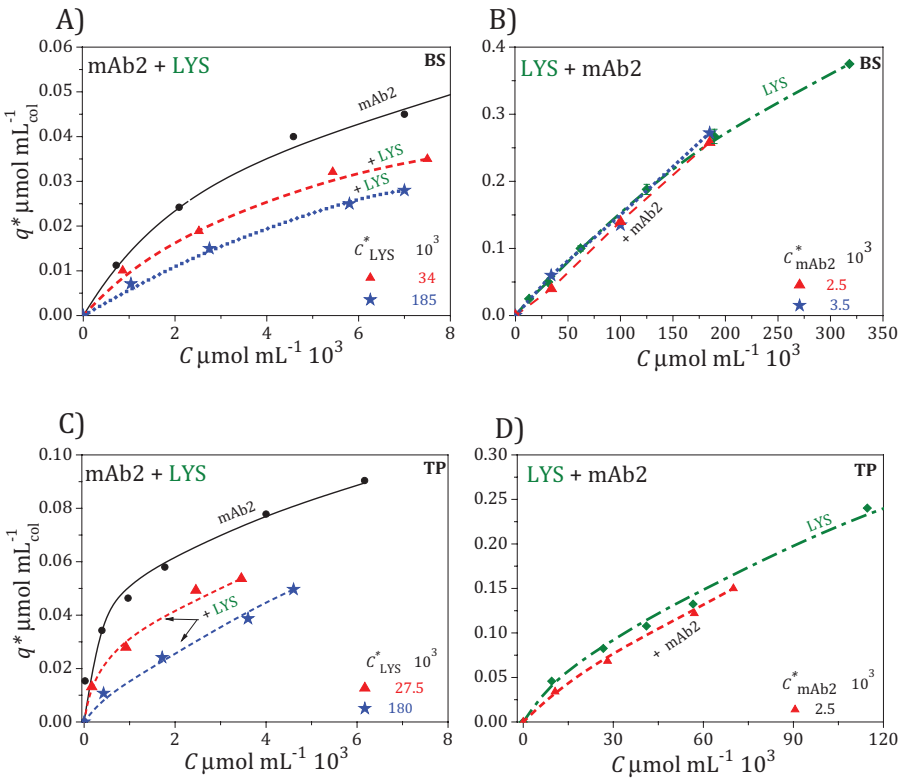


FIGURE 1.13 Adsorption isotherms of binary mixtures of mAb2 and LYZ on the BS resin. A), C) Influence of LYS on the adsorption of mAb2 on the BP (snapshot) and TP resins, respectively; B), D) influence of mAb2 on the adsorption of LYZ on the BP and TP resins, respectively. The salt concentration for BS: 0.765 M AS and for TP—0.595 M AS. *C** is the equilibrium concentration in binary mixtures in µmol mL^{-1}.

From: R. Muca, M. Kołodziej, W. Piątkowski, G. Carta, D. Antos, J. Chromatogr. A, 1625, (2020) 461309 [29].

of LYZ is presented. The favorable shape of the isotherm of BSA diminishes in the presence of LYS and takes a linear form at higher LYS concentrations (above about $180x10^{-3}$ µmol mL^{-1}), whereas increasing concentrations of BSA almost do not affect the adsorption of LYS. This unusual adsorption behavior may result from differences in the adsorption of the native and unfolded forms of BSA. Adsorption of the native form was relatively weak and occurred in a linear isotherm range, whereas the adsorption of the unfolded form was much stronger and was characterized by a favorable non-linear isotherm. It can be expected that the presence of LYS induced the crowding effect, which reduced the unfolding of BSA on the adsorbent surface. This stems from the difference in the molar concentrations of BSA and LYS and the molecular size of the proteins. At a sufficiently high LYS concentration, the native form of BSA prevails in the adsorbed phase.

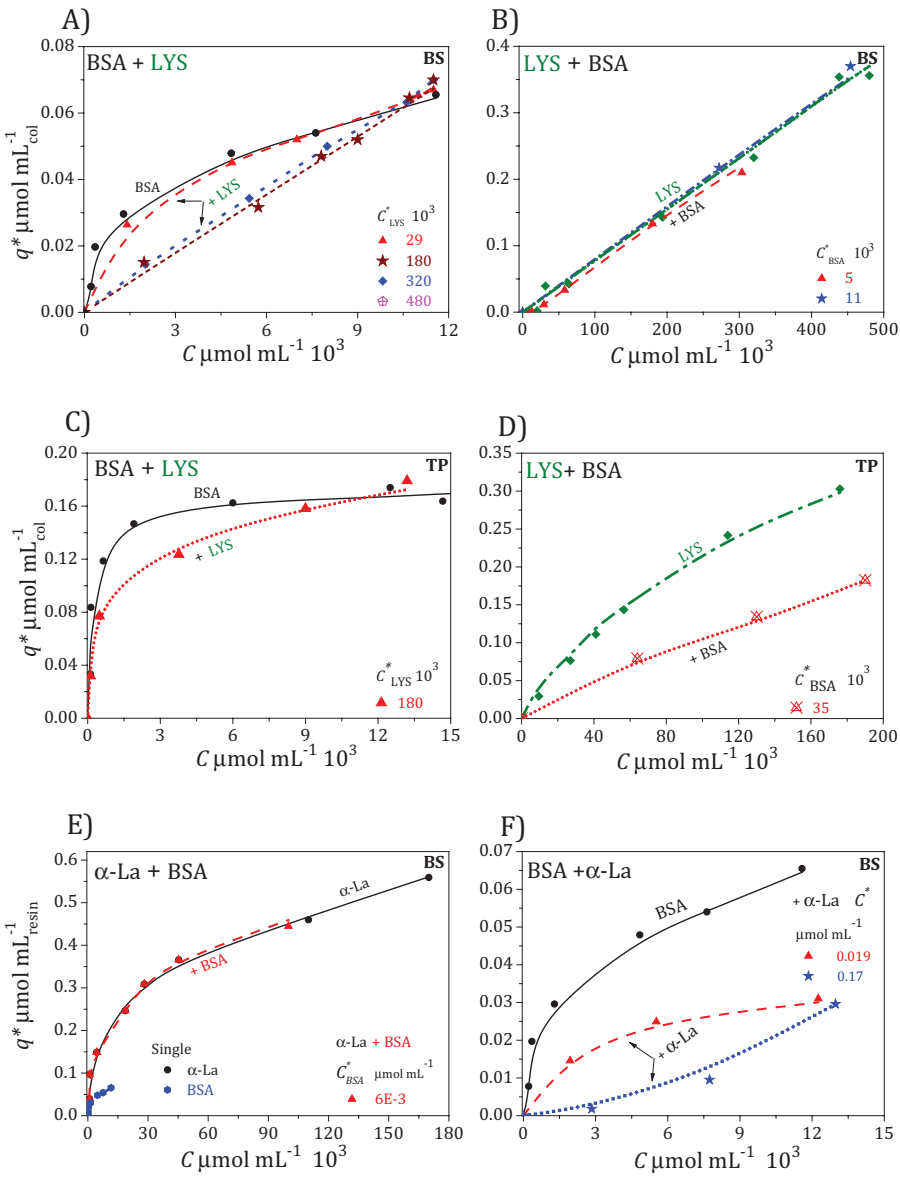


FIGURE 1.14 Adsorption isotherms of binary mixtures of BSA with LYS or BSA with α-La. A), C) Influence of LYS on the adsorption of BSA on the BS and TP resins, respectively, B), D) influence of BSA on the adsorption of LYS on the BS and TP resins, respectively, E) influence of BSA on the adsorption of a-Lac, F) influence of α-Lac on the adsorption of BSA. The salt concentration for BS: C_{salt} = 0.765 M AS and for TP: C_{salt} = 0.595 M AS**.**

From: Muca, M. Kołodziej, W. Piątkowski, G. Carta, D. Antos, J. Chromatogr. A, 1625, (2020) 461309 [29].

Figures 1.14C and 1.14D illustrate the adsorption of BSA and LYS on the TP resin. BSA was less affected by the presence of LYS compared to the BS resin, which could have resulted from the difference in pore size between the BP and TP resins. Moreover, the pores of the TP resin can accommodate BSA molecules better than mAb2; therefore, the adsorption pattern of the pair of BSA and LYS on the TP resin is more characteristic of competitive binding compared with the pair of mAb2 and LYS (Figure 1.13).

The adsorption equilibrium of the pair α-La and BSA on BS is illustrated in Figures 1.14C and 1.14D; an increase in the concentration of α-La caused a strong reduction in the adsorption of BSA, whereas the adsorption of α-La was only slightly affected by BSA within the molar concentration range analyzed. Again, this can be attributed to the difference in the molar concentrations of the proteins as well as the stronger non-linear binding of α-La.

1.6 SOLUBILITY LIMITATIONS

1.6.1 Sample Solvent Effect

The binding capacity of proteins on HIC media increases with increasing concentrations of kosmotropic salt; therefore, a relatively high salt content is usually used in the loading step of the chromatographic process. The presence of a high amount of kosmotropic salt in the solution may cause protein precipitation; therefore, the protein concentration in the loading buffer is restricted by the protein solubility. Thus, to increase the separation throughput, a high volume of dilute protein solution is often loaded into the column, which increases the duration of the loading step and thus impairs the process productivity. To reduce the loading volume, the protein can be dissolved up to a high concentration in the solution with a low salt content and eluted with a salt-reach mobile phase. Since the elution strength of the solutions strongly depends on the salt content, such a procedure can alter the adsorption behavior of proteins and cause band deformation. The sample solvent effect also occurs when the solution to be processed by HIC is an effluent of another separation process whose solvent environment is different from the mobile phase in HIC. A high elution strength of the sample solvent triggers accelerating the migration velocity of chromatographic peaks, which reduces solute retention [56, 57, 83–90]. An increase in the injection volume causes the protein to co-elute with the sample solvent over a longer distance in the column. This results in band broadening, which enhances with increasing injection volume. Moreover, kinetic effects arising from slow rates of mass transport may promote band deformation and cause the protein to co-elute with the sample solvent over the full length of the column.

Figure 1.15A shows a band profile of a protein dissolved in a salt-free solution, injected with a salt-rich loading buffer, and eluted with a salt gradient. The protein band split into two separated peaks. The first appeared at the column outlet along with the sample solvent, whereas the second one was eluted only with the salt gradient. However, when the feed was dosed in small-volume portions, the sample solvent diluted in the mobile phase while migrating through the column, and band splitting was avoided (Figure 1.15B).

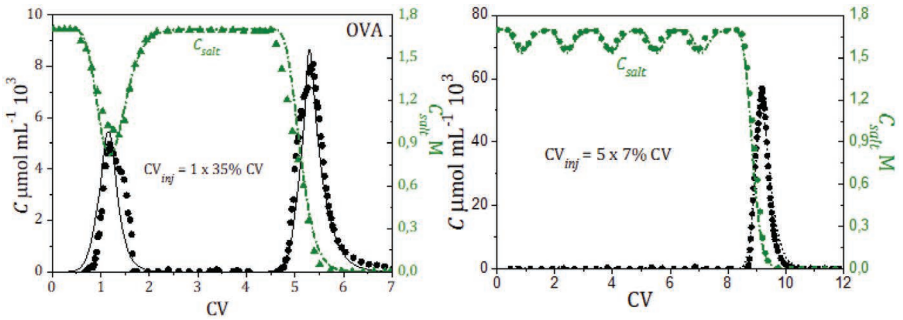


FIGURE 1.15 HIC elution of ovalbumin (OVA) dissolved in a salt-free loading buffer. A) Large injection volume, B) small injection volume in multiple injections.

From: W. Marek, R. Muca, W. Piątkowski, D. Antos, J. Chromatogr. A1218 (2011) 5423–5433 [56].

The extent of the phenomenon also depends on the mobile phase flow rate; a change in flow rate causes the corresponding change in the peak size and in the duration of the contact between the protein and the sample solvent, hence in the contribution of adsorption kinetics to band broadening.

The remedy for undesirable sample-solvent effects might be a reduction in the loading volume and its compensation by an increase in the protein concentration.

1.6.2 Risk of In-Column Precipitation

As mentioned in Section 1.6.1, another problem that can occur when processing the protein in concentrated solutions of kosmotropic salts is the risk of precipitation or crystallization inside the column. The presence of an additional phase in the form of amorphous precipitate, gel, or protein crystals can trigger flow blockage, destruction of elements of the chromatographic system, and failure of the separation process. This phenomenon can occur in the course of protein elution when the protein solubility limit is exceeded due to changes in the salt concentration in the mobile phase. This issue is illustrated in Figure 1.16 for LYS. Figure 1.16A shows a solid–liquid equilibrium (SLE) diagram of LYS in AS solutions. The cloud point line indicates the lower boundary for protein precipitation or gelation, which are fast or instantaneous processes. Between the cloud point line and the solubility line, there is the metastable region in which crystallization of the protein is possible but kinetically inhibited. If the crystallization rate is sufficiently slow, which is often the case in protein crystallization, the protein can be processed in chromatographic columns in the metastable zone without triggering crystallization. Nevertheless, design of such a process requires knowledge of SLE and crystallization kinetics. Figure 1.16B shows the elution of the protein loaded in the mobile phase with a concentration below the solubility limit and desorbed with a salt gradient. The corresponding local concentration levels of the cloud point and the solubility indicate that the protein concentration in the desorption band was far below the cloud point boundary, whereas it markedly exceeded the solubility limit. Figure 1.17A shows the elution of

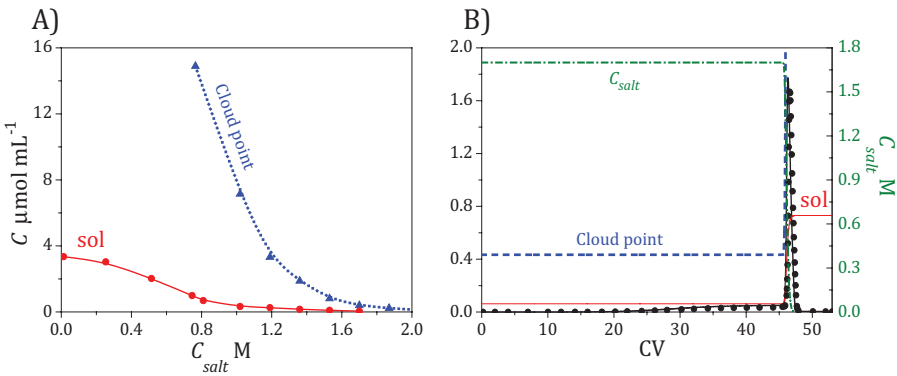


FIGURE 1.16 Illustration of the solubility limitation problem. A) Phase diagram for LYS in AS solutions: cloud point and the solubility (sol) curves, symbols experimental data, lines are a guide for the eye, B) example of course of gradient elution: LYZ dissolved in 1.7 M AS, V_{inj} = 67 mL, C_{inj} = 0.049 μmol mL^{-1} (0.7 mg mL^{-1}), protein load 34 mg mL$^{-1}_{resin}$, and eluted with a step gradient of AS, symbols—experimental profile, lines—simulations by a dynamic model.

From: I. Poplewska, W. Piątkowski, D. Antos, J. Chromatogr. A. 1386 (2015) 1–12 [79].

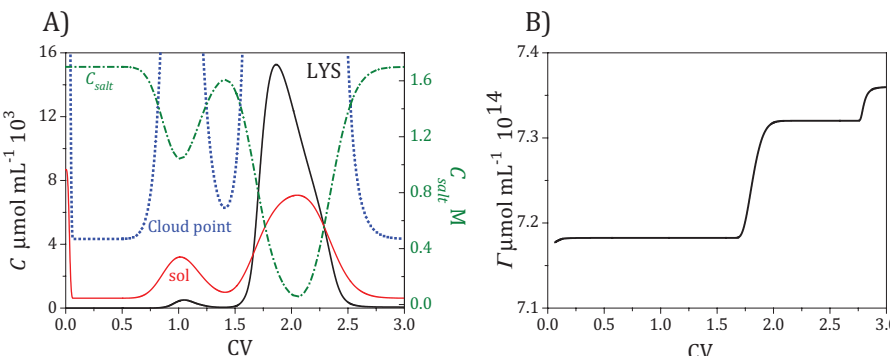


FIGURE 1.17 Concentration profiles of LYS, A) dissolved in a salt-free buffer, C_{inj} = 78 mg mL^{-1}, V_{inj} = 0.2 mL, loaded and washed with 1.7 M AS and eluted with a salt-free buffer along with the corresponding profiles of the salt, local cloud point, and local solubility, (B) simulations of the corresponding concentration of the crystalline protein, Γ.

From: I. Poplewska, W. Piątkowski, D. Antos, J. Chromatogr. A. 1386 (2015) 1–12, [79].

the protein dissolved in a salt-free loading buffer and eluted with the salt gradient. In both cases, the protein concentration locally fell into the metastable zone, but due to the very slow crystallization rate, the crystalline phase was practically not formed. It is shown in Figure 1.17B, where the amount of the crystalline phase formed during the elution process, which was calculated by coupling the dynamic model with crystallization kinetics specific to the protein, is depicted. The concentration of the crystalline protein did not exceed 10^{-14} μmol per mL of the resin.

1.7 BAND DEFORMATION IN THERMALLY HETEROGENEOUS COLUMNS

Since the adsorption properties of proteins on HIC media depend strongly on temperature, temperature gradients might be used to improve separation selectivity or even to replace the salt gradient with the temperature gradient. An example of HIC separation of a model ternary mixture by a combination of temperature and salt gradient is shown in Figure 1.18 [74]. A baseline separation of myoglobin (MYO) and LYZ was achieved by a step change in temperature; the most retained BSA was received by a step change in the salt concentration.

However, a prerequisite for efficient realization of temperature-mediated separations is to ensure uniformity of temperature distribution in radial and axial directions by fast exchange of thermal conditions of the system. Thermal heterogeneity can be caused by viscous friction of the mobile phase or ineffective thermal equilibration of the mobile phase [91–96]. The latter source is of major importance in ultra-HPLC columns packed with submicron porous particles [96–102] but insignificant in low-pressure HIC systems [75]. However, differences in the temperature of the mobile phase and the temperature of the column wall induce the formation of axial and radial temperature gradients, which may result in the departure of the temperature gradient from the desired shape and distortion of protein band profiles. This phenomenon is affected by the mobile phase flow rate, column dimensions, and temperature differences between the column wall and the mobile phase [75].

An example of the radial temperature gradient measured in an HIC column operated at low pressure is presented in Figure 1.19A. The column wall was thermostatted

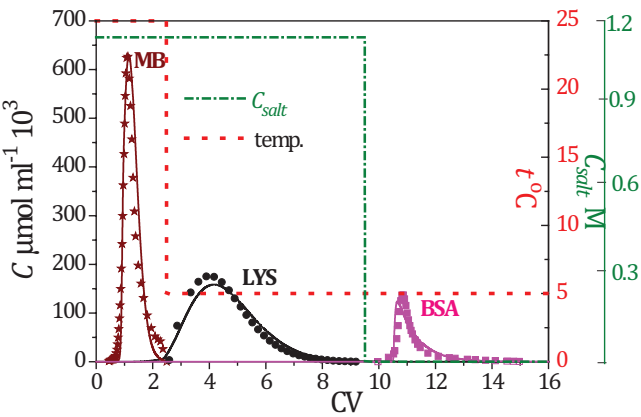


FIGURE 1.18 Elution of a ternary protein mixture of myoglobin (MB), LYS, BSA by a combination of temperature and salt gradients of, $C_{inj,MB}$ = 0.058 μmol mL^{-1} $C_{inj,LYS}$ = 0.070 μmol mL^{-1}, $C_{inj,BSA}$ = 0.015 μmol mL^{-1} ($C_{inj,MB,LYS,BSA}$ = 1.0 mg mL^{-1}), V_{inj} = 0.1 mL, Q = 0.5 mL min^{-1}, symbols—experimental profile, lines—simulation by a dynamic model.

From: R. Muca, W. Piątkowski, D. Antos, J. Chromatogr. A, 1216 (2009), 8712–8721 [74].

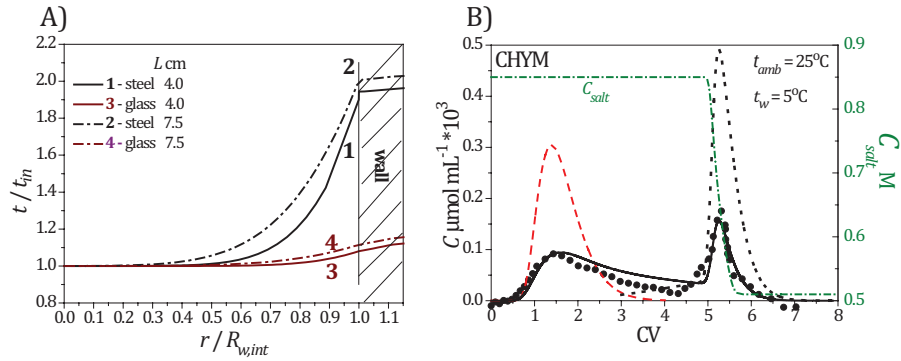


FIGURE 1.19 Illustration of radial temperature distribution in an HIC column with I.D. = 16 mm, the wall at t = 5°C, the mobile phase at the column inlet at t = 26°C, Q = 3 mL min^{-1}. A) Simulations of the radial temperature distribution for different column lengths at a distance of 0.75 of the column length, L, from the inlet, B) band profile of chymotrypsinogen (CHYM) recorded at the column outlet L = 5 cm in the salt gradient, solid line: simulated band profile averaged over the column radius, dashed line: band profile at temperature of the column wall, dotted line: band profile at temperature in the column center, dash-dot line: gradient profile.

From: R. Muca, W. Piątkowski, D. Antos, J. Chromatogr. A, 1216 (2009), 6716–6727 [75].

at a temperature 20°C lower than the mobile phase fed into the column. Heterogeneity of temperature distribution was enhanced with increasing column length. This resulted in the splitting of the protein band profile, as illustrated in Figure 1.19B. A part of the protein migrated faster through the column at the retention corresponding to the wall temperature, whereas the remaining part lagged behind and was eluted by the salt gradient at the retention associated with the temperature of the mobile phase. Still, the uniformity of the temperature distribution can be maintained when the rate of heat transfer from the column wall to the environment is much lower than the rate of heat transfer inside the column. This condition is met when heat exchange between the column wall and the environment occurs by natural convection or when the column is insulated with an insulating material protected from the condensation of humidity [75, 98]. The radial temperature profile for the latter case is illustrated in Figure 1.19A.

1.8 SUMMARY

In this chapter, different thermodynamics and kinetic effects underlying protein adsorption on HIC media have been discussed. The focus was on the examination of less recognized phenomena accompanying elution of proteins in HIC, including protein unfolding and cooperative adsorption along with competitive and synergistic effects, as well as on pitfalls associated with non-isocratic protein elution that are triggered by the sample solvent effect, which may occur while processing supersaturated protein solutions or in temperature-mediated separations.

Protein unfolding can induce undesirable effects in the form of multipeak elution. This phenomenon diminishes with increasing protein loading due to the crowding effect. Yet when the unfolded form tends to aggregate, an increase in protein concentration may accelerate the rate of aggregation, which favors aggregate formation.

The presence of other proteins in a multicomponent solution can also trigger the crowding effect, which is enhanced by size exclusion. Adsorption of a smaller protein with better accessibility to the pore volume of adsorbent and with a higher molar concentration may be preferred over its large-molecule competitors.

At high protein loads, preferential or multilayer adsorption can occur, which manifests in continuously increasing upper parts of the isotherm curves. In the case of strong interactions in the adsorbed phase, protein clusters may form irreversible bounded aggregates. However, cluster formation may be kinetically inhibited and therefore negligible in chromatographic elution. In such cases, the protein retention behavior does not reflect the pattern indicated by the isotherm courses.

Peak deformation may also occur in non-isocratic elution when different solvents are used in the protein feed solution and the mobile phase. The presence of the sample solvent alters the protein adsorption behavior, which can be a cause of band broadening or of peak splitting. This effect weakens at small-volume injections, whereas it can be strongly pronounced when a large volume of the sample is injected.

Non-isocratic elution by salt gradient is accompanied by the concentration of protein solution in desorption bands. When the protein concentration exceeds the solubility limits, precipitation can occur. However, if the kinetic rate of protein crystallization is slow, it is possible to elute the protein at a concentration that falls into the metastable zone without the risk of crystallization.

Another cause of pitfalls in HIC is the thermal heterogeneity of the column, which may occur due to improper column thermostatting in temperature-mediated separations. The radial and axial thermal column heterogeneity implicates the distribution of the mobile phase velocity. Moreover, it also alters protein adsorption behavior in HIC, which is strongly temperature-dependent. This has a detrimental influence on column performance. Nevertheless, this effect can be avoided by proper column insulation and eluent pre-heating.

The occurrence of each of the effects mentioned above can trigger a failure of the separation process or at least a reduction in the yield and purity of the target protein, thus destroying the benefits of the HIC technique. Still, while recognized and quantified, these pitfalls can be avoided, and the separation can be realized with high yield, throughput, and product purity.

REFERENCES

[1] S. Hjerten, Some general aspects of hydrophobic interaction chromatography, *J. Chromatogr. A* 87 (1973) 325–331.

[2] S. Pahlman, J. Rosengren, S. Hjerten, Hydrophobic interaction chromatography on uncharged sepharose derivatives: effects of neutral salts on the adsorption of proteins, *J. Chromatogr. A* 131 (1977) 99–108.

[3] J.A. Queiroz, C.T. Tomaz, J.M.S. Cabral, Hydrophobic interaction chromatography of proteins, *J. Biotech*. 87 (2001) 143–159.

[4] W. Melander, C. Horváth, Salt effects on hydrophobic interactions in precipitation and chromatography of proteins: an interpretation of the lyotropic series. *Arch. Biochem. Biophys.* 183 (1977) 200–215.
[5] E. Boschetti, A. Jungbauer, S. Ahuja, *Handbook of bioseparations.* Academic Press, New York, 535 (2000).
[6] A.L. Grilo, M. Mateus, M.R. Aires-Barros, A.M. Azevedo, Monoclonal antibodies production platforms: an opportunity study of a non-protein-a chromatographic platform based on process economics, *Biotechnol. J.* 12 (2017) 1700260.
[7] N. Kateja, D. Kumar, A. Godara, V. Kumar, A.S. Rathore, Integrated chromatographic platform for simultaneous separation of charge variants and aggregates from monoclonal antibody therapeutic products, *Biotechnol. J.* 12 (2017) 1700133.
[8] L.K. Shekhawat, M. Chandak, A.S Rathore, Mechanistic modeling of hydrophobic interaction chromatography for monoclonal antibody purification: process optimization in the quality by design paradigm, *J. Chem. Technol. Biotechnol.* 92 (2017) 2527–2537.
[9] H. Bo, J. Wang, Q. Chen, H. Shen, F. Wu, H. Shao, S. Huang, Using a single hydrophobic-interaction chromatography to purify pharmaceutical-grade supercoiled plasmid DNA from other isoforms, *Pharm. Biol.* 51 (2013) 42–48.
[10] M.M. Diogo, S. Ribeiro, J.A. Queiroz, G.A. Monteiro, P. Perrin, N. Tordo, D.M.F. Prazeres, Scale-up of hydrophobic interaction chromatography for the purification of a DNA vaccine against rabies, *Biotech. Lett.* 22 (2000) 1397–1400.
[11] M.J. Wilson, C.L. Haggart, S.P. Gallagher, D. Walsh, Removal of tightly bound endotoxin from biological products. *J. Biotech.* 88 (2001) 67–75.
[12] D. Schumacher, C.P.R. Hackenberger, H. Leonhardt, J. Helma, Current status: site-specific antibody drug conjugates, *J. Clin. Immunol.* 36 (2016) 100–107.
[13] S. Andris, J. Hubbuch, Modeling of hydrophobic interaction chromatography for the separation of antibody-drug conjugates and its application towards quality by design, *J. Biotech.* 317 (2020) 48–58.
[14] S.Z. Fekete, J.-L. Veuthey, A. Beck, D. Guillarme, Hydrophobic interaction chromatography for the characterization of monoclonal antibodies and related products, *J. Pharm. Biomed. Anal.* 130 (2016) 3–18.
[15] B. Bobály, S. Fleury-Souverain, A. Beck, J.-L. Veuthey, D. Guillarme, S.Z. Fekete, Current possibilities of liquid chromatography for the characterization of antibody-drug conjugates, *J. Pharm. Biomed. Anal.* 147 (2018) 493–505.
[16] L. Luo, B. Jiang, Y. Cao, L. Xu, M. Shameem, D. Liu, A hydrophobic interaction chromatography method suitable for quantitating individual monoclonal antibodies contained in co-formulated drug products, *J. Pharm. Biomed. Anal.* 193 (2021) 113703.
[17] J. Kim, Y.J. Kim, M. Cao, N. De Mel, K. Miller, J.S. Bee, J. Wang, X. Wang, M. Albarghouthi, Analytical characterization of coformulated antibodies as combination therapy, *Mabs.* 12 (2020) e1738691, Taylor & Francis.
[18] J. Mollerup, Applied thermodynamics: a new frontier for biotechnology, *Fluid Ph. Equilibr.* 241 (2006) 205–215.
[19] T.W. Perkins, D.S. Mak, T.W. Root, E.N. Lightfoot, Protein retention in hydrophobic interaction chromatography: modeling variation with buffer ionic strength and column hydrophobicity, *J. Chromatogr. A* 766 (1997)1–14.
[20] X. Geng, L. Guo, J. Chang, Study of the retention mechanism of proteins in hydrophobic interaction chromatography, *J. Chromatogr. A* 507 (1990) 1–23.
[21] J. Chen, Y. Sun, Modeling of the salt effects on hydrophobic adsorption equilibrium of protein, *J. Chromatogr. A* 992 (2003) 29–40.
[22] J. Chen, S.M. Cramer, Protein adsorption isotherm behavior in hydrophobic interaction chromatography, *J. Chromatogr. A* 1165 (2007) 67–77.
[23] J.M. Mollerup, T.B. Hansen, St. Kidal, A. Staby, Quality by design–thermodynamic modelling of chromatographic separation of proteins, *J. Chromatogr. A* 1177 (2008) 200–206.

[24] R. Deitcher, J. Rome, P. Gildea, J. O'Connel, E. Fernandez, A new thermodynamic model describes the effects of ligand density and type, salt concentration and protein species in hydrophobic interaction chromatography, *J. Chromatogr. A* 1217 (2010) 199–208.
[25] M.R. Mirani, F. Rahimpour, Thermodynamic modelling of hydrophobic interaction chromatography of biomolecules in the presence of salt, *J. Chromatogr. A* 1422 (2015) 170–177.
[26] G. Wang, T. Hahn, J. Hubbuch, Water on hydrophobic surfaces: mechanistic modeling of hydrophobic interaction chromatography, *J. Chromatogr. A* 1465 (2016) 71–78.
[27] R.D. Cramer III, "Hydrophobic interaction" and solvation energies: discrepancies between theory and experimental data, *J. Am. Chem. Soc.* 99 (1977) 5408–5412.
[28] E.E. Meyer, K.J. Rosenberg, J. Israelachvili, Recent progress in understanding hydrophobic interactions, *PNAS* 103 (2006) 15739–15746.
[29] R. Muca, M. Kołodziej, W. Piątkowski, G. Carta, D. Antos, Effects of negative and positive cooperative adsorption of proteins on hydrophobic interaction chromatography media, *J. Chromatogr. A* 1625 (2020) 461309.
[30] R. Hahn, K. Deinhofer, C. Machold, A. Jungbauer, Hydrophobic interaction chromatography of proteins: II. Binding capacity, recovery and mass transfer properties, *J. Chromatogr. B* 790 (2003) 99–114.
[31] B.C.S. To, A.M. Lenhoff, Hydrophobic interaction chromatography of proteins: III. Transport and kinetic parameters in isocratic elution, *J. Chromatogr. A* 1205 (2008) 46.
[32] B.C.S. To, A.M. Lenhoff, Hydrophobic interaction chromatography of proteins. IV. Protein adsorption capacity and transport in preparative mode, *J. Chromatogr. A* 1218 (2011) 427–440.
[33] S.L. Wu, K. Benedek, B.L. Karger, Thermal behavior of proteins in high-performance hydrophobic-interaction chromatography. On-line spectroscopic and chromatographic characterization, *J. Chromatogr. A* 359 (1986) 3–17.
[34] T.T. Jones, E.J. Fernandez, α-Lactalbumin tertiary structure changes on hydrophobic interaction chromatography surfaces, *J. Colloid Interface Sci.* 259 (2003) 27–35.
[35] A. Jungbauer, C. Machold, R. Hahn, Hydrophobic interaction chromatography of proteins: III. Unfolding of proteins upon adsorption, *J. Chromatogr. A* 1079 (2005) 221–228.
[36] E. Haimer, A. Tscheliessnig, R. Hahn, A. Jungbauer, Hydrophobic interaction chromatography of proteins IV. Kinetics of protein spreading, *J. Chromatogr. A* 1139 (2007) 84–94.
[37] A. Rodler, R. Ueberbacher, B. Beyer, A. Jungbauer, Calorimetry for studying the adsorption of proteins in hydrophobic interaction chromatography, *Prep. Biochem. Biotechnol.* 49 (2018) 1–20.
[38] P. Oroszlan, R. Blanco, X.M. Lu, D. Yarmush, B.L. Karger, Intrinsic fluorescence studies of the kinetic mechanism of unfolding of α-lactalbumin on weakly hydrophobic chromatographic surfaces, *J. Chromatogr. A* 500 (1990) 481–502.
[39] M.F.M. Engel, C.P.M. van Mierlo, A.J.W.G. Visser, Kinetic and structural characterization of adsorption-induced unfolding of bovine α-lactalbumin, *J. Biol. Chem.* 277 (2002) 10922–10930.
[40] N.J. Greenfield, Using circular dichroism collected as a function of temperature to determine the thermodynamics of protein unfolding and binding interactions, *Nat. Protoc.* 1 (2006) 2527–2535.
[41] R. Ueberbacher, E. Haimer, R. Hahn, A. Jungbauer, Hydrophobic interaction chromatography of proteins V. Quantitative assessment of conformational changes, *J. Chromatogr. A* 1198–1199 (2008) 154–163.
[42] W.Y. Chen, H.M. Huang, C.C. Lin, F.Y. Lin, Y.C. Chan, Effect of temperature on hydrophobic interaction between proteins and hydrophobic adsorbents: studies by isothermal titration calorimetry and the van't Hoff equation, *Langmuir* 19 (2003) 9395–9403.

[43] R. Ueberbacher, A. Rodler, R. Hahn, A. Jungbauer, Hydrophobic interaction chromatography of proteins: thermodynamic analysis of conformational changes, *J. Chromatogr. A* 1217 (2010) 184–190.
[44] Z.Q. Zhang, D.L. Smith, Determination of amide hydrogen exchange by mass spectrometry: a new tool for protein structure elucidation, *Protein Sci*. 2 (1993) 522–531.
[45] Y.W. Bai, T.R. Sosnick, L. Mayne, S.W. Englander, Protein folding intermediates: native-state hydrogen exchange, *Science* 5221 (1995) 192–197.
[46] J. Buijs, C.C. Vera, E. Ayala, E. Steensma, P. Hakansson, S. Oscarsson, Conformational stability of adsorbed insulin studied with mass spectrometry and hydrogen exchange, *Anal. Chem*. 71 (1999) 3219–3225.
[47] J.L. McNay, E.J. Fernandez, How does a protein unfold on a reversed-phase liquid chromatography surface? *J. Chromatogr. A* 84 9 (1999) 135–148.
[48] Y. Xiao, T. Tibbs-Jones, A.H. Laurent, J.P. O'Connell, T.M. Przybycien, E.J. Fernandez, Protein instability during HIC: hydrogen exchange labeling analysis and a framework for describing mobile and stationary phase effects, *Biotechnol. Bioeng.* 96 (2007) 80–93.
[49] R. Muca, W. Marek, M. Żurawski, W. Piątkowski, D. Antos, Effect of mass over-loading on binding and elution of unstable proteins in hydrophobic interaction chromatography, *J. Chromatogr. A* 1492 (2017) 79–88.
[50] G. Senisterra, I. Chau, M. Vedadi, Thermal denaturation assays in chemical biology, *Assay Drug Dev. Technol*. 10 (2012) 128–136.
[51] A. Stańczak, K. Baran, D. Antos, A high-throughput method for fast detecting unfolding of monoclonal antibodies on cation exchange resins, *J. Chromatogr. A* 1634 (2020) 461688.
[52] Y. Xiao, A.S. Freed, T. Tibbs-Jones, K. Makrodimitris, J.P. O'Connell, E.J. Fernandez, Protein instability during HIC: describing the effects of mobile phase conditions on instability and chromatographic retention, *Biotechnol. Bioeng.* 93 (2006) 1177–1189.
[53] R.W. Deitcher, Y. Xiao, J.P. O'Connell, E.J. Fernandez, Protein instability during HIC: evidence of unfolding reversibility, and apparent adsorption strength of disulfide bond-reduced α-lactalbumin variants, *Biotechnol. Bioeng.* 102 (2009) 1416–1427.
[54] I. Lundström, Models of protein adsorption on solid surfaces, in: B. Lindman, G. Olofsson, P. Stenius (Eds.), *Surfactants, Adsorption, Surface Spectroscopy and Disperse Systems*, Steinkopff, Darmstadt, Germany, 1985.
[55] R. Muca, W.K. Marek, W. Piątkowski, D. Antos, Influence of the sample-solvent on protein retention, mass transfer and unfolding kinetics in hydrophobic interaction chromatography, *J. Chromatogr. A* 1217 (2010) 2812–2820.
[56] W. Marek, R. Muca, W. Piątkowski, D. Antos, Multiple-injection technique for isolating a target protein from multicomponent mixtures, *J. Chromatogr. A* 1218 (2011) 5423–5433.
[57] W.K. Marek, R. Muca, S. Woś, W. Piątkowski, D. Antos, Isolation of monoclonal antibody from a CHO supernatant. II. Dynamics of the integrated separation on IEC and HIC column, *J. Chromatogr. A* 1305 (2013) 64–75.
[58] R. Bochenek, W. Marek, W. Piątkowski, D. Antos, Evaluating the performance of different multicolumn setups for chromatographic separation of proteins on hydrophobic interaction chromatography media by a numerical study, *J. Chromatogr. A* 1301 (2013) 60–72.
[59] J.L. Fogle, J.P. O'Connell, E.J. Fernandez, Loading, stationary phase, and salt effects during hydrophobic interaction chromatography: α-Lactalbumin is stabilized at high loadings, *J. Chromatogr. A* 1121 (2006) 209–218.
[60] R.C. Chatelier, A.P. Minton, Adsorption of globular proteins on locally planar surfaces: models for the effect of excluded surface area and aggregation of adsorbed protein on adsorption equilibria, *Biophys. J.* 71 (1996) 2367–2374.

[61] A.P. Minton, Effects of excluded surface area and adsorbate clustering on surface adsorption of proteins. II. Kinetic models, *Biophys. J.* 80 (2001) 1641–1648.
[62] M. Rabe, D. Verdes, S. Seeger, Understanding protein adsorption phenomena at solid surfaces, *Adv. Colloid Interface Sci.* 162 (2011) 87–106.
[63] T. Arakawa, S.N. Timasheff, Preferential interactions of proteins with salts in concentrated solutions, *Biochemistry* 21 (1982) 6545–6552.
[64] W.R. Melander, D. Corradini, C. Horvath, Salt-mediated retention of proteins in hydrophobic-interaction chromatography: application of solvophobic theory, *J. Chromatogr. A* 317 (1984) 67–85.
[65] A. Katti, Y.F. Maa, C. Horvath, Protein surface-area and retention in hydrophobic interaction chromatography, *Chromatographia* 24 (1987) 646–650.
[66] W.R. Melander, Z. El Rassi, C. Horvath, Interplay of hydrophobic and electrostatic interactions in biopolymer chromatography: effect of salts on the retention of proteins, *J. Chromatogr.* 469 (1989) 3–27.
[67] A. Vailaya, C. Horvath, Retention thermodynamics in hydrophobic interaction chromatography, *Ind. Eng. Chem. Res.* 35 (1996) 2964–2981.
[68] Z. El Rassi, Recent progress in reversed-phase and hydrophobic interaction chromatography of carbohydrate species, *J. Chromatogr. A* 720 (1996) 93–118.
[69] A.C. Dias-Cabral, J.A. Queiroz, N.G. Pinto, Effect of salts and temperature on the adsorption of bovine serum albumin on polypropylene glycol-sepharose under linear and overloaded chromatographic conditions, *J. Chromatogr. A* 1018 (2003) 137–153.
[70] R. Bonomo, L. Minim, J. Coimbra, R. Fontan, L. Mendesdasilva, V. Minim, Hydrophobic interaction adsorption of whey proteins: effect of temperature and salt concentration and thermodynamic analysis, *J. Chromatogr. B.* 844 (2006) 6–14.
[71] A.C. Dias-Cabral, N.G. Pinto, J.A. Queiroz, Studies on hydrophobic interaction adsorption of bovine serum albumin on polypropylene glycol-sepharose under overloaded conditions, *Sep. Sci. Technol.* 37 (2002) 1505–1520.
[72] A.C. Dias-Cabral, A.S. Ferreira, J. Phillips, J.A. Queiroz, N.G. Pinto, The effects of ligand chain length, salt concentration and temperature on the adsorption of bovine serum albumin onto polypropyleneglycol–Sepharose, *Biomed. Chromatogr.* 19 (2005) 606–616.
[73] G.E. Rowe, H. Aomari, T. Chevaldina, M. Lafrance, S. St-Arnaud, Thermodynamics of hydrophobic interaction chromatography of acetyl amino acid methyl esters, *J. Chromatogr. A* 1177 (2008) 243–253.
[74] R. Muca, W. Piątkowski, D. Antos, Altering efficiency of hydrophobic interaction chromatography by combined salt and temperature effects, *J. Chromatogr. A* 1216 (2009) 8712–8721.
[75] R. Muca, W. Marek, W. Piątkowski, D. Antos, Effects of thermal heterogeneity in hydrophobic interaction chromatography, *J. Chromatogr. A* 1216 (2009) 6716–6727.
[76] G. Guiochon, A. Felinger, D.G. Shirazi, *Fundamentals of Preparative and Nonlinear Chromatography*, Academic Press, Amsterdam, The Netherlands, 2006.
[77] W.K. Marek, D. Sauer, A. Dürauer, A. Jungbauer, W. Piątkowski, D. Antos, Prediction tool for loading, isocratic elution, gradient elution and scaling up of ion exchange chromatography of proteins, *J. Chromatogr. A* 1566 (2018) 89–101.
[78] W.K. Marek, W. Piątkowski, D. Antos, Retention behavior of polyethylene glycol and its influence on protein elution on hydrophobic interaction chromatography media, *Chromatographia* 81 (2018) 1641–1648.
[79] I. Poplewska, W. Piątkowski, D. Antos, Overcoming solubility limits in overloaded gradient hydrophobic interaction chromatography, *J. Chromatogr. A* 1386 (2015) 1–12.
[80] J. Chien, St. Y. Sun, Modeling of the salt effects on hydrophobic adsorption equilibrium of protein, *J. Chromatogr. A* 992 (2003) 29–40.

[81] J. Chien, St. M. Cramer, Protein behavior in hydrophobic interaction chromatography, *J. Chromatogr. A* 1165 (2007) 67–77.
[82] Q. Meng, J. Wang, G. Ma, Z. Su, Isotherm type shift of hydrophobic interaction adsorption and its effect on chromatographic behavior, *J. Chromatogr. Sci*. 51 (2012) 173–180.
[83] P. Jandera, G. Guiochon, Effect of the sample solvent on band profiles in preparative liquid chromatography using non-aqueous reversed-phase high-performance liquid chromatography, *J. Chromatogr. A* 588 (1991) 1–10.
[84] T. Fornstedt, G. Guiochon, Comparison between experimental and theoretical profiles of high concentration elution bands and large system peaks in nonlinear chromatography, *Anal. Chem*. 66 (1994) 2686–2693.
[85] K. Gedicke, M. Tomusiak, D. Antos, A. Seidel-Morgenstern, Analysis of applying different solvents for the mobile phase and for sample injection, *J. Chromatogr. A* 1092 (2005) 142–148.
[86] G. Ströhlein, M. Morbidelli, H.K. Rhee, M. Mazzotti, Modeling of modifier-solute peak interactions in chromatography, *AIChE J*. 52 (2006) 565–573.
[87] G. Ströhlein, L. Aumann, L. Melter, K. Buescher, B. Schenkel, M. Mazzotti, M. Morbidelli, Experimental verification of sample-solvent induced modifier—solute peak interactions in biochromatography, *J. Chromatogr. A* 1117 (2006) 146–153.
[88] S. Keunchkarian, M. Reta, L. Romero, C. Castells, Effect of sample solvent on the chromatographic peak shape of analytes eluted under reversed-phase liquid chromatographic condition, *J. Chromatogr. A* 119 (2006) 20–28.
[89] D. Antos, W. Piątkowski, Band deformation in non-isocratic liquid chromatography, *TrAC* 81 (2016) 69–78.
[90] C. Rédei, A. Felinger, Modeling the competitive adsorption of sample solvent and solute in supercritical fluid chromatography, *J. Chromatogr. A* 1603 (2019) 348–354.
[91] H. Poppe, J.C. Kraak, Influence of thermal conditions on the efficiency of high-performance liquid chromatographic columns, *J. Chromatogr. A* 282 (1983) 399–412.
[92] R.G. Wolcott, J.W. Dolan, L.R. Snyder, R. Bakalyar, M.A. Arnold, J.A. Nichols, Control of column temperature in reversed-phase liquid chromatography, *J. Chromatogr. A* 869 (2000) 211–230.
[93] J.D. Thompson, J.S. Brown, P.W. Carr, Dependence of thermal mismatch broadening on column diameter in high-speed liquid chromatography at elevated temperatures, *Anal. Chem*. 73 (2001) 3340–3347.
[94] G.B. Dapremont, G.B. Cox, M. Martin, P. Hilaireau, H. Colin, Effect of radial gradient of temperature on the performance of large-diameter high performance liquid chromatography columns. I. Analytical conditions, *J. Chromatogr. A* 796 (1998) 81–99.
[95] K. Kaczmarski, F. Gritti, G. Guiochon, Prediction of the influence of the heat generated by viscous friction on the efficiency of chromatography columns, *J. Chromatogr. A* 1177 (2008) 92–104.
[96] K. Kaczmarski, F. Gritti, J. Kostka, G. Guiochon, Modeling of thermal processes in high pressure liquid chromatography. II—thermal heterogeneity at very high pressures, *J. Chromatogr. A* 1216 (2009) 6575–6586.
[97] S. Fekete, J. Fekete, D. Guillarme, Estimation of the effects of longitudinal temperature gradients caused by frictional heating on the solute retention using fully porous and superficially porous sub-2 μm materials, *J. Chromatogr. A* 1359 (2014) 124–130.
[98] J. Kostka, F. Gritti, G. Guiochon, K. Kaczmarski, Modeling of thermal processes in very high pressure liquid chromatography for column immersed in a water bath: application of the selected models, *J. Chromatogr. A* 1217 (2010) 4704–4712.
[99] M.M. Fallas, M.R. Hadley, D.V. McCalley, Practical assessment of frictional heating effects and thermostat design on the performance of conventional (3 μm and 5 μm) columns in reversed-phase high-performance liquid chromatography, *J. Chromatogr. A* 1216 (2009) 3961–3969.

[100] D. Åsberg, J. Samuelsson, M. Leśko, A. Cavazzini, K. Kaczmarski, T. Fornstedt, Method transfer from high-pressure liquid chromatography to ultra-high-pressure liquid chromatography. II. Temperature and pressure effects, *J. Chromatogr. A* 1401 (2015) 52–59.

[101] K. Broeckhoven, J. Billen, M. Verstraeten, K. Choikhet, M. Dittmann, G. Rozing, G. Desmet, Towards a solution for viscous heating in ultra-high pressure liquid chromatography using intermediate cooling, *J. Chromatogr. A* 1217 (2010) 2022–2031.

[102] D. Åsberg, M. Chutkowski, M. Leśko, J. Samuelsson, K. Kaczmarski, T. Fornstedt, A practical approach for predicting retention time shifts due to pressure and temperature gradients in ultra-high-pressure liquid chromatography, *J. Chromatogr. A* 1479 (2017) 107–120.

2 A Kinetic Model of Evaporation Based on Gas Chromatographic Retention Index

Environmental and Forensic Applications

Victoria L. McGuffin, Ruth Waddell Smith

CONTENTS

DOI: 10.1201/9781003330080-2

2.1 INTRODUCTION

Evaporation is a fundamental, natural process that has an impact in many diverse areas of science and engineering. For example, in environmental engineering, evaporation is a dominant weathering process that can influence compound distribution following a petroleum or other chemical waste spill [1]. In forensic science, evaporative residues of ignitable liquids such as gasoline may indicate that a fire is intentional rather than accidental [2]. Evaporation also has significance in some manufacturing and industry sectors. For example, flavor and aroma volatiles, whether natural or synthetic, are important in foods and beverages as well as perfumes and fragrances [3]. In addition, organic solvent residues may be present after manufacture, posing a health hazard in pharmaceuticals, supplements, and related products [4]. Finally, in homeland security and law enforcement, evaporation can influence the dispersal of chemical warfare agents, explosives, and their degradation products [5–8]. In these and many other areas, a thorough and comprehensive understanding of the evaporation process is necessary.

Although the evaporation rate can be experimentally measured, it is often time consuming and may be hazardous with the risk of fire, explosion, and exposure to toxic vapors. It is more practical to model the evaporation process based on physical properties, such as boiling point, vapor pressure, or rate constant. Many models of this kind have been developed, particularly for environmental applications. These models vary in their foundation, whether empirically or theoretically derived, and whether based in thermodynamics, kinetics, or mass-transfer theory.

Fingas developed an empirical approach to predict the percent evaporation of a sample as a function of time and temperature [9, 10]. Experimental measurements of the change in mass of the sample were fit to either a logarithmic or square-root function of time using nonlinear regression. This approach was applied to crude oils and refined petroleum fuels at various temperatures, for which extensive information

is available in the literature [11]. However, to apply these models, the identity and source of the fuel must be known and the regression coefficients for that fuel must be available. Fingas also developed two general equations, one with logarithmic and one with square-root dependence on time, that are based on the weight percent of the fuel distilled at 180°C [12]. Although these general equations do not require that the specific fuel source be known, they do require the relevant distillation data and, hence, may require a sample of the fuel.

The analytical models use a mass-transfer approach, often based in thermodynamics, combined with experimental measurements. Mackay and Matsugu's model predicts the evaporation rate, or change in volume with time, based on the vapor pressure and temperature of the sample [13]. For pure liquid samples, such as water and cumene, acceptable agreement was observed between the predicted and experimentally determined evaporation rates. However, for gasoline, the predicted evaporation rate was notably faster than the experimental rate [13].

In the empirical models described above, the fuel is considered as a single component with physical and chemical properties that are a fixed average of all constituents [9, 10, 12, 13]. However, because fuels such as gasoline are complex mixtures in which the properties change as a function of evaporation level, single-component models are prone to prediction errors [14–16]. Other modeling approaches have been used in which the fuel is considered to contain multiple components, whether pure constituents or mixtures of constituents with similar properties (called pseudo-components).

Stiver and Mackay adapted the analytical model to predict the evaporation rate of individual compounds in complex mixtures using a "synthetic oil" composed of normal (*n*-) alkanes [17]. The *n*-alkane mixture was experimentally evaporated using three different methods: tray evaporation, gas stripping, and distillation. The first two methods are isothermal, where tray evaporation is mass-transfer limited and gas stripping is equilibrium controlled. In contrast, distillation is isobaric and non-isothermal, measuring volume distilled as a function of boiling temperature. To predict the evaporation of individual compounds, the net vapor pressure in the analytical model was substituted with the partial vapor pressure, calculated according to Raoult's law [17, 18]. This model was applied to predict evaporation of the *n*-alkane mixture, with generally good agreement between the predicted and experimental evaporation rates for tray evaporation and gas stripping methods. For the distillation method, however, discrepancies were observed between the predicted distillation curve and boiling point data for *n*-alkanes from the literature. This discrepancy was attributed to the design of the still, which was found to have more than one theoretical plate. After correcting for this error, the model was then applied to crude oil samples, with relevant vapor pressure or boiling point obtained from distillation curves, with satisfactory results [17].

The pseudo-component model, which is based on the analytical model of Stiver and Mackay, is the most common method currently used to estimate the extent of fuel evaporation. This model approximates the composition of a complex fuel as several discrete and independent components whose properties are derived from distillation data. The total evaporation of the fuel is based on the sum of the evaporation of the pseudo-components. This allows for a more accurate determination of vapor

pressure and molar volume, but requires additional empirical data and assumptions to implement the model [14–16, 19, 20].

Many of the models described above, especially those based in mass-transfer theory, require a distillation curve to determine the physicochemical and transport properties of the fuel or its pseudo-components. Using traditional distillation methods, there is considerable uncertainty in the estimates of these properties. More recently, Bruno et al. developed the advanced distillation curve (ADC) apparatus and method [21]. In this approach, the liquid- and vapor-phase temperatures, distillate volume, and other parameters are continuously and accurately measured. Trace chemical analysis is performed on each distillate fraction using any suitable analytical technique, such as gas chromatography–mass spectrometry (GC-MS). Using the ADC method, more accurate and thermodynamically meaningful estimates of properties are obtained for a wide range of fuels [21–23].

Jackson and co-workers employed a more classical thermodynamic approach to predict evaporation of a simulated gasoline mixture of 7—9 components, including *n*-alkanes, alkyl benzenes, and naphthalenes [24, 25]. Vapor pressures for each component were calculated using the Antoine equation [26], and the partial and total pressures were calculated using Raoult's law and Dalton's law, respectively [27]. Evaporation was simulated by mathematically removing a small fraction of the vapor phase in a step-wise manner, recalculating the partial and total pressures at each step. Overall, there was good agreement between the predicted and experimental mole fractions of each component remaining in the liquid phase [24, 25]. Although this approach shows promise, at present it can only be applied to a small number of components. It is not yet practical for more complex mixtures, as the identity and mole fraction of all components must be known and their Antoine coefficients must be available in the literature.

Regnier and Scott developed a kinetic model to predict evaporation of crude oil based on the composition of *n*-alkanes [28]. A regression equation was established between the calculated vapor pressure (thermodynamic property) and the measured evaporation rate constant (kinetic property) of the selected *n*-alkanes. From this equation, the rate constant can be predicted for any component if the vapor pressure is known. The predicted rate constants, together with the initial concentrations, enable calculation of the fraction remaining of the *n*-alkanes at given time intervals. The sum of these fractions for the *n*-alkanes agreed well with the fraction remaining of the total crude oil at each time interval and at temperatures of 5, 10, 20, and 30°C. However, practical application of this model requires detailed knowledge of the chemical composition of the crude oil [28].

Okamoto et al. used a similar regression approach to predict the amount of gasoline vapor generated after a spill [29]. Gasoline samples were evaporated at room temperature from 0—70% by mass, in increments of 10%, and the vapor pressure and evaporation rates were determined experimentally. The authors demonstrated that the vapor pressure and evaporation rate were exponentially related to the mass of gasoline lost at constant temperature. In a later study, Okamoto et al. applied similar principles to predict evaporation and diffusion of mixtures of gasoline and kerosene [30]. Both of these studies considered the fuel as a single component rather than a mixture of the individual components. Accordingly, the vapor pressure could not be calculated from standard equations, as in the work of Regnier and Scott [28], but was obtained from experimental measurements.

Although the models described above were developed specifically to predict evaporation of petroleum fuels, in principle, they can be adapted for other samples and other applications. However, there are inherent problems that limit the practical utility and accuracy of each model. The primary difficulty relates to the physicochemical properties (vapor pressure, boiling point, rate constant, etc.) that are needed as input parameters. For models that treat the sample as a single component, these properties are usually measured as a bulk value from the original sample. However, as the sample evaporates, the bulk value of the property changes, and the presumed time dependence (often linear, logarithmic, or square root) leads to uncertainty and inaccuracy. For models that treat the sample as individual components (or pseudocomponents), each component must be identified and its properties must be known, predicted, or measured. This limits the number of components that can realistically be accommodated for complex samples.

To overcome these limitations, the ideal model should be capable of characterizing the evaporation of every component, even in highly complex samples, without implicitly knowing the chemical composition or the properties of the sample. To achieve this goal, McIlroy et al. [31, 32] developed a classical kinetic model in which the evaporation rate constants are empirically related to a surrogate property, the gas chromatographic retention index. In gas chromatography (GC), the separation is based directly on boiling point or vapor pressure when using a nonpolar stationary phase such as polydimethylsiloxane. The high resolution of gas chromatography allows the separation of many, if not all, of the components, which, in turn, allows the retention index to be accurately determined and the evaporation rate constants to be calculated. Finally, the kinetic foundation provides a theoretically established relationship to time and temperature. In this chapter, the fundamental basis of this model is described and its validation is demonstrated with emphasis on environmental and forensic applications.

2.2 THEORY

2.2.1 Kinetic Model of Evaporation

In the irreversible kinetic model of evaporation, the system is assumed to be fully open, and compound X is transferred from the liquid phase (L) to the gas phase (G)

$$X_L \xrightarrow{k} X_G \tag{2.1}$$

where k is the rate constant for evaporation. The rate law, or rate of change in the concentration of X in the liquid phase $\left[X_L\right]$ as a function of time (t), is given by

$$\frac{-d\left[X_L\right]}{d\,t} = k\left[X_L\right] \tag{2.2}$$

Upon separation of variables and integration,

$$\frac{\left[X_L\right]_t}{\left[X_L\right]_0} = \exp\left(-k\,t\right) \tag{2.3}$$

which is the integrated rate law describing evaporation in an irreversible system.

The rate constants were experimentally determined by McIlroy et al. by evaporating thin films of diesel fuel in a chamber with controlled temperature and then analyzing the samples by GC-MS with a nonpolar stationary phase [31, 32]. Approximately 78 compounds were selected for model development and 28 compounds for model validation, including normal alkanes, branched alkanes, cyclic alkanes, alkyl benzenes, and polycyclic aromatic compounds. To illustrate the characteristic decay curves predicted by Equation 2.3, the fraction remaining in the liquid phase ($F = [X_L]_t / [X_L]_0$) is shown as a function of evaporation time in Figure 2.1 for representative *n*-alkanes at 20°C. For *n*-octane (Figure 2.1A), with

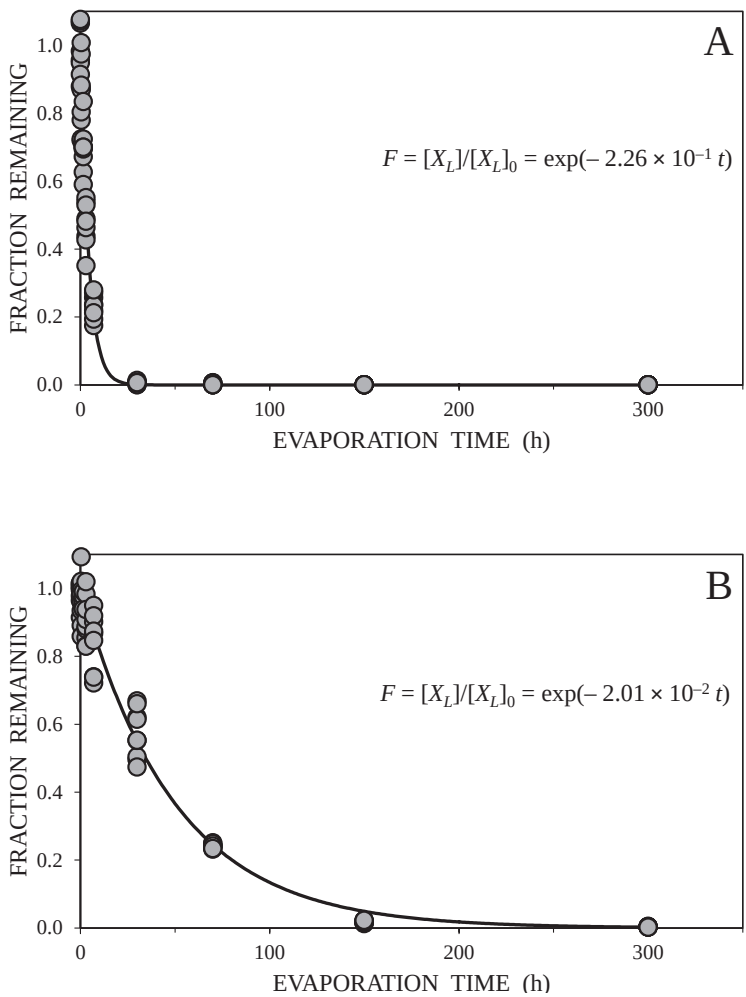


FIGURE 2.1 Fraction remaining in the liquid phase versus evaporation time for (A) *n*-octane, (B) *n*-decane, (C) *n*-dodecane, and (D) *n*-tetradecane at 20°C [31].

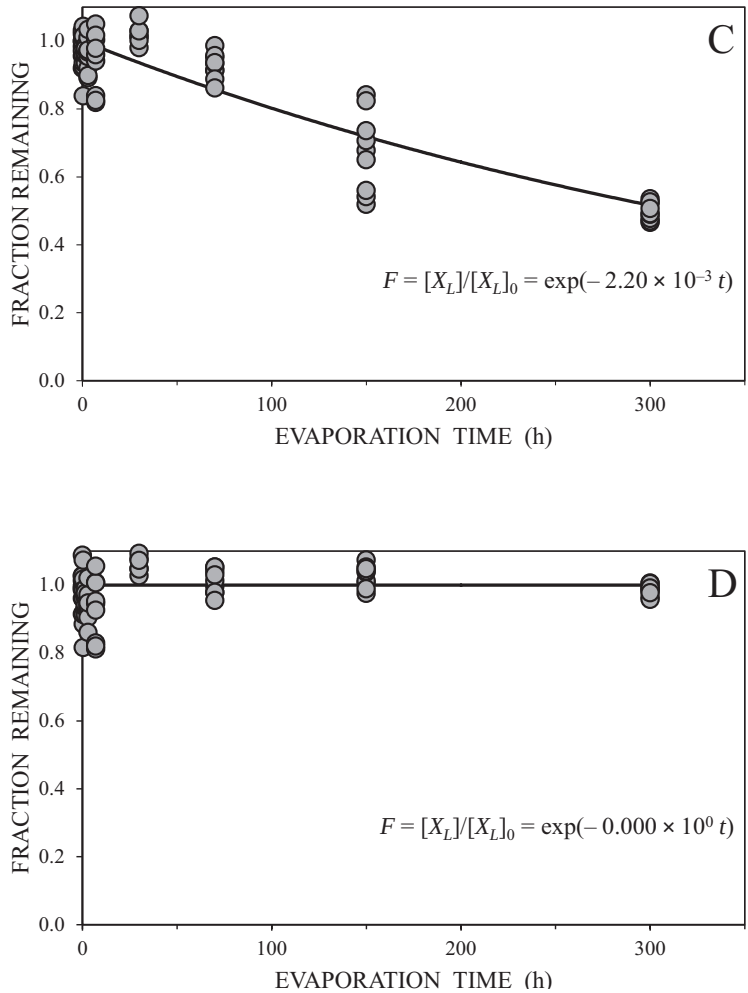


FIGURE 2.1 (Continued)

the highest rate constant (k = 2.26 x 10^{-1} h^{-1}), the fraction remaining decays in a characteristic time $5\tau = 5/k$ = 22 h. At this time, the fraction remaining is $F = \exp(-5) = 0.0067$, which corresponds to 99.33% evaporated (i.e., nearly complete). The rate constant for each *n*-alkane systematically decreases with increasing carbon number: 2.01 x 10^{-2} h^{-1} for *n*-decane (Figure 2.1B), 2.20 x 10^{-3} h^{-1} for *n*-dodecane (Figure 2.1C), and statistically indeterminate for *n*-tetradecane (Figure 2.1D) at 20°C. The characteristic decay time ($5\tau = 5/k$) is approximately 250 h, 2300 h, and indeterminate for *n*-decane, *n*-dodecane, and *n*-tetradecane, respectively. Hence, the addition of each ethylene group (-C_2H_4-) to the *n*-alkane structure causes a 10-fold decrease in the rate constant and a 10-fold increase in the time required for complete evaporation.

The initial work by McIlroy et al. provided evaporation rate constants for compounds in the range from *n*-octane to *n*-tetradecane [31, 32]. However, many samples of environmental and forensic interest contain more volatile compounds whose rate constants exceed this range. To address this issue, Burkhart et al. measured evaporation rate constants for 11 additional compounds in the range from *n*-pentane to *n*-octane [33]. Because of the high volatility of these compounds, their vapor accumulated in the headspace of the evaporation chamber, changing the kinetics from an irreversible first-order reaction given by Equation 2.1 to a reversible reaction given by Equation 2.4

$$X_L \underset{k'}{\overset{k}{\rightleftarrows}} X_G \xrightarrow{k_e} X_e \tag{2.4}$$

where k and k' are the rate constants for evaporation and condensation, respectively. In addition to the liquid–vapor equilibrium, the compound in the vapor phase can exit (X_e) from the system by means of air flow, with a rate constant of k_e. When corrected for condensation, the evaporation rate constants for representative *n*-alkanes at 20°C are as follows: 18.6 h^{-1} for *n*-pentane, 3.86 h^{-1} for *n*-hexane, 0.823 h^{-1} for *n*-heptane, and 0.240 h^{-1} for *n*-octane. The characteristic decay times ($5\tau = 5/k$) are approximately 0.27, 1.3, 6.1, and 21 h for *n*-pentane to *n*-octane, respectively. Again, the addition of each ethylene group (-C_2H_4-) to the *n*-alkane structure causes roughly a 10-fold decrease in rate constant and a 10-fold increase in the time required for complete evaporation. Because these data were collected by a different investigator at a different time, they are validated separately in the discussion below.

