Technologies to accelerate protein purification process development

Hanke, Alex

DOI
10.4233/uuid:0dc8a46c-1963-4f53-a3b3-6d1dc67202c7

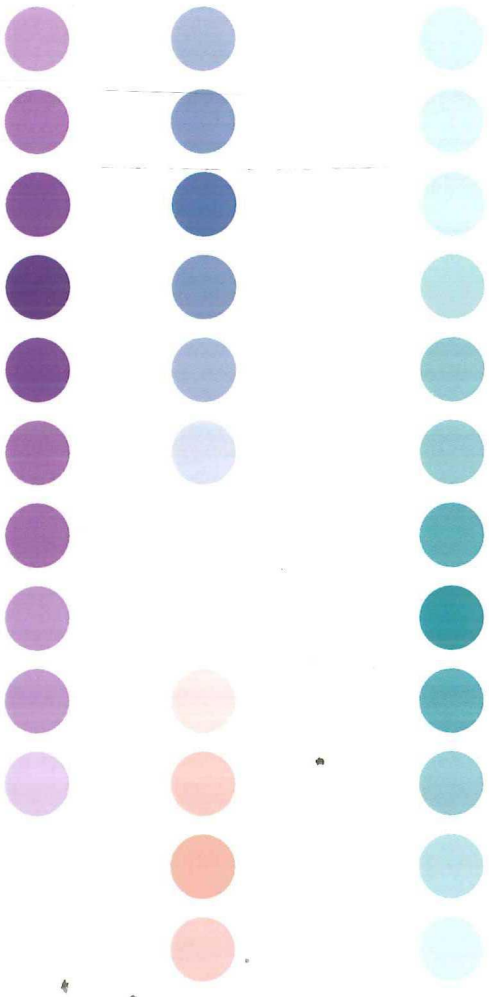
Publication date
2016

Document Version
Final published version

Citation (APA)
Hanke, A. (2016). *Technologies to accelerate protein purification process development*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:0dc8a46c-1963-4f53-a3b3-6d1dc67202c7

Technologies to accelerate protein purification process development

Alexander T. Hanke



Propositions accompanying the thesis

"Technologies to accelerate protein purification process development"

by Alexander Thomas Hanke

1) All modern process development strategies in the biopharmaceutical field, including those based on mechanistic modelling, are trial and error based. (**Chapter 1**)

2) The amount of information that can be drawn from a chromatogram strictly depends on the available prior knowledge. (**Chapter 2**)

3) The greatest challenge of working in a high-throughput laboratory is not to interpret the results, but to deal with technical failures. (**Chapter 3&4**)

4) A predictive model that requires training will never fail, but never really work either. (**Chapter 5**)

5) As soon as your project description mentions deliverables, you're an engineer and not a scientist.

6) When talking about to the production of biopharmaceuticals, the use of the terms 'up- and downstream' in regard to the process stage would make more sense when applied the other way around.

7) Instead of arguing about the problems with traditional scientific publishing, the scientific community should adopt a wiki-style knowledge sharing system.

8) The main purpose of the education system is neither to pass on knowledge nor teach skills.

9) When something is not working, trying harder is always the wrong solution.

10) Universities benefit from creating frustrating conditions for their staff and students.

These propositions are regarded as opposable and defendable, and have been approved as such by the promotor Prof. dr. ir. L.A.M. van der Wielen

Technologies to accelerate protein purification process development

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof. ir. K.C.A.M. Luyben,
voorzitter van het College voor Promoties,

in het openbaar te verdedigen op
maandag 21 maart 2016 om 12:30 uur

door

Alexander Thomas HANKE

Diplom Ingenieur im Bioingenieurwesen
Karlsruhe Institut für Technologie, Duitsland
geboren te Heidelberg, Duitsland

This dissertation has been approved by the promotor:

Prof. dr. ir. L.A.M. van der Wielen and

copromotor: Dr. ir. M. Ottens

Composition of the doctoral committee:

Rector Magnificus,	chairman
Prof. dr. L.A.M. van der Wielen	promotor
Dr. Marcel Ottens	copromotor
Prof. dr. M.H.M. Eppink	Wageningen University Synthon Biopharmaceuticals B.V.

Independent members:

Prof. dr. G.J. Witkamp	Delft University of Technology
Prof. dr. W.R. Hagen	Delft University of Technology
Prof. dr. J.J. Hubbuch	Karlsruhe Institute of Technology, Germany
Prof. dr. H.J. Noorman	DSM Biotechnology Center Delft University of Technology

The research described in this thesis was performed at the Department of Biotechnology, Delft University of Technology, The Netherlands. The research was financially supported by the BE-BASIC foundation, a public private partnership of knowledge institutes, industry and academia.

ISBN: 978-94-6186-544-1

Contents

List of Abbreviations

2D	Two-dimensional
3D	Three-dimensional
AEX	Anion exchange
AS	Ammonium sulphate
ASP	Average surface property
ATPS	Aqueous two phase systems
CEX	Cation exchange
CGE	Capillary gel electrophoresis
CHO	Chinese hamster ovary
CPP	Critical process parameters
CQA	Critical quality attributes
CT	Central tendency
CV	Column volume
DoE	Design of experiments
DRT	Dimensionless retention time
ELSIA	Enzyme-linked immunosorbent assay
EMG	Exponentially modified Gaussian distribution
FDA	Food and drug administration
FFT	Fast Fourier transform
GC	Gas chromatography
gCF	Gradient chromatofocusing
GE	Gel electrophoresis
GRM	General rate model
HCP	Host cell proteins
HIC	Hydrophobic interaction chromatography
HMW	High molecular weight
HTPD	High-throughput process development
HTS	High-throughput screening

IDP	Intrinsically disordered protein
IgG	Immunoglobulin G
IQR	Inter-quartile range
JK	Jackknife
LB	Lower boundary
LC	Liquid chromatography
LF	Lower fence
LMW	Low molecular weight
mAbs	Monoclonal antibodies
MAD	Median absolute dispersion
MALDI	Matrix-assisted laser desorption/ionization
MCR	Multivariate curve resolution
MD	Molecular dynamics
MFAD	Mode fenced average dispersion
MMC	Mixed-mode chromatography
MOAD	Mode absolute dispersion
MOIAD	Mode inter-quartile average dispersion
MS	Mass spectrometry
MSE	Mean square error
NIR	Near infra-red
NMR	Nuclear magnetic resonance
PAGE	Polyacrylamide gel electrophoresis
PAT	Process analytical technology
PCC	Pearson correlation coefficient
PDB	Protein database
PMG	Polynomial modified Gaussian distribution
QbD	Quality by design
QSAR	Quantitative structure activity relationship
QSPR	Quantitative structure property relationship
RI	Refractive index

RPC	Reversed phase chromatography
RSA	Response surface analysis
SASA	Solvent accessible surface area
SEC	Size-exclusion chromatography
SELDI	Surface-enhanced laser desorption/ionization
SMA	Steric mass action
SSC	Separation selection coefficient
SVC	Second virial coefficient
THE	High-throughput experimentation
TM	Trimean
TMAD	Trimean absolute dispersion
TMFAD	Trimean fenced average dispersion
TMIAD	Trimean inter-quartile average dispersion
TOF	Time of flight
UB	Upper boundary
UF	Upper fence
UV	Ultra violet

1

Purifying biopharmaceuticals: knowledge-based chromatographic process development

Abstract

The purification of biopharmaceuticals is commonly considered the bottleneck of the manufacturing process. An increasing product diversity together with growing regulatory and economic constraints raise the need to adopt new rational, systematic and generally applicable process development strategies. Liquid chromatography is the key step in most purification processes and a well understood unit operation, yet this understanding is still rarely effectively utilized during process development. Knowledge of the composition of the mixture, the solutes' molecular properties and how they interact with the resins are required to rationalize the design choices. Here we provide an overview of the advances in the determination and measurement of these properties and interactions, and outline their use throughout the different stages of downstream process development.

Keywords: Process Development; Chromatography; High-throughput; Host Cell Proteins; Mathematical Modelling

Published as: A.T. Hanke and M. Ottens, **Trends Biotechnol**, 32 (2014): 210-220

1.1 Introduction

Biopharmaceuticals have been a major driving force for growth in the pharmaceutical industry in the past years [1]. Over 40% of drugs granted FDA approval in 2012 were biopharmaceuticals, of which therapeutic proteins, including monoclonal antibodies (mAbs), constituted the largest group next to therapeutic peptides [2]. From a manufacturing perspective, the increased product titres achieved over the last decade have long shifted the attention towards the downstream process [3]. Despite increasing competition from non-chromatographic techniques [4], pressure to reduce costs and increase throughput, packed bed chromatography is still the dominant technique in biopharmaceutical purification [5]. This prevalence is mainly due to the high-resolutions that can be achieved even for highly similar components. Advances in resin chemistry have alleviated some of the throughput concerns that packed bed chromatography could handle the production needs in coming years [6]. If not during early process stages, then during product polishing where very high purities are required for therapeutics, it seems unlikely that chromatography will lose its place in biopharmaceutical manufacturing in the close future.

Besides having to handle increasing production volumes, downstream scientists and engineers face a plethora of technical, economic and regulatory challenges. While in the past the dominance of mAbs as a product class allowed to establish platform processes that required relatively minor adaptions from product to product [7], recent trends towards more diverse therapeutic proteins require more generally applicable process development approaches [8]. Increased competition through biosimilars catalysed by abbreviated regulatory pathways have increased the economic pressure on the manufacturing of biotherapeutics [9]. From the regulatory side, the 'Quality by Design' (QbD) and 'Process

Analytical Technology' (PAT) initiatives call for increased process and product understanding to ensure each process consistently meets precisely defined quality attributes [10]. Definition of the critical quality attributes (CQA) and linking them to the underlying critical process parameters (CPP) requires a thorough and systematic characterization of the process parameter space.

From a process development perspective this requirement has rendered trial-and-error based process development and univariate optimization largely obsolete. This has led to the wide-spread adoption of high-throughput screening technologies (HTS) [11]. In this context the relationship between the CPP and CQA is usually of a statistical nature derived from a Response Surface Analysis of a Design of Experiments (DoE) or multivariate data analysis [12,13]. Genetic algorithms are being increasingly employed to identify optima in the design-space [14,15]. The combination of these experimental and data processing techniques has been coined high-throughput process development (HTPD). The statistical relationship allows to rank CPP by significance of impact on the CQA but lacks the ability to predict process performance, limiting their use for process optimization.

The degree to which CPP and CQA can be causally linked reflects the level of process understanding achieved [16,17]. Throughout the biopharmaceutical manufacturing process, both upstream [18], downstream [19] and during formulation [20], there is a trend to gradually replace statistical and empirical correlations with mechanistic models. Mechanistic models typically allow more accurate extrapolation making them very useful as tools for fast and cheap process optimization. As they are derived from fundamental principles, mechanistic models reflect a higher level of process understanding. Most mechanistic models describing chromatographic separations consist of two parts: equations describing the fluid flow and mass-transfer in the column and a

model to describe the interactions between the sample and the resin in the form of adsorption isotherms. Experimental approaches to determine both mass-transfer [21] and resin-interaction parameters [22] have been extensively reviewed, but require very large numbers of experiments to gain the parameters needed to model even simple systems, such as illustrated in Figure 1.1. The availability of these parameters often restricts the use these modelling tools to optimization of very specific separation problems during late stages of process development, when feed compositions are already less complex and many design decisions have already been made.

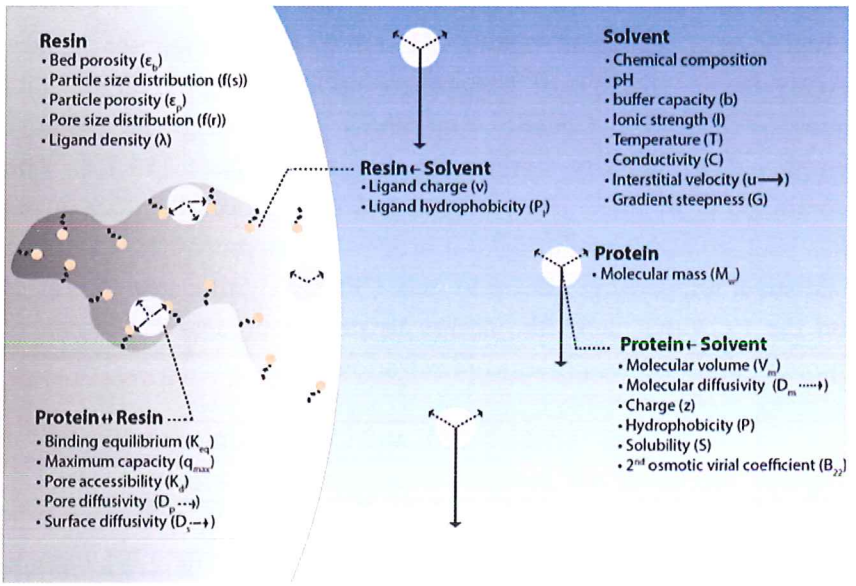


Figure 1.1. Schematic representation of the processes taking place inside a chromatographic column on the microscopic scale and the parameters commonly used to describe protein molecular properties, resin properties and their interactions. Arrows indicate mass-transfer effects, solid arrows convection, dashed arrows diffusion, and the length of each arrow is qualitatively related to the magnitude of the effect. The exact definitions of

To bridge the gap between early stage design choices and advanced stage optimization a variety of process development assisting tools have emerged that utilize experimentally determined or empirically derived component properties. Component properties in this context are considered physico-chemical parameters that describe the components of interest on a macroscopic level without taking resin-specific interactions into account. These parameters are then used to systematically determine a feasible processing pathway, without making quantitative performance predictions of the individual steps.

The approaches discussed so far fundamentally differ in what type of knowledge is generated and how this knowledge is used to make design decisions. The interplay of knowledge generating and process design modules is illustrated in Figure 1.2. Combinations of experimental knowledge generation and model-based process design modules are commonly referred to as hybrid process development approaches and offer many practical advantages compared to purely experimental or model based approaches [23]. Choosing the approach most suitable for a specific project is not always straight forward and depends on the available resources and process development stage. To assist separation scientists and downstream process engineers in choosing which tools to add to their process development toolbox, we review recent developments in the determination of protein properties and resin interaction parameters and how each can be used for rational design decisions.

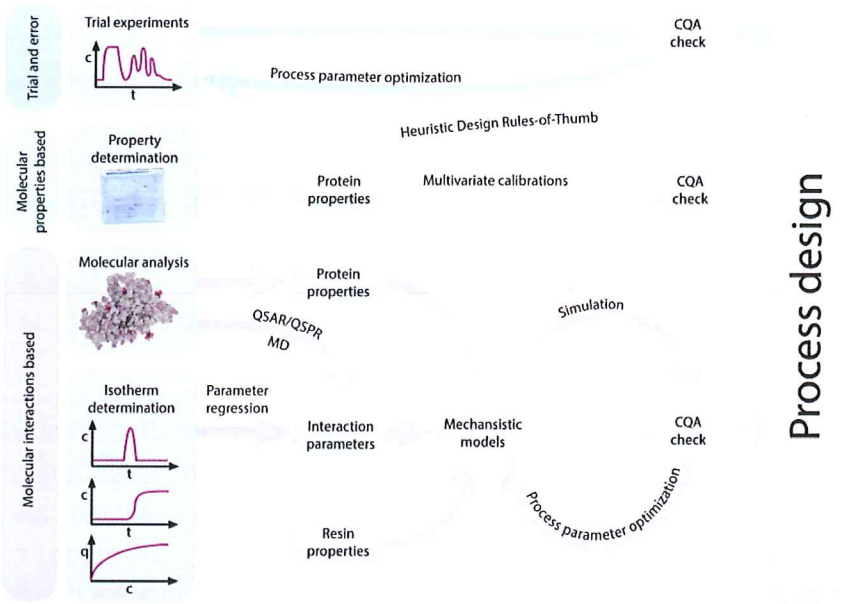


Figure 1.2. A schematic representation of the different downstream process development approaches applied to a single or series of unit operations. The experiments that need to be performed are positioned at the left, generated intermediate data classes in the centre and how they are converted to a process design on the right. Before a design is accepted its ability to meet the CQAs is evaluated. Trial and error process development (yellow) includes what is often referred to as experimental process development and subclasses of this approach can be distinguished by the number of experiments considered and the efficiency of the optimization loop, ranging from univariate optimization to genetic algorithms. The properties based approach (green) most closely corresponds to heuristic process development. The molecular interactions based approach can be viewed as purely model based or an hybrid experimental approach depending on how the data sources are combined. It should be noted that the trial and error and interactions based approach can lead to multiple feasible process options. When multiple processes that comply with CQA requirements are identified, economic performance indicators should be included in the decision process.

Box 1. The Purification Challenge

The purification of a biopharmaceutical is not a simple task. Besides the cell debris, host cell proteins, DNA, endotoxins and viruses that need to be removed, product related impurities such as product moieties that are misfolded, aggregated, carry the wrong post-translational modifications or are otherwise chemically degraded may complicate the purification, due to their high similarity to the target molecule. Achieving the high levels of product purity required for the use as an active pharmaceutical ingredient, requires a complex cascade of unit operations. An example of a relatively simple process is given in Figure 1.3. The complexity of such multi-stage processes poses two major challenges to the downstream process developer. The effects that small variations in the upstream process and the quality of the chemicals and auxiliary materials can have on the performance of unit operations further downstream must be accounted for in the process control strategy. Similarly the dependencies of the single unit operations need to be considered during the earlier stages of process development, when the choice for specific unit operations and their position in the process are decided.

Despite the importance of understanding these dependencies, most process development approaches only consider the optimization of individual unit operations outside of the context of the process in its entirety. In most cases the unit operations are chosen and optimized sequentially in the downstream direction, as the changes in feed composition due to the prior unit operation affect the choice of the subsequent unit operation [24]. In many cases this simplification is necessary, as the exponential increase in possible unit operations sequences with increasing number of unit operations to be considered quickly becomes unmanageable within the timeframe available for process development when time-consuming experimentation is involved. The downside of such a sequential approach is that it poses the risk of missing the global optima by excluding options choosing a worse performing early step followed by a more efficient subsequent step might lead to a more economical process. With increasing availability of powerful supercomputers the time and resource limitation no longer holds for in-silico process

optimization. Given enough computational power all, or a large subset of possible flowsheet options can be optimized and evaluated in parallel.

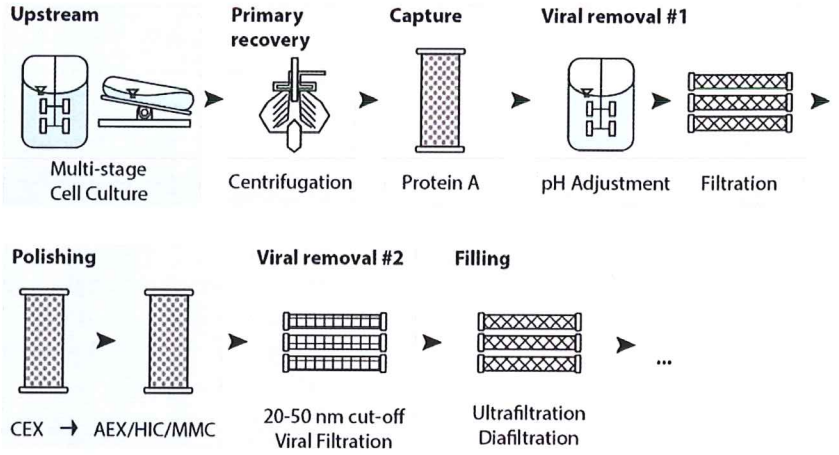


Figure 1.3. Simplified platform process for the production of monoclonal antibodies (MAbs) in bulk form [7]. In reality the cells undergo a series of inoculum steps and seed reactors before reaching the production reactor. The downstream process, usually considered to start after primary recovery, mostly consists of at least two chromatographic separations to reach the desired purity. Regulations require at least two orthogonal methods of virus removal, commonly low pH inactivation and viral filtration. The greatest process diversity lies in the polishing steps where the choice for the combination of Cation Exchange (CEX), Anion Exchange (AEX), Hydrophobic Interaction (HIC) and Mixed-Mode Chromatography (MMC) is based on the characteristics of the impurities to be removed. After rebuffering by diafiltration further processing steps follow for the drug to reach its final formulation.

1.2 Process development based on molecular properties

The thermodynamic interactions and mass-transfer effects occurring during liquid chromatography are very complex in nature.

It has been shown that a proteins shape and heterogeneity of surface characteristics can lead to a non-random distribution of binding orientations during the adsorption process [25]. Nevertheless, Asenjo et al. have demonstrated in numerous cases that fairly simple macroscopic molecular properties can be used to make qualitative predictions towards the efficiency of a separation step [26-28]. The property or combination of properties to consider depends on the principle of separation.

Separation in Size Exclusion Chromatography (SEC) is based on differences in the molecular size, Ion-exchange chromatography (IEC) exploits differences in charge, hydrophobic interaction chromatography (HIC) and Reversed Phase Chromatography (RPC) separate based on surface hydrophobicity and Mixed-Mode Chromatography (MMC) exploit combinations thereof. In all these cases the simplification of the parameters lies in the assumption of uniformity. The probability for the protein's behaviour to deviate from this idealization increases for larger molecules. Within limits such effects can be compensated by considering composite parameters such as the charge density (net charge divided by molecular weight) [27].

1.2.1. Determining protein properties

The key advantage of the simplified macroscopic properties is that they can easily be determined through experiments and do not require knowledge of the respective protein's structure or sequence. An overview of single-parameter experimental techniques is given in Table 1. There are many different experimental techniques available to determine each fundamental parameter. For them to be useable interchangeably in terms of a generalized thermodynamic framework requires careful consideration of the technique specific influence on the parameter [29]. The presented selection includes the solubility and second osmotic virial coefficient. While these parameters are more commonly linked to non-chromatographic operations such as

precipitation and crystallization, they can impact chromatographic separation behaviour when local concentrations become comparatively high e.g. as under overloading conditions, where protein-protein interactions are assumed to make a significant contribution to the free binding energy [30,31].

Many of the experimental techniques mentioned in Table 2, specifically those involving a separation mechanism, can be combined to form multi-dimensional characterization schemes. There are two major advantages to such a setup: combining techniques increases the information gained per volume of sample and the gained increased resolution capacity of applying different, preferably orthogonal, separation techniques allows simultaneous characterization of multiple species directly from a complex sample. This has led to the widespread adoption of such techniques as 2D-gel electrophoresis and 2D-LC-MS in the proteomics field, where the analysis of highly complex samples is a daily occurrence [32].

Table 1.1 Analytical and computational methods to obtain molecular properties of biomolecules

Property	Variable with	Methods to determine	Ref.*
Molecular mass	-	Mass spectrometry, Capillary gel electrophoresis, SDS-PAGE, Primary sequence	
Partial specific volume	Conformation	Size-exclusion chromatography, Analytical centrifugation, Dynamic light scattering, Crystal structure	[33]
Diffusivity	Conformation, Temperature	Dynamic light scattering, Aris-Taylor capillaries Peak parking in non-porous media columns H-cell	[33] [21] [34] [35]
Net Charge	pH	Iso-electric focusing Capillary electrophoresis, Titration, DLVO calculations	[36]
Hydrophobicity	Conformation, Temperature, Solvent chemistry, Ionic strength	Hydrophobic interaction chromatography, Reversed Phase chromatography, Hydrophobic imbalance, Precipitation with ammonium sulphate, Primary sequence, Crystal structure, Aqueous two phase extraction	[37] [38] [39]

Solubility	Conformation, Temperature, pH, Solvent composition, Ionic strength	Addition of Lyophilized Protein, Concentration by Ultrafiltration, Induction of Amorphous Precipitation	[40] [40] [41]
B_{22}/SVC	Conformation, Temperature, pH, Solvent composition, Ionic strength	Static light scattering, Self-interaction chromatography in columns, Self-interaction chromatography in microchannels, Membrane osmometry Sedimentation equilibrium measurements	[42] [43] [44] [45] [42]

*** Unless stated otherwise the technique is described in standard textbooks [46].**

Recent years have seen some of these proteomic principles being applied in a process development context. Glatz et al. characterized a variety of plant-based transgenic feedstocks with a 3D strategy based on partioning in aqueous two-phase systems (ATPS) followed by 2D-GE, thereby providing insight into the proteins' hydrophobicity, charge and size [39,47,48]. Another application of multi-dimensional techniques gaining attention is tracking different HCP levels throughout process development operations. 2D-differential gel-electrophoresis [49] and 2D-PAGE with SELDI-TOF MS [50] have been demonstrated as useful tools to track individual host-cell protein levels throughout various process stages. A variety of immunospecific methods using poly-clonal anti-HCP antibodies including ELISAs western blots have been shown to also be suitable for HCP tracking [51]. These tools could in principle also be combined with many of the parameter determining experiments to create multiplexed variants.

1.2.2. Molecular properties based process development strategies

Due to their relative lack of quantitative performance prediction capability, as compared to approaches applying mechanistic models, the focus of purely property based process development strategies has been on guiding the engineer during early process development decisions, such as choice and sequencing of unit operations. Such guidance can be provided through expert

systems, computerized implementations of design rules-of-thumb that can generate a first draft of process flow sheet [28]. Besides general design rules, such as first removing the contaminants present in greatest concentration, these systems calculate heuristic 'separation selection coefficients (SSC)' from the physicochemical properties governing the unit-operation under consideration. The flow sheet is then generated by choosing the operation that yields the highest SSC. A drawback of such heuristic approaches is that their insufficient investigation of the design-space is expected to lead to suboptimal process efficiency [3]. A second challenge during heuristic flow sheeting arises from the need to predict changes in the stream composition between unit operations. Univariate models predicting retention factors from governing property parameters alone lead to unsatisfactory results [26,28]. To overcome this, the changes in composition can be predicted by mechanistic models such as the general rate model (GRM) [24], this however requires a much more complex dataset.

The lack of purely property based performance prediction capability has recently been addressed through the introduction of a three-dimensional multi-variate calibrations based on the proteins molecular weight, pI and hydrophobicity [52]. Fairly accurate predictions were possible and analysis of parameter correlation allowed gaining some process understanding. Application of the three-dimensional characterization principle coupled to a multi-variate random forest calibration showed mixed results yielding useful predictions only for proteins similar to the calibration set [39].

1.3 Process development based on molecular interactions

Numerous mechanistic models describing the transport phenomena in liquid chromatography in varying detail have been proposed over the years and are extensively discussed elsewhere [53-55]. One of the most comprehensive models that can be efficiently

solved is the general rate model [56,57]. Less complex models such as the Equilibrium Dispersive Model simplify the contribution of non-equilibrium effects while still allowing for a reasonable prediction of system performance [58]. The downside to such simplifications is that the lumped parameters give no insight into the causes of band broadening [21], yet the reduced computational effort is beneficial when very large datasets need to be treated [59], or very large numbers of simulations need to be performed.

1.3.1. Model parameter estimation

As previously discussed, for highly complex biologically produced feedstocks it is often rather the lack of complete parameter sets than models that limit the use of mechanistic modelling during process development. Figure 1.1 illustrates the complexity of a single parameter set required to model the chromatographic behaviour of a single solute in a column packed with a specific resin. The exact definition of each of these parameters may vary depending on the applied model, e.g. when modelling the protein-resin binding the steric mass action (SMA) considers a proteins effective surface instead of the net charge and introduces a steric hindrance factor to account for blocked resin charges [60]. Once a set of suitable models has been chosen, different experimental or computational approaches to gain the required mass-transfer and isotherm parameters can be followed.

To describe the inner-column mass transfer, macroscopic chromatography models require an effective diffusion coefficient, whereas mesoscopic and microscopic models differentiate between pore and surface diffusion. All of these parameters can be scaled to the bulk-diffusion coefficient [21]. Experimental measurement of the effective diffusion coefficient and the scaling factor for a single solute in specific column can be achieved with the Peak Parking method. Peak parking involves the injection of a small defined sample volume and isocratic elution to the half the length of the

column [34]. The external flow is then stopped for a certain parking time, after which the original flow rate is resumed until the sample is completely eluted. Precise analysis of the resulting peak shape allows to regress the effective solute diffusivity in the packed bed. Detailed analysis of intraparticle diffusion fronts through confocal laser scanning microscopy showed that sharp diffusion fronts, as assumed by commonly used uptake models such as the shrinking core model [61], don't hold for non-linear isotherm conditions and surface diffusion becomes increasingly important. On a macroscopic level however, Lenhoff et al. showed that predictions made by a shrinking core model with an adjusted pore diffusivity, derived from confocal laser scanning microscopy experiments, were virtually indistinguishable from uptake predictions of more detailed models and gave a detailed description of how to perform the adjustment [62].

The basic approaches to experimentally determine isotherm parameters have not significantly changed over the last 15 years [22]. In principle any number of existing chromatographic band profiles can be used to gain the parameters necessary to model it through the so called inverse method where the profile is repeatedly simulated by the model with variation of the parameters until an optimal fit between the simulated and experimental profile is achieved [63]. This approach can lead to parameters offering good predictions, but due to their physical significance experimental determination of the parameters is often preferred. The most common dynamic techniques to do this are isocratic pulse [64-66] and breakthrough experiments [67] and linear gradient elutions [68,69] as they can be easily performed on standard laboratory liquid chromatography systems. Static batch experiments in principle do not require specialized equipment [70,71] but have regained popularity through the adoption of high-throughput systems [72,73]. In spite of their static nature batch adsorption experiments can be used to gain insight into the dynamic protein adsorption behaviour

[74,75]. Batch adsorption through resin aliquots in filter plates has proven to be a robust and very versatile method [76] allowing to determine the adsorption isotherms for ion-exchange [77], hydrophobic interaction [78] and mixed-mode resins [79,80]. Although the basic approaches have not changed, the recent years have seen many technological advances in the equipment to perform these experiments. Besides the aforementioned use for static batch experiments, high-throughput systems have been shown to perform column breakthrough experiments [81] and step gradient [82] and linear gradient elutions [83] in miniaturized columns. Even further miniaturization has been demonstrated by packing columns in microfluidic chips [84,85]. The further miniaturization is especially interesting for the investigation of strongly overloaded conditions, as the reduction of resin volume significantly cuts down the consumption of valuable sample. The experimental techniques discussed so far are usually applied to single solute systems. They can be expanded for multi-component mixture characterization by adding analytical techniques that can distinguish between different components that aren't sufficiently resolved by the screening experiment itself. Techniques that have been successfully demonstrated for this purpose include SELDI-TOF for the distinction between mAbs and host cell proteins [86] and selective protein quantification based on UV-spectra [87]. Combining the principles of these basic screening experiments with the proteomic complex sample analytical techniques has recently led to the development of a first multi-dimensional experimental characterization scheme to determine full parameter sets for both the product and many major contaminants directly from complex feedstocks [88].

A powerful complementary technique to experimental parameter determination lies in their calculation through predictive models. Performance prediction through quantitative structure-property relationships (QSPR) is achieved in three steps. First property and structure descriptors are calculated for a large set of

model proteins and resins for which the binding behaviour is known. Then the model is trained through statistical analysis of the relationship between descriptors and binding behaviour. Ultimately the descriptors of proteins or resins not in the training set are calculated and binding behaviour is predicted through extrapolation. Cramer et al. have demonstrated the successful application of QSPR models in the prediction of protein binding behaviour in ion-exchange, hydrophobic interaction, and mixed mode chromatography [89]. A very useful by-product of such an approach is that the examination of the descriptors can give a more detailed insight into poorly understood binding mechanisms, as often the case in mixed mode chromatography, at a microscopic level than purely macroscopic measurement of the binding performance. Better understood interactions, such as in ion-exchange chromatography where the binding is governed by electrostatic interactions could be predicted without model-calibration experiments through an colloidal sphere and plane approximation based on the Derjaguin-Landau-Verwey-Overbeek theory describing the interaction of charged surfaces in a liquid environment by accounting for van der Waals attraction and electrostatic repulsion caused by the formation of ion double layers [36,90]. An approach without the need for geometric approximations employs molecular dynamics simulations leading to very detailed predictions, but at significant computational cost [91]. As illustrated in Figure 1.4, these interaction predictive models lead to exceptional levels of process understanding, but their requirement for the protein 3D structure limits their use for developing separations of complex poorly characterized mixtures.