2.2.2 Retention Index as a Surrogate for Evaporation Rate Constant

In Equation 2.3, the fraction remaining in the liquid phase is calculated from the rate constant for evaporation. Although this approach is theoretically correct and accurate, it requires that all compounds be identified and their relevant properties be known, experimentally determined, or predicted. To overcome this limitation, a surrogate property may be used that is closely related to the rate constant. The gas chromatographic retention index is uniquely well suited for this purpose. When compounds are separated on a nonpolar stationary phase, such as 100% polydimethylsiloxane, they elute in order of increasing boiling point. The retention index under temperature-programmed conditions (I^T) is calculated from the retention time for the compound of interest ($t^T_{R,i}$) and the retention times for the *n*-alkanes of carbon number z that elute immediately before ($t^T_{R,z}$) and after ($t^T_{R,z+1}$).

$$I^T = 100\left[z + \frac{t^T_{R,i} - t^T_{R,z}}{t^T_{R,z+1} - t^T_{R,z}} \right] \tag{2.5}$$

The retention index is independent of GC parameters, such as column length and diameter, stationary phase thickness, flow rate, and temperature program. Thus, it is more broadly applicable than retention time or retention factor, and more

reproducible with different instruments and in different laboratories. Moreover, it is not necessary to know the identity of each compound, only its retention time within the alkane ladder, in order to calculate the retention index.

2.2.2.1 Fixed-Temperature Models

The efficacy of retention index as a surrogate for rate constant is demonstrated in Figure 2.2A for the data collected by McIlroy et al. for various compound classes at

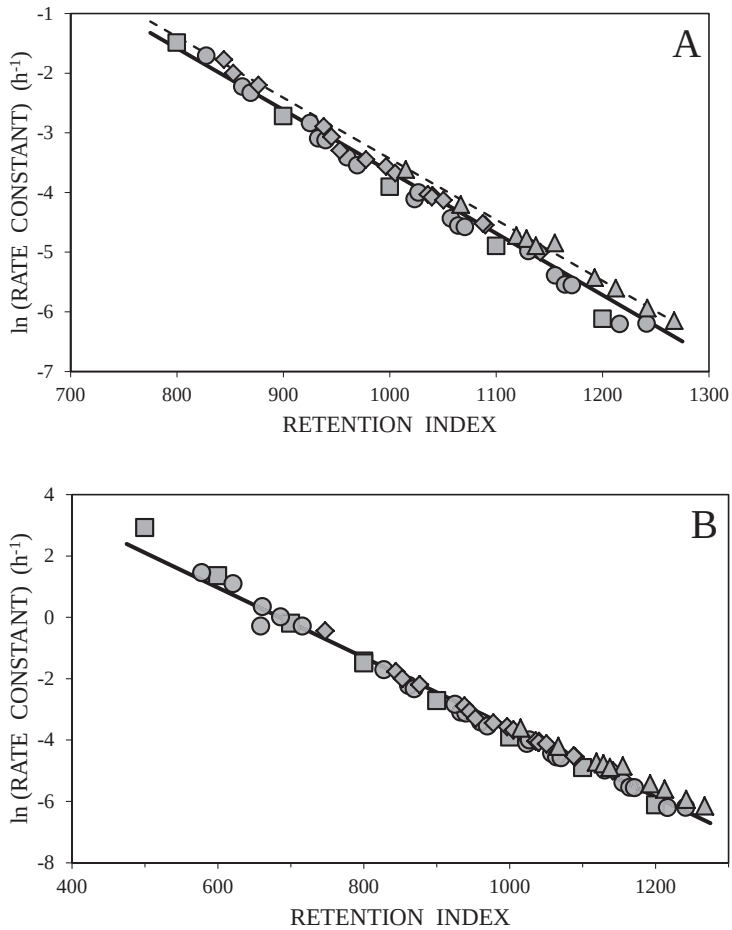


FIGURE 2.2 Logarithm of the rate constant versus retention index for *n*-alkanes (■), branched and cyclic alkanes (●), alkyl benzenes (♦), and polycyclic aromatic compounds (▲) at 20°C. (A) Experimental data of McIlroy et al. [31] with comprehensive fixed-temperature regression model for all compound classes shown as solid line (Equation 2.6, parameters given in Table 2.2) and comprehensive variable-temperature model shown as dashed line (Equation 2.8, parameters given in Table 2.3). (B) Combined experimental data of McIlroy et al. [31] and Burkhart et al. [33] with comprehensive fixed-temperature regression model for all compound classes shown as solid line (Equation 2.6, parameters given in Table 2.2).

TABLE 2.1
Class-specific models developed to predict the evaporation rate constant at 20°C based on retention index (Equation 2.6). For each model, the slope (m_1), intercept (b), and square of the correlation coefficient (R^2) are given, as well as the mean absolute percent error (MAPE) for predicting rate constants for compounds in each class [31].

				MAPE (%) in Rate Constant[a]				
Model Class	m_1	b	R^2	*n - Alkane*	*Branched Alkane*	*Alkyl Benzene*	*Polycyclic Aromatic*	*All Compounds*
n-Alkane	-1.14×10^{-2}	7.61	0.999	4.90	10.2	19.2	42.7	19.0
Branched Alkane	-1.08×10^{-2}	7.05	0.994	12.3	8.21	14.0	32.2	15.2
Alkyl Benzene	-1.08×10^{-2}	7.20	0.992	25.6	17.1	5.70	21.5	15.0
Polycyclic Aromatic	-1.00×10^{-2}	6.47	0.992	43.8	35.6	14.7	4.05	23.3
All Compounds	-1.04×10^{-2}	6.70	0.981	24.3	15.0	6.92	19.7	14.2

[a] MAPE = $\Sigma \mid (k_{pred} - k_{exp})/k_{exp} \mid \times 100/n$, where k_{pred} and k_{exp} are the predicted and experimental rate constants, respectively and n is the number of measurements

20°C [31, 32]. The natural logarithm of the rate constant (ln k) is linearly related to the retention index by

$$\ln k = m_1 I^T + b \tag{2.6}$$

where the best-fit slope (m_1) and intercept (b) are determined by linear regression. For the highest accuracy, separate regression models can be developed for each compound class (Table 2.1). For example, the class-specific model for the *n*-alkanes has slope $m_1 = -1.14 \times 10^{-2}$, intercept $b = 7.61$, and the square of the correlation coefficient $R^2 = 0.999$ [31]. These regression parameters provide the most accurate prediction of rate constants for the *n*-alkanes (4.90% error), but higher error for branched and cyclic alkanes (10.2%), alkyl benzenes (19.2%), and polycyclic aromatic compounds (42.7%). Similarly, the class-specific regression model for alkyl benzenes has lowest error for the alkyl benzenes (5.70%), but higher errors for all other compound classes (17.1—25.6%). While class-specific models are more accurate for compounds within the class, they are less convenient because they require identification of the proper class for each compound of interest. Accordingly, a comprehensive regression model that is suitable for all compound classes is preferable. The comprehensive model for the compounds examined by McIlroy et al. ($m_1 = -1.04 \times 10^{-2}$, $b = 6.70$, $R^2 = 0.981$) has slightly higher error for all classes: *n*-alkanes (24.3%), branched and cyclic alkanes (15.0%), alkyl benzenes (6.92%), and polycyclic aromatic compounds (19.7%) (Table 2.1). However, the overall performance and accuracy of the comprehensive model (14.2% error) is better than that of any class-specific model (15.0—23.0 % error) and is acceptable for practical applications.

The data collected by McIlroy et al. (I^T = 800—1400 [31, 32]) were combined with the data collected by Burkhart et al. (I^T = 500—800 [33]) to extend the range of predicted evaporation rate constants. As shown in Figure 2.2B, the relationship between rate constant and retention index remains linear according to Equation 2.6. However, the regression parameters for the combined data set (m_1 = -1.14 x 10^{-2}, b = 7.79, R^2 = 0.986) are statistically different from those of the original data set (m_1 = -1.04 x 10^{-2}, b = 6.70, R^2 = 0.981) at 20°C. In general, the rate constants for the more volatile compounds of Burkhart et al. lie above the original regression line. Accordingly, the original regression parameters provide the most accurate prediction of rate constants for I^T = 800—1400 (14.2% error), but poorer prediction for I^T = 500—800 (41.3% error), whereas the regression parameters for the combined data set provide more accurate prediction for I^T = 500—800 (24.8% error), but poorer prediction for I^T = 800—1400 (17.5% error) at 20°C. For this reason, the most appropriate regression equation should be chosen based on the retention index range for the samples of interest.

The class-specific and comprehensive models discussed above were developed and validated at 20°C. Similar models were developed at 5, 10, 30, and 35°C by McIlroy et al. [32], and the results for the comprehensive models are summarized in Table 2.2. At all temperatures, the correlation coefficients (R^2 = 0.977—0.989) indicate a good quality of fit to Equation 2.6. It is noteworthy that the slopes and the intercepts show a systematic temperature dependence. As temperature increases, both the slope and intercept increase from -1.09 x 10^{-2} and 6.59 at 5°C to -1.00 x 10^{-2} and 7.70 at 35°C. The

TABLE 2.2

Fixed-temperature models developed to predict the evaporation rate constant based on retention index (Equation 2.6). For each model, the slope (m_1), intercept (b), and square of the correlation coefficient (R^2) are given, as well as the mean absolute percent error (MAPE) for predicting rate constants for all compound classes [32, 33].

	McIlroy Model[a]				Burkhart–McIlroy Model[b]			
Temperature (°C)	m_1	b	R^2	MAPE (%) in Rate Constant[c]	m_1	b	R^2	MAPE (%) in Rate Constant[c]
5	-1.09 × 10^{-2}	6.59	0.977	13.5				
10	-1.04 × 10^{-2}	6.12	0.976	13.8	-1.20 x 10^{-2}	7.86	0.985	20.9
20	-1.04 × 10^{-2}	6.70	0.981	14.2	-1.14 x 10^{-2}	7.79	0.986	18.8
30	-1.02 × 10^{-2}	7.39	0.989	13.1	-1.06 x 10^{-2}	7.83	0.992	13.7
35	-1.00 × 10^{-2}	7.70	0.989	13.5				
Average				13.6				18.4

[a] McIlroy model developed over range I^T = 800—1400, T = 5—35°C

[b] Burkhart–McIlroy model developed over range I^T = 500—1400, T = 10—30°C

[c] MAPE defined in Table 2.1

mean absolute percent error (MAPE) in the predicted rate constant for each fixed-temperature model is 13.1—14.2%, with an average prediction error of 13.6% (Table 2.2).

The more volatile compounds of Burkhart et al. were examined at temperatures of 10, 20, and 30°C [33]. When combined with the data of McIlroy et al. at the same temperatures, the results for the comprehensive models are summarized in Table 2.2. The correlation coefficients for the Burkhart–McIlroy models (R^2 = 0.985—0.992) are generally higher than those for the McIlroy models (R^2 = 0.976—0.989) at the same temperatures. However, the prediction of rate constants is generally more accurate with the McIlroy models than with the Burkhart–McIlroy models, with average errors of 13.6% and 18.4%, respectively.

2.2.2.2 Variable-Temperature Models

The fixed-temperature models described in Section 2.2.2.1 and summarized in Tables 2.1 and 2.2 can accurately predict the rate constants for evaporation [31–33]. However, for environmental and forensic applications, where temperature is rarely constant, it is more convenient and more broadly useful to develop a variable-temperature model. According to the Arrhenius equation [27], the natural logarithm of the rate constant is inversely related to absolute temperature (T),

$$\ln k = \ln A - \frac{E_a}{RT} \tag{2.7}$$

where E_a is the activation energy, R is the gas constant, and A is the pre-exponential factor. By combining Equations 2.6 and 2.7, a variable-temperature model to predict the rate constant can be defined as

$$\ln k = m_1 \, \mathrm{I}^T + \frac{m_2}{T} + b \tag{2.8}$$

where the best-fit slopes (m_1 and m_2) and intercept (b) are determined by multiple linear regression. Using the data of McIlroy et al., the regression coefficients for the comprehensive model are m_1 = -1.02 x 10^{-2}, m_2 = -6147, b = 27.8, and R^2 = 0.975 [32]. The performance of the variable-temperature model is summarized in Table 2.3 at the same temperatures as the fixed-temperature models (5—35°C). In particular, the center column of Table 2.3 contains the retention index range used to develop the McIlroy models (I^T = 800—1400), which can be directly compared. Although the fixed-temperature models are more accurate in predicting rate constants, with errors of 13.1—14.2% (Table 2.2), the variable-temperature model also has acceptable accuracy, with errors of 13.6—27.5% (Table 2.3). There are no noteworthy trends in accuracy with temperature for either the fixed-temperature or the variable-temperature models.

When the data of Burkhart et al. [33] for volatile compounds are combined with those of McIlroy et al. [32], the regression coefficients for the comprehensive model are m_1 = -1.12 x 10^{-2}, m_2 = -6045, b = 28.4, and R^2 = 0.984. Again, there are statistically significant differences in the regression coefficients for the McIlroy model and the Burkhart–McIlroy model, leading to different predictive accuracy.

TABLE 2.3
Variable-temperature models developed to predict the evaporation rate constant based on retention index (Equation 2.8). For each model, the mean absolute percent error (MAPE) for predicting rate constants for all compound classes is given in the retention index range I^T = 500—800, I^T = 800—1400, and I^T = 500—1400 [32, 33].

	MAPE (%) in Rate Constant [a]					
	McIlroy Model [b]			Burkhart–McIlroy Model [c]		
Temperature (°C)	I^T = 500—800	I^T = 800—1400	I^T = 500—1400	I^T = 500—800	I^T = 800—1400	I^T = 500—1400
5		19.9				
10	55.4	17.7	24.9	39.5	17.4	21.6
20	34.2	27.5	28.7	24.8	24.5	24.6
30	17.4	13.6	14.2	24.6	15.7	17.1
35		14.6				
Average	35.7	18.7	22.6	29.6	19.2	21.1

[a] MAPE defined in Table 2.1

[b] McIlroy model: m_1 = -1.02 x 10^{-2}, m_2 = -6147, b = 27.8, and R^2 = 0.975, developed over range I^T = 800—1400, T = 5—35°C

[c] Burkhart–McIlroy model: m_1 = -1.12 x 10^{-2}, m_2 = -6045, b = 28.4, and R^2 = 0.984, developed over range I^T = 500—1400, T = 10—30°C

When compared over the retention index range used to develop the McIlroy model (I^T = 800—1400), the variable-temperature model of McIlroy is slightly more accurate than the Burkhart–McIlroy model, having average errors of 18.7% and 19.2%, respectively (Table 2.3). When compared over the broader retention index range used to develop the Burkhart–McIlroy model (I^T = 500—1400), the variable-temperature Burkhart–McIlroy model is slightly more accurate than the McIlroy model, having average errors of 21.1% and 22.6%, respectively (Table 2.3). Although the predictive accuracy is quite similar over these retention index ranges, there is a significant difference for volatile compounds in the range I^T = 500—800. Herein, the McIlroy model exhibits errors of 17.4—55.4%, with an average error of 35.7%, whereas the Burkhart–McIlroy model has errors of 24.6—39.5%, with an average error of 29.6%. Consequently, the most appropriate variable–temperature model should be chosen based on the retention index range for the samples of interest.

2.2.3 Fraction Remaining and Fraction-Remaining Curves

For the fixed-temperature model, the fraction remaining (F_{I^T}) of an individual compound at retention index I^T can be calculated by substituting Equation 2.6 into Equation 2.3

$$F_{I^T} = \frac{\left[X_L\right]_{I^T,t}}{\left[X_L\right]_{I^T,0}} = \exp\left(-k\,t\right) = \exp\left(-\exp\left(m_1\,I^T + b\right)t\right) \tag{2.9}$$

where $\left[X_L\right]_{I^T,0}$ and $\left[X_L\right]_{I^T,t}$ are the concentrations of compound X in the liquid phase at time 0 (initial) and time t, respectively. Similarly for the variable-temperature model, the fraction remaining of an individual compound can be calculated by substituting Equation 2.8 into Equation 2.3. If the temperature remains constant, then

$$F_{I^T} = \frac{\left[X_L\right]_{I^T,t}}{\left[X_L\right]_{I^T,0}} = \exp\left(-k\,t\right) = \exp\left(-\exp\left(m_1\,I^T + \frac{m_2}{T} + b\right)t\right) \tag{2.10}$$

If the temperature fluctuates as a function of time, an iterative calculation must be performed until the total evaporation time at all temperatures is reached

$$F_{I^T} = \frac{1}{\left[X_L\right]_{I^T,0}} \prod_{j=1}^{n}\left[X_L\right]_{I^T,j-1}\left(\exp\left(-\exp\left(m_1\,I^T + \frac{m_2}{T_j} + b\right)t_j\right)\right) \tag{2.11}$$

Equations 2.9–2.11 can be used to predict the extent of evaporation for an individual compound at retention index I^T as a function of temperature and time.

It is also useful to predict the extent of evaporation for the complete sample, whether as a bulk quantity or as a chromatogram. The total fraction remaining (F_{Total}) can be calculated as a bulk quantity by summation over the retention index range from initial (I_i^T) to final (I_f^T) values

$$F_{Total} = \frac{\sum_{I_i^T}^{I_f^T} F_{I^T}\left[X_L\right]_{I^T,0}}{\sum_{I_i^T}^{I_f^T}\left[X_L\right]_{I^T,0}} = \frac{\sum_{I_i^T}^{I_f^T} F_{I^T}\;A_{I^T,0}}{\sum_{I_i^T}^{I_f^T} A_{I^T,0}} \tag{2.12}$$

where the individual values for fraction remaining (F_{I^T}) at each retention index are obtained from Equation 2.9, 2.10, or 2.11, as appropriate. In Equation 2.12, $\left[X_L\right]_{I^T,0}$ is the initial concentration of the compound at each retention index, which is proportional to the detector response, such as the GC-MS abundance ($A_{I^T,0}$) of that compound in the unevaporated sample. The total fraction remaining represents the sample considered as a single or bulk entity; for example, if $F_{Total} = 0.7$, then 70% of the total sample remains and 30% has been evaporated. F_{Total} does not provide specific information about the individual compounds in the sample.

To observe how individual compounds are distributed, the total fraction remaining can be predicted as a chromatogram. First, a fraction-remaining curve is constructed by plotting Equation 2.9, 2.10, or 2.11, as appropriate, as a function of the retention index.

To illustrate the characteristic shape, a representative fraction-remaining curve using the fixed-temperature model (Equation 2.9) at $T = 20°C$ and $t = 1$ h is shown in Figure 2.3A. From this sigmoidal curve, any compound with $I^T < 500$ (*n*-pentane) is predicted to be completely evaporated ($F_{I^T} = 0$). Compounds with $500 < I^T < 1000$ undergo evaporation to different extents, with an inflection point at approximately $I^T = 700$ (*n*-heptane) where the fraction remaining is $F_{I^T} = 0.5$. Finally, compounds with $I^T > 1000$ (*n*-decane) are relatively unaffected by evaporation ($F_{I^T} = 1$).

The fraction-remaining curve (Figure 2.3A) is then multiplied, point by point at each retention index, with the chromatogram of the unevaporated sample of gasoline (Figure 2.3B) to generate the chromatogram corresponding to a sample evaporated at 20°C for 1 h. The predicted chromatogram of evaporated gasoline (Figure 2.3C) illustrates many of the features discussed above. Compounds with $I^T < 500$ are completely evaporated, *n*-hexane ($I^T = 600$) is reduced by approximately 90%, *n*-heptane ($I^T = 700$) is reduced by approximately 50%, *n*-octane ($I^T = 800$) is reduced by approximately 30%, *n*-nonane ($I^T = 900$) is reduced by approximately 5%, and compounds with $I^T > 1000$ are not significantly affected.

The extent of evaporation in the chromatogram is directly related to the evaporation rate constant (k), as predicted from the McIlroy model or Burkhart–McIlroy model at fixed temperature (Equations 2.6 and 2.9) or variable temperature (Equations 2.8 and 2.10 or 2.11). Hence, the error in the predicted rate constants, as discussed in Sections 2.2.2.1 and 2.2.2.2 above, is reflected in the error in the predicted chromatograms. A compound with a predicted rate constant that is higher than the experimental value

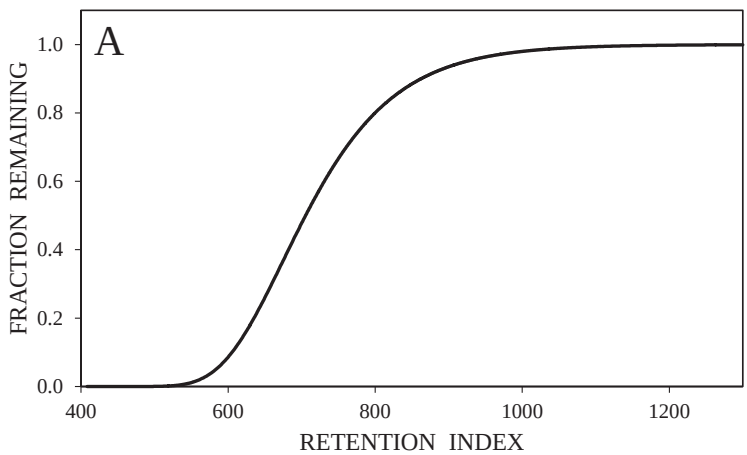


FIGURE 2.3 (A) Fraction-remaining curve calculated from Equation 2.9 using the comprehensive fixed-temperature model of McIlroy et al. [31] at $T = 20°C$ and $t = 1$ h. (B) Experimental chromatogram of unevaporated gasoline. (C) Predicted chromatogram of evaporated gasoline generated by multiplying the chromatogram of unevaporated gasoline by the fraction-remaining curve ($F_{Total} = 0.7$). The *n*-alkanes are labeled, together with selected compounds: (1) toluene, (2) ethylbenzene, (3) *m,p*-xylene, (4) *o*-xylene, (5) ethylmethylbenzene, (6) 1,2,4-trimethylbenzene, (7) 1,2,3-trimethylbenzene, (8) indane, (9) methylindane, (10) naphthalene, and (11) methylnaphthalene.

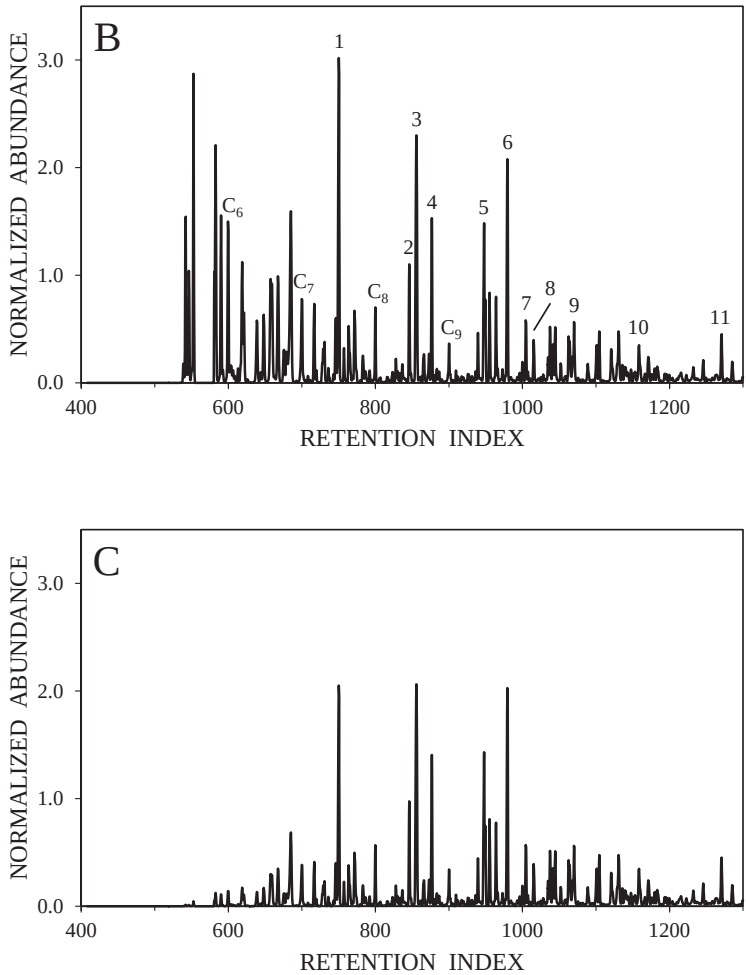


FIGURE 2.3 (Continued)

(i.e., points below the regression line in Figures 2.2A and 2.2B, positive error in rate constant) will have a predicted concentration/abundance that is lower than the experimental value (negative error in concentration/abundance). Conversely, a compound with a predicted rate constant lower than the experimental value will have a predicted concentration/abundance higher than the experimental value. However, these systematic errors arising from the kinetic model can be exceeded and masked by errors arising from experimental and instrumental sources.

In this review chapter, several useful applications of the fixed- and variable-temperature kinetic models are presented for environmental science (Section 2.3) and forensic science (Section 2.4). The accuracy of the models to predict evaporation of individual compounds and bulk samples, as well as chromatograms of those samples, is demonstrated.

2.3 ENVIRONMENTAL APPLICATIONS

Petroleum and petroleum products are a major part of industrial and domestic activities, with approximately 20 million barrels of oil used each day in the United States [34]. With this widespread use, there are unintentional releases of petroleum and petroleum products into the environment from natural and anthropogenic sources. These latter sources can account for the release of approximately 200 million gallons per year through spills, leaks, and other discharges during the processing, transport, and consumption of petroleum and petroleum products. These releases can have a devastating effect on the surrounding environment for years after the release [35].

After an environmental release, petroleum begins undergoing physical, chemical, and biological weathering almost immediately. The weathering processes that occur are dependent on the type of petroleum that has been released, whether a crude oil or a refined product such as gasoline, kerosene, diesel, and heating oils, among many others. They are also dependent on the location of the spill, whether on land, in fresh water, or in salt water, as well as the temperature and many other environmental factors [36, 37]. Among these processes, evaporation is typically the most dominant weathering process, beginning immediately after the spill and continuing throughout the spill remediation [38, 39]. For typical crude oil, the mass loss due to evaporation ranges from 40—75%, whereas for refined products such as gasoline, evaporation can account for 100% of the mass loss [35].

In general, the identity of the petroleum product is known or can be surmised to arise from a source in temporal and spatial proximity to the spill. Accordingly, the primary goal of environmental modeling is to accurately predict evaporation or the fraction remaining of that petroleum product as a function of time and/or temperature. These predictions can be used to establish that exposure levels for humans and for wildlife, including land and aquatic species, are within safe levels. They can also be used to evaluate the effectiveness of remediation efforts by comparing the actual fraction remaining with that predicted for evaporation alone.

In the following sections, the results of several environmental applications of the fixed- and variable-temperature kinetic models are presented. First, the models are used to predict the total fraction remaining of diesel fuel after evaporation under both constant temperature and fluctuating temperature conditions, which is compared with experimental results. This prediction is also performed for kerosene and marine fuel stabilizer. The models are then used to predict the distribution of individual compounds in diesel fuel, kerosene, and marine fuel stabilizer after evaporation, which is compared with experimental results. Next, the models are used to estimate the evaporation time from chromatograms of an unevaporated and an evaporated fuel sample. Finally, the models are used to estimate the time required to evaporate the total fuel or an individual compound, such as benzene, to a specific fraction remaining.

2.3.1 Prediction of Total Fraction Remaining

The total fraction remaining provides a means to assess, at the most basic level, the net or collective environmental impact posed by a petroleum spill. Together with an estimate of the initial amount of the spill, it can be used to calculate the mass or

volume remaining as a function of time. It can also be used to evaluate and compare the net or collective effect of different remediation strategies.

2.3.1.1 Diesel Fuel

To test and validate the fixed- and variable-temperature models of McIlroy et al., three samples of automotive diesel fuel were evaporated at 20°C for a total time of 100 h [31, 32]. Representative chromatograms of diesel fuel before and after evaporation are shown in Figures 2.4A and 2.4B, respectively. Diesel contains a wide range of compounds of varying volatility, including *n*-alkanes ranging from *n*-octane to *n*-docosane or higher, as well as branched and cyclic alkanes, alkyl benzenes, and polycyclic aromatics.

The experimental total fraction remaining was calculated from the average change in mass of diesel fuel before and after evaporation to be $F_{Total} = 0.8176$. The comprehensive fixed-temperature model was utilized to predict the fraction-remaining curve at $T = 20°C$ and $t = 100$ h *via* Equation 2.9, with parameters given in Table 2.2. This fraction-remaining curve (Figure 2.5, solid line) was then multiplied by the normalized abundance at the corresponding retention index in the chromatogram of the unevaporated diesel fuel (Figure 2.4A), and then summed according to Equation 2.12 from $I_i^T = 800$ to $I_f^T = 2200$. The predicted total fraction remaining for the fixed-temperature model was $F_{Total} = 0.8542$, representing an error of 4.49% relative to the experimental value.

Similarly, the comprehensive variable-temperature model was utilized to predict the fraction-remaining curve at $T = 20°C$ and $t = 100$ h *via* Equation 2.10, with parameters given in Table 2.3. This fraction-remaining curve (Figure 2.5, dashed line) was then multiplied by the chromatogram of the unevaporated diesel fuel (Figure 2.4A), and then summed according to Equation 2.12 from $I_i^T = 800$ to $I_f^T = 2200$. The predicted total fraction remaining for the variable-temperature model was $F_{Total} = 0.8347$, representing an error of 2.09% relative to the experimental value.

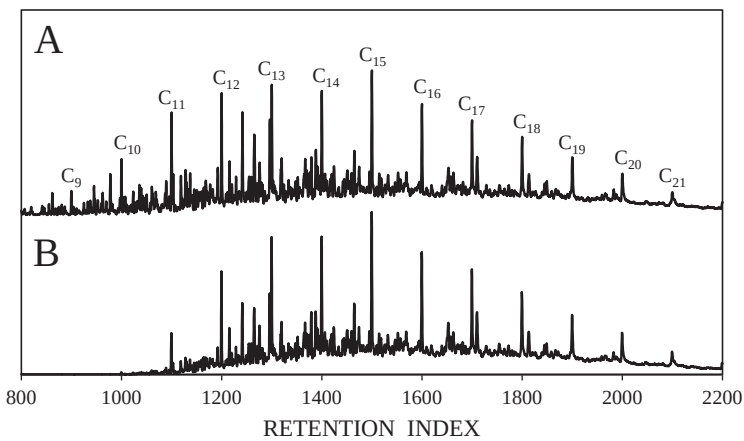


FIGURE 2.4 Experimental chromatograms of diesel fuel (A) prior to evaporation and (B) after evaporation at 20°C for 100 h. The *n*-alkanes are labeled.

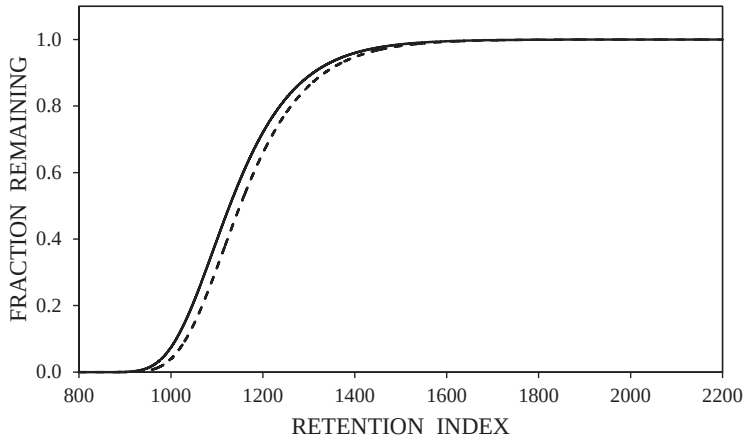


FIGURE 2.5 Fraction-remaining curve calculated using the comprehensive fixed-temperature model (Equation 2.9, solid line) and the comprehensive variable-temperature model (Equation 2.10, dashed line) for evaporation at $T = 20°C$ and $t = 100$ h.

The variable-temperature model was also validated under conditions of fluctuating temperature in order to simulate environmentally relevant diurnal and seasonal variations. The temperature was varied in the range of 12—27°C approximately every 12 h, for a total time of 100 h. The temperature profile, recorded in the evaporation chamber at 2-min intervals, is shown in Figure 2.6A (solid line). Again, three samples of diesel fuel were evaporated. Based on the average change in mass before and after evaporation, the experimental total fraction remaining was $F_{Total} = 0.8253$. To predict the fraction remaining using the variable-temperature model, it is necessary to use the iterative calculations in Equation 2.11, together with the temperature recorded at 2-min intervals. The iterative calculations were performed using an algorithm written in-house [40]. The total fraction remaining was predicted *via* Equation 2.12 to be $F_{Total} = 0.8672$, which represents 5.07% error compared to the experimental value (Table 2.4). This error is slightly greater than that observed above at a constant temperature of 20°C (2.09%). This suggests that the variable-temperature model can predict the fraction remaining over a wide range of fluctuating temperatures with good accuracy, comparable to that at constant temperature.

For many practical environmental applications, such highly accurate and detailed temperature data may not be available. For example, temperature data are available at hourly intervals for many areas in the United States from the National Oceanic and Atmospheric Administration (NOAA) National Climatic Data Center [41]. To simulate more readily available temperature data, profiles with the temperature collected at 1-h, 5-h, and 12-h intervals were also utilized. The temperatures at 5-h intervals (circles) and 12-h intervals (asterisks) are shown in Figure 2.6A. In addition, the running average temperature was calculated (Figure 2.6A, dashed line). The comprehensive variable-temperature model (Equation 2.11, parameters given in Table 2.3) was used to calculate the fraction remaining using each temperature

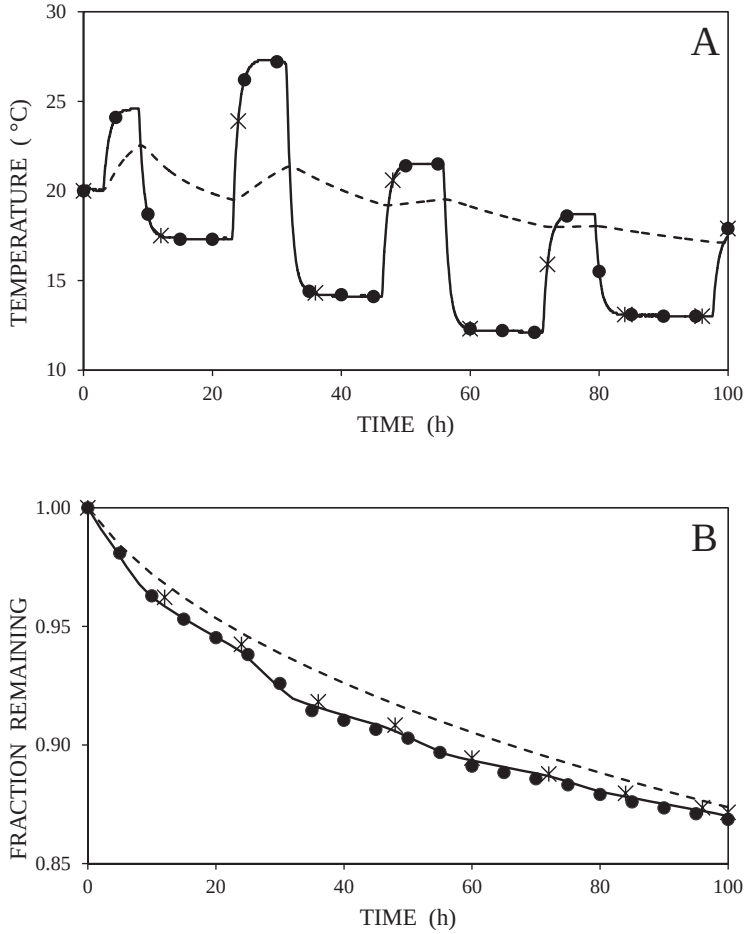


FIGURE 2.6 (A) Temperature profile of the evaporation experiment with fluctuating temperature recorded every two minutes (solid line) and as a running average temperature (dashed line). The temperatures at 5-h intervals (●) and 12-h intervals (⚹) are also shown. (B) Fraction of fuel remaining calculated by using the comprehensive variable-temperature model (Equation 2.11) at 2-min intervals (solid line), 5-h intervals (●), 12-h intervals (⚹), and running average temperature (dashed line).

interval. The predicted fraction remaining over the duration of the 100-h experiment is shown in Figure 2.6B. The predicted fraction remaining for the 5-h and 12-h intervals is very similar to that for the 2-min interval at all evaporation times. When the running average temperature is used, the fraction remaining is slightly higher because the average temperature is less sensitive to the high and low temperature fluctuations (Figure 2.6A, dashed line). However, by 100 h, the predicted fraction remaining using the running average temperature (17.1°C) becomes more similar to that from the other temperature profiles.

TABLE 2.4
The total fraction remaining (F_{Total}) of diesel fuel predicted by using the variable-temperature model with temperature data collected every 2 minutes, every 1 hour, every 5 hours, every 12 hours, and the running average temperature (Figure 2.6A). The experimental fraction remaining of diesel fuel based on the average change in mass was 0.8253. The error between the predicted and experimental F_{Total} values is shown. In addition, the error between the predicted F_{Total} values using the 2-min temperature interval compared to the longer intervals is shown.

Time Interval	*Predicted F_{Total}*	*% Error from Predicted F_{Total} using 2-min Time Interval*	*% Error from Experimental F_{Total}*[a]
2 min	0.8672	0.00	5.07
1 h	0.8672	0.01	5.08
5 h	0.8658	-0.15	4.91
12 h	0.8688	0.19	5.27
100 h Average	0.8711	0.46	5.55

[a] % error = $(F_{Total,pred} - F_{Total,exp})/F_{Total,exp} \times 100$, where $F_{Total,pred}$ and $F_{Total,exp}$ are the predicted and experimental total fraction remaining, respectively.

The predicted total fraction remaining at the end of the 100-h experiment is summarized in Table 2.4. In general, the fraction remaining is similar for all temperature profiles. The 2-min interval is expected to be the most accurate (F_{Total} = 0.8672), since it most closely reflects the actual temperature in the evaporation chamber. The running average temperature is expected to be the least accurate (F_{Total} = 0.8711), yet the difference between these two values is only 0.46%. The error for all time intervals ranges from 4.91% to 5.27% relative to the experimental value, while that for the running average is 5.55%. This suggests that the use of the average temperature over the course of an environmental spill or discharge is a reasonable approximation. This is advantageous because the average temperature is more readily obtained and allows for simpler application of the predictive models.

2.3.1.2 Kerosene and Marine Fuel Stabilizer

One of the advantages of this kinetic model is that the same regression equations (Equations 2.6 and 2.9 for fixed temperature and Equations 2.8 and 2.10 or 2.11 for variable temperature) can be applied, in principle, to any petroleum fuel. To demonstrate this capability, the fixed- and variable-temperature models were applied to predict the total fraction remaining for kerosene and marine fuel stabilizer. Three samples of each fuel were evaporated at a constant temperature of 20°C for 100 h, as discussed for diesel fuel in Section 2.3.1.1. Representative chromatograms of each fuel before and after evaporation are shown in Figures 2.7 and 2.8. Kerosene

has a similar composition and distribution of compounds compared to diesel fuel, but contains more short-chain *n*-alkanes ranging from *n*-nonane to *n*-heptadecane (Figure 2.7). As a result, kerosene is more volatile than diesel fuel. Marine fuel stabilizer contains mostly branched and cyclic alkanes, with very low abundances of *n*-alkanes and aromatic compounds (Figure 2.8). Marine fuel stabilizer is more volatile than either diesel fuel or kerosene.

For kerosene, the experimental fraction remaining based on the average change in mass before and after evaporation was F_{Total} = 0.6171. The predicted total fraction remaining using the fixed-temperature model at 20°C (Equations 2.9 and 2.12, parameters given in Table 2.2) was F_{Total} = 0.7095, which represents 15.0% error compared to the experimental value. The predicted total fraction remaining using the variable-temperature model (Equations 2.10 and 2.12, parameters given in Table 2.3) was F_{Total} = 0.6819, which represents 10.5% error compared to the experimental value. For marine fuel stabilizer, the experimental fraction remaining based on the change in mass was F_{Total} = 0.5576. The predicted fraction remaining using the fixed-temperature model was F_{Total} = 0.5798, representing 3.97% error. The predicted fraction remaining using the variable-temperature model was F_{Total} = 0.5187, representing -6.97% error. These low errors demonstrate the success of the kinetic models in predicting the total fraction remaining for a range of petroleum fuels and products.

2.3.1.3 Comparison to Other Evaporation Models

The accuracy of the fixed- and variable-temperature models developed in this work has been demonstrated in Sections 2.3.1.1 and 2.3.1.2. For further validation, the total fraction remaining predicted by these models was compared to existing evaporation models. Among these, the empirical models of Fingas [11, 36] demonstrate a linear dependence on temperature (*T*, °C) and a square-root dependence on time (*t*,

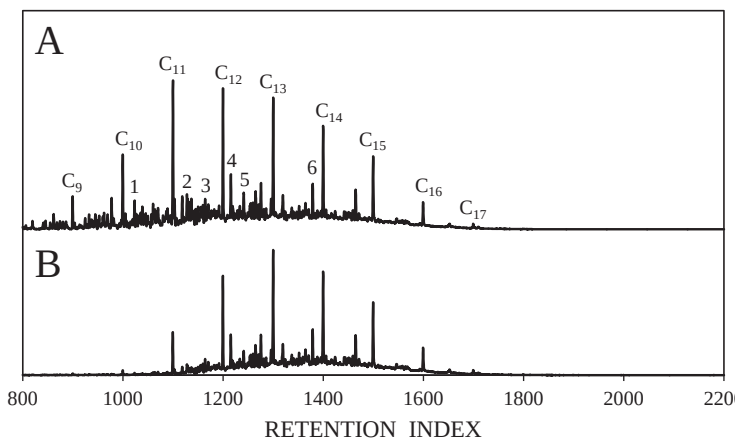


FIGURE 2.7 Experimental chromatograms of kerosene (A) prior to evaporation and (B) after evaporation at 20°C for 100 h. The *n*-alkanes are labeled, together with selected compounds: (1) 4-methyldecane, (2) pentyl cyclohexane, (3) methylundecane, (4) 2,6-dimethylundecane, (5) hexyl cyclohexane, and (6) 2,6,10-trimethyldecane.

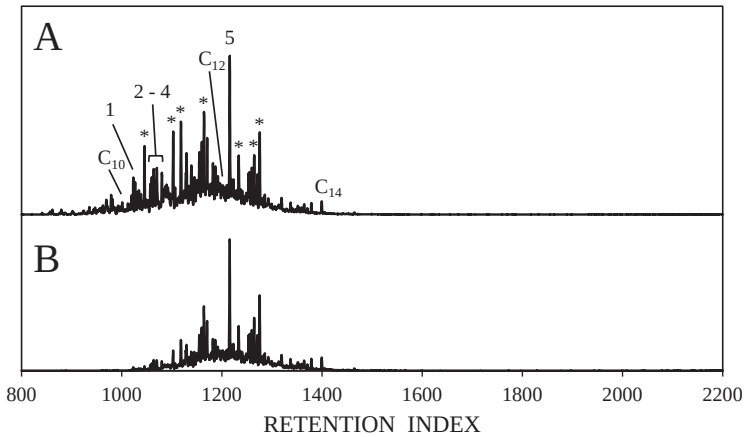


FIGURE 2.8 Experimental chromatograms of marine fuel stabilizer (A) prior to evaporation and (B) after evaporation at 20°C for 100 h. The *n*-alkanes are labeled, together with selected compounds: (1) 4-methyldecane, (2) 5-methyldecane, (3) 2-methyldecane, (4) 3-methyldecane, (5) 2,6-dimethylundecane, and (*) unidentified branched alkanes.

min) for short-term evaporation of diesel fuels. The percent of diesel fuel evaporated (*%Evap*) is given by

$$\%Evap = \left(A + B \times T\right)\sqrt{t} \tag{2.13}$$

where the regression coefficients (*A* and *B*) are determined from experimental measurements. While the empirical models of Fingas are very simple, Jones [20] demonstrated that they provide results similar to those of other common evaporation models, including the analytical model [17, 18] and the pseudo-component model [16, 20].

For diesel fuel evaporated at 20°C for 100 h, the experimental total fraction remaining was $F_{Total} = 0.8176$ (Section 2.3.1.1). Using the empirical models of Fingas (Equation 2.13, parameters given in Table 2.5 [11]), the predicted fraction remaining ranges from $F_{Total} = 0.8141$ for short-term evaporation of southern diesel to $F_{Total} = 0.4036$ for short-term evaporation of northern Anchorage diesel (Table 2.5). The short-term southern diesel model is the most accurate, with error of -0.42% relative to the experimental value. The short-term diesel Mobile 1997 and diesel 2002 models also perform well, with errors of -5.16% and -4.21%, respectively. However, the short-term regular stock diesel and Anchorage diesel models have significantly greater errors of -41.2% and -50.6%, respectively. These results emphasize the importance of knowing the type, source, and/or chemical composition of the fuel for accurate prediction using empirical models.

Using the fixed-temperature model (Equations 2.9 and 2.12, parameters given in Table 2.2), the predicted fraction remaining is $F_{Total} = 0.8542$, representing 4.49% error relative to the experimental value. Using the variable-temperature model (Equations 2.10 and 2.12, parameters given in Table 2.3), the predicted fraction remaining is $F_{Total} = 0.8347$, representing 2.09% error. The error using

the kinetic models is similar to that using the best empirical models of Fingas (Table 2.5). Moreover, the kinetic models do not require information regarding the type or source of the fuel and can be used to predict both short-term and long-term evaporation.

2.3.2 Prediction of Compound Distribution

The fixed- and variable-temperature kinetic models can also be used to predict the chromatogram, or the distribution of individual compounds in the petroleum sample, after evaporation. This distribution provides more detailed and specific information about the chemical composition of the residue, which can aid in the assessment of environmental impact and evaluation of remediation strategies.

2.3.2.1 Diesel Fuel

To test and validate the fixed- and variable-temperature models of McIlroy et al., the evaporation of diesel fuel at 20°C for 100 h, described in Section 2.3.1.1, serves as a representative example. The fraction remaining at each retention index is calculated using the fixed-temperature model (Equation 2.9, parameters given in Table 2.2) and using the variable-temperature model (Equation 2.10, parameters given in Table 2.3). The fraction-remaining curves (Figure 2.5, solid line and dashed line, respectively) are multiplied by the chromatogram of the unevaporated diesel fuel (Figure 2.4A) to generate the predicted distribution of compounds after evaporation. The predicted chromatograms (Figure 2.9) are then compared to the experimental chromatogram (Figure 2.4B) obtained by evaporation of diesel at 20°C for 100 h.

TABLE 2.5

Empirical models of Fingas for predicting short-term (< 5 days) evaporation of diesel fuel [11]. The empirical regression coefficients of Equation 2.13 are given, together with the predicted evaporation (%*Evap*) and corresponding predicted total fraction remaining (F_{Total}). The experimental fraction remaining of diesel fuel based on the average change in mass was 0.8176, and the error between the predicted and experimental F_{Total} values is shown.

Model	*A*	*B*	Predicted %*Evap*	Predicted F_{Total}	% Error from *Experimental* F_{Total} [a]
Diesel Mobile 1997	0.03	0.013	22.46	0.7754	-5.16
Diesel 2002	0.02	0.013	21.69	0.7831	-4.21
Diesel—Regular Stock	0.31	0.018	51.90	0.4810	-41.2
Diesel—Southern	-0.02	0.013	18.59	0.8141	-0.42
Diesel—Anchorage	0.51	0.013	59.64	0.4036	-50.6

[a] % error defined in Table 2.4

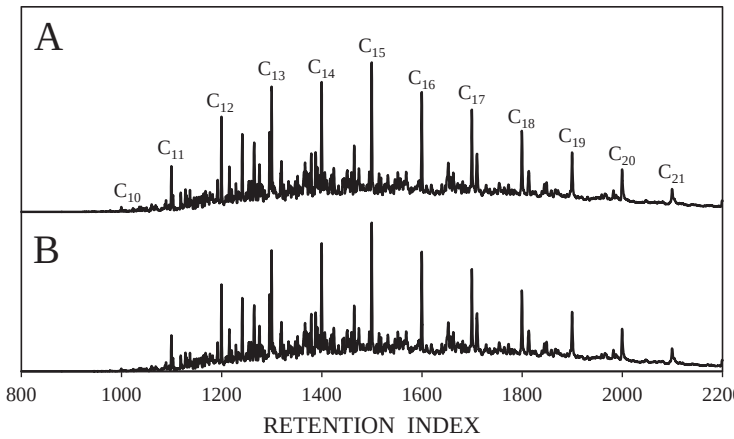


FIGURE 2.9 Predicted chromatograms of diesel fuel after evaporation at 20°C for 100 h calculated by using (A) the fixed-temperature model (Equation 2.9, parameters given in Table 2.2) and (B) the variable-temperature model (Equation 2.10, parameters given in Table 2.3). The *n*-alkanes are labeled.

A visual comparison of the predicted and experimental chromatograms of diesel fuel suggests a relatively high degree of similarity. To quantify the similarity, the Pearson product-moment correlation (PPMC or *r*) coefficient is used

$$r = \frac{\sum_{I_i^T}^{I_f^T}\left[\left(A_{I^T,1} - \overline{A_1}\right)\left(A_{I^T,2} - \overline{A_2}\right)\right]}{\sqrt{\sum_{I_i^T}^{I_f^T}\left(A_{I^T,1} - \overline{A_1}\right)^2 \sum_{I_i^T}^{I_f^T}\left(A_{I^T,2} - \overline{A_2}\right)^2}} \tag{2.14}$$

where $A_{I^T,1}$ and $A_{I^T,2}$ represent the GC-MS abundance at each retention index and $\overline{A_1}$ and $\overline{A_2}$ represent the mean abundance in the two chromatograms being compared. PPMC coefficients in the range of $1.00 \geq |r| \geq 0.80$ indicate strong correlation, $0.80 > |r| \geq 0.50$ indicate moderate correlation, and $0.50 > |r| \geq 0.00$ indicate weak to no correlation [42].

For comparison of the experimental chromatogram (Figure 2.4B) and the predicted chromatogram using the fixed-temperature model (Figure 2.9A), the PPMC coefficient is $r = 0.9962$ over the retention index range $I^T = 800$—2200. Similarly, for comparison of the experimental chromatogram (Figure 2.4B) and the predicted chromatogram using the variable-temperature model (Figure 2.9B), the PPMC coefficient is $r = 0.9981$ over the same retention index range. These high correlation coefficients indicate that the experimental and predicted chromatograms are strongly correlated, and that the fixed- and variable-temperature models can accurately predict the distribution of individual compounds after evaporation.

Upon closer inspection, differences between the experimental and predicted chromatograms can be seen in the retention index range I^T = 1100—1200, where the fraction-remaining curve changes significantly (Figure 2.5). In this retention index range, the PPMC coefficients are r = 0.9897 and 0.9934 for the fixed- and variable-temperature models, respectively. Although the experimental and predicted chromatograms are still strongly correlated, the PPMC coefficients are slightly lower than those over the complete range of I^T = 800—2200. Hence, it is beneficial to examine the accuracy of the predicted concentration or GC-MS abundance of individual compounds throughout the retention index range.

Using the *n*-alkanes as representative compounds, the percent error was calculated from the experimental and predicted chromatograms of diesel fuel and the results are summarized in Table 2.6. The error in concentration/abundance for most *n*-alkanes is

TABLE 2.6
Accuracy of predicted concentration or GC-MS abundance of *n*-alkanes in chromatograms of diesel fuel (Figures 2.4 and 2.9), kerosene (Figures 2.7 and 2.10), and marine fuel stabilizer (Figures 2.8 and 2.11) evaporated at 20°C for 100 h.

		% Error in Concentration/Abundance [a]					
		Diesel		Kerosene		Marine Fuel Stabilizer	
Compound	I^T	Fixed T Model [b]	Variable T Model[c]	Fixed T Model[b]	Variable T Model[c]	Fixed T Model[b]	Variable T Model [c]
n-Undecane	1100	9.98	-13.3	37.6	8.55	24.7	-1.63
n-Dodecane	1200	-7.89	-15.8	2.34	-6.43	-0.79	-9.29
n-Tridecane	1300	-9.10	-10.6	-6.35	-9.45	-2.25	-5.51
n-Tetradecane	1400	-5.91	-7.10	-4.03	-5.24	-4.08	-5.29
n-Pentadecane	1500	-8.09	-8.52	-0.47	-0.94	-2.40	-2.86
n-Hexadecane	1600	-2.11	-2.29	-1.31	-1.48	0.99	0.81
n-Heptadecane	1700	-2.94	-3.00	10.4	10.4		
n-Octadecane	1800	-1.87	-1.90				
n-Nonadecane	1900	-0.11	-0.12				
n-Eicosane	2000	1.79	1.79				
MAPE[d]		4.98	6.44	8.93	6.06	5.60	4.23

[a] % error = $(A_{I^T,pred} - A_{I^T,exp})/A_{I^T,exp} \times 100$, where $A_{I^T,pred}$ and $A_{I^T,exp}$ are the predicted and experimental GC-MS abundances, respectively, at the appropriate retention index (peak apex).

[b] McIlroy fixed-temperature model (Equations 2.6 and 2.9) with parameters given in Table 2.2.

[c] McIlroy variable-temperature model (Equations 2.8 and 2.10) with parameters given in Table 2.3.

[d] Mean absolute percent error (MAPE) = $\Sigma \mid (A_{I^T,pred} - A_{I^T,exp})/A_{I^T,exp} \mid \times 100/n$, where n is the number of measurements.

negative, as would be expected. Because their experimental rate constants are below the regression lines in Figure 2.2A, the rate constants are overestimated by the kinetic models and the concentration/abundance in the chromatogram is underestimated. The error is generally greater for more volatile *n*-alkanes and decreases progressively with increasing carbon number. This trend is consistent with the PPMC coefficients, as discussed above. The errors range from -9.10% to 9.98% for the fixed-temperature model and from -15.8% to 1.79% for the variable-temperature model. Although the range of errors is similar, those for the variable-temperature model are typically more negative (or, equivalently, less positive) than those for the fixed-temperature model. This trend can also be explained by the regression lines shown in Figure 2.2A, where the fixed-temperature model (solid line) is lower than the variable-temperature model (dashed line) at 20°C. The variable-temperature model predicts a higher rate constant and, hence, a faster evaporation rate at each retention index, leading to more negative error in concentration/abundance. The mean absolute percent error is 4.98% and 6.44% for the fixed- and variable-temperature models, respectively. Hence, the predictive accuracy of both kinetic models is appropriate for detailed evaluation of the concentration/abundance of individual compounds in diesel fuels undergoing evaporative weathering.

Finally, the variable-temperature model was tested for evaporation of diesel under conditions of fluctuating temperature, as described in Section 2.3.1.1 [40]. The fraction-remaining curve was calculated at the average temperature (17.1°C) and final time (100 h) for the fluctuating temperature profile shown in Figure 2.6A. As described previously, the fraction remaining curve was multiplied by the chromatogram of the unevaporated diesel fuel to generate the predicted distribution of compounds after evaporation. The predicted chromatogram was then compared to the experimental chromatogram obtained by evaporation of diesel in the fluctuating-temperature experiment. The PPMC coefficient was $r = 0.9878$ over the retention index range I^T = 800—2200, indicating that the predicted and experimental chromatograms were strongly correlated. This demonstrates that the average temperature can be used to predict the distribution of individual compounds under conditions of fluctuating temperature with results similar to those at constant temperature.

2.3.2.2 Kerosene and Marine Fuel Stabilizer

The fixed- and variable-temperature kinetic models can also be used to predict the distribution of compounds in kerosene and marine fuel stabilizer after evaporation at 20°C for 100 h, as described in Section 2.3.1.2. After prediction of the chromatograms, both visual and quantitative comparisons were performed.

For kerosene, the experimental chromatogram (Figure 2.7B) and the predicted chromatograms (Figure 2.10) show good visual agreement. The PPMC coefficients for the fixed- and variable-temperature models are $r = 0.9834$ and 0.9893, respectively, indicating strong correlation to the experimental chromatogram in the retention index range I^T = 800—2200. However, as for diesel fuel, volatile compounds in the range I^T = 1100—1200, where the fraction-remaining curve changes significantly (Figure 2.5), appear to have higher abundance in the predicted chromatograms. In this retention index range, the PPMC coefficients are $r = 0.9535$ and 0.9663 for the fixed- and variable-temperature models, respectively. Although these PPMC values

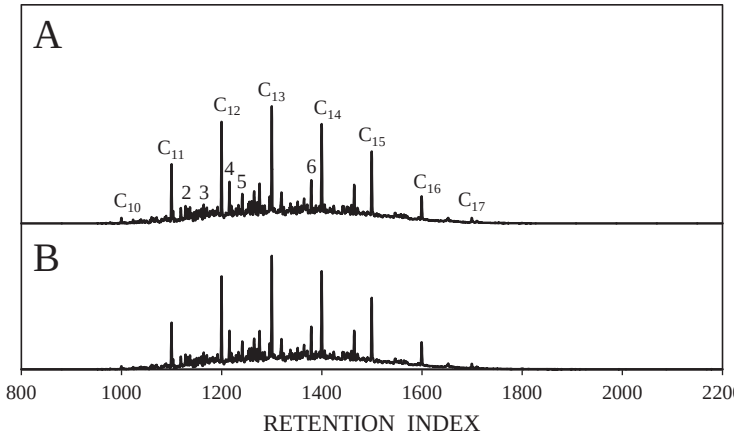


FIGURE 2.10 Predicted chromatograms of kerosene after evaporation at 20°C for 100 h calculated by using (A) the fixed-temperature model (Equation 2.9, parameters given in Table 2.2) and (B) the variable-temperature model (Equation 2.10, parameters given in Table 2.3). The *n*-alkanes are labeled, other compounds identified in Figure 2.7.

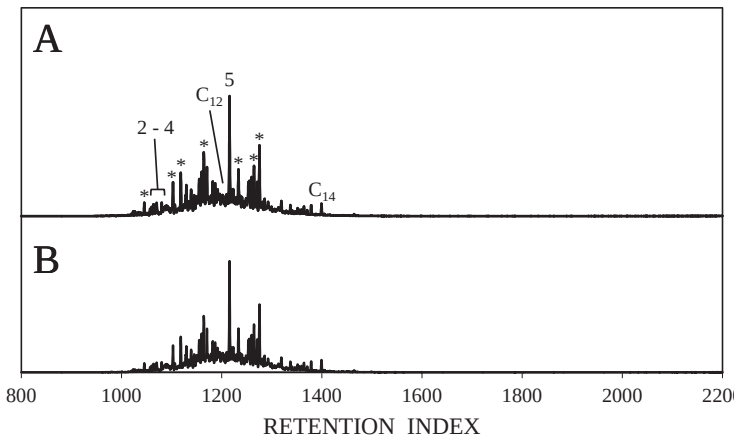


FIGURE 2.11 Predicted chromatograms of marine fuel stabilizer after evaporation at 20°C for 100 h calculated by using (A) the fixed-temperature model (Equation 2.9, parameters given in Table 2.2) and (B) the variable-temperature model (Equation 2.10, parameters given in Table 2.3). The *n*-alkanes are labeled, other compounds identified in Figure 2.8.

are lower than those observed for diesel fuel over the same retention index range, they still indicate strong correlation.

For marine fuel stabilizer, the experimental chromatogram (Figure 2.8B) and the predicted chromatograms (Figure 2.11) also show good visual agreement, with trends similar to those observed for kerosene. The PPMC coefficients for the fixed- and variable-temperature models are $r = 0.9824$ and 0.9867, respectively, again

indicating strong correlation to the experimental chromatogram in the retention index range I^T = 800—2200. For the volatile compounds in the range I^T = 1100—1200, the PPMC coefficients are r = 0.9542 and 0.9610 for the fixed- and variable-temperature models, respectively, similar to those observed for kerosene. Overall, the PPMC coefficients for kerosene and marine fuel stabilizer are comparable to those for diesel fuel (Section 2.3.2.1), confirming that the kinetic models can accurately predict the distribution of compounds in other complex petroleum products.

It is also beneficial to examine the accuracy of predicting the concentration/abundance of individual compounds in these fuels. Using the *n*-alkanes as representative compounds, the percent error was calculated from the experimental and predicted chromatograms of kerosene and marine fuel stabilizer and the results are summarized in Table 2.6. As for diesel fuel, the fixed- and variable-temperature models have similar error. The error is greatest for *n*-undecane and decreases progressively with increasing carbon number, in many cases becoming negative. The mean absolute percent error for the fixed- and variable-temperature models for kerosene (8.93% and 6.06%, respectively) and marine fuel stabilizer (5.60% and 4.23%, respectively) is comparable to that for diesel fuel (4.98% and 6.44%, respectively). These low errors demonstrate the success of the kinetic models in predicting the distribution of compounds and the concentration/abundance of individual compounds in a range of petroleum products.

2.3.3 Prediction of Evaporation Time

The kinetic models developed in this work can also be used to estimate the evaporation time from the chromatograms of an unevaporated and evaporated fuel sample. This is useful in environmental applications to estimate the time at which the spill or discharge occurred, which may be helpful for source identification or apportionment. To do so, a fraction-remaining curve is created sequentially for each possible evaporation time. The fraction-remaining curve is multiplied by the normalized abundance at the corresponding retention index in the chromatogram of the unevaporated fuel sample to generate the predicted chromatogram (as discussed in Section 2.3.2). The predicted chromatogram at each possible evaporation time is compared to the chromatogram of the evaporated fuel sample using PPMC coefficients (Equation 2.14), and a characteristic graph of PPMC versus evaporation time is then prepared. The time at which the PPMC coefficient reaches a maximum value is considered to be the best estimate of the evaporation time.

This approach was tested using three samples of diesel fuel evaporated at 20°C for 100 h [40]. For each replicate, the fixed- and variable-temperature models were used to predict chromatograms of samples evaporated over the time range 0—500 h at 1-h intervals. At each time, the predicted chromatogram was compared to the experimental chromatogram and the PPMC coefficient was calculated. A representative example of the distribution of PPMC coefficients as a function of predicted evaporation time is shown in Figure 2.12.

As the evaporation time is incremented, the PPMC coefficient increases, reaches a maximum value, and then decreases. The curve for the fixed-temperature model is shifted to longer evaporation times than the curve for the variable-temperature model.

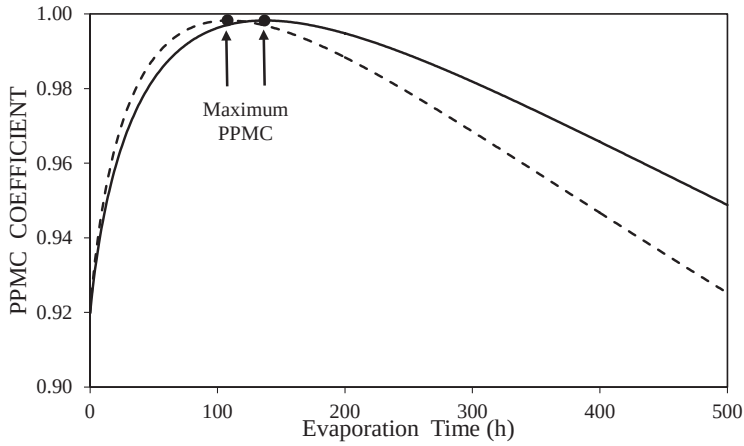


FIGURE 2.12 Pearson product-moment correlation (PPMC) coefficient between an experimental chromatogram of diesel fuel evaporated at 20°C for 100 h and the predicted chromatogram based on the fixed- and variable-temperature models calculated for evaporation time t = 0—500 h at 1-h intervals. The curve for the fixed-temperature model (solid line) maximizes at t_{max} = 137 h (r_{max} = 0.9982), whereas the curve for the variable-temperature model (dashed line) maximizes at t_{max} = 108 h (r_{max} = 0.9982).

This trend can be explained by the regression lines shown in Figure 2.2A, where the fixed-temperature model (solid line) is lower than the variable-temperature model (dashed line) at 20°C. As noted previously, the variable-temperature model predicts a higher rate constant and, hence, a faster evaporation rate at each retention index. Accordingly, this model reaches the maximum PPMC coefficient in a shorter evaporation time. Because the same replicate chromatograms are compared, both models have the same maximum PPMC coefficient (r_{max} = 0.9982), but the variable-temperature model maximizes at t_{max} = 108 h, whereas the fixed-temperature model maximizes at t_{max} = 137 h. As the actual evaporation time was 100 h, the error for this replicate is 8% for the variable-temperature model and 37% for the fixed-temperature model. For all replicates, the average predicted evaporation time is t_{max} = 108 h (8% error), with a range of 100—112 h for the variable-temperature model, and t_{max} = 136 h (36% error), with a range of 127—142 h for the fixed-temperature model. The average PPMC coefficient was r_{max} = 0.9983, with a range of 0.9980—0.9986. This demonstrates the utility of the kinetic models in estimating the length of time a petroleum fuel sample has been evaporating in the environment, given chromatograms of the original unevaporated fuel and the evaporated fuel.

2.3.4 Prediction of Time to Specific Fraction Remaining

The fixed- and variable-temperature kinetic models have been shown to accurately predict the evaporation time (Section 2.3.3) and can, therefore, be used to estimate the time required for the entire fuel to reach a specific fraction remaining. Additionally, they can be used to estimate the time for an individual compound to reach a specific

level, such as a lethal dose (LD_{50}), lethal concentration (LC_{50}), or limit of detection. This information is critical to assess safety at spill or discharge sites and to predict the persistence of an individual compound in the environment.

Using the total fraction remaining (Equation 2.12) with either the fixed-temperature model (Equation 2.9, parameters given in Table 2.2) or variable-temperature model (Equation 2.10, parameters given in Table 2.3), numerical integration can be used to determine the time to reach a specific fraction remaining [40]. A semi-logarithmic plot of the predicted fraction remaining versus evaporation time of diesel at 20°C is shown in Figure 2.13. For both kinetic models, the fraction remaining decreases rapidly for the first day and into the first week, then decreases more slowly. A plot such as this is useful in assessing temporal changes in the fuel due to evaporation. For example, using the variable-temperature model, 25% ($F_{Total} = 0.75$) of the total fuel is predicted to evaporate by approximately 240 h (10 days), and 50% ($F_{Total} = 0.50$) to evaporate by approximately 2400 h (100 days) at 20°C.

A similar calculation can be performed for any individual compound in the fuel sample. For example, the BTEX compounds (benzene, toluene, ethylbenzene, and xylene) are among the most water-soluble components in petroleum fuels. Benzene is of particular interest because it is highly toxic with acute exposure and is carcinogenic and mutagenic with chronic, long-term exposure [43]. Benzene constitutes approximately 1% of commercial gasoline by volume [44]. If 50 L of gasoline (a typical automobile gas tank) were discharged into a small stream, approximately 440 g of benzene would be released into the environment. In a water volume of 25,000 L, the initial concentration of benzene would be 17.5 mg/L. For rainbow trout, the lethal concentration (LC_{50}) is 5.3 mg/L for 96 h [43]. The time until the concentration reaches below the LC_{50} can be solved by using either the fixed-temperature model (Equation 2.9) or variable-temperature model (Equation 2.10) for benzene ($I^T = 650$ [45]). Using the variable-temperature model, the time to reach the LC_{50} ($F_{I^T} = 5.3/17.5 = 0.30$) is approximately

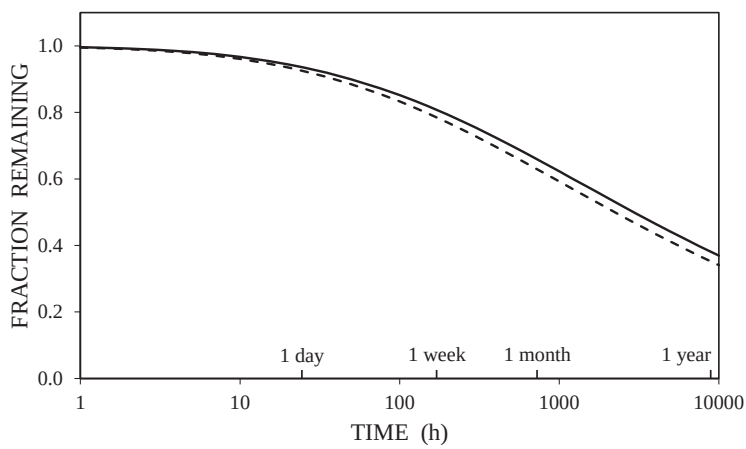


FIGURE 2.13 The total fraction remaining for evaporation of diesel fuel predicted using the fixed-temperature model (solid line) and variable-temperature model (dashed line) over a total time of 10,000 h (approximately 1 year) at an average temperature of 20°C.

1 h and the time to reach 1% remaining ($F_{I^T} = 0.01$) is approximately 4 h. While this is a simple example, it serves to demonstrate the utility of the model in predicting removal of a specific compound from the environment by evaporation.

2.3.5 Summary

The fixed- and variable-temperature kinetic models of McIlroy et al. [31, 32] offer many useful applications for environmental modeling. The rate constant can be utilized to predict the fraction remaining of an individual compound as well as the fraction remaining of the entire fuel. At a constant evaporation temperature of 20°C and time of 100 h, the fraction remaining of diesel fuel was predicted with 2.09% error using the variable-temperature model, which is comparable to the best empirical models of Fingas [11]. However, unlike other models, the kinetic models can be directly applied to a wide range of other petroleum fuels. The fractions remaining of kerosene and marine fuel stabilizer were predicted with 10.5% and -6.97% error, respectively, using the variable-temperature model. More importantly, the fraction remaining can be predicted for fluctuating temperature conditions, an option that is not available with most existing models.

From the fraction remaining of individual compounds, the fuel composition can be predicted after evaporation at a given temperature and time. This prediction is helpful in establishing the loss that would be expected due to evaporation alone, which can then be compared with actual loss to assess the effectiveness of remediation strategies. The models can also be used to estimate the evaporation time, from chromatograms of an unevaporated and evaporated fuel sample, or alternatively, to estimate the original fuel composition from the chromatogram of an evaporated sample and the evaporation time. Both of these predictions may be useful for source identification or apportionment of an environmental spill or discharge. Finally, the models can be used to estimate the evaporation time required to reach a specific fraction remaining for an individual compound or for the entire fuel. This prediction is useful for assessing hazards for cleanup workers or for determining the time to reach safe exposure levels, which is particularly important for toxic volatile compounds.

2.4 FORENSIC APPLICATIONS

The ability to predict evaporation has numerous potential applications in forensic science, such as in the characterization of explosives, chemical warfare agents, and fire debris. Of these applications, fire debris analysis is perhaps the most routinely performed in forensic science laboratories across the country and is the focus in this section.

According to the United States National Fire Protection Association (NFPA), an average of 52,260 intentional fires were set annually in the five-year period 2014—2018 [46]. Of these intentional fires, three in five were set in residential properties and resulted in $815 million in property damage, 950 civilian injuries, and 400 deaths. Ignitable liquids are commonly used as accelerants in intentional fires, with the result that evaporated residues of the liquids may remain in the resulting debris. As such, the goal of forensic fire debris analysis is to identify the presence of extraneous liquids in debris collected from the scene of a suspicious fire.

Ignitable liquid residues in debris samples submitted to a laboratory are first extracted using an acceptable method, such as passive-headspace extraction or solvent extraction [47–49]. Extracts are routinely analyzed by GC-MS and the resulting chromatograms are compared to chromatograms of reference liquids for identification. It is worth noting here that the goal is primarily to identify the chemical class of liquid present rather than a specific liquid. Liquid classes are defined by ASTM International and include the gasoline, aromatic, isoparaffinic, and petroleum distillate classes, among others [50]. Liquids are further classified based on carbon number range with 'light' liquids containing carbon numbers in the range C_4—C_9, 'medium' containing C_8—C_{13}, and 'heavy' containing C_9—C_{20+}. As an example, a liquid containing an approximately Gaussian-shaped distribution of *n*-alkanes across the range C_8—C_{13} with a lower abundance of aromatic compounds is defined as a medium petroleum distillate.

Identification of liquids in debris samples based on direct comparison of the chromatogram to a reference collection is challenging due to the nature of the debris sample. Any liquid initially present is likely to be evaporated to some extent due to the heat of the fire, which results in different chemical composition compared to the corresponding unevaporated reference liquid [2]. The chromatogram of the debris is also further complicated by additional contributions from thermal degradation and pyrolysis of the substrate [2]. Such contributions are not present in the chromatogram of the reference liquid, which complicates direct comparison of the chromatograms.

The challenges of evaporation and substrate contributions are routinely addressed in two ways. Reference collections of ignitable liquids typically include chromatograms of the liquids experimentally evaporated to different levels. Further, in addition to total ion chromatograms (TICs), extracted ion profiles (EIPs) are also compared between the submitted sample and the reference collection. Extracted ion profiles are generated for compound classes commonly present in ignitable liquids and are used to minimize or even eliminate substrate contributions [50].

As with any reference collection comparison, successful identification requires an extensive and sufficiently representative reference collection. Unfortunately, with limited time and resources, it is not feasible to experimentally evaporate all liquids in the reference collection to a range of evaporation levels. As such, evaporated versions of a select number of liquids may be included in the reference collection, which also impacts the number of corresponding EIPs that can be generated for comparison. To avoid the time and expense associated with experimental evaporation of reference liquids, the kinetic model can be used to predict TICs and EIPs corresponding to different evaporation levels for any class of liquid. The ability to predict evaporation as a function of retention index is particularly advantageous for fire debris applications in which the identity of the sample liquid is typically not known initially.

In this section, the kinetic model is first applied to identify liquids experimentally evaporated to different levels [51–55]. The accuracy of the model in predicting evaporation of individual compounds is evaluated based on the mean absolute percent error (MAPE). The accuracy in predicting the chromatogram of the evaporated liquid is subsequently evaluated using PPMC coefficients to assess correlation. The model is then used to create a reference collection containing predicted chromatograms corresponding to various evaporation levels for 10 liquids representing 4

different chemical classes. Both same-source and different-source liquids are compared to the reference collection to demonstrate identification. Finally, to demonstrate more realistic applications, the model is used to identify ignitable liquids in fire debris samples [51, 52]. Given that these samples contain substrate contributions, the predicted reference collection is expanded to include EIPs of relevant compound classes. Thus, identification is based on TICs and EIPs, with the latter being used to increase confidence in identification in the presence of substrate contributions.

2.4.1 Identification of Gasoline

Gasoline is the most common ignitable liquid used as an accelerant in intentional fires [2]. To identify gasoline in a fire debris sample, the chromatographic profile should contain *o*-, *m*-, and *p*-ethylmethylbenzene, 1,3,5-trimethylbenzene, and 1,2,4-trimethylbenzene (C_3-alkylbenzenes), along with naphthalene, 1-methylnaphthalene, 2-methylnaphthalene, and indanes [50]. While *n*-alkanes may be present, the actual content varies according to the brand and grade of gasoline. Nonetheless, when present, the concentration of any alkanes with a carbon number greater than C_7 (*n*-heptane) should be less than that of the aromatic content [50].

2.4.1.1 Evaluation of the Kinetic Model to Predict Evaporation

A gasoline sample (Gas A) was collected from a local service station and first analyzed in the unevaporated state by GC-MS (Figure 2.14A) [53, 54]. Compounds present in the unevaporated sample included *n*-heptane (I^T = 700), *n*-octane (I^T = 800), toluene

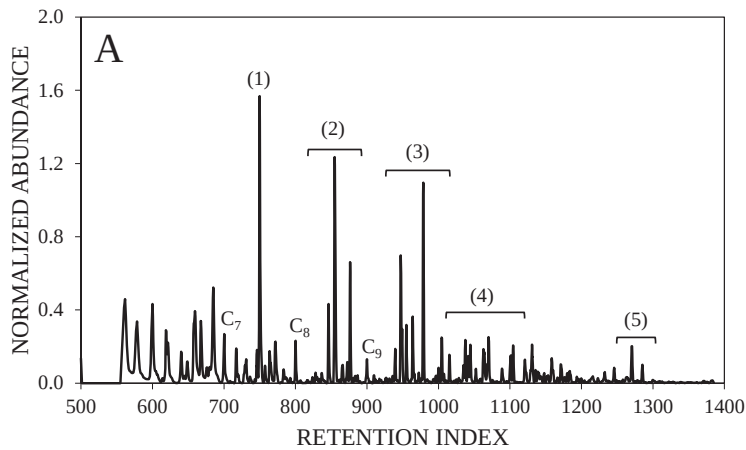


FIGURE 2.14 Representative chromatograms of gasoline (A) unevaporated, (B) evaporated to F_{Total} = 0.5 with experimental chromatogram (top) and predicted chromatogram (bottom), (C) evaporated to F_{Total} = 0.3 with experimental chromatogram (top) and predicted chromatogram (bottom), and (D) evaporated to F_{Total} = 0.1 with experimental chromatogram (top) and predicted chromatogram (bottom). The *n*-alkanes are labeled, together with selected compounds: (1) toluene, (2) C_2-alkylbenzenes, (3) C_3-alkylbenzenes, (4) C_4-alkylbenzenes, and (5) methylnaphthalenes.

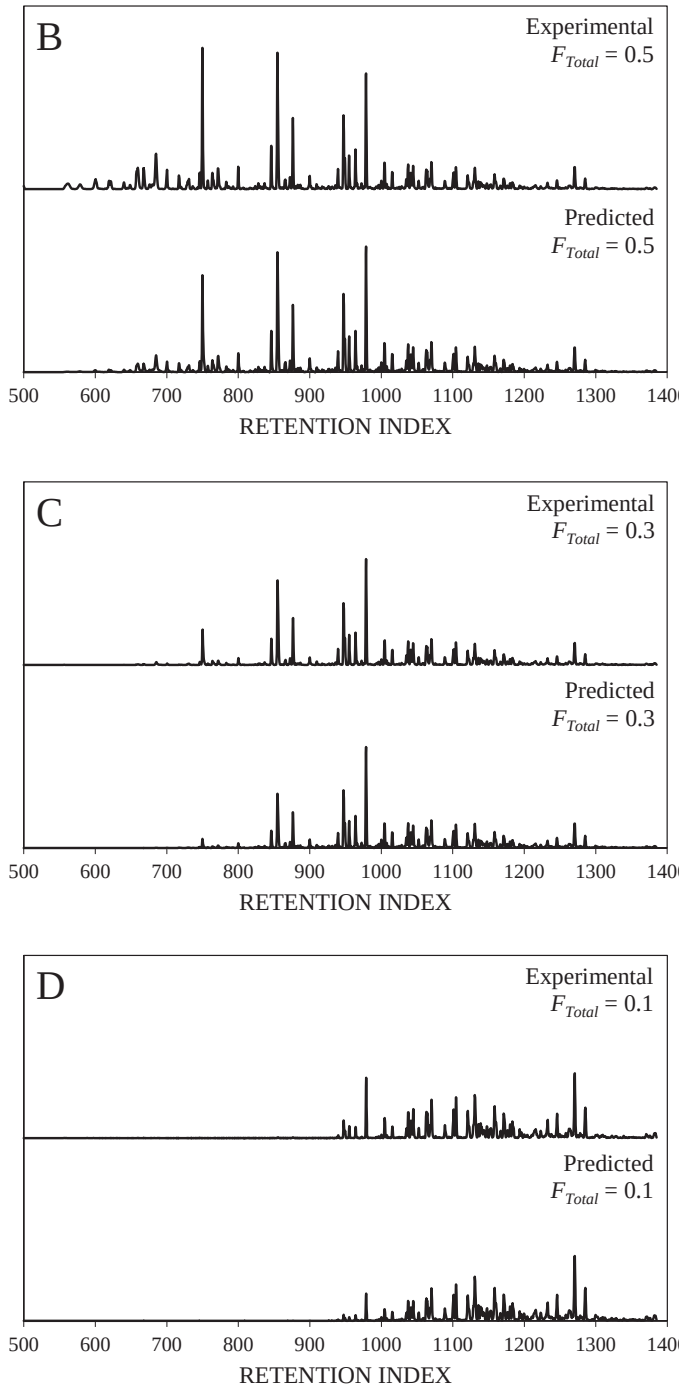


FIGURE 2.14 (Continued)

(I^T = 750), C_2-alkylbenzenes (I^T = 845—876), C_3-alkylbenzenes (I^T = 939—1004), C_4-alkylbenzenes (I^T = 1044—1172), and methylnaphthalenes (I^T = 1270—1285). The unevaporated sample was then experimentally evaporated to nominal F_{Total} = 0.5, 0.3, and 0.1, which correspond to evaporation levels of 50, 70, and 90% by volume, respectively (Figures 2.14B–D, top). At F_{Total} = 0.5, compounds characteristic of gasoline are still present (Figure 2.14B, top); however, at F_{Total} = 0.3, the abundance of compounds with I^T < 900 is substantially reduced (Figure 2.14C, top) and, by F_{Total} = 0.1, all compounds eluting in this range are completely evaporated (Figure 2.14D, top).

Eklund et al. applied the fixed-temperature McIlroy model at 20°C (Equation 2.6, parameters given in Table 2.2) to predict chromatograms corresponding to each of the experimental F_{Total} levels [53]. The actual F_{Total} levels for the evaporated samples were calculated based on the area under the chromatogram relative to the area under the chromatogram of the unevaporated gasoline [53]. As such, for nominal F_{Total} = 0.5, 0.3, and 0.1, the actual F_{Total} values were 0.671, 0.425, and 0.133, respectively. The kinetic model was applied to the chromatogram of the unevaporated gasoline, changing the time t in Equation 2.9 to reach the actual F_{Total} levels (Equation 2.12) corresponding to the experimental samples. For example, to reach F_{Total} = 0.671, 0.425, and 0.133, time was set to t = 2.045 h, 9.13 h, and 87.3 h, respectively.

The predicted chromatograms (Figures 2.14B–D, bottom) were then compared to the corresponding experimental chromatograms (Figures 2.14B–D, top) to evaluate the accuracy of the model in predicting evaporation of gasoline. Chromatograms were evaluated in two ways. First, PPMC coefficients (Equation 2.14) were calculated to assess correlation between the predicted and experimental chromatograms, and second, the mean absolute percent error (MAPE, defined in Table 2.6) in predicting abundance of selected compounds was evaluated. The model accuracy in predicting evaporation of gasoline is summarized in Table 2.7.

TABLE 2.7
Accuracy of McIlroy fixed-temperature model[a] at 20°C in predicting evaporation of gasoline at three different F_{Total} levels.

Nominal F_{Total}	Mean PPMC Coefficient	MAPE (%) [b]
0.5	0.9710 ± 0.0053	18.0 [c]
0.3	0.9713 ± 0.0044	23.3 [d]
0.1	0.9613 ± 0.0049	32.9 [e]

[a] Equation 2.6, parameters given in Table 2.2
[b] MAPE defined in Table 2.6
[c] MAPE calculated for 14 compounds across the range I^T = 600—1159 (*n*-hexane to naphthalene)
[d] MAPE calculated for 13 compounds across the range I^T = 700—1159 (*n*-heptane to naphthalene)
[e] MAPE calculated for 10 compounds across the range I^T = 846—1159 (ethylbenzene to naphthalene). Note that the number of compounds included in MAPE calculation decreases due to evaporation at lower F_{Total} levels.

At each F_{Total} level, there is strong correlation between the experimental and predicted chromatograms, with PPMC coefficients greater than 0.96. The MAPE, which represents the percent error in predicting the abundance of up to 14 compounds in gasoline, ranges from 18.0—32.9% for F_{Total} = 0.5—0.1, respectively. The increase in error is expected as F_{Total} level decreases due to extensive evaporation. At F_{Total} = 0.1, the more volatile compounds have fully evaporated and less volatile compounds are present at substantially reduced abundance, resulting in higher overall error.

These comparisons demonstrate the accuracy of the model to predict evaporation of gasoline, a liquid that is substantially more volatile than diesel and other liquids discussed thus far in this chapter. However, for all of the comparisons shown in this section, the predicted chromatogram was modeled to the same F_{Total} level as the experimental sample. While this approach enables a direct evaluation of predictive accuracy, it does not demonstrate the ability to use the model to identify given liquids, which is of more practical use.

2.4.1.2 Generation and Application of a Predicted Reference Collection

Rather than predicting a chromatogram corresponding to a specific F_{Total} level, the model can be applied to predict chromatograms corresponding to a range of F_{Total} levels. In essence, these predicted chromatograms form the basis of a predicted reference collection to which sample chromatograms can be compared for identification purposes.

To begin demonstrating this application, Eklund et al. collected four additional gasoline samples (Gasolines B—E) from local service stations over a six-month period, and each unevaporated gasoline was analyzed by GC-MS [53, 54]. Despite being distilled and refined from different sources of crude oil, transported in different tankers, and collected from different pumps, there is a high degree of chemical similarity among the gasolines, as evidenced in Figure 2.15. In fact, of the chemical classes defined by ASTM International [50], gasoline is the most chemically similar. The five gasolines all contain the characteristic compounds required for identification (e.g., toluene, C_2-, C_3-, and C_4-alkylbenzenes, and methylnaphthalenes), although there are differences in abundance ratios of specific compounds. For example, in Gas E, the ratio of toluene (I^T = 750) to *m*,*p*-xylene (I^T = 855) is approximately 2:1, whereas in Gas B, these two compounds are present in an approximately 1:1 ratio (Figure 2.15).

Eklund et al. applied the fixed-temperature McIlroy model at 20°C (Equation 2.6, parameters given in Table 2.2) to the chromatograms of the five unevaporated gasolines (Gasolines A—E), changing t in Equations 2.9 and 2.12 to generate predicted chromatograms corresponding to levels of F_{Total} = 0.9—0.1 in 0.1 increments [54]. The resulting 45 predicted chromatograms (nine F_{Total} levels for each of five gasolines) constituted a predicted reference collection to which experimentally evaporated gasolines (same-source and different-source) were compared [53, 54].

Representative chromatograms of the experimentally evaporated Gas A (Figure 2.14B–D) were compared to the full predicted gasoline reference collection. The PPMC coefficient was calculated for each pairwise comparison, with the maximum PPMC coefficient being used to determine the predicted chromatogram deemed most similar to the experimental chromatogram. Figure 2.16 summarizes

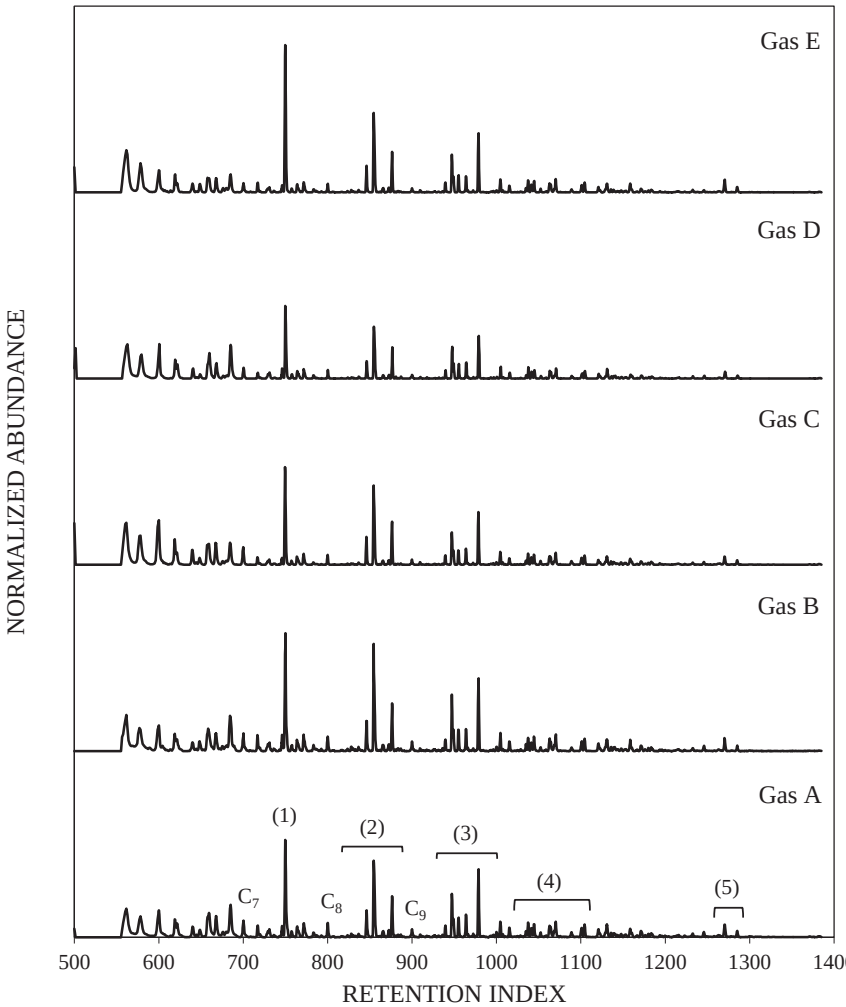


FIGURE 2.15 Representative experimental chromatograms of five gasolines (Gas A—E) in the unevaporated state. The *n*-alkanes are labeled, together with selected compounds: (1) toluene, (2) C_2-alkylbenzenes, (3) C_3-alkylbenzenes, (4) C_4-alkylbenzenes, and (5) methylnaphthalenes.

the PPMC coefficients calculated for comparison of Gas A experimentally evaporated to F_{Total} = 0.5 compared to all predicted chromatograms in the reference collection.

Similar trends are observed for comparison of the experimental chromatogram of Gas A at F_{Total} = 0.5 (Figure 2.14B) to predicted chromatograms of each gasoline. Strong correlation is observed for comparison of the experimentally evaporated liquid to each of the unevaporated liquids (r = 0.84—0.96 at F_{Total} = 1.0, highlighted in gray dashed box in Figure 2.16). As the F_{Total} represented by the predicted

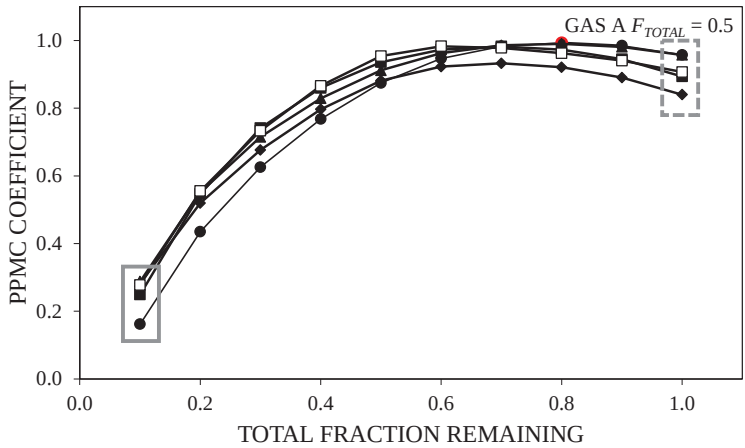


FIGURE 2.16 Comparison of experimentally evaporated Gas A (F_{Total} = 0.5) to reference collection of gasolines A—E containing predicted chromatograms corresponding to F_{Total} = 0.9—0.1 for each gasoline. Gasolines are labeled as follows: Gas A (●), Gas B (▲), Gas C (■), Gas D (♦), and Gas E (□). Gray dashed box highlights correlation between the chromatogram of each unevaporated gasoline and the experimentally evaporated Gas A. Gray solid box highlights correlation between the predicted chromatogram corresponding to F_{Total} = 0.1 for each gasoline to the experimentally evaporated Gas A. Red highlighting indicates the maximum correlation observed, which occurs for comparison of the experimental Gas A to the predicted chromatogram corresponding to Gas A at F_{Total} = 0.8.

chromatograms decreases to F_{Total} = 0.6—0.7, correlation to the experimentally evaporated sample increases. The maximum correlation (r = 0.9931) is observed for comparison to the predicted chromatogram of Gas A at F_{Total} = 0.8 (highlighted with red outline in Figure 2.16). As the F_{Total} level represented by the predicted chromatograms decreases further to F_{Total} = 0.1, correlation decreases (r = 0.16—0.28, highlighted in gray solid box in Figure 2.16), indicating weak to no correlation.

For Gas A experimentally evaporated to F_{Total} = 0.3 (Figure 2.14C), the maximum PPMC coefficient (r = 0.9916) occurs for comparison to predicted Gas A at F_{Total} = 0.5. Further, for Gas A experimentally evaporated to F_{Total} = 0.1 (Figure 2.14D), the maximum PPMC coefficient (r = 0.9808) occurs for comparison to predicted Gas A at F_{Total} = 0.2. Thus, for all of the experimental chromatograms, the correct liquid (Gas A) is identified, even with four additional gasolines in the reference collection.

While successful identification of the experimentally evaporated gasoline was achieved, this example demonstrates same-source comparison; that is, the reference collection contains chromatograms that were predicted from the same gasoline that was used for the experimental evaporations. From a practical standpoint, the liquid present in a submitted sample is unlikely to be from the same source as the corresponding liquid in the reference collection. As such, it is important to investigate different-source comparisons to demonstrate the practical utility of the predicted reference collection.

To demonstrate the potential for different-source comparisons, a sixth gasoline (Gas F) was experimentally evaporated for comparison to the gasoline predicted reference collection [54]. For Gas F experimentally evaporated to $F_{Total} = 0.5$ and 0.3, the highest correlation ($r = 0.9838$ and 0.9863) is observed to Gas B at F_{Total} levels = 0.9 and 0.7, respectively. For Gas F experimentally evaporated to $F_{Total} = 0.1$, the highest correlation ($r = 0.9812$) is observed to predicted Gas A at $F_{Total} = 0.2$. Thus, the predicted gasoline reference collection can be used to identify an evaporated gasoline originating from a source not included in the reference collection.

2.4.2 Identification of Liquids from Different Chemical Classes

2.4.2.1 Evaluation of the Kinetic Model to Predict Evaporation

Although gasoline is the most commonly encountered accelerant, the kinetic model can also be applied to identify evaporated ignitable liquids from different chemical classes. Capistran et al. selected nine additional liquids representing the aromatic, isoparaffinic, oxygenated, and petroleum distillate classes and analyzed the liquids in the unevaporated state by GC-MS [51, 52]. Representative chromatograms of one liquid (paint remover, paint thinner, lacquer thinner, and torch fuel) from each class are shown in Figure 2.17.

Based on the criteria in ASTM E1618 for liquid class identification, paint remover (Figure 2.17A) is considered an aromatic liquid due to the presence of only aromatic compounds [50]. The liquid contains ethylbenzene, *m,p*-xylene, and *o*-xylene, that elute across the range $I^T = 800$—900, further sub-classifying the liquid as a light aromatic [50]. Paint thinner (Figure 2.17B) contains branched alkanes eluting across the range $I^T = 900$—1200 and is classed as a medium isoparaffinic liquid [50]. Lacquer thinner (Figure 2.17C) contains aromatic compounds (substituted benzenes) but also

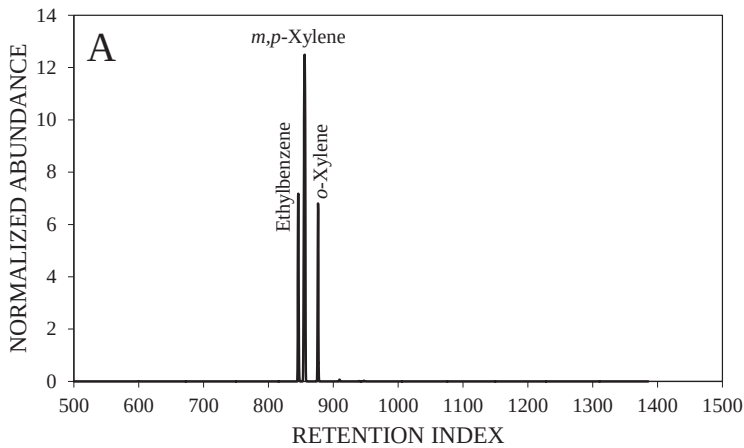
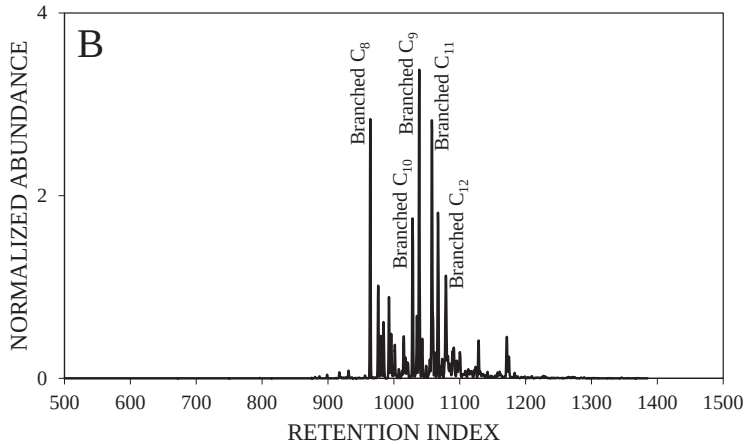
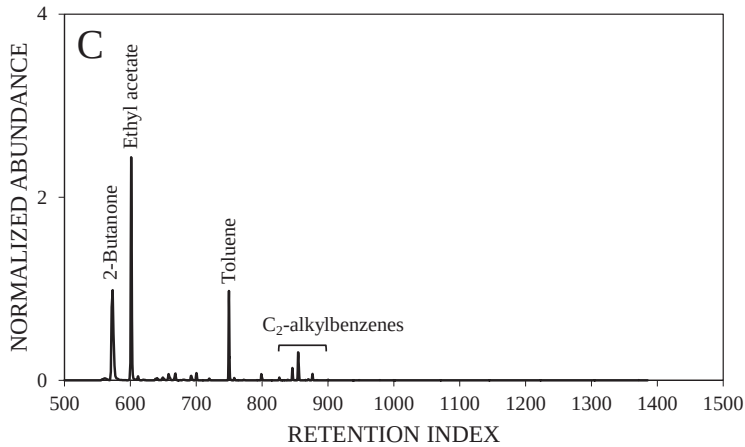


FIGURE 2.17 Representative experimental chromatograms of unevaporated liquids representing different chemical classes (A) paint remover, aromatic class, (B) paint thinner, isoparaffinic class, (C) lacquer thinner, oxygenated class, and (D) torch fuel, petroleum distillate class.





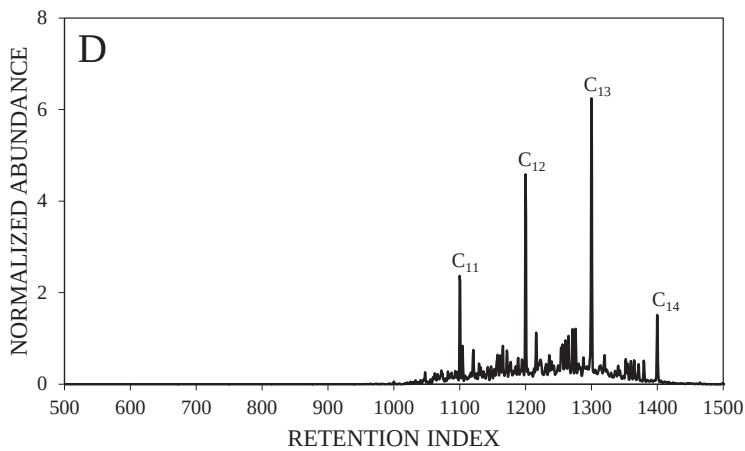


FIGURE 2.17 (Continued)

contains two oxygenated compounds (2-butanone and ethyl acetate) that elute before the substituted benzenes. As such, this liquid is a member of the oxygenated rather than aromatic class [50]. Finally, torch fuel (Figure 2.17D) contains *n*-alkanes that elute across the range I^T = 1000—1500 in an approximately Gaussian-shaped distribution. Based on this composition, torch fuel is defined as a heavy petroleum distillate [50].

Each of the nine liquids was experimentally evaporated to levels of F_{Total} = 0.5, 0.3, and 0.1 and analyzed by GC-MS [51, 52]. The kinetic model was used to predict chromatograms for each liquid corresponding to these F_{Total} levels. The predicted chromatograms were then compared to the experimental chromatograms to evaluate the model accuracy in predicting evaporation of liquids from different chemical classes. Representative comparisons are summarized in Table 2.8. For all comparisons of experimental and predicted chromatograms, strong correlation is observed with PPMC coefficients of r = 0.8372—0.9904. In general, for each liquid, correlation decreases slightly as F_{Total} level decreases. This trend is attributed to more extensive evaporation resulting in lower abundance of the remaining compounds, which increases error in prediction.

The lowest correlation (r = 0.8372) is observed for comparison of experimental and predicted chromatograms of lacquer thinner (oxygenated class) at F_{Total} = 0.3. The correlation is lower because the model overpredicts the extent of evaporation of the two oxygenated compounds, which are present at significant abundance in the experimental chromatogram. However, it is worth noting that the comprehensive kinetic model was developed based on experimentally determined rate constants for compounds representing four different chemical classes (*n*-alkanes, branched alkanes, alkyl benzenes, and polycyclic hydrocarbons) and did not include oxygenated compounds [31]. As such, the discrepancy in predicting evaporation of these

TABLE 2.8
Mean PPMC coefficient for comparison of experimental to predicted chromatograms for liquids representing different ASTM classes.

ASTM Class	Representative Ignitable Liquid	Nominal F_{Total}	Mean PPMC Coefficient
Aromatic	Paint remover	0.5	0.9758 ± 0.0032
		0.3	0.9710 ± 0.0004
		0.1	0.9400 ± 0.0174
Isoparaffinic	Paint thinner	0.5	0.9855 ± 0.0027
		0.3	0.9904 ± 0.0005
		0.1	0.9661 ± 0.0049
Oxygenated	Lacquer thinner	0.5	0.9351 ± 0.0004
		0.3	0.8372 ± 0.0134
		0.1	0.9026 ± 0.0029
Petroleum Distillate	Torch fuel	0.5	0.9480 ± 0.0006
		0.3	0.9807 ± 0.0012
		0.1	0.9432 ± 0.0057

two compounds is due to inaccuracy in modeling compound classes not included in model development. Nonetheless, strong correlation is still observed between predicted and experimental chromatograms for lacquer thinner. The expansion of the model to include additional compound classes is the focus of future work in this area.

The predictive accuracy demonstrated here highlights a unique advantage of the kinetic model. Chromatograms are predicted as a function of I^T such that the identity of compounds within the liquid does not need to be known in advance. As a result, and as demonstrated here, the model is readily applicable to chemically diverse liquids with similar predictive accuracy.

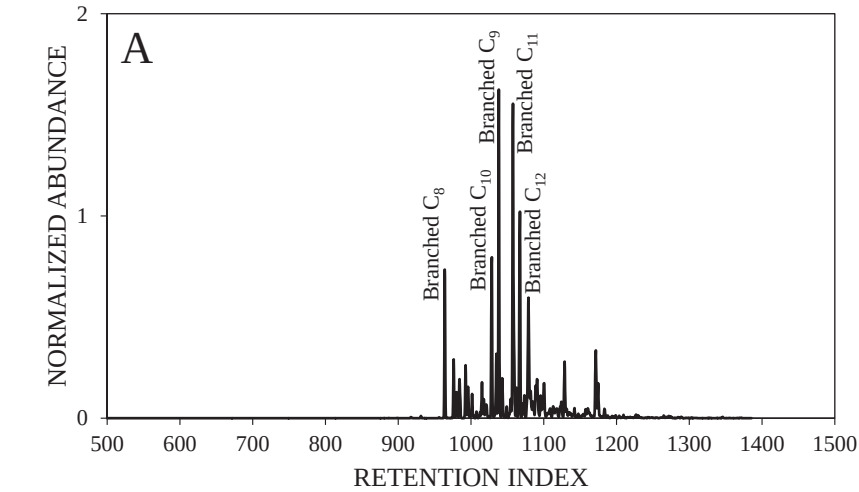
2.4.2.2 Generation and Application of a Predicted Reference Collection

Having demonstrated accuracy in predicting evaporation of a series of chemically diverse ignitable liquids, the model was then applied to generate an extensive reference collection of predicted chromatograms [51, 52]. Following the procedure described in Section 2.4.1.2, the model was applied to chromatograms of 10 unevaporated liquids (one gasoline and nine liquids from different ASTM classes) to generate predicted chromatograms corresponding to levels of F_{Total} = 0.9—0.1 in increments of 0.1. Thus, the complete predicted reference collection contained 90 chromatograms representing 10 liquids at nine different F_{Total} levels.

Four additional single-blind samples (Liquids A—D) were experimentally evaporated to levels of F_{Total} = 0.5, 0.3, and 0.1 to evaluate the potential of the predicted reference collection for identification purposes. The liquids were selected to evaluate different-source comparisons and to provide a more realistic evaluation of the utility of the reference collection. Chromatograms of the experimentally evaporated liquids were compared to each predicted chromatogram (all liquids at all F_{Total} levels), calculating PPMC coefficients for each comparison. The evaporated liquid was subsequently identified as belonging to the class to which the maximum PPMC coefficient was observed.

To illustrate, consider the chromatogram of Liquid A experimentally evaporated to F_{Total} = 0.5 (Figure 2.18A). This liquid primarily contains branched alkanes, ranging from C_8—C_{12} and eluting across the range I^T = 800—1200. On comparison to the reference collection, Liquid A at F_{Total} = 0.5 exhibits strong correlation to predicted chromatograms of paint thinner, which is an isoparaffinic liquid (Figure 2.18B). Specifically, the maximum correlation (r = 0.9800) is observed for comparison of the evaporated liquid to the predicted chromatogram at F_{Total} = 0.6 (Table 2.9). Weak to no correlation is observed for comparison of Liquid A to all other liquids at all F_{Total} levels in the predicted reference collection. Similarly, at levels of F_{Total} = 0.3 and 0.1, maximum correlation is observed for comparison to predicted chromatograms of paint thinner (r = 0.9844 and 0.9638 for F_{Total} = 0.3 and 0.1, respectively). As such, Liquid A is identified as an isoparaffinic liquid. This is the correct class identification, as Liquid A is a different-brand paint thinner that is chemically similar to the paint thinner represented in the reference collection.

The chromatogram of Liquid B at F_{Total} = 0.5 is dominated by *n*-alkanes (C_{12}—C_{15}, I^T = 1200—1500) that elute in an approximately Gaussian-shaped distribution (Figure 2.19A). At each experimental F_{Total} level, maximum correlation is observed



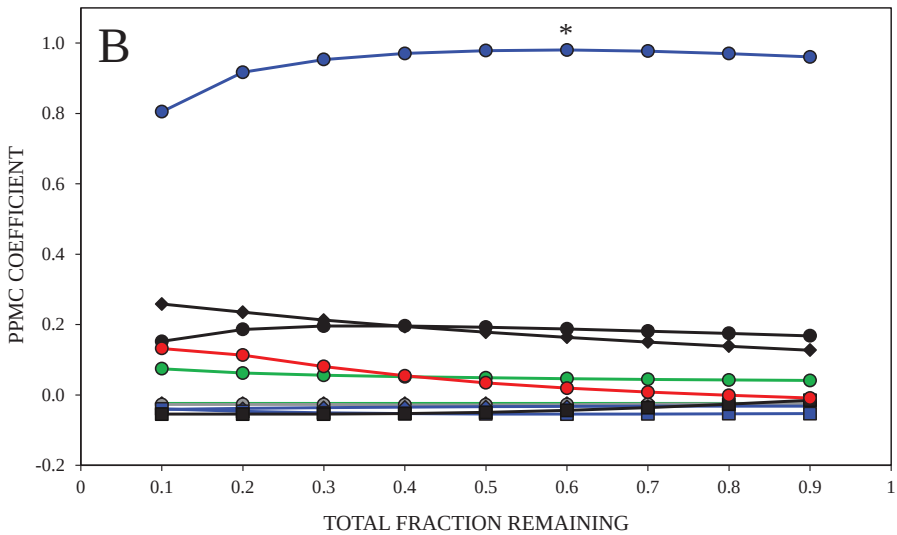


FIGURE 2.18 Identification of Liquid A (A) experimental chromatogram corresponding to F_{Total} = 0.5 and (B) comparison of experimental chromatogram to the predicted reference collection. In (B), liquids are denoted as follows: aromatic class (green), with fruit tree spray (●) and paint remover (♦); oxygenated class (gray), with lacquer thinner (●); isoparaffinic class (blue), with paint thinner (●), fabric protector (♦), and lighter fluid (■); petroleum distillate class (black), with charcoal lighter fluid (●), paint thinner (♦), and torch fuel (■); and gasoline class (red, ●). Maximum correlation indicated by *.

TABLE 2.9
Identification of experimentally evaporated liquids through comparison to predicted reference collection (different-source comparisons).

Experimentally Evaporated Liquid	Nominal F_{Total} Level	Comparison to Predicted Reference Collection			
		Max. PPMC Coefficient	Predicted F_{Total} Level	Liquid with Maximum Correlation	Class Identification
Liquid A	0.5	0.9800	0.6	Paint thinner	Isoparaffinic
	0.3	0.9844	0.4/0.3	Paint thinner	Isoparaffinic
	0.1	0.9638	0.1	Paint thinner	Isoparaffinic
Liquid B	0.5	0.9545	0.3	Torch fuel	Petroleum distillate
	0.3	0.9199	0.1	Torch fuel	Petroleum distillate
	0.1	0.8094	0.1	Torch fuel	Petroleum distillate
Liquid C	0.5	0.8977	0.2	Paint remover	Aromatic
	0.3	0.8452	0.1	Lacquer thinner	Oxygenated
	0.1	0.7858	0.1	Paint remover	Aromatic
Liquid D	0.5	0.8805	0.2	Lighter fluid	Isoparaffinic
	0.3	0.9065	0.2	Lighter fluid	Isoparaffinic
	0.1	0.9293	0.1	Lighter fluid	Isoparaffinic

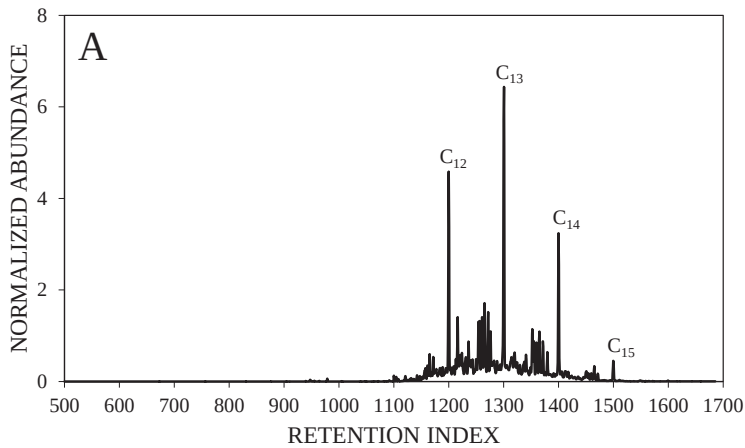


FIGURE 2.19 Identification of Liquid B (A) experimental chromatogram corresponding to F_{Total} = 0.5, (B) comparison of experimental chromatogram to the predicted reference collection, and (C) chromatogram of same-brand paint thinner included in reference collection, highlighting compositional differences. Color and symbol designations in (B) are described in Figure 2.18. Maximum correlation indicated by *.

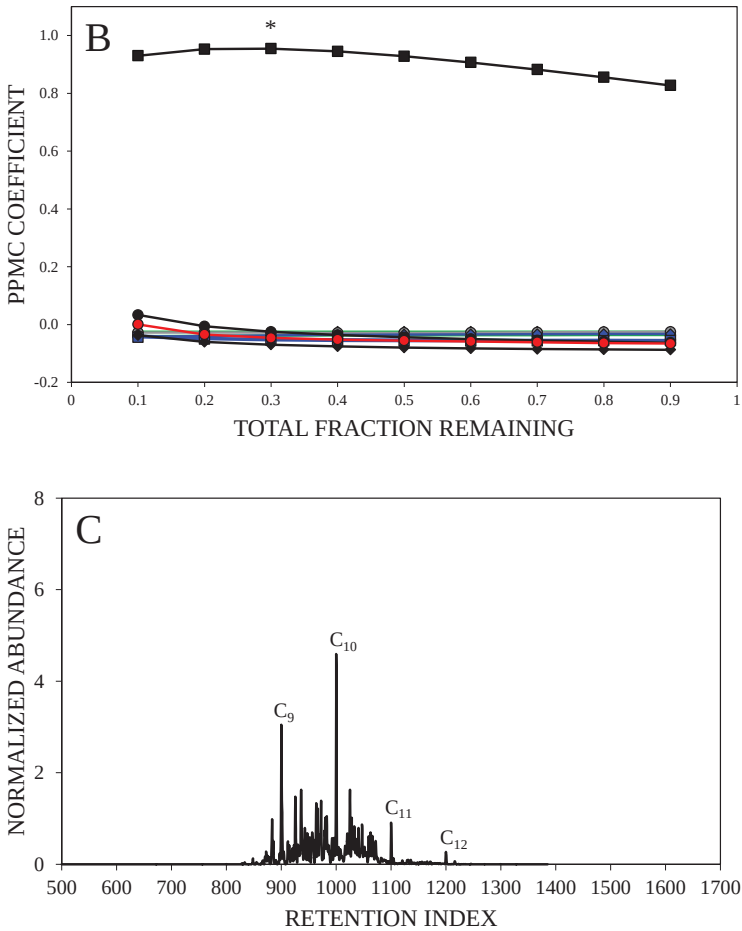


FIGURE 2.19 (Continued)

for comparison to predicted chromatograms of torch fuel, which is a petroleum distillate (Table 2.9). Little to no correlation is observed for comparisons to any other liquid in the reference collection (Figure 2.19B). Thus, based on these comparisons, Liquid B is identified as a petroleum distillate.

Although this is the correct classification, it raises an important point. Liquid B is a paint thinner sample, and the same-brand paint thinner was included in the reference collection. However, the chemical profiles of the two paint thinners are markedly different. While both liquids are defined as petroleum distillates, compounds in Liquid B elute over the range I^T = 1200—1500, whereas those in the reference collection sample elute across the range I^T = 900—1200 (Figure 2.19C). As such, Liquid B is defined as a heavy petroleum distillate while the reference collection sample is defined as a medium petroleum distillate. These two samples were purchased in Michigan and New Hampshire. Despite being the same brand with similar

packaging, differences in composition may be due to differences in chemical regulation between the two states. It is also worth noting that the strongest correlation was observed between Liquid B and torch fuel, the only heavy petroleum distillate represented in the reference collection.

The chromatogram of Liquid C at F_{Total} = 0.5 contains only ethylbenzene, *m,p*-xylene, and *o*-xylene that elute in the range I^T = 800—900 (Figure 2.20A). Maximum correlation (r = 0.8977) is observed for comparison to the predicted chromatogram of paint remover at F_{Total} = 0.2 (Figure 2.20B and Table 2.9). Paint remover is an aromatic liquid that contains the same C_2-alkylbenzenes present in similar ratios as Liquid C, resulting in strong correlation between the two liquids. It is worth noting that there is no correlation between Liquid C and the predicted chromatograms

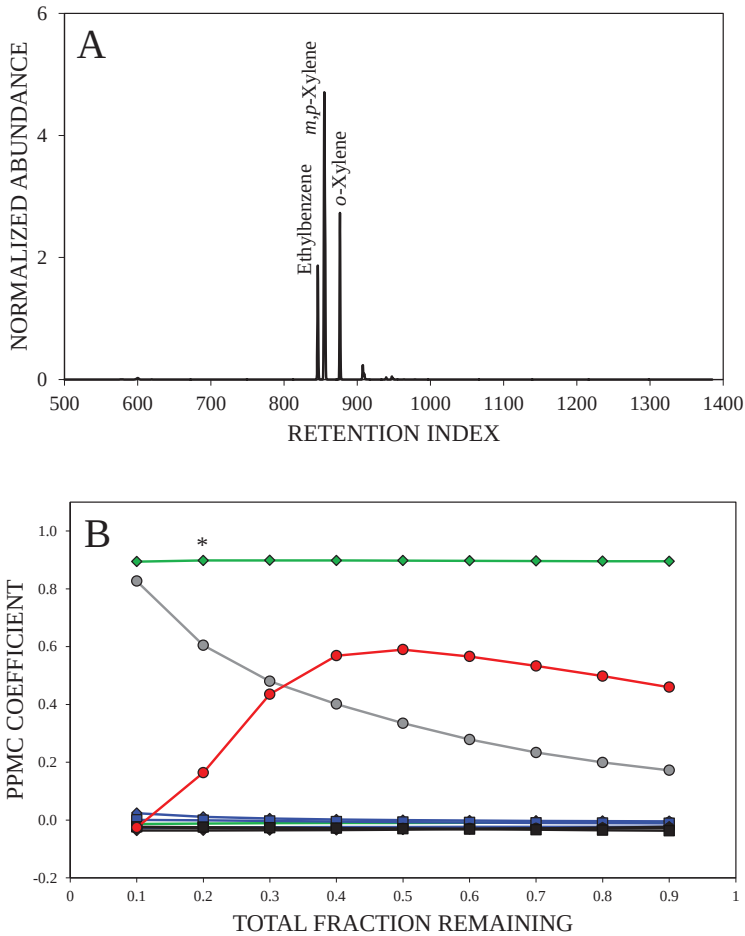


FIGURE 2.20 Identification of Liquid C (A) experimental chromatogram corresponding to F_{Total} = 0.5 and (B) comparison of experimental chromatogram to the predicted reference collection. Color and symbol designations in (B) are described in Figure 2.18. Maximum correlation indicated by *.

of fruit tree spray, which is the other aromatic liquid in the reference collection. However, fruit tree spray contains C_3- and C_4-alkylbenzenes that elute across the range I^T = 900—1100. Given that there is no similarity in the compounds present, there is no correlation between these two liquids.

From Figure 2.20B, there is strong correlation (r = 0.8265) between Liquid C and the predicted chromatogram of lacquer thinner at F_{Total} = 0.1. However, correlation to the predicted lacquer thinner chromatograms decreases as F_{Total} level increases (r = 0.1722 at F_{Total} = 0.9). While lacquer thinner does contain C_2-alkylbenzenes, this liquid also contains 2-butanone and ethyl acetate (I^T = 572 and 601, respectively), resulting in classification as an oxygenated rather than aromatic liquid. At higher F_{Total} levels, the presence of 2-butanone and ethyl acetate decreases correlation to Liquid C. However, as F_{Total} level decreases (greater extent of evaporation), 2-butanone and ethyl acetate are evaporated, leaving only the C_2-alkylbenzenes in lacquer thinner and, therefore, increasing correlation to Liquid C. Similarly, there is moderate correlation between Liquid C and predicted chromatograms of gasoline at F_{Total} = 0.8—0.4. The correlation is again due to the presence of the C_2-alkylbenzenes in both liquids; however, correlation is only moderate ($0.80 > |r| \geq 0.50$) due to the presence of additional compounds in gasoline (e.g., toluene, C_3-alkylbenzenes, C_4-alkylbenzenes, etc.) that are not present in Liquid C. There is no correlation between Liquid C and any other liquid in the predicted reference collection at any F_{Total} level (Figure 2.20B).

Similar trends are observed when chromatograms of Liquid C evaporated to F_{Total} = 0.3 and 0.1 are compared to the predicted reference collection. There is strong correlation to paint remover (aromatic class), strong correlation to lacquer thinner (oxygenated class) at F_{Total} = 0.1, moderate correlation to gasoline at F_{Total} = 0.8—0.4, and no correlation to any other liquid in the reference collection. While comparison of Liquid C at F_{Total} = 0.3 yields strong correlation to both paint remover and lacquer thinner (r = 0.8342 and 0.8452, respectively), the maximum correlation is to lacquer thinner. In cases such as this, where there is similar high correlation to liquids representing different classes, the trend in PPMC coefficients can be evaluated further. For lacquer thinner, PPMC coefficients consistently decrease, indicating moderate to weak correlation as F_{Total} levels increase from 0.1—0.9. This trend suggests the presence of additional volatile compounds in lacquer thinner that decrease correlation to the liquid of interest. In contrast, for paint remover, strong correlation is observed across all F_{Total} levels. Thus, based on these comparisons and evaluation of the PPMC coefficients, Liquid C is correctly classed as an aromatic liquid at each F_{Total} level investigated.

The chromatogram of Liquid D at F_{Total} = 0.5 contains n-alkanes, branched alkanes, and cyclic alkanes that elute across the range I^T = 700—900 (Figure 2.21A). At each F_{Total} level, maximum correlation is observed for comparisons to lighter fluid (r = 0.8805, 0.9065, and 0.9293 for F_{Total} = 0.5, 0.3, and 0.1, respectively) (Figure 2.21B for F_{Total} = 0.5). Lighter fluid is an isoparaffinic liquid that contains primarily branched alkanes. However, Liquid D is naphtha, which is a naphthenic-paraffinic liquid due to the additional presence of cyclic alkanes. Because this class is not represented in the reference collection, maximum correlation occurs

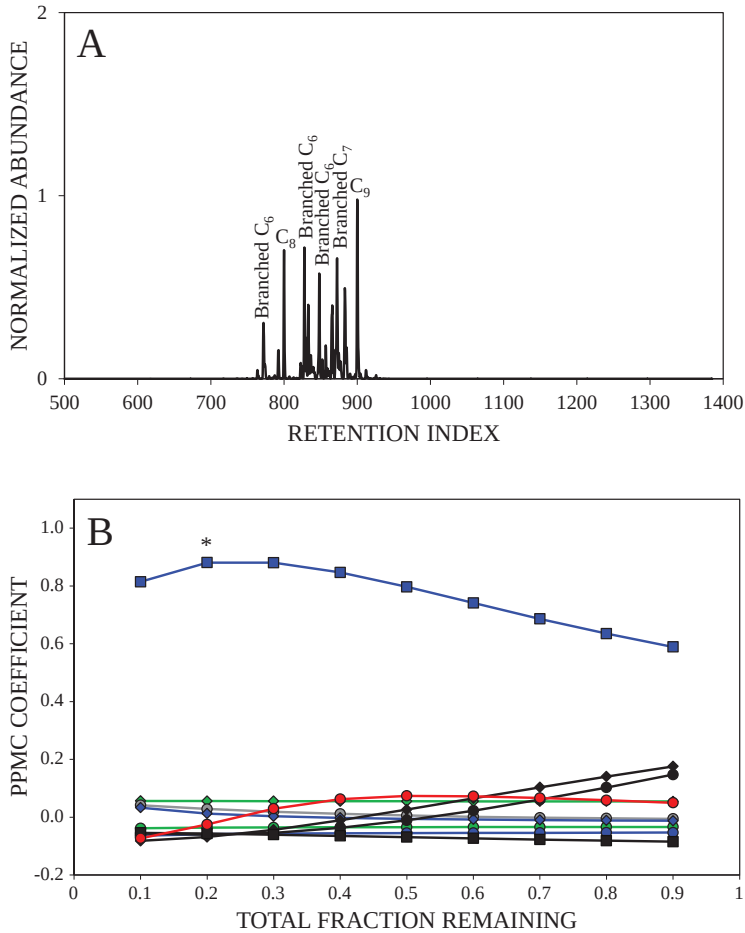


FIGURE 2.21 Identification of Liquid D (A) experimental chromatogram corresponding to F_{Total} = 0.5 and (B) comparison of experimental chromatogram to the predicted reference collection. Color and symbol designations in (B) are described in Figure 2.18. Maximum correlation indicated by *.

to the most chemically similar class, which in this case is the isoparaffinic class. This example is included to demonstrate that while chemical information can be obtained, successful class identification requires extensive and representative reference collections.

2.4.3 Identification of Liquids in Fire Debris Samples

Application of the kinetic model to generate a predicted reference collection of TICs for the identification of liquids has been demonstrated in the previous sections. However, these examples focused only on the identification of experimentally

TABLE 2.10.
Fragment ions characteristic of major compound classes in ignitable liquids used to generate extracted ion profiles.

Compound Class	Characteristic Ions
Alkane	*m/z* 57, 71, 85, 99
Aromatic	*m/z* 91, 105, 119
Cycloalkane	*m/z* 55, 69, 83, 97
Indane	*m/z* 117, 131, 145, 159
Polynuclear aromatic	*m/z* 128, 142, 156

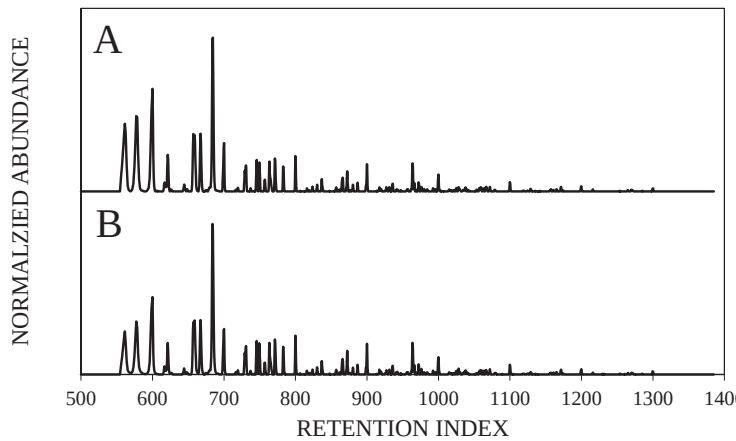


FIGURE 2.22 Extracted ion profiles (EIPs) representing the alkane class in gasoline (A) experimental EIP and (B) predicted EIP.

evaporated liquids. To be of practical use in forensic laboratories, the identification of liquids in the presence of substrate contributions must also be considered, as demonstrated by Capistran et al. [51, 52]. Due to the chromatographic complexity of typical fire debris samples, EIPs are often considered in addition to the TIC for identification purposes. Profiles representing major compound classes present in ignitable liquids are generated from the TIC using the characteristic ions defined in ASTM E1618 [50], which are summarized in Table 2.10. The EIPs offer increased selectivity and, depending on the chemical nature of the substrate, may minimize or even eliminate substrate contributions. The kinetic model can be applied to predict EIPs of the characteristic compound classes and thereby generate a predicted EIP reference collection similar to that described for TICs (Section 2.4.2.2).

2.4.3.1 Evaluation of the Kinetic Model to Predict Extracted Ion Profiles of Characteristic Compound Classes

Before developing reference collections of predicted EIPs, it is important to again evaluate the predictive accuracy of the kinetic model for this purpose [51, 52]. As an example, consider the experimental alkane EIP of gasoline in Figure 2.22. This profile was generated from the TIC of unevaporated gasoline using the ions characteristic of the alkane class (*m/z* 57, 71, 85, and 99, Table 2.10).

Capistran et al. applied the fixed-temperature McIlroy model at 20°C (Equation 2.6, parameters given in Table 2.2) to this EIP to generate predicted alkane EIPs for gasoline [51, 52]. As before, EIPs corresponding to different F_{Total} levels were generated by changing the time t in Equations 2.9 and 2.12. Here, time was set to t = 0.7, 1.9, and 33.8 h^{-1} to generate alkane EIPs corresponding to nominal F_{Total} = 0.5, 0.3, and 0.1, respectively. Alkane profiles were also generated from the TICs of each experimentally evaporated gasoline corresponding to the same nominal F_{Total} levels. The predicted and experimental EIPs were then compared using PPMC coefficients to evaluate correlation (Figure 2.22B). Extracted ion profiles representing the aromatic, cycloalkane, indane, and polynuclear aromatic compound classes within gasoline were also generated and compared to the corresponding experimental profiles in a similar manner (Table 2.11).

For each EIP and across all F_{Total} levels, strong correlation is observed between predicted and experimental profiles, with correlation in the range of r = 0.9512—0.9973 (Table 2.11). The lowest correlation is observed for comparisons of the cycloalkane profile. However, this compound class is present at low abundance, and,

TABLE 2.11
Comparison of predicted and experimental extracted ion profiles representing different compound classes within gasoline.

Compound Class	F_{Total} Level	Mean PPMC Coefficient
Alkane	0.5	0.9733 ± 0.0103
	0.3	0.9943 ± 0.0007
	0.1	0.9661 ± 0.0100
Aromatic	0.5	0.9753 ± 0.0164
	0.3	0.9983 ± 0.0013
	0.1	0.9941 ± 0.0003
Cycloalkane	0.5	0.9512 ± 0.0190
	0.3	0.9647 ± 0.0073
	0.1	–
Indane	0.5	0.9928 ± 0.0036
	0.3	0.9940 ± 0.0026
	0.1	0.9918 ± 0.0043
Polynuclear aromatic	0.5	0.9952 ± 0.0017
	0.3	0.9951 ± 0.0015
	0.1	0.9973 ± 0.0005

— Indicates no significant profile at this F_{Total} level

in fact, gasoline at F_{Total} = 0.1 did not contain a significant cycloalkane profile for comparison. Nonetheless, strong correlation is observed, which demonstrates the accuracy of the model in predicting EIPs in addition to TICs.

With the accuracy in predicting EIPs demonstrated, the next step was to generate a predicted reference collection of EIPs. As such, alkane, aromatic, indane, and polynuclear aromatic profiles were generated from the TIC of each unevaporated liquid used in the reference collection described in Section 2.4.2.2. It is worth noting here that not all liquids have EIPs for all compound classes. For example, paint remover, which is an aromatic liquid, does not have an associated alkane profile due to the lack of this compound class within the liquid. The final EIP reference collection contained a total of 24 profiles, each at nine different F_{Total} levels (F_{Total} = 0.9—0.1 in 0.1 increments), across four compound classes.

2.4.3.2 Application of Predicted Reference Collections

The utility of the predicted TIC and EIP reference collections was evaluated using chromatograms of debris collected from large-scale burns, which were conducted at the New England Fire Investigation Seminar at St. Anselm College in Manchester, New Hampshire [51, 52]. Two large containers (approximately 2.4 m x 4.9 m) were furnished with carpeting, drapes, and furniture to resemble a small living space. Approximately 5 mL of gasoline was poured in several locations throughout each container and then set alight. The fire was allowed to burn for approximately 10 minutes until past flashover and then extinguished with water. Samples of debris were collected from each burn cell in unlined metal paint cans and transported to the laboratory for analysis [51, 52].

Ignitable liquid residues were extracted from the collected debris samples following the passive-headspace extraction method detailed in ASTM E1412 [49]. Extracts were then analyzed by GC-MS to generate the TIC and, using the characteristic ions listed in Table 2.10, to generate the EIPs of relevant compound classes. The TIC and EIPs were then compared to the relevant predicted reference collection to identify the class of liquid present, which was determined based on the strongest correlation.