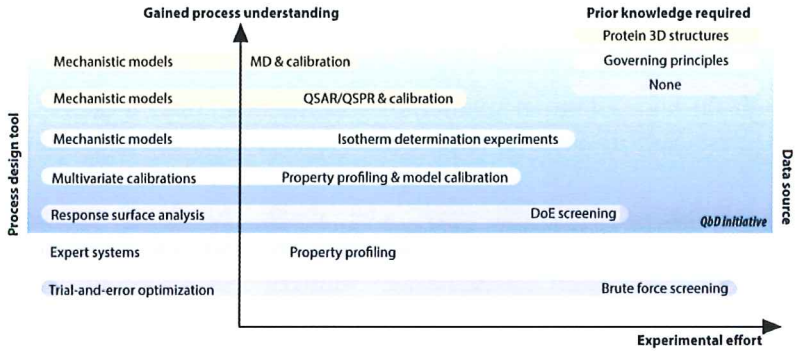


Figure 1.4. Overview of possible combinations process design tools and suitable input data sources, sorted by the level of process understanding gained from them. Certain combinations (shown in purple) can only be applied when the 3D structures of the proteins to be considered are available. Others (shown in green) require at least a prior understanding of the governing principles of separation process, whereas others (shown in yellow) can be applied without prior knowledge. However, also these approaches become more efficient when only parameters of significant impact are included within the optimization. the data sources are combined. It should be noted that the trial and error and interactions based approach can lead to multiple feasible process options. When multiple processes that comply with CQA requirements are identified, economic performance indicators should be included in the decision process.

Box 2. Hybrid process development approaches

Process development methods that combine techniques such as high-throughput experimentation and mathematical models have been classified as Hybrid approaches. These combinations can be advantageous to the efficiency of the approach when the combinations are chosen such that each individual technique compensates the disadvantage of the other. For the example of combining HTE with mathematical models, the scalability issues arising from the miniaturization associated with HTE approaches are alleviated through the incorporation of scaling effects into the model, whereas the large number of parameters required for model-based process development can only be generated through very efficient experimentation. In principle, multiple data sources could be combined to generate the required parameter sets. For instance, the production system related contaminants could be characterized through experiments applying principles from proteomics and HTE, whereas the product parameters could at the same time be generated thorough predictive approaches such as QSPR or MD simulations. In a similar manner, when a process is planned to include unit operations for which no suitable model is available, this step may be optimized and evaluated by tools requiring less prior knowledge. The ideal is to always choose the techniques that lead to the highest level of process understanding.

1.3.2. Mechanistic models in process development

Once the mass-transfer and isotherm models for the chromatographic unit operations to be considered have been chosen and all necessary parameters have been estimated, the models need to be subjected to rigorous validation. While high-throughput systems are definitely suitable to produce a lot of data, the process developer needs to investigate the uncertainties in the parameters he is going to use [92]. In this context, it is important to note that a low residual in the fit of the model to the experimental data does not necessarily imply that accurate parameters are estimated as long as there is uncertainty in the experimental conditions under which the

data was generated [93]. To be able to judge what level of parameter quality is acceptable requires a detailed analysis of the parameter sensitivities of the model [94]. Finally, taking all parts together, the models predictions must be compared to the performance actually achieved under the simulated conditions [95].

Once the combination of models and parameters have been found to give sufficiently accurate predictions, the ability to simulate process performance can be useful during many stages of process development. One of the first choices that has to be made is a selection of resins to be used for the process. Model-based approaches have been demonstrated to allow rational and fair comparison of resin separation performance under ideal conditions for each resin, while reducing the experimental load compared to conventional column scouting [58].

The most common application of models is probably the optimization of process conditions, once a resin has been chosen and the position within the process cascade has been decided. Examples for a variety of cases including various chromatographic modes have been published [96-98]. An advantage of optimizing a chromatographic step through mechanistic models instead of response surface analysis is that the mechanistic model can usually still make good predictions outside of its calibration [99]. This is an important requirement when considering using models for the design of adaptive design spaces to manage resin lot-to-lot variations, as it would be impossible to calibrate the model for all possible variations [100]. When validated models for multiple unit operations are available, they can be combined to cascaded models to allow process flow sheet optimization [101]. Finally, a validated model allows analysing the process towards robustness against disturbances. Demonstrated robustness then serves as the basis to register a larger design space with the regulatory authorities [102].

Table 1.2 The areas of application for state-of-the-art and novel techniques during various stages of biopharmaceutical downstream process development

Development Stage	Tasks	State of the Art	Alternatives	Benefits
1) Initial stability screening	• Identification of conditions causing product aggregation • Define process condition limitations	• High-throughput screening with multi-angle light scattering and/or SEC analysis	Measurement of self-interaction coefficients	• Lower material requirements • May be used for long-term stability prediction
2) Resin screening	• Determine resin selectivities • Quantify resin capacities	• High-throughput batch adsorption experiments	Techno-economic driven selection based on (optimized) productivity estimates	• Allows balancing of technical and economical properties
			Structure based calculations or simulations (QSPR / MD)	• Fewer experiments • May provide information on binding site
3) Column scouting	• Evaluate column performance under dynamic conditions	• Automated scouting programs on lab-scale LC-systems • RoboColumn screenings	Peak parking to determine dynamic effects and simulation of various process conditions	• Fewer experiments • Faster
4) Process cascade development	• Determine a cascade of unit operations that comply to the process productivity and purity requirements	• Adaption of existing platform designs • New approach Heuristic design rules • Trial-and-Error tests	Expert systems	• Simple guiding rules • Applicable to non-platform problems
			Parallel cascaded evaluation model	• Highest probability of identifying an optimal process • Applicable to non-platform problems
5) Single unit optimization	• Optimize the single unit operation and to increase the overall process performance • Replace single unit operations if overall	• Laboratory scale experiments according to an experimental design (DoE) followed by a Response surface analysis (RSA)	Application of optimization algorithms to mechanistic models	• Allows a much higher throughput of conditions to be tested • Higher chance of finding a process optimum

	process performance targets cannot be met		Trial and error optimizations with genetic algorithms	• Faster convergence to process optimum • Less likely to optimize towards local performance optimum
			Performance models based on multivariate calibrations	• Applicable to very complex feedstocks
6) Lab-scale trials	• Identify critical intermediate performance parameters and inter-unit operation dependencies	• Lab-scale modules provided by manufacturers	Incorporation of scaling effects into mechanistic models	• Fewer experiments necessary
7) Scale up	• Transfer the process design to production scale and solve issues arising from scaling effects	• Stepwise increase of equipment scale • Dimensionless process analysis (π-Theorem)	Integration of scaling issues within mechanistic models	• Fewer large scale trials necessary • Less likely to result in suboptimal process designs due to retroactive adjustments due to scaling problems
8) Optimization	• Further optimize the process parameters to increase the process scale performance	• Process scale experiments according to an experimental design (DoE) followed by a Response surface analysis (RSA)	Application of optimization algorithms to mechanistic models	• Fewer large scale trials necessary
9) Validation	• Deliver final proof on process robustness • Deliver final report on process control strategy and risk-management	• Risk assessment • Spiked recovery studies • Report response surfaces for CPP	Replace response surfaces with validated mechanistic models	• Demonstrates a higher level of process understanding to the regulatory authorities

1.4 Concluding Remarks

The introduction of high-throughput technology revolutionized how chromatographic process development is approached by enabling thorough and systematic investigation of which parameters influence the process and how they can be optimized. Throughout the biotechnology industry there is a trend to move towards the next level of process understanding, reflected in the increasing adoption of mechanistic models. The inherent complexity of biopharmaceuticals and the systems they are produced in pose a unique set of challenges on this path. This has led to the development of a wide variety of powerful new techniques, models and analytical tools to aid process developers. An overview of how to apply them during process development is given in Table 2. A unique opportunity arises from the complementary character of these exciting new tools. Downstream scientist and engineers will be able make a selection from various process development modules to meet their current project needs. Models can be connected to multiple data sources ranging from experimentally determined parameters to molecular dynamics based predictions, while poorly understood operations can continue to be optimized empirically. Creating a universal module interface and a common processing parameter database will be the key challenges of the future.

References

[1] M. Goodman, Market Watch: Pharma industry performance metrics: 2007-2012E, Nat. Rev. Drug Discov. 7 (2008) 795-795.

[2] R. Osborne, Fresh from the biotech pipeline--2012, Nat. Biotechnol. 31 (2013) 100-103.

[3] G. Guiochon, L.A. Beaver, Separation science is the key to successful biopharmaceuticals, J. Chromatogr. A 1218 (2011) 8836-8858.

[4] T.M. Przybycien, N.S. Pujar, L.M. Steele, Alternative bioseparation operations: life beyond packed-bed chromatography, Curr. Opin. Biotechnol. 15 (2004) 469-478.

[5] U. Gottschalk, K. Brorson, A.A. Shukla, The need for innovation in biomanufacturing, Nat. Biotechnol. 30 (2012) 489-492.

[6] S.M. Cramer, M.A. Holstein, Downstream bioprocessing: recent advances and future promise, Current Opinion in Chemical Engineering 1 (2011) 27-37.

[7] A.A. Shukla, J. Thommes, Recent advances in large-scale production of monoclonal antibodies and related proteins, Trends Biotechnol. 28 (2010) 253-261.

[8] G.R. Bolton, B.N. Violand, R.S. Wright, S.J. Sun, K.M. Sunasara, K. Watson, J.L. Coffman, C. Gallo, R. Godavarti, Addressing the Challenges in Downstream Processing Today and Tomorrow, Biopharm International 24 (2011) S8-S15.

[9] I. Ahmed, B. Kaspar, U. Sharma, Biosimilars: impact of biologic product life cycle and European experience on the regulatory trajectory in the United States, Clin. Ther. 34 (2012) 400-419.

[10] A.S. Rathore, Roadmap for implementation of quality by design (QbD) for biotechnology products, Trends Biotechnol. 27 (2009) 546-553.

[11] R. Bhambure, K. Kumar, A.S. Rathore, High-throughput process development for biopharmaceutical drug substances, Trends Biotechnol. 29 (2011) 127-135.

[12] L.X. Yu, Pharmaceutical quality by design: product and process development, understanding, and control, Pharm. Res. 25 (2008) 781-791.

[13] A.O. Kirdar, K.D. Green, A.S. Rathore, Application of multivariate data analysis for identification and successful resolution of a root cause for a bioprocessing application, Biotechnol. Prog. 24 (2008) 720-726.

[14] K.M. Lacki, High-throughput process development of chromatography steps: advantages and limitations of different formats used, Biotechnol. J. 7 (2012) 1192-1202.

[15] K. Treier, P. Lester, J. Hubbuch, Application of genetic algorithms and response surface analysis for the optimization of batch chromatographic systems, Biochem. Eng. J. 63 (2012) 66-75.

[16] C. Julien, W. Whitford, A New Era for Bioprocess Design and Control, Part 1: The Basic Concepts, BioProcess Int 6 (2008) 16.

[17] C. Julien, W. Whitford, A New Era for Bioprocess Design and Control, Part 2: Harmonizing Concepts, BioProcess Int 6 (2008) 24.

[18] K.V. Gernaey, A.E. Lantz, P. Tufvesson, J.M. Woodley, G. Sin, Application of mechanistic models to fermentation and biocatalysis for next-generation processes, Trends Biotechnol. 28 (2010) 346-354.

[19] K. Westerberg, E.B. Hansen, T.B. Hansen, M. Degerman, B. Nilsson, Model-assisted process development for preparative chromatography applications, BioProcess International 9 (2011) 48-56.

[20] Z. Hamrang, N.J. Rattray, A. Pluen, Proteins behaving badly: emerging technologies in profiling biopharmaceutical aggregation, Trends Biotechnol. 31 (2013) 448-458.

[21] F. Gritti, G. Guiochon, Mass transfer kinetics, band broadening and column efficiency, J. Chromatogr. A 1221 (2012) 2-40.

[22] A. Seidel-Morgenstern, Experimental determination of single solute and competitive adsorption isotherms, J. Chromatogr. A 1037 (2004) 255-272.

[23] B.K. Nfor, P.D. Verhaert, L.A. van der Wielen, J. Hubbuch, M. Ottens, Rational and systematic protein purification process development: the next generation, Trends Biotechnol. 27 (2009) 673-679.

[24] G. Sandoval, C. Shene, B.A. Andrews, J.A. Asenjo, Extension of the selection of protein chromatography and the rate model to affinity chromatography, J. Mol. Recognit. 23 (2010) 609-617.

[25] F. Dismer, J. Hubbuch, A novel approach to characterize the binding orientation of lysozyme on ion-exchange resins, J. Chromatogr. A 1149 (2007) 312-320.

[26] J.A. Asenjo, B.A. Andrews, Protein purification using chromatography: selection of type, modelling and optimization of operating conditions, J. Mol. Recognit. 22 (2009) 65-76.

[27] J.A. Asenjo, B.A. Andrews, Is there a rational method to purify proteins? From expert systems to proteomics, J. Mol. Recognit. 17 (2004) 236-247.

[28] M.E. Lienqueo, J.A. Asenjo, Use of expert systems for the synthesis of downstream protein processes, Comput. Chem. Eng. 24 (2000) 2339-2350.

[29] T. Ahamed, M. Ottens, B.K. Nfor, G.W.K. van Dedem, L.A.M. van der Wielen, A generalized approach to thermodynamic properties of biomolecules for use in bioseparation process design, Fluid Phase Equilib. 241 (2006) 268-282.

[30] B. Guelat, G. Strohlein, M. Lattuada, L. Delegrange, P. Valax, M. Morbidelli, Simulation model for overloaded monoclonal antibody variants separations in ion-exchange chromatography, J. Chromatogr. A 1253 (2012) 32-43.

[31] X.L. Su, Y. Sun, Thermodynamic model for nonlinear electrostatic adsorption equilibrium of protein, AIChE J. 52 (2006) 2921-2930.

[32] X. Zhang, A.Q. Fang, C.P. Riley, M. Wang, F.E. Regnier, C. Buck, Multi-dimensional liquid chromatography in proteomics-A review, Anal. Chim. Acta 664 (2010) 101-113.

[33] A. Hawe, W.L. Hulse, W. Jiskoot, R.T. Forbes, Taylor dispersion analysis compared to dynamic light scattering for

the size analysis of therapeutic peptides and proteins and their aggregates, Pharm. Res. 28 (2011) 2302-2310.

[34] K. Miyabe, J. Nagai, G. Guiochon, Peak parking-moment analysis: A strategy for the measurement of molecular diffusivity in liquid phase, Chem. Eng. Sci. 65 (2010) 3859-3864.

[35] E. Hausler, P. Domagalski, M. Ottens, A. Bardow, Microfluidic diffusion measurements: The optimal H-cell, Chem. Eng. Sci. 72 (2012) 45-50.

[36] B. Guelat, G. Strohlein, M. Lattuada, M. Morbidelli, Electrostatic model for protein adsorption in ion-exchange chromatography and application to monoclonal antibodies, lysozyme and chymotrypsinogen A, J. Chromatogr. A 1217 (2010) 5610-5621.

[37] J.C. Salgado, I. Rapaport, J.A. Asenjo, Predicting the behaviour of proteins in hydrophobic interaction chromatography. 1: Using the hydrophobic imbalance (HI) to describe their surface amino acid distribution, J. Chromatogr. A 1107 (2006) 110-119.

[38] J.C. Salgado, I. Rapaport, J.A. Asenjo, Predicting the behaviour of proteins in hydrophobic interaction chromatography. 2. Using a statistical description of their surface amino acid distribution, J. Chromatogr. A 1107 (2006) 120-129.

[39] R.K. Swanson, R. Xu, D. Nettleton, C.E. Glatz, Proteomics-based, multivariate random forest method for prediction of protein separation behavior during cation-exchange chromatography, J. Chromatogr. A 1249 (2012) 103-114.

[40] S.R. Trevino, J.M. Scholtz, C.N. Pace, Measuring and increasing protein solubility, J. Pharm. Sci. 97 (2008) 4155-4166.

[41] B.K. Nfor, N.N. Hylkema, K.R. Wiedhaup, P.D. Verhaert, L.A. van der Wielen, M. Ottens, High-throughput protein precipitation and hydrophobic interaction chromatography: salt effects and thermodynamic interrelation, J. Chromatogr. A 1218 (2011) 8958-8973.

[42] D.J. Winzor, M. Deszczynski, S.E. Harding, P.R. Wills, Nonequivalence of second virial coefficients from

sedimentation equilibrium and static light scattering studies of protein solutions, Biophys. Chem. 128 (2007) 46-55.

[43] N. Rakel, K. Schleining, F. Dismer, J. Hubbuch, Self-interaction chromatography in pre-packed columns: a critical evaluation of self-interaction chromatography methodology to determine the second virial coefficient, J. Chromatogr. A 1293 (2013) 75-84.

[44] K.S. Deshpande, S. Kuddannaya, J. Staginus, P.C. Thune, L.C.P.M. de Smet, J.H. ter Horst, L.A.M. van der Wielen, M. Ottens, Biofunctionalization and self-interaction chromatography in PDMS microchannels, Biochem. Eng. J. 67 (2012) 111-119.

[45] C.A. Haynes, K. Tamura, H.R. Korfer, H.W. Blanch, J.M. Prausnitz, Thermodynamic properties of aqueous .alpha.-chymotrypsin solution from membrane osmometry measurements, The Journal of Physical Chemistry 96 (1992) 905-912.

[46] R. Kellner, F. Lottspeich, H.E. Meyer, Microcharacterization of proteins, Wiley-VCH, Weinheim; New York, 1999.

[47] O. Aguilar, M. Rito-Palomares, C.E. Glatz, Coupled Application of Aqueous Two-Phase Partitioning and 2D-Electrophoresis for Characterization of Soybean Proteins, Sep. Sci. Technol. 45 (2010) 2210-2225.

[48] O. Aguilar, C.E. Glatz, M. Rito-Palomares, Characterization of green-tissue protein extract from alfalfa (Medicago sativa) exploiting a 3-D technique, J. Sep. Sci. 32 (2009) 3223-3231.

[49] J.K. Grzeskowiak, A. Tscheliessnig, M.W. Wu, P.C. Toh, J. Chusainow, Y.Y. Lee, N. Wong, A. Jungbauer, Two-dimensional difference fluorescence gel electrophoresis to verify the scale-up of a non-affinity-based downstream process for isolation of a therapeutic recombinant antibody, Electrophoresis 31 (2010) 1862-1872.

[50] A.S. Tait, C.E.M. Hogwood, C.M. Smales, D.G. Bracewell, Host cell protein dynamics in the supernatant of a mAb producing CHO cell line, Biotechnol. Bioeng. 109 (2012) 971-982.

[51] A.L. Tscheliessnig, J. Konrath, R. Bates, A. Jungbauer, Host cell protein analysis in therapeutic protein bioprocessing - methods and applications, Biotechnol. J. 8 (2013) 655-670.

[52] L. Xu, C.E. Glatz, Predicting protein retention time in ion-exchange chromatography based on three-dimensional protein characterization, J. Chromatogr. A 1216 (2009) 274-280.

[53] S. Javeed, S. Qamar, W. Ashraf, G. Warnecke, A. Seidel-Morgenstern, Analysis and numerical investigation of two dynamic models for liquid chromatography, Chem. Eng. Sci. 90 (2013) 17-31.

[54] G. Guiochon, Preparative liquid chromatography, J. Chromatogr. A 965 (2002) 129-161.

[55] X. Du, Q. Yuan, J. Zhao, Y. Li, Comparison of general rate model with a new model--artificial neural network model in describing chromatographic kinetics of solanesol adsorption in packed column by macroporous resins, J. Chromatogr. A 1145 (2007) 165-174.

[56] H. Gao, B. Lin, A simplified numerical method for the General Rate model, Comput. Chem. Eng. 34 (2010) 277-285.

[57] E. von Lieres, J. Andersson, A fast and accurate solver for the general rate model of column liquid chromatography, Comput. Chem. Eng. 34 (2010) 1180-1191.

[58] B.K. Nfor, D.S. Zuluaga, P.J. Verheijen, P.D. Verhaert, L.A. van der Wielen, M. Ottens, Model-based rational strategy for chromatographic resin selection, Biotechnol. Prog. 27 (2011) 1629-1643.

[59] K. Westerberg, E.B. Hansen, M. Degerman, T.B. Hansen, B. Nilsson, Model-Based Process Challenge of an Industrial Ion-Exchange Chromatography Step, Chem. Eng. Technol. 35 (2012) 183-190.

[60] A.A. Shukla, S.S. Bae, J.A. Moore, K.A. Barnthouse, S.M. Cramer, Synthesis and characterization of high-affinity, low molecular weight displacers for cation-exchange chromatography, Ind. Eng. Chem. Res. 37 (1998) 4090-4098.

[61] S.J. Traylor, X. Xu, A.M. Lenhoff, Shrinking-core modeling of binary chromatographic breakthrough, J. Chromatogr. A 1218 (2011) 2222-2231.

[62] A.M. Lenhoff, Multiscale modeling of protein uptake patterns in chromatographic particles, Langmuir 24 (2008) 5991-5995.

[63] A. Osberghaus, S. Hepbildikler, S. Nath, M. Haindl, E. von Lieres, J. Hubbuch, Determination of parameters for the steric mass action model-A comparison between two approaches, J. Chromatogr. A 1233 (2012) 54-65.

[64] B.C. To, A.M. Lenhoff, Hydrophobic interaction chromatography of proteins. I. The effects of protein and adsorbent properties on retention and recovery, J. Chromatogr. A 1141 (2007) 191-205.

[65] B.C. To, A.M. Lenhoff, Hydrophobic interaction chromatography of proteins. II. Solution thermodynamic properties as a determinant of retention, J. Chromatogr. A 1141 (2007) 235-243.

[66] J.M. Mollerup, T.B. Hansen, S. Kidal, L. Sejergaard, E. Hansen, A. Staby, Use of Quality by the Design for the Modelling of Chromatographic Separations, J. Liq. Chromatogr. Relat. Technol. 32 (2009) 1577-1597.

[67] B.D. Bowes, A.M. Lenhoff, Protein adsorption and transport in dextran-modified ion-exchange media. II. Intraparticle uptake and column breakthrough, J. Chromatogr. A 1218 (2011) 4698-4708.

[68] T. Ishihara, T. Kadoya, S. Yamamoto, Application of a chromatography model with linear gradient elution experimental data to the rapid scale-up in ion-exchange process chromatography of proteins, J. Chromatogr. A 1162 (2007) 34-40.

[69] S. Yamamoto, A. Kita, Rational design calculation method for stepwise elution chromatography of proteins, Food and Bioproducts Processing 84 (2006) 72-77.

[70] E.X.P. Almodovar, Y.Y. Tao, G. Carta, Protein Adsorption and Transport in Cation Exchangers with a Rigid Backbone Matrix with and without Polymeric Surface Extenders, Biotechnol. Prog. 27 (2011) 1264-1272.

[71] M.C. Stone, Y. Tao, G. Carta, Protein adsorption and transport in agarose and dextran-grafted agarose media for ion exchange chromatography: Effect of ionic strength and protein characteristics, J. Chromatogr. A 1216 (2009) 4465-4474.

[72] K. Treier, A. Berg, P. Diederich, K. Lang, A. Osberghaus, F. Dismer, J. Hubbuch, Examination of a genetic algorithm for the application in high-throughput downstream process development, Biotechnol. J. 7 (2012) 1203-1215.

[73] A. Susanto, E. Knieps-Grunhagen, E. von Lieres, J. Hubbuch, High Throughput Screening for the Design and Optimization of Chromatographic Processes: Assessment of Model Parameter Determination from High Throughput Compatible Data, Chem. Eng. Technol. 31 (2008) 1846-1855.

[74] G. Carta, Predicting protein dynamic binding capacity from batch adsorption tests, Biotechnol. J. 7 (2012) 1216-1220.

[75] T. Bergander, K. Nilsson-Vaelimaa, K. Oberg, K.M. Lacki, High-throughput process development: Determination of dynamic binding capacity using microtiter filter plates filled with chromatography resin, Biotechnol. Prog. 24 (2008) 632-639.

[76] J.L. Coffman, J.F. Kramarczyk, B.D. Kelley, High-throughput screening of chromatographic separations: I. Method development and column modeling, Biotechnol. Bioeng. 100 (2008) 605-618.

[77] B.D. Kelley, M. Switzer, P. Bastek, J.F. Kramarczyk, K. Molnar, T. Yu, J. Coffman, High-throughput screening of chromatographic separations: IV. Ion-exchange, Biotechnol. Bioeng. 100 (2008) 950-963.

[78] J.F. Kramarczyk, B.D. Kelley, J.L. Coffman, High-throughput screening of chromatographic separations: II. Hydrophobic interaction, Biotechnol. Bioeng. 100 (2008) 707-720.

[79] D.L. Wensel, B.D. Kelley, J.L. Coffman, High-throughput screening of chromatographic separations: III. Monoclonal antibodies on ceramic hydroxyapatite, Biotechnol. Bioeng. 100 (2008) 839-854.

[80] B.K. Nfor, M. Noverraz, S. Chilamkurthi, P.D. Verhaert, L.A. van der Wielen, M. Ottens, High-throughput isotherm determination and thermodynamic modeling of protein adsorption on mixed mode adsorbents, J. Chromatogr. A 1217 (2010) 6829-6850.

[81] M. Wiendahl, P.S. Wierling, J. Nielsen, D.F. Christensen, J. Krarup, A. Staby, J. Hubbuch, High throughput screening for the design and optimization of chromatographic processes - Miniaturization, automation and parallelization of breakthrough and elution studies, Chem. Eng. Technol. 31 (2008) 893-903.

[82] S.K. Hansen, E. Skibsted, A. Staby, J. Hubbuch, A label-free methodology for selective protein quantification by means of absorption measurements, Biotechnol. Bioeng. 108 (2011) 2661-2669.

[83] A. Osberghaus, K. Drechsel, S. Hansen, S.K. Hepbildikler, S. Nath, M. Haindl, E. von Lieres, J. Hubbuch, Model-integrated process development demonstrated on the optimization of a robotic cation exchange step, Chem. Eng. Sci. 76 (2012) 129-139.

[84] M.S. Shapiro, S.J. Haswell, G.J. Lye, D.G. Bracewell, Design and characterization of a microfluidic packed bed system for protein breakthrough and dynamic binding capacity determination, Biotechnol. Prog. 25 (2009) 277-285.

[85] M.S. Shapiro, S.J. Haswell, G.J. Lye, D.G. Bracewell, Microfluidic Chromatography for Early Stage Evaluation of Biopharmaceutical Binding and Separation Conditions, Sep. Sci. Technol. 46 (2011) 185-194.

[86] P.S. Wierling, R. Bogumil, E. Knieps-Grunhagen, J. Hubbuch, High-throughput screening of packed-bed chromatography coupled with SELDI-TOF MS analysis: monoclonal antibodies versus host cell protein, Biotechnol. Bioeng. 98 (2007) 440-450.

[87] S.K. Hansen, B. Jamali, J. Hubbuch, Selective high throughput protein quantification based on UV absorption spectra, Biotechnol. Bioeng. 110 (2013) 448-460.

[88] B.K. Nfor, T. Ahamed, M.W. Pinkse, L.A. van der Wielen, P.D. Verhaert, G.W. van Dedem, M.H. Eppink, E.J. van de

Sandt, M. Ottens, Multi-dimensional fractionation and characterization of crude protein mixtures: toward establishment of a database of protein purification process development parameters, Biotechnol. Bioeng. 109 (2012) 3070-3083.

[89] T. Yang, C.M. Breneman, S.M. Cramer, Investigation of multi-modal high-salt binding ion-exchange chromatography using quantitative structure-property relationship modeling, J. Chromatogr. A 1175 (2007) 96-105.

[90] B. Guelat, L. Delegrange, P. Valax, M. Morbidelli, Model-based prediction of monoclonal antibody retention in ion-exchange chromatography, J. Chromatogr. A 1298 (2013) 17-25.

[91] F. Dismer, J. Hubbuch, 3D structure-based protein retention prediction for ion-exchange chromatography, J. Chromatogr. A 1217 (2010) 1343-1353.

[92] A. Osberghaus, P. Baumann, S. Hepbildikler, S. Nath, M. Haindl, E. von Lieres, J. Hubbuch, Detection, Quantification, and Propagation of Uncertainty in High-Throughput Experimentation by Monte Carlo Methods, Chem. Eng. Technol. 35 (2012) 1456-1464.

[93] N. Borg, K. Westerberg, N. Andersson, E. von Lieres, B. Nilsson, Effects of uncertainties in experimental conditions on the estimation of adsorption model parameters in preparative chromatography, Comput. Chem. Eng. 55 (2013) 148-157.

[94] A. Puttmann, S. Schnittert, U. Naumann, E. von Lieres, Fast and accurate parameter sensitivities for the general rate model of column liquid chromatography, Comput. Chem. Eng. 56 (2013) 46-57.

[95] B.K. Nfor, J. Ripic, A. van der Padt, M. Jacobs, M. Ottens, Model-based high-throughput process development for chromatographic whey proteins separation, Biotechnol. J. 7 (2012) 1221-1232.

[96] D. Karlsson, N. Jakobsson, A. Axelsson, B. Nilsson, Model-based optimization of a preparative ion-exchange step for antibody purification, J. Chromatogr. A 1055 (2004) 29-39.

[97] D. Nagrath, F. Xia, S.M. Cramer, Characterization and modeling of nonlinear hydrophobic interaction chromatographic systems, J. Chromatogr. A 1218 (2011) 1219-1226.

[98] G. Sandoval, B.A. Andrews, J.A. Asenjo, Elution relationships to model affinity chromatography using a general rate model, J. Mol. Recognit. 25 (2012) 571-579.

[99] A. Osberghaus, S. Hepbildikler, S. Nath, M. Haindl, E. von Lieres, J. Hubbuch, Optimizing a chromatographic three component separation: a comparison of mechanistic and empiric modeling approaches, J. Chromatogr. A 1237 (2012) 86-95.

[100] E. Close, D.G. Bracewell, E. Sorensen, in: K. Andrzej, T. Ilkka (Eds.), Computer Aided Chemical Engineering, Elsevier, 2013, p. 115-120.

[101] B.K. Nfor, T. Ahamed, G.W.K. van Dedem, P.D.E.M. Verhaert, L.A.M. van der Wielen, M.H.M. Eppink, E.J.A.X. van de Sandt, M. Ottens, Model-based rational methodology for protein purification process synthesis, Chem. Eng. Sci. 89 (2013) 185-195.

[102] M. Degerman, K. Westerberg, B. Nilsson, A Model-Based Approach to Determine the Design Space of Preparative Chromatography, Chem. Eng. Technol. 32 (2009) 1195-1202.

2

Fourier transform assisted deconvolution of skewed peaks in complex multi-dimensional chromatograms

Abstract

Lower order peak moments of individual peaks in heavily fused peak clusters can be determined by fitting peak models to the experimental data. The success of such an approach depends on two main aspects: the generation of meaningful initial estimates on the number and position of the peaks, and the choice of a suitable peak model. For the detection of meaningful peaks in multidimensional chromatograms, a fast data scanning algorithm was combined with prior resolution enhancement through the reduction of column and system broadening effects with the help of two-dimensional fast Fourier transforms. To capture the shape of skewed peaks in multi-dimensional chromatograms a formalism for the accurate calculation of exponentially modified Gaussian peaks, one of the most popular models for skewed peaks, was extended for direct fitting of two-dimensional data. The method is demonstrated to successfully identify and deconvolute peaks hidden in strongly fused peak clusters. Incorporation of automatic analysis and reporting of the statistics of the fitted peak parameters and calculated properties allows to easily identify in which regions of the chromatograms additional resolution is required for robust quantification.

Keywords: Comprehensive two-dimensional chromatography; Deconvolution; Fourier transform;Exponentially modified Gaussian; Peak model; Non-linear curve fitting;

Published as: A.T. Hanke, P.D.E.M. Verhaert, L.A.M. van der Wielen, M.H.M. Eppink, E.J.A.X. van de Sandt and M. Ottens, **J Chrom A**, 1394 (2015): 54-61

2.1 Introduction

Chromatography is one of the most common techniques in analytical laboratories, especially for the analysis of mixtures of larger organic molecules. Its output is typically presented in the form of chromatograms, the intensity of a detector signal over the time of a separation in which each component is represented by a peak. The amplitude of the peak reflects the concentration of the components in relation to the detector sensitivity. The shape of the peak on the other hand is determined by the complex interplay of mass-transfer and adsorption phenomena occurring in the column and in the system dead-volume [1]. In the ideal case, i.e. where the components of interest are fully resolved, the interpretation of these chromatograms is relatively straightforward as lower statistical peak moments, such as the area (0^{th}-moment) and average retention time (1^{st} moment), can be calculated accurately by simple integrators or even graphically [2]. Where peaks are not fully resolved, straightforward approaches such as perpendicular drop or tangent skim, may still lead to reasonable results for symmetrical peaks with limited overlap [3]. In practice peaks may often be skewed due to slow mass transfer or extra-column effects, that can lead to large errors during chromatogram analysis [3].

One of the most wide-spread approaches to solving the problem of overlapping skewed peaks is multivariate curve resolution (MCR) [4]. MCR utilizes the bilinear character of spectroscopic chromatograms [5], i.e. that the recorded chromatogram is a linear combination of the concentration profiles of the present species and their respective absorption properties. MCR has been demonstrated to be both effective for mixtures where the single components absorption spectra are known [6] and unknown [7], though in the latter the statistical uncertainties of the obtained results increase with the number of components present. In contrast to MCR, hard-modeling techniques focus only on the

concentration profiles and require only univariate data. The two techniques are highly compatible and can compensate for the shortcomings of each other [8].

In this study we introduce a hard-modelling approach for the deconvolution of complex two-dimensional chromatograms. A special focus lies on the generation of good initial estimates with the help of Fourier transforms. The results are subjected to rigorous statistical analysis to identify the regions where the hard-modelling approach by itself can lead to sufficiently robust results and where the multivariate techniques might be necessary.

2.2 Theory

Over the years a huge library of peak models has been developed, many of which can account for peak asymmetries [9]. Probably one of the most popular models for the description of asymmetrical chromatographic peaks is the exponentially modified Gaussian distribution (EMG). Its popularity is, at least, partly based on the relative physical significance of its parameters: the variance can be related to the peak broadening caused by axial dispersion, whereas the exponential decay is a reasonable model to capture dead-volume effects. An additional advantage of the EMG is that it requires a relatively small number of parameters to be able to describe a large variety of peak shapes, from almost perfectly Gaussian to heavily tailing peaks with sharp fronts. The EMG can be expressed in many mathematically equivalent ways that may lead to large errors when calculated numerically for certain parameter ranges. Kalambet et al. [10] introduced a simple decision parameter z to guide in the selection of the form of the EMG to use for accurate numerical calculation. For a single peak this decision factor can be expressed as

$$z = \frac{1}{\sqrt{2}} \cdot \left(\frac{\mu - x}{\sigma} + \frac{\sigma}{\tau} \right) \tag{1}$$

where x is the input variable, for chromatography typically the elution time or volume. The mode of the Gaussian constituent is given as μ, the Gaussian variance σ and the relaxation parameter of the exponential decay as τ. The equation best suited for numerical calculation of the EMG is then for $z < 0$

$$F(x) = h \cdot \sqrt{\frac{\pi}{2}} \cdot \frac{\sigma}{\tau} \cdot \exp\left(\frac{\mu - x}{\tau} + \frac{\sigma^2}{2\tau^2}\right) \cdot \operatorname{erfc}\left(\frac{1}{\sqrt{2}} \cdot \left(\frac{\mu - x}{\sigma} + \frac{\sigma}{\tau}\right)\right) \qquad (2)$$

and for $z \geq 0$

$$F(x) = h \cdot \sqrt{\frac{\pi}{2}} \cdot \frac{\sigma}{\tau} \cdot \exp\left(\frac{-(\mu - x)^2}{2\sigma^2}\right) \cdot \operatorname{erfcx}\left(\frac{1}{\sqrt{2}} \cdot \left(\frac{\mu - x}{\sigma} + \frac{\sigma}{\tau}\right)\right) \qquad (3)$$

with h being the height of the unmodified Gaussian.

Once such a suitable peak model has been identified, numerical optimizers have been shown to be able to fit them to experimental data [11]. The EMG has been demonstrated to be suitable for optimizer based fitting and deconvolution of most chromatographic peaks providing the observed peak tailing is not too pronounced. In these cases the polynomial modified Gaussian (PMG) shows better fitting capability [12]. Besides the suitability of the peak shape, knowledge of the number of peaks fused in the chromatogram and their relative positons were identified as critical parameters in the success of optimizer based deconvolution [13]. Depending on the complexity of the chromatograms, identifying the number and positions of possible peaks is not a trivial task. To avoid operator to operator variation, especially in quality control environments, it is preferable to have this operation performed by peak detection algorithms. Two popular peak detection approaches are simple local maxima search algorithms that closely resemble a human looking for visually distinguishable peaks, and analysis of higher-order derivatives of the measurement signal. The latter has been shown to be able to recognize more peaks; especially such

hidden in shoulders of larger peaks, but is highly sensitive to noise in the original signal [14,15].

A more robust approach to increase the probability to observe well resolved peaks is to increase the system's peak capacity [16]. This can be achieved by increasing the efficiency of the used columns and reduction of extra-column effects, but most effectively by increasing the number of orthogonal separation dimensions [17]. The principles for the interpretation of these multi-dimensional chromatograms remain the same. For practical reasons, the dimensionality of comprehensive separations is often limited to two orthogonal methods, even when performed in offline mode [18]. As a peak model for two-dimensional chromatography the Kalambet et al. [10] system of equations for the description of EMG shaped peaks can be extended by a second dimension. The general equation to describe a fused set of n two-dimensional EMG distributions (2D-mEMG) is then given by

$$F(x,y) = \sum_{1}^{n} \left\{ h_i \cdot \frac{\pi}{2} \cdot \frac{\sigma_{x,i} \cdot \sigma_{y,i}}{\tau_{x,i} \cdot \tau_{y,i}} \cdot co_{x,i} \cdot co_{y,i} \right\} \tag{4}$$

where $co_{x,i}$ and $co_{y,i}$ are co-factors that change depending on the peak and parameter range. Similar to the one-dimensional case the equation for the accurate calculation of the cofactors can be chosen by decision variables:

$$z_{x,i} = \frac{1}{\sqrt{2}} \cdot \left(\frac{\mu_{x,i} - x}{\sigma_{x,i}} + \frac{\sigma_{x,i}}{\tau_{x,i}} \right) \tag{5}$$

$$z_{y,i} = \frac{1}{\sqrt{2}} \cdot \left(\frac{\mu_{y,i} - y}{\sigma_{y,i}} + \frac{\sigma_{y,i}}{\tau_{y,i}} \right) \tag{6}$$

Similar to the one-dimensional case the co-factors for the first dimension are for $z_{x,i} < 0$

$$co_{x,i} = \exp\left(\frac{\mu_{x,i} - x}{\tau_{x,i}} + \frac{{\sigma_{x,i}}^2}{2{\tau_{x,i}}^2} \right) \cdot \operatorname{erfc}(z_{x,i}) \tag{7}$$

and for $z_{x,i} \geq 0$

$$co_{x,i} = \exp\left(\frac{-(\mu_{x,i} - x)^2}{2{\sigma_{x,i}}^2}\right) \cdot \mathrm{erfcx}(z_{x,i}) \tag{8}$$

Analog to the first dimension, the cofactors for the second dimension are for $z_{y,i} < 0$

$$co_{y,i} = \exp\left(\frac{\mu_{y,i} - y}{\tau_{y,i}} + \frac{{\sigma_{y,i}}^2}{2{\tau_{y,i}}^2}\right) \cdot \mathrm{erfc}(z_{y,i}) \tag{9}$$

and for $z_{y,i} \geq 0$

$$co_{y,i} = \exp\left(\frac{-(\mu_{y,i} - y)^2}{2{\sigma_{y,i}}^2}\right) \cdot \mathrm{erfcx}(z_{y,i}) \tag{10}$$

When viewing the system of Eq. 4-10 it becomes apparent that there is no built in correlation between the first and second dimensions. As a result the peak model should preferably be used to describe systems where the dimensions consist of orthogonal methods. This restriction to the application of the peak model is deemed acceptable, as orthogonality of the separation dimensions is an important part of the design paradigm of multi-dimensional chromatography systems [19]. It should also be noted that fitting multiple peaks to a single chromatogram approach also assumes that the chromatogram is the result of the linear addition of the single component contributions to the final recorded chromatogram, a condition only met when the used detector is strictly operated within its linear response range. When it is no longer feasible to improve the separation system on a technical level, there is the possibility to virtually reduce the contribution of band-broadening and extra column effects. This effect can be achieved with the help of Fourier transformations [20]. Deconvolution by means of the Fourier transforms has been shown to have a suitable sharpening effect on chromatograms with EMG shaped peaks [21]. The characteristics of applying the Fourier transformation in the form of fast Fourier transform (FFT) algorithms to real experimental chromatograms has

been studied thoroughly [22]. Due to the introduction of artefacts such as small negative side-lobes, and slight shifts in peak retention patterns, research on the use of Fourier transforms in chromatography has since focused more on complete Fourier Analysis of the chromatograms, rather than on resolution enhancement [23-26].

For some applications of multi-dimensional separations, such as the regression of thermodynamic parameters [27,28], it is important to determine the peak properties of the unaltered chromatograms. In this context, the proposed sample and data processing scheme outlined in Figure 2.1 was developed. It constructs two-dimensional chromatograms from independently recorded single dimension chromatograms and utilizes Fourier transforms to increase the efficiency of the subsequent peak detection algorithm. The detected peak properties are passed on to an optimizer that fits the data to the described two-dimensional EMG. The resulting fits are used for calculation of the lower rank statistical peak moments.

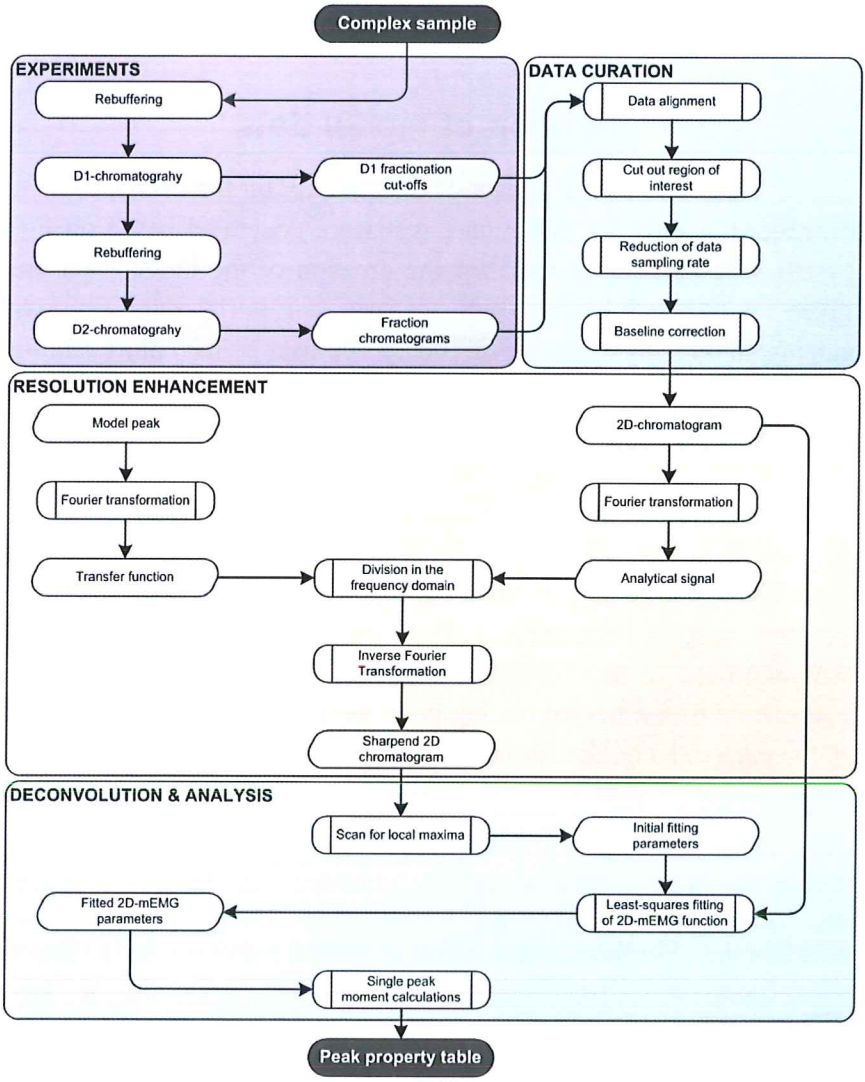


Figure 2.1. Flowchart of the experimental and data processing steps for the deconvolution and analysis of complex multidimensional chromatograms.

2.3 Experimental and computational methods

2.3.1. Creation of model data

To analyze the impact of data quality on the deconvolution procedure a series of model data sets were generated based on Eq. (4-10). The parameters used for the creation of the model data are given in Table 2.1. For each scenario, an initial data set was calculated on a regular mesh of 100 by 100 data points before adding artificial imperfections. Two types of noise were investigated: a normally distributed pseudo-random background noise across the entire chromatogram with an amplitude dependent on the maximum signal amplitude, and a normally distributed pseudo-random measurement error dependent on the local signal strength. A total of nine linear combinations of noise and error were tested for each scenario ranging from 0-2% of the respective amplitudes. Each test was repeated at least three times. The variance and relaxation parameters of the deconvoluting peak were chosen to be a factor of 0.9 smaller than of the simulated peaks to avoid the introduction of artefacts.

Table 2.1 Peak parameters of the simulated data for the statistical evaluation of the deconvolution procedure. Scenario 1 represents two overlapping identical peaks with different retention in a single dimension, and Scenario 2 with different retention in two dimensions.

Scenario 1			Scenario 2		
	Peak 1	Peak 2		Peak 1	Peak 2
h	1	1	h	1	1
μ_x	0.2	0.2	μ_x	0.2	0.3
μ_y	0.2	0.3	μ_y	0.2	0.3
σ_x	0.04	0.04	σ_x	0.04	0.04
σ_y	0.04	0.04	σ_y	0.04	0.04
τ_x	0.1	0.1	τ_x	0.1	0.1
τ_y	0.1	0.1	τ_y	0.1	0.1

2.3.2. Experimental apparatus and procedures

Prior to the first chromatographic separation all samples were transferred into the appropriate starting buffer through size-exclusion chromatography with disposable PD-10, Sephadex G-25 columns (GE-Healthcare, Uppsala, Sweden) according to the manufacturer protocol. A long-range cation exchange gradient chromatofocusing step carried out on an Äkta Explorer 10 system equipped with a Mono S 4.6/100 column (GE-Healthcare, Uppsala, Sweden) according to a protocol described elsewhere [29] served as the first separation dimension. A total of 96 fractions of 200 µl were collected at regular intervals over the course of the gradient. The frequency of the fractionation was based on the anticipated peak width to increase the chance of the relevant peaks to be sampled multiple times.

All collected fractions were sealed and transferred to a WPS-3000TXRS In-Line Split-Loop Autosampler (Thermo Fisher Scientific, Breda, The Netherlands) cooled to 4°C. The second chromatographic separation was size-exclusion chromatography performed on an Acquity BEH 200 column (Waters, Etten-Leur, The Netherlands) mounted in a TCC-3000RS Thermostat heated to 30°C. The column dimensions were 4.6x150mm, preceded by a 4.6x50mm guard column of the same resin and a 0.2µm stainless steel in-line filter (Waters, Etten-Leur, The Netherlands). The mobile phase was a 100 mM sodium phosphate buffer at pH 6.8, driven by a LPG-3400RS Quaternary Gradient Pump. Samples of 10 µl were injected sequentially in 12min intervals. The UV absorption at the column outlet was monitored by a VWD-3400RS detector set to 280 nm and a 100 Hz sampling rate.

2.3.3. Data curation

To construct a two-dimensional chromatogram from the individual second dimension datasets, a vector with a number of elements equal to the number of fractions analysed in the second dimension was created. Each element of this vector was assigned a value equal to the centroid of the corresponding fractionation interval.

The output signals of the second dimension were aligned to share a common time base and stored in a matrix. This matrix was then cut to only include measurements during the time of interest. In the case of size-exclusion chromatography this meant to remove data points collected during the lag-phase and after the elution of very small molecules. To reduce the burden on the optimization algorithms the sampling rate of the chromatograms was reduced by averaging over constant intervals to achieve a resolution of 200 data points. This value was chosen as a compromise between loss of resolution and reduction of computational time. To compensate for offsets between the second dimension measurements each row of the second dimension is corrected for a linear baseline, so that the first and last column of the measurement matrix equal zero. This step needs to be closely controlled as inaccuracies in the baseline correction introduces noise and can lead to the introduction of false peaks.

A vector with the same length as the second dimension of the measurement matrix is constructed and filled with the second dimension time-base values corresponding to the measurement intervals after resampling. Both first and second dimension time-base vectors are repeated in the opposite dimensions to form two time-base matrices of the same dimensions as the measurement matrix.

2.3.4. Fourier Transform assisted sharpening of the peak profiles

The parameters of the transfer function were estimated by fitting the mEMG function to a series of well resolved model protein chromatograms generated under the conditions of the experiment (data not shown). For each parameter the lowest found value was chosen and multiplied with a safety margin of 0.9 to prevent the introduction of false peaks and negative side lobes. The deconvolution signal matrix was then calculated with the time-base matrices as input and, the estimated variance and relaxation parameters. The modes were chosen for the peak to be roughly located in the middle of the respective dimension time interval. The elements of the resulting response matrix were then shifted for the maximum to be located at the position corresponding to the origin of the chromatogram. Elements left empty after the shift were filled with zeros.

The discrete Fourier transforms of the measurement matrix and deconvolution matrix were computed by the two-dimensional fast Fourier transforms algorithm implemented in the signal processing toolbox of Matlab 2013b (Mathworks, Natick, USA). In the frequency domain the transformed measurement matrix was divided by the transformed deconvolution matrix. The resulting matrix was transformed back into the time domain by the two-dimensional inverse fast Fourier transform algorithm of the same toolbox, yielding a version of the chromatogram with increased resolution. To remove noise introduced by the transformations the resulting matrix was smoothed by a digital Gaussian low-pass filter. To compensate for changes in signal amplitude caused by the procedure, the entire matrix is multiplied by a scalar factor determined by the ratio of the signal maxima of the original and deconvoluted measurement matrices.

2.3.5. Peak fitting

After sharpening of the chromatogram the resulting matrix was scanned for local maxima. To be identified as a local maximum each data point was checked for the following criteria: its value needed to be greater than 2% of the maximum signal intensity and no data point within 2% of the matrix dimensions is larger than the considered data point. The coordinates of points fulfilling both criteria were used to extract the corresponding retention times from the time-base matrices, to serve as initial estimates for the modes of each peak. The variances and relaxation parameters of the deconvolution peak were used as the initial estimates of the corresponding parameters. The combination of all parameter sets determined this way were rearranged into a vector that served as the starting values for the optimization algorithm.

The optimization was carried out by the trust-region-reflective algorithm of Matlab 2013b (Mathworks, Natick, USA), with zero as the lower bound on all parameters. No upper bounds were set. The optimizer was set to terminate when either the function or the norm of the step was smaller than 10E-6 or when a maximum of 500 iterations had been performed.

2.3.6. Statistical output analysis

Once the optimizer had reached one of the termination criteria, the output vector was split into single peak parameter sets. For each parameter set the lower rank moments were calculated numerically. The standard error $s_{\hat{\beta},j}$ for each parameter j was estimated by Eq. (11), where n is the number of data points, p the number of parameters and RSS the sum of squared residuals of the fit.

$$s_{\hat{\beta},j} = \sqrt{\frac{RSS}{n-p}} \cdot \sqrt{c_{jj}} \tag{11}$$

The matrix c of which the j^{th}-diagonal elements are extracted is calculated by Eq. (12) from the Jacobian matrix J, as returned by the optimizer algorithm.

$$c = (J' \cdot J)^{-1} \tag{12}$$

For the simulated data, the deviations from the expected parameters were calculated as an additional measure for the quality of the deconvolution. The errors $s_{\hat{\beta},M}$ for the numerically calculated peak moments M were estimated by Eq. (13), where P_j is the j^{th}-parameter of the fit.

$$s_{\hat{\beta},M} = \sqrt{\sum \left(\left(\frac{\partial M}{\partial P_j} \right)^2 \cdot {s_{\hat{\beta},j}}^2 \right)} \tag{13}$$

The partial derivatives were evaluated numerically by varying each parameter by ±0.1% of its value and calculating the inclination over the resulting interval.

2.4 Results and discussion

2.4.1. Importance of resolution

Two model scenarios were analyzed to evaluate the robustness of the deconvolution procedure. The difference between the scenarios lies in the resolution achieved by the multidimensional separation. In Scenario 1 the two peaks only differ in their retention in a single dimension, causing them to appear completely fused, as depicted in Figure 2.2. Detection of two separate peaks by the local maxima approach is only possible after sharpening of the chromatogram with the help of Fourier Transforms. In the Scenario 2, the two peaks show an equal difference in retention in both dimension, together sufficient for two local maxima to be detectable

even prior to Fourier Transform sharpening. In both cases the described approach leads to the identification of the correct number of peaks and the determined parameters reproducibly deviate by less than 1.5% of their value from the parameters used for the creation of the model data.

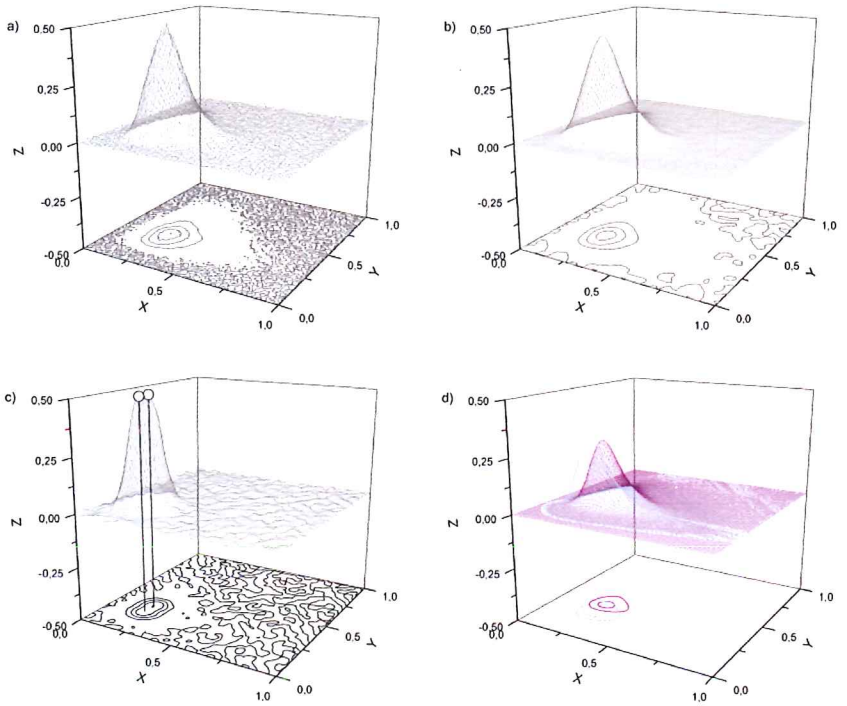


Figure 2.2. Example data during various stages of the deconvolution procedure: a) raw simulated data of two overlapping peaks with a 2% normal distributed measurement error and 2% of the maximum signal as background noise. b) after smoothing with a Gaussian low-pass filter. c) sharpening of the peak profile with Fourier Transforms reveals two distinct local peak maxima above the rejection threshold. d) Non-linear fitting of the 2D-mEMG function to the raw data allows reconstruction of the single peaks.

A noteworthy difference between the two scenarios only becomes apparent when the statistical errors of the determined parameters are considered, as in Fig. 2.3 and 2.4. While the error on each parameter is less than 2% when the resolution is sufficient for each peak to show a local maximum in the raw data, some parameters show large statistical errors when that is not the case. The peak heights and the relaxation parameters of the dimension in which the peaks do not differ in their retention appear to be most strongly affected, as can clearly be seen in Fig. 2.3 b-d. The reason for both parameters being associated with a large statistical error for the case of completely fused peaks lies in their correlated influence on the observed peak amplitude. When the tailing part of one peak is completely hidden within another peak, it becomes difficult to determine whether the observed signal intensity is due to tailing of the first, or height of the second peak. This ambiguity is then reflected in comparatively large statistical errors on these parameters. Although the difference is far more subtle for other parameters, it can be seen in Fig. 2.3 that errors for parameters associated with a dimension in which no separation occurs are generally larger than their counterparts. For a symmetrical case such as Scenario 2 there is no difference between the errors on any dimension.

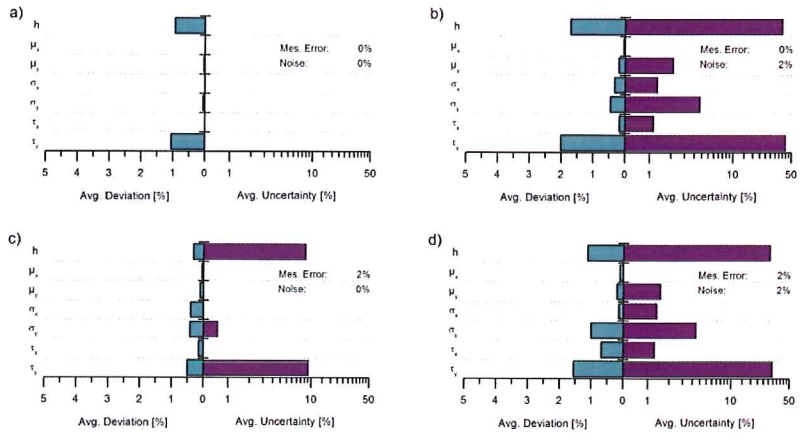


Figure 2.3. Overview of the impact of artificial normal distributed pseudo-random background noise and measurement errors on the peak parameters determined by the proposed deconvolution procedure. All reported values are averages from at least three repetitions. The base case was calculated with the two-dimensional mEMG function and the Scenario 1 parameters of Table 2.1.

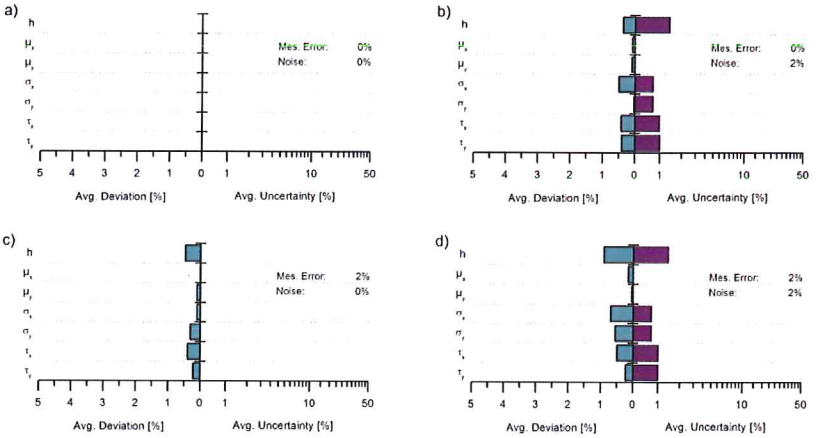


Figure 2.4. Impact of artificial background noise and measurement errors on the peak parameters determined by the proposed deconvolution procedure, applied to the model Scenario 2, given by the parameters in Table 2.1.

2.4.2. Influence of data quality

The influence of a constant background noise and a measurement error dependent on the local signal intensity were investigated separately. A constant background noise had on average a stronger negative influence on the statistical error of the fitted parameters than the artificial measurement errors. This is inherent to how the statistical errors are calculated. The background noise is present across the entire chromatogram, rather than only the area covered by the fused peak. As a consequence the background noise has a systematically larger contribution to the residuals of the fit than the measurement error. For the same reason, no significant errors are reported when fitting to perfect data, even when the determined parameters do not perfectly match the expected values. Noise and measurement error levels of 1% and larger than 2% were also tested, as were their combinations. For 1% imperfection levels the trends were the same as reported in Fig. 3 and Fig.4, but with lower errors corresponding to a lower contribution of the imperfections to the residuals of the fit. Higher levels of imperfections sometimes lead to the detection of false peaks. This effect can be suppressed by applying stricter filters, but as knowledge of the expected number of peaks is unrealistic for real applications, the procedure is not recommended for too noisy data.

2.4.3. Characterization of complex mixtures

Application of the described approach to experimental data collected during multi-dimensional chromatography of an IgG-1 producing Chinese hamster ovary (CHO) cell-culture supernatant of unknown composition gave results consistent with the behavior of the simulated cases. Sharpening of the chromatogram showed an increase in local maxima detected for broad and highly fused peak clusters, such as peaks 4, 5 an 7in Fig. 5. Peaks that were relatively well resolved on the other hand simply appear to slightly change in

shape without revealing any hidden peaks. An overall trend in the data is that position related peak parameters such as their Gaussian modes and calculated first moments, are mostly statistically better defined than parameters related to quantification, such as peak heights, variances and volumes (zeroth moments), except for small peaks surrounded by larger peaks from multiple sides, or peaks that appear to not fit the EMG shape. An example for this case is given by peak 2 in Fig. 5, where a secondary local maxima is present, but below the peak acceptance threshold. Should this occur unacceptably often the peak rejection criteria should be revised and use of another peak model more suitable for the particular case should be considered. The uncertainty of parameters may also increase when small peaks are in the tailing zone of larger ones, such as peak 8.

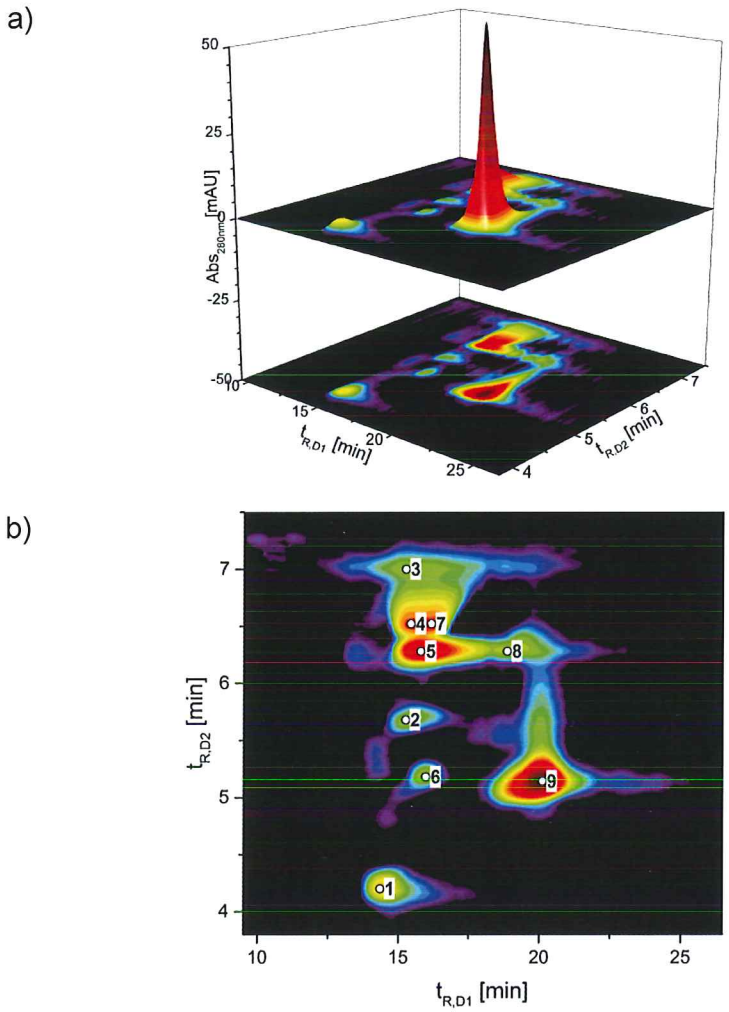


Figure 2.5. Example data during various stages of the deconvolution procedure: a) raw simulated data of two overlapping peaks with a 2% normal distributed measurement error and 2% of the maximum signal as background noise. b) after smoothing with a Gaussian low-pass filter. c) sharpening of the peak profile with Fourier Transforms reveals two distinct local peak maxima above the rejection threshold. d) Non-linear fitting of the 2D-mEMG function to the raw data allows reconstruction of the single peaks.

Table 2.2 Fitted EMG parameters and numerically calculated peak moments of the peaks detected during the two-dimensional chromatographic analysis of the IgG-1 containing CHO cell culture supernatant.