2.4.3.2.1 Burn Sample A

In the first burn cell, a sample of burned carpet with no ignitable liquid present was collected in addition to the debris sample. The burned carpet sample contains primarily branched and cyclic alkanes that elute across the range I^T = 800—1500, in addition to styrene (I^T = 872) and acetophenone (I^T = 1033) (Figure 2.23A). The TIC for Burn Sample A indicates the presence of toluene (I^T = 750), C_2-alkylbenzenes (I^T = 845—875), C_3-alkylbenzenes (I^T = 938—1004), 2-methylnaphthalene (I^T = 1271), and 1-methylnaphthalene (I^T = 1284) (Figure 2.23B). These compounds are characteristic of gasoline identification; however, the TIC also indicates the presence of styrene (I^T = 872) and estragole (I^T = 1172) that originate from the fire debris substrate. Styrene is present in the TIC of the burned carpet, while estragole is a component in turpentine oil, which is used for furniture and other wood treatments.

The TICs of the burned carpet and of Burn Sample A were compared to the predicted reference collection generated in Section 2.4.2.2. For the burned carpet, PPMC coefficients for comparison to the predicted reference collection are less than

0.4 (Figure 2.23C), indicating weak correlation to all liquids at all F_{Total} levels. In contrast, for Burn Sample A, the maximum correlation ($r = 0.8389$) is observed for comparison to the predicted chromatogram of gasoline at $F_{Total} = 0.2$ (Figure 2.23D). As F_{Total} increases further, correlation to gasoline decreases to $r = 0.3742$ at $F_{Total} = 0.9$. This trend is consistent with evaporation of the more volatile compounds present in gasoline, particularly toluene and C_2-alkylbenzenes. At $F_{Total} = 0.9$, these compounds are present at relatively high abundance in the predicted chromatogram, and as F_{Total} decreases, so too does the abundance of these compounds. The fire debris

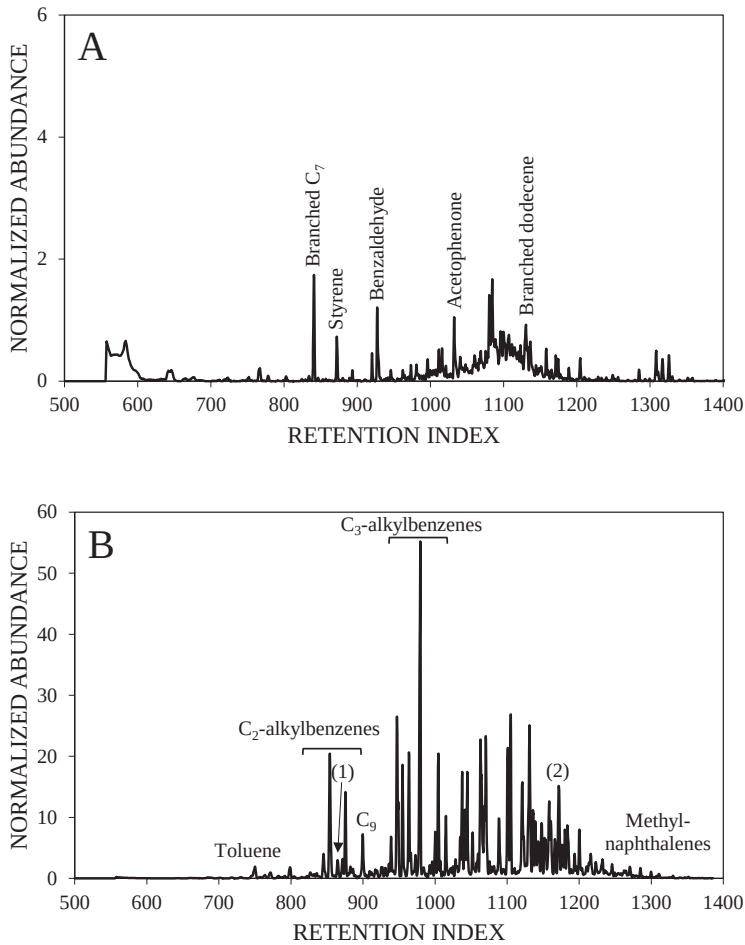


FIGURE 2.23 Identification of liquid in Burn Sample A based on total ion chromatograms (TICs) (A) experimental chromatogram of burned carpet, (B) experimental chromatogram of Burn Sample A, (C) comparison of burned carpet to predicted reference collection, and (D) comparison of Burn Sample A to predicted reference collection. In (B), substrate contributions are denoted as follows: (1) styrene and (2) estragole. Color and symbol designations in (C) and (D) are described in Figure 2.18. Maximum correlation indicated by *.

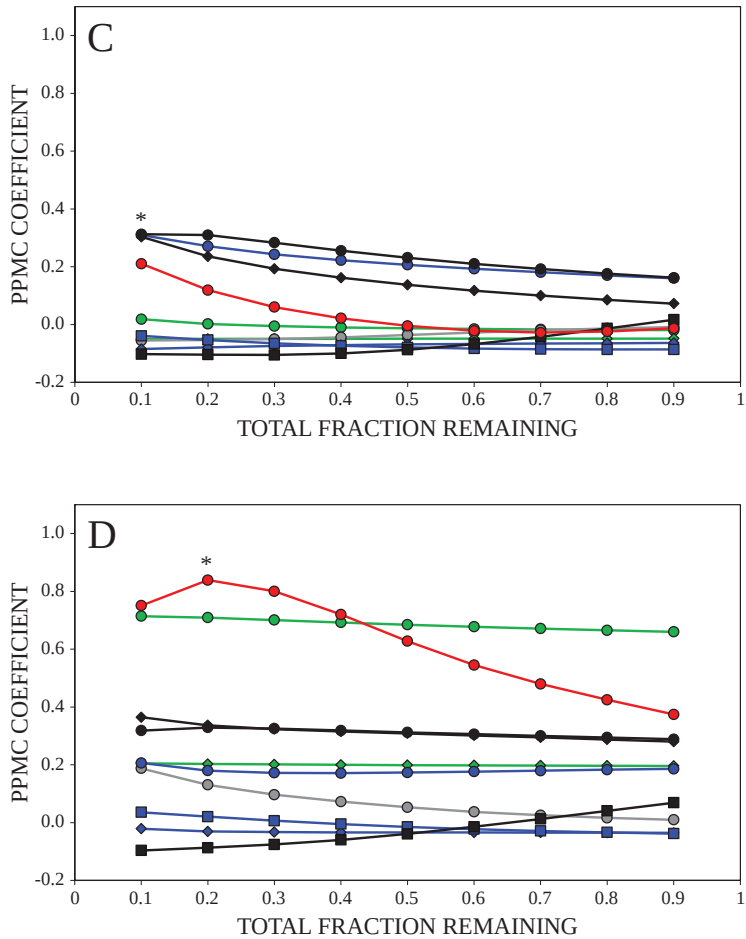


FIGURE 2.23 (Continued)

sample contains relatively low abundance of these compounds, such that the highest correlation is observed for comparison to a predicted chromatogram representing an advanced state of evaporation (F_{Total} = 0.2). It is also worth noting here that the maximum coefficient is lower than observed in Section 2.4.1.2 for different-source comparisons of gasoline (e.g., r = 0.9812—0.9863 for comparison of Gas F to the gasoline reference collection). Lower correlation is expected here due to the additional presence of substrate contributions (i.e., styrene and estragole) and combustion products in the fire debris sample that are not present in the reference collection chromatograms.

Burn Sample A also indicates moderate correlation to fruit tree spray, which is an aromatic liquid (Figure 2.23D). However, in contrast to the trend described above, correlation to this liquid remains relatively constant across the F_{Total} range (r = 0.7093 at F_{Total} = 0.2 to r = 0.6601 at F_{Total} = 0.9). The moderate correlation

observed is due to the common presence of C_3-alkylbenzenes eluting across the range I^T = 900—1100 in the fire debris sample and fruit tree spray. Due to their lower volatility, these compounds are less affected by evaporation than toluene and the C_2-alkylbenzenes. As a result, there is consistent, albeit lower, correlation of the fire debris sample to fruit tree spray than to gasoline. Comparison of Burn Sample A with all other liquids at all F_{Total} levels in the predicted reference collection yields PPMC coefficients less than 0.4, indicating weak to no correlation.

While comparison to the predicted TIC reference collection indicates the presence of gasoline in Burn Sample A, the EIPs can be used to increase confidence in

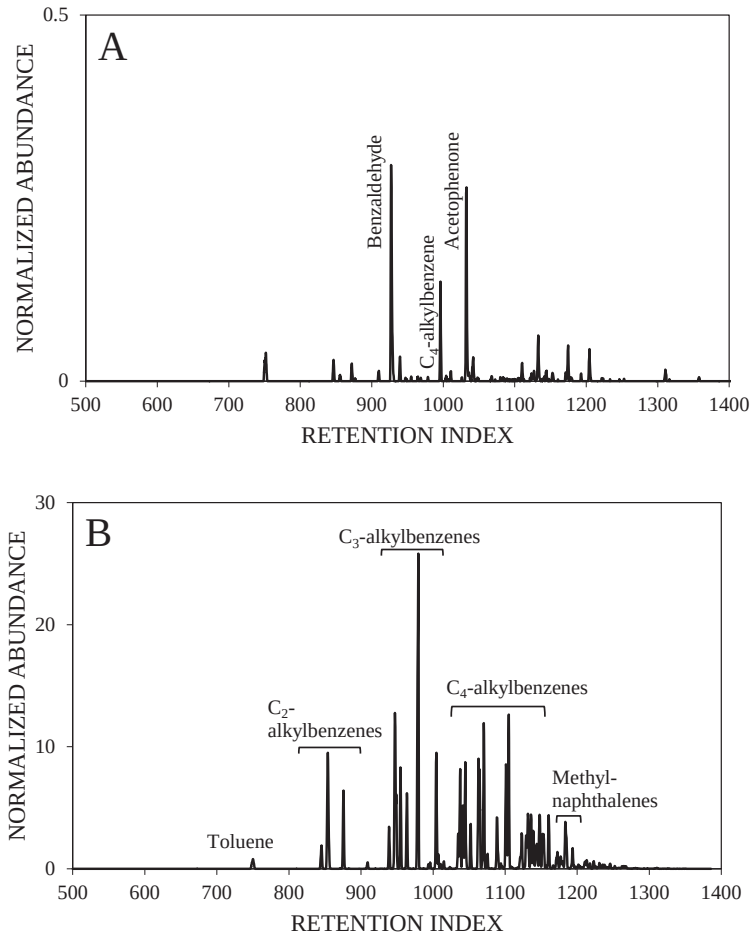


FIGURE 2.24 Identification of liquid in Burn Sample A based on extracted ion profiles (EIPs) (A) experimental aromatic EIP of burned carpet, (B) experimental aromatic EIP of Burn Sample A, (C) comparison of burned carpet aromatic EIP to predicted reference collection, and (D) comparison of Burn Sample A aromatic EIP to predicted reference collection. Color and symbol designations in (C) and (D) are described in Figure 2.18. Maximum correlation indicated by *.

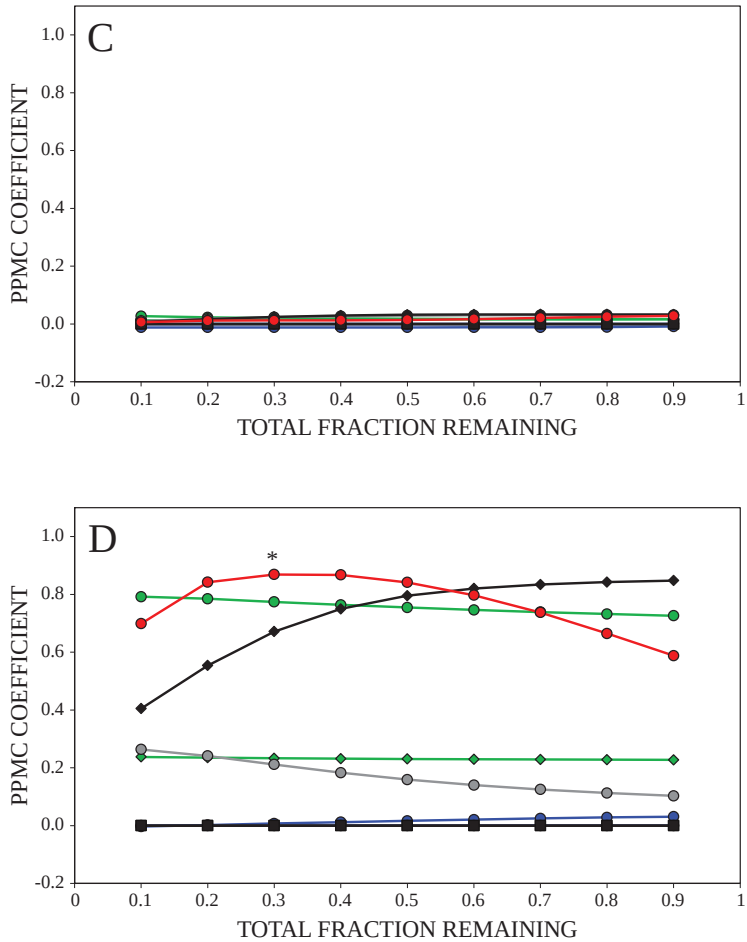


FIGURE 2.24 (Continued)

the identification. As an example, consider the aromatic EIP of both the burned carpet and of Burn Sample A (Figures 2.24A and B). The aromatic EIP of the burned carpet does not display a significant aromatic content and, when compared to the reference collection, displays no correlation with the aromatic EIP of any liquid at any F_{Total} level (Figure 2.24C). In contrast, the aromatic EIP of Burn Sample A displays the highest correlation to gasoline ($r = 0.8686$) predicted at $F_{Total} = 0.3$, indicating advanced evaporation in the burn sample (Figure 2.24D). Correlation between the burn sample and gasoline decreases as F_{Total} increases, which is consistent with evaporation of the volatile compounds, as noted previously. Although not strictly necessary, evaluation of the aromatic EIP serves to increase confidence in the identification of gasoline in Burn Sample A. It should also be noted here that the gasoline used in this burn cell was a different source than that included in the reference

collection, further demonstrating the broad applicability of the predicted reference collections.

2.4.3.2.2 Burn Sample B

Total ion chromatograms of an unburned wood flooring sample and a debris sample collected in this burn cell are shown in Figure 2.25. The unburned wood flooring sample contains primarily a substituted cycloalkene (I^T = 930), β-pinene (I^T = 969), along with hexanal, nonanal, and decanal (I^T = 775, 1083, and 1174, respectively)

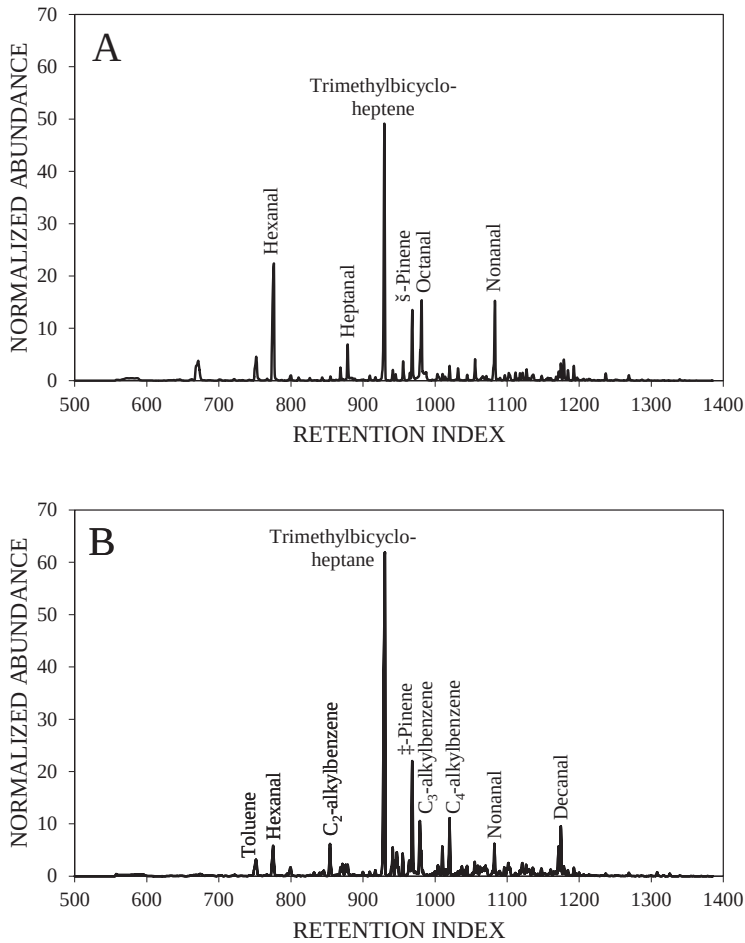


FIGURE 2.25 Identification of liquid in Burn Sample B based on total ion chromatograms (TICs) (A) experimental chromatogram of unburned wood subflooring, (B) experimental chromatogram of Burn Sample B, (C) comparison of unburned wood subflooring to predicted reference collection, and (D) comparison of Burn Sample B to predicted reference collection. Color and symbol designations in (C) and (D) are described in Figure 2.18. Maximum correlation indicated by *.

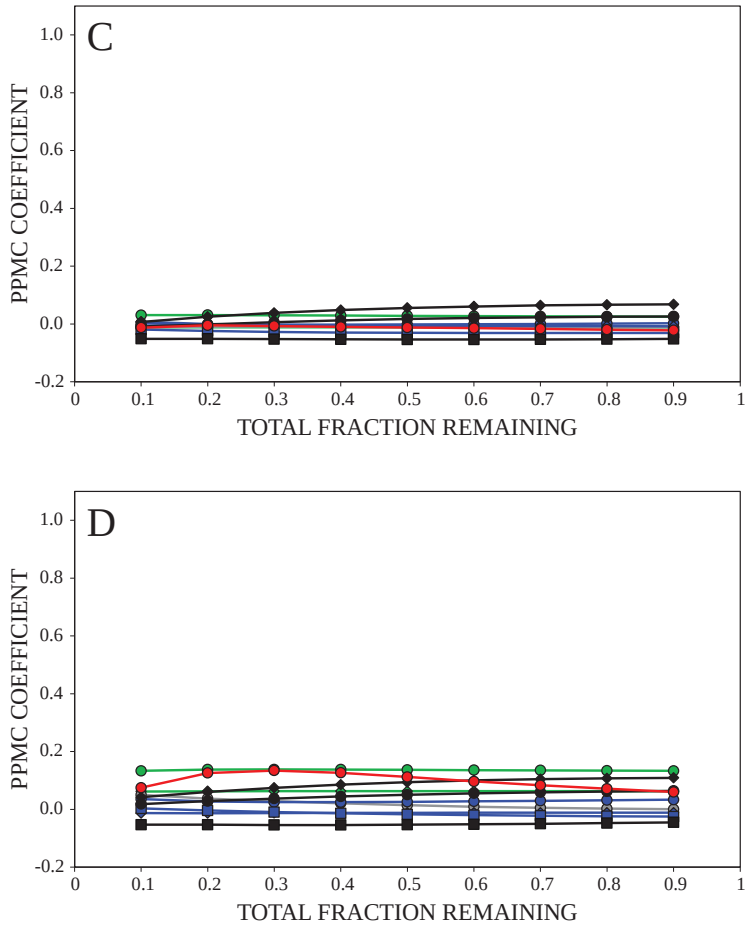


FIGURE 2.25 (Continued)

(Figure 2.25A). The TIC of Burn Sample B is remarkably similar and is dominated by these substrate contributions (Figure 2.25B). The burn sample also contains toluene, C_2-, C_3-, and C_4-alkylbenzenes, although these compounds are not in the ratios typically observed in gasoline. When the TICs of the unburned wood flooring and the burn sample are compared to the predicted reference collection (Figures 2.25C and D, respectively), there is no correlation with any liquid at any F_{Total} level. As such, given the extensive substrate interferences in the TIC of Burn Sample B, identification of any liquid present is not possible.

Extracted ion profiles corresponding to the alkane, aromatic, and indane classes were subsequently generated from the TIC of the burn sample. Compounds in the alkane EIP include hexanal, nonanal, and decanal, along with the *n*-alkanes C_8—C_{12} (Figure 2.26A). The aromatic profile contains compounds consistent with gasoline

(toluene, C_2-, C_3, and C_4-alkylbenzenes) (Figure 2.26B), while the indane profile contains indane, methylindane, and branched indanes (Figure 2.26C).

The alkane EIP from Burn Sample B was first compared to the predicted alkane EIPs (Figure 2.26D). For most comparisons, PPMC coefficients are less than 0.5, which indicates weak to no correlation. Weaker correlation is expected for these comparisons due to the extensive substrate contributions present in the sample. Despite such contributions, the highest correlation ($r = 0.5336$) is observed for comparison of the burn sample profile to the alkane EIP for gasoline corresponding to $F_{Total} = 0.1$.

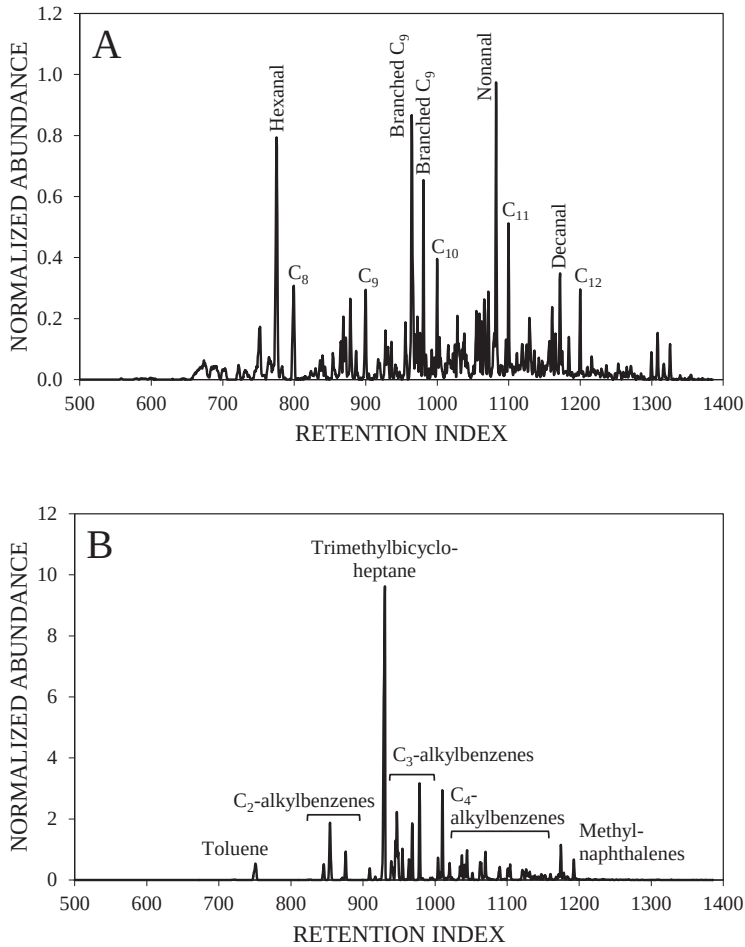


FIGURE 2.26 Identification of liquid in Burn Sample B using extracted ion profiles (EIPs) (A) alkane EIP of Burn Sample B, (B) aromatic EIP of Burn Sample B, (C) indane EIP of Burn Sample B, (D) comparison of alkane EIP to predicted alkane EIP reference collection, (E) comparison of aromatic EIP to predicted aromatic EIP reference collection, and (F) comparison of indane EIP to predicted indane EIP reference collection. Color and symbol designations in (D), (E), and (F) are described in Figure 2.18. Maximum correlation indicated by *.

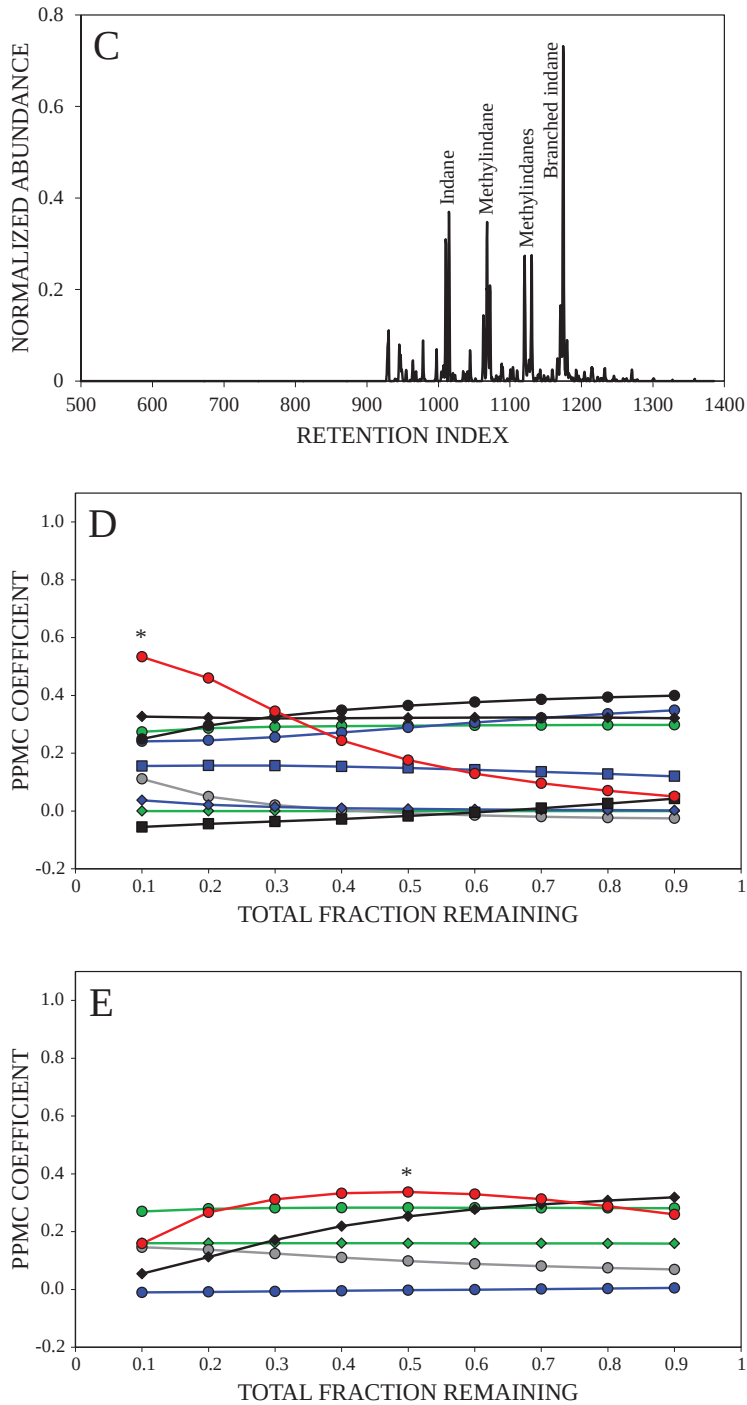


FIGURE 2.26 (Continued)

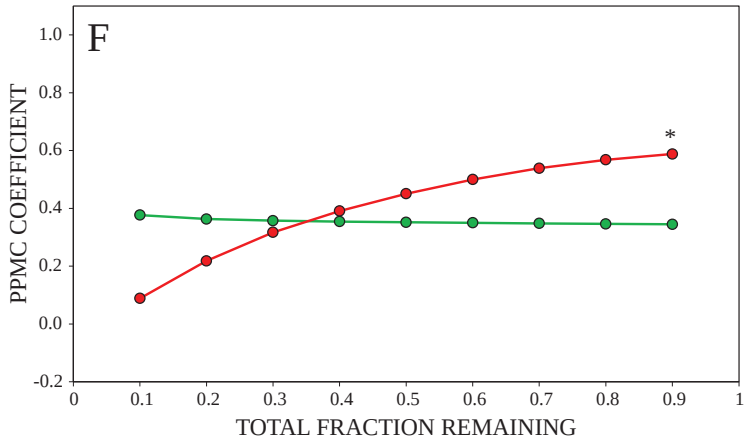


FIGURE 2.26 (Continued)

As F_{Total} increases, correlation decreases, which is consistent with evaporation of the more volatile compounds in gasoline, as noted previously. For the other liquids in the predicted reference collection, there is consistently weak to no correlation to the burn sample profile across all F_{Total} levels (Figure 2.26D).

The aromatic profile of burn sample B also shows weak to no correlation when compared to the aromatic EIP predicted reference collection (Figure 2.26E). The highest correlation ($r = 0.3370$) occurs for comparison to gasoline at $F_{Total} = 0.5$. As before, correlation to gasoline decreases as F_{Total} increases, consistent with evaporation of more volatile compounds. Higher correlation is also observed for comparisons to fruit tree spray (aromatic class, $r = 0.2829$ at $F_{Total} = 0.5$) and to paint thinner (petroleum distillate class, $r = 0.3186$ at $F_{Total} = 0.9$) than to any other liquid in the reference collection. However, correlation to fruit tree spray remains relatively constant across the F_{Total} range (Figure 2.26E). In contrast, correlation to paint thinner increases as F_{Total} increases due to higher aromatic content in this liquid at higher F_{Total} values.

Only two liquids in the predicted reference collection contain a significant abundance of indanes: gasoline and fruit tree spray (Figure 2.26F). The indane profile of Burn Sample B displays weak yet consistent correlation to the corresponding profile of fruit tree spray across all F_{Total} levels ($r = 0.3767$ at $F_{Total} = 0.1$ to $r = 0.3447$ at $F_{Total} = 0.9$). However, for gasoline, increasing correlation is observed as F_{Total} increases, reaching a maximum correlation of $r = 0.5880$ at $F_{Total} = 0.9$. This trend is opposite to that described above for comparison of the aromatic profiles. However, in this case, the indanes are considerably less volatile than the aromatic compounds and, thus, are less affected by evaporation. As such, these compounds remain in relatively high abundance in the burn sample, resulting in higher correlation to predicted profiles corresponding to higher F_{Total} levels.

Overall, despite extensive substrate interferences present in Burn Sample B, comparisons to the predicted EIP reference collections give some indication of the presence of gasoline. Such identification was not apparent based on the TIC alone, for which there was no correlation with any liquid in the reference collection. The alkane and aromatic EIP comparisons provide evidence for the possible presence of

gasoline, with correlation trends consistent with those expected due to evaporation of volatiles. However, the indane profile is perhaps the most informative. Of all 10 liquids in the reference collections, only two generated significant indane profiles. And, of these two liquids, higher correlation was observed for comparison to gasoline. In addition, the correlation trends observed are consistent with those expected for less volatile compounds. Thus, all three EIPs indicate the presence of gasoline in Burn Sample B and demonstrate the potential to exploit the chemical selectivity offered in a predicted EIP reference collection for identification purposes.

2.4.4 Summary

In this section, the fixed-temperature kinetic model of McIlroy et al. was demonstrated to accurately predict the evaporation of chemically diverse ignitable liquids [31]. Total ion chromatograms and extracted ion profiles were predicted and successfully used to identify liquids in fire debris samples collected from large-scale burns. Applying the model in this manner offers several advantages for future implementation in forensic fire debris analysis. Evaporation is predicted as a function of retention index, such that the chemical composition of the liquid need not be known in advance. The model is used to predict chromatograms corresponding to any F_{Total} level based only on chromatograms of unevaporated liquids. As such, predicted TICs and EIPs can be generated rapidly and used to create extensive and representative reference collections in a time- and resource-efficient manner. Finally, sample chromatograms are compared to each predicted chromatogram in the reference collection using correlation coefficients, thereby offering an objective, statistically based evaluation of the chromatograms for identification. Thus, in forensic science, the kinetic model can be employed to identify highly evaporated ignitable liquids in fire debris samples to indicate intentional rather than accidental fires.

2.5 CONCLUSIONS

Many existing models of evaporation rely on physical properties, such as boiling point, vapor pressure, or rate constant. For models that consider the sample as a single component, these properties must be experimentally measured as a bulk value from the original sample and the change in these properties with time must be estimated. For models that consider the sample as individual components or pseudocomponents, each component must be identified and its properties must be known, predicted, or measured. This limits the number of components that can realistically be accommodated for complex samples.

The kinetic model of evaporation developed in this work represents an important and timely advance. Because the model is based on the gas chromatographic retention index, a surrogate property, it is not necessary to identify the individual compounds and to determine their physical properties. Moreover, the kinetic foundation provides an accurate time and temperature basis. The kinetic model can be applied in several different modes. First, the regression parameters can be evaluated for a specific compound class (e.g., *n*-alkanes, branched and cyclic alkanes, alkyl benzenes, or polycyclic aromatics) or they can be evaluated comprehensively for all classes. Second, the regression

parameters can be evaluated for fixed-temperature or for variable-temperature conditions. The specific mode is selected according to the complexity of the modeling problem and the desired accuracy, allowing greater flexibility in the application.

Despite the advantages of using GC retention indices mentioned above, there are limitations to this approach. Accurate modeling requires that all components in the sample be sufficiently volatile for analysis within the normal range of GC operating temperatures. For nonpolar stationary phases such as polydimethylsiloxane, the temperature range is approximately 35—325°C, allowing for analysis of compounds in the range of *n*-pentane to *n*-triacontane or higher. Although this includes a wide range of refined petroleum products, it does not cover the entire range needed for crude oils. To extend the range to less volatile compounds, it will be necessary to use high-temperature GC columns and instrumentation to achieve temperatures of 400°C or higher. Similarly, to extend the range to more volatile compounds, it will be necessary to use cryogenic GC instrumentation to achieve temperatures less than 35°C.

In addition to the extended volatility range, it is important to explore a broader range of chemical compounds. Thus far, the applications have focused on modeling evaporation of petroleum fuels, whose primary constituents are hydrocarbons. The inclusion of more polar compounds, such as those containing heteroatoms (oxygen, nitrogen, sulfur), is necessary for industrial applications to foods, beverages, and fragrances as well as for homeland security and law enforcement applications to explosives and chemical warfare agents. These extensions will require models for specific compound classes, combined with suitable extracted ion profiles from the GC-MS data. With these modifications, the kinetic model can be extended to a wider range of potential applications.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the contributions to this research by former graduate students in the Department of Chemistry and the Forensic Science Program, School of Criminal Justice, at Michigan State University: Dr. John W. McIlroy (Drug Enforcement Administration, South Central Laboratory, Dallas, TX), Dr. Amanda L. Burkhart (Chemistry and Physics Department, University of Tennessee-Martin, Martin, TN), Dr. Briana A. Capistran (National Institute of Standards and Technology, Gaithersburg, MD), and Natasha K. Eklund (Georgia Bureau of Investigation, Decatur, GA).

Parts of this work were supported by Award No. 2018-DU-BX-0225, awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect those of the Department of Justice.

REFERENCES

[1] Fingas, M. 2017. *Oil Spill Science and Technology.* 2nd ed. Cambridge, MA: Elsevier.
[2] Stauffer, E., J.A. Dolan, and R. Newman. 2008. *Fire Debris Analysis*. Burlington, MA: Academic Press.

[3] Maarse, H. 1991. *Volatile Compounds in Foods and Beverages*. New York: Marcel Dekker.
[4] Witschi, C., and E. Doelker. 1997. Residual solvents in pharmaceutical products: acceptable limits, influences on physicochemical properties, analytical methods and documented values. *European Journal of Pharmaceutics and Biopharmaceutics* 43:215–242.
[5] Brevett, C.A.S., K.B. Sumpter, J. Pence, et al. 2009. Evaporation and degradation of VX on silica sand. *Journal of Physical Chemistry C* 113:6622–6633.
[6] Columbus, I., D. Waysbort, I. Marcovitch, L. Yehezkel, and D.M. Mizrahi. 2012. VX fate on common matrices: evaporation versus degradation. *Environmental Science & Technology* 46:3921–3927.
[7] Kudryashova, O.B., A.A. Pavlenki, S.S. Titov, and A.B. Vorozhtsov. 2021. A mathematical model for evaporation of explosive thin film. *Journal of Energetic Materials* 39:246–254.
[8] Nyholm Westin, S., S. Winter, E. Karlsson, A. Hin, and F. Oeseburg. 1998. On modeling of the evaporation of chemical warfare agents on the ground. *Journal of Hazardous Materials A* 63:5–24.
[9] Fingas, M.F. 1997. Studies on the evaporation of crude oil and petroleum products: I. The relationship between evaporation rate and time. *Journal of Hazardous Materials* 56:227–236.
[10] Fingas, M.F. 1998. Studies on the evaporation of crude oil and petroleum products: II. Boundary layer regulation. *Journal of Hazardous Materials* 57:41–58.
[11] Fingas, M. 2015. Oil and petroleum evaporation. In: *Handbook of Oil Spill Science and Technology*. Edited by M. Fingas. Hoboken, NJ: John Wiley & Sons.
[12] Fingas, M.F. 2004. Modeling evaporation using models that are not boundary-layer regulated. *Journal of Hazardous Materials* 107:27–36.
[13] Mackay, D., and R.S. Matsugu. 1973. Evaporation rates of liquid hydrocarbon spills on land and water. *The Canadian Journal of Chemical Engineering* 51:434–439.
[14] Berry, A., T. Dabrowski, and K. Lyons. 2012. The oil spill model OILTRANS and its application to the Celtic Sea. *Marine Pollution Bulletin* 64:2489–2501.
[15] Jones, R.K. 1996. Method for estimating boiling temperatures of crude oils. *Journal of Environmental Engineering* 122:761–763.
[16] Lehr, W., R. Jones, M. Evans, D. Simecek-Beatty, and R. Overstreet. 2002. Revisions of the ADIOS oil spill model. *Environmental Modelling & Software* 17:191–199.
[17] Stiver, W., and D. Mackay. 1984. Evaporation rate of spills of hydrocarbons and petroleum mixtures. *Environmental Science and Technology* 18:834–840.
[18] Stiver, W., W.Y. Shiu, and D. Mackay. 1989. Evaporation times and rates of specific hydrocarbons in oil spills. *Environmental Science and Technology* 23:101–105.
[19] Berry, A. 2011. *The Atlantic Regions' Coastal Pollution Response (ARCOPOL)*. Development of OILTRANS Model Code. Available at http://www.arcopol.eu/?/=/section/resources/sub/r_modelling_decision_support_tools/resource/41 (accessed February 2022).
[20] Jones, R.K. 1997. A simplified pseudo-component oil evaporation model. *Arctic and Marine Oilspill Program Technical Seminar*. Vancouver, BC, Canada.
[21] Bruno, T.J., L.S. Ott, B.L. Smith, and T.M. Lovestead. 2010. Complex fluid analysis with the advanced distillation curve method. *Analytical Chemistry* 82:777–783.
[22] Bruno, T.J., T.M. Lovestead, and M.L. Huber. 2011. Prediction and preliminary standardization of fire debris constituents with the advanced distillation curve method. *Journal of Forensic Sciences* 56:S192–S202.
[23] Bruno, T.J., and S. Allen. 2013. Weathering patterns of ignitable liquids with the advanced distillation curve method. *Journal of Research of the National Institute of Standards and Technology* 118:29–51.

[24] Birks, H.L., A.R. Cochran, T.J. Williams, and G.P. Jackson. 2017. The surprising effect of temperature on the weathering of gasoline. *Forensic Chemistry* 4:32–40.
[25] Willis, I.C., Z. Fan, J.T. Davidson, and G.P. Jackson. 2020. Weathering of ignitable liquids at elevated temperatures: a thermodynamic model, based on laws of ideal solutions, to predict weathering in structure fires. *Forensic Chemistry* 18:100215.
[26] Yaws, C.L., P.K. Narasimhan, and C. Gabbula. 2009. *Yaws' Handbook of Antoine Coefficients of Vapor Pressure* (electronic edition). 2nd ed. Knovel. Available at https://app.knovel.com/kn/resources/kpYHACVPEH/toc (accessed February 2022).
[27] Atkins, P., J. de Paula, and J. Keeler. 2018. *Atkins' Physical Chemistry*. 11th ed. New York: Oxford University Press.
[28] Regnier, Z.R., and B.F. Scott. 1975. Evaporation rates of oil components. *Environmental Science and Technology* 5:469–472.
[29] Okamoto, K., N. Watanabe, Y. Hagimoto, K. Miwa, and H. Ohtani. 2009. Changes in evaporation rate and vapor pressure of gasoline with progress of evaporation. *Fire Safety Journal* 44:756–763.
[30] Okamoto, K., M. Hiramatsu, H. Miyamoto, et al. 2012. Evaporation and diffusion behavior of fuel mixtures of gasoline and kerosene. *Fire Safety Journal* 44:47–61.
[31] McIlroy, J.W., A.D. Jones, and V.L. McGuffin. 2014. Gas chromatographic retention index as a basis for predicting evaporation rates of complex mixtures. *Analytica Chimica Acta* 852:257–266.
[32] McIlroy, J.W., R. Waddell Smith, and V.L. McGuffin. 2018. Fixed- and variable-temperature kinetic models to predict evaporation of petroleum distillates for fire debris applications. *Separations* 5:47–65.
[33] Burkhart, A.L., R. Waddell Smith, and V.L. McGuffin. 2021. Measuring evaporation rate constants of highly volatile compounds for use in predictive kinetic models. *Analytica Chimica Acta* 1182:338932.
[34] United States Energy Information Administration. *US Product Supplies for Crude Oil and Petroleum Products*. Available at https://www.eia.gov/dnav/pet/pet_cons_psup_dc_nus_mbblpd_a.htm (accessed February 2022).
[35] National Research Council. 2003. *Oil in the Sea III: Inputs, Fates, and Effects*. Washington, DC: National Research Council. Available at https://doi.org/10.17226/10388 (accessed February 2022).
[36] Fingas, M. 2017. Introduction to spill modeling. In: *Oil Spill Science and Technology*. Edited by M. Fingas. 2nd ed. Cambridge, MA: Elsevier.
[37] American Petroleum Institute. 1999. *Fate of Spilled Oil in Marine Waters*. Washington, DC: American Petroleum Institute.
[38] Fingas, M.F. 2013. *The Basics of Oil Spill Cleanup*. 3rd ed. Boca Raton, FL: CRC Press.
[39] Hayes, M.O., R. Hoff, J. Michel, D. Scholz, and G. Shigenaka. 1992. An introduction to coastal habitats and biological resources. *Hazardous Materials Response and Assessment Division*. National Oceanic and Atmospheric Administration. Seattle, WA.
[40] McIlroy, J.W. 2014. *Kinetic Models for the Prediction of Weathering of Complex Mixtures on Natural Waters* [Ph.D. Dissertation]. Department of Chemistry, Michigan State University, East Lansing, MI.
[41] Access, National Centers for Environmental Information (NCEI). *National Oceanic and Atmospheric Administration (NOAA)*. Available at https://www.ncei.noaa.gov/access (accessed February 2022).
[42] DeVore, J.L. 1990. *Probability and Statistics for Engineering and the Sciences*. Belmont, CA: Duxbury Press.
[43] Safety Data Sheet: Benzene. 2021. *Sigma-Aldrich*. Available at https://www.sigmaaldrich.com/US/en/sds/SIGALD/319953 (accessed February 2022).

[44] US Environmental Protection Agency. 2021. *Gasoline Mobile Source Air Toxics*. Available at https://www.epa.gov/gasoline-standards/gasoline-mobile-source-air-toxics (accessed February 2022).
[45] Wallace, W.E. 2022. Retention indices. In *NIST Chemistry WebBook, NIST Standard Reference Database Number 69*. Edited by P.J. Linstrom and W.G. Mallard. Gaithersburg, MD: National Institute of Standards and Technology. Available at https://doi.org/10.18434/T4D303 (accessed February 2022).
[46] Campbell, R. 2021. *Intentional Structure Fires*. National Fire Protection Association. Available at https://www.nfpa.org//-/media/Files/News-and-Research/Fire-statistics-and-reports/US-Fire-Problem/Fire-causes/osintentional.pdf (accessed February 2022).
[47] ASTM E1386-15. 2015. *Standard Practice for Separation of Ignitable Liquid Residues from Fire Debris Samples by Solvent Extraction*. ASTM International, West Conshohocken, PA.
[48] ASTM E2154-15a. 2015. *Standard Practice for Separation and Concentration of Ignitable Liquid Residues from Fire Debris Samples by Passive Headspace Concentration with Solid Phase Microextraction (SPME)*. ASTM International, West Conshohocken, PA.
[49] ASTM E1412-19. 2019. *Standard Practice for Separation of Ignitable Liquid Residues from Fire Debris Samples by Passive Headspace Concentration with Activated Charcoal*. ASTM International, West Conshohocken, PA.
[50] ASTM E1618-19. 2019. *Standard Test Method for Ignitable Liquid Residues in Extracts from Fire Debris Samples by Gas Chromatography-Mass Spectrometry*. ASTM International, West Conshohocken, PA.
[51] Capistran, B.A. 2020. *Kinetically Modeling Total Ion Chromatograms and Extracted Ion Profiles to Identify Ignitable Liquids for Fire Debris Applications* [M.S. Thesis]. Forensic Science Program, Michigan State University, East Lansing, MI.
[52] Capistran, B.A., V.L. McGuffin, and R. Waddell Smith. 2021. Application of a kinetic model to predict extracted ion profiles for the identification of evaporated ignitable liquids. *Forensic Chemistry* 24:100340.
[53] Eklund, N.K., B.A. Capistran, V.L. McGuffin, and R. Waddell Smith. 2020. Improvements in a kinetic-based model to predict evaporation of gasoline. *Forensic Chemistry* 17:100194.
[54] Eklund, N.K. 2019. *Further Investigation of a Kinetic Model to Accurately Predict Evaporation of Gasoline* [M.S. Thesis]. Forensic Science Program, Michigan State University, East Lansing, MI.
[55] Waddell Smith, R., R.J. Brehe, J.W. McIlroy, and V.L. McGuffin. 2016. Mathematically modeling chromatograms of evaporated ignitable liquids for fire debris applications. *Forensic Chemistry* 2:37–45.

3 Advanced QSRR Modeling in β-CD-Modified RP-HPLC System

Nevena Djajić, Ana Protić

CONTENTS

Prospective application of developed models in predicting retention, complex stability constants, and thermodynamic parameters as alternatives to experimentation

DOI: 10.1201/9781003330080-3

3.1 INTRODUCTION

High-performance liquid chromatography (HPLC) is the most widely applied analytical technique in contemporary drug quality control (QC) laboratories. Numerous positive features characterizing the technique contribute to its broad field of usage, although it possesses two main drawbacks: the time-consuming analysis and the consumption of high amounts of toxic organic solvents in comparison to other analytical techniques (1). The separation is mainly performed in the reversed-phase mode (RP-HPLC). Among different organic solvents, acetonitrile is labeled a gold standard in pharmaceutical analysis due to its outstanding physicochemical properties and chromatographic efficiency. On the other hand, acetonitrile is toxic, volatile, and flammable. Hence, the ecological acceptability of HPLC methods is an important issue that should be improved through the so-called process of "greening" the method. Developing green RP-HPLC methods should be considered an obligation owing to the negative influence of toxic organic solvents on human health and nature itself, especially when working in the pharmaceutical industry, which is designated to be in the service of human health. To exclude or reduce the amount of toxic organic solvents, different strategies could be used. One of the available approaches, which is relatively frequently used at present, is the addition of cyclodextrin (CD) in the mobile phase. In this way, CD-modified HPLC systems are created. CDs can form inclusion complexes with various analytes, improving their solubility in the aqueous phase and thus reducing their retention times. Consequently, the consumption of the toxic organic solvent is reduced, increasing the ecological acceptability of the developed HPLC method (2–5). This concept is also beneficial considering the overall price of the analytical methods and the fact that toxic organic solvents are partially or fully substituted with ecologically acceptable CD, a substance of semi-natural origin.

The presented concept could be extended to incorporate sustainability in the field of separation science, from the stage of method development to routine analysis, through the quantitative structure retention relationship (QSRR) modeling approach. The QSRR models represent the methods of mathematical modeling, commonly built by employing different types of machine learning algorithms. Earlier, the QSRR models linked solely the molecular characteristics (molecular descriptors) of the examined analytes to their retention. Using this approach, the retention could be predicted toward molecular descriptors at only one defined set of experimental conditions. When working with chromatographic methods, different experimental conditions largely influence retention and should therefore be included in the modeling. The QSRR models that include both experimental conditions and molecular descriptors at the same time

are denoted as mixed models, characterized by great predictive ability and utility in separation science (6, 7). In that respect, this book chapter is aimed at presenting the possibilities of building and employing QSRR models in CD-modified HPLC methods, followed by investigating their potential utility in different areas of application. When dealing with CD-modified RP-HPLC, challenges arise from its complexity. Namely, CDs from the mobile phase can be adsorbed onto the stationary phase and/or form inclusion complexes with analytes in the mobile phase. The CD-analyte complex is dynamic and characterized by certain equilibrium and stability constant values, depending on the type of CD and the analyte's structure. Taking into consideration all the presented facts, it is known that the solute can be distributed between the bulk mobile phase, stationary phase, CD in the mobile phase, and CD adsorbed onto the stationary phase, making the modeling in these kinds of chromatographic systems more complicated in comparison to regular RP-HPLC (8, 9). The complexity of the chromatographic system indicates the associated complexity of the modeling, which is reflected in the need for descriptors able to describe the formed inclusion complexes. These descriptors, labeled as complex association constants, could be included in the QSRR model along with experimental parameters and molecular descriptors (10). The apparent utility of the proposed model is in predicting the retention of examined model substances that can be further used in the selection of the optimal chromatographic conditions and the calculation of stability constant values of formed inclusion complexes (10, 11). In this way, the *in silico* approach is implemented in the development of CD-modified HPLC methods, replacing the time-consuming experimentation prior to verification of selected chromatographic conditions and method validation. Special attention should be paid to the possibility of utilizing these models in the determination of complex stability constants and accompanying thermodynamic parameters in general. Moreover, these properties could be of great importance in other scientific fields, apart from chromatography. For example, a CD forms inclusion complexes with various compounds, and in this way, the chemical stability of compounds could be improved as well as its water solubility, odor, taste, and other undesired properties. Therefore, CDs are employed in different segments of chemistry, pharmacy, food industry, etc. (12).

Sections 1 and 2 of the book chapter provide a concise overview of CD structures and present their influence on the formation of the inclusion complexes with different kinds of hydrophobic analytes. Analytes can encompass inorganic and organic molecules as well as bio-molecules. Special focus was placed on the practical and theoretical knowledge regarding the retention mechanisms and equilibria existing in CD-modified HPLC systems as a convenient introduction to retention modeling in this kind of chromatographic system, which is explained in Section 4.4. Prior to modeling, an overview of the analytical techniques used in the characterization of CD inclusion complexes is given in Section 3. Furthermore, Section 4.1 provides a detailed presentation of the molecular descriptors mostly applicable in modeling retention in HPLC, as well as complex association constants that can properly describe the formed inclusion complexes in CD-modified HPLC. In addition, the machine learning techniques were presented and discussed in terms of their utility and contribution in building good predictive QSRR models. One of the last sections (Section 5) summarizes the key experimental findings regarding the utility of the

obtained QSRR models in retention prediction and evaluation of optimal conditions of the chromatographic method, as well as thermodynamic parameters of complexation. A special theoretical discussion of possible benefits, drawbacks, and future perspectives of the presented approach is also presented.

This is only the beginning of the investigation in this broad, cutting-edge scientific field in which more questions have been asked than answered so far. Therefore, a new direction in research is potentially open. Using the QSRR models to predict the optimal chromatographic conditions and thermodynamic parameters of CD-analyte complexation is bringing a new dimension of greening and sustainability into separation science. The procurement of the retention factors by *in silico* methods, instead of performing the experiments, offers a great possibility of time, nature, and cost savings.

3.2 PROPERTY ASSESSMENT OF CD, FORMED INCLUSION COMPLEXES, AND CD-MODIFIED CHROMATOGRAPHIC SYSTEMS

3.2.1 Structure and Properties of CD

3.2.1.1 Brief Historical Overview

CDs were discovered by chance in 1891 by Viller. He isolated a material that "forms beautiful radiate crystals" after starch digestion with *B. amylobacter*. At that moment, it was named "cellulosine." Later, in 1903, Schardinger described the process of digesting starch with microorganisms, which resulted in the formation of two different crystalline products, dextrin A and B, characterized by a lack of reducing power similar to the "beautiful radiate crystals" described by Viller. Schardinger named the crystalline products "crystallized dextrin α" and "crystallized dextrin β" and discovered that they formed characteristic iodine adducts after the addition of iodine-iodide solution. Structural insight was provided by Freudenberg and colleagues in the 1930s. They discovered that crystalline structures contain only α-1,4-glycosidic bounds and further on, in 1936, postulated the cyclic structure of the dextrins. The exact determination of the structure of α-dextrin and β-dextrin was enabled in 1942, employing X-ray crystallography. It revealed that α-CD and β-CD were formed of 6 and 7 glucopyranosyl units, respectively. One of the most important discoveries regarding CDs was in 1948, when Freudenberg et al. recognized the possibility of CDs to form inclusion complexes. Soon, in 1953, Freudenberg, Camer, and Plieinger utilized this ability and applied CDs in drug formulations for protection from oxidation, enhancement of solubility, and stabilization of volatile substances. This was a historical moment, and it demonstrated the non-toxicity of CDs. Afterwards, the utilization of CDs was carried forward, and their areas of application were simultaneously broadened. At the same time, other types of CDs were found, including γ-CD, δ—CD, ζ—CD, ξ—CD, η-CD, consisting of 9 to 12 glucopyranosyl units. These CDs with a larger degree of polymerization were named large-ring cyclodextrins (LR-CD) and at first consisted of up to 13, while afterwards up to 45 glucopyranosyl units. All the described CDs are of semi-natural origin. Synthetic derivatives were later obtained in order to highlight the desired properties of CD. In that sense, hydroxypropyl-*β*-CD and *γ*-CD, randomly

methylated *α*- and *β*-CDs, maltosyl-*β*-CD, acetylated CD, and others were synthesized and widely used in different application areas (13, 14).

From 1950 to 1970, the research was oriented toward the investigation of CD structures and their inclusion complexes, along with their application in catalysis and enzyme models. Since 1970, CDs have been applied in different sectors of industries, such as the pharmaceutical and food industries. In these industries, CDs found different purposes, from odor and taste masking ability to enantioselective catalysts, drug carriers, and additives in separation science. Among separation techniques, CDs are widely used in gas chromatography (GC), HPLC, supercritical fluid chromatography (SFC), capillary electrophoresis (CE), and capillary electrochromatography (CEC). One of its prominent characteristics used in separation science is the so-called chiral or molecular recognition. Therefore, it is used in the separation of chiral compounds, in the first-line enantiomers. CDs are recognized as universal chiral selectors and utilized both as stationary phase modifiers (commercially available from 1984) and as mobile phase additives (12, 15, 16). However, one of its newest applications is to contribute to green liquid chromatography method development if used as RP-HPLC mobile phase modifiers (4).

3.2.1.2 Categorization

The need to improve the physicochemical properties of CDs was recognized over time, so CD derivatives and branched CDs were successfully synthesized by chemical or enzymatic modifications. In this way, the solubility, complex formation efficiency, chemical stability, and various other properties were improved. Along with the improved properties, the applicability of CDs was slowly extended. The free hydroxyl groups of glucopyranosyl units represent a convenient position for the potential structural modification. For each glucopyranosyl unit, there are two secondary and one primary free hydroxyl groups; therefore, reactions of modification could occur by substituting the hydrogen atom or the whole hydroxyl group.

Based on their structure, all currently known CDs can be divided into the following categories (13):

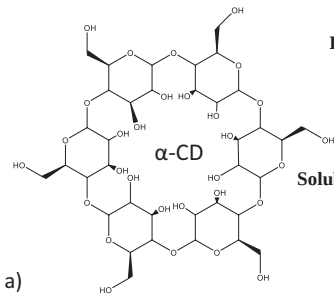
- *Small natural CDs*
  - This group consists of α-CD, β-CD and γ-CD.
- *CD derivatives*
  - CD derivatives are obtained by amination, esterification, or etherification. In this way, numerous functional groups could be incorporated in the structure of CD, namely methyl, sulfate, nitrate, phosphate, acetyl, benzoyl, propionyl, carbamoyl, hydroxypropyl, hydroxyethyl, etc. CD derivatives have modified solubility and hydrophobic cavity volume. In that respect, it could be concluded that structural modifications influence the inclusion process of the guest molecule into the cavity. The solubility and chemical stability of the guest molecule are also altered.
  - Furthermore, homogeneous and heterogeneous CD derivatives are distinguished depending on whether the CD hydroxyl groups are modified with the same or diverse functional groups.

- *Branched CDs*
  - Branched CDs, obtained if glucose, maltose, galactose, mannose, or any other oligosaccharide units are bound to the native CD through its hydroxyl groups, are labeled as the second generation of CD. If only glucose or malto-oligosaccharide units are added, homogenous branched CDs are obtained. Conversely, when galactose or mannose units are added, heterogeneous branched CDs are acquired. If only one unit is added to the native CD, a single branched CD is obtained, while adding two or more units leads to multiple branched CDs. The solubility of these kinds of CDs is higher in comparison to native CDs, although the improvement in solubility depends on the degree of derivatization. Apart from changes in solubility, the availability of the internal hydrophobic cavity is altered as well.
- *Large ring CDs*
  - Large ring CDs (LR-CDs) consist of more than eight glucopyranosyl units (13).

3.2.1.3 Structure and Physicochemical Characteristics of CDs

α-, β- and γ-CDs consist of 6, 7, and 8 glucopyranosyl units (Figure 3.1a, 3.1b, and 3.1c), bound together with α-1,4-glycosidic link in a truncated cone-shaped structure characterized by a hydrophobic cavity delimited with two edges, a narrow and a wider one. Both oxygen atoms from the α-1,4-glycosidic bridge (ether-like oxygen) and hydrogen atoms (apolar C-3 and C-5 hydrogen, Figure 3.1d) form the CD cavity. The non-bound electron pairs originating from ether-like oxygen atoms are oriented toward the inside of the cavity. In this way, the interior of the cavity is hydrophobic and possesses the base characteristics, enabling hydrophobic and electrostatic interactions with analytes and solvents. In fact, the CD cavity is rarely empty, and molecules of water, acetonitrile, methanol, and other solvents, or their mixture, fill its interior. This property is one of the most significant CD characteristics, because almost all CD applications involve the formation of inclusion complexes with various relatively hydrophobic compounds, which is thermodynamically favorable regarding the water-CD interactions (13, 17). The prominent property related to the ability of CD to form inclusion complexes is the hydrophobic cavity diameter. The internal cavity diameter increases with the increase in the number of glycoside units incorporated in the structure of the CD molecule. The values of the internal cavity diameters for α-, β- and γ-CDs together with the remaining significant CD characteristics are presented in Figure 3.1.

The edges of the CDs consist of free primary and secondary hydroxyl groups. The primary hydroxyl groups are situated at the narrower edge, while the secondary hydroxyl groups are positioned at the wider edge of the cavity. This kind of disposition is not randomly determined but rather specified with free rotation of the latter that reduces the effective diameter of the edge. Since hydroxyl groups are hydrophilic and positioned outside the cavity, they enable CDs' solubility in water. Generally, molecules with a larger number of glucopyranosyl units possess greater solubility in water. In fact, this is true when LR-CDs are considered but is not when dealing with small native CDs. The solubility of β-CDs is significantly lower compared to α- and γ-CDs. This abnormality can be explained by the so-called second



a)

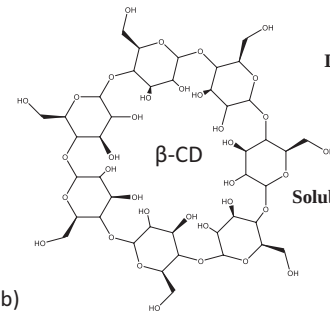
Diameter of the hydrophobic cavity - 500

6 water molecules in the cavity

Solubility in water (at room temperature, g/100mL) – 14.5

Slightly distorted simmetry

Consisted of 6 glucopyranosyl units



b)

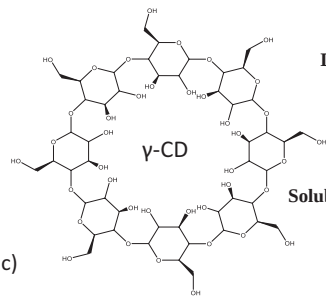
Diameter of the hydrophobic cavity - 620

11 water molecules in the cavity

Solubility in water (at room temperature, g/100mL) – 1.85

Perfect simmetry

Consisted of 7 glucopyranosyl units



c)

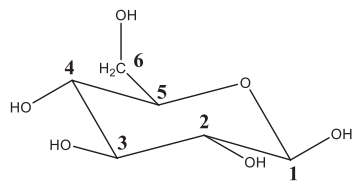
Diameter of the hydrophobic cavity - 800

17 water molecules in the cavity

Solubility in water (at room temperature, g/100mL) – 23.2

Slightly distorted simmetry

Consisted of 8 glucopyranosyl units



d)

FIGURE 3.1 CD structures and significant characteristics

1a: α-CD
1b: β-CD
1c: γ-CD
1d: glucopyranosyl unit

belt of CDs and water-β-CD thermodynamic properties. Namely, the secondary hydroxyl groups on the C-2 of the glucopyranosyl unit can form a hydrogen bond with the C-3-OH group of the neighboring glucopyranosyl unit. In this way, the H-bonds form a belt around the wider edge, which is called a secondary belt. In the case of α-CD, the secondary belt is not completely formed, since one glucopyranosyl unit is in a distorted position. This is a reason for the formation of four rather than six possible H-bonds, leaving more hydroxyl groups to interact with the molecules of water and enhance solubility. The β-CD forms a complete secondary belt, exhibiting the lowest solubility in water. The lower water solubility of β-CD can be explained with water-β-CD interactions, when a favorable enthalpy followed by the unfavorable entropy of solutions is occurring, as well (18). Further, the β-CD is insoluble in most organic solvents, though the solubility is improved in the water-organic solvent mixtures. The general rule is that solubility of CDs decreases with an increase in the amount of organic solvent. This is not the case for ethanol, propanol, and acetonitrile, where the solubility reaches its maximum of 20–30% of this organic solvent in water (13). Conversely, the γ-CD has a non-coplanar, more flexible structure that is freely soluble in water (13, 17).

On the other hand, though LR-CDs possess a large number of glucopyranosyl units, they do not exhibit wider cavity diameters, due to conformational changes of these molecules. When dealing with CDs consisting of 9 glucopyranosyl units (CD-9), the molecular shape looks like a boat. For CDs consisting of 10 to 14 glucopyranosyl units, the molecular shape resembles a saddle, and for molecules possessing 20 glucopyranosyl units and lager, the cavity structure is similar to the helicoidal-like channels. In that manner, these CDs do not possess wider cavities, though they are largely soluble in water, as expected (13, 17).

Flexibility is also considered a valuable CD characteristic. Though CDs are built from rigid glucopyranosyl units, numerous experimental and theoretical data indicate that ether-like oxygen bonds that link those rigid units possess a low barrier to internal rotation equal to 1 kcal mol^{-1}. This discernment was supported by the "model molecular mechanics calculation on α-CD showing that planar structure does not correspond to the energy minimum and the energy hypersurface exhibits several energy minima separated by low barriers." The concept of a rigid structure is insupportable with the extensive formation of inclusion complexes with analytes of different structures. The studies of the complexation process and complexes implied the effective fitting of the host and guest molecules to each other. Experimental evidence of CDs' non-rigid structure came from NMR studies, both in solution and in solid state. The most rigid structure is related to β-CD due to the formation of a complete secondary belt. In fact, it could be noticed that all phenomena related to β-CD's properties are associated with the described secondary belt (13, 17, 19, 20).

CDs are stable to different degradation reactions but are susceptible to hydrolysis and oxidation. Generally, CDs are more resistant to hydrolysis than starch, stable when exposed to bases, but strong acids (like hydrochloric acid and sulfuric acid) could cause their hydrolysis, resulting in the formation of a mixture of oligosaccharides. The hydrolysis can occur at the α-(1,4)-glycosidic bound and open ring down to glucose. The rate of hydrolysis increases simultaneously with the increase in temperature and concentration of acid. On the other hand, organic acids and weak acids

are not able to significantly influence hydrolysis. The stability of LR-CDs is jeopardized, since the number of α-D-1,4 linkages as decomposition points increases. In this way, LR-CDs are vulnerable to hydrolysis and form a large number of degradation products. The susceptibility of CDs to the oxidation process is reflected in glucose ring oxidation, and even perforation of the glucose ring can easily occur. The literature survey revealed that strong oxidation agents, along with elevated temperatures around 50°C, is needed to induce degradation, though hydrogen peroxide and weak oxidation agents can cause oxidation to a smaller extent and less rapidly (13, 21).

3.2.2 Inclusion Complexes Formed Between CD and Various Guest Molecules

The ability to form inclusion complexes is highlighted as one of the most important CD properties since it is responsible for most of its significant applications. Upon complexation, the guest molecule is temporarily locked or caged inside the CD cavity, being partially protected from the environment, which provides the entrapped molecules with different physicochemical characteristics. In that respect, the solubility in water could be improved when dealing with hydrophobic analytes, the stability toward oxygen, light and heat could be increased, the undesired odor and taste could be masked and the separation and isolation of analytes enabled. For all those reasons, it is not surprising that the complexation process is constantly intriguing scientists worldwide and forcing them to dive deeper into the field (12, 16, 21, 22).

The guest molecule could be fully or partially incorporated into the CD cavity. Inclusion complexes are formed in both solid state and solution, though in solutions complexes prevail in the rapid equilibrium of free CD and guest molecule. In most cases, it is approximated that one molecule of CD is complexed with one guest molecule. However, the stoichiometry of CD and guest molecule can differ, from 1:1 to 1:n or n:1. In other words, CDs with larger cavities, such as γ-CD, can accommodate more than one guest molecule at the same time, while several CDs can be linked by different parts of a large guest molecule. Though the assumption of 1:1 stoichiometry is frequently present in research papers, this stoichiometry cannot be taken for granted (8, 16).