Peak	h	μ_x	μ_y	σ_x	σ_y	τ_x	τ_y	M_0	$M_{1,x}$	$M_{1,y}$
1	9.0 ±1.3	14.2 ±0.0	4.2 ±0.0	0.2 ±0.0	0.1 ±0.0	0.8 ±0.1	0.1 ±0.0	1.0 ±0.2	15.0 ±0.1	4.3 ±0.0
2	3.6 ±1.9	15.0 ±0.1	5.7 ±0.0	0.2 ±0.1	0.1 ±0.0	1.2 ±0.2	0.1 ±0.1	0.4 ±0.3	16.1 ±0.2	5.8 ±0.1
3	3.9 ±1.2	14.7 ±0.1	7.0 ±0.0	0.3 ±0.1	0.1 ±0.0	1.8 ±0.3	0.1 ±0.1	1.0 ±0.5	16.5 ±0.3	7.1 ±0.1
4	14.4 ±3.5	15.2 ±0.0	6.5 ±0.0	0.2 ±0.0	0.2 ±0.0	0.8 ±0.2	0.1 ±0.1	2.5 ±0.7	15.9 ±0.2	6.6 ±0.1
5	26.7 ±2.3	15.5 ±0.0	6.3 ±0.0	0.3 ±0.0	0.1 ±0.0	0.7 ±0.0	0.1 ±0.0	2.4 ±0.3	16.3 ±0.0	6.4 ±0.0
6	2.6 ±1.0	15.9 ±0.1	5.2 ±0.0	0.2 ±0.0	0.1 ±0.0	0.2 ±0.1	0.1 ±0.0	0.2 ±0.1	16.1 ±0.1	5.3 ±0.0
7	4.6 ±1.0	15.9 ±0.0	6.6 ±0.0	0.3 ±0.0	0.3 ±0.0	0.7 ±0.1	0.0 ±0.0	2.7 ±0.7	16.6 ±0.1	6.6 ±0.0
8	4.1 ±1.4	18.6 ±0.1	6.3 ±0.0	0.2 ±0.1	0.1 ±0.0	1.8 ±0.2	0.1 ±0.1	0.7 ±0.4	20.3 ±0.2	6.4 ±0.1
9	94.0 ±0.8	20.0 ±0.0	5.2 ±0.0	0.3 ±0.0	0.1 ±0.0	0.1 ±0.0	0.1 0.0	10.9 ±0.1	20.2 ±0.0	5.3 ±0.0

2.5 Conclusions

The comparison of the two simulated scenarios makes a strong case for the use of multi-dimensional techniques for the analysis of complex mixtures. The increased system peak capacity strongly enhances the chance for the components to be sufficiently resolved for showing local maxima, leading to significantly better statistics on the parameters determined by fitting peak models to the experimental data with the help of algorithmic optimizers. Good initial estimates of the number and positions of the peaks are crucial to the success of such fitting approaches. Local peak maxima detection in combination with enhancing the resolution of the chromatogram through removal of system broadening effects by Fourier transforms, was shown to be a robust and computationally efficient means to generate these starting estimates. As the fit is performed to the original data, signal distortion, noise amplification and generation of artefacts such as negative side lobes usually associated with Fourier transform chromatogram sharpening are not of major concern here.

The comparatively small errors on the peak locations and first moments make the technique well suitable for applications such as component identification or for the regression of thermodynamic parameters [30]. Robust quantitative results are achieved when resolution is sufficient for local maxima to be detected prior to resolution enhancement. When quantitative analysis of all components is the main objective, the detection of a highly fused peak cluster should lead to optimization of the experimental protocol to increase the resolution in that zone of the chromatogram. When further chromatographic separation is not feasible, the introduction of multichannel detectors [31] and expansion of the approach to a hybrid MCR should be considered.

Acknowledgements

This work was supported by the BE-Basic foundation, a public private partnership of knowledge institutes, industry and academia, under the project FS2.003. We want to thank out industrial partners for financial support and valuable input. We acknowledge Synthon Biopharmaceuticals B.V. for providing the CHO cell culture supernatant analysed in this project.

References

[1] O. Kaltenbrunner, A. Jungbauer, S. Yamamoto, Prediction of the preparative chromatography performance with a very small column, J. Chromatogr. A 760 (1997) 41-53.

[2] J.P. Foley, Equations for Chromatographic Peak Modeling and Calculation of Peak Area, Anal. Chem. 59 (1987) 1984-1987.

[3] A.N. Papas, T.P. Tougas, Accuracy of Peak Deconvolution Algorithms within Chromatographic Integrators, Anal. Chem. 62 (1990) 234-239.

[4] C. Ruckebusch, L. Blanchet, Multivariate curve resolution: a review of advanced and tailored applications and challenges, Anal. Chim. Acta 765 (2013) 28-36.

[5] A. de Juan, J. Jaumot, R. Tauler, Multivariate Curve Resolution (MCR). Solving the mixture analysis problem, Analytical Methods 6 (2014) 4964.

[6] D.W. Osten, B.R. Kowalski, Multivariate Curve Resolution in Liquid-Chromatography, Anal. Chem. 56 (1984) 991-995.

[7] F. Dismer, S. Hansen, S.A. Oelmeier, J. Hubbuch, Accurate retention time determination of co-eluting proteins in analytical chromatography by means of spectral data, Biotechnol. Bioeng. 110 (2013) 683-693.

[8] L. Blanchet, C. Ruckebusch, J.P. Huvenne, A. de Juan, Hybrid hard- and soft-modeling applied to difference spectra, Chemom. Intell. Lab. Syst. 89 (2007) 26-35.

[9] V.B. Di Marco, G.G. Bombi, Mathematical functions for the representation of chromatographic peaks, J. Chromatogr. A 931 (2001) 1-30.

[10] Y. Kalambet, Y. Kozmin, K. Mikhailova, I. Nagaev, P. Tikhonov, Reconstruction of chromatographic peaks using the exponentially modified Gaussian function, J. Chemom. 25 (2011) 352-356.

[11] R.A. Vaidya, R.D. Hester, Deconvolution of Overlapping Chromatographic Peaks Using Constrained Non-Linear Optimization, J. Chromatogr. 287 (1984) 231-244.

[12] J.R. Torres-Lapasió, J.J. Baeza-Baeza, M.C. García-Alvarez-Coque, A Model for the Description, Simulation,

and Deconvolution of Skewed Chromatographic Peaks, Anal. Chem. 69 (1997) 3822-3831.

[13] P. Nikitas, A. Pappa-Louisi, A. Papageorgiou, On the equations describing chromatographic peaks and the problem of the deconvolution of overlapped peaks, J. Chromatogr. A 912 (2001) 13-29.

[14] G. Vivo-Truyols, J.R. Torres-Lapasio, A.M. van Nederkassel, Y. Vander Heyden, D.L. Massart, Automatic program for peak detection and deconvolution of multi-overlapped chromatographic signals part II: peak model and deconvolution algorithms, J. Chromatogr. A 1096 (2005) 146-155.

[15] G. Vivo-Truyols, J.R. Torres-Lapasio, A.M. van Nederkassel, Y. Vander Heyden, D.L. Massart, Automatic program for peak detection and deconvolution of multi-overlapped chromatographic signals part I: peak detection, J. Chromatogr. A 1096 (2005) 133-145.

[16] J.C. Giddings, Sample dimensionality: a predictor of order-disorder in component peak distribution in multidimensional separation, J. Chromatogr. A 703 (1995) 3-15.

[17] G. Vivo-Truyols, S. van der Wal, P.J. Schoenmakers, Comprehensive study on the optimization of online two-dimensional liquid chromatographic systems considering losses in theoretical peak capacity in first- and second-dimensions: a Pareto-optimality approach, Anal. Chem. 82 (2010) 8525-8536.

[18] J.N. Fairchild, K. Horvath, G. Guiochon, Theoretical advantages and drawbacks of on-line, multidimensional liquid chromatography using multiple columns operated in parallel, J. Chromatogr. A 1216 (2009) 6210-6217.

[19] G. Guiochon, N. Marchetti, K. Mriziq, R.A. Shalliker, Implementations of two-dimensional liquid chromatography, J. Chromatogr. A 1189 (2008) 109-168.

[20] N.A. Wright, D.C. Villalanti, M.F. Burke, Fourier-Transform Deconvolution of Instrument and Column Band Broadening in Liquid-Chromatography, Anal. Chem. 54 (1982) 1735-1738.

[21] A. Felinger, Deconvolution of Overlapping Skewed Peaks, Anal. Chem. 66 (1994) 3066-3072.

[22] A. Economou, P.R. Fielden, A.J. Packham, Deconvolution of overlapping chromatographic peaks by means of fast Fourier and Hartley transforms, Analyst 121 (1996) 97-104.

[23] F. Dondi, A. Betti, L. Pasti, M.C. Pietrogrande, A. Felinger, Fourier-Analysis of Multicomponent Chromatograms - Application to Experimental Chromatograms, Anal. Chem. 65 (1993) 2209-2222.

[24] F. Dondi, M.C. Pietrogrande, A. Felinger, Decoding complex multicomponent chromatograms by Fourier Analysis, Chromatographia 45 (1997) 435-440.

[25] A. Felinger, L. Pasti, F. Dondi, Fourier-Analysis of Multicomponent Chromatograms - Theory and Models, Anal. Chem. 62 (1990) 1846-1853.

[26] A. Felinger, L. Pasti, P. Reschiglian, F. Dondi, Fourier-Analysis of Multicomponent Chromatograms - Numerical Evaluation of Statistical Parameters, Anal. Chem. 62 (1990) 1854-1860.

[27] B.K. Nfor, T. Ahamed, M.W. Pinkse, L.A. van der Wielen, P.D. Verhaert, G.W. van Dedem, M.H. Eppink, E.J. van de Sandt, M. Ottens, Multi-dimensional fractionation and characterization of crude protein mixtures: toward establishment of a database of protein purification process development parameters, Biotechnol. Bioeng. 109 (2012) 3070-3083.

[28] F. Kroner, D. Elsasser, J. Hubbuch, A high-throughput 2D-analytical technique to obtain single protein parameters from complex cell lysates for in silico process development of ion exchange chromatography, J. Chromatogr. A 1318 (2013) 84-91.

[29] F. Kroner, A.T. Hanke, B.K. Nfor, M.W. Pinkse, P.D. Verhaert, M. Ottens, J. Hubbuch, Analytical characterization of complex, biotechnological feedstocks by pH gradient ion exchange chromatography for purification process development, J. Chromatogr. A 1311 (2013) 55-64.

[30] A.T. Hanke, M. Ottens, Purifying biopharmaceuticals: knowledge-based chromatographic process development, Trends Biotechnol. 32 (2014) 210-220.

[31] R.F. Lacey, Deconvolution of Overlapping Chromatographic Peaks, Anal. Chem. 58 (1986) 1404-1410.

3

3D-liquid chromatography as a complex mixture characterization tool for knowledge-based downstream process development

Abstract

Knowledge-based development of chromatographic separation processes requires efficient techniques to determine the physicochemical properties of the product and the impurities to be removed. These characterization techniques are usually divided into approaches that determine molecular properties, such as charge, hydrophobicity and size, or molecular interactions with auxiliary materials, commonly in the form of adsorption isotherms. In this study we demonstrate the application of a three-dimensional liquid chromatography approach to a clarified cell homogenate containing a therapeutic enzyme. Each separation dimension determines a molecular property relevant to the chromatographic behaviour of each component. Matching of the peaks across the different separation dimensions and against a high-resolution reference chromatogram allows to assign the determined parameters to pseudo-components, allowing to determine the most promising technique for the removal of each impurity. More detailed process design using mechanistic models requires isotherm parameters. For this purpose, the second dimension consists of multiple linear gradient separations on columns in a high-throughput screening compatible format, that allow regression of isotherm parameters with an average standard error of 8%.

Keywords: Multi-dimensional chromatography, host cell proteins, process development, feedstock characterization

3.1 Introduction

The technical and regulatory challenges of biopharmaceutical downstream processing lead to the widespread adoption of high-throughput screening (HTS) techniques [1]. In this context, process performance and the screening results are typically correlated through the use of statistical tools [2,3]. To gain a higher level of process understanding, there is an ongoing trend to utilize both the knowledge of the molecular properties of the components to be separated and the specific interaction of each of these molecules with auxiliary materials such as chromatography resins [4]. The molecular properties are used to rationally identify promising separation strategies through the use of heuristic design rules [5], whereas the interaction parameters are mostly used for detailed process designs and unit operation optimizations using mechanistic models [6-8].

A key challenge in applying these approaches to real purification problems is finding experimental techniques that are able to determine the necessary property and interaction parameters in a time and material efficient manner. The proteins in the mixture should be characterized towards their charge, hydrophobicity and size properties, as these are the fundamental properties utilized during most chromatographic purifications [9], other than specific biological affinities [10,11]. Combinations of 2D-gels with other techniques have been demonstrated to be effective tools at determining these parameters for multiple components of complex mixtures [12]. To be able to predict process performance from these parameters alone is possible, but requires extensive model calibration [13].

Recent years have seen the development of multi-dimensional techniques to allow the application of proven parameter estimation strategies such as regression from the retention times in multiple linear salt gradients directly to complex samples [14,15].

The downside to these techniques is that they still require extensive manual labour in terms of experimentation and data processing. In this study we demonstrate a three-dimensional experimental approach that characterizes the major components of a complex protein mixture towards their charge, hydrophobicity and size. The approach is designed to allow for a high degree of automation in both experimentation and data analysis. The second dimension is designed to be a HTS compatible format that also allows the regression of adsorption isotherm parameters.

3.2 Materials and methods

3.2.1. Approach overview

The complex sample used for this study is an extract of an intracellularly produced therapeutic enzyme clarified after homogenization, provided by industrial partners (DSM, The Netherlands). An overview of the experimental approach taken to characterize the chromatographic behaviour of the product and major protein impurities in the sample is shown in Figure 1. Gradient chromatofocusing (gCF), or linear pH-gradient chromatography, on a strong anion exchange column was used as a universal first separation dimension. Linear gradient experiments with gradient lengths varying from 12 to 36 column volumes (CV) were performed on each of the collected fractions, to allow regression of isotherm parameters. To assess if this step can be performed in a HTS compatible format, these experiments were carried out on 200 µl RoboColumns (Atoll, Germany) packed with a resin for Hydrophobic interaction chromatography (HIC). To compensate for the inherently low resolution of such small columns, the collected fractions of these gradient experiments were subjected to high-resolution size-exclusion chromatography. As gradient experiments in HIC require adjusting the fractions to relatively high salt concentrations which may induce precipitation, a reference data set skipping the salt-gradient separation is performed. The two data sets

are compared to see if any major contamination peaks are lost due to precipitation or if any new detectable aggregates are formed.

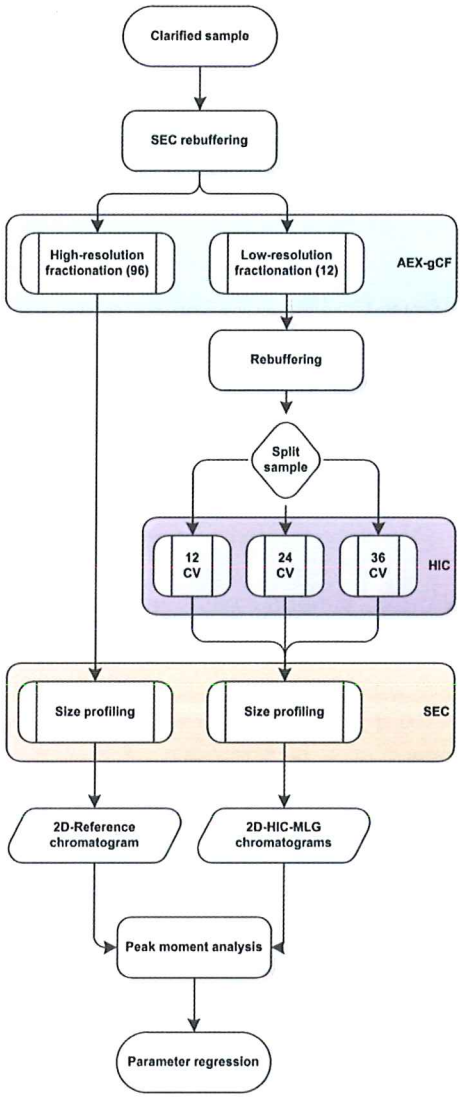


Figure 3.1 Schematic overview of the procedure for producing the three-dimensional data for isotherm parameter regression together with the two-dimensional reference chromatogram.

3.2.2. pH-gradient anion-exchange prefractionation

Prior to gCF, the cell free clarified extract was transferred into the appropriate starting buffer through SEC on disposable PD-10, Sephadex G-25 columns (GE-Healthcare, Uppsala, Sweden) according to the manufacturer recommendations. The gCF was performed on a Mono Q 4.6/100 column (GE-Healthcare, Sweden) with a gradient length of 15 column volumes (CV) on an Äkta Explorer 10 liquid chromatography system (GE-Healthcare, Sweden). The buffers for the pH-gradient were made according to Kröner et al. for a pH-range of 10.5 to 3.5 for AEX-gCF [16]. For creation of the reference data set a total of 96 fractions were collected during the gradient, whereas 12 fractions were collected for three-dimensional analysis.

3.2.3. Size exclusion chromatography

All SEC experiments were performed on a UHPLC$^+$ (Thermo Fisher Scientific, MA, USA) equipped with a WPS-3000TXRS in-line split-loop autosampler, TCC-3000RS column thermostat set to 30 °C, LPG-3400RS Quaternary Gradient Pump. The UV absorption at the column outlet was monitored by a VWD-3400RS detector ay 230nm. The stationary phase was a 4.6 x 150mm Acquity BEH 200 column, preceded by a 4.6 x 50mm guard column of the same resin and a 0.2 µm stainless steel in-line filter Waters, MA, USA). The used mobile phase was a 100mM sodium phosphate buffer at pH 6.8 applied at a flowrate of 0.3 ml/min. The injection volume per measurement was 10 µl. The SEC system was calibrated with gel filtration chromatography standard mixture for a molecular weight range of 1.3 – 670 kDa (Bio-Rad, CA, USA). Instead of calibrating towards the molecular weight of each protein (M_r), the system was calibrated towards their hydrodynamic radius (R_{Hydro}) which was estimated through Eq. (1) [17]:

$$R_{Hydro,i,Prot} = 0.081 \cdot (M_r)^{\frac{1}{3}} \quad (1)$$

3.2.4. Mini-column characterization

The mini columns used for this study were 200 µl RoboColumns packed with Cellufine Phenyl resin. To allow recording of continuous signal chromatograms the columns were mounted on an Äkta Explorer 10 with a custom made adapter. The column packing and porosity were evaluated through inverse size exclusion chromatography (ISEC) with a range of different molecular weight Dextran standards (American Polymer Standards, OH, USA). A full list of the used Dextran standards is given in Table 1. The peak molecular weight (M_p) of each Dextran was related to its hydrodynamic radius (R_{Hydro}) according to the empirical correlation in Eq. (2), as reported by To et al. [18].

$$R_{Hydro,i,Dex} = 0.0271 \cdot (M_p)^{0.498} \quad (2)$$

Dextran standard solutions were prepared at a concentration of 2 g/l. The injection volume of all ISEC experiments was 10 µl and the flowrate was set to 0.156 ml/min. To correct for system contributions, all standards were measured once with a column attached and once with a bridge capillary with the same dead volume as the RoboColumn adapter. The refractive index was measured at the column outlet by a 1100 series RID (Agilent, CA, USA) From the corrected retention values the smallest and largest Dextrans were used to calculate the column bed (ϵ_b), total (ϵ_t) and particle (ϵ_p) porosities. The pore accessibility coefficients (K_D) for the intermediate Dextrans were calculated in relation to those values. For interpolation a logistic function was fitted through the calculated K_D in relationship to the probe hydrodynamic radii follow Eq. (3), where p and r_m are empirical fitting coefficients.

$$K_{D,i} = \frac{1}{1 + \left(\frac{R_{Hydro,i}}{r_m}\right)^p} \quad (3)$$

Table 3.1 Peak molecular weights and empirically calculated hydrodynamic radii of Dextran standards used for ISEC experiments.

Mp [Da]	Rh [nm]
180	0.36
18500	3.51
29200	4.54
85000	7.72
410000	16.91
970000	24.97
2825000	44.22
2900000	44.80
6300000	65.92

3.2.5. Multi-linear gradient experiments on small columns

Linear salt gradient HIC experiments from 1000 mM to 0 mM Ammonium Sulphate (AS) over 12, 24 and 36 CV were performed on the same liquid chromatography setup and column as the ISEC experiments. The pH was kept constant at 7.5 during all HIC experiments. To facilitate detection in the third dimension the injection volume was increased to 50 µl per experiments. Prior to sample injection the samples were rebuffered into 25 mM Tris buffer containing 1000 mM AS with Amicon spin filters with a nominal molecular weight cut-off of 3 kDa (Millipore, MA, USA). The rebuffering was controlled to achieve a concentration factor of at least 2.5 to allow tracing low concentration contaminants. Fractions were collected at regular intervals of 275 µl throughout the HIC gradient experiments.

3.2.6. Multi-dimensional peak analysis

Composite multidimensional chromatograms were constructed for the two-dimensional AEX-gCF-SEC reference data, as well as for each HIC gradient for each fraction with a significant UV-detectable protein content. The peaks in each complex composite chromatogram were detected, deconvoluted and fitted with a Fourier-transform assisted technique reported elsewhere [19]. To increase the methods sensitivity the absorption signal at 230 nm was used instead of 280 nm. The peak acceptance threshold for the reference chromatogram was set to 3 mAU and 0.5 mAU for the HIC gradient data.

3.2.7. Parameter regression

Peaks with the same hydrodynamic radius and a similar volume across the different salt-gradient chromatograms of the same pH-gradient fraction were paired together and assigned an identity based on the largest similarity to a peak from the reference chromatogram. To regress isotherm parameters from changes in elution volume according to changes in gradient length, a derivation of the formalism of Parente and Wetlaufer was used [20]. The regression model is based on Mollerup's description of the initial slope (A_i) of a hydrophobic interaction isotherm of a component 'i' [21].

$$A_i = \widetilde{K}_i \left(\frac{\Lambda}{c}\right)^n \tilde{\gamma}_i \qquad (4)$$

Here $\widetilde{K}_i$ is the equilibrium constant grouped together with the symmetric activity coefficient of the component in an infinite dilution of water, Λ is the ligand density, c the molarity of the solution, n the average number of binding sites and $\tilde{\gamma}_i$ the assymetric activity coefficent. When performing experiments only within the linear part of the isotherm, some parameters require lumping to avoid correlation errors during the regression. As Λ, c and n are

constants within the area of interest they can be lumped with the equilibrium constant:

$$K_{B,i} = \tilde{K}_i \left(\frac{\Lambda}{c}\right)^n \tag{5}$$

The asymmetric activity coefficient is given by:

$$\tilde{\gamma}_i = \exp(K_{s,i} c_s + K_{p,i} c_i) \tag{6}$$

with $K_{s,i}$ and $K_{p,i}$ being parameters dependent on the salt, the protein and the pH. The salt concentration is c_s and $c_{p,i}$ is the protein concentration. For low protein concentrations this is very close to the standard state activity coefficient:

$$\tilde{\gamma}_{i,0} = \exp(K_{s,i} c_s) \tag{7}$$

This further reduces the complexity of the isotherm model to:

$$ln(A_i) = \ln(K_{B,i}) + K_{s,i} c_s \tag{8}$$

This is mathematically identical to the model for which Chen et al. demonstrated very good correlation between isocratic and gradient retention in HIC, except that Ai takes the place of the retention coefficient k'_1 [22] . As discussed by Mollerup, the two are similar but not interchangeable and are related to each other by [21]:

$$k'_i = \frac{(1-\varepsilon_b)\varepsilon_p K_{D,i}}{\varepsilon_b}(1 + A_i) \tag{9}$$

While more accurate, this definition does not allow for an analytical solution of the Parente and Wetlaufer relation. Assuming that the differences in retention caused only by the size-exclusion effect of the resin are minimal and retention is dominated by hydrophobic effects (Ai >> 1) leads to an approximation of the retention coefficient by:

$$k_i' \approx \frac{(1-\varepsilon_b)\varepsilon_p K_{D,i}}{\varepsilon_b} A_i \tag{10}$$

The analytical solution for the dependence of the gradient retention volume ($V_{R,g,i}$), corrected for instrumental dwell and gradient delay volumes, under these assumptions is:

$$V_{R,g,i} = \frac{V_G}{-K_{s,i}(c_{s,f}-c_{s,i})} \ln\left(1 + V_{column}(1-\varepsilon_b)\varepsilon_p K_{D,i} \frac{-K_{s,i}(c_{s,f}-c_{s,i})}{V_G} K_{B,i} e^{K_{S,i}c_{s,i}}\right) \tag{11}$$

with $c_{s,f}$ and $c_{s,i}$ being the final and initial salt concentrations of the gradient, V_{column} being the nominal volume of the column and V_G being the length of the salt gradient. Eq. (11) was used for weighted regression of the parameters $K_{B,i}$ and $K_{s,I}$ for each peak that could be matched across the different gradients by Matlab (Mathworks, MA, USA).

3.3 Results and discussion

3.3.1. Prefractionation reference data

Creating a high-resolution comprehensive two-dimensional chromatogram, by skipping the second screening dimension and instead directly analysing each fraction in the designated third dimension serves three purposes. The main purpose is to create a reference size to signal profile of the mixture, to allow the detection of changes in that profile that might be caused by the experimental conditions applied in the second dimension. This includes the complete disappearance or reduction in relative signal of certain peaks due to precipitation of that contaminant or the emergence of unexpected peaks due to aggregation effects. As skipping the second dimension avoids the dilution that can be caused by those experiments, it also allows detecting some of the trace contaminants of which the peak would otherwise drop below the detection limit. Both of these goals could be met by only skipping the second

dimension, but fractionating at the same resolution as when producing fractions for three-dimensional analysis. The higher fractionation resolution applied while creating the reference chromatogram serves to be able to determine the first moment of each peak in the first separation dimension, which corresponds to its elution-pH more accurately. In the three-dimensional experiments, each fraction roughly corresponds to a half pH unit step in the pH gradient, which would in most cases not allow to determine the elution-pH of each peak more accurately than within 0.5 pH units. The elution-pH of a product and its major impurities is a valuable asset for designing ion-exchange based separation of that mixture [23]. In some cases this information alone may already be enough to design a separation process achieving sufficient purity for some applications [24].

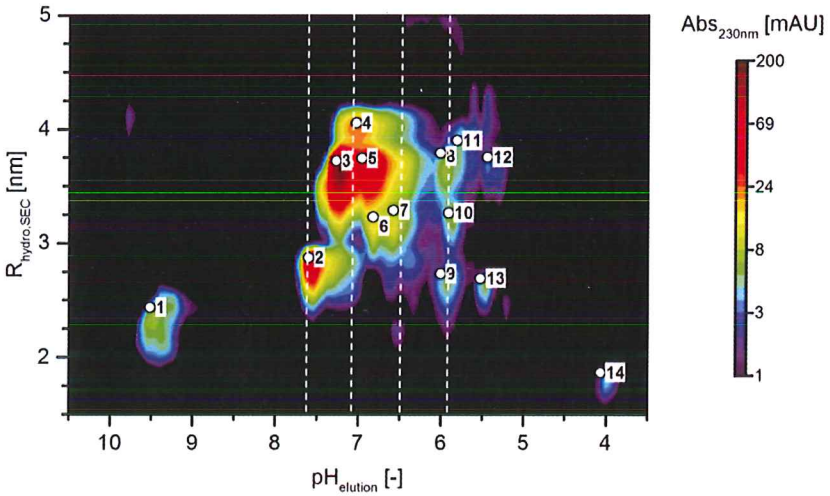


Figure 3.2 Two-dimensional reference chromatogram of the sample used in this study. The first separation dimension is a linear downwards gradient pH on a strong anion exchanger. The hydrodynamic radii ($R_{Hydro,SEC}$) are calculated from the retention times in SEC by Eq. (1). Peak labels are positioned at the Gaussian modes of the fitted peaks. The collection intervals of the fractions tested in the HIC dimension are also shown(dashed white)..

The high-resolution composite reference AEX-gCF-SEC chromatogram for the sample used in this study is shown in Figure 2. Each individual peak that could be detected and deconvoluted by the applied algorithm is labelled. These labels serve as the identities of the components associated with those peaks. Despite the high-resolution of the technique and deconvolution algorithm, each of these peaks might still in fact be the result of several overlapping peaks, but as they must be highly similar in two fundamental properties to elute so closely together, they will be treated as one pseudo-component in this study. The properties of each peak as calculated by the deconvolution algorithm are given in Table 2. The largest peak, both in height and volume, is peak 3 and is associated with the target product. The second largest peak, peak 5, appears to be of very similar size, but elutes 0.4 pH units later than the product. Components 2, 4, 6 and 7 also elute close enough to the product to be considered critical contaminants, whereas components 1 and 9-14 should be easily removed by an appropriate ion-exchange step. Some unlabelled peaks below the 3 mAU cut-off are also visible, but won't be considered further in this study, as they will fall below the detection limit in the three-dimensional analysis.

Table 3.2 Peak characteristics determined by Fourier transform assisted peak deconvolution of the two-dimensional AEX-gCF SEC reference chromatogram. The chromatogram is shown in Figure 2 and the peak IDs correspond to the labels assigned there. The pH of peak elution ($pH_{Elution}$) and the hydrodynamic radii ($R_{hydro,SEC}$) relate to the first moment of the peak in the corresponding separation dimension.

ID	$pH_{Elution}$ [-]	$R_{Hydro,SEC}$ [nm]	Peak volume [mAU*min^2]
1	9.4 ± 0.1	2.23 ± 0.06	4.2 ± 1.4
2	7.5 ± 0.0	2.74 ± 0.01	9.2 ± 0.8
3	7.2 ± 0.0	3.55 ± 0.01	25.9 ± 1.7
4	6.9 ± 0.0	3.77 ± 0.34	6.6 ± 4.9
5	6.8 ± 0.0	3.57 ± 0.06	17.4 ± 6.0

6	6.5 ± 0.9	2.88 ± 0.14	3.0 ± 7.7
7	6.4 ± 0.6	2.88 ± 0.27	1.9 ± 5.9
8	5.9 ± 0.7	3.44 ± 1.48	2.7 ± 9.0
9	5.9 ± 0.5	2.47 ± 0.23	1.1 ± 2.4
10	5.7 ± 0.5	2.82 ± 0.82	0.9 ± 3.5
11	5.5 ± 1.0	3.52 ± 2.33	1.0 ± 7.7
12	5.3 ± 0.4	3.37 ± 1.00	0.5 ± 1.8
13	5.4 ± 0.2	2.38 ± 0.17	1.0 ± 1.8

3.3.2. Column properties

For the characterization of the packing and particle properties, the retention measurements of the smallest and largest dextran are clearly the most important, as both the porosity and pore volume accessibility calculations depend on them. When working with very small column volumes, it becomes necessary to accurately account for system effects that otherwise may be neglected. This can clearly be seen in Figure 2a), where the difference in retention volume between smallest and largest Dextran lies at 130 µl without a column. With a column attached, the uncorrected difference in retention volume between these two markers is 260 µl, which would lead to physically impossible negative bed porosity values. Calculation of the column bed porosity (ε_b) assumes that the used marker is fully excluded from the pore volume. As the two largest Dextrans used in this study show the same retention volume despite a significant difference in their hydrodynamic radii this criterion is clearly satisfied. Additional corrections are made for the flow distributors and capillary connections, which amount to an additional 16 µl of hold-up volume in the frits and 14 µl in the connections. With these corrections in place, the bed porosity was calculated to be 0.30 and the particle porosity (ε_p) to be 0.93, both values similar to those of other HIC columns despite the smaller column size [18]. The dependence of the accessible pore volume for different probe sizes is shown in Figure 2b). Despite there being a clear impact on

the retention of larger probes, it appears that in the range up to 10 nm, in which most proteins are expected to lie, the size of the solute does not influence the pore accessibility to more than 10% and could be neglected in this particular case.

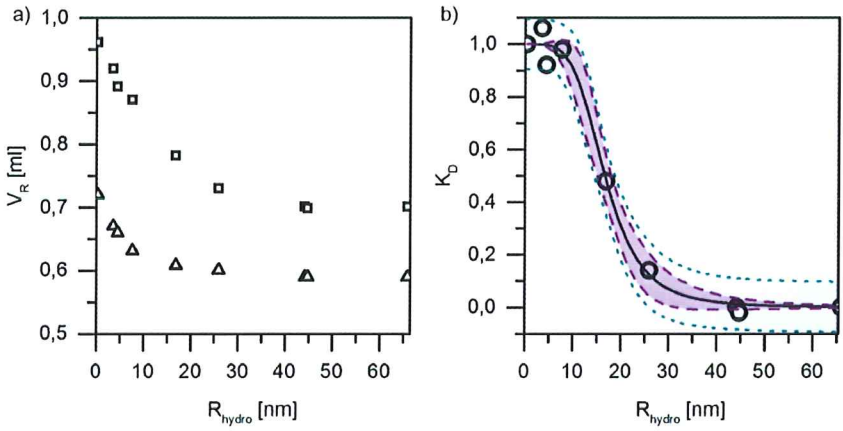


Figure 3.3 a) Uncorrected retention volumes of Dextran standards with column(□) and without column (Δ). b) Calculated K_D values of the Dextran standards (O) together the fitted K_D curve (solid black), its 95% confidence interval (dashed purple) and prediction interval (dotted teal).

3.3.3. Linear salt-gradient experiments

All collected fractions were subjected to linear-salt gradients of different gradient lengths. A HIC resin was chosen for this study, to serve as an orthogonal separation dimension to the charge-based prefractionation and size based third dimension, but in principle, any chromatography resin for which the interaction parameters should be determined could be used, if the gradient conditions are adapted accordingly. An exemplary chromatogram of a salt gradient of the different fractions is shown in Figure 4. There is no detectable peak in the fractions above pH 8, indicating that component 1 might have precipitated during the rebuffering step. Fractions collected between pH 7.6 and 5.5 show significant

absorbance peaks at 230 nm, which is in accordance with the results of the reference experiments. Unlike the reference data, each fraction appears to only contain a single albeit broad peak. The low apparent resolution of the separation is not unexpected, considering the columns have a bed height of only 5 mm.

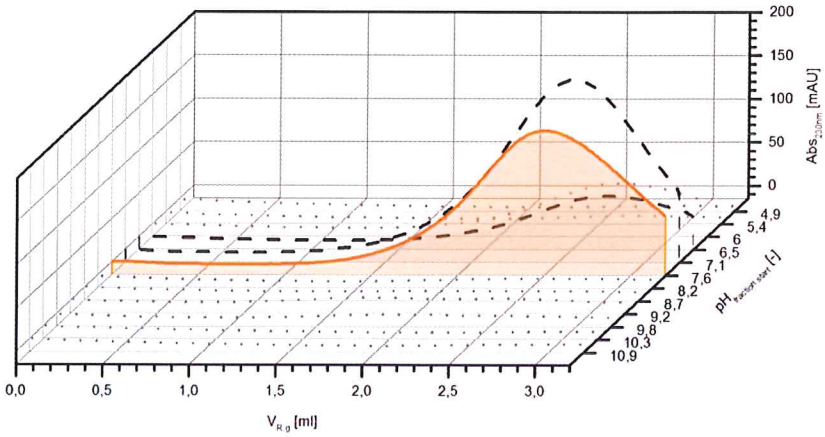


Figure 3.4 Linear gradient chromatograms of pH-gradient fractions on 200µl Cellufine Phenyl RoboColumns. The reported pH values correspond to the measured pH at the beginning of fraction collection. The gradient retention volumes in HIC (HIC-$V_{R,g}$) are corrected for instrument dwell, column dead volume and gradient delay. The first fraction to show a significant response in UV absorption (solid orange) is expected to contain most of the product. The following to fractions also show significant peaks (black dashed) and were also analysed in the third dimension. The chromatograms of the remaining fractions (grey dotted) were not analysed by SEC.

3.3.4. Multidimensional peak tracking

The criteria for matching peaks between different chromatograms was the hydrodynamic radius derived from the peaks first moment in the analytical SEC dimension. This choice was based on the property being almost concentration independent and the data

quality being very high, as the chromatograms are continuously recorded at a high sampling rate. The results of the peak matching are shown in Table 3. The standard deviation of the hydrodynamic radius between peaks that were matched to each other never exceeds 2% indicating a high degree of matching certainty. Based on this hydrodynamic radius and similarities in the relative peak volume the peaks were assigned identities corresponding to the peak labels in the reference chromatogram. In two cases this could not be done without some ambiguity, as indicated by the attribution of the peaks to two identities. Analysis of the fractionation range in the reference data reveals that in both cases both components are expected to be present, so the observed peaks are indeed most likely an unresolved combination of both components peaks. Matching peaks between different gradient lengths was possible without ambiguity, as even for the peaks of component 7 and 10 in the fraction collected between pH 6.5 and 6.0, that appear very similar both in radius and peak volume, the radius is still determined within a significant difference. Component 10 forms an exception, in that its peak was not detected in the 36 CV chromatogram. As the peak volume in the detected chromatograms is very small, it is most likely that it was diluted to below the detection limit in the longest gradient. While mass-transfer limitations apply, a longer gradient is generally expected to lead to a higher resolution. A higher resolution allows the peak fitting algorithm to determine peak properties with a higher certainty [19]. This is clearly seen for the standard errors of the calculated peak volumes, that drop from an average error of 81% for the 12 CV gradient, to 40% for the 24 CV gradient, and down to 26% for the 36 CV gradient. This trend would suggest that increasingly longer gradients should be used for parameter determination. On the downside this would decrease the methods throughput. This is especially true for multi-dimensional techniques, as longer gradients require collection of more fractions to achieve the same sampling resolution which would exponentially increase the time required in the subsequent analysis of those fractions.

Table 3.3 Peak tracking results across multiple linear gradients. The fraction range refers to the span of pH units during the linear pH-gradient. The hydrodynamic radii ($R_{Hydro,SEC}$) is the weighted average of the peak at different gradient lengths reported with the standard deviation of the measurements. The peaks are matched to peak identities (ID*) in the reference chromatogram by similarity in $R_{Hydro,SEC}$ and elution pH. The gradient retention volumes in HIC (HIC-$V_{R,g}$) are corrected for instrument dwell, column dead volume and gradient delay. The peak volumes were determined by a nonlinear least-squares fit to a two-dimensional EMG peak model. Peak volumes and retentions are reported with their standard error.

Fraction	$R_{Hydro,SEC}$	ID*	HIC-$V_{R,g}$			Peak volume		
[pH]	[nm]		[ml]			[mAU*min^2]		
			12 CV	24 CV	36 CV	12 CV	24 CV	36 CV
7.6 - 7.1	2.68 ± 0.00	2	2.54 ± 0.03	4.01 ± 0.04	5.29 ± 0.42	9.1 ± 0.6	9.4 ± 1.5	7.1 ± 0.7
7.6 - 7.1	3.47 ± 0.01	3	2.91 ± 0.03	4.66 ± 0.05	6.28 ± 0.09	15.5 ± 1.0	15.6 ± 0.7	13.5 ± 0.8
7.6 - 7.1	3.90 ± 0.08	4	2.65 ± 0.70	4.45 ± 0.32	6.22 ± 0.37	0.8 ± 1.21	0.6 ± 0.7	1.4 ± 0.5
7.1 - 6.5	3.78 ± 0.01	4	2.78 ± 0.05	4.45 ± 0.11	5.93 ± 0.14	4.8 ± 4.7	5.2 ± 1.2	5.7 ± 2.8
7.1 - 6.5	3.51 ± 0.01	5	2.81 ± 0.06	4.58 ± 0.06	6.05 ± 0.08	20.3 ± 4.4	17.1 ± 2.0	16.8 ± 2.4
7.1 - 6.5	2.64 ± 0.03	2	2.57 ± 0.36	4.13 ± 3.23	5.34 ± 3.70	2.3 ± 1.1	1.6 ± 1.9	2.0 ± 1.6
7.1 - 6.5	2.94 ± 0.01	6+7	3.26 ± 0.37	5.28 ± 0.26	7.12 ± 0.12	2.2 ± 1.4	2.1 ± 0.9	2.2 ± 0.4
6.5 - 6.0	2.97 ± 0.01	7	3.37 ± 0.12	5.23 ± 0.05	6.91 ± 0.26	1.2 ± 0.4	1.3 ± 0.1	1.0 ± 0.2
6.5 - 6.0	3.52 ± 0.02	5+8	2.82 ± 0.03	4.72 ± 0.05	6.33 ± 0.10	2.7 ± 2.5	1.8 ± 0.3	0.8 ± 0.1
6.5 - 6.0	3.00 ± 0.00	10	2.70 ± 0.72	4.31 ± 0.64	- -	0.8 ± 1.0	0.5 ± 0.3	- -
6.5 - 6.0	3.81 ± 0.03	4	2.81 ± 0.06	4.69 ± 0.09	6.14 ± 0.56	1.3 ± 3.2	1.5 ± 0.4	1.9 ± 0.3

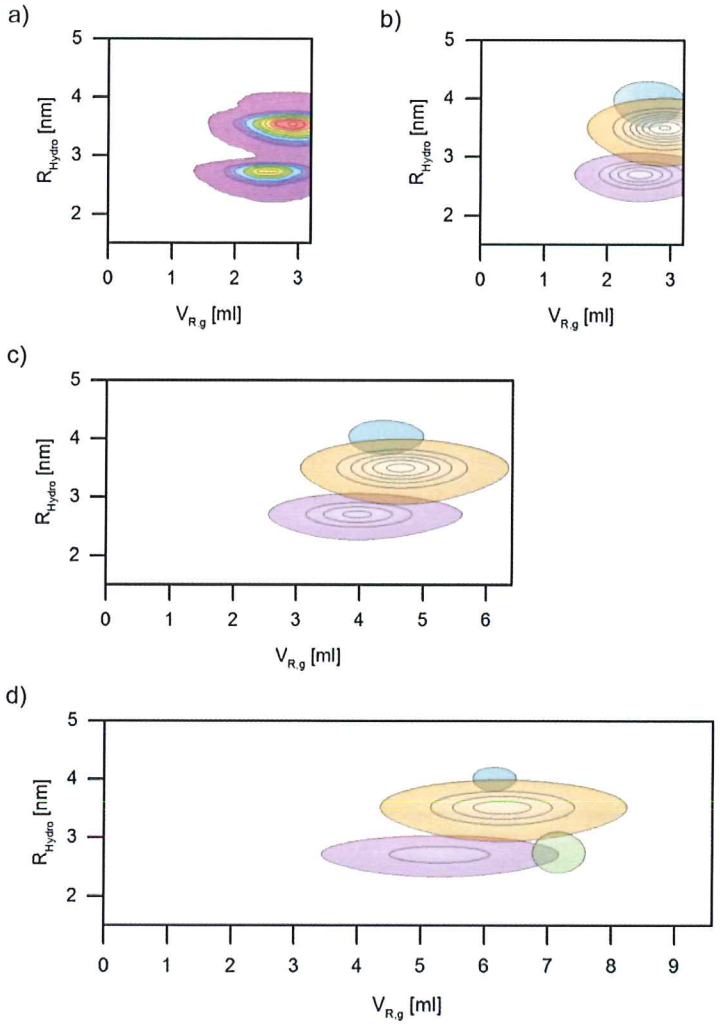


Figure 3.5 Deconvolution and tracking of peaks across multiple linear gradient experiments for the example of the fraction collected between pH 7.6-7.1. a) raw gradient composite chromatogram. b-d) deconvoluted peaks of the different gradient length experiments of the same fraction. Peaks with matching properties were assigned the same colour an matched to a peak in the reference data. Contour lines are drawn at steps of 1 mAU of 230nm UV absorption.

3.3.5. Isotherm parameter regression

The parameters regressed for the matched retention volumes of the peaks in Table 3 are shown in Table 4. The standard errors of the regressed parameters are on average 8% of the parameters nominal value for both the K_B and K_S parameters, with maximum errors of 16% for K_B and 13% for K_S. No standard error could be calculated for the matched component 10, as only two data points were available for fitting. These low errors indicate that the additional separation dimension allows regression of reliable isotherm parameters for individual components, despite the initial low resolution of the small column. Even though the components in the reference chromatogram were defined only by similarity in size and charge, introduction of hydrophobic interaction as a third orthogonal separation dimension did not reveal any additional peaks, allowing easy matching of detected peaks to the reference data. Under these conditions, such a reference chromatogram appears to be a useful tool that could allow easy comparison of different resins used in the second dimension. Should the second separation dimension offer significantly higher resolution, mass spectrometric techniques might be necessary to allow peak matching with high certainty [15,24]. As the K_D coefficients are close to unity for all compounds of interest and the retention volumes in the second dimension clearly being mostly dependent on the strength of the hydrophobic interactions, the assumptions made to allow parameter regression from the analytical solution of the Parente and Wetlaufer formalism appear to have been reasonable. An alternative route to gain these parameters when these simplifications should not apply would be derive them through inverse-modelling [15,25,26]. In such a scenario the parameters determined from the analytical solution could serve as a starting point for further fitting, to reduce the computational effort.

Table 3.4 HIC linear isotherm parameters regressed for the components tracked across multiple linear gradients.

ID*	K_{eq} [-]	K_S [M^{-1}]
2	6.13 ± 0.36	3.31 ± 0.14
3	9.36 ± 1.41	3.43 ± 0.44
4	7.23 ± 1.13	3.90 ± 0.45
5	8.57 ± 0.29	3.42 ± 0.09
6+7	15.45 ± 1.28	3.18 ± 0.32
7	17.32 ± 1.72	2.53 ± 0.33
5+8	8.31 ± 0.07	3.89 ± 0.03
10	7.93 -	3.06 -

3.3.6. Implications for process development

The ultimate goal of such profiling techniques as demonstrated in this study is to build a large enough database of isotherm parameters to allow full in-silico development of a multistep separation process [6,27]. Successful optimization of chromatographic separations based on data determined on such small columns has been demonstrated for model mixtures [28]. The proposed multidimensional technique appears to be a promising step to making this type of HTS compatible format also useful for parameter determination directly from complex mixtures.

As the proposed technique simultaneously characterizes the components of the mixture towards charge, hydrophobicity and size, the data is also useful for identifying potentially feasible purification routes through the use of expert systems or heuristic design rules [5,29,30]. As mentioned before, the large differences in elution-pH indicate that components 1 and 8-14 can all easily be removed by most charge based separations, leaving components 2 and 5-7 as critical contaminants. Component 2, 6 and 7 show significant

differences in hydrophobicity to the product, 2 being less hydrophobic and 6 and 7 significantly more hydrophobic, suggesting that under the right conditions a separation between these components and the products could be achieved. Hydrophobic interaction does not appear to be a suitable mechanism for the separation of components 4 or 5 from the product, as in both cases the isotherm parameters of these components lie within the error margin of the product's parameters. Size is rarely a feasible mode of separation, but component 5 is so similar that removal of this component by SEC does not appear to be possible at all.

The differences in elution-pH between components 4 and 5 from the product are 0.3 and 0.4 respectively. For development of ion-exchange operations the elution-pH serves as a more reliable replacement parameter for the pI [31]. Such small differences are regarded as challenging separation problems [32], but systematic model-based optimization of the process conditions has been shown to be capable of identifying conditions leading to sufficient process performance for a similar separation challenge [33].

3.4 Concluding remarks

The proposed multi-dimensional profiling approach is demonstrated to allow direct characterization of a complex mixture of proteins towards their behaviour in ion-exchange, hydrophobic and size-exclusion chromatography. In a first instance, knowledge of these properties allows determining which contaminants will be most challenging to remove, and gives an indication of what separation mechanism is most promising for the removal of each major contaminant. Through increasing the dimensionality of the separation, it was demonstrated that isotherm interaction parameters can be determined within reasonable certainty directly from the complex mixtures while using a HTS compatible column format. These parameters may be useful for the in-silico optimization of the more challenging separation tasks. The creation of a two-

dimensional high-resolution reference chromatogram allowed assigning the peaks of different experiments to pseudo-components, avoiding the need for time-intensive mass-spectrometry based techniques. This is of particular importance when aiming to operate the second dimension in a true HTS environment working with large numbers of resins and operation conditions.

Acknowledgements

This project was financed by BE-Basic (Bio-based Ecologically Balanced Sustainable Industrial Chemistry), a public-private NWO-ACTS programme (ACTS = Advanced Chemical Technologies for Sustainability). The authors thank DSM for providing the complex sample feedstock used in this study.

References

[1] R. Bhambure, K. Kumar, A.S. Rathore, High-throughput process development for biopharmaceutical drug substances, Trends Biotechnol. 29 (2011) 127-135.

[2] L.X. Yu, Pharmaceutical quality by design: product and process development, understanding, and control, Pharm. Res. 25 (2008) 781-791.

[3] A.O. Kirdar, K.D. Green, A.S. Rathore, Application of multivariate data analysis for identification and successful resolution of a root cause for a bioprocessing application, Biotechnol. Prog. 24 (2008) 720-726.

[4] A.T. Hanke, M. Ottens, Purifying biopharmaceuticals: knowledge-based chromatographic process development, Trends Biotechnol. 32 (2014) 210-220.

[5] J.A. Asenjo, B.A. Andrews, Is there a rational method to purify proteins? From expert systems to proteomics, J. Mol. Recognit. 17 (2004) 236-247.

[6] B.K. Nfor, T. Ahamed, G.W.K. van Dedem, P.D.E.M. Verhaert, L.A.M. van der Wielen, M.H.M. Eppink, E.J.A.X. van de Sandt, M. Ottens, Model-based rational methodology for protein purification process synthesis, Chem. Eng. Sci. 89 (2013) 185-195.

[7] J.A. Asenjo, B.A. Andrews, Protein purification using chromatography: selection of type, modelling and optimization of operating conditions, J. Mol. Recognit. 22 (2009) 65-76.

[8] A. Osberghaus, S. Hepbildikler, S. Nath, M. Haindl, E. von Lieres, J. Hubbuch, Optimizing a chromatographic three component separation: a comparison of mechanistic and empiric modeling approaches, J. Chromatogr. A 1237 (2012) 86-95.

[9] T. Ahamed, M. Ottens, B.K. Nfor, G.W.K. van Dedem, L.A.M. van der Wielen, A generalized approach to thermodynamic properties of biomolecules for use in bioseparation process design, Fluid Phase Equilib. 241 (2006) 268-282.

[10] W.H. Scouten, Affinity chromatography for protein isolation, Curr. Opin. Biotechnol. 2 (1991) 37-43.

[11] D.S. Hage, Affinity chromatography: a review of clinical applications, Clin. Chem. 45 (1999) 593-615.

[12] O. Aguilar, C.E. Glatz, M. Rito-Palomares, Characterization of green-tissue protein extract from alfalfa (Medicago sativa) exploiting a 3-D technique, J. Sep. Sci. 32 (2009) 3223-3231.

[13] R.K. Swanson, R. Xu, D. Nettleton, C.E. Glatz, Proteomics-based, multivariate random forest method for prediction of protein separation behavior during cation-exchange chromatography, J. Chromatogr. A 1249 (2012) 103-114.

[14] F. Kroner, D. Elsasser, J. Hubbuch, A high-throughput 2D-analytical technique to obtain single protein parameters from complex cell lysates for in silico process development of ion exchange chromatography, J. Chromatogr. A 1318 (2013) 84-91.

[15] B.K. Nfor, T. Ahamed, M.W. Pinkse, L.A. van der Wielen, P.D. Verhaert, G.W. van Dedem, M.H. Eppink, E.J. van de Sandt, M. Ottens, Multi-dimensional fractionation and characterization of crude protein mixtures: toward establishment of a database of protein purification process development parameters, Biotechnol. Bioeng. 109 (2012) 3070-3083.

[16] F. Kroner, J. Hubbuch, Systematic generation of buffer systems for pH gradient ion exchange chromatography and their application, J. Chromatogr. A 1285 (2013) 78-87.

[17] Y. Nozaki, N.M. Schechter, J.A. Reynolds, C. Tanford, Use of gel chromatography for the determination of the Stokes radii of proteins in the presence and absence of detergents. A reexamination, Biochemistry (Mosc.) 15 (1976) 3884-3890.

[18] B.C. To, A.M. Lenhoff, Hydrophobic interaction chromatography of proteins. I. The effects of protein and adsorbent properties on retention and recovery, J. Chromatogr. A 1141 (2007) 191-205.

[19] A.T. Hanke, P.D. Verhaert, L.A. van der Wielen, M.H. Eppink, E.J. van de Sandt, M. Ottens, Fourier transform assisted deconvolution of skewed peaks in complex multi-dimensional chromatograms, J. Chromatogr. A 1394 (2015) 54-61.

[20] E.S. Parente, D.B. Wetlaufer, Relationship between Isocratic and Gradient Retention Times in the High-Performance Ion-Exchange Chromatography of Proteins - Theory and Experiment, J. Chromatogr. 355 (1986) 29-40.

[21] J.M. Mollerup, The thermodynamic principles of ligand binding in chromatography and biology, J. Biotechnol. 132 (2007) 187-195.

[22] J. Chen, T. Yang, S.M. Cramer, Prediction of protein retention times in gradient hydrophobic interaction chromatographic systems, J. Chromatogr. A 1177 (2008) 207-214.

[23] T. Ahamed, S. Chilamkurthi, B.K. Nfor, P.D. Verhaert, G.W. van Dedem, L.A. van der Wielen, M.H. Eppink, E.J. van de Sandt, M. Ottens, Selection of pH-related parameters in ion-exchange chromatography using pH-gradient operations, J. Chromatogr. A 1194 (2008) 22-29.

[24] F. Kroner, A.T. Hanke, B.K. Nfor, M.W. Pinkse, P.D. Verhaert, M. Ottens, J. Hubbuch, Analytical characterization of complex, biotechnological feedstocks by pH gradient ion exchange chromatography for purification process development, J. Chromatogr. A 1311 (2013) 55-64.

[25] A. Osberghaus, S. Hepbildikler, S. Nath, M. Haindl, E. von Lieres, J. Hubbuch, Determination of parameters for the steric mass action model-A comparison between two approaches, J. Chromatogr. A 1233 (2012) 54-65.

[26] T. Hahn, P. Baumann, T. Huuk, V. Heuveline, J. Hubbuch, UV absorption-based inverse modeling of protein chromatography, Eng. Life Sci. (2015) n/a-n/a.

[27] T.C. Huuk, T. Hahn, A. Osberghaus, J. Hubbuch, Model-based integrated optimization and evaluation of a multi-step ion exchange chromatography, Sep. Purif. Technol. 136 (2014) 207-222.

[28] A. Osberghaus, K. Drechsel, S. Hansen, S.K. Hepbildikler, S. Nath, M. Haindl, E. von Lieres, J. Hubbuch, Model-integrated process development demonstrated on the optimization of a robotic cation exchange step, Chem. Eng. Sci. 76 (2012) 129-139.

[29] M.E. Lienqueo, J.A. Asenjo, Use of expert systems for the synthesis of downstream protein processes, Comput. Chem. Eng. 24 (2000) 2339-2350.

[30] E.W. Leser, J.A. Asenjo, Rational design of purification processes for recombinant proteins, J. Chromatogr. 584 (1992) 43-57.

[31] T. Ahamed, B.K. Nfor, P.D. Verhaert, G.W. van Dedem, L.A. van der Wielen, M.H. Eppink, E.J. van de Sandt, M. Ottens, pH-gradient ion-exchange chromatography: an analytical tool for design and optimization of protein separations, J. Chromatogr. A 1164 (2007) 181-188.

[32] L. Pedersen, J. Mollerup, E. Hansen, A. Jungbauer, Whey proteins as a model system for chromatographic separation of proteins, J. Chromatogr. B Analyt. Technol. Biomed. Life. Sci. 790 (2003) 161-173.

[33] B.K. Nfor, J. Ripic, A. van der Padt, M. Jacobs, M. Ottens, Model-based high-throughput process development for chromatographic whey proteins separation, Biotechnol. J. 7 (2012) 1221-1232.

4

Multi-dimensional fractionation and characterization of crude protein mixtures: high-throughput parameter determination

Abstract

The vast experimental space that needs to be explored during the development of a biopharmaceutical purification process has led to the widespread adaption of high-throughput technologies. For chromatographic separations, one of the most popular formats besides batch adsorption studies in micro titre plates, are miniaturized packed chromatography columns that are operated on robotic liquid handling systems. The practical limitations resulting from the use of liquid handling systems instead of conventional liquid chromatography setups influence both the design of experiments and how their data needs to be processed. To minimize the shortcomings of the fractionation system, we introduce a new technique for the meniscus sensitive estimation of single well liquid volumes in micro titre plates with less than 5% deviation. With this improvement in place, we explore how such a high-throughput system can be utilized in the context of a multi-dimensional fractionation scheme for the regression of isotherm parameters directly from crude mixtures, using an IgG-1 containing CHO cell-culture supernatant as a case study. Applying a two-dimensional strategy already allowed to regress the equilibrium constants of nine pseudo-components with an average standard error of 21%, with the potential introduction of further dimensions expected to further improve these results.

Keywords: Process development parameters; Crude protein mixtures; High-throughput chromatography; Mechanistic models; Parameter database;

4.1 Introduction

The ambition of the biopharmaceutical field to move towards knowledge-based process development principles has created the need for efficient experimental means to determine the parameters required for mechanistic modelling of the separation steps [1]. For complex samples, multi-dimensional separation techniques have been demonstrated to allow the simultaneous regression for a whole set of practically defined pseudo-components [2].

In Chapter 3, we have demonstrated that comprehensive multi-dimensional separations can be performed around miniaturized chromatography columns packed with industrial grade resins, allowing to determine isotherm interaction parameters from complex mixtures, in spite of the low-resolution of the screening step itself. In that study the columns were operated on a conventional liquid chromatography system, with synchronized double dual-piston pumps for gradient generation and in-line detectors. Such a system is too complex to economically increase its throughput by parallelisation.

More economic parallelization can be achieved by operating the columns within a robotic liquid handling system [3]. These systems are neither equipped with dual-piston pumps, nor with inline detectors. Instead single piston pumps apply a liquid flow, fractions are collected at the column outlet by a 96 well plate placed on a motorized shuttle, and analysis takes place offline. These mechanical simplifications require some adaptions to the experimental approach, to allow generation of data that is straightforward comparable to experiments performed on traditional systems.

One of the main technical challenges in the operation of RoboColumns on a conventional liquid handling system, is that the fractionation intervals, the moments at which the collection plate

shuttle moves from one column of wells to the next, are defined in relation to the syringe motor position that applies flow to the columns. As there is no reliable mechanism to synchronize the falling of drops from the column outlet, and the size of the drops themselves may vary with changes in buffer composition and protein content, the volume that actually ends up in each well may vary significantly, especially when the target fraction volume is small. It is therefore necessary to measure the volume of each well in order to reduce the experimental noise that would be caused by assuming a constant fraction volume [3]. So far this was either performed by detection of the liquid level by probing with the pipetting needles [3], or by correlation of the transmission path with the near-infrared-red (NIR) adsorption of the buffer [4,5]. Both approaches have been demonstrated to be suitable for the normalization of absorption measurements towards the transmission path, but both lack the ability to detect and quantify the shape of the meniscus in each well, limiting their ability to accurately measure the total volume of liquid in a well. To overcome this limitation an extension of the NIR absorption based volume detection technique is introduced in Section 4.3.3.

The lack of two pumps per column prevents inline generation of salt or pH gradients. Offline generation of pre-mixed small steps to simulate a gradient has been demonstrated to be a viable option [6], but unless as time-consuming and difficult to realize conductivity measurement of each well is built into the offline fraction analysis, the exact resulting gradient shape and elution volume are unknown. This may cause additional uncertainties during data interpretation. To avoid these issues, isocratic elution conditions were investigated.

4.2 Theory and Models

The retention coefficient (k'_i) of a species 'i' in chromatography is defined as [7]:

$$k_i' = \frac{V'_{R,i}}{V_M} \quad (1)$$

where $V'_{R,I}$ is the adjusted retention volume, i.e. the observed retention volume corrected for the systems hold-up volume (V_M). Both values can easily be determined by simple pulse injection experiments; the retention volume by injection of the species itself and the hold-up volume by an inert tracer molecule that is fully excluded from the resin pore volume. The first moment of each peak is used as the value for calculation of the retention factor. For single component systems the retention factor can directly be related to the columns properties and the species thermodynamic and mass-transfer properties [8]:

$$k_i' = \frac{(1-\varepsilon_b)\varepsilon_p K_{D,i}}{\varepsilon_b}(1+A_i) \quad (2)$$

where ε_b is the column's bed porosity, ε_p the particle porosity, $K_{D,i}$ the fraction of the total pore volume accessible to the species and A_i the slope of the species adsorption isotherm. In this form the relation is particularly useful as a basis for parameter regression, as it can be used for all types of liquid chromatography. For size-exclusion chromatography, only the pore accessibility coefficient needs to be experimentally determined. This can either be achieved by measuring the retention of the species under non-binding conditions [9], or can be interpolated from a resin-specific calibration line, provided the hydrodynamic radius ($R_{hydro,i}$) of the species is known [10]. A simple pore accessibility model that can be calibrated with a set of inert tracer of known hydrodynamic radius is given by:

$$K_{D,i} = \frac{1}{1+\left(\frac{R_{Hydro,i}}{r_m}\right)^p} \quad (3)$$

where r_m and p are resin specific parameters. The thermodynamic interactions of the solute with the resin can be described by a suitable isotherm model. One of the most commonly

used isotherm models for ion-exhange chromatography is the Steric Mass-Action model [11]. Here the initial slope of the isotherm is given by [9]:

$$A_i = K_{eq,i}\left(\frac{\Lambda}{z_s c_s}\right)^{\nu_i} \quad (4)$$

with the two component specific parameters being the equilibrium constant ($K_{eq,i}$) and charge (ν_i), the ionic capacity of the resin (Λ) and the concentration (c_s) and charge (z_s) of the counter-ions in the elution buffer. Similar isotherms following the same formalism have also been proposed for hydrophobic interaction [12] and mixed mode chromatography [13].

4.3 Materials and methods

4.3.1. Gradient chromatofocusing prefractionation

The complex sample used for this study is a clarified CHO cell culture supernatant containing a monoclonal IgG-1. Prior to use, the samples were rebuffered using disposable PD-10 columns, following the manufacturers protocol (GE-Healthcare, Sweden). As a first separation dimension, the samples were fractionated by linear pH-gradient chromatography on a Mono Q 4.6/100 strong anion exchange column (GE Healthcare, Sweden) following the same protocol as described in Section 3.2.2. To facilitate the interpretation of the high-throughput chromatograms, a two-dimensional reference such described in Section 3.3.1 was created following the same protocols.

4.3.2. High-throughput isocratic chromatography

The high-throughput liquid chromatography experiments were performed on a Freedom Evo 200 liquid handling workstation

equipped with an 8-channel liquid handling arm fitted with 1 ml syringes and Te-Chrom station (Tecan Switzerland). The columns were 200 µl RoboColumns (Atoll Bio, Germany), packed with POROS 50 HS strong cation exchange resin (Thermo Fisher Scientific, The Netherlands). The porosity and pore accessibility of these columns was analysed on an Äkta Explorer 10 (GE Healthcare, Sweden) equipped with a 1100 series refractive index detector (Agilent, CA, USA) following the protocol described in Section 3.2.4. Prior to each chromatographic experiment, a sufficient volume of buffer for both column equilibration and elution was mixed from stock solutions by the liquid handling system. The two stock solutions prepared for this step were 25 mM Acetic acid in Milli Q (low salt) titrated to pH 4.5 (low salt) and the same buffer containing 1 M sodium chloride added prior to titration (high salt). The mixing ratios were chosen to result in eight different final sodium chloride concentrations ranging from 0 to 0.5 M. Samples collected from the prefractionation gradient were transferred into low salt buffer through at least 3 buffer exchange cycles in Amicon spin filters with a nominal molecular weight cut-off of 3 kDa (Millipore, USA) following the manufacturer recommend protocol. After rebuffering, each sample was split into eight aliquots and appropriate volumes of low and high salt buffer were added to result in eight samples of equal protein content and pH, but with salt concentrations corresponding to the eight prepared elution buffers.

Prior to injection each column was equilibrated with 5 column volumes (CV) of elution buffer. The volume of salt concentration adjusted sample injected to each column was 20 µl. The samples were eluted with a total of 15 CV of elution buffer at a flowrate of 0.15 ml/min per column. During the isocratic elution a total of 22 samples were collected from each column. The first twelve fractions had a target volume of 75 µl and were collected in a half area UV-star plate (Greiner-Bio One, The Netherlands). Afterwards six additional fractions with a target volume of 150 µl

were collected in a full area UV-Star plate (Greiner Bio-One, The Netherlands), followed by four more with a target volume of 300 µl. This staggered fractionation strategy was chosen as a compromise, to provide high resolution at the beginning of the experiment where sharp and narrow peaks were expected while simultaneously keeping the total number low. The columns were subsequently cleaned with 5 CV of washing buffer of which the first 600 µl were collected in two fractions with a target volume of 300 µl each. Once this step had been completed both fractionation plates were passed on to the plate reader for analysis. Prior to the next experiment each column was sanitized with 5 CV of sanitation buffer.