The driving forces that govern the complexation process are still not completely understood, although years of research have been committed to this topic. In order to accomplish the inclusion complex formation, it is necessary to galvanize the equilibrium transfer in the direction of inclusion complex formation. This could be induced by some of the following interactions:

- Extrusion of water molecules from hydrophobic CD cavity
- Increased formation of hydrogen bonds with supplanted water molecules
- Reduction of repulsive interactions between the hydrophobic guest molecule and aqueous environment
- Hydrophobic interaction increase upon entrance of the guest molecule in the hydrophobic CD cavity (23–25).

Consequently, influential factors toward inclusion complex formation include CD type and its cavity dimensions, temperature, as well as pH affecting the ionization form of the investigated guest molecules (26). Although the initial equilibrium for inclusion complex formation is quite fast, reaching the final equilibrium usually takes much more time, mostly due to conformational adjustments of the guest molecule within the cavity. On the other hand, dissociation of the formed complex driven by the increase in surrounding water molecules is very fast. Therefore, in dynamic systems, there are certain difficulties for the guest molecule to find another CD and regenerate the inclusion complex, thus leaving it free in the solution (23). Equilibrium is characterized as one of the prominent association properties, and there is a simple rule saying that the faster the complexation is, the faster the dissociation will be.

During CD-guest complexation, different kinds of non-covalent interactions occur, ranging from van der Waals interactions to hydrogen bonds, dipole-dipole interactions and London dispersion forces. Most of them are hydrophobic and are responsible for stable inclusion complex formation. Different forces can be involved in the complexation process, but their number and type cannot be predicted. In light of this fact, it is difficult to estimate how well a particular guest will accommodate within the CD cavity, though this information would be very useful, especially when choosing between natural and chemically modified CDs with different cavity diameters (12, 22).

The efficiency of complexation depends largely on the structural characteristics of both the CD and the guest molecule supported by their mutual geometric compatibility. Although the structure of β-CD is the most rigid in relation to other native CDs, it is the most commonly used CD in pharmaceutical formulations. The CD cavity is characterized by its height and internal diameter determined by the number of glucose units. In that respect, the internal diameter of α-CD is lower than β-CD and γ-CD thus being able to incorporate low molecular weight compounds with aliphatic chains. Moreover, β-CD would accommodate heterocyclic and aromatic compounds, covering a broad range of pharmaceutical active compounds, while γ-CD would accommodate complex macrocycles and steroids (23). Another reason for the preferred use of β-CD in RP-HPLC is based on its characteristics in terms of weak adsorption onto C18 columns. Therefore, column performances would remain intact, β-CD could be easily washed and thus cause less damage to the column in comparison to the other CD (10).

Other structural characteristics, such as polarity and charge, also play an important role in the complexation process. Analytes in their ionized state are less susceptible to complexation and decomposition. On the other hand, analytes in their neutral form are readily complexed compared to their ionized forms. To form inclusion complexes, the analytes should be less polar than water. Moreover, analytes with more pronounced hydrophobic properties are able to form complexes with significant stability. Complexation efficiency is also highly dependent on the medium in which the complexation process takes place. In theory, the complexation process does not require any specific solvent, but it requires a small amount of water to galvanize the thermodynamics. The co-solvents can enter the cavity, inducing the threefold complex formation. Generally speaking, the complexation process is

influenced by enthalpic and entropic energies, both dependent on CD–guest molecule fit, solvents used, and other factors involved in this challengeable process of complexation (12, 21).

The described complexation process is valid when dealing with small CDs. LR-CDs have not been extensively studied as small CDs, so a certain gap in knowledge is present. As already described, LR-CDs do not have a regular cylindrical shape. In fact, their cavity has an irregular collapsed shape, inducing a smaller cavity diameter compared to γ-CD. In order to form inclusion complexes, LR-CDs must be highly flexible. Since these kinds of CDs consist of many more glucopyranose units than other CDs, there are numerous water molecules inside their non-polar cavity (12).

The complexation ability of CDs might serve for molecular recognition of enantiomers and closely related compounds, and counting geometrical and structural isomers. During the formation of inclusion complexes, CD can recognize slight differences in the structural characteristics of the guest molecules that are exhibited through non-covalent interactions. For example, β-CD is 100% selective for β-naphthalenesulfonate and α-naphthalenesulfonate. The molecular recognition and selective interactions of CDs with analytes are of great importance in separation science and offer a great possibility for developing highly selective analytical methods. If working with chiral analytes, natural CDs are known as substances with limited molecular recognition capacity. In that sense, it is advisable to utilize synthetic CD derivatives, with introduced functional groups, to enhance the binding selectivity through π-π interactions and π-CH hyperconjugation (8, 12).

CD-analyte inclusion complexes are usually spontaneously formed. However, there are methods proposed to induce the inclusion complex formation process, such as co-precipitation, slurry, paste and dry mixing, damp mixing, heating, and extrusion methods. All of them could be characterized as waterless. However, water is significant and should be a part of the medium in which both CD and analyte are dissolved. Although representing the driving force for complexation, water is sometimes essential to maintain the complex integrity. Water is not only important when dealing with solutions but also with crystal forms of the complexes, where they can form a bridge between the hydroxyl groups of the adjacent molecules of the CD. In the solid state, the magnitude of the crystal forces is comparable to the forces keeping the complex together. To reveal the structure of the formed complexes, different analytical methods have been proposed in the literature (12, 21).

3.2.3 CD-Modified RP-HPLC Systems

RP-HPLC systems modified by addition of CD in the mobile phase are considered dynamic and rather complicated since the solute could be distributed between the stationary phase, mobile phase, and CD dissolved in the mobile phase (8). Under certain conditions, the adsorption of CD onto the stationary phase surface could cause the formation of a so-called pseudo-stationary phase. For example, methylated β-CD could enable chiral resolution due to its strong adsorption onto the hydrophobic stationary phase (27). Depending on the investigated solute, any CD could possibly be used as a mobile phase additive. However, when dealing with pharmaceutical compounds, the use of β-CD is preferred over other CDs,

since it is able to accommodate most heterocyclic and aromatic compounds. As mentioned in section 2.2, β-CD is weakly adsorbed onto the stationary phase surface; thus, it is often used as an RP-HPLC additive. Throughout the literature, when assessing the apparent stability constants of inclusion complexes formed with various drug molecules, both native and modified β-CDs are mostly applied (28–30). The complexity of CD-modified RP-HPLC systems arises from the possibility for the formation of multiple interactions, while resolution and separation efficiency depend on different experimental conditions, such as type and concentration of applied CD, mobile and stationary phase characteristics, as well as column temperature. As already explained, inclusion complex formation in aqueous solutions is driven by the release of enthalpy-rich water molecules surrounding the CD from its cavity. These water molecules are supplanted with more hydrophobic molecules, so energetically more favorable non-polar interactions are established. So far, much work has been dedicated to revealing the structure of inclusion complexes and retention behavior in these kinds of chromatographic systems (10, 31, 32). However, although it is known how the solute is distributed, researchers are still not completely familiar with the influence of the structure of analytes and experimental parameters on retention mechanisms occurring in these RP-HPLC systems (11). CD modified RP-HPLC systems are still not investigated enough to be sure which retention mechanism would prevail and lead to the retention.

3.3 CHARACTERIZATION OF CD INCLUSION COMPLEXES IN SOLUTION

CDs are often considered to be challenging molecules from the analytical characterization point of view, but they are worth the effort because of their wide range of applications. However, to fully exploit their potential, the analytical technique for their comprehensive characterization should be readily available (33, 34). The first step in the characterization of inclusion complexes is the determination of stoichiometry and complex stability constant (K), as a quantitative parameter of the binding strength (35). Determining the complex stability constant is an inevitable step since different complexation-related effects depend on the corresponding inclusion complex stability.

Often, it is necessary to employ more than one analytical technique in order to provide insight into the interactions established during the complexation. In this way, the results of different techniques are combined and various complexation features are assessed. Incorporation of a guest molecule into the CD cavity causes alterations in the physical or chemical properties of the given guest molecule. Detecting the suitable changes occurring upon complexation provides a basis for determining the stability of the formed complex. Additionally, the observed change needs to be large enough to enable measurement precision (33).

In general, the methods used to characterize inclusion complexes can be divided into several groups, with certain subcategories (33, 35). Spectroscopic methods consist of UV-Vis spectroscopy (36, 37), fluorescence spectroscopy (38), and nuclear magnetic resonance (NMR) spectroscopy (36, 39, 40). When determining K with

UV/Vis spectroscopy, the guest molecule is used in fixed concentration, while CD is added in increasing concentrations and the change in absorbance peak of the guest molecule is monitored. The analysis is rather simple due to the non-absorbable nature of CD. The obtained absorbance values are then fitted to binding models based on the Benesi-Hildebrand, Scott, or Scatchard methods in order to calculate K from the slope and intercept of the constructed plots. One of the recognized disadvantages of these linear approaches is that they assume that equilibrium and initial CD concentration are the same. Additionally, the changes in absorbance are assumed to be proportional to the concentration of the formed complex, which in this case would be fully formed at the end of the titration. Moreover, a unique wavelength for all spectra should be carefully chosen due to bathochromic or hypsochromic shift of the maximum absorbance wavelength of the guest molecule upon its insertion in the CD cavity (35). Currently, non-linear regression is gaining precedence over linear transformations due to the development of algorithms able to postulate K values and compare them to experimentally obtained ones (35). The method established by Landy et al. enabled the determination of K using derivatives of the spectra, which helped diminish experimental error and/or spectral variations related to the rarely seen weak absorbance of CD (41).

On the other hand, if the guest molecule is present in low concentration and shows fluorescence, fluorescence spectroscopy could be used for characterizing the formed inclusion complexes. NMR is also used in the investigation of the inclusion complexation phenomenon, mainly due to its ability to elucidate the conformational accommodation of the guest inside the cavity and assess K values at the same time. K values are assessed based on the changes in the chemical shifts occurring when the concentration of the guest and/or CD is changed. Benefits brought with NMR, unlike other available methods, are reflected in the possibility of discovering the structure of the formed complex through detected variations (35).

If the guest molecule is electroactive, electroanalytical techniques, especially polarographic and voltammetric, are extensively used (42, 43). Furthermore, isothermal titration calorimetry (ITC) is the only method that simultaneously provides information on both thermodynamic aspects of complexation as well as K (33, 35, 44). Upon binding, a heat flow is generated, allowing a real measurement of binding interactions, determination of complex stoichiometry and K, as well as thermodynamic parameters of binding, such as enthalpy or entropy changes. ITC also has advantages shown through a shorter analysis time and a smaller sample in comparison to other methods (33).

Separation techniques, such as HPLC and capillary electrophoresis (CE) (45) or affinity capillary electrophoresis (ACE) (46), are also used in the analytical characterization of CD inclusion complexes and consequent calculation of K. In CE, the difference in the ion mobility or affinity of charged/uncharged molecules to charged electrolytes is the basis for separation, while ACE complements analyses of affinity effects, such as electrostatic interactions, hydrogen bonding, and van der Waals interactions. CE also finds its purpose in analyzing CD inclusion complexes with charged guest molecules (33). However, all separation techniques lack the ability to provide direct structural information on formed inclusion complexes (33).

In recent years, total organic carbon (TOC), a method usually known for testing water quality, has been used to analyze CD inclusion complexes. This method is not specific; thus, it finds its purpose in analyzing compounds lacking chromophores or fluorophores in their structure (47).

3.3.1 Determination of K by Means of HPLC

HPLC has the potential to be used in K determination if the mobile phase is modified with CD. The guest molecule is accommodated within the CD cavity and eluted from the column by the order of the highest inclusion complex stability. To be able to use HPLC to determine K, stationary phase properties should remain intact; namely, adsorption of CD onto the stationary phase surface should be very weak. However, in practice, free analyte, free CD, and CD-analyte complex could be adsorbed on the stationary phase, creating a miscellaneous environment. Although powerful for determining stoichiometry and K in solution, HPLC often requires extensive sample preparation as well as strict control of experimental conditions, enabling data reproducibility (23, 48, 49).

If K is determined in the HPLC environment, the difference in retention factor of the guest molecule upon complexation is followed. K is calculated based on the following equation (3.1):

$$\frac{1}{k}=\frac{1}{k_0}+\frac{K\left[CD^x\right]}{k_0} \tag{3.1}$$

where k is retention factor of the guest molecule forming an inclusion complex with CD, k_0 is retention factor of the guest molecule if CD is not present in the mobile phase, [CD] is CD concentration in the mobile phase, x is previously determined complex stoichiometry, and K is complex stability constant. The presented Equation 3.1 was developed for the chromatographic assessment of K, and it is extensively used throughout the literature (28, 29, 48–53). To determine K, 1/k versus [CD] graph is constructed. If the obtained graph is linear, K is calculated from the slope and intercept of the constructed graph. This approach is beneficial due to the possibility of simultaneous assessment of thermodynamic parameters of complexation by varying the temperature while conducting the experiments (50).

The chromatographic approach was successfully applied to determine the stability of *trans*-resveratrol:β-CD inclusion complex (50). It was concluded that the complex stability is largely affected by mobile phase composition, as well as guest molecule structure. β-CD concentration was increased up to 2.5 mM, and as expected, the retention time of the guest molecule simultaneously decreased due to inclusion complex formation, regardless of the amount of organic modifier in the mobile phase. However, it is observed that the decrease in retention time is higher with the lowest content of organic modifier. In this study, methanol was chosen as an organic modifier since its affinity toward β-CD is not substantial; specifically, methanol forms the association of 0.32 M^{-1} with β-CD. However, if the methanol concentration in the mobile phase is high, it is competing with *trans*-resveratrol for β-CD complexation (50).

Ravelet et al. studied the inclusion complex stability of nimesulide with both native and modified β-CD in HPLC (29). Phenyl silica gel, as a weak non-polar stationary phase, was used to preclude the use of higher amounts of mobile phase organic modifiers and their competitiveness with nimesulide for binding to β-CD. Modified β-CD established stronger interactions with nimesulide in comparison to the native one. In addition, nimesulide was investigated under one pH value, only in its ionized form, which could also affect the interactions forming with β-CD. Furthermore, in the case of bupivacaine inclusion complexes with modified β-CD, there were differences in K values obtained in HPLC at varying pH, indicating the importance of pH and consequently the ionization form of the analyte (28).

De Melo et al. managed to determine the K of inclusion complexes formed between nitroheterocyclic compounds and β-CD using HPLC (53). Experiments were conducted on the C18 stationary phase, with 20% (v/v) acetonitrile as the organic modifier and β-CD concentration in the aqueous part of the mobile phase increasing up to 30 mM. As expected, the retention times of the investigated solutes decreased with an increase in β-CD concentration (53).

HPLC approach was also successfully applied in assessing the stability of inclusion complexes formed between geraniol and α-terpineol, volatile and water-insoluble compounds, with β-CD, using ethanol as the mobile phase organic modifier (48).

Gazpio et al. applied a chromatographic approach to determine K of inclusion complexes formed between pindolol and other indole derivatives with different types of CD (52). The investigation was undertaken using the C18 stationary phase and with low amounts of methanol as the mobile phase modifier. K values for complexes formed between indole derivatives and β-CD were in accordance with the literature data (52).

El-Barghouthi et al. used phase-solubility studies to determine K values for complexes that risperidone forms with various CDs (54). The study shows that both β-CD and hydroxyl-propyl β-CD form 1:1 and 1:2 complexes with risperidone, with risperidone:β-CD 1:2 complex reaching saturation at 7 mM β-CD concentration. Also, K values are slightly higher for inclusion complexes with β-CD in comparison to hydroxyl-propyl β-CD due to the hydrophobic effect. The results also show that K is affected by pH; namely, K values are higher if the inclusion complex is formed with non-ionized risperidone species (54).

In the authors' previous research, a chromatographic approach was employed to determine K for risperidone and its three structurally related impurities, as well as olanzapine and its two structurally related impurities inclusion complexes with β-CD (11). The experimental setting was composed of varying acetonitrile content, the pH of the aqueous part of the mobile phase in order to investigate the analytes in both ionized and non-ionized form, β-CD concentration in the aqueous part of the mobile phase, and the column temperature. The addition of an organic modifier in the mobile phase can diminish the stability of the formed β-CD inclusion complex. Increasing the organic solvent content provides a less polar mobile phase, which also becomes a more comfortable environment for non-polar solute. As a result, the non-polar solute is soluble in the non-polar mobile phase; thus, there is no driving force attracting the solute to β-CD cavity (52). Therefore, the acetonitrile content was kept at 15% (v/v) or 20% (v/v). Nevertheless, the solute and organic modifier

could compete for β-CD binding sites, affecting the solutes' complexation process, even if it is proven that the solvent binds weakly (52). The choice of the type of organic modifier is also important, besides its content in the mobile phase, not only from the complexation perspective but the stationary phase microenvironment as well. Although methanol is preferred over acetonitrile when assessing the stability of β-CD inclusion complexes throughout the literature, acetonitrile was chosen in the previous authors' study due to its higher elution strength (11). The general rule that an increase in β-CD concentration in the mobile phase leads to a decrease in retention factor value was not strictly followed in the entire experimental space. For that reason, in this research, K could not be calculated by applying the aforementioned formula (3.1) for all substances under all of the examined experimental conditions (11). The authors observed that the changes in retention factor values are affected by pH and acetonitrile content. pH also influences the ionization of stationary phase free silanol groups, so secondary interactions with the stationary phase could be weakened if free silanol groups are non-ionized at pH lower than 3.0. When solutes' secondary interactions with the stationary phase are diminished, it is left to acetonitrile and β-CD in the mobile phase to compete for interaction with the solute and determine whether retention will be governed by complexation mechanisms or driven by acetonitrile. The authors hypothesized that for these reasons, stability constants could be calculated for all examined compounds when pH was set to 2.0 and acetonitrile content to 15% (v/v), and for most of them if acetonitrile content was 20% (v/v) (11).

HPLC experiments are time-consuming and require both chemical and human resources; therefore, a need for *in silico* tools able to replace extensive HPLC experiments exists. Applying *in silico* tools could help in defining the experimental space within which change in retention factor values would comply with changes in β-CD concentration. This led the authors to think that the QSRR model developed to describe retention behavior in β-CD modified RP-HPLC could predict conditions under which interactions leading to complexation would be able to outperform remaining interactions in a dynamic system with an increased level of complexity.

3.3.2 Determination of Complexation-Related Thermodynamic Parameters in β-CD-Modified RP-HPLC

Stability constants provide information about the mutual affinity between molecules but could be insufficient to reflect the real stability of the formed complex without the accompanying thermodynamic parameters (55). Van der Waals and hydrophobic interactions are dominant in the CD complexation process, while hydrogen bonding and steric effects also play a certain role (56). Complexation thermodynamic parameters can be considered as results of the weighted contributions of the aforementioned interactions. Thermodynamic parameters include Gibbs free energy (ΔG°), standard molar enthalpy (ΔH°) and standard molar entropy (ΔS°) and provide additional information about binding mechanism dynamics in the microenvironment of the CD cavity (57).

Different methods could be used to determine thermodynamic parameters of CD complexation. There are papers reporting the utilization of microcalorimetry

to determine thermodynamic parameters of complexation between catechin and paeonol with β-CD (58, 59). Further, UV/Vis and fluorescence spectroscopy were used to assess thermodynamic parameters of β-CD complexation with ibuprofen and indole chalcones (60, 61). Among the spectroscopic methods, NMR spectroscopy can be used for the same purpose (39). However, methods such as liquid chromatography, CE, pH potentiometry, and many others have been reported as solutions to determine the thermodynamic aspects of β-CD complexation (56).

When the chromatographic approach is used for determining inclusion complex stability, the aforementioned thermodynamic parameters could be assessed at the same time if the temperature is varied during the experiments. ΔH° and ΔS° could be easily calculated from the following equation (3.2):

$$\ln K = \frac{-\Delta H^\circ}{RT} + \frac{\Delta S^\circ}{R} \tag{3.2}$$

where R stands for the universal gas constant (8.314 J mol^{-1} K^{-1}), while T (K) is the varying column temperature. Van't Hoff plot of lnK versus 1/T is constructed and the slope of the obtained curve equals -ΔH°/R, while ΔS°/R represents the intercept.

ΔG° is calculated employing ΔH° and ΔS° in the following manner (3.3) (29, 55, 62):

$$\Delta G^\circ = \Delta H^\circ — T\Delta S^\circ \tag{3.3}.$$

Upon addition of CD to the mobile phase ΔH° and ΔS° values are increasing, indicating the unlikeliness of transfer of the guest molecule from mobile to stationary phase, most certainly due to inclusion complex formation. The rise in thermodynamic parameter value is associated with inclusion complex formation between the guest molecule and CD in the mobile phase. It is reported in the literature that stationary phase influence is negligible if the CD concentration in the mobile phase is lower than 1 mM (63).

It is well known that classical hydrophobic interactions are associated with positive enthalpy and entropy changes. Within deep insertion into the CD cavity, the carboxylic group could stay outside the cavity and form a hydrogen bond with functional groups at the outer rim of the cavity, explaining the negative enthalpy values (63).

As in the case of stability constants, there have been attempts to predict complexation thermodynamic parameters by means of Quantitative Structure-Property Relationship (QSPR) models (64, 65). This kind of approach offers great savings in terms of cost and time. Maljurić et al. developed a novel *in silico* approach to derive the stability and thermodynamic parameters of solute:β-CD inclusion complexes, based on the QSRR model employing a machine learning algorithm (11). The proposed methodology, which will be discussed later, provides insight into the inclusion complexation behavior of selected analytes.

3.3.3 Stoichiometry of Formed Inclusion Complexes

The first step in the analysis of inclusion complexes formed between guest molecules and CD is the determination of stoichiometry. Through a literature search, it

can often be seen that authors assume 1:1 stoichiometry between the guest molecule and CD. This assumption can easily lead to errors in consequently calculated inclusion complex stability (66). For that reason, it is always advisable to experimentally determine the molar ratio between the guest molecule and CD in the formed inclusion complex. Among available analytical techniques, Danel et al. used 1H NMR spectroscopy with a continuous variation method to determine the stoichiometry of risperidone and 9-hydroxyrisperidone with CD hosts, as one of the most applied methods for the intended purpose (39). Further, the continuous variation method using fluorescence spectroscopy was also used in analyzing the stoichiometry of ibuprofen: CD inclusion complexes (67). The continuous variation method is not the first choice when investigating stoichiometry, due to its duration and the large amount of substance needed for sample preparation.

Further, mass spectrometry with the electrospray ionization technique (ESI-MS) is also used in assessing complex stoichiometry. ESI-MS is a mild ionization procedure capable of studying interactions between solutes and CD. However, its main advantage is the possibility of transferring ions from the solution into the gas phase without disrupting non-covalent interactions forming the inclusion complex (68, 69).

The authors used ESI-MS in their previous paper to determine the stoichiometry of inclusion complexes formed between risperidone and its related impurities with β-CD, as well as olanzapine and its related impurities with β-CD. Mass range from 100—3000 *m/z*, which was the upper limit according to instrument setting, was analyzed, and it was concluded that all inclusion complexes are formed in a 1:1 ratio (11). Figure 3.2 presents the signals corresponding to the investigated solutes, β-CD and formed inclusion complexes for each of the examined substances (11).

3.4 QUANTITATIVE STRUCTURE-RETENTION RELATIONSHIP MODELING

Investigating the retention of an analyte, as well as the resolution in the given RP-HPLC system, usually requires an experimental approach. The experimental approach has a deficiency reflected in the substantial amount of time and resources needed to come to valid conclusions. If the theoretical approach capable of predicting retention to a certain extent existed, RP-HPLC method development would be faster and more efficient (70). The idea of predicting chromatographic behavior on the basis of molecular structure induced the development of the Quantitative Structure-Retention Relationships (QSRRs) methodology. First, the Quantitative Structure-biological Activity Relationships (QSARs) methodology was developed and applied the same pattern of thinking to the analysis of chromatographic data-enabled development of QSRRs (71).

The QSRRs represent mathematical relationships between chromatographic parameters determined for a series of analytes in a given chromatographic system and numerical values accounting for structural differences between the investigated analytes, denoted as molecular descriptors (72, 73). In most cases, when using QSRR, the target retention parameter represents the dependent variable of the linear equation obtained as a result

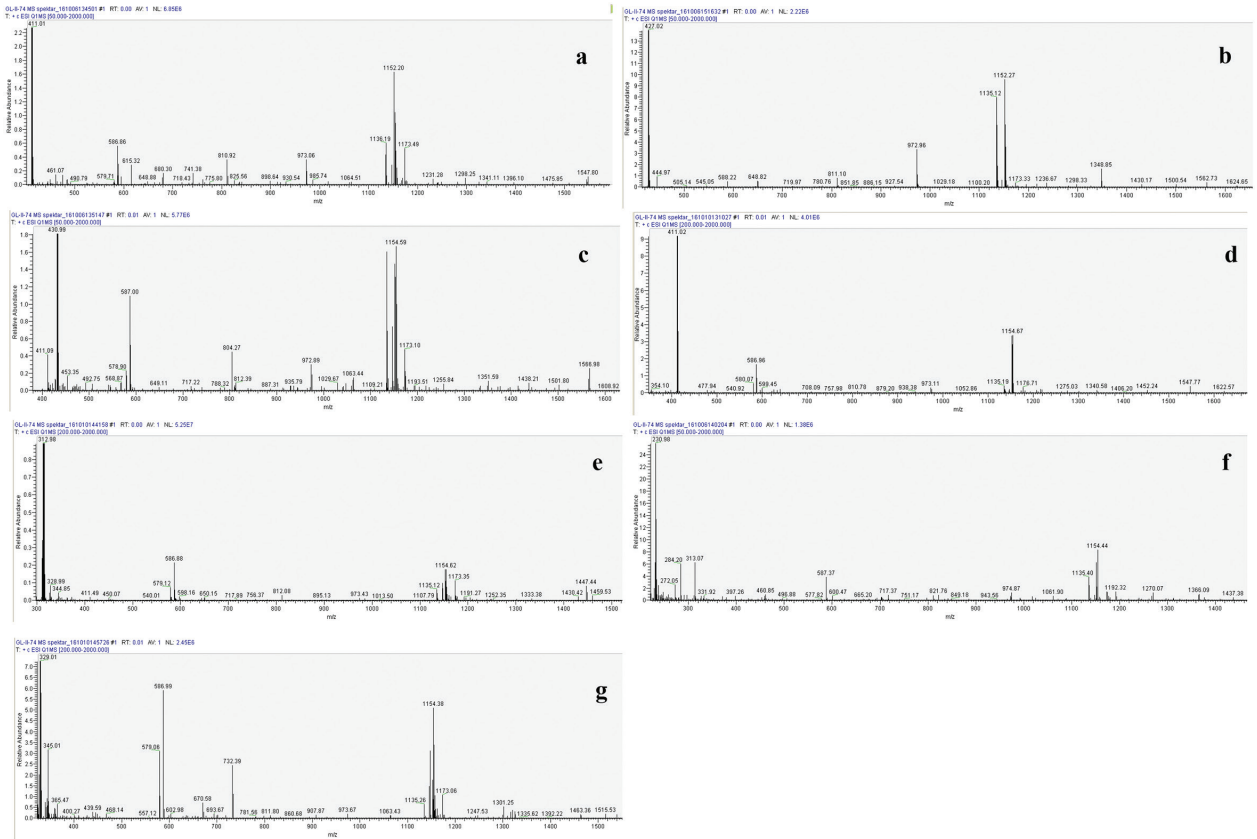


FIGURE 3.2 Full scan positive ESI mass spectra of:

2a: Risperidone—β-CD complex; 2b: Risperidone impurity 1—β-CD complex; 2c: Risperidone impurity 2—β-CD complex; 2d: Risperidone impurity 3—β-CD complex; 2e: Olanzapine—β-CD complex; 2f: Olanzapine impurity B—β-CD complex; 2g: Olanzapine impurity C—β-CD complex;

From Maljurić, N. et al., Journal of Chromatography A, 1619, 460971, 2020., reused with publisher's permission.

of multivariate regression of retention data in dependence of descriptors describing the investigated analytes (74).

In the last few decades, QSRRs have been used to:

(i) Characterize chromatographic columns by quantitative comparison of their separation capabilities
(ii) Identify structural descriptors of the utmost importance
(iii) Investigate separation mechanisms under certain chromatographic conditions at the molecular level
(iv) Evaluate complex physicochemical properties of analytes other than chromatographic
(v) Predict retention of a new compound or identify an unknown compound (75)

There are several types of classical QSRR models. The oldest one relates the retention factor with the logarithm of the octanol-water partition coefficient (logP), calculated on the basis of analyte structure with the aid of commercially available computer programs (72, 76). This type of QSRR is also considered the simplest. The second type of QSRR is based on the fundamental theory of liquid chromatography, which assumes that retention is governed by intermolecular interactions. Namely, this approach relies on Abraham's linear solvation energy relationships (LSER) theory (77). The general LSER equation in HPLC is as follows (3.4):

$$\log k = \log ko + rR^2 + vVx + s\pi_2^H + a\sum\alpha_2^H + b\sum\beta_2^H \tag{3.4}$$

where R_2 is the excess molar refraction of the analyte, Vx is its molecular volume calculated from the McGowan algorithm (78), π_2^H is dipolarity/polarizability descriptor, $\Sigma\alpha_2^H$is a measure of the ability of the analyte to donate a hydrogen bond, $\Sigma\beta_2^H$ is an analogous parameter of hydrogen bond accepting potency. Log k_0 is a constant, while r, v, s, a, and b are regression coefficients accounting for the total complementary properties of the chromatographic system (79). The aforementioned theory enables determination of cavity terms, dipolar terms, and hydrogen bonding terms as contributing mechanism types used to explain retention behavior. The cavity term is a measure of the free energy necessary for separating the solvent molecules and providing a cavity of a suitable size for the solute. In the case in which π* is defined as a measure of the dipolarity-polarizability of the investigated species, the dipolar term may be explained as a product of the solute and the solvent π* interactions. Finally, the hydrogen bonding terms show the hydrogen donating and accepting capabilities of a given solute (80).

The third QSRR type relates retention factor values to structural descriptors obtained by applying computational chemistry (72). This type of QSRR model can be represented via the following equation (3.5):

$$t_R = k_1' + k_{2\mu}' + k_{3\delta Min}' + k_4' A_{WAS} \tag{3.5},$$

where $k_1' - k_4'$ are regression coefficients, μ accounts for dipole-dipole or dipole-induced dipole interactions between analyte and mobile or stationary phase components, δ_{Min} represents analytes' fragment polarity and consequently the ability of analyte to participate in polar interactions with phases such as dipole-dipole, charge transfer and hydrogen bonding, while A_{WAS} characterizes the strength of London dispersion bond forming between analyte and molecules forming chromatographic phases (76).

The classical QSRR approach solely links molecular descriptors to the observed response. Therefore, experiments are conducted under only one defined set of conditions, which constrains the practical applicability of the model. Consequently, future usage of the developed model is limited to the concrete values of the factors (6). For these reasons, interest has arisen in the so-called mixed modeling relating both experimental parameters and molecular descriptors to the observed response. Incorporating all influential factors in the model increases the percentage of explained variance and improves the model's predictive performance (81).

3.4.1 Molecular Descriptor Selection

A molecular descriptor could be defined as a result of a logical and mathematical procedure that transforms chemical information into a useful number or results of a standardized experiment, enabling a better understanding of the various characteristics of a molecule (82). Definitions of molecular descriptors could rely on different theories, while their simple calculation is enabled by the development and application of algorithms. There are several groups of molecular descriptors, namely physicochemical descriptors, clearly connected to retention but often susceptible to errors. On the other hand, quantum chemical descriptors have a weak correlation with retention, which is one of their flaws. However, quantum chemical descriptors provide detailed information on retention mechanisms in chromatography at the molecular or submolecular levels (83).

There are also theoretical descriptors that are easy to calculate, but their correlation with a certain retention phenomenon is not always obvious (84). Theoretical descriptors are generally divided into zero- (0D), one- (1D), two- (2D), three- (3D) and four-dimensional (4D) descriptors. Descriptors denoted as 0D are usually derived from molecular formulas and represent the type and number of atoms, molecular mass, etc. In the case of 1D descriptors, molecular functional groups or their substituents are also taken into account. 2D descriptors consider molecular topology, while 3D descriptors also take into account spatial conformation. Further, 4D descriptors are calculated on the basis of the molecular representation of properties indicating the interaction of a molecule with the surrounding space (85).

Molecular descriptors contribute to a more comprehensive understanding of various molecular characteristics. The development of different algorithms enabled their simple calculation, but the problem remained that there is a practically unlimited number of structural descriptors that can be assigned to one analyte. Therefore, the key to successful QSRR model building is the proper selection of the most informative molecular descriptors for a given set of analytes from a large pool of mutually correlated descriptors (83). For all these reasons, in order to properly utilize

molecular descriptors, a certain prior knowledge in statistics, chemometrics, QSRR principles, as well as the characteristics of the field from which the problem arises, is needed. This indicates that the field of investigation and utilization of molecular descriptors is highly interdisciplinary (82, 84).

3.4.2 Selection of Experimental Parameters

Experimental design methodology represents an efficient procedure for planning the experiments and defining the experimental space. Applying experimental design methodology enables testing of the influence of the change of one or more independent variables on the system's response. Consequently, valid conclusions about certain behaviors in the system can be made.

The benefits of this multifactorial approach are reflected in the reduction in the number of experiments and cost savings in comparison to the traditional One Factor At a Time (OFAT) approach. The OFAT considers changing one factor in the experimental setting and investigating its influence on the observed response, while other factors are kept constant. Additionally, the experimental design approach takes interactions between variables into account and provides information about system behavior in the entire experimental space, while the quality of this information is always higher in comparison to the OFAT approach. Within the defined experimental space, multiple linear regressions are used to fit a mathematical model to the experimental data obtained. Prior to its use, validation of the obtained mathematical models needs to be performed. The quality of the model is usually evaluated by analysis of variance (ANOVA), providing information about the significance of the model's coefficients and not a significant lack of fit test if the model is reliable. In most cases, during the optimization, quadratic models explaining the relationship between investigated variables and selected responses are the most suitable, allowing the construction of 3D response surface plots. The general expression of one quadratic model for investigation of the influence of two variables on the selected response is as follows: $y = b_0 + b_1x_1 + b_2x_2 + b_{12}x_1x_2 + b_{11}x_1^2 + b_{22}x_2^2$, where y is the response, x_1 and x_2 are investigated variables, x_1x_2 represents the interaction term, x_1^2 and x_2^2 are quadratic terms, b_0 is an intercept, while b_1, b_2, b_{12}, b_{11} and b_{22} are the model's coefficients. Based on the relationship within the equation, 3D response surface plots could be constructed, enabling the selection of optimal separation conditions. (86–88).

3.4.3 Techniques for QSRR Model Building

When the input and output variables of the QSRR model are selected, the next step is to find an appropriate technique to correlate them. The choice of a regression technique to correlate descriptors and/or experimental parameters with chromatographic retention is crucial to the prediction ability of the developed QSRR (89). With respect to the studied problem, strategies for QSRR model building are based on either regression or classification. The most commonly applied regression technique relating molecular descriptors to chromatographic retention in the

first QSRR models was multiple linear regression (MLR). Moreover, when there is a higher number of molecular descriptors, partial least squares (PLS) regression was more frequently used (85). Although easily interpreted, MLR was not able to follow the progress in molecular descriptors theory, which required the use of techniques capable of handling a large number of a model's inputs and dealing with non-linear dependencies between input and output variables (90).

Machine learning algorithms (MLA) take advantage of simple modeling techniques due to their ability to combine attributes in an advanced way. In addition, MLAs are capable of dealing with a large number of model inputs. Among the available MLA, artificial neural networks (ANN) and support vector regression (SVR) aid in computer-assisted retention prediction (89, 90).

3.4.3.1 Artificial Neural Networks

ANNs represent a chemometric tool for solving multivariate chemical problems based on computer techniques inspired by simulating neurological processes of a human brain (91). There are numerous definitions of ANNs, but the simplest one presents ANN as a black box with various inputs entering and capable of producing different outputs. ANN's strength is reflected in the utilization of rather simple mathematical operations to solve complicated, ill-defined, or non-linear problems (91, 92). An additional advantage of ANN over classical statistical methods is that ANN does not require prior knowledge of the mathematical relationship between the examined variables (93).

A multilayer feedforward network is an ANN type mostly recognized in pharmaceutical analysis, due to its clear architecture and the relatively simple back-propagation algorithm used for its training. ANN architecture considers organizing nodes into layers and linking layers of neurons with modifiable weighted interconnections. The number of neurons in the input layer equals the number of input variables. The same stands for the output layer, while the input and output layers are connected with a variable number of hidden layers consisting of an optimized number of neurons (94).

ANN training consists of adjusting the weights to achieve certain optimal values. Initially, weights are randomly assigned, and input-output pairs are represented in the non-trained network. The output predicted by the network is compared to the desired output value, and the differences serve to adjust the weights. These cycles are called epochs, and they are repeated until the desired error values are obtained. The optimal duration of training is defined based on the minimum validation error through the defined number of epochs.

To evaluate how successful the model is in terms of its predictive ability, it is necessary to determine the coefficient of determination between responses obtained experimentally and predicted by the network (R^2), root mean square error (RMSE), and the correlation coefficient between experimentally obtained and predicted values of the observed response. R^2 is calculated according to the following formula (2.6):

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2}{\sum_{i=1}^{n} \left(y_i - \bar{y} \right)^2} \tag{2.6}$$

While RMSE is calculated according to the formula presented below (2.7):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_1\right)^2}{n}} \tag{2.7}$$

After optimization, the real prediction ability of the network is evaluated on a new set of data previously unseen to the network (test set or external validation) (95, 96).

There are two approaches to model validation: internal or cross-validation and independent or external validation. Cross-validation starts with removing a compound or group of compounds from the training set in order to make them temporarily test set, while regression is repeated on a segregated set of data applying Leave-One-Out or Leave-Many-Out strategies. The predictive ability of the model is evaluated based on the cross-validation coefficient of determination (Q^2) calculated according to the formula presented below (2.8):

$$Q^2 = 1 - \frac{\sum\left(y_{exp} - y_{LOO}\right)^2}{\sum\left(y_{exp} - \overline{y_{exp}}\right)^2} \tag{2.8}$$

where y_{LOO} is the response predicted by the model.

External validation ensures the possibility of applying the developed model to predict the behavior of untested compounds. However, there are circumstances, such as a low amount of data, under which external validation cannot be performed. Throughout the literature, the authors state that the model could be highly predictive, even if validation was not undertaken on an external set of data (97–101).

3.4.4 Development of the QSRR in β-CD Modified RP-HPLC

As previously mentioned, β-CD modified RP-HPLC systems are dynamic and rather complicated, thus making retention modeling more challenging in comparison to regular RP-HPLC. Additionally complicated by the joint effect of complexation and adsorption equilibrium on retention, they require a specific methodological approach for retention modeling. The complexity of the system outreaches the capabilities of mixed QSRR models but requires the introduction of an additional type of descriptors, so-called association constants, introduced in the paper published by Maljurić et al. The association constants would characterize the formed inclusion complexes with β-CD. Together with molecular descriptors and experimental parameters, association constants would be considered QSRR model inputs (10).

To calculate complex association constants, the inclusion complex structures should first be obtained. To anticipate the extensive experimental procedure, an *in silico* approach in terms of molecular docking was conducted. A docking study is performed with the aim of predicting the most certain structures of inclusion complexes. To perform a docking study, it is necessary to label β-CD as a receptor (host), while investigated analytes in their minimum energy conformations across

the investigated experimental conditions are marked as ligands (guests). Docking studies consider discovering energetically the most favored type of binding between the host and the guest molecule. Therefore, it is necessary to determine the ligand variables that would uniquely define the binding mode. If the ligand is rigid, the variables include position, orientation, and conformation. Docking methods also require certain search functions, such as the Lamarckian genetic algorithm, which enables prediction of free binding energies and consequently the association constants of incorporated ligand molecules. Generally, docking studies are expected to provide information on the structure of the complex formed between ligand and receptor, as well as its stability expressed through the calculated energy of binding (102, 103). The binding energy of the formed inclusion complexes, as well as the difference between the heat of formation of the inclusion complex and the heat of formation of the free binding molecule, was calculated and used to choose the most stable conformation of the inclusion complex. In general, a more thermodynamically favorable pathway of inclusion complex formation is associated with a lower binding energy of the suggested complex structures. Based on the predicted preferred inclusion complex structures formed between β-CD and selected analytes (risperidone, olanzapine, and their structurally related impurities), the following complex association constants were calculated: energy of binding (BE), estimated inhibition constant (IC), interaction energy (IE), electrostatic energy (EE), final intermolecular energy consisting of van der Waals energy, hydrogen bonding, desolvation energy (VDW-HB-DE), final total internal energy (TI), torsion energy (TORE) and unbound system energy (UE).

Preliminary selection of molecular descriptors in β-CD modified RP-HPLC was done in accordance with Abraham's LSER theory, as previously explained, and it included the octanol/water partition coefficient (logP), polarizability (POL), the sum of the hydrogen atoms connected to the hydrogen bond donating atoms (H-don), the sum of the ion pairs on the hydrogen bond acceptor atoms (H–acc), dipole-dipole energy (DEN), charge dipole-dipole energy (CDEN), Conolly solvent accessible area (SAS), solvent-excluded volume (SEV), 1,4 van der Waals energy (VDW), non-1,4 van der Waals energy (NON VDW), molecular area (MA), molar refractivity (MR), highest occupied molecular orbital (HOMO) and lowest occupied molecular orbital (LUMO) (10).

The main role of each molecular descriptor, independent of its way of calculation, is to encode the analyte's chemical structure. However, in a particular task of modeling, not all molecular descriptors are equally important. For that reason, in the initial phase, it is important to come to the rational number of descriptors while simultaneously acquiring as much chemical and structural information as possible (104).

Prior to the QSRR model development, the correlation between descriptors was performed via multiple linear regressions to remove redundant descriptors and reduce the system load. In this way, mutually non-correlated descriptors would be kept for model building. The correlation was performed within molecular descriptor and complex association constants based on the determined cut-off value for descriptor correlation coefficients of 0.990. POL, SAS, MA, MR, VDW, D, LUMO, and HOMO highly correlated with each other, so POL was included in the model as it showed the highest correlation with all the other molecular descriptors. When dealing with

complex association constants, only IE and VDW-HB-DE correlated with each other; therefore, it was decided to retain VDW-HB-DE in the model. Another confirmation that complex association constants are important in model building in this kind of RP-HPLC system lies in the fact that there was no correlation between molecular descriptors and complex association constants, as they carry different information equally important for a given system (10).

Experimental space for the QSRR model construction was properly covered by creating a plan of experiments applying central composite design (CCD). Preliminary experiments revealed the parameters with the highest influence on retention. Consequently, the following parameters were included in the CCD: acetonitrile content in the mobile phase (%), pH of the aqueous part of the mobile phase, β–CD concentration in aqueous mobile phase, and column temperature.

To sum up, mutually uncorrelated molecular descriptors and complex association constants along with experimental parameters were included as model inputs toward retention factors. The QSRR model was built with the aid of ANN, a machine learning algorithm capable of solving complicated modeling problems (7). A multilayer perceptron with three layers (input, hidden and output), 11-8-1 topology was constructed and trained with a backpropagation algorithm (Figure 3.3). Good predictive performance of the networks was confirmed through low RMSE values for all datasets accompanied by high R^2 (10).

3.5 DEVELOPMENT OF COMPUTATIONAL MODELS TO PREDICT CD COMPLEXATION BEHAVIOR

The pharmaceutical industry prioritizes usage of β-CDs among available native and modified CDs, with respect to their cavity size complementary to most drug-size molecules. There are pharmaceutical formulations with β-CD in the role of solubilizer of otherwise insoluble drugs (105, 106). The existence of such formulations

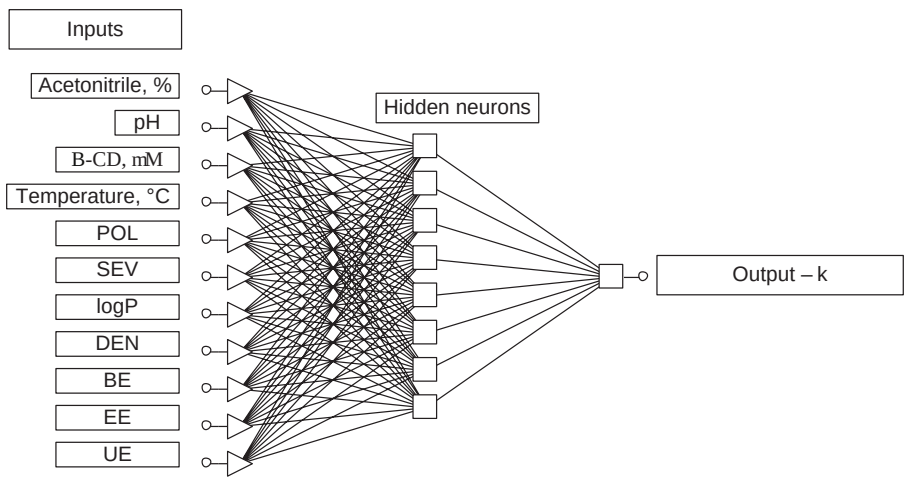


FIGURE 3.3 ANN with 11-8-1 topology trained with back propagation algorithm

on the market encouraged predicting thermodynamic properties of the complexation process, due to its potential influence on new pharmaceutical formulation development.

Determining the binding constant and accompanying thermodynamic parameters of CD complex formation experimentally is not an easy task, mainly due to low solubility of drug molecules in the aqueous solutions and therefore substantial amounts of time needed. Major interaction types revealed so far include van der Waals interactions, hydrophobic interactions, hydrogen bonding, and relaxation by extrusion of energy-rich water molecules by the guest molecule (107). To avoid extensive experiments, there are computational methods capable of predicting stability constants, thermodynamic parameters, and interactions leading to complex formation. Further, the direction of computational methods has started to rise relatively recently but is already widely used in clarifying the factors involved in the CD complexation process (107, 108).

Steffen et al. investigated the predictability of three thermodynamic properties in relation to CD complex formation with various guest molecules (109). The QSPR models were built with principal component regression (PCR), partial least squares regression (PLSR), and support vector machine regression (SVMR), and their ability to accurately predict $\Delta G°$, $\Delta H°$ and $\Delta S°$ were evaluated. The study results showed that the poor predictability of $\Delta S°$ to a larger extent, followed by $\Delta H°$, is related to its strong dependence on the structure of the formed complex (109). In that respect, Katritzky et al. developed QSPR models with an aim to predict free energies of complexation process occurring between guest molecules and CDs employing fragmental descriptors and the CODESSA-PRO program, which uses different geometrical, topological, quantum chemical, and thermodynamic molecular descriptors derived from molecular structural without need for experiments (65). CODESSA descriptors are relatively difficult to interpret, unlike fragmental descriptors. Nevertheless, a physical phenomenon occurring during host–guest interaction can be a good basis for selecting the fragments for modeling. However, if the QSPR model uses fragments, it includes more variables than models using traditional descriptors, which is one of the disadvantages of this approach (107). Then, Perez-Garrido et al. constructed a QSPR regression-based model to predict and correlate stability constants of 233 organic compounds toward β-CD with known stability constant values (107). For the first time, the models correlated TOPS-MODE descriptors with β-CD complexation ability, and it was concluded that the leading complexation forces are hydrophobicity and van der Waals interactions. For this reason, the complexation process is promoted by the presence of the hydrophobic group and voluminous species in the guest molecule structure (107).

As a sequel of their previous research in which host-guest interactions were modeled (110, 111), Ahmadi et al. developed a 3D-QSAR model with the aid of Grid-INdependent Descriptors (GRIND). The selection of important variables was performed with a genetic algorithm, while the selected descriptors were correlated to complex stability constants of 126 organic compounds with β-CD using PLSR. The validated 3D-QSAR model provided information about the importance of hydrogen bond acceptor and/or donor groups in the molecule structure with respect to unfavorable complexation. The stability of the formed complexes with β-CD is affected by

the size and shape of the complexed guest molecules (108). In addition, according to these study results, steric and hydrophobic interactions are labeled as the major driving forces in the β-CD complexation process. In another study, Ghasemi et al. constructed QSAR models for prediction of stability constant of mono and 1,4-di-substituted benzenes with α-CD applying methods of comparative molecular field analysis region focusing (CoMFA-RF) and VolSurf. CoMFA fields were combined with physicochemical descriptors to improve the model predictability (112). The study results showed that inclusion complexation of benzene derivatives with α-CD is mostly influenced by electrostatic and hydrophobic effects as well as molecular shape (112).

Li et al. focused on modeling β-CD binding behavior of structurally diverse drug molecules with poor solubility. They used K values obtained from a literature search and established a model employing multiple linear regression (113). As previously reported, the hydrophobic effect also appeared to be the most important in the drug β-CD binding. Larger binding constants were always followed by higher drug hydrophobicity. The inclusion of a drug molecule into the cavity was also influenced by the attractive forces of one transient dipole to another. The developed *in silico* model elucidated the most important driving forces in the complexation process, namely hydrophobic interactions, electrostatic interactions, van der Waals interactions, and hydrogen bonding (113). Hydrophobic interactions and van der Waals interactions are the main driving forces in the process of binding, while hydrogen bonding and electrostatic interactions play a role in stabilizing the formed inclusion complex by establishing and maintaining the binding and dissociation equilibrium. However, the presented model has its weaknesses, mainly in its dataset size, which needs to be larger in order to generalize the conclusions. In addition, there is a skewed distribution of the observed stability constant values, so R^2 values are relatively lower in comparison to models obtained for organic compounds found in the literature (107, 113).

Further, Merzlikine et al. developed machine learning models based on Cubist and Random Forest to evaluate the complexation between small organic molecules and β-CD (114).

In order to preclude the calculation of the optimal molecular geometry and develop the QSPR model in a shorter time frame, Veselinović et al. constructed a model using SMILES attributes as a representation of the molecular structure and Monte Carlo simulation method (115). The study results labeled the Monte Carlo method as promising computational methods in the QSPR. A few years later, SMILES strings were used in combination with non-linear MARSplines (multivariate adaptive regression splines) methodology for the purpose of quantifying the stability constant values of a variety of molecules toward β-CD (116). As in the previously discussed papers, the hydrophobic nature of the CD cavity, as well as the importance of hydrophobic effects, was confirmed. Apart from predicting the affinity of guest molecules to β-CD cavity, the developed QSPR model could be applied to classify the compounds into types according to the Biopharmaceutical Classification System, since it is known that permeability of the drug is affected by its hydrophobicity (116). When discussing different application fields of *in silico* models, it is worth mentioning the recent study performed by Ling et al. Ling et al. developed a QSAR model to predict

the adsorption behavior of micropollutants from water to CD as adsorbent during the water treatment process (117).

However, when dealing with β-CD usage in chromatography, Šoškić et al. reported the development of QSPR models relating physicochemical and structural attributes to retention factors of 31 indole derivatives. Retention factors were obtained by HPLC with a stationary phase composed of immobilized β-CD, so that inclusion complexes are mainly formed between solute and stationary phase (118). The results show that the stability of the formed inclusion complexes is affected by the joint influence of hydrophobic interactions and hydrogen bonds formed between β-CD in the stationary phase and indole derivatives.

On the other hand, as previously mentioned, Maljurić et al. reported the development of a QSRR model in RP-HPLC modified by the addition of β-CD in the mobile phase. Risperidone with its structurally related impurities, as well as olanzapine with its structurally related impurities, represented a model mixture for investigating β-CD complexation process. Although the QS(P)RR methodology has been extensively used in characterizing β-CD complexation with various compounds, Maljurić et al. were the first to use molecular descriptors, complex association constants, and experimental parameters as inputs of the QSRR model. Introducing complex association constants as descriptors of the formed inclusion complexes highly contributed to the predictive power of the developed models and their future applicability (10).

3.5.1 Applying the QSRR in Green RP-HPLC Method Development

Adding CDs to the mobile phase leads to the development of green chromatographic methods. Moreover, an additional component contributing to the greening concept is the construction of QSRR models able to explain the retention behavior in a given system. Therefore, such models could be used in the HPLC method development for separation of the investigated model mixture. In that sense, the HPLC method is additionally improved in terms of eco-friendly character since experiments are replaced with the modeling approach. The scientifically based approach in method development results in continuous improvement and operational excellence. It would be best if the ecologically acceptable method and potential optimal separation conditions could be selected by using an *in silico* approach, without experimentation (119).

The development of QSRR models explaining the retention behavior of risperidone, olanzapine, and structurally related impurities in β-CD modified RP-HPLC performed by Maljurić et al. (10) was previously explained in subsection 3.4. The obtained models enabled optimization of separation conditions for a given set of analytes by means of response surfaces. Certain goals regarding the retention factor and resolution values were defined, while response surface plots were constructed based on the developed network. When selecting the region of factors fulfilling the defined goals, the error of the constructed network should not be neglected. A combination of factors should be chosen in a manner to give solutions better than the defined acceptance minimum.

The constructed network enabled discussion of how each of the investigated experimental parameters, as well as molecular descriptors and association constants, influence the retention of the investigated compounds. Response surface plots in

Figure 3.4 show the predicted retention factor of a compound against the most influential experimental parameters, namely acetonitrile content in the mobile phase and pH of the aqueous part of the mobile phase. These parameters were considered critical in terms of separation. Detailed analyses of the obtained response surface plots (Figure 3.4) enable speed-up and efficient selection of optimal separation conditions for both risperidone and its impurities and olanzapine and its impurities, thus encouraging efficient and rational method development. Moreover, to determine the influence of β-CD concentration in the mobile phase, response surface plots of the predicted retention factor of each of the investigated compounds against β-CD concentration and pH (Figure 3.5), as well as β-CD concentration and acetonitrile percentage (Figure 3.6) were constructed. Response surface plots revealed similar retention behaviors among risperidone and its related impurities. Retention factors increased with a lower β-CD concentration across the whole investigated pH range (Figure 3.5). There is an obvious interaction between the influence of β-CD concentration and acetonitrile, explaining its joint or competitive effect on retention (Figure 3.6). Retention factors of olanzapine and its impurity C are lower if pH is low, but with high concentrations of β-CD in the mobile phase (Figure 3.5). However, the 3D response surface also shows variable retention behavior of impurity B at different pH in dependence of β-CD, which was unexpected due to its non-ionized form across the investigated pH range (Figure 3.5). When discussing the changes in the retention factor against acetonitrile content and β-CD concentration in the mobile phase, it can be seen that in the case of olanzapine and its related impurities, high acetonitrile content and β-CD concentration are associated with lower retention factors (Figure 3.6) (11).

POL, SEV, DEN, and log P, which were labeled as the most influential molecular descriptors, showed the same effect on retention (Figure 3.7). Retention factor values decrease with an increase in each of these parameters. Compounds characterized as lipophilic (high log P values) are retained longer on the stationary phase than hydrophilic compounds, but it is also known that a hydrophobic CD cavity is an excellent microenvironment for hydrophobic compounds. Therefore, the increase in log P values, namely the hydrophobicity of the compound, favors the binding with CD and decreases its retention time. POL is a tensor known as an important electronic parameter with an impact on chemical interactions (120), so it was expected that a higher value would lead to a decrease in retention factor value. The same trend can be seen in the case of DEN. As dipole–dipole interactions are one of the recognized driving forces in complexation, higher DEN values would encourage the inclusion complex formation, thus reducing the retention factor of a complexed solute. Solvent exclusion effect is one of the major components of hydrophobic effect (121), thus higher SEV values contribute to hydrophobic interactions, also labeled as one of the most important interactions in the complexation process. When analyzing the relation of the complex association constant to the retention of the investigated solutes, the retention factor is directly proportional to the values of BE and UE, while inversely proportional to the values of BE. Thermodynamically stable inclusion complexes are those with the lowest BE; therefore, an increase in BE causes the instability of inclusion complexes and thus an increase in retention factor is expected. Nevertheless, the authors hypothesize that the increase in EE

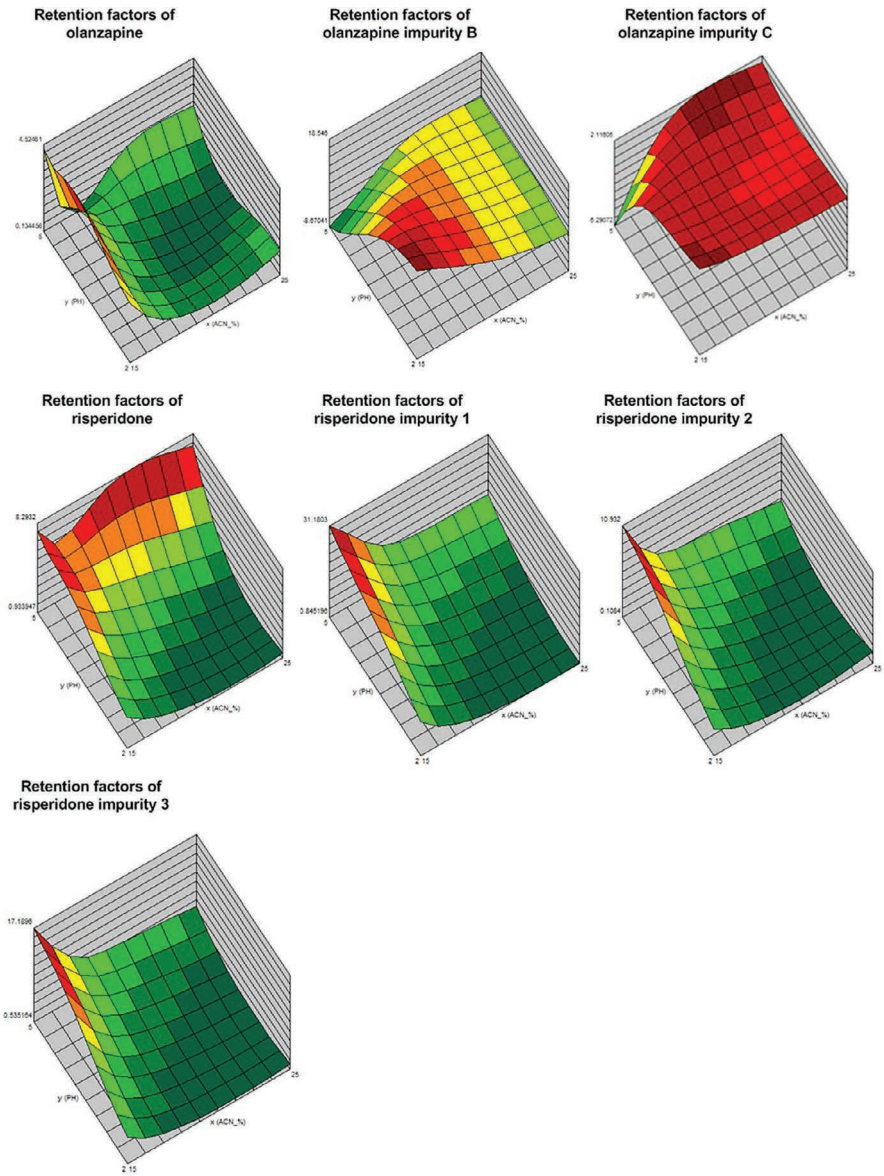


FIGURE 3.4 3D response surface plots of predicted retention factor against acetonitrile content in the mobile phase and pH

From Maljurić, N. et al., Analytical and Bioanalytical Chemistry, 410, 2533—2550, 2018., reused with publisher's permission.

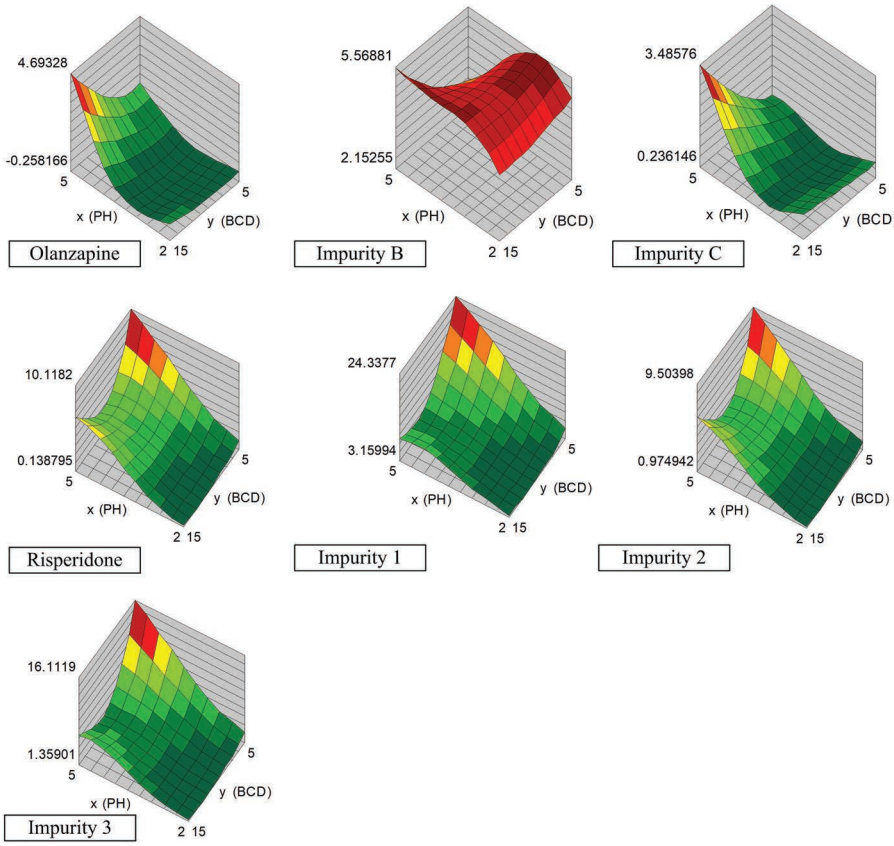


FIGURE 3.5 3D response surface plots of predicted retention factor against β-CD concentration in the mobile phase and pH

From Maljurić, N. et al., Journal of Chromatography A, 1619, 460971, 2020., reused with publisher's permission.

and UE encourages the inclusion of complex formation and consequent reduction in retention factor.

3.5.2 QSRR Model as a Potential Tool in the Chromatographic Determination of Stability Constants and Accompanying Thermodynamic Parameters

The QSRR model developed by Maljurić et al. was used to predict retention factor values of each substance from the model mixture under the defined ranges of β-CD concentration and varying column temperature. Predicting the change in retention factor value is the basis for proposing a novel *in silico* approach to assess complex stability constants and accompanying thermodynamic parameters in RP-HPLC. The

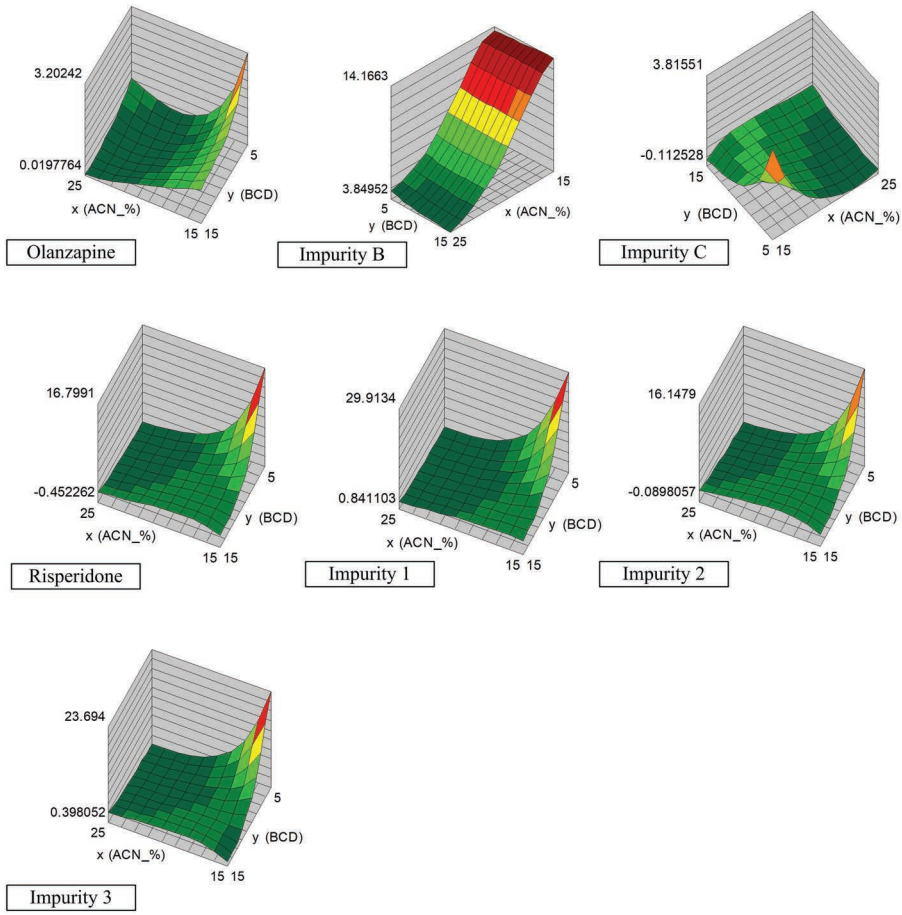


FIGURE 3.6 3D response surface plots of predicted retention factor against β-CD concentration and acetonitrile content in the mobile phase

From Maljurić, N. et al., Journal of Chromatography A, 1619, 460971, 2020., reused with publisher's permission.

applied range of acetonitrile was quite narrow, since values lower than 15% (v/v) could unreasonably prolong the retention time, while percentages above 20% (v/v) could compromise the complexation equilibrium.