4.3.3. Fraction volume estimation

All optical measurements in 96 well plates were performed in an infinite M200 plate reader (Tecan, Switzerland). The absorption values at 600 nm, 900 nm and 997 nm wavelengths were measured at the geometric well centre. The adsorption at 600 nm is measured at an additional 20 points, evenly distributed along a circle around the geometric well centre using the built-in multiple reads per well function of the plate reader. The minimum distance of these measurement points from the well walls was set to 330 µm. These measurements are used in combination with knowledge of the well geometry as provided by the plate manufacture to estimate the volume of each well. An overview of the geometric parameters of the well that are used for these calculations is given in Figure 4.1 b).

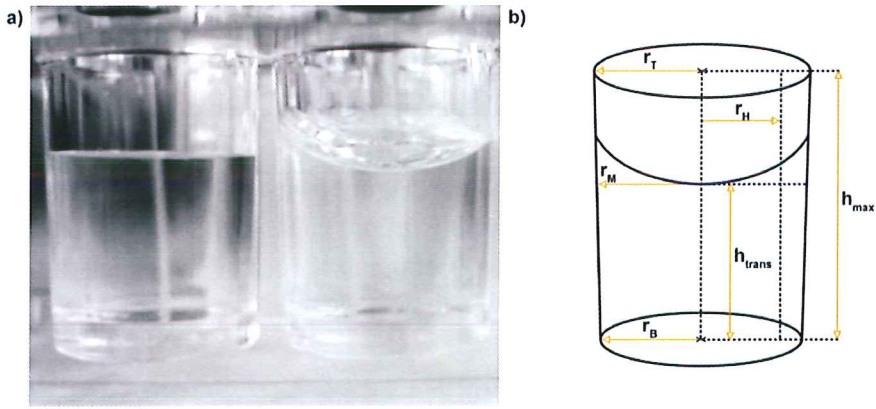


Figure 4.1 Impact of liquid distribution in a well on the transmission path in the geometric well centre. a) two wells with equal liquid volume of buffer with no protein present (left) and 0.1 g/l of lysozyme (right). b) schematic illustration of the geometry of a well showing the height and radius parameters used for the estimation of the liquid volume.

In accordance with the Lambert-Beer law a linear correlation between the transmission path (h_{trans}) and the corrected NIR absorption (ΔNIR) of the buffer is assumed, with a specific transmission coefficient ($\tau_{NIR,buffer}$) related to the density of the buffer.

$$h_{trans} = \tau_{NIR,buffer} \cdot \Delta NIR \tag{7}$$

The ΔNIR is the difference between the absorption at 997 nm and 900 nm. The walls of the used 96 well plates are slightly slanted. The radius of the wells cross section at the height of the bottom of the meniscus (r_M) is calculated from h_{trans} and the wells upper (r_T) and lower radius (r_B) by:

$$r_M = (r_T - r_B)\frac{h_{trans}}{h_{max}} + r_B \tag{8}$$

with h_{max} being the total height of the well. For a perfectly flat meniscus, such as in the left well shown in Figure 4.1 a) the volume can now be estimated by the formula for the volume of a

circular truncated cone. For wells with a more pronounced meniscus, such as the right one in Figure 4.2 a), an extra term needs to be added. The 600 nm measurements are corrected for the value at the centre of the well and summed up (Σ_{halo}). For a flat meniscus this value is close to zero. For more pronounced menisci the value exponentially increases, so a correction factor (C_{Vm}) based on its natural logarithm is introduced, leading to the following equation for the estimation of liquid volume (V_{est}) in a well:

$$V_{est} = \frac{1}{3}\pi({r_B}^2 + r_B r_M + {r_M}^2)h_{trans} + C_{Vm} \cdot \ln(\Sigma_{halo}) \quad (9)$$

The method is calibrated with both a half and full area plate containing known volumes ranging from 0 µl to the maximum well capacity, of both protein free buffer and buffer with addition of a small concentration (~0.1 g/l) of model proteins, such as bovine serum albumin or lysozyme. The buffer NIR extinction coefficient $\epsilon_{NIR,buffer}$ is assumed to be identical for half and full area plates, whereas the meniscus coefficient ($C_{M,\Sigma halo}$) is determined separately for each plate geometry. The coefficients are determined by a least-square regression of Eq. (7-9) in Matlab 2013b (Mathworks, USA). Afterwards the coefficients are validated against a second set of plates with a different distribution of sample volumes. The accuracy of each measurement was calculated by:

$$Acc(V_{est}) = \left(1 - \frac{V_{est} - V_{nominal}}{V_{nominal}}\right) \cdot 100[\%] \quad (10)$$

4.3.4. Reconstruction of high-throughput chromatograms

As high-throughput chromatography systems, such as the Te-Chrom used in this study, do not possess in-line detection systems chromatograms need to be reconstructed from the measurements performed on the collected fractions. The transmission path and total well volume of each collected fraction

were calculated as according to the approach outlined in Section 4.3.3. To reduce the noise in the absorption signals each value is corrected for the absorption at 330 nm and normalized against the estimated transmission path. To determine the positon of each normalized absorption in the reconstructed chromatogram, the volume of all preceding fractions is summed up and added to half the volume of the corresponding fraction.

4.3.5. Deconvolution and peak moment calculations

To estimate the number of peaks in each chromatogram, each data set was scanned for data points fulfilling the following criteria: they had to have a normalized 230 nm absorption of at least 0.1 mAU/cm and this value needed to be larger than both the neighbouring fractions. For practical purposes related to the small number of available data points per chromatogram only the largest four points fulfilling this criteria were considered for further analysis. The heights and positions of the local maxima identified by this algorithm were used as initial guesses for a least-squares based fitting of peak model to the reconstructed chromatogram.

The function chosen for fitting was based on a one-dimensional adaption of the model for multiple superimposed exponentially modified Gaussian peaks described in Section 2.2. Instead of minimizing the squares between the measured data point and the curve described by the peak model, the average of the model curve was calculated over each fraction interval, and the squares between this value and the measurement were minimized. The fitting was carried out in Matlab 2013b (Mathworks, MA, USA), using the built in lsqcurvefit function. The areas and first moments of the fitted peaks were calculated together with their standard errors of regression following the same principles as described in Section 2.3.6.

4.4 Results and discussion

4.4.1. Prefractionation reference data

The resulting two-dimensional map of the high-resolution reference experiments is shown in Figure 4.2 and the corresponding peak properties are presented in Table 4.1. Peak 2 corresponds to the IgG-1, the protein of interest. The most abundant contaminants, Peaks 1,3,4,6 and 7, appear to have very similar charge properties, as does at least one high molecular weight (HMW) contaminant, marked as Peak 5. The large difference in elution-pH of peaks 8-17 in relation to protein of interest, indicates that these contaminants could easily be removed by an appropriate anion-exchange step [14]. The similarity in size of the contaminant constituting peak 7 makes size-based separation unfeasible. Therefore an additional orthogonal purification step on will be necessary to remove these critical contaminants.

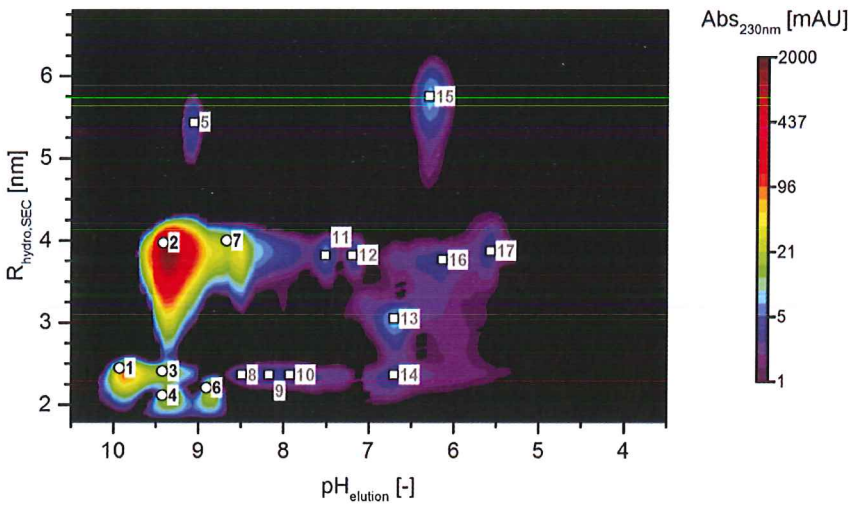


Figure 4.2 Two-dimensional reference chromatogram of the CHO-IgG supernatant used in this study. The hydrodynamic radii ($R_{Hydro,SEC}$) are estimated from the SEC retention times. Fitted peaks are marked by white dots, whereas peaks that could be detected, but were below the fitting threshold ware marked by white squares.

Table 4.1 Peak characteristics determined by Fourier transform assisted peak deconvolution of the two-dimensional AEX-gCF SEC reference chromatogram. The chromatogram is shown in Figure 4.2 and the peak IDs correspond to the labels assigned there. The pH of peak elution ($pH_{Elution}$) and the hydrodynamic radii ($R_{hydro,SEC}$) are estimated from the peaks first moments in the corresponding dimensions for fitted peaks, and the peak maxima positons for peaks that were not fitted.

ID	$pH_{Elution}$ [-]	$R_{Hydro,SEC}$ [nm]	Peak volume [mAU*min^2]
1	9.8 ± 0.0	2.33 ± 0.01	20.1 ± 2.3
2	9.3 ± 0.0	3.76 ± 0.00	297.7 ± 1.8
3	9.3 ± 0.0	2.31 ± 0.74	4.8 ± 6.7
4	9.3 ± 0.0	2.03 ± 0.07	2.9 ± 4.4
5	9.1 *	5.43 *	-
6	8.8 ± 0.0	2.12 ± 0.06	4.1 ± 1.9
7	8.6 ± 0.1	3.78 ± 0.04	11.4 ± 2.7
8	8.5 *	2.36 *	-
9	8.2 *	2.36 *	-
10	7.9 *	2.36 *	-
11	7.5 *	3.82 *.	-
12	7.2 *	3.82 *	-
13	6.7 *	3.05 *	-
14	6.7 *	2.36 *	-
15	6.3 *	5.75 *	-
16	6.1 *	3.76 *	-
17	5.6 *	3.87 *	-

4.4.2. Well volume measurement

The regressed transmission coefficients for the used buffer system ($\tau_{NIR,buffer}$) was determined to be 0.640 ± 0.001 mm/AU. As the absorption in this wavelength is dominated by the water content of the buffer, it is practically the same for all aqueous buffer systems, provided that their density is still close to pure water. The meniscus correction coefficients (C_{Vm}) were determined to be 1.76 ± 0.30 µl/ln(Σ_{halo}) for the half area plates and 12.72 ± 0.35 µl/ln(Σ_{halo}) for the full area plates. The choice of model protein used to induce the formation of the meniscus was not found to have a significant effect on these parameters. Figure 4.2 a) shows the relation between the volume that is not accounted for by the truncated cone volume and therefor attributed to the meniscus, and the logarithm of the sum of the measurement around the well. The highly linear relationship for both well geometries supports the choice for a simple linear model. While the meniscus volume in the half area plates is in the range of 0-5 µl, it can account for up to 35 µl in the full area plates, at least 10% of the total well volume. With the meniscus correction in place the accuracy of the volume detection is improved to better than 3% for the full area plates and 5% for the half area plates. Figure 4.2 b) shows the techniques accuracy over a range of different volumes in both plate geometries. While the accuracy appears to be largely volume independent in the half area plates, it there is clearly a negative effect caused by low volumes in full area plates. This is caused by the tendency of small volumes to not evenly distribute across the well in full area plates. As a result, it is recommended to use half area plates for collection of fraction volumes in the range of 50 to 125 µl and full area plates for volumes exceeding 125 µl.

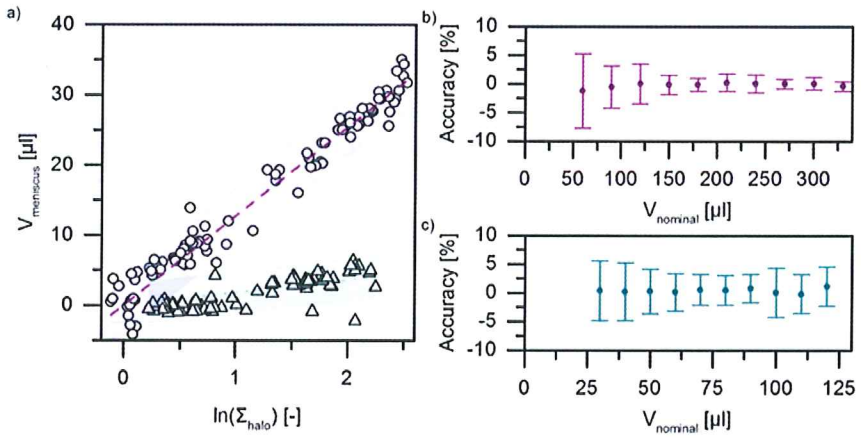


Figure 4.2 Calibration of the meniscus-sensitive volume detection method. a) Linear correlation between the volume hidden by the meniscus and the natural logarithm of the sum of the absorption values on the measurement halo together with the 95% prediction bands for both full area plates (dashed purple line and circles) and half area plates (dotted teal line and triangles). b) and c) Average volume estimation accuracy of method in full area plates (b) and half area plates (c). The error bars correspond to twice the standard deviation across at least 8 measurements.

4.4.3. Column properties

The POROS 50 HS RoboColumns were characterized towards by pulse injections of Dextran standards. The measured retention volumes of these standards are shown in Figure 4.3 a). From the differences in retention of the smallest and largest Dextran the bed porosity (ε_b) was calculated to be 0.4. As the retention of the three largest tracers were all the same, the assumption that this corresponds to a full exclusion from the particle pore volume can be considered as correct. Knowledge of the columns void volume allowed calculation of the particle porosity (ε_p), resulting in a value of 0.51. Based on this the pore accessibility was calculated for the remaining tracers, the results shown in Figure 4.3 b). The coefficients of the curve were regressed to be 11.85 ± 0.96 r_m and

2.76 ± 0.43 for p. The reference data revealed that most components in the sample have a hydrodynamic radius of less than 6 nm. From the pore accessibility calibration curve it can be seen that the K_D coefficients for the protein of interest and most critical contaminants are expected to be in the range of 1.0 to 0.9, but may range down to 0.7 in the case of HMW contaminants.

Knowledge of the ligand density allows regression of thermodynamically more meaningful parameters [8] according to the equations outlined in Section 4.2. For POROS 50 HS this data is available in literature [15]. The charge molar equivalent charge density per dry weight of resin is reported to be 0.41 mol_{eq}/g_{dry}. Given a swollen and packed resin density of 0.32 g_{dry}/ml the total charge capacity of the RoboColumns can be calculated to 26.2 $mmol_{eq}$.

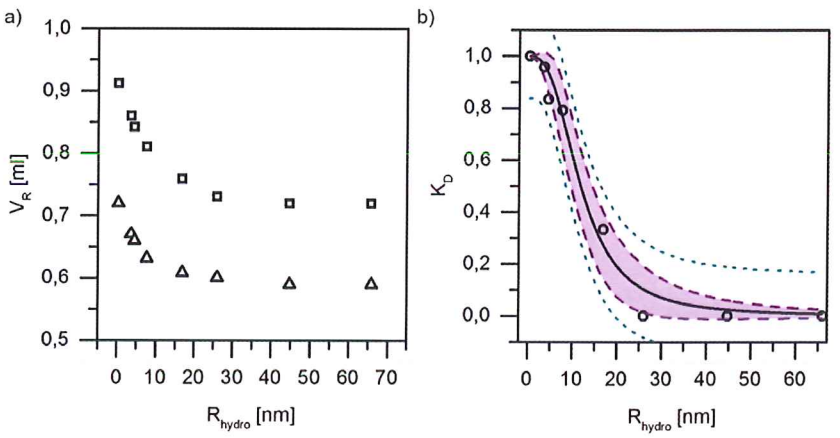


Figure 4.3 a) Uncorrected retention volumes of Dextran standards with column(□) and without column (Δ). b) Calculated K_D values of the Dextran standards (**O**) together the fitted K_D curve (solid black), its 95% confidence interval (dashed purple) and prediction interval (dotted teal).

4.4.4. High-throughput pulse injection experiments

Following the procedure described in Section 4.3.4 results in data sets that closely resemble traditional online recorded chromatograms. Figure 4.5 shows examples of such reconstructed chromatograms for the isocratic elution at different salt concentrations of the fraction containing the main protein of interest. These chromatograms can be interpreted in two ways: the ratio between absorbance during the elution step and cleaning step can be compared, which allows to determine at which salt concentrations a selective separation takes place, or the position of individual peaks can be attributed to pseudo-components for which sets of isotherm parameters are regressed.

Taking the fraction collected between pH 9.8 and 9.0 as example, based on the prefractionation reference chromatogram we expect it to contain the protein of interest, a HMW contaminant and a variety of LMW contaminants. As can be seen in Figure 4.5 a), no detectable peaks elute within the elution phase of the experiment with an elution buffer salt concentration of 100 mM NaCl, whereas a large absorbance signal is present for the first fraction of the washing step. The rapid decline of that signal in the second wash fraction indicates that the ionic strength of the washing step is strong enough to generate non-binding conditions for all the bound components. Increasing the elution buffer salt concentration to 200 mM NaCl, shows an increase in the absorption signal during the elution phase, while the signal of the wash phase appears the same. Without an additional orthogonal detection method it remains ambiguous, but the size of the peak and the resemblance of a Gaussian peak shape suggest that this is the result of a the elution of single weakly bound contaminant. The ratio of material eluting during the elution phase versus the washing phase significantly changes when reaching a salt concentration of 300 mM NaCl, shown in Figure 4.5 c), reducing the

signal in the wash to almost zero. The chromatogram shows a broad distribution of significant absorbance signal across many fractions. As opposed to the chromatogram of 200 mM, the signal does not follow the shape of a single model peak as smoothly, indicating that it is the result of multiple components being retained to different degrees. Reaching a salt concentration of 400 mM NaCl, narrows the distribution of absorption signal to first 0,5 mL of the chromatogram. The lack of absorption signal in all subsequent fractions including the washing step shows that neither the protein of interest, nor any of the contaminants experience strong retention under these conditions. Further increases of salt concentration would therefore only lead to a further loss in selectivity, eventually causing all components to elute at their specific void volumes.

The overall trend observed in this data follows the expectation, that increasing counter-ion shift the equilibrium for the proteins to remain in the mobile phase, thereby lowering their retention factors. Following this principle, combined with the knowledge of sample composition gained from the prefractionation reference chromatogram, and that the peaks under isocratic condition will closely resemble an exponentially modified Gaussian peak, allows deconvolution of the high-throughput chromatogram into some pseudo-components. Detected peaks were assigned to one pseudo-component if their areas are within the standard error of each other, and their retention factor decreases within the tolerance of its standard error with increasing salt concentration. Following these rules allowed distinguishing between a total of nine pseudo-components for the three prefractionation fractions containing the protein of interest and the major contaminants. For five components the peak could be detected in at least 3 instances, allowing for the calculations of standard errors on the regressed parameters. The results of these regressions are listed in Table 4.2.

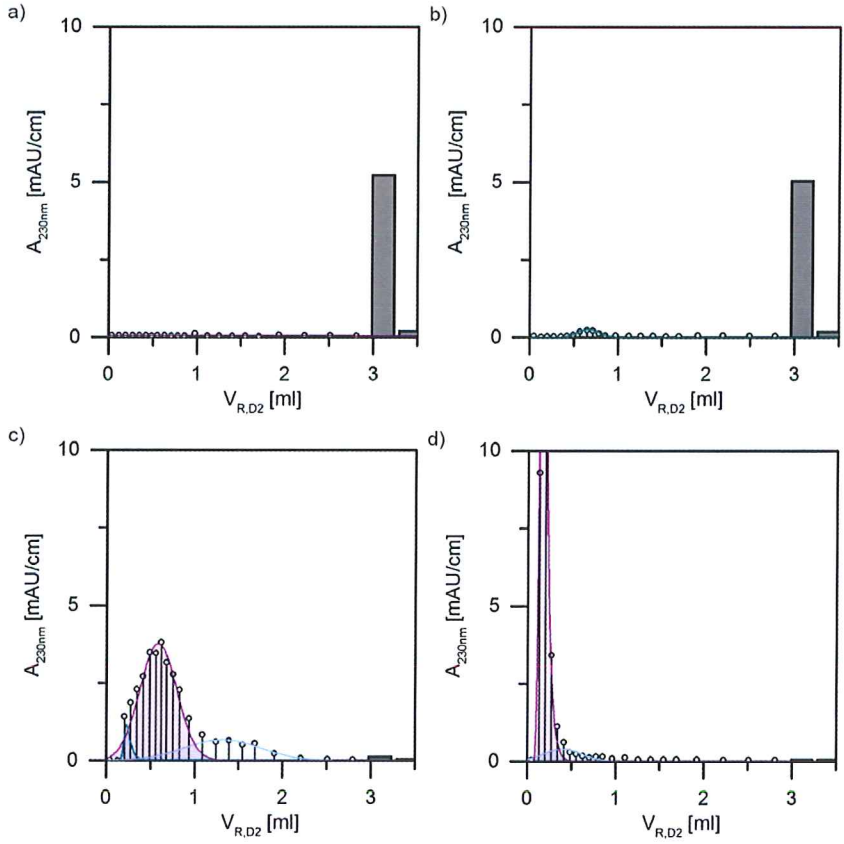


Figure 4.5 Examples of reconstructed high-throughput chromatograms and fitted peaks of the fraction collected between pH 9.8 and 9.0 in the anion exchange pH gradient for different salt concentrations: a) 100 mM b) 200 mM NaCl, c) 300 mM and d) 400 mM NaCl. The transmission path normalized absorption at 230 nm (**O**) is plotted at the average retention volume of each fraction. The normalized absorptions of the last two collected fractions which correspond to the column cleaning steps are shown as grey bars. Fitted peaks are shown by solid lines and are coloured according to assumed identities.

Table 4.2 Regressed ion-exchange isotherm parameters for the peaks found in the high-throughput chromatograms of the three main fractions. Standard errors are provided where more than two occurrences of the corresponding peak could be detected. The areas correspond to the average area of the peaks associated to the pseudo-component and are reportet together with their standard deviation.

Fraction [pH]	Area [(mAU*ml)/cm]		$\ln(K_{eq})$ [-]		v [-]	
9.8 – 9.0	181.8	± 16.4	8.2	± 2.8	2.8	± 1.0
9.8 – 9.0	13.6	± 5.8	2.3	± 1.1	0.5	± 0.5
9.8 – 9.0	43.5	± 18.9	15.5	± 0.6	5.1	± 0.2
9.0 – 8.5	25.4	± 5.6	3.0	± 0.5	0.9	± 0.2
9.0 – 8.5	2.4	± 1.2	6.2	-	1.6	-
8.5 – 8.0	29.5	± 0.8	13.3	-	6.1	-
8.5 – 8.0	37.8	± 2.5	8.2	-	3.1	-
8.5 – 8.0	149.8	± 16.3	2.8	± 0.1	0.9	± 0.1
8.5 – 8.0	48.1	± 27.7	33.0	-	12.0	-

The equilibrium constants were regressed with an average standard error of 21%, with a maximum of 48%. The effective protein charge was determined with slightly less certainty, resulting in an average standard error 38% for this parameter, with a maximum of 100%. The largest error for the effective protein charge coincides with the instance for which the lowest value was determined. Based on these regressed parameters, the retention coefficients can be inter and extrapolated to the salt concentrations that were not measured, or where the peak was not detected. The predicted retention curves are shown in Figure 4.6.

In some cases the predicted retention curve is less steep than expected, as they predict retentions of lower than 3 ml, yet no peaks are detected eluting at salt concentrations below 100 mM. This under prediction is partly the result of the weighted regression. The errors on the first moment tend to increase with increasing retention, as peaks with a larger residence time get broader. To avoid this tendency would require testing at more low salt concentration conditions in which the peaks first moment of the peak can still be

accurately determined. Practically this is difficult to realize, given the exponential character of the salt dependence. In this context a gradient regime as described in Chapter 3 can has the benefit that less prior knowledge is necessary to perform the experiments under conditions were the components elute before a the washing step. The advantage of an isocratic regime as used here on the other hand, is that the relationship used for parameter regression relies on less assumptions and simplifications than the gradient based formalism.

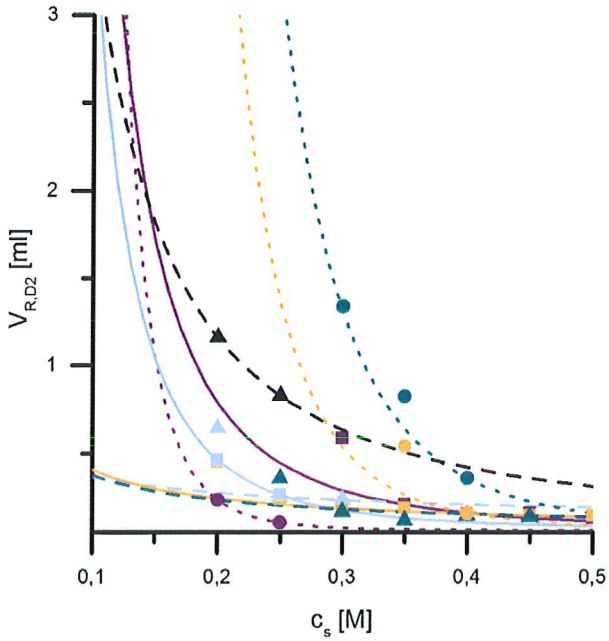


Figure 4.6 Experimental (symbols) and model predicted (lines) second dimension retention volumes of the of CHO IgG-1 supernatant component peaks listed in Table 4.2 for POROS 50 HS at pH 4.5 in dependence of the molar concentration of sodium chloride (c_s).

Introduction of an orthogonal third separation dimension to increase the data density might also suffice to reduce the error on the first moments of the later eluting peaks to a sufficient degree to supress this under prediction. The major technical challenge to

overcome in this context is the large sample number associated with performing these experiments in a true high-throughput fashion. The system described in this study is capable of screening 16 sample and condition combinations in the course of 2 hours. With the current fractionation scheme that is already optimized towards collecting as few fractions as necessary, this already results in a total of 384 fractions that would need to be analysed by a third dimension. In order to not introduce an analytical bottleneck, the throughput of the third dimension would need to be less than 20 seconds per sample. The only analytical techniques able to differentiate a large number of proteins at the same time to offer this kind of throughput currently available are MALDI-TOF mass-spectrometry and capillary gel electrophoresis (CGE). Despite their quick measurement time per sample, both techniques require extensive sample preparation, including a desalting step, and addition of some reagents. These steps are straight-forward in terms of automation and can be realized with liquid handling systems equipped with disposable solid phase extraction modules, but introducing such a large number of additional sample processing steps may increase the risk of technical faults occurring during the screening.

4.5 Conclusions

In this study we demonstrate that the principles of feedstock profiling using multi-dimensional chromatographic separations introduced in Chapter 3 can also be applied in a true high-throughput fashion, using robotic liquid handling systems. Introduction of a novel absorption based meniscus volume correction allowed to significantly reduce the largest source of experimental noise associated with the use of such high-throughput systems and reduce the maximum error of volume estimation to about 3%.

Conducting the screening operations in isocratic mode, allowed regression of ion-exchange isotherm parameters without relying on as many assumptions as an analytical solution of a

gradient based relationship would require. A further advantage of the isocratic regime is that it can be most closely mimicked by the liquid handling systems, as these systems tend to lack dual pumps that are necessary to create true linear gradients.

The lack of an analytical third dimension capable of robustly handling the large amount of samples generated with such an high-throughput approach was identified as the greatest drawback of the approach, increasing the first moment errors of overlapping and preventing pseudo-component matching with the same degree of certainty. Nevertheless the approach is shown to be able to produce results following the expected trends and can quickly screen for conditions that show the highest selectivity. For samples of relative low complexity it can yield ion-exchange isotherm parameters with reasonable degrees of uncertainty.

Acknowledgements

This work was supported by the BE-Basic foundation, a public private partnership of knowledge institutes, industry and academia, under the project FS2.003. We want to thank out industrial partners for financial support and valuable input. We acknowledge Synthon Biopharmaceuticals B.V. for providing the CHO cell culture supernatant analysed in this project.

References

[1] A.T. Hanke, M. Ottens, Purifying biopharmaceuticals: knowledge-based chromatographic process development, Trends Biotechnol. 32 (2014) 210-220.

[2] B.K. Nfor, T. Ahamed, M.W. Pinkse, L.A. van der Wielen, P.D. Verhaert, G.W. van Dedem, M.H. Eppink, E.J. van de Sandt, M. Ottens, Multi-dimensional fractionation and characterization of crude protein mixtures: toward establishment of a database of protein purification process development parameters, Biotechnol. Bioeng. 109 (2012) 3070-3083.

[3] M. Wiendahl, P.S. Wierling, J. Nielsen, D.F. Christensen, J. Krarup, A. Staby, J. Hubbuch, High throughput screening for the design and optimization of chromatographic processes - Miniaturization, automation and parallelization of breakthrough and elution studies, Chem. Eng. Technol. 31 (2008) 893-903.

[4] M. Devices, Making optical density measurements automatically corrected to a 1 cm pathlength in the SPECTRAmax® PLUS microplate spectrophotometer, Application Note (2010).

[5] A. Osberghaus, K. Drechsel, S. Hansen, S.K. Hepbildikler, S. Nath, M. Haindl, E. von Lieres, J. Hubbuch, Model-integrated process development demonstrated on the optimization of a robotic cation exchange step, Chem. Eng. Sci. 76 (2012) 129-139.

[6] P.S. Wierling, R. Bogumil, E. Knieps-Grunhagen, J. Hubbuch, High-throughput screening of packed-bed chromatography coupled with SELDI-TOF MS analysis: monoclonal antibodies versus host cell protein, Biotechnol. Bioeng. 98 (2007) 440-450.

[7] D.o.C. Nomenclature, S.R.I.U.o. Pure, A. Chemistry, M. Nic, J. Jirat, B. Kosata, IUPAC goldbook, IUPAC, 2006.

[8] J.M. Mollerup, Applied thermodynamics: A new frontier for biotechnology, Fluid Phase Equilib. 241 (2006) 205-215.

[9] L. Pedersen, J. Mollerup, E. Hansen, A. Jungbauer, Whey proteins as a model system for chromatographic separation

of proteins, J. Chromatogr. B Analyt. Technol. Biomed. Life. Sci. 790 (2003) 161-173.

[10] B.C. To, A.M. Lenhoff, Hydrophobic interaction chromatography of proteins. I. The effects of protein and adsorbent properties on retention and recovery, J. Chromatogr. A 1141 (2007) 191-205.

[11] C.A. Brooks, S.M. Cramer, Steric Mass-Action Ion-Exchange - Displacement Profiles and Induced Salt Gradients, AIChE J. 38 (1992) 1969-1978.

[12] J.M. Mollerup, The thermodynamic principles of ligand binding in chromatography and biology, J. Biotechnol. 132 (2007) 187-195.

[13] B.K. Nfor, M. Noverraz, S. Chilamkurthi, P.D. Verhaert, L.A. van der Wielen, M. Ottens, High-throughput isotherm determination and thermodynamic modeling of protein adsorption on mixed mode adsorbents, J. Chromatogr. A 1217 (2010) 6829-6850.

[14] T. Ahamed, S. Chilamkurthi, B.K. Nfor, P.D. Verhaert, G.W. van Dedem, L.A. van der Wielen, M.H. Eppink, E.J. van de Sandt, M. Ottens, Selection of pH-related parameters in ion-exchange chromatography using pH-gradient operations, J. Chromatogr. A 1194 (2008) 22-29.

[15] P.R. Levison, C. Mumford, M. Streater, A. BrandtNielsen, N.D. Pathirana, S.E. Badger, Performance comparison of low-pressure ion-exchange chromatography media for protein separation, J. Chromatogr. A 760 (1997) 151-158.

5

Prediction of protein retention times in hydrophobic interaction chromatography by robust statistical characterization of their atomic-level surface properties

Abstract

The correlation between the dimensionless retention times (DRT) of proteins in hydrophobic interaction chromatography (HIC) and their surface properties were investigated. A ternary atomic-level hydrophobicity scale was used to calculate the distribution of local average hydrophobicity across the proteins surfaces. These distributions were characterized by robust descriptive statistics to reduce their sensitivity to small changes in the three-dimensional structure. The applicability of these statistics for the prediction of protein retention behaviour was looked into. A linear combination of robust statistics describing the central tendency, heterogeneity and frequency of highly hydrophobic clusters was found to have a good predictive capability (R2 = 0.78), when combined a factor to account for protein size differences. The achieved error of prediction was 35% lower than for a similar model based on a description of the protein surface on an amino acid level. This indicates that a robust and mathematically simple model based on an atomic description of the protein surface can be used for the prediction of the retention behaviour of globular proteins with a well determined 3D structure in HIC.

Keywords: Hydrophobic interaction chromatography; Protein surface properties; Retention time prediction; Atomic-level surface description; Robust statistics;

Published as: A.T. Hanke, P.D.E.M. Verhaert, L.A.M. van der Wielen, M.H.M. Eppink, E.J.A.X. van de Sandt and M. Ottens, **Biotechnol. Prog.**, 1394
doi:10.1002/btpr.2219

5.1 Introduction

Hydrophobic interaction chromatography (HIC) is one of the fundamental chromatographic techniques applied for the separation of complex biological mixtures but defining the design space for a HIC step is not a straightforward task [1]. With computational power getting cheaper and high-throughput technology becoming increasingly wide-spread, science and industry are moving towards knowledge-based approaches for downstream process design [2]. To reduce the experimental burden even further, there is a growing interest in either relating the chromatographic behaviour of proteins in HIC to simpler experiments [3-5] or to predict them directly from information on either the protein's amino acid composition [6] or 3D structure [7,8]. The tools used to make predictions range from relatively simple correlations to more complex quantitative structure property relationship (QSPR) models using support vector machines [9] or binding site identification through molecular docking experiments [10].To achieve good results, these approaches implicitly require the used 3D structure to closely resemble the structure of the protein under the conditions for which the behaviour is being predicted. This makes them inherently unsuited for predicting the behaviour of inherently disorder proteins (IDPs) [11]. While molecular dynamics simulation have great potential to increase our understanding of the hydrophobic interactions on an atomic level and, due to their dynamic nature, have a greater have a greater potential to be able to deal with flexible molecules and IDPs , they are still too computationally expensive to be applied to large proteins [12]. Although this limitation might be overcome in within the next years, until this is achieved, computationally inexpensive models based on descriptors of the whole molecule, or just its surface, remain the more viable option.

There are many descriptors that can be calculated to describe complex molecules [10]. However, the interactions in HIC

are thought to be mainly driven by protein surface hydrophobicity [13] and therefore descriptors of the protein surface are of particular interest for the prediction of HIC retention behaviour. Surface property attributes that have been considered to be correlated with retention times in hydrophobic interaction chromatography are the overall hydrophobicity, the heterogeneity of the protein surface and the hydrophobicity of the most extreme patches [13]. These properties can be described by a wide variety of statistical measures. Commonly used statistical measures are the average to capture the overall hydrophobicity, the standard deviation to describe the heterogeneity and the maximum value to quantify the strength of the patch with the highest affinity. Salgado et al. demonstrated that predictive models with good performance could be formulated on the basis of these statistics, provided that the values are calculated on basis of a suitable hydrophobicity scale [13]. As the interpretation of these statistics is unambiguous, they are a good starting point for the investigation of which kind of surface properties influence the retention times of proteins.

A potential downside to using these statistics as a basis for a predictive model is that they are very sensitive to small changes in the data set and outlying data points. While such sensitivities might be considered advantageous when the goal is to predict the differences in the behaviour of proteins with point mutations, there are some practical considerations to be made concerning their general use. Protein surface descriptors are typically calculated from three-dimensional structures determined in solution by NMR spectroscopy [14] or more commonly by crystal X-Ray diffraction [15]. As noted by Petsko and Ringe, despite the impression given by the structures gained form crystallography, proteins in solution are far from rigid molecules. Not only can small conformational changes affect which parts of the protein are more or less exposed, the numerical calculation of their solvent accessible surface areas itself is merely an approximation [16] and the choice of

algorithm and settings for the calculation can easily change the obtained values by 10% [17]. The reliability of a predictive model might therefore increase when surfaces descriptors with a certain robustness towards small changes in the protein structure are chosen.

The robustness of a statistical measure can be described by its breakdown point; the fraction of data that can be given an arbitrary value before the measure itself assumes an arbitrary value. The value of the average for example, can in principle change to any value based on the inclusion of a single outlying data point. The average therefore has a breakdown point of 0. When using single data point surface descriptors such as the maximum and minimum of the surface property distribution similar sensitivity problems arise. For comparison: to cause a significant change in the median of a data set 50% of the data needs to change. Due to this property the median has a break-down point of 0.5 and is considered a robust statistic.

The discussion on the robustness of descriptive statistics is not new and has been extensively covered in literature [18]. The aim of this study is to investigate if a computationally inexpensive model based on robust descriptive statistics of the atomic protein surface property distributions can achieve a reasonable quality of retention time prediction.

5.2 Methodology

5.2.1. 3D structure curation

The three-dimensional structures of the proteins in Table 1 were retrieved from the protein database PDB [19]. The missing hydrogen atoms were added and all present water molecules deleted. After this step the resulting structures were compared to the proposed biological assemblies and any excess copies of the molecule present in the asymmetric cell were removed, as were any molecules considered not to be part of the protein.

Table 1. An overview of the proteins used in this study together, with information on used structures. The experimental dimensionless retention were reported by Mahn, Lienqueo and Salgado [6,33]. Structures for which a resolution range is reported were refined from multiple datasets collected at different resolutions.

Name	Organism	PDB ID	Source	DRT (-)	Mass (kDa)	r_q (Å)	Structure curation
Cytochrome C	*equus caballus*	1HRC	1.94Å XRD	0.002	12.3	12.6	• Associated Haem group still present
Ribonuclease A	*bos taurus*	1AFU	1.7Å XRD	0.352	13.6	14.3	• Removed molecule copy B from cell
Ribonuclease T1	*aspergillus oryzae*	1RGC	10-2Å XRD	0.371	11.1	12.4	• Removed of all 3gp • Removed molecule copy B from structure file • Associated Ca++ ion still present
Metmyoglobin	*equus caballus*	1YMB	6.0-1.9 Å XRD	0.373	17.5	15.1	• Removed SO4-- ion • Haem group included
Ribonuclease T1	*aspergillus oryzae*	1TRP	2.3-2.4 Å XRD	0.482	11.1	12.4	• Removed all 2gp • Removed molecule copy from cell • Associated Ca++ ion still present
Ovotransferrin	*gallus gallus*	1OVT	2.4 Å XRD	0.504	75.6	29.5	• Kept Fe++ ions • Kept both Co++ ions
Ovalbumin	*gallus gallus*	1OVA	6.0-1.9 Å XRD	0.570	84.5	27.7	• Removed all Nag from the cell • Removed molecule copies C and D from cell • Delete one Ca++ ion
Hen egg-white Lysozyme	*gallus gallus*	2LYM	2 Å XRD	0.603	14.3	14.0	
Thaumatin Isoform A	*thaumatococcus daniellii*	1THV	2.6-1.7 Å XRD	0.663	22.1	16.3	
α-chymotrypsin A	*bos taurus*	2CHA	2 Å XRD	0.694	48.2	21.9	• Delete peptide fragment from molecule copy A and E • Left Tsu molecules in place

β-lactoglobulin A	*bos taurus*	1CJ5	NMR	0.734	18.3	14.5	• Deleted copies B-J
α-amylase	*bacillus licheniformis*	1BLI	XRD	0.753	55.1	24.0	• Kept three Ca++ ions in place • Kept one Na+ ion in place
α-chymotrypsin A	*bos taurus*	4CHA	1.68 Å XRD	0.774	47.9	21.9	• Removed peptides A and D from cell
Ribonuclease S	*bos taurus*	1RBC	XRD	0.829	11.6	14.8	• Removed molecule S from cell • Kept S04- ions in place
α-lactalbumin	*homo sapiens*	1A4V	1.8 Å XRD	0.936	14.1	14.3	• Kept both Ca++ ions in place

5.2.2. A simple atomic hydrophobicity scale

There is an abundance of residue-level scales available to rank amino acids by their relative hydrophobicity. Most of them are either derived from experimental measurement of their physicochemical behaviour, such as their partition coefficients between water and various organic solvents, chromatographic retention or influence on surface tension, or related to their probability to be located within the non-solvent accessible core of proteins in their folded state. An extensive comparison of such scales found that while the positions of several amino acids may vary greatly depending on the scale, some agreement for the relative hydrophobicity of many amino acids could be found [20]. While experimentally determined hydrophobicity scales, especially those based on the measurement of their chromatographic behaviour, appear to be attractive options to be used in a model to predict chromatographic retention times, choosing such a scale fundamentally limits the resolution with which the protein surface can be described as it is based on a residue-level. Even though the contribution of each amino acid can be normalized according to its contribution to the solvent accessible surface area, this level does not take into account which part of the residue is exposed. To overcome this limitation Kapcha et al. proposed a simple binary atomic-level

hydrophobicity scale [21]. Classification of atoms as either hydrophilic or hydrophobic was based on their calculated partial charges, where a partial charge magnitude of greater than 0.25 is to be treated as hydrophilic and less or equal than 0.25 as hydrophobic. Despite its simplicity, a good agreement with the ranking of residues based on this scale and other hydrophobicity scales was found. Even better agreement could be achieved by adopting a ternary scale further classifying atoms as charged. Atoms are classified as charged when they belong to a terminal group where the magnitude of the sum of the partial charges is greater than 0.5. The proposed ternary scale assigned a value of -0.5 to hydrophobic atoms, 1 to hydrophilic atoms and 2 to charged atoms. As the values themselves are by definition arbitrary, any scaling with the same weighting would lead to the same ranking of residues. In this study we adopted a ternary scale with the same classification criteria but with values scaled to 0 for charged atoms, 0.4 for hydrophilic atoms and 1 for hydrophobic atoms.

5.2.3. Calculation of average surface atom neighbourhood properties

The hydrophobicity of a protein surface can be described by its average surface property (ASP). The ASP can be calculated for either the entire protein surface at once or for a set of neighbourhoods around defined reference points, as described by Lienqueo et al. [22]. In this study the ASPs of the proteins listed in Table 1 were calculated for all atoms (i) contributing to the solvent accessible surface area (SASA) within neighbourhoods (N) containing the atoms (nϵN) around them:

$$ASP(i) = \frac{\sum_{n \epsilon N(i)} SASA(n) \cdot \varphi(n)}{\sum_{n \epsilon N(i)} SASA(n)} \quad (1)$$

where φ(n) is the hydrophobicity value of the atom n, assigned to it according to the atomic ternary scale described in

Section 5.2.2. Neighbourhood radii from of 5-15 Å were considered. This range was selected so each neighbourhood would always contain multiple atoms but not greatly exceed the radius of gyration of the protein. The SASAs of the single atoms and their neighbourhoods were determined with YASARA 14.6.23 (YASARA Biosciences GmBH, Vienna, Austria). The SASAs were calculated based on probe molecules with a radius of 1.4 Å with a Gaussian smoothing of the surface, as illustrated by Figure 2. This probe size was chosen to reflect the surface accessibility of a water molecule, which has an approximate diameter 2.8 Å. Water was chosen as probe molecule as HIC always takes place in aqueous solution. The distance between the geometric centres of the atoms was calculated to decide whether an atom was within a neighbourhood. For an accurate calculation of the SASA it is important that positions of the atoms in the used 3D structures are determined to a sufficient degree of certainty. Structures determined by X-Ray Diffraction with a resolution of approximately 2 Å are usually deemed sufficient [10]. Working with structures determined at lower resolutions are anticipated to have a negative impact on the quality of predictions based on the derived SASA values, as they might not accurately reflect reality. The SASA and neighbourhood atom lists generated by YASARA were imported into Matlab2013b (Mathworks, Natick, MA, USA) for calculation of the ASP values. This results in an ASP value for each neighbourhood radius for each surface atom of each protein. As these results typically consist of a large numbers of unique values, they needed to be binned before their distribution could be described by descriptive statistics. Two methods of data binning were tested. The first binning method was to control the size of the bins, regardless of the data. Bin sizes between 0.001 and 0.1 were tested. For the second method, the number of bins was controlled. In this case the range of the values in divided into an equal number of bins, resulting in a different bin size for each protein surface property distribution.

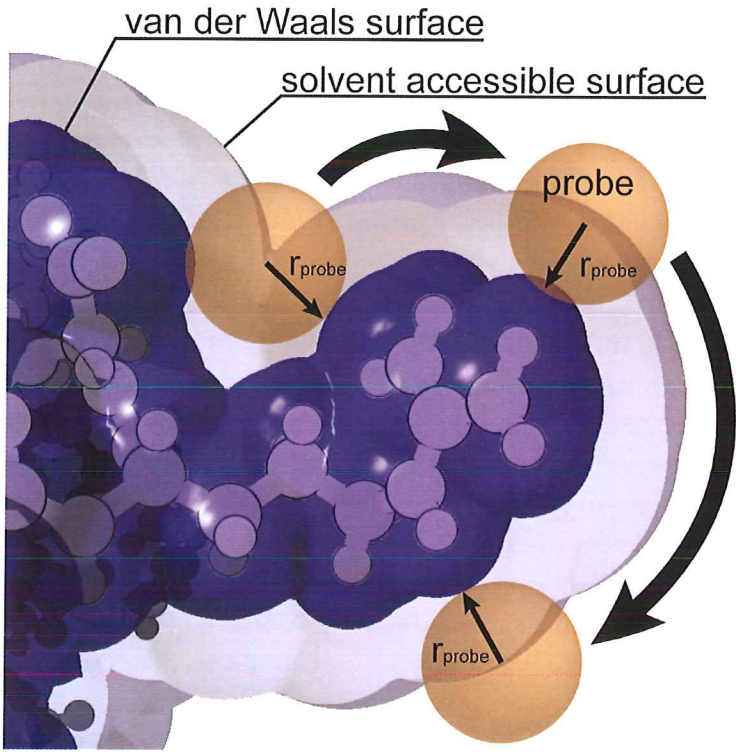


Figure 5.1. The first step in the approximation of the solvent accessible surface area (SASA) of a full protein, or neighbourhood patch, is the construction of the van der Waals surface of the molecule by creating spheres around each atom's geometric centre, corresponding to their van der Waals volumes. A probe molecule of defined radius (r_{probe}) is then rolled across the molecule's outer van der Waals surface. The points that are traced by the centre of this probe molecule are then interpolated, which results in an estimated SASA.

5.2.4. Calculation of surface property distribution statistics

To reduce the complexity of the data, descriptive statistics were calculated for each of the surface property distributions. The three main attributes of the surfaces to be captured by the descriptive

statistics were the central tendency, the heterogeneity and the relative frequency of high hydrophobicity clusters. The workflow for the calculation and evaluation of these molecular descriptors is shown in Figure 1.

Four statistics were calculated to represent the central tendency. The median (M) of each distribution was chosen as it has an asymptotic break-down point of 0.5, the maximum possible value and is therefore a very robust statistic [23]. The second investigated central tendency statistic chosen was the trimean (TM). The trimean is the weighted mean of the three quartiles of the distribution calculated by:

$$TM = \frac{Q_1 + 2Q_2 + Q_3}{4} \quad (2)$$

The three quartiles are the values that split the data into four populations of equal size. The value of Q_2 is equal to that of the median. In contrast to the median, the TM better reflects the central tendency of asymmetric distributions. However, by including this information its sensitivity increases and lowers its breakdown point to 0.25 [24]. The third central tendency measure tested was the mode. The mode represents the most frequently occurring value of the distribution. Although the mode does not always satisfy all requirements to be considered an accurate measure of location [25], it has been noted for its intuitive appeal as it represents the value with the highest probability and its insensitivity to outliers [26]. The ASP distributions in this study were not assumed to be unimodal. For distributions containing multiple modes of the same relative frequency the mean of these modes was used instead. A fourth central tendency descriptive statistic included in this study is the mode weighted by its relative frequency (MOFM).

The second attribute set out to be captured by descriptive statistics was the heterogeneity of the protein surface. In terms of the surface property distributions, more heterogeneous surfaces should

lead to a higher dispersion in ASP values and therefore a broader distribution. The broadness of a distribution can either be described by a range statistic (i.e. a statistic that reflects the span of possible values a property can take) or the average deviation of a property value from the central tendency of the distribution. A robust range statistic chosen for the quantification of the range is the inter-quartile range (IQR), which is the distance, between the first and third quartiles (Q_1 and Q_3). As it is purely defined by the quartile locations the IQR has a break-down point of 0.25, similar to the TM. Trimming of extreme values is an important tool for increasing the robustness of dispersion descriptive statistics [27]. Quartiles are an efficient tool for trimming distributions but they trim a large fraction of the data. A common measure to distinguish outliers is the introduction of a lower fence (LF) and upper fence (UF):

$$\mathrm{LF} = Q_1 - (1.5 \cdot IQR) \tag{3}$$

$$\mathrm{UF} = Q_3 + (1.5 \cdot IQR) \tag{4}$$

This leads to three dispersion statistics per central tendency statistic (CT). The average dispersion that does not exclude extreme data points (AD) was calculated by:

$$CTAD = \frac{1}{N}\sum_{n=1}^{N}(x_n - CT)^2 \tag{5}$$

It was calculated for the median, mode and trimean resulting in the median absolute dispersion (MAD), mode absolute dispersion (MOAD) and trimean absolute dispersion (TMAD).

Bounded average dispersion statistics were also calculated for the mode and trimean by:

$$CTBAD = \frac{1}{N_{(LB<x<UB)}}\sum_{n=1}^{N_{(LB<x<UB)}}\left(x_{n,(LB<x<UB)} - CT\right)^2 \tag{6}$$

Use of the first and second quartile as lower boundary (LB) and upper boundary (UB) lead to the mode and trimean interquartile average dispersions (MOIAD and TMIAD), whereas use of the LF and UF as boundaries lead to the mode and trimean fenced average dispersions (MFAD and TMFAD).

To quantify the presence of high-hydrophobicity regions (H-regions) the relative frequencies of neighbourhoods with ASP values 0.5-0.75, the third quarter of the hydrophobicity scale reflecting a mild hydrophobicity was calculated (F3), as well as the relative frequency of values in the range of 0.75-1.00 (F4). The total relative frequencies of overall hydrophobic neighbourhoods were also taken into consideration (F3+F4).

5.2.5. Size factors

In their landmark study of underlying mechanism of hydrophobic interaction chromatography To and Lenhoff noted that differences in protein retention in HIC could not be attributed to protein surfaces alone but that protein rigidity also had a large influence [28-31]. As protein rigidity is not a simple property to predict from 3D structures, the size of proteins was considered as an easy to calculate proxy measure. A simple measure for the size of a protein is its radius of gyration (r_g) that can be calculated by:

$$r_g(i) = \sqrt{\frac{1}{N} \cdot \sum_{n=1}^{N} \left(\overrightarrow{R_n} - \vec{C}\right)^2} \text{ [Å]} \qquad (7)$$

C is the geometric centre of mass of the protein. A problem with using the radius of gyration as parameter for a predictive model is that it cannot be scaled to values between 0 and 1 easily, as normalizing the values towards the maximum or inverse of the minimum would introduce an unwanted dependence of scaling on the used training set. A common measure to scale the influence of protein size on their chromatographic behaviour is the calculation of

the pore accessibility solute distribution coefficient (K_D) calculated from the extended Ogston model [32]:

$$K_D(i) = \exp\left(-ln\left(\frac{1}{1-\phi_p}\right)\cdot\left(1+\frac{r_g}{r_p}\right)^2\right) \quad (8)$$

The resin specific pore radius (r_p) and pore volume fraction (Φ_p) were taken from literature [31].

5.2.6. Selection of neighbourhood radii and binning approach

While the descriptive statistics outlined in Section 5.2.4 were calculated for all considered neighbourhood radii and binning procedures, it is not feasible to include all possible combinations for integration into a multi-parameter model. As objective selection criteria, the linear correlation between the descriptive statistics and the experimental dimensionless retention times of the proteins (DRT_{exp}) was considered. The DRT of a protein (i) is:

$$DRT(i) = \frac{t_{R,i} - t_g}{t_G - t_g} \quad (9)$$

where $t_{R,i}$ is its measured retention time, t_g the time of the beginning of the salt gradient and t_G the end of the salt gradient. The DRT values used in this study are listed in Table 1. They were reported by Lienqueo and Mahn for ammonium sulphate (AS) gradient elution experiments on 1ml Sephadex Phenyl FF columns with 2M AS starting concentration [22,33]. Despite being dimensionless, DRTs are influenced by the system specific gradient delay and may therefore vary depending on the liquid chromatographic system they were determined on [34]. The degree of linear correlation between the descriptive statistics (S) and the experimental DRTs was quantified by their Pearson correlation coefficients (PCC) calculated by:

$$PCC = \frac{cov(S, DRT_{exp})}{\sigma_S \cdot \sigma_{DRTexp}} \quad (10)$$

where σ_S and $\sigma_{DRT,exp}$ are the standard deviations of the descriptive statistics and the experimentally determined DRTs. For each statistic the neighbourhood radius and binning method leading to the highest positive correlation were identified and selected to calculate candidate predictors for multivariate models.

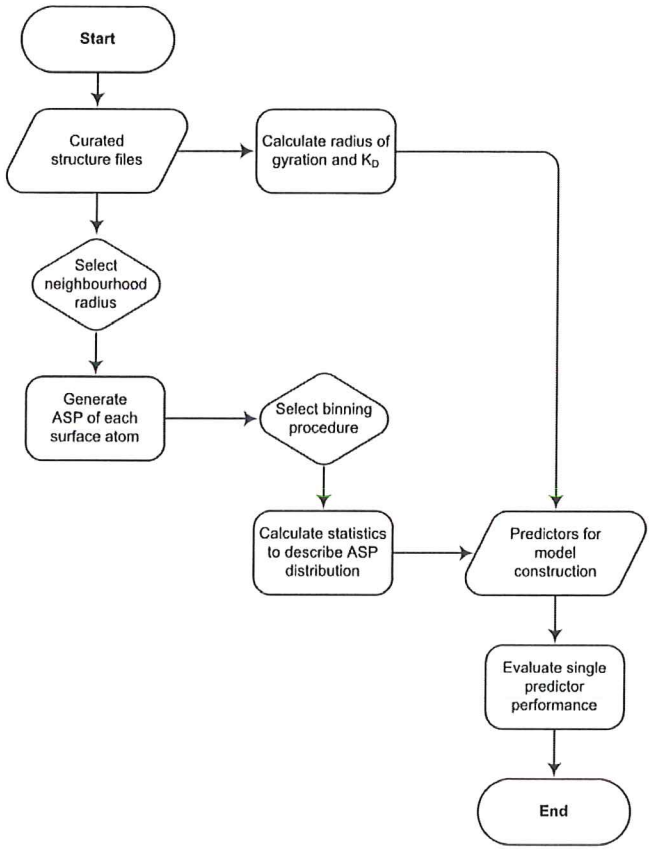


Figure 5.2. A schematic representation of the workflow for the calculation of surface property distribution statistics from 3D protein structure files. The curation of the structures files is a manual step that involves editing the scene to represent the proposed biological assembly according to the comments in Table 1.

5.2.7. Multivariate model construction and performance evaluation

Linear combinations of the descriptive statistics with the strongest positive correlation to the observed DRT were considered as predicative multivariate models. All possible linear combinations were tested, with a few restrictions: each model was only allowed to contain a single descriptive statistic for each fundamental surface distribution attribute and statistics calculated from bin size controlled distributions could not be combined with statistics calculated from bin number controlled distributions. For each model and each training set the coefficients were optimized by the lsqcurvefit algorithm of Matlab 2013b [35,36]. The performance of the models was evaluated both by the mean square error (MSE):

$$\mathrm{MSE} = \frac{1}{N} \cdot \sum_{n=1}^{N} (\widehat{Y}_n - Y_n)^2 \tag{11}$$

and the coefficient of determination:

$$\mathrm{R}^2 = 1 - \frac{\sum_{n=1}^{N} (\widehat{Y}_n - Y_n)^2}{\sum_{n=1}^{N} (\widehat{Y}_n - Y_n)} \tag{12}$$

of the predictions. The reported performance indicators and predicted DRTs are the results of Jack Knife cross validation [35,36]. The key concept is to only consider the predictions made for proteins, when the protein is not part of the training set for the model. A schematic overview of the workflow is presented in Figure 4.

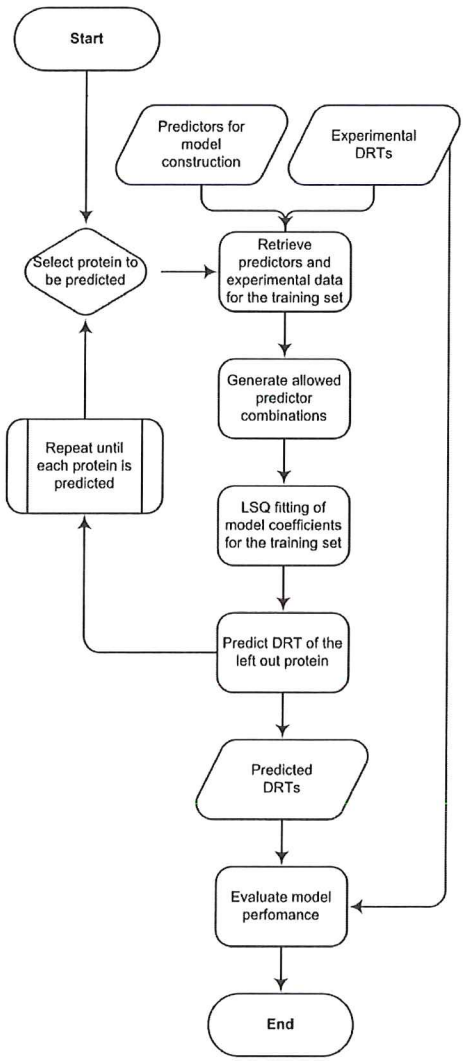


Figure 5.3. A schematic representation of the workflow applied for the creation and training of the multivariate models. The predictors considered for the model where the same as in Table 2. To predict the retention time of a protein, each model is trained with the data of the other proteins. This is repeated until each protein was predicted once, without being part of the training set. Only the fully predicted DRTs are used for evaluation of the model performance (Jack Knife cross-validation).

5.3 Results and Discussion

5.3.1. The influence of the neighbourhood radius and binning on the ASP distributions

The observed ASP distribution is a strong function of both the selected neighbourhood radius and method of binning. Figure 3 shows the ASP distributions of Cytochrome C from equine heart for two neighbourhood radii and two binning approaches. Comparison of Figures 3a) and b) show that an increase in neighbourhood radius decreases the dispersion of the distribution. This behaviour is universal, as a larger neighbourhood represents a larger fraction of the total protein area. The central tendency of the distribution must therefore converge towards to average total surface property for large neighbourhood radii, whereas the dispersion of the distribution must converge to zero. Table 2 shows the radii and binning settings for which each statistic had the highest positive correlation with the experimental data. The majority of statistics performs best for small radii in the range between 6-9Å, the only outliers to this trend being the frequency of the weighted mode (MOFM) and H-region frequencies that include mildly hydrophobic values (F3).

Table 5.2 A summary of the highest Pearson correlation coefficients (PCC) found between the calculated individual surface property statistics described in Section 2.4 and the experimental data presented in Table 5.1. The binning column yields information on the binning method used during the calculation of the underlying statistic, where 's' stands for a controlled bin size and 'n' for a controlled bin number. The neighbourhood radius that lead to the statistic with the highest correlation coefficient is noted as $r_{N\_opt.}$

	Statistic	Binning	$r_{N\ opt}$	PCC
Central tendency	M	-	6 Å	0,40
	TM	-	8 Å	0,37
	MO	s: 0,01	6 Å	0,63
		n: 10	7 Å	0,74
	MOFM	s: 0,01	13 Å	0,29
		n: 27	13 Å	0,62
	IQR	-	9 Å	0,34
	MAD	-	8 Å	0,36
	MOAD	s: 0,001	9 Å	0,48
		n: 30	8 Å	0,52
	MOIAD	s: 0,1	8 Å	0,59
		n: 10	8 Å	0,63
	MOFAD	s: 0,001	9 Å	0,54
		n: 30	8 Å	0,57
Heterogeneity	TMAD	-	8 Å	0,18
	TMIAD	-	8 Å	0,41
	TMFAD	-	8 Å	0.29
H-region freq.	F3	-	13 Å	0,07
	F4	-	7 Å	0,49
	F3+F4	-	12 Å	0,37

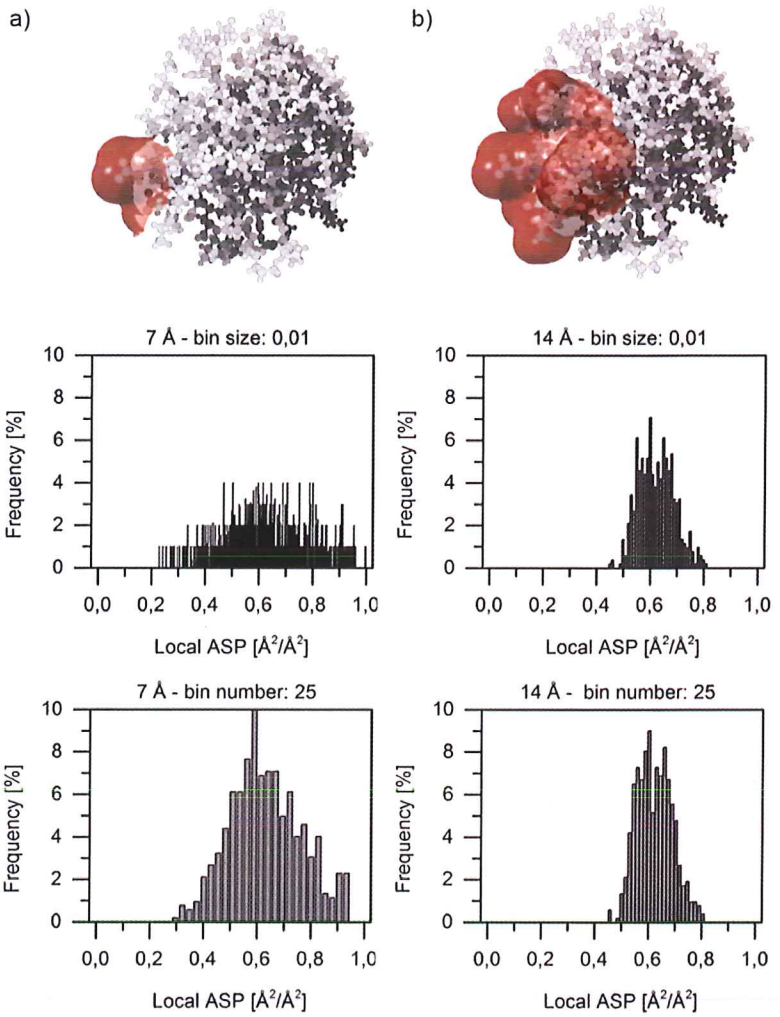


Figure 5.4. An example of the influence of the selected neighbourhood radius and binning method on the observed surface property distributions for the example of Cytochrome C from equine heart (1HRC). a) The surface neighbourhood of atom 56 with a neighbourhood cut-off radius of 7 Å around the geometric atom centre together with the ASP distribution for the entire surface controlled for a bin size of 0.01 and a bin number of 25. b) The same as a) but with a 14 Å neighbourhood cut-off radius. Comparison of a) an b) shows how the distribution converges towards its central tendency for an increasing neighbourhood radius.

Statistics that are defined in relation to the mode are further influenced by the binning method. A representative example for the influence of the choice of neighbourhood radius and binning method on the performance of dispersion statistic related to the mode is shown in Figure 4. Comparison of Figure 4 a) and b) shows that neighbourhood radii of 6-8 Å have the best positive correlation with the experimental DRT for all tested binning approaches. The influence of the binning method shows less clear trends as often very similar values appear at irregular intervals, especially when controlling the number of bins. This is why the MOIAD and MOFAD, two statistics with very similar definitions, appear to be best correlated at opposing ends of the tested binning approaches. The bin sizes and bin numbers controlled binning methods show similar trends across neighbourhood radii for small bin sizes and large bin numbers, but diverge for large bin sizes and small bin numbers. This is related to dispersion of the distribution converging towards the central tendency for large neighbourhood radii. For large bin sizes a reduction of the range of the values leads to an inability to distinguish small differences in dispersion. This does not occur when controlling the number of bins, as the bin size is then scaled to the dispersion of the values. A downside to controlling the bin number however is that it does not allow comparing frequencies of similar values across proteins.