The developed model successfully predicted the retention factors of all investigated compounds under the examined conditions. In this way, its applicability in predicting the retention change occurring upon complexation was confirmed. However, the general rule for decreasing the retention factor with an increase in β-CD concentration was not strictly followed. In that manner, under certain conditions, stability constants for certain compounds could be calculated by tracking the decrease in retention factor

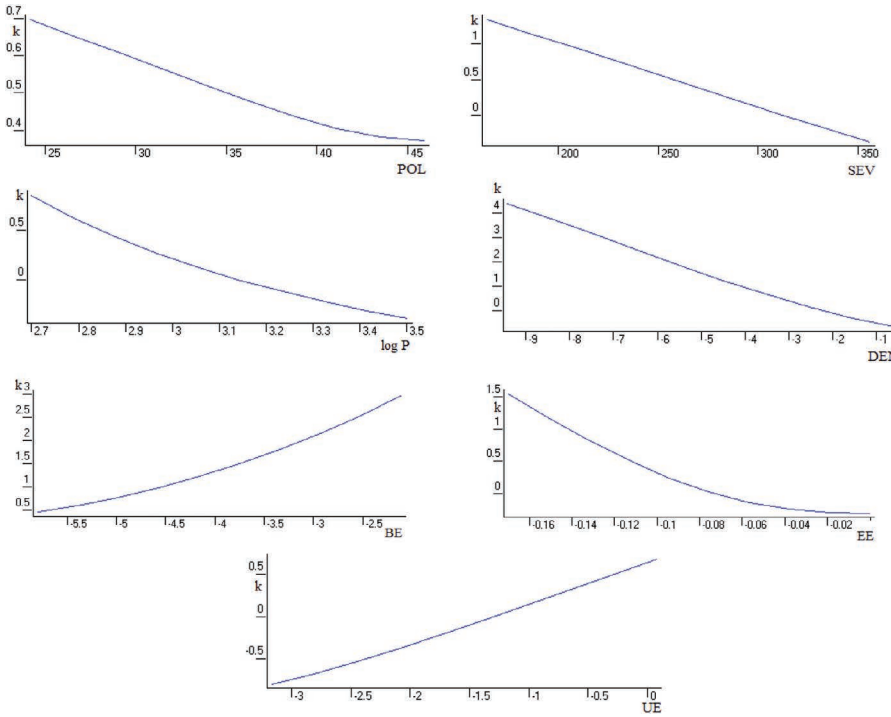


FIGURE 3.7 The effect of the most influential molecular descriptors on retention factor

values, but there were conditions under which an increase in β-CD concentration in the mobile phase led to the simultaneous increase in retention factor value. Then the authors tried to find reasons for such behavior by detailed analyses of experimental conditions and their relation to the complexation process. It was observed that the percentage of acetonitrile, as well as the pH of the aqueous phase, plays an important role. Namely, if the acetonitrile content in the mobile phase is 15% (v/v), stability constants of inclusion complexes formed between β-CD and risperidone and its structurally related impurities could be calculated on the basis of QSRR predicted retention factor change, regardless of the pH values of the aqueous phase. However, this is not the case if the acetonitrile percentage is increased to 20% (v/v). In such an experimental setting, the retention factor of risperidone decreased with an increase in β-CD concentration only if pH was 2, while retention factors of risperidone impurities increased if the β-CD concentration was 10 mM or higher. Consequently, if the retention factors of the investigated analytes increase with an increase in β-CD concentration, the assessment of complex stability constants is disabled (11).

In the case of olanzapine inclusion complexes with β-CD, only one combination of experimental parameters enabled the expected trend in retention factor values and consequent calculation of complex stability constants. It was proven by both HPLC experiments and the developed QSRR model that stability constants could

be determined if acetonitrile is set to 15% (v/v) and pH to 2.0. Further, the stability constants for complexes formed between olanzapine impurity B and β-CD were successfully calculated at any pH value if the acetonitrile content in the mobile phase is 15% (v/v). On the other hand, if acetonitrile content is 20% (v/v), complex stability constants could not be assessed if pH is set to 5. This observation was unexpected, since olanzapine impurity B is in its non-ionized form across the investigated pH range. However, it helped to create the opinion that analyte structure is not the only factor influencing the complexation and retention behavior in such a complicated chromatographic system.

Although in the case of all previously mentioned analytes, it seemed that acetonitrile influence is stronger than pH, in the case of impurity C:β-CD inclusion complexes, stability constants could be calculated if pH is set to 2.0 or 3.5 regardless of the acetonitrile content in the mobile phase (11).

In regular RP-HPLC systems, the solute's retention behavior depends on the interactions of its molecular structure with either the mobile or stationary phase, while in β-CD modified RP-HPLC there is an additional component, the dissolved β-CD. Therefore, there is a possibility for the dissolved β-CD to form interactions with either solute or other components of the chromatographic system. There are different factors influencing the binding process between the guest molecule and β-CD cavity, such as the molecular structure of the compound, the type of additives applied and/or steric factor. Additionally, there is a need for a favorable net energetic drive force, which would be able to allocate the equilibrium in the direction of the complex formation (23).

Based on the results of this study, the authors hypothesized that mobile phase pH influences retention behavior to a large extent, since it determines the ionization form of the compound on the one hand and the ionization form of free silanol groups on the surface of the stationary phase on the other hand. If the mobile phase pH is lower than 3, stationary phase free silanol groups are non-ionized, thus not establishing secondary interactions with the solute. If the solute is in its non-ionized form, it is retained more by the stationary phase; therefore, faster elution is accomplished by changing the pH and transferring the solute to the ionized form. This served the authors to explain the possibility of assessing stability constants if pH is 2 and acetonitrile content in the mobile phase is 15% (v/v). Under such conditions, secondary interactions with the stationary phase are reduced. Moreover, lower acetonitrile content is favorable in terms of complex formations, since the mobile phase remains polar and not a favorable environment to the apolar solute.

The size match between the investigated solutes and β-CD cavity also appeared as an important factor in complexation, but not the leading one, since in the author's previous research it was shown that risperidone and its impurities are only partially incorporated into the β-CD cavity (10).

To provide proof of concept for the presented methodology, the authors performed certain verification experiments. As recognized methods for inclusion complex stability constant assessment, HPLC and UV/Vis experiments were performed and calculated complex stability constants were compared to those obtained by utilizing the QSRR methodology. Figures 3.8 and 3.9 show the comparison between stability

constant values obtained by different methods. Stability constants obtained by HPLC experiments were in compliance with the QSRR predicted values under the same experimental setting. The UV/Vis method enabled confirmation of complex formation, even under the conditions precluding stability constant assessment with HPLC and QSRR approaches. Figures 3.8 and 3.9 show that although the numerical values of complex stability constants are not the same, their trend equals across different methodologies (11). Although the numerical values of complex stability constants should be unconstrained by the methodological approach applied, the literature search showed the opposite (32).

Apart from its ability to predict retention change caused by inclusion complexation, the QSRR model was also used to provide information about thermodynamic

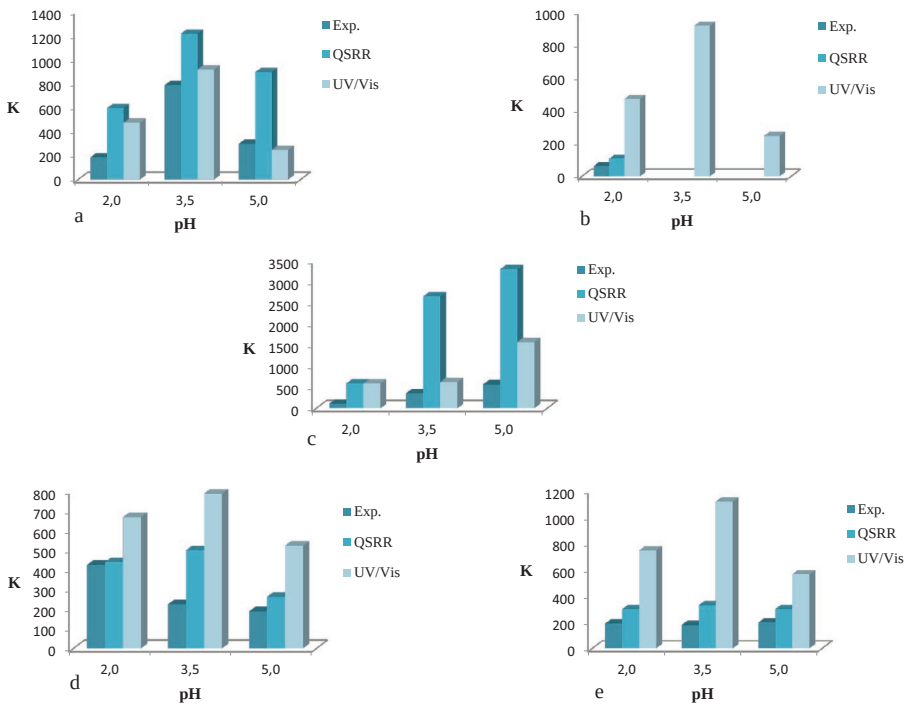


FIGURE 3.8 Complex stability constants calculated by HPLC experiments, QSRR model, and UV/Vis spectroscopy for complexes formed between risperidone and its impurities and β-CD under varying acetonitrile content in the mobile phase.

8a: Risperidone—β-CD complex, if content of acetonitrile in the mobile phase is 15% (v/v)
8b: Risperidone—β-CD complex, if content of acetonitrile in the mobile phase is 20% (v/v)
8c: Risperidone impurity 1—β-CD complex, if content of acetonitrile in the mobile phase is 15% (v/v)
8d: Risperidone impurity 2—β-CD complex, if content of acetonitrile in the mobile phase is 15% (v/v)
8e: Risperidone impurity 3—β-CD complex, if content of acetonitrile in the mobile phase is 15% (v/v)

From Maljurić, N. et al., Journal of Chromatography A, 1619, 460971, 2020., reused with publisher's permission.

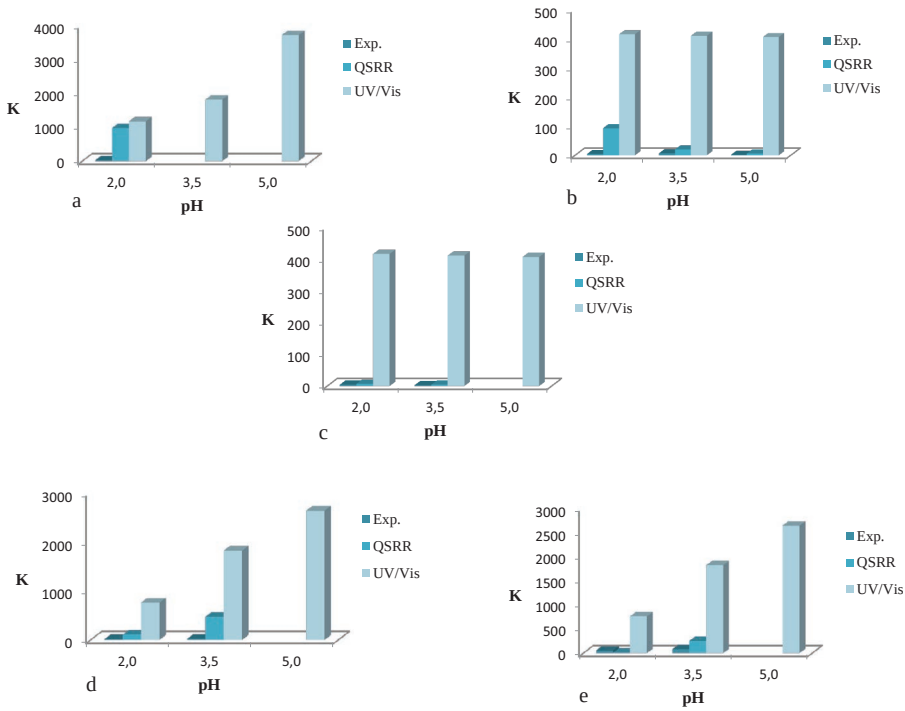


FIGURE 3.9 Complex stability constants calculated by HPLC experiments, QSRR model, and UV/Vis spectroscopy for complexes formed between olanzapine and its impurities and β-CD under varying acetonitrile content in the mobile phase

9a: Olanzapine—β-CD complex, if content of acetonitrile in the mobile phase is 15% (v/v)
9b: Olanzapine impurity B—β-CD complex, if content of acetonitrile in the mobile phase is 15% (v/v)
9c: Olanzapine impurity B—β-CD complex, if content of acetonitrile in the mobile phase is 20% (v/v)
9d: Olanzapine impurity C—β-CD complex, if content of acetonitrile in the mobile phase is 15% (v/v)
9e: Olanzapine impurity C—β-CD complex, if content of acetonitrile in the mobile phase is 20% (v/v)

From Maljurić, N. et al., Journal of Chromatography A, 1619, 460971, 2020., reused with publisher's permission.

parameters under experimental conditions, enabling the calculation of complex stability constants. High negative values of ΔH° surpassing negative values of ΔS° characterizing the complexation of risperidone and/or its impurities with β-CD are indicating the formation of van der Waals interactions. There was only one exception for impurity 3:β-CD complex under one set of experimental conditions, where ΔS° are highly negative but could be attributed to intermolecular hydrogen bonding. Olanzapine complexation with β-CD is also enthalpy driven, with negative ΔG values referring to the spontaneous flow of the process. However, in the case of olanzapine impurities complexation ΔS° values are positive, while ΔH° is highly negative. This means that the formed inclusion complexes are not destabilized by hydrogen bonding and impurities are better accommodated in the cavity in comparison to olanzapine (11).

3.6 FUTURE PERSPECTIVES

The difficulties in modeling retention in systems, such as β-CD modified RP-HPLC, arise from multiple interactions in which a solute is capable of forming, with the stationary phase, mobile phase and β-CD dissolved in the mobile phase. In addition, the possibility of a solute to interact with β-CD adsorbed onto the stationary phase surface should not be neglected. Using risperidone, olanzapine, and their structurally related impurities as model substances, Maljurić et al. successfully developed a QSRR able to reveal retention behavior of given solutes in such a complicated RP-HPLC system. The novelty of the modeling approach Maljurić et al. applied was reflected in introducing complex association constants as descriptors characterizing the formed inclusion complexes, as inputs of the QSRR model. After verifying the model's validity and predictive ability, it was successfully used in the green chromatography RP-HPLC method development. The RP-HPLC methods for separation of given analytes were sped up; thus, the organic solvent consumption was reduced and methods were labeled as eco-friendly.

Apart from applying the model in green RP-HPLC method development, it could be used as an *in silico* tool in the assessment of inclusion complex stability and accompanying thermodynamic parameters. The benefits of this approach account for savings in terms of time and costs since the *in silico* approach could successfully replace the experimental one. Although relatively similar results were obtained with RP-HPLC experiments and QSRR model prediction, the suitability of retention factor change as a measure of inclusion complex stability should be reassessed, since the expected decrease in retention factor upon the increase in β-CD concentration was not observed under a broad range of experimental parameters. However, these observations also brought another benefit of applying the QSRR model, which is to define the experimental space within which interactions leading to complexation would be the dominant ones and thus stability constants and thermodynamic parameters could be calculated.

The presented results provide a good basis for further research, but the conclusions cannot be generalized due to the limited number of analyzed compounds. In order to obtain more general conclusions, the model mixture should be complemented with compounds of varying ability to form inclusion complexes, specifically those with different physicochemical characteristics, especially lipophilicity and ionization ability. Since retention behavior is dependent not only on experimental parameters but also on analyte characteristics, a broad range of characteristics should be covered to provide the rationale for the observed phenomena, which is not in line with the recognized theory of the abovementioned β-CD modified RP-HPLC systems.

Throughout the literature, the adsorption of the CD onto the stationary phase surface was considered negligible. However, the adsorption of a free CD, as well as formed inclusion complexes, could influence the retention behavior. For that reason, in future research, the adsorption of the CD onto the stationary phase surface should be evaluated and incorporated in the QSRR model. Therefore, its influence on retention behavior, as well as joint influence with other parameters, could be observed. In this way, researchers could obtain additional information, enabling them to immerse themselves more deeply into the occurring retention phenomenon. When

all the components of this complicated chromatographic system are well assessed, the proper application of developed QSRR models could be recommended and help both the research community and industries in replacing extensive experiments with an *in silico* approach.

3.7 CONFLICT OF INTEREST

The authors declare no conflicts of interest.

3.8 ACKNOWLEDGMENTS

These results are part of Project no. 451-03-68/2020-14/200161, financed by the Ministry of Education, Science and Technological Development of the Republic of Serbia.

REFERENCES

[1] Cielecka-Piontek J, Zalewski P, Jelińska A, Garbacki P. UHPLC: the greening face of liquid chromatography. *Chromatographia*. 2013;76(21–22):1429–1437.
[2] Armenta S, Garrigues S, De la Guardia M. Green analytical chemistry. *TrAC Trends in Analytical Chemistry*. 2008;27(6):497–511.
[3] González-Ruiz V, León AG, Olives AI, Martin MA, Menéndez JC. Eco-friendly liquid chromatographic separations based on the use of cyclodextrins as mobile phase additives. *Green Chemistry*. 2011;13(1):115–126.
[4] Płotka J, Tobiszewski M, Sulej AM, Kupska M, Górecki T, Namieśnik J. Green chromatography. *Journal of Chromatography A*. 2013;1307:1–20.
[5] Fifere A, Marangoci N, Maier S, Coroaba A, Maftei D, Pinteala M. Theoretical study on β-cyclodextrin inclusion complexes with propiconazole and protonated propiconazole. *Beilstein Journal of Organic Chemistry*. 2012;8:2191.
[6] Čolović J, Kalinić M, Vemić A, Erić S, Malenović A. Investigation into the phenomena affecting the retention behavior of basic analytes in chaotropic chromatography: joint effects of the most relevant chromatographic factors and analytes' molecular properties. *Journal of Chromatography A*. 2015;1425:150–157.
[7] Golubović J, Birkemeyer C, Protić A, Otašević B, Zečević M. Structure–response relationship in electrospray ionization-mass spectrometry of sartans by artificial neural networks. *Journal of Chromatography A*. 2016;1438:123–132.
[8] Cserháti T, Forgács E. *Cyclodextrins in chromatography*. Royal Society of Chemistry, Cambridge, UK (2003).
[9] Spencer B. *Separation of isomers and structurally related compounds using cyclodextrins as mobile phase and buffer additives in high performance liquid chromatography and capillary electrophoresis*. McGill University Libraries; 1996.
[10] Maljurić N, Golubović J, Otašević B, Zečević M, Protić A. Quantitative structure–retention relationship modeling of selected antipsychotics and their impurities in green liquid chromatography using cyclodextrin mobile phases. *Analytical and Bioanalytical Chemistry*. 2018:1–18.
[11] Maljurić N, Otašević B, Malenović A, Zečević M, Protić A. Quantitative structure retention relationship modeling as potential tool in chromatographic determination of stability constants and thermodynamic parameters of β-cyclodextrin complexation process. *Journal of Chromatography A*. 2020:460971.
[12] Dodziuk H. *Cyclodextrins and their complexes: chemistry, analytical methods, applications*. John Wiley & Sons; 2006.

[13] Jin Z-Y. *Cyclodextrin chemistry: preparation and application*. World Scientific; 2013.
[14] Crini G. A history of cyclodextrins. *Chemical Reviews*. 2014;114(21):10940–10975.
[15] Morin-Crini N, Fourmentin S, Fenyvesi É, Lichtfouse E, Torri G, Fourmentin M, et al. History of cyclodextrins. In *The history of cyclodextrins*. Springer; 2020. p. 1–93.
[16] Chankvetadze B. *Liquid chromatographic separation of enantiomers*. Liquid Chromatography. Elsevier; 2017. p. 69–86.
[17] Dodziuk H. Molecules with holes—cyclodextrins. *Cyclodextrins and Their Complexes*. 2006:1–30.
[18] Zabel V, Saenger W, Mason SA. Topography of cyclodextrin inclusion complexes. Part 23. Neutron diffraction study of the hydrogen bonding in. beta.-cyclodextrin undecahydrate at 120 K: from dynamic flip-flops to static homodromic chains. *Journal of the American Chemical Society*. 1986;108(13):3664–3673.
[19] Dodziuk H, Nowiński K. Structure of cyclodextrins and their complexes: Part 2. Do cyclodextrins have a rigid truncated-cone structure? *Journal of Molecular Structure: THEOCHEM*. 1994;304(1):61–68.
[20] Koshland Jr D. The lock-and-key principle and the induced-fit theory. *Angewandte Chemie International Edition in English*. 1994;33:2475–2478.
[21] Szejtli J. Introduction and general overview of cyclodextrin chemistry. *Chemical Reviews*. 1998;98(5):1743–1754.
[22] Shieh WJ, Hedges A. Properties and applications of cyclodextrins. *Journal of Macromolecular Science, Part A: Pure and Applied Chemistry*. 1996;33(5):673–683.
[23] Del Valle EM. Cyclodextrins and their uses: a review. *Process Biochemistry*. 2004;39(9):1033–1046.
[24] Mazzobre M, Elizalde B, dos Santos C, Cevallos PP, Buera M. Nanoencapsulation of food ingredients in cyclodextrins: effect of water interactions and ligand structure. *Functional Food Product Development*. 2010;2:24.
[25] Abarca RL, Rodriguez FJ, Guarda A, Galotto MJ, Bruna JE. Characterization of beta-cyclodextrin inclusion complexes containing an essential oil component. *Food Chemistry*. 2016;196:968–975.
[26] Marques HMC. A review on cyclodextrin encapsulation of essential oils and volatiles. *Flavour and Fragrance Journal*. 2010;25(5):313–326.
[27] Schneiderman E, Stalcup AM. Cyclodextrins: a versatile tool in separation science. *Journal of Chromatography B: Biomedical Sciences and Applications*. 2000;745(1): 83–102.
[28] Moraes CM, Abrami P, de Paula E, Braga AF, Fraceto LF. Study of the interaction between S (–) bupivacaine and 2-hydroxypropyl-β-cyclodextrin. *International Journal of Pharmaceutics*. 2007;331(1):99–106.
[29] Ravelet C, Geze A, Villet A, Grosset C, Ravel A, Wouessidjewe D, et al. Chromatographic determination of the association constants between nimesulide and native and modified β-cyclodextrins. *Journal of Pharmaceutical and Biomedical Analysis*. 2002;29(3):425–430.
[30] Shuang S, Choi MM. Retention behaviour and fluorimetric detection of procaine hydrochloride using carboxymethyl-β-cyclodextrin as an additive in reversed-phase liquid chromatography. *Journal of Chromatography A*. 2001;919(2):321–329.
[31] Singh R, Bharti N, Madan J, Hiremath S. Characterization of cyclodextrin inclusion complexes—a review. *Journal of Pharmaceutical Science and Technology*. 2010;2(3):171–183.
[32] Loftsson T, Másson M, Brewster ME. Self-association of cyclodextrins and cyclodextrin complexes. *Journal of Pharmaceutical Sciences*. 2004;93(5):1091–1099.
[33] Mura P. Analytical techniques for characterization of cyclodextrin complexes in aqueous solution: a review. *Journal of Pharmaceutical and Biomedical Analysis*. 2014;101:238–250.

[34] Szente L, Szemán J, Sohajda T. Analytical characterization of cyclodextrins: history, official methods and recommended new techniques. *Journal of Pharmaceutical and Biomedical Analysis*. 2016;130:347–365.
[35] Kfoury M, Landy D, Fourmentin S. Characterization of cyclodextrin/volatile inclusion complexes: a review. *Molecules*. 2018;23(5):1204.
[36] Leyva E, Moctezuma E, Loredo-Carrillo SE, Espinosa-González CG, Cárdenas-Chaparro A. Determination of the structure of quinolone-γ-cyclodextrin complexes and their binding constants by means of UV–Vis and 1 H NMR. *Journal of Inclusion Phenomena and Macrocyclic Chemistry*. 2018;91(3–4):211–218.
[37] Vashi P, Cukrowski I, Havel J. Stability constants of the inclusion complexes of ß-cyclodextrin with various adamantane derivatives. A UV-Vis study. *South African Journal of Chemistry*. 2001;54.
[38] Zhou X, Liang JF. A fluorescence spectroscopy approach for fast determination of β-cyclodextrin-guest binding constants. *Journal of Photochemistry and Photobiology A: Chemistry*. 2017;349:124–128.
[39] Danel C, Azaroual N, Brunel A, Lannoy D, Vermeersch G, Odou P, et al. Study of the complexation of risperidone and 9-hydroxyrisperidone with cyclodextrin hosts using affinity capillary electrophoresis and 1 H NMR spectroscopy. *Journal of Chromatography A*. 2008;1215(1):185–193.
[40] Recio R, Elhalem E, Benito JM, Fernández I, Khiar N. NMR study on the stabilization and chiral discrimination of sulforaphane enantiomers and analogues by cyclodextrins. *Carbohydrate Polymers*. 2018;187:118–125.
[41] Landy D, Fourmentin S, Salome M, Surpateanu G. Analytical improvement in measuring formation constants of inclusion complexes between β-cyclodextrin and phenolic compounds. *Journal of Inclusion Phenomena and Macrocyclic Chemistry*. 2000;38(1–4):187–198.
[42] Ibrahim M, Shehatta I, Al-Nayeli A. Voltammetric studies of the interaction of lumazine with cyclodextrins and DNA. *Journal of Pharmaceutical and Biomedical Analysis*. 2002;28(2):217–225.
[43] Yáñez C, Núñez-Vergara L, Squella J. Differential pulse polarographic and UV-Vis spectrophotometric study of inclusion complexes formed by 1, 4-dihydropyridine calcium antagonists, nifedipine and nicardipine with β-Cyclodextrin. *Electroanalysis: An International Journal Devoted to Fundamental and Practical Aspects of Electroanalysis*. 2003;15(22):1771–1777.
[44] Hansen LD, Fellingham GW, Russell DJ. Simultaneous determination of equilibrium constants and enthalpy changes by titration calorimetry: methods, instruments, and uncertainties. *Analytical Biochemistry*. 2011;409(2):220–229.
[45] Sursyakova VV, Maksimov NG, Levdansky VA, Rubaylo AI. Combination of phase-solubility method and capillary zone electrophoresis to determine binding constants of cyclodextrins with practically water-insoluble compounds. *Journal of Pharmaceutical and Biomedical Analysis*. 2018;160:12–18.
[46] Lancioni C, Keunchkarian S, Castells CB, Gagliardi LG. Determination of thermodynamic binding constants by affinity capillary electrophoresis. *Talanta*. 2019;192:448–454.
[47] Kfoury M, Auezova L, Greige-Gerges H, Fourmentin S. Development of a total organic carbon method for the quantitative determination of solubility enhancement by cyclodextrins: application to essential oils. *Analytica Chimica Acta*. 2016;918:21–25.
[48] Ceborska M, Szwed K, Asztemborska M, Wszelaka-Rylik M, Kicińska E, Suwińska K. Study of β-cyclodextrin inclusion complexes with volatile molecules geraniol and α-terpineol enantiomers in solid state and in solution. *Chemical Physics Letters*. 2015;641:44–50.
[49] Asztemborska M, Bielejewska A, Duszczyk K, Sybilska D. Comparative study on camphor enantiomers behavior under the conditions of gas–liquid chromatography and

reversed-phase high-performance liquid chromatography systems modified with α-and β-cyclodextrins. *Journal of Chromatography A*. 2000;874(1):73–80.
[50] López-Nicolás JM, Núñez-Delicado E, Pérez-López AJ, Barrachina ÁC, Cuadra-Crespo P. Determination of stoichiometric coefficients and apparent formation constants for β-cyclodextrin complexes of trans-resveratrol using reversed-phase liquid chromatography. *Journal of Chromatography A*. 2006;1135(2):158–165.
[51] López-Nicolás JM, García-Carmona F. Rapid, simple and sensitive determination of the apparent formation constants of trans-resveratrol complexes with natural cyclodextrins in aqueous medium using HPLC. *Food Chemistry*. 2008;109(4):868–875.
[52] Gazpio C, Sánchez M, García-Zubiri IX, Vélaz I, Martínez-Ohárriz C, Martín C, et al. HPLC and solubility study of the interaction between pindolol and cyclodextrins. *Journal of Pharmaceutical and Biomedical Analysis*. 2005;37(3):487–492.
[53] de Melo NF, Grillo R, Rosa AH, Fraceto LF. Interaction between nitroheterocyclic compounds with β-cyclodextrins: phase solubility and HPLC studies. *Journal of Pharmaceutical and Biomedical Analysis*. 2008;47(4–5):865–869.
[54] El-Barghouthi M, Masoud N, Al-Kafawein J, Zughul M, Badwan A. Host–guest interactions of risperidone with natural and modified cyclodextrins: phase solubility, thermodynamics and molecular modeling studies. *Journal of Inclusion Phenomena and Macrocyclic Chemistry*. 2005;53(1–2):15–22.
[55] Rozou S, Michaleas S, Antoniadou-Vyza E. Study of structural features and thermodynamic parameters, determining the chromatographic behaviour of drug–cyclodextrin complexes. *Journal of Chromatography A*. 2005;1087(1):86–94.
[56] Rekharsky MV, Inoue Y. Complexation thermodynamics of cyclodextrins. *Chemical Reviews*. 1998;98(5):1875–1918.
[57] Connors KA. The stability of cyclodextrin complexes in solution. *Chemical Reviews*. 1997;97(5):1325–1358.
[58] Kríž Z, Koča J, Imberty A, Charlot A, Auzély-Velty R. Investigation of the complexation of (+)-catechin by β-cyclodextrin by a combination of NMR, microcalorimetry and molecular modeling techniques. *Organic & Biomolecular Chemistry*. 2003;1(14):2590–2595.
[59] Sun D-Z, Li L, Qiu X-M, Liu F, Yin B-L. Isothermal titration calorimetry and 1H NMR studies on host–guest interaction of paeonol and two of its isomers with β-cyclodextrin. *International Journal of Pharmaceutics*. 2006;316(1–2):7–13.
[60] Manzoori JL, Amjadi M. Spectrofluorimetric study of host–guest complexation of ibuprofen with β-cyclodextrin and its analytical application. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*. 2003;59(5):909–916.
[61] Saroj MK, Payal R, Jain SK, Sharma N, Rastogi RC. Investigation of indole chalcones encapsulation in β-cyclodextrin: determination of stoichiometry, binding constants and thermodynamic parameters. *Journal of Inclusion Phenomena and Macrocyclic Chemistry*. 2018;90(3–4):305–320.
[62] Liu Y, Jin L, Zhang H-Y. Inclusion complexation thermodynamics of acridine red and rhodamine B by natural and novel oligo (ethylenediamine) tethered schiff base β-cyclodextrin. *Journal of Inclusion Phenomena and Macrocyclic Chemistry*. 2002;42(1–2):115–120.
[63] Aljhni R, Andre C, Lethier L, Guillaume YC. An HPLC chromatographic framework to analyze the β-cyclodextrin/solute complexation mechanism using a carbon nanotube stationary phase. *Talanta*. 2015;144:226–232.
[64] Solovev A, Solov'ev V. 3D molecular fragment descriptors for structure–property modeling: predicting the free energies for the complexation between antipodal guests and β-cyclodextrins. *Journal of Inclusion Phenomena and Macrocyclic Chemistry*. 2017;89(1–2):167–175.
[65] Katritzky AR, Fara DC, Yang H, Karelson M, Suzuki T, Solov'ev VP, et al. Quantitative structure–property relationship modeling of β-Cyclodextrin complexation free

energies. *Journal of Chemical Information and Computer Sciences*. 2004;44(2): 529–541.
[66] Armstrong DW, Nome F, Spino LA, Golden TD. Efficient detection and evaluation of cyclodextrin multiple complex formation. *Journal of the American Chemical Society*. 1986;108(7):1418–1421.
[67] Oh I, Lee M-Y, Lee Y-B, Shin S-C, Park I. Spectroscopic characterization of ibuprofen/2-hydroxypropyl-β-cyclodextrin inclusion complex. *International Journal of Pharmaceutics*. 1998;175(2):215–223.
[68] Ahn S, Ramirez J, Grigorean G, Lebrilla CB. Chiral recognition in gas-phase cyclodextrin: amino acid complexes—Is the three point interaction still valid in the gas phase? *Journal of the American Society for Mass Spectrometry*. 2001;12(3):278–287.
[69] Dotsikas Y, Loukas YL. Efficient determination and evaluation of model cyclodextrin complex binding constants by electrospray mass spectrometry. *Journal of the American Society for Mass Spectrometry*. 2003;14(10):1123–1129.
[70] Hancock T, Put R, Coomans D, Vander Heyden Y, Everingham Y. A performance comparison of modern statistical techniques for molecular descriptor selection and retention prediction in chromatographic QSRR studies. *Chemometrics and Intelligent Laboratory Systems*. 2005;76(2):185–196.
[71] Kaliszan R. Quantitative structure-retention relationships. *Analytical Chemistry*. 1992;64(11):619A–631A.
[72] Kaliszan R, van Straten MA, Markuszewski M, Cramers CA, Claessens HA. Molecular mechanism of retention in reversed-phase high-performance liquid chromatography and classification of modern stationary phases by using quantitative structure–retention relationships. *Journal of Chromatography A*. 1999;855(2):455–486.
[73] Héberger K. Quantitative structure–(chromatographic) retention relationships. *Journal of Chromatography A*. 2007;1158(1–2):273–305.
[74] Aschi M, D'Archivio AA, Maggi MA, Mazzeo P, Ruggieri F. Quantitative structure-retention relationships of pesticides in reversed-phase high-performance liquid chromatography. *Analytica Chimica Acta*. 2007;582(2):235–242.
[75] Goryński K, Bojko B, Nowaczyk A, Buciński A, Pawliszyn J, Kaliszan R. Quantitative structure–retention relationships models for prediction of high performance liquid chromatography retention time of small molecules: endogenous metabolites and banned compounds. *Analytica Chimica Acta*. 2013;797:13–19.
[76] Bączek T, Kaliszan R, Novotná K, Jandera P. Comparative characteristics of HPLC columns based on quantitative structure–retention relationships (QSRR) and hydrophobic-subtraction model. *Journal of Chromatography A*. 2005;1075(1–2):109–115.
[77] Sadek PC, Carr PW, Doherty RM, Kamlet MJ, Taft RW, Abraham MH. Study of retention processes in reversed-phase high-performance liquid chromatography by the use of the solvatochromic comparison method. *Analytical Chemistry*. 1985;57(14):2971–2978.
[78] Abraham MH, McGowan J. The use of characteristic volumes to measure cavity terms in reversed phase liquid chromatography. *Chromatographia*. 1987;23(4):243–246.
[79] Kaliszan R, Bączek T, Buciński A, Buszewski B, Sztupecka M. Prediction of gradient retention from the linear solvent strength (LSS) model, quantitative structure-retention relationships (QSRR), and artificial neural networks (ANN). *Journal of Separation Science*. 2003;26(3–4):271–282.
[80] Hsieh M-M, Dorsey JG. Accurate determination of log k′ w in reversed-phase liquid chromatography: implications for quantitative structure—retention relationships. *Journal of Chromatography A*. 1993;631(1–2):63–78.
[81] Schilling K, Krmar J, Maljurić N, Pawellek R, Protić A, Holzgrabe U. Quantitative structure-property relationship modeling of polar analytes lacking UV chromophores

to charged aerosol detector response. *Analytical and Bioanalytical Chemistry*. 2019;411(13):2945–2959.
[82] Todeschini R, Consonni V. *Handbook of molecular descriptors*. John Wiley & Sons; 2008.
[83] Bodzioch K, Durand A, Kaliszan R, Bączek T, Vander Heyden Y. Advanced QSRR modeling of peptides behavior in RPLC. *Talanta*. 2010;81(4):1711–1718.
[84] Karelson M, Lobanov VS, Katritzky AR. Quantum-chemical descriptors in QSAR/QSPR studies. *Chemical Reviews*. 1996;96(3):1027–1044.
[85] Todeschini R, Consonni V. *Molecular descriptors for chemoinformatics: Volume I: Alphabetical listing/volume II: Appendices, references*. John Wiley & Sons; 2009.
[86] Leardi R. Experimental design in chemistry: a tutorial. *Analytica Chimica Acta*. 2009;652(1–2):161–172.
[87] Ferreira SLC, Bruns RE, da Silva EGP, Dos Santos WNL, Quintella CM, David JM, et al. Statistical designs and response surface techniques for the optimization of chromatographic systems. *Journal of Chromatography A*. 2007;1158(1–2):2–14.
[88] Bezerra MA, Santelli RE, Oliveira EP, Villar LS, Escaleira LA. Response surface methodology (RSM) as a tool for optimization in analytical chemistry. *Talanta*. 2008;76(5):965–977.
[89] Goodarzi M, Jensen R, Vander Heyden Y. QSRR modeling for diverse drugs using different feature selection methods coupled with linear and nonlinear regressions. *Journal of Chromatography B*. 2012;910:84–94.
[90] Krmar J, Vukićević M, Kovačević A, Protić A, Zečević M, Otašević B. Performance comparison of nonlinear and linear regression algorithms coupled with different attribute selection methods for quantitative structure-retention relationships modelling in micellar liquid chromatography. *Journal of Chromatography A*. 2020:461146.
[91] Graupe D. *Principles of artificial neural networks*. World Scientific; 2013.
[92] Marini F, Bucci R, Magrì A, Magrì A. Artificial neural networks in chemometrics: history, examples and perspectives. *Microchemical Journal*. 2008;88(2):178–185.
[93] Zupan J. Introduction to artificial neural network (ANN) methods: what they are and how to use them. *Acta Chimica Slovenica*. 1994;41:327.
[94] Smits J, Melssen W, Buydens L, Kateman G. Using artificial neural networks for solving chemical problems: Part I. Multi-layer feed-forward networks. *Chemometrics and Intelligent Laboratory Systems*. 1994;22(2):165–189.
[95] Carlucci G, D'Archivio AA, Maggi MA, Mazzeo P, Ruggieri F. Investigation of retention behaviour of non-steroidal anti-inflammatory drugs in high-performance liquid chromatography by using quantitative structure–retention relationships. *Analytica Chimica Acta*. 2007;601(1):68–76.
[96] Golubović J, Protić A, Zečević M, Otašević B, Mikić M, Živanović L. Quantitative structure–retention relationships of azole antifungal agents in reversed-phase high performance liquid chromatography. *Talanta*. 2012;100:329–337.
[97] Veerasamy R, Rajak H, Jain A, Sivadasan S, Varghese CP, Agrawal RK. Validation of QSAR models-strategies and importance. *International Journal of Drug Design & Discovery*. 2011;3:511–519.
[98] Roy K, Kar S, Das RN. *A primer on QSAR/QSPR modeling: fundamental concepts*. Springer; 2015.
[99] Bordás B, Kömíves T, Szántó Z, Lopata A. Comparative three-dimensional quantitative structure−activity relationship study of safeners and herbicides. *Journal of Agricultural and Food Chemistry*. 2000;48(3):926–931.
[100] Fan Y, Shi LM, Kohn KW, Pommier Y, Weinstein JN. Quantitative structure-antitumor activity relationships of camptothecin analogues: cluster analysis and genetic algorithm-based studies. *Journal of Medicinal Chemistry*. 2001;44(20):3254–3263.

[101] Suzuki T, Ide K, Ishida M, Shapiro S. Classification of environmental estrogens by physicochemical properties using principal component analysis and hierarchical cluster analysis. *Journal of Chemical Information and Computer Sciences*. 2001;41(3):718–726.
[102] Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, et al. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *Journal of Computational Chemistry*. 1998;19(14):1639–1662.
[103] Morris GM, Lim-Wilby M. Molecular docking. In *Molecular modeling of proteins*. Springer; 2008. p. 365–382.
[104] Consonni V, Todeschini R, Pavan M. Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3D molecular descriptors. *Journal of Chemical Information and Computer Sciences*. 2002;42(3):682–692.
[105] Jansook P, Kurkov SV, Loftsson T. Cyclodextrins as solubilizers: formation of complex aggregates. *Journal of Pharmaceutical Sciences*. 2010;99(2):719–729.
[106] Jansook P, Ogawa N, Loftsson T. Cyclodextrins: structure, physicochemical properties and pharmaceutical applications. *International Journal of Pharmaceutics*. 2018;535(1–2):272–284.
[107] Pérez-Garrido A, Helguera AM, Cordeiro MND, Escudero AG. QSPR modelling with the topological substructural molecular design approach: β-cyclodextrin complexation. *Journal of Pharmaceutical Sciences*. 2009;98(12):4557–4576.
[108] Ahmadi P, Ghasemi JB. 3D-QSAR and docking studies of the stability constants of different guest molecules with beta-cyclodextrin. *Journal of Inclusion Phenomena and Macrocyclic Chemistry*. 2014;79(3–4):401–413.
[109] Steffen A, Apostolakis J. On the ease of predicting the thermodynamic properties of beta-cyclodextrin inclusion complexes. *Chemistry Central Journal*. 2007;1(1):1–11.
[110] Ghasemi J, Saaidpour S. QSPR modeling of stability constants of diverse 15-crown-5 ethers complexes using best multiple linear regression. *Journal of Inclusion Phenomena and Macrocyclic Chemistry*. 2008;60(3–4):339–351.
[111] Ghasemi JB, Rofouei M, Salahinejad M. A quantitative structure–property relationships study of the stability constant of crown ethers by molecular modelling: new descriptors for lariat effect. *Journal of Inclusion Phenomena and Macrocyclic Chemistry*. 2011;70(1–2):37–47.
[112] Ghasemi JB, Salahinejad M, Rofouei M, Mousazadeh M. Docking and 3D-QSAR study of stability constants of benzene derivatives as environmental pollutants with α-cyclodextrin. *Journal of Inclusion Phenomena and Macrocyclic Chemistry*. 2012;73(1–4):405–413.
[113] Li H, Sun J, Wang Y, Sui X, Sun L, Zhang J, et al. Structure-based in silico model profiles the binding constant of poorly soluble drugs with β-cyclodextrin. *European Journal of Pharmaceutical Sciences*. 2011;42(1–2):55–64.
[114] Merzlikine A, Abramov YA, Kowsz SJ, Thomas VH, Mano T. Development of machine learning models of β-cyclodextrin and sulfobutylether-β-cyclodextrin complexation free energies. *International Journal of Pharmaceutics*. 2011;418(2):207–216.
[115] Veselinović AM, Veselinović JB, Toropov AA, Toropova AP, Nikolić GM. In silico prediction of the β-cyclodextrin complexation based on Monte Carlo method. *International Journal of Pharmaceutics*. 2015;495(1):404–409.
[116] Cysewski P, Przybyłek M. Predicting value of binding constants of organic ligands to beta-cyclodextrin: application of MARSplines and descriptors encoded in SMILES string. *Symmetry*. 2019;11(7):922.
[117] Ling Y, Klemes MJ, Steinschneider S, Dichtel WR, Helbling DE. QSARs to predict adsorption affinity of organic micropollutants for activated carbon and β-cyclodextrin polymer adsorbents. *Water Research*. 2019;154:217–226.

[118] Šoškić M, Porobić I. Interactions of indole derivatives with β-Cyclodextrin: a quantitative structure-property relationship study. *PLoS ONE*. 2016;11(4):e0154339.
[119] Lawrence XY, Kopcha M. The future of pharmaceutical quality and the path to get there. *International Journal of Pharmaceutics*. 2017;528(1–2):354–359.
[120] Tandon H, Ranjan P, Chakraborty T, Suhag V. Polarizability: a promising descriptor to study chemical–biological interactions. *Molecular Diversity*. 2020:1–14.
[121] Richmond TJ. Solvent accessible surface area and excluded volume in proteins: analytical equations for overlapping spheres and implications for the hydrophobic effect. *Journal of Molecular Biology*. 1984;178(1):63–89.

4 Comprehensive Two-Dimensional Chromatography with Chemometric Data Analysis

Caitlin N. Cain, Timothy J. Trinklein, Sonia Schöneich, Grant S. Ochoa, Sarah C. Rutan, Robert E. Synovec

CONTENTS

4.1 INTRODUCTION

Use of one-dimensional (1D) chromatography is ubiquitous in a variety of fields, such as forensics, environmental science, metabolomics, fuel quality, and food

DOI: 10.1201/9781003330080-4

analysis, for its ability to resolve complex mixtures into their individual components. Specifically, gas (GC) and liquid (LC) chromatography are commonly used to characterize the respective volatile and non-volatile components of a sample. Depending on the analysis goal, either chromatographic platform can be used to fully resolve a specific analyte (or set of analytes) or provide a fingerprint of the entire sample. Hence, these chromatographic platforms must have a high resolving power to effectively achieve these goals. The resolving power for a separation can be measured through its peak capacity, which defines the maximum number of resolvable peaks in a separation. In this context, peaks refer to distinguishable concentration pulses that may be composed of one analyte or multiple analytes. As a result, the number of peaks observed for a given chromatogram is always less than or equal to the number of analytes (i.e., components) in the sample matrix. In 1983, Davis and Giddings developed statistical overlap theory to describe the relationship between analyte overlap and peak capacity [1]. This theory suggests that the peak capacity of 1D separations is inadequate for the separation of multicomponent mixtures [1]. For example, if the number of analytes in a sample equals the peak capacity of the 1D chromatographic instrument (i.e., a saturation factor of 1), then the maximum number of resolvable peaks equals 37% of the peak capacity [1]. Even worse, the maximum number of pure, single-analyte peaks equals 18% of the peak capacity under the same conditions [1]. Therefore, technological advancements in 1D separations have focused on providing higher peak capacities to increase the probability of resolving complex samples containing hundreds of components. However, despite the improvements in peak capacity for 1D chromatography, most of the compounds in these highly saturated samples remain unresolved.

Considering the instrumental and statistical limitations of 1D chromatography, the development of comprehensive two-dimensional (2D) chromatography (such as GC × GC and LC × LC) provides an intriguing means of generating high peak capacities. Comprehensive 2D separations occur when the entire effluent from the primary separation is continuously sampled, reinjected, and separated on a secondary column. While any two separation techniques can be coupled together to yield a comprehensive 2D separation, LC × LC and GC × GC are the prominent methods employed in the literature and thus the focus of this chapter. The first LC × LC separation was demonstrated by Erni and Frei in 1978 [2], which was later improved upon by Bushey and Jorgenson in 1990 [3]. Meanwhile, the first GC × GC instrument was developed by Liu and Phillips in 1991 [4]. These initial reports demonstrated that comprehensive 2D chromatography was a powerful technique for the separation of complex samples due to its increased peak capacity and increased selectivity relative to 1D chromatography. For example, the ideal peak capacity of a comprehensive 2D separation is simply the product of the peak capacities on both dimensions. Klee et al. showed that an ideal GC × GC separation provides a ~ 10-fold increase in peak capacity over its 1D counterpart [5]. Likewise, Stoll et al. demonstrated that for a separation longer than 10 min, the peak capacity of LC × LC is far superior to a fully optimized 1D-LC separation [6]. To achieve these near-ideal increases in peak capacity, the separations in each dimension should be complementary and sufficiently independent, which in some cases can provide compound group-type separations that aid in the interpretation of these chromatograms. Given the increased

resolving power, the use of comprehensive 2D chromatography in the fields mentioned earlier is growing.

While the instrumental technology and operation have improved, the general designs of these instruments are still similar to their initial reports [2–4]. Indeed, comprehensive 2D chromatography relies upon the use of a modulator. The modulator serves as an injection interface that periodically samples the first dimension (^{1}D) column effluent and reinjects it onto the head of the second dimension (^{2}D) column in a sharp pulse. For a separation to be sufficiently comprehensive, each ^{1}D peak must be sampled a minimum of 2–4 times [7,8]. The time between sampling events is denoted by the modulation period, P_M. The use of a short P_M ensures that the resolution of the separation that was achieved on ^{1}D is not seriously degraded because of undersampling. For GC × GC (Figure 4.1), fast ^{2}D separations are achieved with short columns and fast temperature programming, which allows for the P_M to generally range from 1–6 s. Similarly, for LC × LC, the use of shorter columns, smaller particle sizes, higher temperatures, and fast mobile phase gradients allow for the P_M to generally range from 15–30 s. Note that while a longer P_M for LC × LC separations ensures that the unique selectivity of the ^{2}D column will be utilized, it can still result in undersampling. Given the fast ^{2}D separation, the detector must be fast enough to record the signal, whereby the general aim is to obtain at least 10–20 data points across the ^{2}D peak width to produce high-quality data. The resulting data array, which is monitored on either a univariate or multivariate detector, can be transformed into a 2D chromatogram by cutting the data array at the time of each P_M sampling event (Figure 4.1). The data can be visualized as a 2D contour plot, where the x-axis and y-axis describe the ^{1}D and ^{2}D separation times, respectively.

Despite the obvious analytical advantages of applying comprehensive 2D separations, the information-rich data produced from these platforms can be large (approximately 50–500 MB per file) and complex. Conventional data analysis approaches involve an analyst manually identifying, quantifying, and determining the significance of every peak in the chromatogram. Because each chemical compound typically results in 2–4 peaks in sequential ^{2}D chromatograms, this means that conventional analysis is often not feasible for comprehensive 2D separations. Co-eluting compounds can complicate the ability to obtain an accurate identification and quantitation for an analyte. Furthermore, hands-on data analysis can be nearly impossible for the size and complexity of a 2D chromatogram, much less for multiple replicates or different samples. Therefore, advanced computational algorithms (i.e., chemometrics) can be used to obtain the same, if not better, results with less analyst intervention compared to manual approaches in a shorter amount of time. Here, chemometrics refers to the use of linear algebra and statistical methods to extract meaningful chemical information from analytical data sets. Traditionally, the application of chemometrics to chromatographic data has been limited to "in-house" analysis, involving data manipulation in a programming software like MATLAB, R, or Python. While the need for chemometrics implementation with comprehensive 2D separation data has been recognized for nearly two decades [9], the specialized expertise needed to analyze data in these programs has been a major obstacle for the widespread adoption of comprehensive 2D chromatography and chemometrics. Fortunately, chemometric methods are becoming more incorporated into

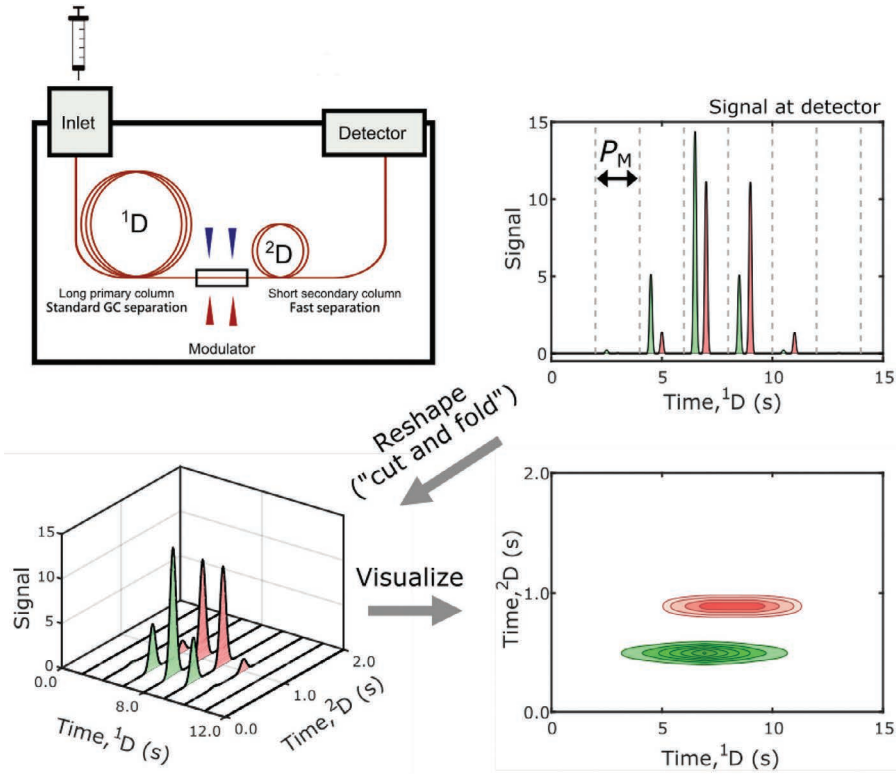


FIGURE 4.1 Schematic of a typical GC × GC instrument and the data collected. The modulator is used to collect fractions of the ^{1}D effluent and reinject those fractions onto the ^{2}D column. Either a univariate or multivariate detector records the analytical signal for each analyte in a vectorized form. The vectorized chromatogram can be cut into segments, based on the modulation period (P_M), and stacked side-by-side. From here, the chromatogram can either be visualized as a three-dimensional plot or a contour plot.

commercial and open-source software packages for GC × GC and LC × LC analysis, providing equitable and easier use [10].

Given these recent developments, this chapter aims to describe the capability, performance, and limitations of common chemometric methods for chromatographic data analysis. First, instrumentation and preprocessing considerations for obtaining and handling GC × GC and LC × LC data will be discussed since these platforms are routinely used for comprehensive 2D chromatography. Furthermore, these experimental decisions can have the greatest effect on the success of chemometric models. The chemometric methods discussed for the analysis of GC × GC and LC × LC data will be broken down into two categories, targeted and non-targeted, based upon the analytical objective (Figure 4.2). Targeted analysis refers to the identification and quantitation of pre-selected (known) analytes of interest. Generally, if the analytes of interest are well-resolved in the chromatogram, or effectively so due to the selectivity

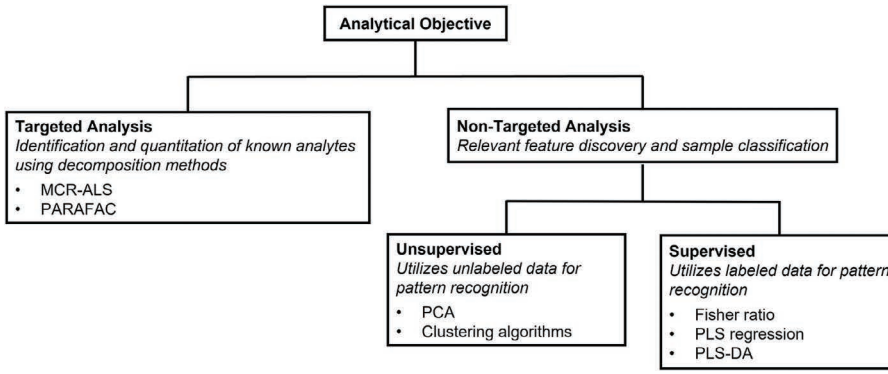


FIGURE 4.2 Overview of chemometric methods based on their approach and analytical goal.

provided by the detector employed, then advanced chemometrics is unnecessary. However, if the analyte is in a region with low chromatographic resolution, traditional identification and quantitation efforts will be severely hindered. Chemometric decomposition of this region can be beneficial in extracting the pure signal for the specified analyte. Herein, two decomposition methods, multivariate curve resolution-alternating least squares (MCR-ALS) and parallel factor analysis (PARAFAC), will be discussed. On the other hand, non-targeted analysis, also referred to as discovery-based analysis, aims to categorize samples and discover compounds responsible for sample differentiation. Non-targeted chemometric techniques can be further categorized as unsupervised or supervised, depending upon whether the method leverages class-based information based upon the experimental design. Unsupervised, non-targeted methods discussed herein are principal components analysis (PCA) and clustering algorithms, while supervised methods like Fisher-ratio analysis, partial least squares (PLS) regression, and partial least squares-discriminant analysis (PLS-DA) will be covered. While this chapter will discuss these methods in their respective groups (Figure 4.2), these methods can be applied in any order, depending upon the analytical objective. For example, the resulting quantitative information gained from targeted methods can be used to build non-targeted models, or in contrast, targeted methods can be developed to identify and quantify analytes based upon the findings of non-targeted methods.

4.2 INSTRUMENTATION AND DATA PREPROCESSING

The selection of modulator, detector, and experimental design will result in a data structure for a single sample run to be either second-order or third-order, where the respective data structure can be described as a matrix or cube (Figure 4.3). In general, use of a univariate detector in the instrumental design will produce second-order data, describing the separation time on both dimensions. Likewise, coupling a multivariate detector to a comprehensive 2D separation produces third-order data, where the three axes describe the separation time on both dimensions and the

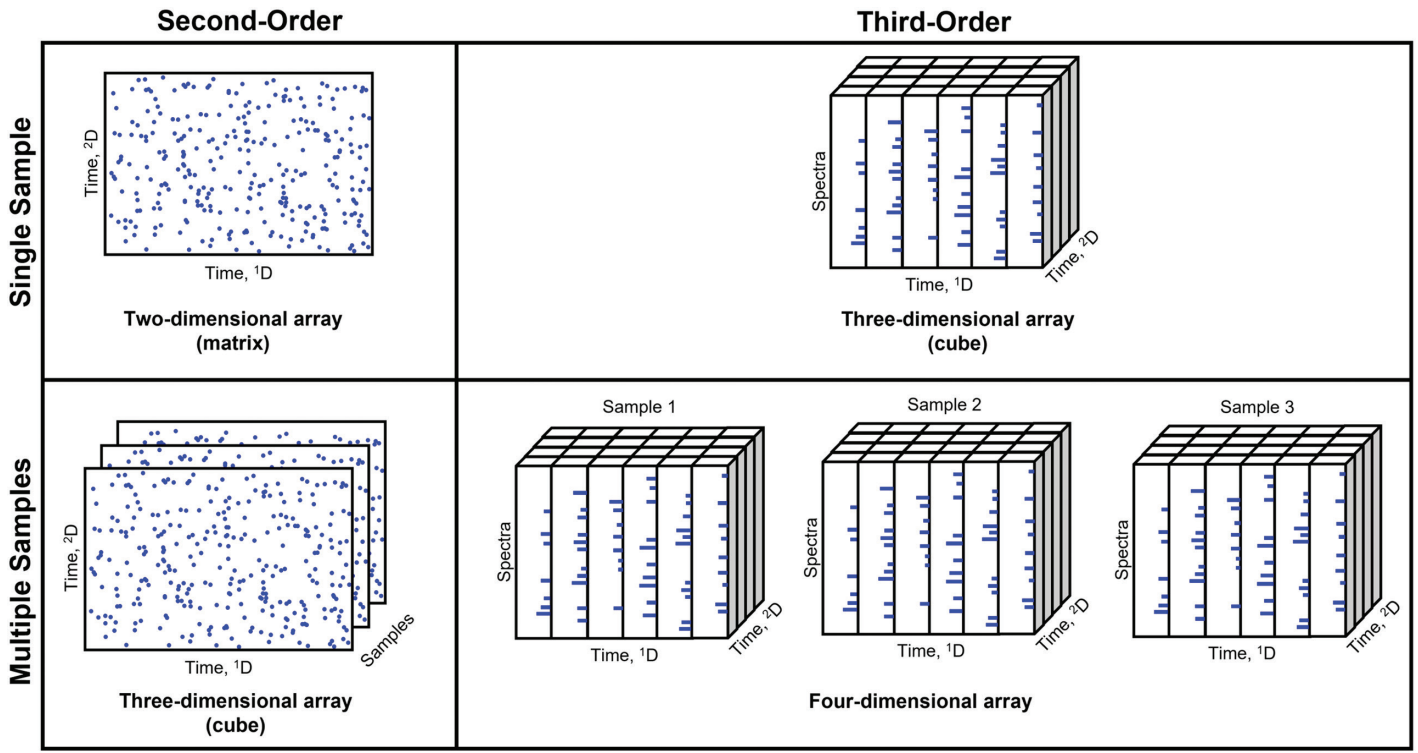


FIGURE 4.3 Schematic of the different dimensionalities for comprehensive 2D chromatographic data. Second-order and third-order data are produced when comprehensive 2D chromatograms are collected with univariate and multivariate detectors, respectively. The dimensionality of the data can be increased by analyzing multiple samples simultaneously.

spectra recorded. Higher-order data can be achieved either by analyzing multiple samples together or using multiple detectors (Figure 4.3). The order and quality of the data produced from a comprehensive 2D separation is the most decisive step in selecting the appropriate chemometric method(s) for later analysis. However, while the instrumental design can produce higher-order data, its operation may not allow for the analyst to fully utilize second-order and third-order chemometric advantages. To realize these advantages, the data must be either bilinear or trilinear, meaning that each data dimension (two for second-order data or three for third-order data) is linearly independent and analytes have sufficiently reproducible peak shapes, retention times, and concentration-dependent signals. Thus, prior to defining a chemometric objective (Figure 4.2), the first goal for the analyst is to produce high-quality bilinear or trilinear data. This section will discuss how modulators, detectors, and preprocessing methods can affect and improve the quality of the data.

4.2.1 GC × GC Instrumentation

Modulators for GC × GC can be broadly classified into two groups: thermal modulators and flow modulators, also known as valve-based or pneumatic modulators. Thermal devices change the temperature on a section of column or a dedicated trapping capillary to modulate analyte introduction to the ^{2}D column. Since the introduction of the first GC × GC instrument [4], thermal modulation has enjoyed significant development and commercialization, leading to reliable and effective designs such as the jet-cooled cryogen modulator designed by Ledford [11]. Although a powerful and effective approach, thermal modulators, especially those that use cryogens, have high capital and operating costs. Conversely, flow modulators use one or more valves to divert, collect, or temporarily stop the ^{1}D gas flow and thus effect modulation. Flow modulation was introduced in 1998 [12] as a simpler, more cost-effective alternative to thermal modulation. Following extensive research and development [13–16], flow modulation has been demonstrated as a reliable and effective alternative to thermal modulation, and many designs are now commercially available. The disadvantages of flow modulation relative to thermal modulation include more challenging method optimization, the generation of high flow rates on the ^{2}D, and in some cases, only partial transfer of the ^{1}D material onto the ^{2}D (duty cycle < 100%). For further discussion on the operation of thermal and flow modulators, the reader is directed to a recent review [17].

While thermal or flow modulation can both be used to generate GC × GC separations, the specific design and operation of a given modulator can affect the degree to which the data follows a bilinear (or trilinear) structure. For each P_M, thermal modulators focus an analyte in the ^{1}D effluent into a narrow pulse that, when injected onto the ^{2}D column, generally results in detected peak profiles that are more likely to fit a bilinear model. In contrast, flow modulators sample slices of the ^{1}D peaks, and due to the absence of a thermal focusing stage, the concentration pulse delivered to the ^{2}D column can maintain to some extent the shape of the fraction of the original ^{1}D peak profile [18]. As a result, the ^{2}D peak maximum may appear to shift between modulations, which deviates from the bilinear model. The deleterious effect of non-focusing modulation can be largely mitigated by sampling each ^{1}D peak many times or by

employing chemometric methods that do not have a strict bilinearity requirement for the ^{1}D or ^{2}D. Another deviation from bilinearity (or trilinearity) occurs when the ^{2}D retention time of a given analyte decreases between successive modulations because of the slight increase in oven temperature between modulations [19]. This effect is most pronounced for peaks with narrow ^{2}D widths and is mitigated by employing fast modulation periods (P_M < ~3 s) with short ^{2}D columns with thin stationary phase films, which ensures that the change in oven temperature between modulations is minimized and that the ^{2}D retention factors of analytes are kept relatively low [19,20].

As a result of the fast ^{2}D column separations, the ^{2}D peak widths for GC × GC range from tens to hundreds of milliseconds. Therefore, suitable detectors for GC × GC are limited to those that can both quickly respond to the changing input chemical concentration(s) and then rapidly convert the chemical concentration into an electronic signal measurement. The former is the primary constraint, which requires small detector volumes for ionization-based detectors or fast mass-to-charge (*m/z*) acquisition speeds, for mass spectrometry (MS) detection. Flame ionization detection (FID) is perhaps the most used detector in 1D-GC and, given its fast ionization mechanism and small internal volumes, is an excellent option for univariate GC × GC detection. However, with FID, analyte identification is fully reliant on matching retention times of sample peaks to a known standard. When high selectivity and component identification are desired, MS detection is indispensable. Time-of-flight mass spectrometry (TOFMS) is the ideal choice for GC × GC since entire mass spectra can be acquired within microseconds [21]. Both high resolution (HR-TOFMS) and different MS ionization methodologies have also been incorporated into GC × GC instruments for increased selectivity and identification capabilities [22,23]. Fast-scanning quadrupole analyzers are occasionally used for GC × GC, due to their low cost relative to TOFMS [24]. However, unlike TOFMS, quadrupole analyzers do not scan all *m/z* simultaneously. As a result, spectral skew occurs, wherein later *m/z* are measured at a different analyte concentration than earlier *m/z*. Therefore, *m/z* which belong to the same analyte reach their respective maxima at different times. This effect decreases the bilinear or trilinear quality of GC × GC-MS data and can incur error in chemometric methods unless the skew is corrected [25].

4.2.2 LC × LC Instrumentation

Modulation techniques for LC × LC can be classified as either passive or active, depending on how the ^{1}D effluent is modified before injection onto the ^{2}D column. LC × LC separations have conventionally utilized passive modulation, where the ^{1}D sample fraction was transferred without any change in volume or concentration [26]. The interface for passive modulation consists of either an 8- or 10-port valve with two identical storage loops that collect the ^{1}D effluent in one loop as the contents of the second loop are injected onto the ^{2}D column. Both of the original LC × LC instruments utilized passive modulation [2,3]. However, peak deformation and/or splitting can occur with passive modulation due to the incompatibility of the mobile phases in both dimensions (i.e., the ^{1}D effluent is the strong solvent for the ^{2}D) [27]. To overcome this solvent mismatch issue, recent work has focused on developing active modulation techniques for LC × LC separations.

Active modulation methods adjust the matrix of the ^{1}D sample fraction to prevent solvent mismatch with the ^{2}D separation [26]. One popular active modulation method is stationary phase assisted modulation (SPAM), which uses low-volume trap columns connected to the valve instead of the storage loops used with passive modulation [28]. The use of trap columns retains analytes from the ^{1}D fraction as the mobile phase passes through to the waste. Once the valve switches, the gradient program for the ^{2}D separation elutes the analytes from the trap columns as focused, concentrated bands. Another modulation approach aimed at resolving solvent incompatibilities is known as active solvent modulation (ASM) [29]. Like passive modulation, ASM utilizes an 8-port/2-position valve with storage loops to collect ^{1}D fractions. However, the ^{1}D effluent becomes diluted with the ^{2}D mobile phase prior to the ^{2}D separation. Both SPAM and ASM have been shown to improve detection sensitivity for LC × LC and allow for the ^{2}D separation to occur on narrower columns, allowing for fast separations in order to overcome undersampling [26,28,29]. Further descriptions of SPAM, ASM, and other active modulation techniques can be found in the literature [26,28,29].

Since new modulator and column technology has decreased the run time for the ^{2}D separations [26,30], detectors for LC × LC instruments must also have fast acquisition speeds to adequately sample and record the resulting narrow peaks. Primarily, ultraviolet-visible (UV) detectors and mass spectrometers are used to collect the analytical signal from LC × LC separations given their fast acquisition speeds. UV detectors, which have acquisition rates of up to 100 scans/s, can either measure analyte absorbance at a single wavelength or over a range of wavelengths. The latter, known as a diode array detector (DAD), provides ample UV spectra across the width of each chromatographic peak to aid in the analyte identification effort. These detectors provide both low detection sensitivity and high reproducibility while being relatively inexpensive compared to MS [30,31]. However, peak identification using LC × LC-DAD data is more reliant upon the chromatographic separation since compounds with similar functionalities can have similar UV spectra. On the other hand, quadrupole-time-of-flight MS (QTOF-MS) and tandem MS (MS/MS or MS^2) with electrospray ionization (ESI) have been incorporated into LC × LC instruments because of their high sensitivity, selectivity, and ability to obtain a pure *m/z* for analyte identification even at low chromatographic resolution [32–34]. However, the main drawbacks of these MS detectors are the high costs associated with instrumentation, upkeep, and signal suppression effects [35].

4.2.3 Data Preprocessing

Prior to analyzing GC × GC or LC × LC data, some degree of data preprocessing is generally required to remove chemically irrelevant variations in the signal to improve chemometric performance. Baseline correction, smoothing, normalization, and retention time alignment methods are commonly used for data preprocessing. Low-frequency detector noise (i.e., baseline drift) can be removed using baseline correction methods, which commonly subtract a fitted curve from the entire chromatogram or sections of the chromatogram. Smoothing methods, such as a Savitzky-Golay filter, are used to reduce high-frequency noise and increase the *S/N*. These

two preprocessing methods must be carefully applied to prevent the loss of chromatographic signal and introduction of new artifacts, which can negatively impact chemometric performance. When comparing multiple replicates and/or samples, normalization and retention time alignment methods must be applied to reduce the inevitable variation from sample preparation and instrument operation. The use of internal standards or total area normalization, where the sum of the baseline corrected signal acts as the normalization factor, are the most common normalization methods. Retention time alignment programs ensure that variables that correlate to the same peak are correctly compared and that the bilinear (or trilinear) nature of the data is preserved. A variety of retention time alignment programs have been developed for comprehensive 2D separations, such as piecewise alignment [36], 2D correlation optimized warping [37], and dynamic time warping [38]. For data collected with a multivariate detector, algorithms can also use the collected spectra to improve the retention time alignment results [39,40]. Another method to resolve misaligned 2D chromatographic data is to average (i.e., bin) the data along both separation axes. Here, the appropriate bin size should be large enough to encompass the peak widths on both dimensions as well as the observed shifting [41]. As a result, proper binning increases the *S/N* while reducing the overall size of the data, which improves computational speed and performance. However, if the chromatogram is not appropriately binned, then a loss in chromatographic resolution can be observed.

For data collected with a high-resolution MS (i.e., HR-TOFMS for GC × GC and QTOF-MS for LC × LC), additional preprocessing steps are necessary prior to data analysis. Since these detectors can have a resolution of 0.0001 amu, only mass channels (*m/z*) with measurable intensities are recorded, leading to unequally spaced masses for each scan. This data richness also leads to file sizes per each sample separation that can be minimally 500 MB but often much larger, which can be unmanageable for computers with limited memory. Two data compression approaches have been developed to convert raw data into a usable data format containing an equidistant *m/z* dimension. The first method is to bin the *m/z* dimension to a lower resolution, often down to unit resolution, making the analysis of the entire chromatogram more computationally manageable [42,43]. Based on these results, the analyst can then further interrogate selected retention time or *m/z* regions using the higher resolution data, keeping a light computational load [42,43]. The bin size must be chosen to ensure that smaller peaks are not overlapped by larger coeluting peaks in the *m/z* dimension, causing the lower intensity peaks to disappear. The second strategy searches for "regions of interest" (ROIs) in the chromatogram based on three parameters: *S/N* threshold, minimum number of successive data points, and allowable mass deviation [33,44–46]. This approach ensures that the new data file only contains the discovered ROIs, producing a chromatogram where each scan has the same measured *m/z* and the high resolution of that dimension has been preserved [33,44–46].

4.3 TARGETED ANALYSIS

The primary goal of any comprehensive 2D separation is to identify and quantify analytes responsible for the similarities and differences in a data set. In targeted studies, the identity of these analytes of interest is known beforehand, and the comprehensive 2D separation is designed to chromatographically resolve each targeted

compound to the greatest extent possible. After the data are collected, the identity of the target analytes is confirmed via spectrum library matching and/or retention time indexing with analyte standards. Then, the peak heights/areas of each modulated ^{2}D peak are summed together and analyte concentration is determined using the standard addition method, external standards, or internal standards [47,48]. However, the experimental design cannot always be optimized for all the target compounds of interest to be fully resolved, causing overlapped interferent signals to challenge identification and quantitation. Therefore, chemometric decomposition methods can be used to obtain pure chromatographic peak profiles and spectra. It is important to note that chemometric decomposition has also been referred to as deconvolution in the literature. This section will focus on two popular decomposition methods: multivariate curve resolution-alternating least squares (MCR-ALS) and parallel factor analysis (PARAFAC). The operation of both methods is similar even though they have different data structure requirements. For example, both methods are traditionally applied to relatively small 2D regions of the chromatogram instead of the entire chromatogram to lighten their computational load. Both MCR-ALS and PARAFAC also require the analyst to provide an estimate of the number of mixture components separated in the selected time window (i.e., the rank of the data). Generally, the number of components is taken as the number of analytes present plus additional component(s) for the baseline/background noise. This estimate of the rank will remove background and noise from the pure component profiles without the need for baseline correction steps. These methods then leverage information in each data dimension to mathematically resolve target and interferent signals. To develop an accurate decomposition model, the experimental design must ensure that the chromatograms adhere to either a bilinear or trilinear data structure.

4.3.1 Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS)

MCR-ALS is a bilinear decomposition method that extracts the pure component information for each dimension of second-order data [49–51]. Given the bilinearity requirement, MCR-ALS can be applied to comprehensive 2D separation data collected with either univariate or multivariate detection. Comprehensive 2D chromatograms collected with univariate detectors (e.g., FID or UV at a single wavelength) are naturally second-order. However, in this case, the chromatographic data from univariate detectors must be aligned to minimize retention time shifting to ensure data bilinearity. To apply MCR-ALS to comprehensive 2D separations collected with multivariate detection (e.g., DAD or MS), the dimensionality of the data must be reduced prior to MCR-ALS. This data reduction can be achieved by analyzing individual modulations (i.e., the ^{2}D separation) or unfolding the time dimension (i.e., concatenating each ^{2}D separation together) while maintaining the spectra dimension. A benefit of applying MCR-ALS to chromatograms collected with multivariate detection is that alignment is not necessary because data bilinearity is supported by the reproducibility of the spectra dimension [52]. The MCR-ALS model can also be extended to simultaneously analyze multiple samples or replicates. For these higher-ordered arrays, the time dimension for each sample would be unfolded and then those samples would be augmented together along the time axis (Figure 4.4).

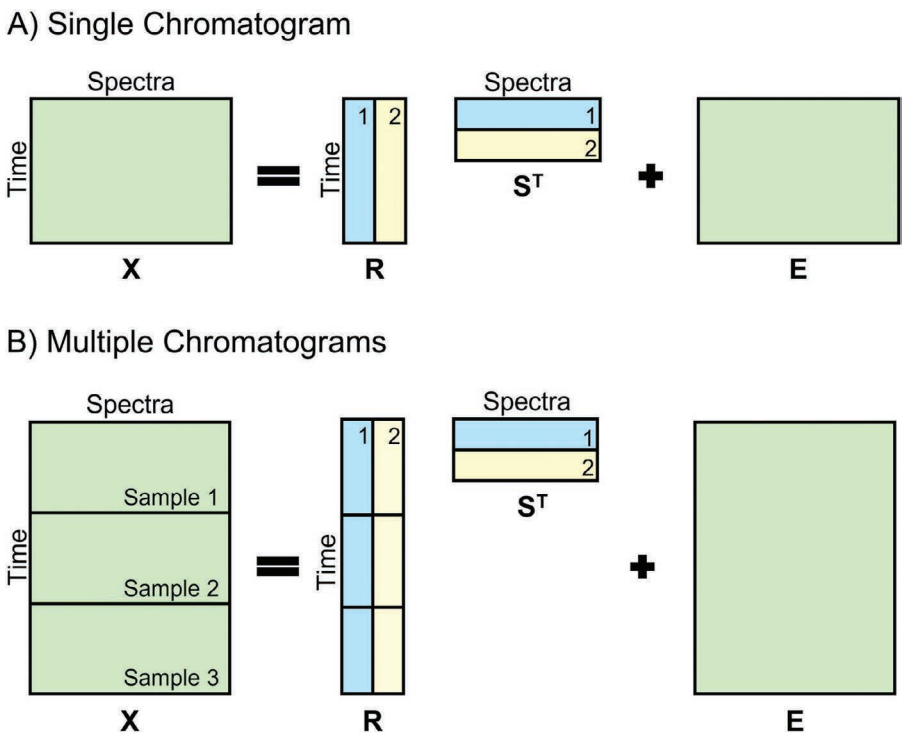


FIGURE 4.4 Illustration of a two-component MCR-ALS model for comprehensive 2D chromatographic data collected with multichannel detection. Components 1 and 2 are highlighted in blue and yellow, respectively. MCR-ALS models can be constructed for single chromatograms (A) or chromatographic data sets with multiple samples (B). To fit the bilinear model, the time dimension of the chromatographic data was unfolded.

The MCR-ALS model is represented as

$$\mathbf{X} = \mathbf{R}\mathbf{S}^{\mathbf{T}} + \mathbf{E} \tag{4.1}$$

where **X** is the chromatographic data matrix, **R** and **S** are matrices containing the pure instrumental responses, and **E** is a matrix containing the residual errors [49–51]. In the context of a comprehensive 2D separation, the matrix **R** generally represents the resolved chromatographic elution profiles (i.e., time dimension) for each modeled component. Likewise, the matrix **S** normally contains the pure spectra for each component in the model. However, for data collected with univariate detection, the final matrices for **R** and **S** will contain the pure ^{1}D and ^{2}D chromatographic profiles for each component modeled. Since MCR-ALS is an iterative method, the algorithm will alternate between the results in **R** and **S** to minimize the errors in **E**. Figure 4.4 illustrates the format of a two-component MCR-ALS model to analyze either a single chromatogram (A) or multiple chromatograms

simultaneously (B). The ease of obtaining pure information from the experimental data is dependent on the number of estimated components in the subsection of the chromatogram along with the *S/N* of the target analyte, its relative intensity, and extent of overlap with interferents. Using the model outputs, the pure elution profile and spectrum for each component can be used for quantitation and identification, respectively.

While MCR-ALS is a flexible decomposition model for various types of comprehensive 2D chromatographic data, it is possible that different solutions can be produced for the same matrix input and those solutions can fit the data equally well. This uncertainty is referred to as "rotational ambiguity," and the extent of this uncertainty can be evaluated by finding all possible, feasible solutions [53,54]. Proper initialization and selection of constraints can reduce the number of possible solutions, improving the fit of the MCR-ALS model. MCR-ALS initialization refers to providing the model of an initial estimate of a data dimension for each component. Typically, for chromatographic applications, initial estimates are provided for the spectrum of each modeled component. These initial estimates can either come from prior knowledge (e.g., the pure spectrum for the target analyte) or from algorithms designed to select the most dissimilar spectra in the original data. The most common initialization methods include simple-to-use self-modeling analysis (SIMPLISMA) [55], orthogonal projection approach (OPA) [56], and key set factor analysis (KSFA) [57]. Both OPA and KSFA can be performed in an iterative manner for further refinement of the spectra to be used as initial estimates [58–60]. Additionally, constraints place mathematical conditions on the fit of **R** and **S** during the iterative optimization of the MCR-ALS model. The most commonly applied constraints for chromatographic data are non-negativity, ensuring that the elution profiles have non-negative concentrations, and unimodality, ensuring only one peak maximum per component. Defining regions in the chromatographic data with an absence of analytes (i.e., local rank constraints), concatenating replicates prior to decomposition, or using hard modeling can also be implemented to mitigate rotational ambiguities [54]. Additionally, application of a trilinearity constraint can be used to obtain essentially the same unique solution as higher-order decomposition models like PARAFAC (discussed in the next section) [54].

Decomposition methods such as MCR-ALS can readily improve the resolving power of a comprehensive 2D chromatogram. For example, Bailey et al. applied MCR-ALS to a subsection of a LC × LC-DAD separation of a human urine sample [61]. Figure 4.5A shows that this region of the chromatogram contains eight overlapped peaks (labeled), with some peaks having the same ^{1}D and ^{2}D retention times. Using an eight-component MCR-ALS model, the pure chromatographic profiles and spectra were obtained (Figure 4.5B). Visual inspection of the first, third, and fifth components were all found to contain pure peak profiles while the background was modeled by the second and fourth components (Figure 4.5B). Interestingly, the sixth through eighth components contained contributions from multiple peaks (Figure 4.5B), hindering quantitation efforts [61]. These components were unable to be decomposed into their pure profiles due to larger interferent signals drowning out lower intensity peaks (i.e., dynamic range issues) and the overlapped peaks having similar spectra (i.e., rank deficiencies) [61].

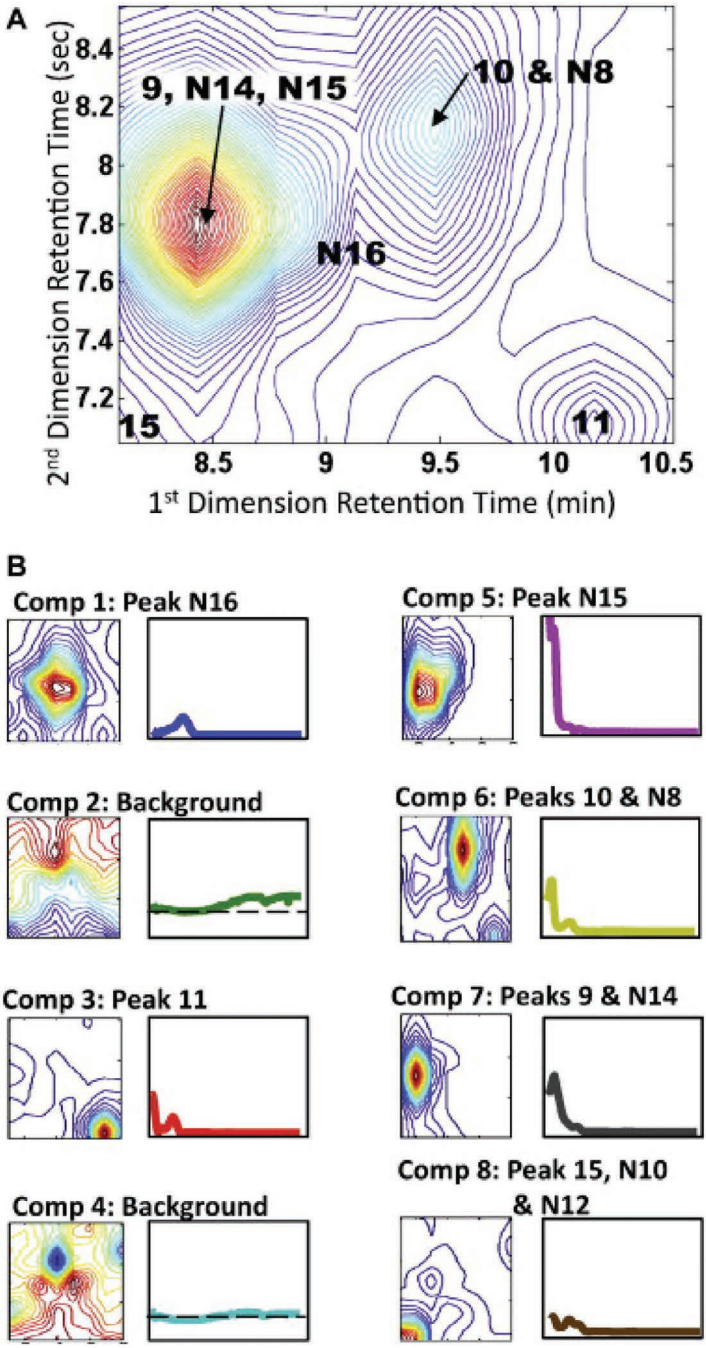


FIGURE 4.5 MCR-ALS decomposition performed on a low chromatographic resolution region from a LC × LC-DAD separation of human urine. (A) A 2D contour plot of this section at 216 nm with the peaks of interest labeled. (B) MCR-ALS resolved chromatographic profile

FIGURE 4.5 (Continued)
and spectrum for components in the model. The chromatographic axes in (B) are the same as the axes in (A). The wavelength range shown in (B) is 200–700 nm.

This figure was taken with permission from H.P. Bailey, S.C. Rutan, P.W. Carr, Factors that affect quantification of diode array data in comprehensive two-dimensional liquid chromatography using chemometric data analysis, J. Chromatogr. A 1218 (2011) 8411–8422. https://doi.org/10.1016/j.chroma.2011.09.057.

While dynamic range issues will always persist in real, complex samples, rank deficiencies can be resolved with the use of a detector with complementary selectivity. Chemometric analysis of both LC × LC-MS and LC × LC-DAD data sets found that a higher number of components could reliably be discovered with MS given its ability to provide a more selective response for each analyte [62]. For illustration, Navarro-Reig et al. showed that MCR-ALS was able to resolve and aid in the identification of different triacylglycerol (TAG) isomers in a LC × LC-MS separation of corn oil [63]. Note, since TAGs are abbreviated according to the three fatty acids bonded to glycerol, this example will discuss TAGs composed of stearic acid (S), oleic acid (O), linoleic acid (L), and palmitic acid (P). Figure 4.6A shows the total ion current (TIC) chromatogram for a region of the separation, where two pairs of positional isomers (SLO/SOL and PLO/POL) are overlapped with one another. Despite four analytes being separated in this region, only three peaks can be seen (Figure 4.6A). Non-chemometric based (traditional) identification efforts for these peaks would be hindered by the low chromatographic resolution and similarity in their mass spectra. Therefore, MCR-ALS was used to produce the pure versions of their unfolded peak profiles and spectra in Figure 4.6B-C. The resolved mass spectra (Figure 4.6C) were then compared to a set of reference spectra (Figure 4.6D) to confidently identify the different isomers. Given the high-quality spectra produced, peaks 1 and 2 were identified as SOL (black) and SOL (blue), respectively, while peak 3 was determined to be the coelution of POL (red) and PLO (green) [63].

4.3.2 Parallel Factor Analysis (PARAFAC)

PARAFAC is a trilinear decomposition method that can extract the pure instrumentally obtained responses from third- or higher-ordered data sets [64]. In terms of comprehensive 2D separations, PARAFAC is commonly performed on chromatograms collected with multivariate detectors since the data structure is naturally third-order (^{1}D time × ^{2}D time × spectrum). For data collected using univariate detectors, multiple samples are required to create a third-order data structure (^{1}D time × ^{2}D time × samples). Compared to MCR-ALS, PARAFAC is advantageous for decomposition of comprehensive 2D separations since it does not require a reduction in data dimensionality and the final solution in the model is unique. However, data submitted for PARAFAC modeling must be sufficiently trilinear whereas MCR-ALS only requires bilinear data. The strict trilinear structure condition requires all the ^{2}D peaks for a single ^{1}D peak to be reproducible in terms of peak shape, width, and retention time (or at least not deviate greatly) and chemically selective information must be present

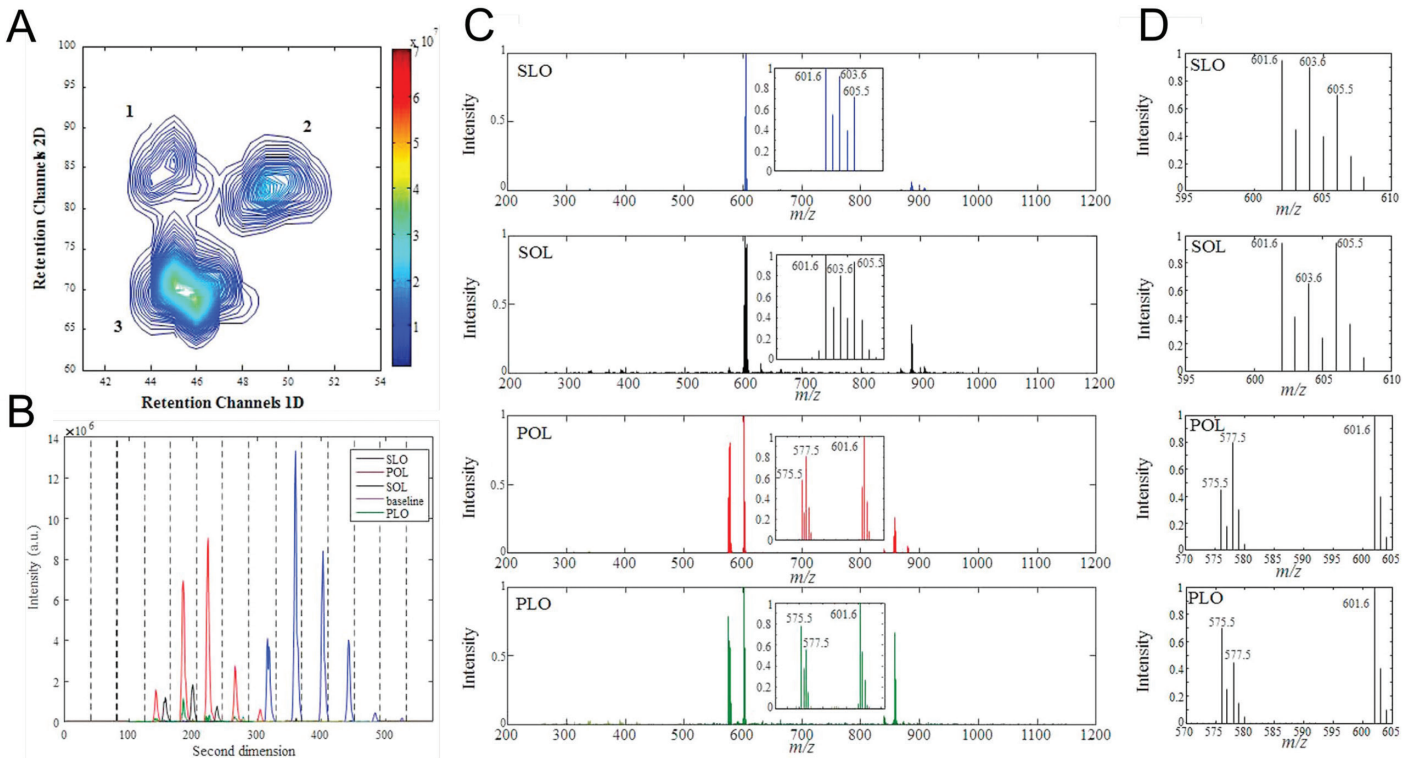


FIGURE 4.6 MCR-ALS decomposition of four triacylglycerols from a corn oil sample that was separated using LC × LC-MS. (A) A 2D contour plot of the region of interest. (B) MCR-ALS resolved chromatographic profiles of the four triacylglycerols: SLO (blue), POL (red), SOL (black), and PLO (green). (C) MCR-ALS resolved mass spectra for the four triacylglycerols. Inserts: Zoom-in from 565–610 *m/z* to show the key ion fragments necessary for identification. (D) Reference mass spectra for each analyte, which was used in identification.

This figure was taken with permission from M. Navarro-Reig, J. Jaumot, T.A. van Beek, G. Vivó-Truyols, R. Tauler, Chemometric analysis of comprehensive LC × LC-MS data: Resolution of triacylglycerol structural isomers in corn oil, Talanta 160 (2016) 624–635. https://doi.org/10.1016/j.talanta.2016.08.005.

in at least two of the data dimensions [19]. Given a chromatographic data cube **X**, consisting of elements x_{ijk} and F number of components (i.e., the rank of the data), the PARAFAC model can be expressed as

$$x_{ijk} = \sum_{f=1}^{F} a_{if} b_{jf} c_{kf} + e_{ijk} \tag{4.2}$$

where a_{if}, b_{jf}, and c_{kf} are the elements of matrices **A**, **B**, and **C** containing the pure instrumental responses for each component and e_{ijk} are the elements of a three-way array, **E**, representing the residual error [64]. For comprehensive 2D data collected with multichannel detection, the columns of the matrices **A**, **B**, and **C** (the loadings) correspond to the chromatographic profiles in both dimensions (^{1}D and ^{2}D) and the spectrum for each component modeled, respectively. Figure 4.7 demonstrates the construction of a two-component PARAFAC model to analyze a single comprehensive 2D chromatogram collected with multichannel detection. Likewise, the PARAFAC loadings for multiple replicates collected with univariate detection will correspond to the ^{1}D, ^{2}D, and sample dimensions. The model is achieved by using initial estimates for two dimensions and then applying alternating least squares to fit the remaining mode to obtain a solution where the residuals, e_{ijk}, are minimized.

Selection of the number of components to model, initialization and stoppage values, and constraints are all necessary to execute PARAFAC modeling. Like MCR-ALS, the number of components to model should equal the number of suspected analytes plus one or more for the background contributions. If too many components are used in the PARAFAC model, then the computational speed decreases and true analyte signals can be modeled by multiple components (i.e., splitting). Hoggard and Synovec defined the appropriate number of components to model as one fewer than the PARAFAC model with the observed splitting [65]. This method successfully created PARAFAC models for target analytes across a wide range of signal intensities (overloaded to low *S/N*) and could be applied in an automated fashion [65]. Split-half experiments can also help determine the correct number of components to model [66]. Here, the chromatographic region is divided into two sections and PARAFAC models are created for both sections. If the number of components is selected correctly, the same loadings should be evident in the models of both data

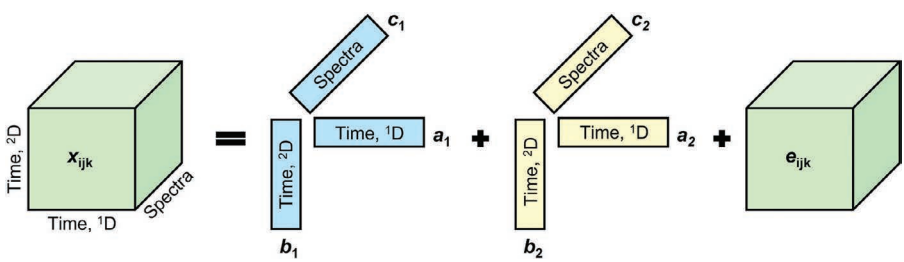


FIGURE 4.7 Illustration of a two-component PARAFAC model for comprehensive 2D chromatographic data collected with multichannel detection. Components 1 and 2 are highlighted in blue and yellow, respectively.

sets. Typically, random values or initial estimates from trilinear decomposition are used for initialization while the minimum of the residuals array (**E**) is a stoppage criterion. Unimodality and orthogonality constraints can help stabilize the solution while non-negativity ensures the loading vector should have positive signal [64].

Using the loadings matrices from a PARAFAC model, the traditional targeted analysis workflow of analyte identification and quantitation follows. For visualization, Sinha et al. developed a PARAFAC model to resolve the analytical signal of trimethylsilylated (TMS) vanillic acid from the matrix components of human urine in a GC × GC-TOFMS chromatogram [67]. Visual inspection of the region surrounding vanillic acid (TMS) reveals that this analyte is overlapped with at least four interfering analytes (Figure 4.8A). The resulting PARAFAC model decomposed the chromatographic region into the pure ^{1}D peak profiles (Figure 4.8B), ^{2}D peak profiles (Figure 4.8C), and mass

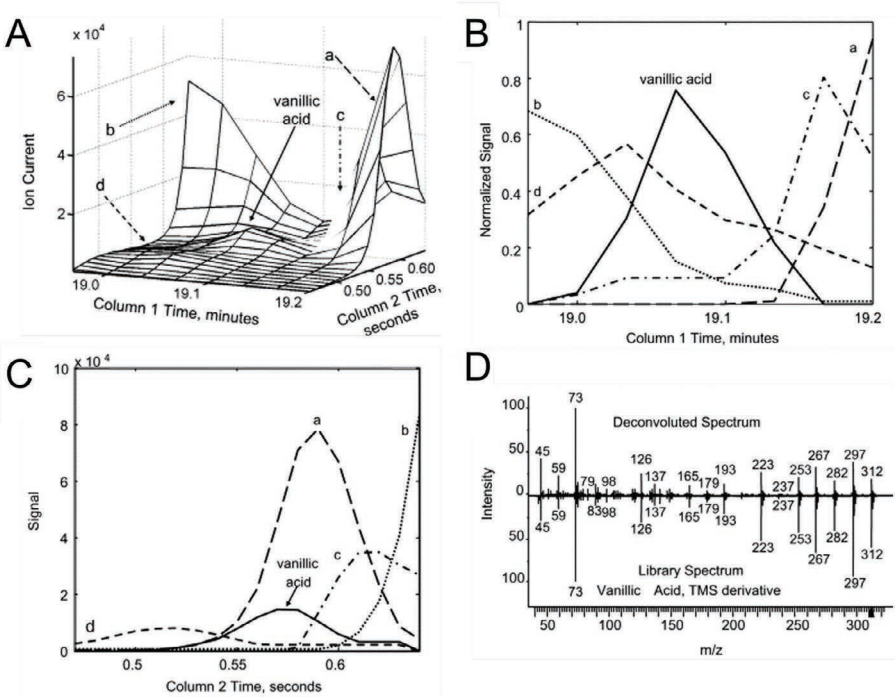


FIGURE 4.8 PARAFAC decomposition of TMS-derivatized vanillic acid in human urine. (A) GC × GC-TOFMS chromatogram at *m/z* 73, highlighting the low chromatographic resolution between vanillic acid (TMS) and four interfering components (labeled a–d). (B) PARAFAC resolved ^{1}D chromatographic profiles for this region. (C) PARAFAC resolved ^{2}D chromatographic profiles for this region. (D) PARAFAC resolved mass spectrum for vanillic acid (TMS) compared to its library reference spectrum.

This figure was taken with permission from A.E. Sinha, J.L. Hope, B.J. Prazen, E.J. Nilsson, R.M. Jack, R.E. Synovec, Algorithm for locating analytes of interest based on mass spectral similarity in GC × GC–TOF-MS data: analysis of metabolites in human infant urine, J. Chromatogr. A 1058 (2004) 209–215. https://doi.org/10.1016/j.chroma.2004.08.064.

spectrum (Figure 4.8D) for TMS derivatized vanillic acid. The resolved mass spectrum for the vanillic acid (TMS) component was then compared to the library spectrum to confirm its identification (Figure 4.8D). For quantitation, the peak profile for vanillic acid (TMS) or any other component could be reproduced by the outer product of the vectors corresponding to the component of each loading matrix.

While the previous example considered the targeted analysis of a single three-way chromatogram ($^1D \times {}^2D \times m/z$), PARAFAC modeling can also be applied for the analysis of four-way data sets if the data is quadrilinear. In these cases, the data dimensions are the ^1D and ^2D chromatographic peak profiles, the signal collected from a multichannel detector, and the sample dimensions. The four loading matrices produced from the PARAFAC model will also represent each dimension. For example, Porter et al. applied PARAFAC to a four-way data set of metabolites in mutant and wild-type maize [31]. Figure 4.9A shows an overlay of the LC × LC-DAD chromatograms at 220 nm for representative mutant (blue), wild-type (red), and indole standards (green). Given the complexity of the data analyzed, the chromatograms were divided into the sections outlined by the black boxes. For the purposes of this discussion, the four-way PARAFAC model is demonstrated on the section indicated by the arrow in Figure 4.9A. PARAFAC successfully decomposed this section into the pure ^1D peak profiles (Figure 4.9B), ^2D peak profiles (Figure 4.9C), spectra (Figure 4.9D), and concentrations measured for component in each sample (Figure 4.9E). By looking at the concentration profiles in Figure 4.9E, similarities and differences between the mutant (M1 and M2) and wild-type (Wt1 and Wt2) maize samples can be made. For instance, multiple components were found in higher abundance in the mutant samples instead of the wild-type samples (Figure 4.9E) [31]. Similarly, two compounds present in the standard (St) chromatogram (5-hydroxytryptamine in orange and indole-3-acetyl-l-lysine in pink) were also found in low abundance in the mutant samples (Figure 4.9E) [31].

Along with identification and quantitation, PARAFAC can also be used to evaluate the trilinearity (or higher) of the chromatographic data structure [68–70]. The magnitude of retention time shifting between modulations to ^2D peak width, referred to as the trilinearity deviation ratio (TDR), can predict the accuracy of PARAFAC models for quantitation [19,20]. Application of PARAFAC to non-trilinear data was shown to cause a negative bias, where true analytical signal that does not fit the model has been removed [19,20]. Furthermore, the trilinear nature can be assessed by comparing the loadings from PARAFAC to the experimental data by calculating two metrics, the lack-of-fit (*LOF*) and percent of explained variance (R^2) [68–70]. Ideally, if the chromatographic data is trilinear, then the measured *LOF* and R^2 will be 0% and 100%, respectively. As stated earlier, the experimental conditions have the greatest influence on the trilinear (or bilinear) nature of the chromatographic data. For example, Prebihalo et al. demonstrated that GC × GC-TOFMS data is sufficiently trilinear (i.e., small TDRs and PARAFAC quantitation errors) with a small P_M (~ 1–2 s) compared to relatively longer P_M (~ 5–8 s), which are typically used [20]. In cases where the experimental design was not optimized to ensure a trilinear data structure, retention time alignment algorithms should be used to make the data more amenable to PARAFAC [40,68,71]. For instance, Allen and Rutan demonstrated that alignment improved the accuracy and reproducibility of quantitative PARAFAC

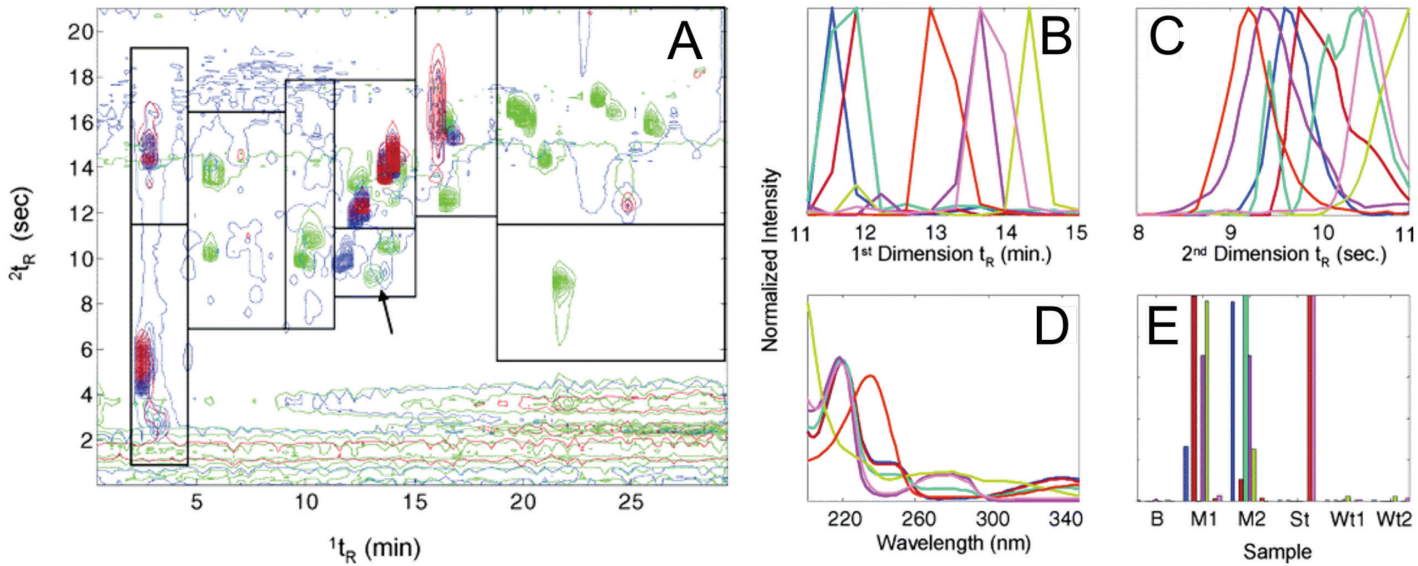


FIGURE 4.9 PARAFAC decomposition of metabolites in maize samples collected with LC × LC-DAD. (A) Overlaid chromatograms of representative mutant (blue), wild-type (red), and standard (green) samples at 220 nm. PARAFAC results for the black box indicated by an arrow will be shown. (B) PARAFAC resolved ^{1}D chromatographic profiles. (C) PARAFAC resolved ^{2}D chromatographic profiles. (D) PARAFAC resolved spectrum for each component. (E) PARAFAC resolved concentration profiles for each component in the background (B), mutant (M1 and M2), standard, and wild-type (Wt1 and Wt2) samples.

This figure was taken with permission from S.E.G. Porter, D.R. Stoll, S.C. Rutan, P.W. Carr, J.D. Cohen, Analysis of four-way two-dimensional liquid chromatography-diode array data: Application to metabolomics, Anal. Chem. 78 (2006) 5559–5569. https://doi.org/10.1021/ac0606195.

models for phenytoin in wastewater samples [40]. PARAFAC2 can also be used to analyze three-way chromatographic data that do not follow a trilinear nature. As a modified version of PARAFAC, PARAFAC2 is less sensitive to misalignment, the main deviation from trilinearity, while still producing unique solutions for all three data dimensions [72].

4.4 UNSUPERVISED, NON-TARGETED ANALYSIS

While targeted chemometric methods are beneficial in the identification and quantitation of previously known and anticipated analytes of interest, non-targeted approaches seek to discover relevant chemical features that describe the similarities and/or differences across multiple chromatograms. Non-targeted chemometric methods can be described as either supervised or unsupervised, where supervised methods depend upon *a priori* knowledge of sample classification. Supervised approaches (discussed later) are appropriate for handling classification and regression problems since these methods leverage class labels. In contrast, unsupervised models do not require knowledge of class memberships. Therefore, unsupervised approaches are suitable for exploratory data analysis, where the user aims to discover patterns and detect outliers in the data set. These unsupervised, non-targeted methods are typically the first step in a chemometric workflow because they are simple, computationally inexpensive, and provide visualization of the main attributes of the data. This section will cover three common unsupervised techniques for chromatographic data analysis: principal components analysis (PCA), hierarchical cluster analysis (HCA), and *k*-means clustering. PCA allows for the identification of features that accurately represent relationships between samples, whereas HCA and *k*-means clustering allows for the discovery of inherent groupings in unlabeled data.

4.4.1 Principal Components Analysis (PCA)

PCA is possibly the most applied exploratory data analysis technique because it reduces the chromatographic data down to only the variables that represent the variation and correlations in the data set. This data reduction is achieved by projecting the possibly correlated variables (i.e., peaks) in the data onto a new set of linearly uncorrelated variables called principal components (PCs) [73]. After the orthogonal transformation, these PCs are then ranked in descending order of explained variance. Therefore, PC 1 explains the maximum variability in the data, PC 2 explains the maximum variance not explained by PC 1, and so on. This process continues until all the variance in the data set has been explained or until an algorithmic stopping point has been reached [74]. In practice, only the first couple of PCs that explain a proportion of the total variance in the model will be kept and utilized for interpretation.

The output of PCA is a decomposition model, which can be described as

$$\mathbf{X} = \mathbf{TP} + \mathbf{E} \tag{4.3}$$

where $\mathbf{X}$ is the original chromatographic data matrix, $\mathbf{T}$ is the scores matrix, $\mathbf{P}$ is the loadings matrix, and $\mathbf{E}$ represents the unaccounted signal that remains. The data

matrix (**X**) must be a two-way array, where the rows represent the sample dimension and the columns are the variables. The variables can be either the data points for the completely unfolded chromatograms or tabulated peak areas. Note that multiway PCA [75] develops the same decomposition model as PCA but for third-order data, preserving chemical information that can be lost when reducing the dimensionality of 2D chromatographic data. An examination of the scores and loadings matrices can provide information on the similarities and/or differences between samples and chemically relevant analytes, respectively. The scores matrix (**T**) describes the coordinates for each sample on the PC axis and the loadings matrix (**P**) highlights the peaks responsible for the variation described by each PC. Plotting the scores on PC 2 versus the scores on PC 1, termed a scores plot, illustrates the relationship between samples in a data set. Ideally, similar samples should have similar scores while dissimilar samples should be separated from one another on the scores plot. The separation between these clusters can be quantified using a variety of metrics such as a degree-of-class separation (DCS) metric [36,41], the Mahalanobis distance [76], and construction of confidence ellipses [77,78]. Furthermore, investigation of the loadings for each PC can determine the peaks responsible for the sample separation on the scores plot. For a given PC, peaks with positive loadings are more abundant in samples with positive scores while peaks with negative loadings are more abundant in samples with negative scores.

Since PCA is extremely sensitive to all sources of variance, applying preprocessing methods is essential for attaining chemically meaningful results. Along with using baseline correction and normalization techniques, retention time shifting should also be reduced. If chromatographic misalignment is not corrected for, then the first few PCs can capture the variance due to shifting instead of the sample-related variances that are of interest. The loading plots for those PCs will also show first derivative Gaussian-like signals instead of chromatographic peaks due to retention time shifting [79]. Therefore, a strategy to mitigate chromatographic misalignment must be employed prior to PCA. The use of peak tables is a common approach for overcoming retention time shifting. These tables are typically generated by commercial software, which attempts to identify and quantify the analytes present in each chromatogram before aligning the tables. However, this approach may not be successful for highly saturated chromatograms where multiple chemical species overlap. Therefore, the analyst may want to either apply an alignment algorithm to the data set or bin the data to overcome misalignment. Application of a retention time alignment algorithm to pixel-level data (i.e., every data point in the unfolded chromatograms) was shown to increase the DCS between different gasoline samples [36]. Binning the pixel-level data can also minimize chromatographic misalignment while increasing the *S/N*. Sudol et al. showed that a maximum DCS between two fuel classes on a scores plot occurs at an optimal level of binning [41]. However, when examining the five adjacent fuel pairs in the study, each had a different optimum bin size due to the number of chemical differences between the fuels [41]. This result indicates that the analyst must select a bin size that balances the *S/N* improvement while minimizing misalignment and maintaining chemical selectivity.

PCA is typically the first chemometric technique applied to GC × GC and LC × LC applications because of the data reduction and visualization provided. For

example, Alexandrino and Augusto performed PCA on a GC × GC-MS data set of crude oils extracted from lacustrine (L) and marine (M) environments to discover the respective geochemical fingerprints [80]. Prior to performing pixel-based PCA, the authors minimized retention time shifting through the use of an in-house piecewise peak alignment algorithm [80]. Visual inspection of the scores plot in Figure 4.10A shows that PC 1, representing 39.15% of the total variance, separates the crude oil samples based upon their environmental source. The loadings for PC 1 at *m/z* 177 and 191 in Figure 4.10B indicate that crude oil from lacustrine environments have higher concentrations of tri- and tetracyclic terpanes and $\alpha\beta$-hopanes (positive loadings) while oils from marine sources have higher concentrations of norneohopanes, homohopanes, and gammacerane (negative loadings). Likewise, the PC 1 loadings at *m/z* 217, 231, and 259 in Figure 4.10C also show that marine crude oils have an increased abundance of diasteranes, $\alpha\alpha\alpha$-steranes, methyl steranes, and triaromatic steroids. Therefore, the sample clustering on PC 1 was due to the ratio of steranes to $\alpha\beta$-hopanes, a common diagnostic ratio for geochemical investigations [80].

4.4.2 Hierarchical Cluster Analysis (HCA)

Along with PCA, HCA is also used to quantify and visualize the similarities and differences between samples using either the unfolded chromatogram or tabulated peak areas. The similarity between samples is commonly calculated using an agglomerative (i.e., bottom-up) approach, which starts with each sample as an individual cluster member and merges similar samples into clusters [81,82]. Note that a divisive (i.e., top-down) approach, where all observations start in one cluster and then are split into smaller clusters, can also be used [81,82]. HCA requires the analyst to define both the distance metric and linkage algorithm. While a variety of distance metrics can be defined [82], either the Euclidean or Manhattan distance is used in practice to determine the similarity between samples. The Euclidean distance measures the straight-line distance between points (root-mean-squared differences in the coordinates) while the Manhattan (city block) distance quantifies the distance in right angle steps between points (absolute value of differences in the coordinates). Various linkage algorithms, which define the criteria used to group clusters together, have also been described in the literature [82]. However, the most popular algorithms are single-link, complete-link, average-link, and Ward's methods. The single-link algorithm finds the minimum distance between two members belonging to different clusters, whereas the complete-link approach considers the maximum distance between two members belonging to different clusters. The average-link method groups clusters together based on the average distance between all member pairs in the two clusters, which results in a minimization of within-cluster variance. Ward's method also produces this result by finding the pair of clusters that minimizes the increase in total within-cluster variance after merging [83].

The resulting arrangement of the clusters for HCA can then be plotted on a dendrogram. Figure 4.11 provides an example dendrogram after HCA was performed on quantified volatile organic compounds discovered in the headspace of aging blood samples with GC × GC-TOFMS [84]. The right-most nodes on the dendrogram (i.e., the leaf nodes) represent each blood sample that was analyzed. Every other node on

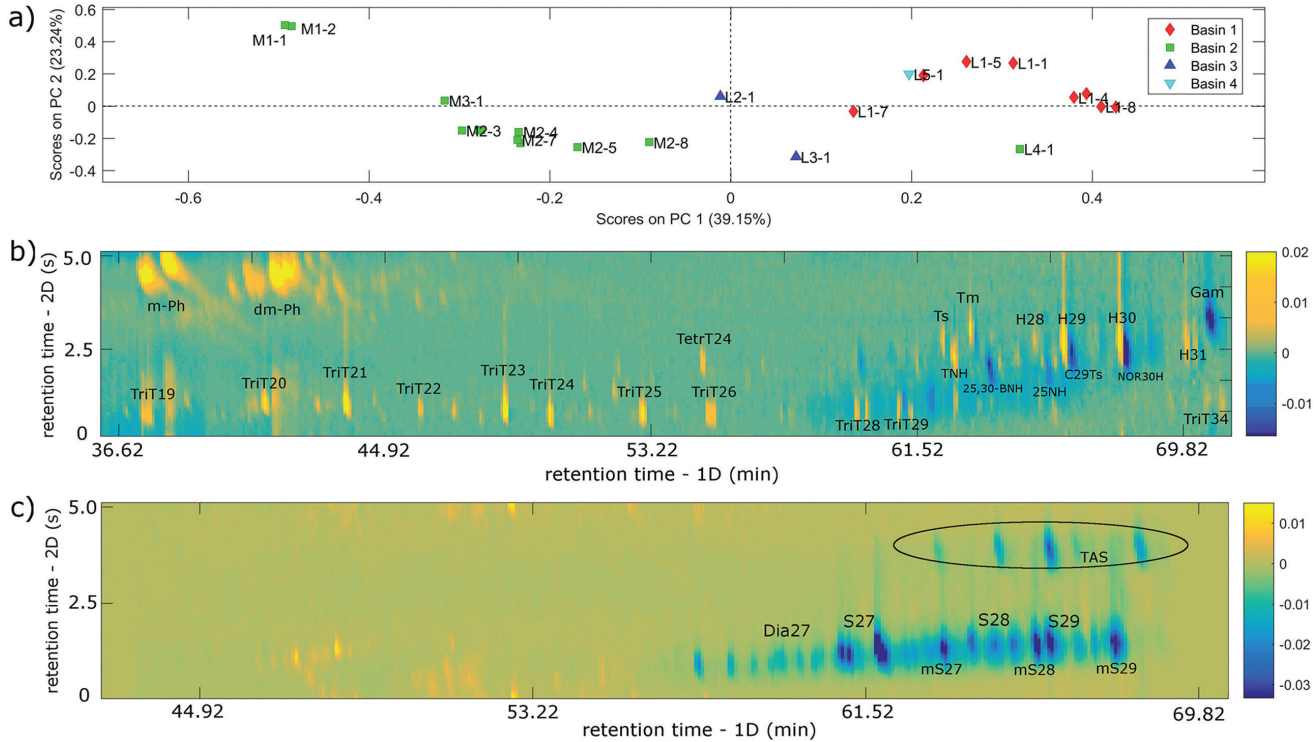


FIGURE 4.10 PCA results of crude oil samples extracted from marine (M) and lacustrine (L) environments The data set was created by concatenating *m/z* 177 191, 217, 231, and 259 to target the following compound classes: terpanes, steranes, hopanes, and triaromatic steroids. (A) The scores plot for this data set shows that the marine and lacustrine samples are mainly separated on PC 1. (B) PC 1 loadings for *m/z* 177 and 191. Yellow regions represent compounds with high abundance in samples with positive scores while blue regions highlight peaks with high abundance in samples with negative scores. (C) PC 1 loadings for *m/z* 217, 231, and 259.

This figure was taken with permission from G.L. Alexandrino, F. Augusto, Comprehensive two-dimensional gas chromatography-mass spectrometry/selected ion monitoring (GC × GC-MS/SIM) and chemometrics to enhance inter-reservoir geochemical features of crude oils, Energy & Fuels 32 (2018) 8017–8023. https://doi.org/10.1021/acs.energyfuels.8b00230.

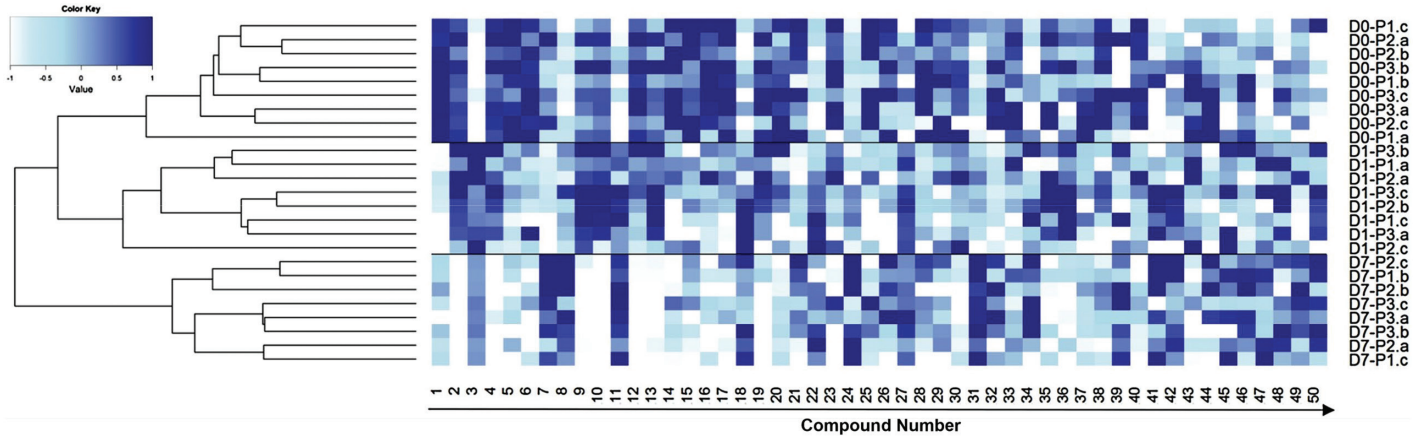


FIGURE 4.11 HCA performed on a GC × GC-TOFMS data set of volatile analytes present in blood at collection, after one day, and after one week. Samples, which are listed horizontally, are labeled according to the day of measurement (D), individual (P), and replicates (a–c). Volatiles discovered from supervised analysis are numbered and listed vertically in the heat map.

This figure was taken with permission from L.M. Dubois, K.A. Perrault, P.-H. Stefanuto, S. Koschinski, M. Edwards, L. McGregor, J.-F. Focant, Thermal desorption comprehensive two-dimensional gas chromatography coupled to variable-energy electron ionization time-of-flight mass spectrometry for monitoring subtle changes in volatile organic compound profiles of human blood, J. Chromatogr. A 1501 (2017) 117–127. https://doi.org/10.1016/j.chroma.2017.04.026.

the dendrogram represents an identified cluster, where all samples connected to that node are members of that cluster. To determine the significant clusters in the data set, the dendrogram would be "cut" at the distance corresponding to the longest branches. For the dendrogram in Figure 4.11, the longest branches separate the three ages of blood analyzed (day 0, day 1, and 1 week). Interpretation of the dendrogram can also be enhanced with a heat map of the relative abundance of each analyte. For example, the heat map in Figure 4.11 shows that limonene (compound 1), camphene (compound 11), and α-pinene (compound 16) were cluster-discriminating analytes [84].

4.4.3 Partitional Clustering Analysis

In contrast to HCA, the partitional clustering algorithm simultaneously assigns all samples into *k* clusters, where *k* represents the number of specified groups in the data set. These algorithms iteratively relocate the samples between clusters until the sum of squared distances is minimized. The most popular partitional clustering algorithm is *k*-means clustering, which defines each cluster centroid as the mean of all samples assigned to that cluster [85]. The algorithm randomly assigns *k* cluster centroids and the distances between sample and cluster centroids are calculated. Either the Euclidean or Manhattan distance can be used for this calculation. During the cluster assignment and centroid update steps, the samples are grouped to their closest cluster centroid and the algorithm determines the new cluster centroids. The algorithm then recalculates the sample-to-centroid distances and repeats the cluster assignment and centroid update steps. This method is repeated until cluster memberships do not change, or a maximum number of iterations is reached. Due to the random selection of centroids, the resulting cluster assignments are not reproducible between algorithm runs. Therefore, the *k*-means algorithm is commonly performed multiple times with different initial centroids and the model with the smallest sum-of-squared distances is selected as the appropriate model [86]. Other variations of *k*-means clustering, which heuristically select the initial centroids or limit the cluster centroids to be a member of the cluster, have also been proposed to improve model reproducibility [86,87].

Appropriate selection of the number of clusters, *k*, to model is imperative to achieving useful cluster assignments. In practice, cluster assignments at different values of *k* are compared using a clustering validity index [82,86]. Numerous index calculations have been described in the literature with the goal of quantifying within-cluster compactness and between-cluster separation [82]. However, the silhouette index [88] and Davies-Bouldin index [89] are most commonly used. The silhouette index provides a measure of how similar a sample is to others in its own cluster relative to other samples in a neighboring cluster. The resulting metric is termed a silhouette value, which ranges from -1 (not well clustered) to 1 (well clustered). The appropriate number of clusters, *k*, can be determined by selecting the clustering solution that has an average silhouette value closest to 1. For example, silhouette results revealed that GC × GC-TOFMS chromatograms of different mask materials clustered into three groups, corresponding to their relative abundance of branched alkanes and olefins [90]. Similarly, the Davies-Bouldin index

measures the ratio of the within-cluster variance to between-cluster distances. The optimal clustering solution has the smallest Davies-Bouldin index value. Instead of using a clustering validity index, recent work by Cain et al. showed that the clustering solution with the smallest probability of occurring by chance is most likely to be due to underlying chemical differences in the chromatographic data set [91].

4.5 SUPERVISED, NON-TARGETED ANALYSIS

While unsupervised approaches are appropriate for initial investigations into a chromatographic data set, supervised approaches are well suited for studying cause and effect experiments by leveraging *a priori* information. Supervised algorithms utilize target variables such as class labels or independently measured sample properties to discover features, build regression models, and/or classify samples. Feature discovery, also known as feature selection, finds a subset of the original chromatographic data that is highly correlated with the target variable(s). For chromatographic data sets, Fisher ratio (F-ratio) analysis is typically used to discover class-distinguishing analytes. It is important to note that unsupervised methods like PCA and HCA are commonly used to visualize the results obtained from non-targeted, supervised feature selection methods. Along with identifying significant analytes (feature selection), methods used for property prediction and sample classification fall under the umbrella of supervised analysis techniques. Note that in this context, property refers to either a chemical or physical quantity that was collected separately from the chromatographic data set. The most common property prediction method is partial least squares (PLS) regression, which develops a multivariate calibration model to discover which analytes correlate with the sample property that is being modeled. The extension of this algorithm, known as partial least squares-discriminant analysis (PLS-DA), can be used for classification applications. This section will detail the principles and preprocessing considerations for each supervised technique.

4.5.1 Fisher Ratio (F-Ratio) Analysis

F-ratio analysis is a popular feature selection technique for chromatographic data because it inherently provides a degree of data reduction, focusing the overall data analysis before performing further targeted and non-targeted chemometric methods. This feature selection method utilizes the analysis of variance (ANOVA) statistical hypothesis test, which compares the variance of observations and discovers significant differences between groups. The total variance, defined as the squared standard deviation, can be partitioned into two contributions: variance *between* classes of samples and *within* classes of samples. The between class (BC) variance, which describes how each class mean varies from the grand mean, is defined as

$$\sigma_{\mathrm{BC}}^2 = \frac{1}{k-1}\sum\left(\bar{x}_{\mathrm{i}} - \bar{\bar{x}}\right)^2 n_{\mathrm{i}} \tag{4.4}$$

where k is the number of classes, n_i is the number of measurements in the ith class, $\bar{x}_i$ is the mean of the ith class, and $\bar{x}$ is the grand mean. The within class (WC) variance, which indicates how much each measurement varies from its class mean, is

$$\sigma^2_{WC} = \frac{1}{N-k}\sum\sum\left(x_{ij} - \bar{x}_i\right)^2 \tag{4.5}$$

where N is the total number of measurements, and x_{ij} is the jth measurement of the ith class. Finally, the F-ratio is then obtained by taking the ratio of these two quantities:

$$F - ratio = \frac{\sigma^2_{BC}}{\sigma^2_{WC}} \tag{4.6}$$

The results from the F-ratio analysis are compiled in a "hit list," which ranks the F-ratio values in descending order. The analyst then mines the hit list in a top-down approach, identifying and quantifying peaks with larger F-ratios, since a high F-ratio generally corresponds to class-distinguishing analytes. A class-distinguishing analyte is an analyte whose concentration is statistically different between classes, which is typically based upon a t-test having a p-value < 0.05 (95% confidence limit). Thus, the results of the t-test show that the concentration ratio between classes sufficiently differs from one. These class-distinguishing analytes are commonly referred to as true positives. However, this data mining approach can be hindered by the presence of both false positives and negatives. A false positive refers to the discovery of an analyte that is not statistically different between classes, while a false negative is the inability to discover a class-distinguishing analyte. The presence of false positives and/or negatives can be present in all implementations of F-ratio analysis (peak table, pixel-based, and tile-based). Therefore, along with discussing the different approaches for F-ratio analysis, strategies to reduce the false discovery rate will be introduced in this section.

Peak table-based [92–94] and pixel-based [95–98] are the most straightforward approaches for F-ratio analysis. A peak table-based approach typically uses the instrumental software to baseline correct and quantify signals for each peak in the chromatograms. These peak tables are then aligned based upon user-defined time windows and mass spectrum match criteria prior to F-ratio analysis [92]. While this approach aids in limiting the pool of potential features to only those with measurable signals, it is important to note that this approach is limited by the capability of the peak finding software and thus, potentially class-distinguishing analytes with low *S/N* can be missed. A pixel-based approach, on the other hand, calculates an F-ratio for every data point in the chromatogram (i.e., at every 2D retention time on every detector channel). The advantage of a pixel-based approach is that it utilizes the entire data set; however, the number of false positives can increase rapidly if a retention time alignment algorithm is not applied. Even if misalignment is mitigated, random detector fluctuations can artificially inflate F-ratios and increase the number of false positives [99].

To address this challenge, a new tile-based approach for F-ratio analysis of comprehensive 2D chromatographic data was developed for GC × GC-TOFMS data

[99,100]. Here, the chromatogram is divided into regular repeating rectangular sections called "tiles," where the encapsulated signal is summed (binned) to a single value. This tiling procedure not only reduces the overall size of the chromatogram but also reduces the effect of retention time shifting. To ensure that peaks are not split between tiles, F-ratio analysis is performed using four different tiling grids, where each grid is offset from the original tile grid by half a tile width in either dimension. A four-grid tile scheme ensures that each peak in the chromatogram is adequately captured by at least one tile. The resulting tiled data sets are then compared using F-ratio analysis on a per-*m/z* basis. Next, a pin approximating the location of the peak for each hit is calculated using the maximum signal difference for each tile, and pins with similar retention times are clustered together. This pinning and clustering step ensures that each feature in the hit list is only represented by one pin with the highest F-ratio.

Proper reduction of false positives/negatives using tile-based F-ratio analysis first requires optimization of the tile size for both dimensions, ^{1}D and ^{2}D. Ideally, the tile dimensions should encompass the average peak width along with any misalignment [99,100]. If the tile size is too large, then interferent signals could mask the signal of true positives at low concentrations [101]. Conversely, the hit list generated after using a tile smaller than optimal can result in numerous redundant hits, which may be interpreted as false positives [101]. Other strategies to reduce the presence of false positives require optimization of different parameters involved in calculating F-ratios and ranking the discovered features. Reaser et al. demonstrated that the occurrence of false positives can be mitigated by using a *S/N* of 10 and ranking F-ratios based on the average of the top 10 *m/z* [102]. Likewise, Sudol et al. showed that in part by ranking the hit list using only the *m/z* that produced the top F-ratio for a given hit allowed for the discovery of features at concentrations as low as 1 ppm [101]. For data sets consisting of control and treatment classes, F-ratios can be calculated using solely the variance of the control class in the denominator [103]. This calculation, termed control-normalized F-ratio, was shown to discover class-distinguishing features that were initially missed by traditional F-ratio analysis due to their non-uniform/high variance in the treatment class [103]. An analyst can evaluate the effectiveness of changing these different parameters on the true and false positive rates with receiving operator characteristic (ROC) curves [102]. The area under the curve (AUC) for the ROC curve can then quantify the optimization of different parameters, where higher AUCs indicate more true positives appear at the top of the hit list [102].

Along with optimizing the calculation and ranking parameters, a reduction of the hit list can be achieved by selecting an F-ratio threshold, deeming any feature above the threshold to be important for further analysis. Traditionally, a manual cutoff can be selected by determining the point where the frequency of false positives increases. However, this method is both time-consuming and subjective toward the false positive tolerance limit defined by the analyst. An F-critical threshold can also be applied, but this approach generally picks small cutoffs that keep too many false positives. Null distribution analysis has been demonstrated to be a more robust method for cutoff determination [100]. This method develops false positive F-distributions from repeated F-ratio analyses on classes built from redistributing

samples evenly and randomly into classes. A distribution of F-ratios from the null comparisons is developed at a desired null probability. The F-ratio threshold is then chosen by identifying the F-ratio that corresponds to a user-defined confidence level. Ultimately, null distribution analysis allows for the objective determination of an F-ratio threshold, which is unique and accurate because it naturally considers the underlying noise for a given data set.

Figure 4.12 illustrates the operation of tile-based F-ratio analysis for the comparison of yeast cell extracts grown under different conditions [104]. Using cystathionine as a representative peak in the chromatogram, the chosen tile size was 6s (four modulations) on ^{1}D and 300 ms on ^{2}D (Figure 4.12A). This tile size was deemed optimal because it not only encompassed signal from the entire peak on both dimensions, but it also compensated for the retention time shifting seen in ^{1}D and ^{2}D (Figure 4.12B-C). Using the optimal tile size, approximately 1700 potential features were discovered via F-ratio analysis [104]. Since mining this hit list would be labor-intensive, null distribution analysis was utilized to determine the F-ratio threshold. Since this comparison contains six samples in each class, there are 200 null permutations of class membership for analysis [104]. A null F-ratio distribution can be developed by plotting the F-ratio values from those 200 permutations at a null probability of 0.1% (Figure 4.12D; blue). This null probability was chosen because it corresponds to a false discovery rate (FDR) of 1 out of 1000 features. From this null F-distribution, an F-ratio threshold of 15 was determined, because this value provides 95% certainty that a 1 in 1000 FDR will be achieved when truncating the true class comparison. Application of this threshold to the F-distribution of the true class comparison (Figure 4.12D; black) resulted in the discovery of 94 features, nearly a 20-fold reduction in the original hit list. The 94 discovered features are represented by circles on the GC × GC-TOFMS chromatogram (Figure 4.12E). The discovered features were then processed using PARAFAC to obtain pure compound profiles and spectra [104]. Overall, this example illustrates how tile-based F-ratio analysis coupled with null distribution analysis can be effectively used as a data reduction tool for supervised, non-targeted analysis.

4.5.2 Partial Least Squares (PLS) Regression

Partial least squares (PLS) is a multivariate regression method that correlates the information in a data matrix (**X**) to information in another matrix (**Y**), which may be a single column vector (i.e., PLS1), or a multi-column matrix (i.e., multiway PLS; n-PLS) [105]. PLS is often used to predict some property in **Y** that is difficult or expensive to obtain by the reference method, using chromatographic data (**X**). In essence, PLS analysis is based on performing PCA individually on **X** and **Y**. In PCA, the direction of the loadings in **X** and **Y** maximizes the variance within each of the respective matrices. However, in PLS, the direction of loadings of **X** and **Y** is chosen to maximize the covariance between these two matrices. Analogous to PCs, the variation in the **X**-block is described by a series of orthogonal linear latent variables (LVs). A visual illustration of PLS model generation and optimization on unfolded comprehensive 2D chromatograms is provided in Figure 4.13. Note that PLS can be a computationally expensive technique and thus is seldom performed on pixel-level comprehensive 2D

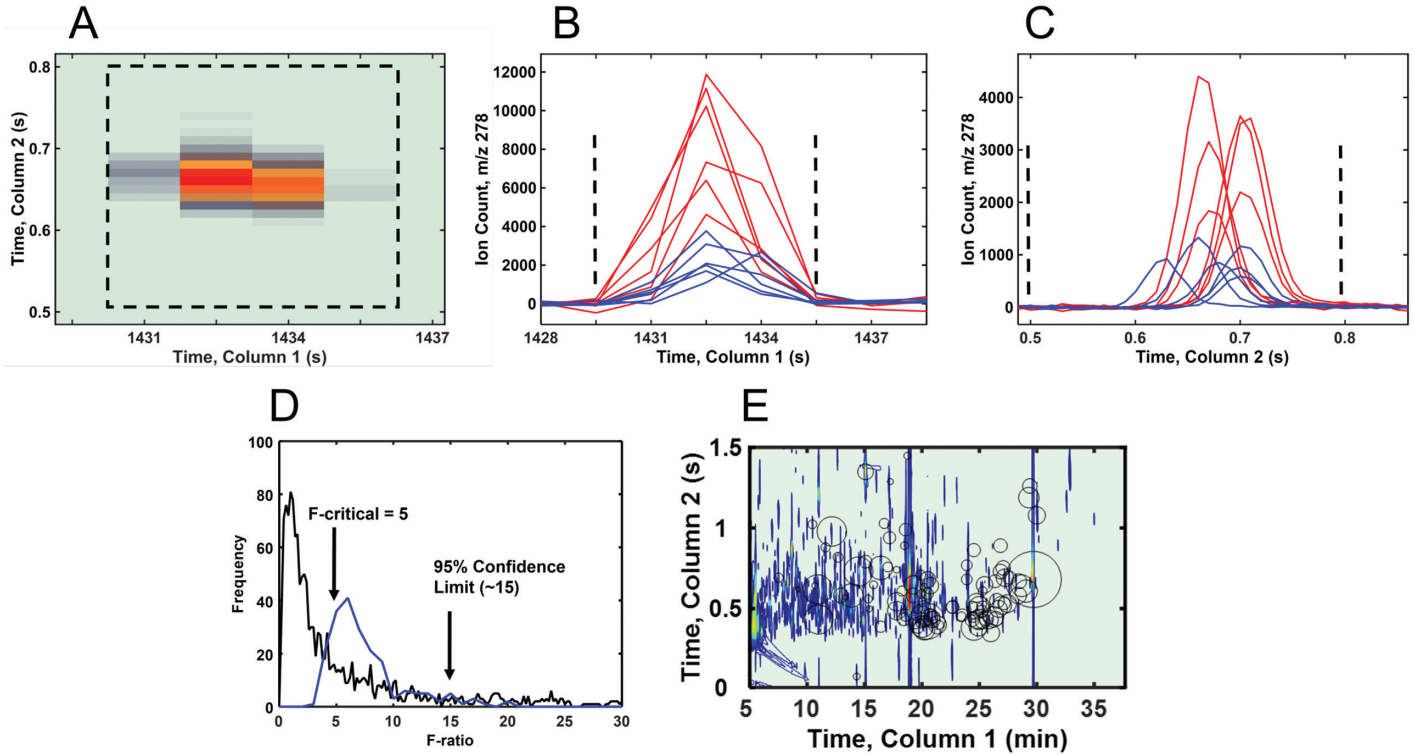


FIGURE 4.12 Results from tile-based F-ratio analysis on a GC × GC-TOFMS data set of yeast samples grown under derepressed (DR) and repressed (R) conditions. (A) Zoom-in on cystathionine at *m/z* 278 to illustrate the selection of the tile size for F-ratio analysis. (B) The ^{1}D peak profile for cystathionine in (A), where the black dashed lines represent the ^{1}D tile size. (C) The ^{2}D peak profile for cystathionine in (A), where the black dashed lines represent the ^{2}D tile size. (D) Resulting F-ratio distribution for the true class comparison (black) and 0.1% null F-distribution (blue). Both the F-critical threshold and threshold from null distribution analysis are labeled. (E) Total ion current chromatogram for DR class with analytes discovered by F-ratio analysis circled. The size of the circles corresponds to the calculated F-ratio for the peak.

This figure was taken with permission from N.E. Watson, B.A. Parsons, R.E. Synovec, Performance evaluation of tile-based Fisher Ratio analysis using a benchmark yeast metabolome dataset, J. Chromatogr. A 1459 (2016) 101–111. https://doi.org/10.1016/j.chroma.2016.06.067.

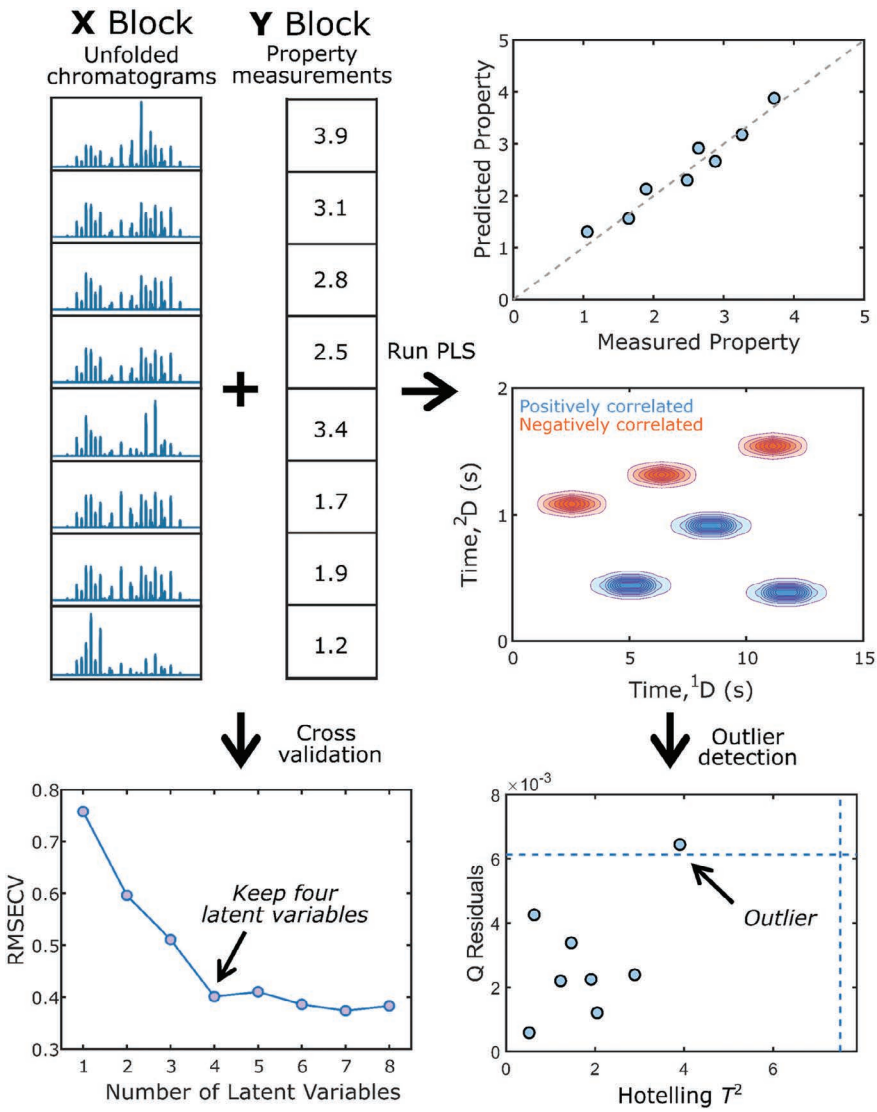


FIGURE 4.13 Schematic illustrating PLS regression analysis for comprehensive 2D separations. The **X**-block contains the chromatograms in their vectorized form, and the **Y**-block contains the corresponding property measurements. Following PLS, a regression model and loadings (or linear regression vectors; LVRs). The regression model is a plot of the predicted versus measured property values. The LVRs highlight the features that are both positively (blue) and negatively correlated (red) to the given property. Cross-validation is performed to determine the number of latent variables retained in the PLS model. Furthermore, a plot of the Q residuals versus Hotelling T^2 can be used for outlier detection.

chromatographic data. Therefore, data reduction techniques such as removing uninformative chromatographic regions or *m/z* values, binning [106,107], and feature selection (e.g., F-ratio analysis) [108] can be used before constructing the **X**-block.

The number of LVs to include is an important step in the development of a PLS model. This is generally determined via leave-one-out-cross-validation (LOOCV), wherein one row of **X** (e.g., an unfolded chromatogram) is excluded from the model, and the model is rebuilt. After repeating this for every row of **X**, the root-mean-square error of cross-validation (RMESCV) is computed, which measures the difference in the cross-validation predicted value of the samples and their measured values from **Y**. Then, the number of LVs to include is selected from the model with the lowest RMESCV, or after the change in RMESCV upon adding additional LVs becomes negligible (Figure 4.13). However, LOOCV is computationally expensive for large data sets and can result in model overfitting. Hence, the cross-validation method may employ one or more sub-validation experiments in which a subset of samples, rather than a single sample, is removed from the **X**-block to generate a validation set. Examples of sub-validation methods include Venetian blinds, contiguous blocks, and random subsets [109]. These sub-validation methods differ in how the samples to remove from **X** are selected. The proper selection of a sub-validation model(s) depends on the nature of the data set and the analysis goals.

The primary outcome of a PLS model is a regression plot showing the correlation of the predicted property to the measured property (Figure 4.13). For n-PLS, a regression plot is generated for each column of **Y**. Ideally, the regression plots should have a correlation coefficient close to 1, demonstrating that the property (or properties) of interest is (are) being accurately represented by the chromatographic data. Furthermore, interpretation of the linear regression vectors (LRV) can specify which variables of **X** are positively correlated, negatively correlated, or non-correlated to **Y**. Each LRV can be refolded to produce a plot that visually appears like a comprehensive 2D chromatogram (Figure 4.13). Positive variables in the LRV will correspond to chromatographic variables that are positively correlated with the predicted variables, whereas anticorrelated variables will be negative. Regions of little or no intensity in the LRVs correspond to chromatographic variables that do not correlate with the predicted property. Outliers in the model can also be detected by examining a plot of the Q residuals versus Hotelling's T^2 statistic (Figure 4.13). The Q residuals are a measure of the difference between the original and modeled data while the Hotelling's T^2 statistic calculates the variation of each sample within the model [110]. Therefore, samples with a high Q residual or Hotelling's T^2 could be considered possible outliers because the samples deviated greatly from the predictions made by the model.

Berrier et al. demonstrated the benefits of PLS regression for the creation of models that can predict a variety of performance properties (e.g., density, kinematic viscosity, and net heat of combustion) for aerospace kerosene fuels [107]. Figure 4.14A shows the TIC chromatogram for a representative fuel that was collected on a GC × GC-TOFMS instrument with a reversed-column format (polar ^{1}D column and non-polar ^{2}D column). From this TIC chromatogram, three chemical classes can be seen: alkanes (purple), cycloalkanes (orange), and aromatics (green). Figure 4.14B shows the PLS regression plot of viscosity for the fuel data set after the chromatograms were binned. This regression plot shows that the samples are closely clustered

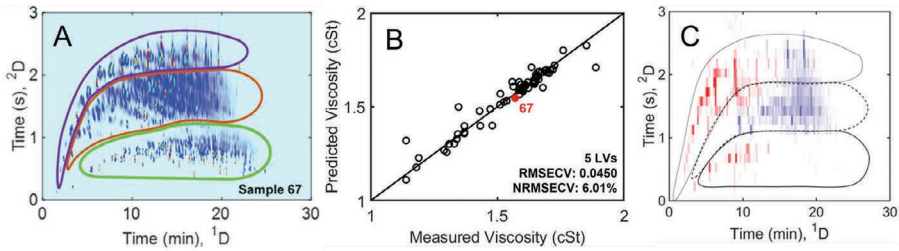


FIGURE 4.14 PLS regression analysis performed on a GC × GC-TOFMS data set of kerosene-based rocket fuels. (A) TIC chromatogram of a representative fuel sample with compound classes primarily separated on the ^{2}D. The compounds outlined in purple are alkanes, orange are cycloalkanes, and green are aromatics. Prior to analysis, each sample chromatogram was binned to improve computational speed. (B) PLS regression model of viscosity with the number of latent variables (LVs) and prediction errors provided. (C) The LVR for the PLS model shown in (B) with the locations of the three compound classes overlaid: alkanes (dotted line), cycloalkanes (dashed line), and aromatics (solid line). Blue regions indicate bins with positive values for the LVR while red regions indicate bins with negative values.

This figure was taken with permission from K.L. Berrier, C.E. Freye, M.C. Billingsley, R.E. Synovec, Predictive modeling of aerospace fuel properties using comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry and partial least squares analysis, Energy & Fuels 34 (2020) 4084–4094. https://doi.org/10.1021/acs.energyfuels.9b04108.

around the 1:1 line and the model has a low normalized RMSECV, indicating that the model utilizing the chemical composition data accurately characterizes the experimental property data. Furthermore, the LVR (Figure 4.14C) shows that analytes with higher boiling points (e.g., long-chain alkanes and cycloalkanes) positively correlate with increasing viscosity while analytes with lower boiling points (e.g., short-chain alkanes and small aromatics) correlate with lower viscosities. Thus, PLS models can not only be used for property prediction of new samples but can also discover analytes (or compound classes) in the separation that relate to those properties.

4.5.3 Partial Least Squares-Discriminant Analysis (PLS-DA)

PLS-DA is a non-targeted classification and feature discovery method that is analogous to a supervised version of PCA but is mathematically similar to PLS. Given the similarity between PLS and PLS-DA, the reader is instructed to refer to the earlier PLS section and the following references for an in-depth description of its mathematical operation [110–112]. However, it is important to note that the **Y**-block for PLS-DA is not composed of continuous variables, as in PLS. Instead, PLS-DA regresses the chromatographic data in the **X**-block against categorical variables denoting class membership (e.g., 0 and 1 for a two-class model) in the **Y**-block. Furthermore, like PLS and PCA, the **X**-block in PLS-DA seldom contains pixel-level chromatograms because of the computational burden and undesirable effects of retention time shifting. If the data is well-aligned, one *m/z* may be used for

model development or multiple PLS-DA models may be built for different *m/z* [113]. Alternatively, PLS-DA can be performed on peak tables [33,114] or chromatograms after applying data reduction techniques (i.e., binning or feature selection) [115] to mitigate the effects of misalignment and reduce computational expense. Analogous to a PCA scores plot, the main output from PLS-DA is a scores plot that plots the scores on LV 2 versus LV 1 to visualize the similarities/differences between samples.

It is important to note that model overfitting is common for data sets that contain many more variables than samples, which is often the case when using comprehensive 2D separations. In these instances, variables that do not correspond to meaningful chemical difference may be excellent sample classifiers solely due to random chance [116]. Therefore, the PLS-DA scores plot may suggest an excellent classification model even though the underlying data does not support the same conclusion. Several validation and optimization methods can be employed to prevent model overfitting. Prior to model development, the data set can be split between a training set and external validation set. The training set is used to generate the PLS-DA model, and the prediction success of the PLS-DA model is verified with the external validation set [117]. For a definitive approach, the analyst may permute the **Y**-block, such that the label attached to each sample is randomized [118]. Following many permutations, the results of the permuted data set are compared to the results of the model with true class labels. Furthermore, ROC curves are frequently employed as an additional validation step to assess the probability of correctly classifying a sample across a range of classification thresholds [118,119].

In addition to classification, PLS-DA can be used to perform feature selection. If true chemical variance is driving class membership, then chromatographic variables of high intensity in the loadings matrix may correspond to peaks in the original GC × GC or LC × LC data that are class-distinguishing analytes. To verify that these highly loaded peaks correspond to meaningful chemical differences, the concentration change should be independently tested for statistical significance (i.e., a *t*-test or one-way ANOVA at the desired confidence interval). Furthermore, several statistical metrics can be used to perform feature selection using the PLS-DA outputs, such as the selectivity ratio (SR) and variable importance in projection (VIP). The SR, measured as the ratio of explained to residual variance, can be used to rank the impact of each chromatographic variable in descending order [120]. Similarly, VIP scores quantify the influence of every individual chromatographic variance on the overall PLS-DA model using its loadings weight [121]. As a simple rule of thumb, variables with VIP scores greater than one are retained as potential class-distinguishing features [121]. For example, Navarro-Reig et al. collected LC × LC-QTOF-MS chromatograms (Figure 4.15A) of rice samples to study the effects of watering on rice metabolism [33]. Due to the complexity of the data set, each chromatogram underwent both ROI and wavelet compression prior to identifying and quantifying analytes with MCR-ALS [33]. A PLS-DA model for these analytes shows a clear separation between the non-watered (orange) and watered (blue) rice samples (Figure 4.15B). The VIP scores in Figure 4.15C demonstrate that flavonoids and glycosides are responsible for this differentiation in the PLS-DA model, since these were found in higher abundance for the non-watered rice samples [33].

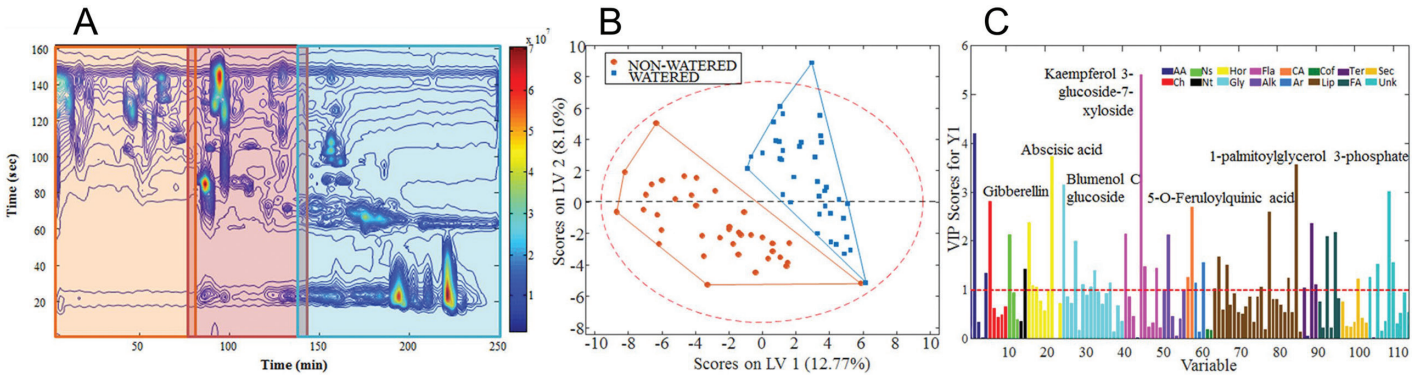


FIGURE 4.15 PLS-DA performed on a LC × LC-QTOF-MS data set investigating the rice metabolism under different watering strategies. (A) Chromatogram collected for a representative rice sample. Given the size of the data, the chromatogram was split into three different time regions (highlighted in orange, red, and blue) and the data was compressed using the regions of interest approach. (B) PLS-DA scores plot of non-watered (orange) and watered (blue) samples using the peak areas obtained from MCR-ALS analysis of the compressed chromatogram in (A). (C) A plot of the variable importance in projection (VIP) scores measured for each metabolite quantified in the chromatograms (i.e., the chromatographic variable). Each metabolite is colored according to its compound class, and metabolites responsible for the sample separation on the PLS-DA scores plot (i.e., have a high VIP value) are labeled.

The figure was taken with permission from M. Navarro-Reig, J. Jaumot, A. Baglai, G. Vivó-Truyols, P.J. Schoenmakers, R. Tauler, Untargeted comprehensive two-dimensional liquid chromatography coupled with high-resolution mass spectrometry analysis of rice metabolome using multivariate curve resolution, Anal. Chem. 89 (2017) 7675–7683. https://doi.org/10.1021/acs.analchem.7b01648.

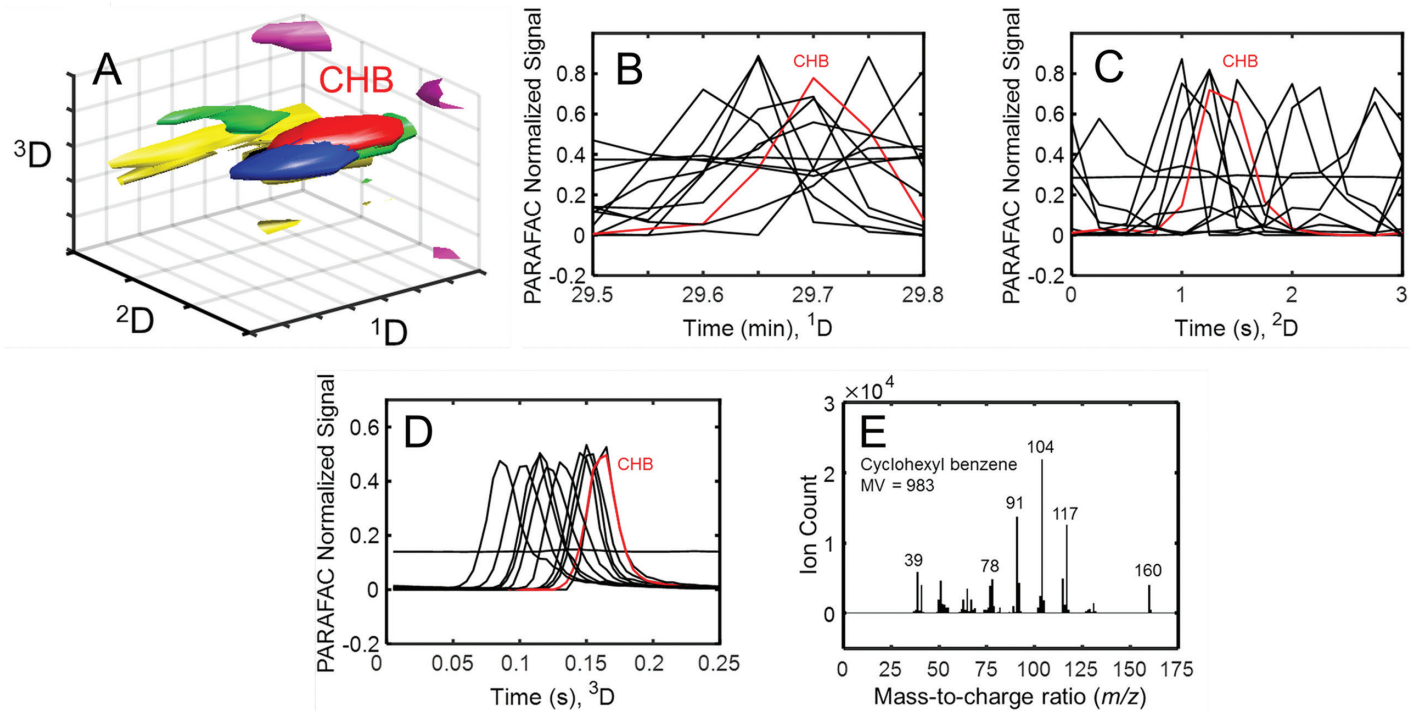


FIGURE 4.16 PARAFAC decomposition of cyclohexyl benzene (CHB—red peak profile) in diesel separated using GC^3-TOFMS. (A) Isosurface plot of *m/z* 57 (yellow), 81 (black), 104 (red), 118 (blue), 131 (green), and 132 (magenta) to show the analytes overlapped with CHB (red retention ellipse). (B) PARAFAC resolved ^{1}D chromatographic profiles. The peak for CHB is shown in red while all other peaks are shown in black. (C) PARAFAC resolved ^{2}D chromatographic profiles. (D) PARAFAC resolved ^{3}D chromatographic profiles. (E) PARAFAC resolved mass spectrum for CHB. The match value (MV) to a library spectrum is also provided.

This figure was taken with permission from N.E. Watson, S.E. Prebihalo, R.E. Synovec, Targeted analyte deconvolution and identification by four-way parallel factor analysis using three-dimensional gas chromatography with mass spectrometry data, Anal. Chim. Acta 983 (2017) 67–75. https://doi.org/10.1016/j.aca.2017.06.017.

4.6 FUTURE PROSPECTUS

Comprehensive 2D chromatography has been proven to be advantageous in a wide variety of applications due to its increased resolving power and commercialization. However, the large file sizes produced from these instruments compel the need for chemometric analysis to maximize the amount of extracted chemical information in a timely and accurate manner. Hence, chemometric analysis is becoming a mainstay in advanced data handling for comprehensive 2D separations. Therefore, knowledge of common chemometric algorithms (i.e., their purpose, operation, and limitations) employed in the literature and incorporated into commercial or open-source software is necessary. This chapter has covered the fundamentals for targeted decomposition methods (MCR-ALS and PARAFAC) along with unsupervised (PCA, HCA, and *k*-means clustering) and supervised (F-ratio, PLS regression, and PLS-DA) nontargeted methods. This chapter has also covered the pertinent instrumental considerations and preprocessing methods required to produce high-quality data for these chemometric analyses.

The use of comprehensive 2D separations will certainly continue to evolve and find new applications. As stated in the introduction, the use of comprehensive 2D separations is growing, and researchers are becoming interested in analyzing these large and very information-rich data sets. Along with the size of these data sets, more complex analytical questions are being proposed. Automated workflows and machine learning algorithms are expected to continue to push the boundaries of chromatographic data analysis with minimal human intervention [122,123]. Furthermore, current research has focused on the development and implementation of three-dimensional (3D) separations [124–128], which has created new opportunities to test and improve chemometric methods [129–131]. For example, Watson et al. used both comprehensive 3D gas chromatography with TOFMS (GC^3-TOFMS) and PARAFAC to resolve and quantify both native and non-native compounds in diesel fuel [130]. Figure 4.16A provides an example of a non-native analyte, cyclohexyl benzene (red retention ellipse), which co-eluted with several native diesel compounds. PARAFAC decomposition of this four-way data set successfully resolved cyclohexyl benzene (red peak profile) into its pure ^{1}D (Figure 4.16B), ^{2}D (Figure 4.16C), ^{3}D (Figure 4.16D), and mass spectrum (Figure 4.16E). Therefore, higher dimensional instruments will allow analysts not only to benefit from the increase in peak capacity and selectivity, but also to access higher-order chemometric methods. As this chapter has shown, the use of comprehensive 2D separations and chemometrics will continue to flourish for years to come.

REFERENCES

[1] J.M. Davis, J.C. Giddings, Statistical theory of component overlap in multicomponent chromatograms, *Anal. Chem.* 55 (1983) 418–424. https://doi.org/10.1021/ac00254a003.
[2] F. Erni, R.W. Frei, Two-dimensional column liquid chromatographic technique for resolution of complex mixtures, *J. Chromatogr. A* 149 (1978) 561–569. https://doi.org/10.1016/S0021-9673(00)81011-0.
[3] M.M. Bushey, J.W. Jorgenson, Automated instrumentation for comprehensive two-dimensional high-performance liquid chromatography/capillary zone electrophoresis, *Anal. Chem.* 62 (1990) 978–984. https://doi.org/10.1021/ac00209a002.

[4] Z. Liu, J.B. Phillips, Comprehensive two-dimensional gas chromatography using an on-column thermal modulator interface, *J. Chromatogr. Sci*. 29 (1991) 227–231. https://doi.org/10.1093/chromsci/29.6.227.
[5] M.S. Klee, J. Cochran, M. Merrick, L.M. Blumberg, Evaluation of conditions of comprehensive two-dimensional gas chromatography that yield a near-theoretical maximum in peak capacity gain, *J. Chromatogr. A* 1383 (2015) 151–159. https://doi.org/10.1016/j.chroma.2015.01.031.
[6] D.R. Stoll, X. Wang, P.W. Carr, Comparison of the practical resolving power of one- and two-dimensional high-performance liquid chromatography analysis of metabolomic samples, *Anal. Chem*. 80 (2008) 268–278. https://doi.org/10.1021/ac701676b.
[7] R.E. Murphy, M.R. Schure, J.P. Foley, Effect of sampling rate on resolution in comprehensive two-dimensional liquid chromatography, *Anal. Chem*. 70 (1998) 1585–1594. https://doi.org/10.1021/ac971184b.
[8] J.V. Seeley, Theoretical study of incomplete sampling of the first dimension in comprehensive two-dimensional chromatography, *J. Chromatogr. A* 962 (2002) 21–27. https://doi.org/10.1016/S0021-9673(02)00461-2.
[9] R.E. Synovec, B.J. Prazen, K.J. Johnson, C.G. Fraga, C.A. Bruckner, Chemometric analysis of comprehensive two-dimensional separations, in: P.R. Brown, E. Grushka (Eds.), *Adv. Chromatogr.*, Marcel Dekker, New York, 2003: pp. 1–42. https://doi.org/10.1201/9780203911266.ch1.
[10] M.J. Wilde, B. Zhao, R.L. Cordell, W. Ibrahim, A. Singapuri, N.J. Greening, C.E. Brightling, S. Siddiqui, P.S. Monks, R.C. Free, Automating and extending comprehensive two-dimensional gas chromatography data processing by interfacing open-source and commercial software, *Anal. Chem*. 92 (2020) 13953–13960. https://doi.org/10.1021/acs.analchem.0c02844.
[11] E.B. Ledford, C. Billesbach, Jet-cooled thermal modulator for comprehensive multidimensional gas chromatography, *J. High Resolut. Chromatogr.* 23 (2000) 202–204. https://doi.org/10.1002/(SICI)1521-4168(20000301)23:3<202::AID-JHRC202>3.0.CO;2–5.
[12] C.A. Bruckner, B.J. Prazen, R.E. Synovec, Comprehensive two-dimensional high-speed gas chromatography with chemometric analysis, *Anal. Chem*. 70 (1998) 2796–2804. https://doi.org/10.1021/ac980164m.
[13] J.V. Seeley, F. Kramp, C.J. Hicks, Comprehensive two-dimensional gas chromatography via differential flow modulation, *Anal. Chem*. 72 (2000) 4346–4352.
[14] J.V. Seeley, N.J. Micyus, J.D. McCurry, S.K. Seeley, Comprehensive two-dimensional gas chromatography with a simple fluidic modulator, *Am. Lab.* 38 (2006) 24–26.
[15] P.Q. Tranchida, G. Purcaro, A. Visco, L. Conte, P. Dugo, P. Dawes, L. Mondello, A flexible loop-type flow modulator for comprehensive two-dimensional gas chromatography, *J. Chromatogr. A* 1218 (2011) 3140–3145. https://doi.org/10.1016/j.chroma.2010.11.082.
[16] T.J. Trinklein, D.V. Gough, C.G. Warren, G.S. Ochoa, R.E. Synovec, Dynamic pressure gradient modulation for comprehensive two-dimensional gas chromatography, *J. Chromatogr. A* 1609 (2020). https://doi.org/10.1016/j.chroma.2019.460488.
[17] H.D. Bahaghighat, C.E. Freye, R.E. Synovec, Recent advances in modulator technology for comprehensive two dimensional gas chromatography, *TrAC Trends Anal. Chem*. 113 (2019) 379–391. https://doi.org/10.1016/j.trac.2018.04.016.
[18] P.M.A. Harvey, R.A. Shellie, Factors affecting peak shape in comprehensive two-dimensional gas chromatography with non-focusing modulation, *J. Chromatogr. A* 1218 (2011) 3153–3158. https://doi.org/10.1016/j.chroma.2010.08.029.
[19] D.K. Pinkerton, B.A. Parsons, T.J. Anderson, R.E. Synovec, Trilinearity deviation ratio: A new metric for chemometric analysis of comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry data, *Anal. Chim. Acta* 871 (2015) 66–76. https://doi.org/10.1016/j.aca.2015.02.040.
[20] S.E. Prebihalo, D.K. Pinkerton, R.E. Synovec, Impact of comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry experimental design on

data trilinearity and parallel factor analysis deconvolution, *J. Chromatogr. A* 1605 (2019) 460368. https://doi.org/10.1016/j.chroma.2019.460368.

[21] M. Van Deursen, J. Beens, J. Reijenga, P. Lipman, C. Cramers, J. Blomberg, Group-type identification of oil samples using comprehensive two-dimensional gas chromatography coupled to a time-of-flight mass spectrometer (GC × GC-TOF), *J. High Resolut. Chromatogr.* 23 (2000) 507–510. https://doi.org/10.1002/1521-4168(20000801)23:7/8<507::aid-jhrc507>3.0.co;2-n.

[22] J.D. Byer, K. Siek, K. Jobst, Distinguishing the C3 vs SH4 mass split by comprehensive two-dimensional gas chromatography-high resolution time-of-flight mass spectrometry, *Anal. Chem.* 88 (2016) 6101–6104. https://doi.org/10.1021/acs.analchem.6b01137.

[23] P.Q. Tranchida, I. Aloisi, B. Giocastro, L. Mondello, Current state of comprehensive two-dimensional gas chromatography-mass spectrometry with focus on processes of ionization, *TrAC Trends Anal. Chem.* 105 (2018) 360–366. https://doi.org/10.1016/j.trac.2018.05.016.

[24] L. Mondello, A. Casilli, P.Q. Tranchida, G. Dugo, P. Dugo, Comprehensive two-dimensional gas chromatography in combination with rapid scanning quadrupole mass spectrometry in perfume analysis, *J. Chromatogr. A* 1067 (2005) 235–243. https://doi.org/10.1016/j.chroma.2004.09.040.

[25] W.G. Pool, J.W. de Leeuw, B. van de Graaf, A rapid routine to correct for skewing in gas chromatography/mass spectrometry, *J. Mass Spectrom.* 31 (1996) 213–215. https://doi.org/10.1002/(SICI)1096-9888(199602)31:2<213::AID-JMS284>3.0.CO;2–6.

[26] B.W.J. Pirok, D.R. Stoll, P.J. Schoenmakers, Recent developments in two-dimensional liquid chromatography: Fundamental improvements for practical applications, *Anal. Chem.* 91 (2019) 240–263. https://doi.org/10.1021/acs.analchem.8b04841.

[27] P. Jandera, T. Hájek, P. Česla, Effects of the gradient profile, sample volume and solvent on the separation in very fast gradients, with special attention to the second-dimension gradient in comprehensive two-dimensional liquid chromatography, *J. Chromatogr. A* 1218 (2011) 1995–2006. https://doi.org/10.1016/j.chroma.2010.10.095.

[28] R.J. Vonk, A.F.G. Gargano, E. Davydova, H.L. Dekker, S. Eeltink, L.J. De Koning, P.J. Schoenmakers, Comprehensive two-dimensional liquid chromatography with stationary-phase-assisted modulation coupled to high-resolution mass spectrometry applied to proteome analysis of saccharomyces cerevisiae, *Anal. Chem.* 87 (2015) 5387–5394. https://doi.org/10.1021/acs.analchem.5b00708.

[29] D.R. Stoll, K. Shoykhet, P. Petersson, S. Buckenmaier, Active solvent modulation: A valve-based approach to improve separation compatibility in two-dimensional liquid chromatography, *Anal. Chem.* 89 (2017) 9260–9267. https://doi.org/10.1021/acs.analchem.7b02046.

[30] D.R. Stoll, X. Li, X. Wang, P.W. Carr, S.E.G. Porter, S.C. Rutan, Fast, comprehensive two-dimensional liquid chromatography, *J. Chromatogr. A* 1168 (2007) 3–43. https://doi.org/10.1016/j.chroma.2007.08.054.

[31] S.E.G. Porter, D.R. Stoll, S.C. Rutan, P.W. Carr, J.D. Cohen, Analysis of four-way two-dimensional liquid chromatography-diode array data: Application to metabolomics, *Anal. Chem.* 78 (2006) 5559–5569. https://doi.org/10.1021/ac0606195.

[32] P. Donato, F. Rigano, F. Cacciola, M. Schure, S. Farnetti, M. Russo, P. Dugo, L. Mondello, Comprehensive two-dimensional liquid chromatography–tandem mass spectrometry for the simultaneous determination of wine polyphenols and target contaminants, *J. Chromatogr. A* 1458 (2016) 54–62. https://doi.org/10.1016/j.chroma.2016.06.042.

[33] M. Navarro-Reig, J. Jaumot, A. Baglai, G. Vivó-Truyols, P.J. Schoenmakers, R. Tauler, Untargeted comprehensive two-dimensional liquid chromatography coupled with high-resolution mass spectrometry analysis of rice metabolome using multivariate curve resolution, *Anal. Chem.* 89 (2017) 7675–7683. https://doi.org/10.1021/acs.analchem.7b01648.

[34] R. Karongo, T. Ikegami, D.R. Stoll, M. Lämmerhofer, A selective comprehensive reversed-phase × reversed-phase 2D-liquid chromatography approach with multiple complementary detectors as advanced generic method for the quality control of synthetic and therapeutic peptides, *J. Chromatogr. A* 1627 (2020) 461430. https://doi.org/10.1016/j.chroma.2020.461430.

[35] C.G. Enke, A predictive model for matrix and analyte effects in electrospray ionization of singly-charged ionic analytes, *Anal. Chem*. 69 (1997) 4885–4893. https://doi.org/10.1021/ac970095w.

[36] K.M. Pierce, J.L. Hope, K.J. Johnson, B.W. Wright, R.E. Synovec, Classification of gasoline data obtained by gas chromatography using a piecewise alignment algorithm combined with feature selection and principal component analysis, *J. Chromatogr. A* 1096 (2005) 101–110. https://doi.org/10.1016/j.chroma.2005.04.078.

[37] D. Zhang, X. Huang, F.E. Regnier, M. Zhang, Two-dimensional correlation optimized warping algorithm for aligning GCxGC-MS data, *Anal. Chem*. 80 (2008) 2664–2671. https://doi.org/10.1021/ac7024317.

[38] J. Vial, H. Noçairi, P. Sassiat, S. Mallipatu, G. Cognon, D. Thiébaut, B. Teillet, D.N. Rutledge, Combination of dynamic time warping and multivariate analysis for the comparison of comprehensive two-dimensional gas chromatograms: Application to plant extracts, *J. Chromatogr. A* 1216 (2009) 2866–2872. https://doi.org/10.1016/j.chroma.2008.09.027.

[39] B. Wang, A. Fang, J. Heim, B. Bogdanov, S. Pugh, M. Libardoni, X. Zhang, DISCO: Distance and spectrum correlation optimization alignment for two-dimensional gas chromatography time-of-flight mass spectrometry-based metabolomics, *Anal. Chem*. 82 (2010) 5069–5081. https://doi.org/10.1021/ac100064b.

[40] R.C. Allen, S.C. Rutan, Semi-automated alignment and quantification of peaks using parallel factor analysis for comprehensive two-dimensional liquid chromatography-diode array detector data sets, *Anal. Chim. Acta* 723 (2012) 7–17. https://doi.org/10.1016/j.aca.2012.02.019.

[41] P.E. Sudol, D.V. Gough, S.E. Prebihalo, R.E. Synovec, Impact of data bin size on the classification of diesel fuels using comprehensive two-dimensional gas chromatography with principal component analysis, *Talanta* 206 (2020) 120239. https://doi.org/10.1016/j.talanta.2019.120239.

[42] E. Gorrochategui, J. Jaumot, S. Lacorte, R. Tauler, Data analysis strategies for targeted and untargeted LC-MS metabolomic studies: Overview and workflow, *TrAC Trends Anal. Chem*. 82 (2016) 425–442. https://doi.org/10.1016/j.trac.2016.07.004.

[43] M.M. Sinanian, D.W. Cook, S.C. Rutan, D.S. Wijesinghe, Multivariate curve resolution-alternating least squares analysis of high-resolution liquid chromatography-mass spectrometry data, *Anal. Chem*. 88 (2016) 11092–11099. https://doi.org/10.1021/acs.analchem.6b03116.

[44] R. Stolt, R.J.O. Torgrip, J. Lindberg, L. Csenki, J. Kolmert, I. Schuppe-Koistinen, S.P. Jacobsson, Second-order peak detection for multicomponent high-resolution LC/MS data, *Anal. Chem*. 78 (2006) 975–983. https://doi.org/10.1021/ac050980b.

[45] R. Tauler, E. Gorrochategui, J. Jaumot, R. Tauler, A protocol for LC-MS metabolomic data processing using chemometric tools, *Protoc. Exch*. (2015) 1–41. https://doi.org/10.1038/protex.2015.102.

[46] M. Pérez-Cova, C. Bedia, D.R. Stoll, R. Tauler, J. Jaumot, MSroi: A pre-processing tool for mass spectrometry-based studies, *Chemom. Intell. Lab. Syst*. 215 (2021) 104333. https://doi.org/10.1016/j.chemolab.2021.104333.

[47] O. Amador-Muñoz, P.J. Marriott, Quantification in comprehensive two-dimensional gas chromatography and a model of quantification based on selected summed modulated peaks, *J. Chromatogr. A* 1184 (2008) 323–340. https://doi.org/10.1016/j.chroma.2007.10.041.

[48] D.W. Cook, M.L. Burnham, D.C. Harmes, D.R. Stoll, S.C. Rutan, Comparison of multivariate curve resolution strategies in quantitative LC x LC: Application to the quantification of furanocoumarins in apiaceous vegetables, *Anal. Chim. Acta* 961 (2017) 49–58. https://doi.org/10.1016/j.aca.2017.01.047.
[49] 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.
[50] S.C. Rutan, A. de Juan, R. Tauler, Introduction to multivariate curve resolution, in: S.D. Brown, R. Tauler, B. Walczak (Eds.), *Compr. Chemom*., Vol. 2, Elsevier, Oxford, 2009: pp. 249–259.
[51] A. de Juan, J. Jaumot, R. Tauler, Multivariate curve resolution (MCR): Solving the mixture analysis problem, *Anal. Methods* 6 (2014) 4964–4976. https://doi.org/10.1039/C4AY00571F.
[52] H. Parastar, J.R. Radović, M. Jalali-Heravi, S. Diez, J.M. Bayona, R. Tauler, Resolution and quantification of complex mixtures of polycyclic aromatic hydrocarbons in heavy fuel oil sample by means of GC × GC-TOFMS combined to multivariate curve resolution, *Anal. Chem*. 83 (2011) 9289–9297. https://doi.org/10.1021/ac201799r.
[53] R. Tauler, Calculation of maximum and minimum band boundaries of feasible solutions for species profiles obtained by multivariate curve resolution, *J. Chemom*. 15 (2001) 627–646. https://doi.org/10.1002/cem.654.
[54] J. Jaumot, R. Tauler, MCR-BANDS: a user friendly MATLAB program for the evaluation of rotation ambiguities in Multivariate Curve Resolution, *Chemom. Intell. Lab. Syst*. 103 (2010) 96–107. https://doi.org/10.1016/j.chemolab.2010.05.020.
[55] W. Windig, J. Guilment, Interactive self-modeling mixture analysis, *Anal. Chem*. 63 (1991) 1425–1432.
[56] F. Cuesta Sánchez, J. Toft, B. Van den Bogaert, D.L. Massart, Orthogonal projection approach applied to peak purity assessment, *Anal. Chem*. 68 (1996) 79–85. https://doi.org/10.1021/ac950496g.
[57] E.R. Malinowski, Obtaining the key set of typical vectors by factor analysis and subsequent isolation of component spectra, *Anal. Chim. Acta* 134 (1982) 129–137. https://doi.org/10.1016/S0003-2670(01)84184-2.
[58] K.J. Schostack, E.R. Malinowski, Investigation of window factor analysis and matrix regression analysis in chromatography, *Chemom. Intell. Lab. Syst*. 20 (1993) 173–182. https://doi.org/10.1016/0169-7439(93)80013-8.
[59] H.P. Bailey, S.C. Rutan, Chemometric resolution and quantification of four-way data arising from comprehensive 2D-LC-DAD analysis of human urine, *Chemom. Intell. Lab. Syst*. 106 (2011) 131–141. https://doi.org/10.1016/j.chemolab.2010.07.008.
[60] D.W. Cook, S.C. Rutan, D.R. Stoll, P.W. Carr, Two dimensional assisted liquid chromatography—a chemometric approach to improve accuracy and precision of quantitation in liquid chromatography using 2D separation, dual detectors, and multivariate curve resolution, *Anal. Chim. Acta* 859 (2015) 87–95. https://doi.org/10.1016/j.aca.2014.12.009.
[61] H.P. Bailey, S.C. Rutan, P.W. Carr, Factors that affect quantification of diode array data in comprehensive two-dimensional liquid chromatography using chemometric data analysis, *J. Chromatogr. A* 1218 (2011) 8411–8422. https://doi.org/10.1016/j.chroma.2011.09.057.
[62] M. Pérez-Cova, R. Tauler, J. Jaumot, Chemometrics in comprehensive two-dimensional liquid chromatography: A study of the data structure and its multilinear behavior, *Chemom. Intell. Lab. Syst*. 201 (2020). https://doi.org/10.1016/j.chemolab.2020.104009.
[63] M. Navarro-Reig, J. Jaumot, T.A. van Beek, G. Vivó-Truyols, R. Tauler, Chemometric analysis of comprehensive LC × LC-MS data: Resolution of triacylglycerol structural isomers in corn oil, *Talanta* 160 (2016) 624–635. https://doi.org/10.1016/j.talanta.2016.08.005.

[64] R. Bro, PARAFAC. Tutorial and applications, *Chemom. Intell. Lab. Syst.* 38 (1997) 149–171. https://doi.org/10.1016/S0169-7439(97)00032-4.
[65] J.C. Hoggard, R.E. Synovec, Parallel factor analysis (PARAFAC) of target analytes in GC × GC-TOFMS data: Automated selection of a model with an appropriate number of factors, *Anal. Chem.* 79 (2007) 1611–1619. https://doi.org/10.1021/ac061710b.
[66] R.A. Harshman, M.E. Lundy, PARAFAC: Parallel factor analysis, *Comput. Stat. Data Anal.* 18 (1994) 39–72. https://doi.org/10.1016/0167-9473(94)90132-5.
[67] A.E. Sinha, J.L. Hope, B.J. Prazen, E.J. Nilsson, R.M. Jack, R.E. Synovec, Algorithm for locating analytes of interest based on mass spectral similarity in GC × GC–TOF-MS data: Analysis of metabolites in human infant urine, *J. Chromatogr. A* 1058 (2004) 209–215. https://doi.org/10.1016/j.chroma.2004.08.064.
[68] R.C. Allen, S.C. Rutan, Investigation of interpolation techniques for the reconstruction of the first dimension of comprehensive two-dimensional liquid chromatography-diode array detector data, *Anal. Chim. Acta* 705 (2011) 253–260. https://doi.org/10.1016/j.aca.2011.06.022.
[69] Y. Izadmanesh, E. Garreta-Lara, J.B. Ghasemi, S. Lacorte, V. Matamoros, R. Tauler, Chemometric analysis of comprehensive two dimensional gas chromatography–mass spectrometry metabolomics data, *J. Chromatogr. A* 1488 (2017) 113–125. https://doi.org/10.1016/j.chroma.2017.01.052.
[70] S. Schöneich, D.V. Gough, T.J. Trinklein, R.E. Synovec, Dynamic pressure gradient modulation for comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry detection, *J. Chromatogr. A* 1620 (2020) 460982. https://doi.org/10.1016/j.chroma.2020.460982.
[71] T. Skov, J.C. Hoggard, R. Bro, R.E. Synovec, Handling within run retention time shifts in two-dimensional chromatography data using shift correction and modeling, *J. Chromatogr. A* 1216 (2009) 4020–4029. https://doi.org/10.1016/j.chroma.2009.02.049.
[72] J.M. Amigo, T. Skov, R. Bro, J. Coello, S. Maspoch, Solving GC-MS problems with PARAFAC2, *TrAC Trends Anal. Chem.* 27 (2008) 714–725. https://doi.org/10.1016/j.trac.2008.05.011.
[73] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, *Chemom. Intell. Lab. Syst.* 2 (1987) 37–52. https://doi.org/10.1016/0169-7439(87)80084-9.
[74] D.A. Jackson, Stopping rules in principal components analysis: A comparison of heuristical and statistical approaches, *Ecology* 74 (1993) 2204–2214. https://doi.org/10.2307/1939574.
[75] R. Henrion, N-way principal component analysis theory, algorithms and applications, *Chemom. Intell. Lab. Syst.* 25 (1994) 1–23. https://doi.org/10.1016/0169-7439(93)E0086-J.
[76] R. De Maesschalck, D. Jouan-Rimbaud, D.L. Massart, The Mahalanobis distance, *Chemom. Intell. Lab. Syst.* 50 (2000) 1–18. https://doi.org/10.1016/S0169-7439(99)00047-7.
[77] N.A. Sinkov, J.J. Harynuk, Cluster resolution: A metric for automated, objective and optimized feature selection in chemometric modeling, *Talanta* 83 (2011) 1079–1087. https://doi.org/10.1016/j.talanta.2010.10.025.
[78] B. Worley, S. Halouska, R. Powers, Utilities for quantifying separation in PCA/PLS-DA scores plots, *Anal. Biochem.* 433 (2013) 102–104. https://doi.org/10.1016/j.ab.2012.10.011.
[79] G. Malmquist, R. Danielsson, Alignment of chromatographic profiles for principal component analysis: A prerequisite for fingerprinting methods, *J. Chromatogr. A* 687 (1994) 71–88. https://doi.org/10.1016/0021-9673(94)00726-8.
[80] G.L. Alexandrino, F. Augusto, Comprehensive two-dimensional gas chromatography-mass spectrometry/selected ion monitoring (GC × GC-MS/SIM) and chemometrics to enhance inter-reservoir geochemical features of crude oils, *Energy Fuels* 32 (2018) 8017–8023. https://doi.org/10.1021/acs.energyfuels.8b00230.

[81] F. Murtagh, P. Contreras, Algorithms for hierarchical clustering: An overview, *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* 2 (2012) 86–97. https://doi.org/10.1002/widm.53.
[82] M. Charrad, N. Ghazzali, V. Boiteau, A. Niknafs, Nbclust: An R package for determining the relevant number of clusters in a data set, *J. Stat. Softw.* 61 (2014) 1–36. https://doi.org/10.18637/jss.v061.i06.
[83] J.H. Ward, Hierarchical grouping to optimize and objective function, *J. Am. Stat. Assoc.* 58 (1963) 236–244. https://doi.org/10.1198/016214503000000468.
[84] L.M. Dubois, K.A. Perrault, P.-H. Stefanuto, S. Koschinski, M. Edwards, L. McGregor, J.-F. Focant, Thermal desorption comprehensive two-dimensional gas chromatography coupled to variable-energy electron ionization time-of-flight mass spectrometry for monitoring subtle changes in volatile organic compound profiles of human blood, *J. Chromatogr. A* 1501 (2017) 117–127. https://doi.org/10.1016/j.chroma.2017.04.026.
[85] J.B. MacQueen, Some methods for classification and analysis of multivariate observations, in: L.M. Le Cam, J. Neyman (Eds.), *Proc. Fifth Berkeley Symp. Math. Stat. Probab.*, University of California Press, Berkeley, 1967: pp. 281–297.
[86] A.K. Jain, Data clustering: 50 years beyond K-means, *Pattern Recognit. Lett.* 31 (2010) 651–666. https://doi.org/10.1016/j.patrec.2009.09.011.
[87] D. Arthur, S. Vassilvitskii, k-means++: The advantages of careful seeding, in: *Proc. Eighteenth Annu. ACM-SIAM Symp. Discret. Algorithms*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 2007: pp. 1027–1035.
[88] P.J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, *J. Comput. Appl. Math.* 20 (1987) 53–65. https://doi.org/10.1016/0377-0427(87)90125-7.
[89] D.L. Davies, D.W. Bouldin, A cluster separation measure, *IEEE Trans. Pattern Anal. Mach. Intell. PAMI-1* (1979) 224–227. https://doi.org/10.1109/TPAMI.1979.4766909.
[90] K.A. Favela, M.J. Hartnett, J.A. Janssen, D.W. Vickers, A.J. Schaub, H.A. Spidle, K.S. Pickens, Nontargeted analysis of face masks: Comparison of manual curation to automated GCxGC processing tools, *J. Am. Soc. Mass Spectrom.* 32 (2021) 860–871. https://doi.org/10.1021/jasms.0c00318.
[91] C.N. Cain, P.E. Sudol, K.L. Berrier, R.E. Synovec, Development of variance rank initiated-unsupervised sample indexing for gas chromatography-mass spectrometry analysis, *Talanta* 233 (2021) 122495. https://doi.org/10.1016/j.talanta.2021.122495.
[92] H.D. Bean, J.E. Hill, J.M.D. Dimandja, Improving the quality of biomarker candidates in untargeted metabolomics via peak table-based alignment of comprehensive two-dimensional gas chromatography-mass spectrometry data, *J. Chromatogr. A* 1394 (2015) 111–117. https://doi.org/10.1016/j.chroma.2015.03.001.
[93] P.H. Stefanuto, K.A. Perrault, L.M. Dubois, B. L'Homme, C. Allen, C. Loughnane, N. Ochiai, J.F. Focant, Advanced method optimization for volatile aroma profiling of beer using two-dimensional gas chromatography time-of-flight mass spectrometry, *J. Chromatogr. A* 1507 (2017) 45–52. https://doi.org/10.1016/j.chroma.2017.05.064.
[94] F. Magagna, A. Guglielmetti, E. Liberto, S.E. Reichenbach, E. Allegrucci, G. Gobino, C. Bicchi, C. Cordero, Comprehensive chemical fingerprinting of high-quality cocoa at early stages of processing: Effectiveness of combined untargeted and targeted approaches for classification and discrimination, *J. Agric. Food Chem.* 65 (2017) 6329–6341. https://doi.org/10.1021/acs.jafc.7b02167.
[95] K.J. Johnson, R.E. Synovec, Pattern recognition of jet fuels: Comprehensive GC × GC with ANOVA-based feature selection and principal component analysis, *Chemom. Intell. Lab. Syst.* 60 (2002) 225–237. https://doi.org/10.1016/S0169-7439(01)00198-8.
[96] K.M. Pierce, J.C. Hoggard, J.L. Hope, P.M. Rainey, A.N. Hoofnagle, R.M. Jack, B.W. Wright, R.E. Synovec, Fisher ratio method applied to third-order separation data to identify significant chemical components of metabolite extracts, *Anal. Chem.* 78 (2006) 5068–5075. https://doi.org/10.1021/ac0602625.

[97] R.E. Mohler, K.M. Dombek, J.C. Hoggard, K.M. Pierce, E.T. Young, R.E. Synovec, Comprehensive analysis of yeast metabolite GC × GC-TOFMS data: Combining discovery-mode and deconvolution chemometric software, *Analyst* 132 (2007) 756–767. https://doi.org/10.1039/b700061h.

[98] H.P. Bailey, S.C. Rutan, Comparison of chemometric methods for the screening of comprehensive two-dimensional liquid chromatographic analysis of wine, *Anal. Chim. Acta* 770 (2013) 18–28. https://doi.org/10.1016/j.aca.2013.01.062.

[99] L.C. Marney, W. Christopher Siegler, B.A. Parsons, J.C. Hoggard, B.W. Wright, R.E. Synovec, Tile-based Fisher-ratio software for improved feature selection analysis of comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry data, *Talanta* 115 (2013) 887–895. https://doi.org/10.1016/j.talanta.2013.06.038.

[100] B.A. Parsons, L.C. Marney, W.C. Siegler, J.C. Hoggard, B.W. Wright, R.E. Synovec, Tile-based fisher ratio analysis of comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry (GC × GC–TOFMS) data using a null distribution approach, *Anal. Chem.* 87 (2015) 3812–3819. https://doi.org/10.1021/ac504472s.

[101] P.E. Sudol, G.S. Ochoa, R.E. Synovec, Investigation of the limit of discovery using tile-based Fisher ratio analysis with comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry, *J. Chromatogr. A* 1644 (2021). https://doi.org/10.1016/j.chroma.2021.462092.

[102] B.C. Reaser, B.W. Wright, R.E. Synovec, Using receiver operating characteristic curves to optimize discovery-based software with comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry, *Anal. Chem.* 89 (2017) 3606–3612. https://doi.org/10.1021/acs.analchem.6b04991.

[103] S.E. Prebihalo, G.S. Ochoa, K.L. Berrier, K.J. Skogerboe, K.L. Cameron, J.R. Trump, S.J. Svoboda, J.K. Wickiser, R.E. Synovec, Control-normalized fisher ratio analysis of comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry data for enhanced biomarker discovery in a metabolomic study of orthopedic knee-ligament injury, *Anal. Chem.* 92 (2020) 15526–15533. https://doi.org/10.1021/acs.analchem.0c03456.

[104] N.E. Watson, B.A. Parsons, R.E. Synovec, Performance evaluation of tile-based Fisher Ratio analysis using a benchmark yeast metabolome dataset, *J. Chromatogr. A* 1459 (2016) 101–111. https://doi.org/10.1016/j.chroma.2016.06.067.

[105] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: A basic tool of chemometrics, *Chemom. Intell. Lab. Syst.* 58 (2001) 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1.

[106] B. Kehimkar, J.C. Hoggard, L.C. Marney, M.C. Billingsley, C.G. Fraga, T.J. Bruno, R.E. Synovec, Correlation of rocket propulsion fuel properties with chemical composition using comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry followed by partial least squares regression analysis, *J. Chromatogr. A* 1327 (2014) 132–140. https://doi.org/10.1016/j.chroma.2013.12.060.

[107] K.L. Berrier, C.E. Freye, M.C. Billingsley, R.E. Synovec, Predictive modeling of aerospace fuel properties using comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry and partial least squares analysis, *Energy Fuels* 34 (2020) 4084–4094. https://doi.org/10.1021/acs.energyfuels.9b04108.

[108] V. Abrahamsson, N. Ristic, K. Franz, K. Van Geem, Comprehensive two-dimensional gas chromatography in combination with pixel-based analysis for fouling tendency prediction, *J. Chromatogr. A* 1501 (2017) 89–98. https://doi.org/10.1016/j.chroma.2017.04.021.

[109] V. Consonni, G. Baccolo, F. Gosetti, R. Todeschini, D. Ballabio, A MATLAB toolbox for multivariate regression coupled with variable selection, *Chemom. Intell. Lab. Syst.* 213 (2021) 104313. https://doi.org/10.1016/j.chemolab.2021.104313.

[110] D. Ballabio, V. Consonni, Classification tools in chemistry. Part 1: Linear models. PLS-DA, *Anal. Methods* 5 (2013) 3790–3798. https://doi.org/10.1039/c3ay40582f.

[111] P.S. Gromski, H. Muhamadali, D.I. Ellis, Y. Xu, E. Correa, M.L. Turner, R. Goodacre, A tutorial review: Metabolomics and partial least squares-discriminant analysis—a marriage of convenience or a shotgun wedding, *Anal. Chim. Acta* 879 (2015) 10–23. https://doi.org/10.1016/j.aca.2015.02.012.

[112] L.C. Lee, C.Y. Liong, A.A. Jemain, Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data: A review of contemporary practice strategies and knowledge gaps, *Analyst* 143 (2018) 3526–3539. https://doi.org/10.1039/c8an00599k.

[113] C. Quiroz-Moreno, M.F. Furlan, J.R. Belinato, F. Augusto, G.L. Alexandrino, N.G.S. Mogollón, RGCxGC toolbox: An R-package for data processing in comprehensive two-dimensional gas chromatography-mass spectrometry, *Microchem. J*. 156 (2020) 104830. https://doi.org/10.1016/j.microc.2020.104830.

[114] S. Wang, L. Qiao, X. Shi, C. Hu, H. Kong, G. Xu, On-line stop-flow two-dimensional liquid chromatography-mass spectrometry method for the separation and identification of triterpenoid saponins from ginseng extract, *Anal. Bioanal. Chem*. 407 (2015) 331–341. https://doi.org/10.1007/s00216-014-8219-4.

[115] L.A. Adutwum, J.K. Kwao, J.J. Harynuk, Unique ion filter—A data reduction tool for chemometric analysis of raw comprehensive two-dimensional gas chromatography-mass spectrometry data, *J. Sep. Sci*. 44 (2021) 2773–2784. https://doi.org/10.1002/jssc.202001127.

[116] R.G. Brereton, G.R. Lloyd, Partial least squares discriminant analysis: Taking the magic away, *J. Chemom*. 28 (2014) 213–225. https://doi.org/10.1002/cem.2609.

[117] A.C. Paiva, L.W. Hantao, Exploring a public database to evaluate consumer preference and aroma profile of lager beers by comprehensive two-dimensional gas chromatography and partial least squares regression discriminant analysis, *J. Chromatogr. A* 1630 (2020) 461529. https://doi.org/10.1016/j.chroma.2020.461529.

[118] K.K. Pasikanti, K. Esuvaranathan, Y. Hong, P.C. Ho, R. Mahendran, L. Raman Nee Mani, E. Chiong, E.C.Y. Chan, Urinary metabotyping of bladder cancer using two-dimensional gas chromatography time-of-flight mass spectrometry, *J. Proteome Res*. 12 (2013) 3865–3873. https://doi.org/10.1021/pr4000448.

[119] N. Di Giovanni, M.A. Meuwis, E. Louis, J.F. Focant, Untargeted serum metabolic profiling by comprehensive two-dimensional gas chromatography-high-resolution time-of-flight mass spectrometry, *J. Proteome Res*. 19 (2020) 1013–1028. https://doi.org/10.1021/acs.jproteome.9b00535.

[120] T. Rajalahti, R. Arneberg, F.S. Berven, K.M. Myhr, R.J. Ulvik, O.M. Kvalheim, Biomarker discovery in mass spectral profiles by means of selectivity ratio plot, *Chemom. Intell. Lab. Syst*. 95 (2009) 35–48. https://doi.org/10.1016/j.chemolab.2008.08.004.

[121] I.G. Chong, C.H. Jun, Performance of some variable selection methods when multicollinearity is present, *Chemom. Intell. Lab. Syst*. 78 (2005) 103–112. https://doi.org/10.1016/j.chemolab.2004.12.011.

[122] J.R. Montenegro-Burke, A.E. Aisporna, H.P. Benton, D. Rinehar, M. Fang, T. Huan, B. Warth, E. Warth, B.T. Abe, J. Ivanisevic, D.W. Wolan, L. Teyton, L. Lairson, G. Siuzdak, Data streaming for metabolomics: Accelerating data processing and analysis from days to minutes, *Anal. Chem*. 89 (2017) 1254–1259. https://doi.org/10.1021/acs.analchem.6b03890.

[123] R. Houhou, T. Bocklitz, Trends in artificial intelligence, machine learning, and chemometrics applied to chemical data, *Anal. Sci. Adv*. 2 (2021) 128–141. https://doi.org/10.1002/ansa.202000162.

[124] E.B. Ledford, C.A. Billesbach, Q. Zhu, GC3: Comprehensive three-dimensional gas chromatography, *J. High Resolut. Chromatogr*. 23 (2000) 205–207. https://doi.org/10.1002/(SICI)1521-4168(20000301)23:3<205::AID-JHRC205>3.0.CO;2-U.

[125] R. Edam, J. Blomberg, H.G. Janssen, P.J. Schoenmakers, Comprehensive multidimensional chromatographic studies on the separation of saturated hydrocarbon ring structures in petrochemical samples, *J. Chromatogr. A* 1086 (2005) 12–20. https://doi.org/10.1016/j.chroma.2005.02.048.

[126] M. Zoccali, P.Q. Tranchida, L. Mondello, On-line combination of high performance liquid chromatography with comprehensive two-dimensional gas chromatography-triple quadrupole mass spectrometry: A proof of principle study, *Anal. Chem.* 87 (2015) 1911–1918. https://doi.org/10.1021/ac504162a.

[127] N.E. Watson, H.D. Bahaghighat, K. Cui, R.E. Synovec, Comprehensive three-dimensional gas chromatography with time-of-flight mass spectrometry, *Anal. Chem*. 89 (2017) 1793–1800. https://doi.org/10.1021/acs.analchem.6b04112.

[128] T.J. Trinklein, S. Schöneich, P.E. Sudol, C.G. Warren, D.V. Gough, R.E. Synovec, Total-transfer comprehensive three-dimensional gas chromatography with time-of-flight mass spectrometry, *J. Chromatogr. A* 1634 (2020) 461654. https://doi.org/10.1016/j.chroma.2020.461654.

[129] N.E. Watson, W.C. Siegler, J.C. Hoggard, R.E. Synovec, Comprehensive three-dimensional gas chromatography with parallel factor analysis, *Anal. Chem*. 79 (2007) 8270–8280. https://doi.org/10.1021/ac070829x.

[130] N.E. Watson, S.E. Prebihalo, R.E. Synovec, Targeted analyte deconvolution and identification by four-way parallel factor analysis using three-dimensional gas chromatography with mass spectrometry data, *Anal. Chim. Acta* 983 (2017) 67–75. https://doi.org/10.1016/j.aca.2017.06.017.

[131] T.J. Trinklein, S.E. Prebihalo, C.G. Warren, G.S. Ochoa, R.E. Synovec, Discovery-based analysis and quantification for comprehensive three-dimensional gas chromatography flame ionization detection data, *J. Chromatogr. A* 1623 (2020) 461190. https://doi.org/10.1016/j.chroma.2020.461190.

Index

Note: Page numbers in *italics* indicate a figure and page numbers in **bold** indicate a table on the corresponding page

D

E

F

G

H

I

J

K

L

M

N

O

P

S

T