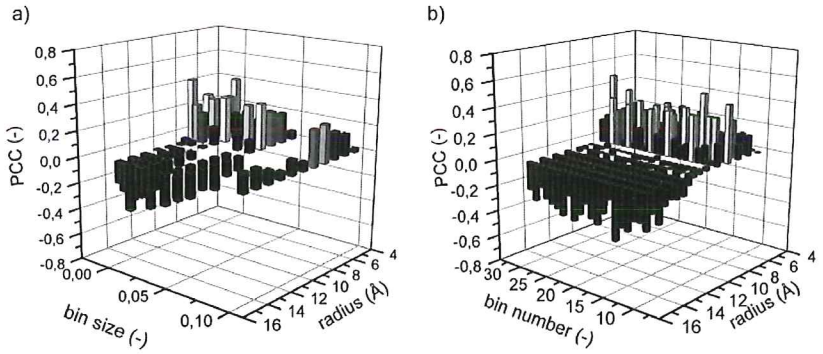


Figure 5.5. The influence of the neighbourhood cut-off radius and binning method on the Pearson correlation coefficient (PCC) of the mode average dispersion (MOAD) of the average surface property distribution and the experimentally determined dimensionless retention times from Table 5.1 a) shows the data when the distribution of each protein is controlled for a constant bin size whereas b) shows the effect when the distributions are controlled to each use the same number of bins and the size of the bins is calculated based on the range of the distribution data for each protein.

5.3.2. Performance of the multivariate models

The linear combination model built from the statistics in Table 5.2 and the calculated r_g and K_D values that gave the best predictive performance was:

$$DRT(i) = 17.0(\pm 4.2) \cdot \mathrm{MOFM}_{13\text{Å},i} + 14.3(\pm 2.4) \cdot \mathrm{MOAD}_{8\text{Å},i} + 2.7(\pm 1.0) \cdot (\mathrm{F3} + \mathrm{F4})_{12\text{Å},i} - 10.8(\pm 2.8) \cdot K_{D,i} \quad (13)$$

It should be noted, that even though coefficients were not restricted to positive values, each statistic that had a positive correlation retained its positive contribution in the multivariate model. Its R^2_{JK} is 0.78 and the Jack Knife cross-validation mean square error (MSE_{JK}) is $12.4*10^{-3}$. The 95% confidence intervals for

the determined coefficients never exceed 40% of their value, meaning that they are quite well determined considering the size of the data set. The fully predicted retention times and distribution of error are shown in Figure 6. As can be seen in Figure 6 b), the errors are fairly evenly distributed across the range of investigated proteins. While this would be just a systematic consequence of using least-square optimizers when looking at the fitted responses of the training sets, in terms of the fully predicted retention times, it is a good indicator that the model is not biased towards strongly or weakly binding species.

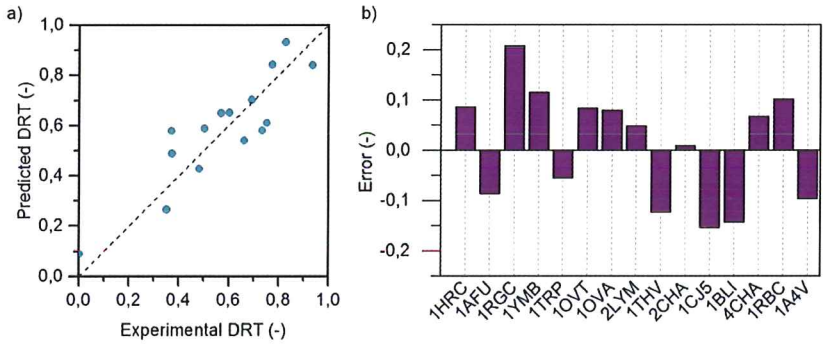


Figure 5.6. The performance of the best multivariate model. a) The predicted dimensionless retention time (DRT) against the experimentally determined values. b) The absolute error in prediction between the DRT predicted by the best multivariate model and the experimental value for each protein species. The PDB codes correspond to the proteins in Table 1.

In agreement with the concepts discussed in Section 5.2.5, a negative coefficient was found for the size-dependent factor K_D. Smaller proteins have a larger K_D, so of two proteins with equal surface properties but different size, the smaller protein would exhibit less retention in HIC. Whether this is due to increased flexibility of larger proteins, or due to an increased possibility of multiple binding sites to interact with the resin at the same time, cannot be determined at this point. Even though the K_D values are

derived from the r_g, models containing the K_D performed universally performed better than their r_g, counterparts. This indicates that it is the property of the K_D to not pass through the origin, rather than the scaling of the values that increases its performance.

Of the five best performing models, all used the MOFM as their central tendency statistics and all were built from statistics controlled by bin size rather than number. This is remarkable, as the bin number controlled statistics tended to show slightly higher correlations with the experimental data when considered individually. Similarly, the bin size controlled MOFM was the central tendency statistic with the overall worst correlation, yet present in each best performing model. The follow up models all substituted either the dispersion statistics for one of its bounded counter parts, or the H-region statistics for F4, but overall showed the same trends in coefficients.

To investigate the importance of each parameter in the final model, the impact on the predictive performance of the model after removal of each of the selected variables was performed. An overview of the results is presented in Figure 7. The F3+F4 statistic describing the relative frequency of neighbourhoods that are considered hydrophobic is least important variable, yet its removal still increases the error of prediction by a factor of 2.2 compared to the base case. Removal of any of the other variables increases the error of prediction at least 3.7 fold, with the dispersion statistic MOAD being the most important with a 5 fold increase. While model simplicity is an important quality, a more than twofold decrease of prediction quality indicates that none of the proposed parameters should be removed from the proposed model.

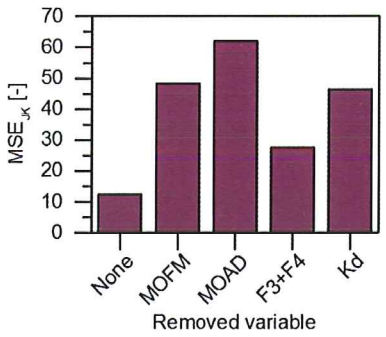


Figure 5.7. Change in the predictive performance of the multivariate model when selected variables are removed. The reported mean square error (MSE_{JK}) is the result of jack-knife cross validation over the full reported protein set.

In comparison to a similar model by Salgado et al. that calculated the local ASP distributions on an amino acid level and used classical statistics descriptors a reduction of the MSE_{JK} by 35% could be achieved, although it should be considered that this model uses one parameter less [13]. However the parameters in the model proposed here are overall better defined, indicating that the model is not over parameterized. Compared to a model based on only the amino acid sequence, rather than defined 3D structures, no increase in predictive performance was achieved [6]. That approach however requires a very large training data set to be able to estimate the degree of surface exposition of each amino acid in the sequence. A previously reported linear multivariate model that only used global ASPs had a 40% larger MSE_{JK} [36]. The trend towards increased predictive performance by increasing the level of surface description detail reported by Salgado et al. seems to hold true for the transition to an atomic description of the protein surface [13].

5.4 Conclusions

A computationally inexpensive and mathematically simple model with well-defined parameters and reasonable predictive performance was found to be able to predict the dimensionless retention times of globular proteins with well-determined 3D structures in HIC. The use of simple descriptive statistics combined with a size factor as predictors allow for a straightforward interpretation of the results. Performing surface property calculations at an atomic level increased the resolution with which the protein surface could be characterized and allowed for an easy definition a neighbourhood. The use of this scale, as opposed to a more complicated hydrophobicity scale, did not restrict the usefulness of the model. Atomic hydrophobicity scales are still an active research area[37]. Further refinement of the used hydrophobicity scale might lead to further improvement of the model. As all used descriptive statistics are derived from large fractions of the distribution data, the model promises increased robustness towards small deviations in ternary structure. In accordance with the findings of To et al., the surface properties alone were not sufficient to fully explain the experimental behaviour of the investigated proteins. Inclusion of a penalizing factor for small proteins showed an improvement of model performance. Further research into more complex size and flexibility descriptors are obvious areas for future improvement. The same holds true for the choice of statistics for the description of the protein surface. The proposed statistical procedures are robust but work best for close to normally distributed data. Further research will be necessary to determine suitable statistics for this type of data. Nevertheless, while there is no limitation on how many molecular descriptors can be calculated from any static structure, one should consider their robustness towards naturally occurring small changes when striving to predict the behaviour of proteins.

Acknowledgements

This work was supported by the BE-Basic foundation, a public private partnership of knowledge institutes, industry and academia, under the project FS2.003. We also acknowledge Prof. G.J. Witkamp and Prof. W.R. Hagen for their valuable comments during this project.

References

[1] C. Jiang, L. Flansburg, S. Ghose, P. Jorjorian, A.A. Shukla, Defining process design space for a hydrophobic interaction chromatography (HIC) purification step: application of quality by design (QbD) principles, Biotechnol. Bioeng. 107 (2010) 985-997.

[2] A.T. Hanke, M. Ottens, Purifying biopharmaceuticals: knowledge-based chromatographic process development, Trends Biotechnol. 32 (2014) 210-220.

[3] B.K. Nfor, N.N. Hylkema, K.R. Wiedhaup, P.D. Verhaert, L.A. van der Wielen, M. Ottens, High-throughput protein precipitation and hydrophobic interaction chromatography: salt effects and thermodynamic interrelation, J. Chromatogr. A 1218 (2011) 8958-8973.

[4] J. Chen, T. Yang, Q. Luo, C.M. Breneman, S.M. Cramer, Investigation of protein retention in hydrophobic interaction chromatographic (HIC) systems using the preferential interaction theory and quantitative structure property relationship models, React. Funct. Polym. 67 (2007) 1561-1569.

[5] J. Chen, S.M. Cramer, Protein adsorption isotherm behavior in hydrophobic interaction chromatography, J. Chromatogr. A 1165 (2007) 67-77.

[6] J.C. Salgado, I. Rapaport, J.A. Asenjo, Prediction of retention times of proteins in hydrophobic interaction chromatography using only their amino acid composition, J. Chromatogr. A 1098 (2005) 44-54.

[7] A. Mahn, J.A. Asenjo, Prediction of protein retention in hydrophobic interaction chromatography, Biotechnol. Adv. 23 (2005) 359-368.

[8] A. Mahn, M.E. Lienqueo, J.A. Asenjo, Effect of surface hydrophobicity distribution on retention of ribonucleases in hydrophobic interaction chromatography, J. Chromatogr. A 1043 (2004) 47-55.

[9] J. Chen, T. Yang, S.M. Cramer, Prediction of protein retention times in gradient hydrophobic interaction chromatographic systems, J. Chromatogr. A 1177 (2008) 207-214.

[10] M.E. Lienqueo, A. Mahn, G. Navarro, J.C. Salgado, T. Perez-Acle, I. Rapaport, J.A. Asenjo, New approaches for predicting protein retention time in hydrophobic interaction chromatography, J. Mol. Recognit. 19 (2006) 260-269.

[11] V.N. Uversky, Proteins without unique 3D structures: biotechnological applications of intrinsically unstable/disordered proteins, Biotechnol. J. 10 (2015) 356-366.

[12] S. Amrhein, S.A. Oelmeier, F. Dismer, J. Hubbuch, Molecular dynamics simulations approach for the characterization of peptides with respect to hydrophobicity, J. Phys. Chem. B 118 (2014) 1707-1714.

[13] J.C. Salgado, I. Rapaport, J.A. Asenjo, Predicting the behaviour of proteins in hydrophobic interaction chromatography. 2. Using a statistical description of their surface amino acid distribution, J. Chromatogr. A 1107 (2006) 120-129.

[14] K. Wuthrich, Protein structure determination in solution by NMR spectroscopy, J. Biol. Chem. 265 (1990) 22059-22062.

[15] G.A. Petsko, D. Ringe, Fluctuations in protein structure from X-ray diffraction, Annu. Rev. Biophys. Bioeng. 13 (1984) 331-371.

[16] T.J. Richmond, Solvent accessible surface area and excluded volume in proteins. Analytical equations for overlapping spheres and implications for the hydrophobic effect, J. Mol. Biol. 178 (1984) 63-89.

[17] L. Cavallo, J. Kleinjung, F. Fraternali, POPS: A fast algorithm for solvent accessible surface areas at atomic and residue level, Nucleic Acids Res. 31 (2003) 3364-3366.

[18] P.J. Huber, E.M. Ronchetti, Wiley, Hoboken, 2009.

[19] H.M. Berman, The Protein Data Bank, Nucleic Acids Res. 28 (2000) 235-242.

[20] G. Trinquier, Y.H. Sanejouand, Which effective property of amino acids is best preserved by the genetic code? , Protein Engineering Design and Selection 11 (1998) 153-169.

[21] L.H. Kapcha, P.J. Rossky, A simple atomic-level hydrophobicity scale reveals protein interfacial structure, J. Mol. Biol. 426 (2014) 484-498.

[22] M.E. Lienqueo, A. Mahn, J.A. Asenjo, Mathematical correlations for predicting protein retention times in hydrophobic interaction chromatography, J. Chromatogr. A 978 (2002) 71-79.

[23] P. Huber, in: M. Lovric (Ed.), International Encyclopedia of Statistical Science, Springer Berlin Heidelberg, 2014, p. 1248-1251.

[24] B.M. Brown, Symmetric Quantile Averages and Related Estimators, Biometrika 68 (1981) 235-242.

[25] D.R. Bickel, Robust and efficient estimation of the mode of continuous data: The mode as a viable measure of central tendency, Journal of Statistical Computation and Simulation 73 (2003) 899-912.

[26] D.R. Bickel, R. Fruhwirth, On a fast, robust estimator of the mode: Comparisons to other robust estimators with applications, Computational Statistics & Data Analysis 50 (2006) 3500-3530.

[27] J.W. Tukey, D.H. McLaughlin, Less Vulnerable Confidence and Significance Procedures for Location Based on a Single Sample: Trimming/Winsorization 1, Sankhyā: The Indian Journal of Statistics, Series A (1961-2002) 25 (1963) 331-352.

[28] B.C. 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.

[29] B.C. To, A.M. Lenhoff, Hydrophobic interaction chromatography of proteins III. Transport and kinetic parameters in isocratic elution, J. Chromatogr. A 1205 (2008) 46-59.

[30] B.C. To, A.M. Lenhoff, Hydrophobic interaction chromatography of proteins. II. Solution thermodynamic properties as a determinant of retention, J. Chromatogr. A 1141 (2007) 235-243.

[31] B.C. To, A.M. Lenhoff, Hydrophobic interaction chromatography of proteins. I. The effects of protein and adsorbent properties on retention and recovery, J. Chromatogr. A 1141 (2007) 191-205.

[32] D.A. Horneman, M. Wolbers, M. Zomerdijk, M. Ottens, J.T. Keurentjes, L.A. van der Wielen, Surfactant-aided size exclusion chromatography, J. Chromatogr. B Analyt. Technol. Biomed. Life. Sci. 807 (2004) 39-45.

[33] A. Mahn, G. Zapata-Torres, J.A. Asenjo, A theory of protein-resin interaction in hydrophobic interaction chromatography, J. Chromatogr. A 1066 (2005) 81-88.

[34] M.A. Quarry, M.A. Stadalius, T.H. Mourey, L.R. Snyder, General-Model for the Separation of Large Molecules by Gradient Elution - Sorption Versus Precipitation, J. Chromatogr. 358 (1986) 1-16.

[35] Y.D. Cai, G.P. Zhou, K.C. Chou, Support vector machines for predicting membrane protein types by using functional domain composition, Biophys. J. 84 (2003) 3257-3263.

[36] J.C. Salgado, I. Rapaport, J.A. Asenjo, Predicting the behaviour of proteins in hydrophobic interaction chromatography. 1: Using the hydrophobic imbalance (HI) to describe their surface amino acid distribution, J. Chromatogr. A 1107 (2006) 110-119.

[37] D.V. Nicolau, Jr., E. Paszek, F. Fulga, D.V. Nicolau, Mapping hydrophobicity on the protein molecular surface at atom-level resolution, PLoS One 9 (2014) e114042.

6

Outlook

The previous chapters of this thesis introduced a series of ideas and technologies that aim towards improving our ability to develop effective processes for the purification of biotechnologically produced drugs. The idea behind all of them is quite simple: the more we know about how the composition and behaviour of the components in our mixture behave, the better we can exploit their differences to achieve a separation. To gain this knowledge, we have developed both experimental techniques to measure what we need to know, and correlative techniques, that allow us to make predictions of how molecules of known structure will behave, based on past experience with similar components.

One of the main lessons of the experimental sections of this thesis, is that when working with mixtures of unknown composition, adding additional degrees of distinction, be it through addition of detectors or separation dimensions, ultimately improves the analyst's ability to interpret the data as a whole. While the addition of separation dimensions may dramatically increases the experimental time required per sample, the addition of detectors may allow distinguishing similar components, without increasing the overall analytical burden. Diode array detectors, multi-angle light scattering cells and refractive index detectors are all examples of additional data streams that could recorded during a single chromatographic separation and fed into a chromatogram deconvolution algorithm to reduce its errors, by reducing ambiguities. In the long term, the

ongoing miniaturization and parallelization driven by research in microfluidics, will also make it feasible to include many chemical and biological assays into the characterization platform.

While expanding the complexity and throughput of such experimental platforms, data management and processing become increasingly important. Already at the small scale of an academic research laboratory, such a setup requires setting up databases, that can pool and organize the raw data output. While designing such a system, flexibility and scalability need to carefully considered. Once such a database has gained a critical size, it will open new possibilities for data mining and machine learning. As mentioned earlier, the predictive techniques discussed in this thesis, need to be trained with data from known scenarios. As such, they will be among the first techniques to benefit from the availability of large data sets, such as the ones generated by experimental characterization platforms. It is unlikely that these kinds of predictions will fully replace experiments in the near future. A more likely application of these techniques will be to a priori determine the most interesting parameter combinations for experimental testing. Given the ever increasing number of potential resin and condition combinations that can be considered, such a reduction would be a highly beneficial step.

A similar effect could be achieved by establishing correlations for adsorption behaviour of molecules across different resins. Ideally, these correlations would be based on a mechanistic understanding of the adsorption process of very large molecules. At this point in time, molecular dynamic simulations appear to be the most promising route to gain this understanding, but they still need to make the transition towards the simulation of actual adsorptions, rather than short time energy calculations.

In summary we can say that it will take a lot more work for us to figure out how to properly work less.

Summary

Biopharmaceuticals, and among them therapeutic proteins, are becoming an increasingly important class of drugs. The complexity of the biological systems these compounds are produced in, together with conformational and other stability issues not present for typical small molecule drugs, make developing an efficient purification process for these molecules a challenging task. Liquid chromatography is one of the most versatile and commonly applied separation techniques in this field. It is a relatively well understood unit operation, and the phenomena occurring during its operation can be well described by mechanistic models. Increasing economic and regulatory constraints are driving the ambition to utilize this knowledge during process development **(Chapter 1)**.

One of the greatest challenges in introducing the use of mechanistic models into a process development workflow under real conditions is the lack of efficient technologies that allow determining all the necessary parameters that need to be provided to these models. Practical limitations dictate that these measurements cannot be carried out separately for every molecule present in the complex streams from which the product is to be isolated. An efficient way to determine the parameters in a multiplexed fashion is to regress them from observations made on the behaviour of the complex mixture. This step requires isolating the signal related to a single component from the complex observation. For chromatograms this can be achieved by least-squares fitting of a peak model to the observed chromatogram. When multiple components show similar behaviour in one separation dimension, the errors of such a regression become too large for the parameters to still be useful for process development. Too overcome the resolution limitations of a single chromatographic separation, multidimensional separations can be carried out. To facilitate the deconvolution of these multi-dimensional chromatograms an algorithm was developed that uses a

Fourier-transform of the chromatogram to generate the initial guess for the fitting procedure **(Chapter 2).** This technique is shown to reduce the errors of the fit by up to two orders of magnitude when compared to the single-dimensional analysis.

To go beyond optimization of a single unit operation, towards model-based development of a multi-stage downstream process requires building parameter database for all the impurities present in the original stream. A three-dimensional fractionation and characterization approach is introduced, that allows regressing isotherm parameters with small standard errors **(Chapter 3)**. Keeping the first and last separation dimensions constant provides an easy and practical mean to compare results from different resins and match parameters to pseudo-identities to allow building an interaction database without the necessity of sophisticated mass-spectrometry based identification of each impurity.

The number of potential chromatography resins that could be used for purification purposes has become so large, that testing even just a significant portion of them is no longer feasible without the use of high-throughput technology. Miniaturized chromatography columns that can be operated in a an automated high-throughput environment have been available for some years already, but technical differences between the used liquid handling systems and conventional liquid chromatography systems require adaptions to the experimental protocols to ensure generation of data with a comparable quality. Implementation of a novel meniscus sensitive single well volume detection method is shown to reduce the experimental noise of the system, while adoption of a an isocratic operation mode is shown to allow the regression of isotherm parameters while working around the technical limitations introduced by the simplified pump systems of the commonly available liquid handling robots **(Chapter 4)**.

Even with high-throughput technologies available, it is desirable to reduce the experimental burden on process development departments. During drug development, the structure of the product is usually known. As the behaviour of a molecule is related to its structure, correlative models can be trained to predict their behaviour based on descriptors related to their structure, such as their surface property distributions. As proteins aren't rigid molecules these descriptors need to be robust to a certain degree of fluctuations. For this purpose a set of robust surface property distribution descriptors are introduced and demonstrated to allow the prediction of the retention time of model proteins in hydrophobic interaction chromatography **(Chapter 5)**.

Samenvatting

Biofarmaceutica, en met name therapeutische eiwitten, worden een steeds belangrijker klasse van geneesmiddelen. De complexiteit van de biologische systemen waarin deze verbindingen worden geproduceerd, tezamen met conformationele en stabiliteitsproblemen welke niet voor kleine moleculen gelden, maakt de ontwikkeling van een efficiënt zuiveringsproces voor deze grote moleculen een uitdagende taak. Vloeistofchromatografie is één van de meest veelzijdige en meest toegepaste scheidingstechnieken op dit gebied. Het is een betrekkelijk goed begrepen scheidingsproces en de verschijnselen die optreden tijdens de uitvoering kunnen goed worden beschreven door mechanistische modellen. Toenemende economische en regelgevende beperkingen sturen de ambitie om deze kennis tijdens de procesontwikkeling te gebruiken (**Hoofdstuk 1**).

Eén van de grootste uitdagingen bij het gebruik van mechanistische modellen in een procesontwikkeling "workflow" onder reële omstandigheden, is het gebrek aan efficiënte technieken om alle noodzakelijke parameters voor deze modellen te verkrijgen. Praktische beperkingen maken het onmogelijk om deze metingen apart voor elk molecuul aanwezig in de complexe, multi-component stromen van waaruit het product geïsoleerd wordt, uit te voeren. Een efficiënte manier om deze parameters te bepalen is via regressie uit multidimensionale experimenten aan deze complexe mengsels. Deze stap vereist het isoleren van een signaal van één component vanuit een complexe waarneming. Voor chromatogrammen kan dit bereikt worden door het fitten m.b.v. de kleinste kwadraten methode van een geschikt piekmodel uit het waargenomen chromatogram. Wanneer meerdere bijdragen in een scheidingdimensie een soortgelijk gedrag vertonen, worden de fouten van een dergelijke fitmethode te groot om de verkregen parameterinformatie nog steeds voor procesontwikkeling te gebruiken. Om de resolutie van een

chromatografische scheiding te vergroten, kan de scheiding in meerdere dimensies worden uitgevoerd. Om de deconvolutie van individuele overlappende pieken in deze meerdimensionale chromatogrammen te vergemakkelijken, is een algoritme ontwikkeld gebaseerd op Fourier-transformatie van het chromatogram voor de initiële schatting van de parameters in het nu multidimensionale piekmodel (**Hoofdstuk 2**). Deze techniek reduceert de fitonnauwkeurigheid met tot twee orden van grootte in vergelijking met een eendimensionale analyse.

Modelgebaseerde ontwikkeling van een meerstaps zuiveringsproces vereist de constructie van een parameter databank voor alle in het originele materiaal aanwezige onzuiverheden. Hiervoor wordt een driedimensionale fractionering en karakterisering aanpak geïntroduceerd, die het mogelijk maakt isothermparameters met slechts kleine standaard fouten te bepalen (**Hoofdstuk 3**). Het gebruiken van steeds dezelfde eerste en laatste scheidingsstap biedt een eenvoudige praktische aanpak om de resultaten van verschillende beschikbare chromatografische harsen te vergelijken en om parameters aan pseudo-componenten toe te wijzen voor de opbouw van deze parameter databank. Dit alles zonder de noodzaak voor gecompliceerde massaspectrometrische identificatie van elke onzuiverheid.

Het aantal beschikbare chromatografieharsen dat gebruikt kan worden voor zuiveringsdoeleinden is zo groot, dat het testen van zelfs maar een klein deel daarvan niet langer mogelijk is zonder het gebruik van "high-throughput" technologie. Geminiaturiseerde chromatografiekolommen, welke in een geautomatiseerde "high-throughput"-omgeving kunnen worden gebruikt, zijn reeds enkele jaren beschikbaar. De technische verschillen echter, tussen de gebruikte robot vloeistofhandelingssystemen en conventionele vloeistofchromatografie systemen vereisen aanpassingen aan de experimentele protocollen om de generatie van parametergegevens

met een vergelijkbare kwaliteit te waarborgen. De implementatie van een nieuwe meniscusgevoelige volumedetectiemethode voor afzonderlijke "wells" vermindert de experimentele fout van het robot vloeistofhandelingssysteem aanzienlijk. Dit maakt het uiteindelijk mogelijk om isocratische chromatogrammen te genereren, nodig voor eerder genoemde parameterbepaling, op robot vloeistofhandelingssystemen, welke normaal door technische beperkingen van vereenvoudigde pompsystemen van algemeen beschikbare "liquid handling" robots onmogelijk is. (**Hoofdstuk 4**).

Zelfs met "high-throughput" technologieën is het wenselijk om de experimentele belasting voor biofarmaceutische industriële procesontwikkeling afdelingen te verminderen. Een mogelijkheid daartoe is het gebruik van modellen. Bij het ontwikkeling van een productieproces voor geneesmiddelen is de structuur van het product meestal bekend. Daar de structuur van een molecuul betrekking heeft op zijn gedrag, zouden correlatieve modellen ontwikkeld kunnen worden om het gedrag op basis van zgn. structuurdescriptoren te voorspellen. Een voorbeeld zijn oppervlakte-eigenschapsverdelingen. Eiwitten zijn geen starre moleculen, en dus moeten deze descriptoren robuust zijn tegen kleine fluctuaties. Hiertoe worden een reeks robuuste oppervlakte-eigenschap descriptoren geïntroduceerd en wordt aangetoond dat deze de retentietijden van modeleiwitten in hydrofobe interactiechromatografie kunnen voorspellen (**Hoofdstuk 5**).

Acknowledgements

Finally I want to thank all the people that in some way or another have been a part of the story behind this little book. There have been so many points, at which I regretted my decision to go down this road, but considering all the great people I have met throughout these years and all the exciting places it has taken me, I must admit that it was worth it all.

First off, I want to thank Luuk for giving me a shot at this elaborate name extension scheme by letting me join the BPE group, back when it was still sailing under the BST flag. I really enjoyed our (rare) meetings, especially in the first years. Speaking in front of a hall full of people is infinitely easier when you know the most critical questions have already been asked.

Marcel, I cannot thank you enough for all the doors you have opened for me in the years we've been working together. I really enjoyed our scientific and strategic conversations over the years, and not just those in the white train with the red stripe, although they might have been my favourites. I sincerely hope we find a way to let that tradition live on in some form or another.

A big thank you to everyone who worked with me as part of the BE-Basic project: Hoon, Beckley, Stefano, Jörg, Tangir and Silvia, as well as Peter, Martijn, Michel, Xiaonan, Ruud, Emile and Han. Almost seems weird to no longer meet up for a quarterly sandwich feast. Both thanks and somewhat apologies go to all the BSc, MSc, and PDEng students that at some point got caught up in this: Pilar, Mark, Victor, Thomas, Emin, Jorike, Marieke, Eleni, Lutz and Diogo. You were great and I wish you all the best for your future careers.

I want to thank all the members of the BPE group I had the pleasure of working with. Special thanks to Stef and Max, the always

helpful backbones of the labs. A big thank you goes to all the real coffee breakers: Arjan, Susana, Camilo, Carlos, Maria, Adrie, Rob and Peter and all the others that joined or left throughout the years. There is no way I would have survived these years without that daily dose of insanity. Cheers to all the other great people who were part of the crazy Kluyver life, the Curry Wednesday Gang, Thursday Moofers, Symposium Survivors, Monday Throners, and Fryday Keldertje Alcoholics: Jelle, Deborah, Hugo, Joana, Nantia, Erik, Ema, Andy, Diederick, Ana, Mar, Daniel, Camilo, Angel, Marcelo, Christina, Aljoscha, Robin, Nick, Linda, Laura, Kawieta, Eline, Jenny, Sjaak, Rob, Jos, Hans, Nayyar and everyone I might have forgotten.

Thanks to Jon and Roderick and the extended world of Delft civil engineers for being the best people to have a drink with whenever I needed a break from the BT world. Same goes for Jovana, my favourite witch and nootjes consumer.

David – my penguin and infamous work wife. Thank you for suffering through my endless list of nicknames. You were truly always there for me and there is no way I could have done this without you.

Dickes Dankeschön auch an Frieder, treuer Pinguin, Wegbereiter und immer wiederkehrender Kollege. Egal in welcher Stadt sich unsere Wege kreuzen, es is immer ein riesen Spass - erst recht wenn es regnet.

Einen riesen Dank an Eva, für die vielen Jahre Engelsgeduld mit mir. Du gibst mir den Mut und die Kraft immer wieder in fremden Städten und Ländern neu anzufangen.

And finally the biggest 'Thank you' goes to my parents and family for 30 years of unconditional love and support. You are truly the best.

List of Publications

Journal Articles

Kröner, F.C., **Hanke, A.T.**, Nfor, B.K., Pinkse, M.W.H., Verhaert, P.D.E.M., Ottens, M., Hubbuch, J., "Analytical characterization of complex, biotechnological feedstocks by pH gradient ion exchange chromatography for purification process development", Journal of Chromatography A, (**2013**), 1311, 55-64

Hanke, A.T., Ottens, M. "Purifying biopharmaceuticals: knowledge-based chromatographic process development", Trends in Biotechnoly, 32 (**2014**), 210-220

Hanke, A.T., Verhaert, P.D.E.M., van der Wielen, L.A.M., van de Sandt, E.J.A.X., Eppink, M.H.M, Ottens, M. (**2015**), "Fourier transform assisted deconvolution of skewed peaks in complex multi-dimensional chromatograms", Journal of Chromatography A, (2015), 1394, 64-61

Rho, H.S., Yang, Y., **Hanke, A.T.**, Ottens, M., Terstappen L.W.M.M., Gardeniers, H. (**2016**), "Progammable v-type valve for cell and particle manipulation in microfluidic devices", Lab Chip, (2016), 16, 305-311

Hanke, A.T., Klijn, M.E., Verhaert, P.D.E.M., van der Wielen, L.A.M., van de Sandt, E.J.A.X., Eppink, M.H.M, Ottens, M. (**2015**), "Prediction of protein retention times in hydrophobic interaction chromatography by robust statistical characterization of their atomic-level surface properties", Biotechnology Progress, (2016), doi:10.1002/btpr.2219

Oral Presentations

Hanke, A.T., Kröner, F.C., Pinkse, M.W.H., Verhaert, P.D.E.M., van der Wielen, L.A.M., Hubbuch, J., Ottens, M. (**2012**),

"High-throughput crude feedstock profiling for model-based bioseparation process development", 8th International PhD Seminar on Chromatographic Separation Science, Freudenstadt-Lauterbad, Germany

Hanke, A.T., Kröner, F.C., Pinkse, M.W.H., Verhaert, P.D.E.M., van der Wielen, L.A.M., Hubbuch, J., Ottens, M. (**2012**), "High-Throughput Acquisition of Physicochemical Parameters from Crude Feedstocks for Model-based Protein Purification Process Development", 243rd ACS National Meeting, San Diego, CA, USA

Hanke, A.T., Kröner, F.C., Pinkse, M.W.H., Verhaert, P.D.E.M., van der Wielen, L.A.M., Hubbuch, J., Ottens, M. (**2012**), "High-throughput crude feedstock profiling for model-based bioseparation process development", 14th Netherlands Biotechnology Congress, Ede, The Netherlands

Hanke, A.T., Nfor, B.K., Kröner, F.C., Pinkse, M.W.H., Verhaert, P.D.E.M., van der Wielen, L.A.M., Hubbuch, J., Ottens, M. (**2012**), "High-throughput crude feedstock profiling for model-based bioseparation process development", 32nd International Symposium on the Separation of Proteins, Peptides and Polynucleotides, Istanbul, Turkey

Hanke, A.T., Ramirez Vazquez, M.d.P., Verhaert, P.D.E.M., van der Wielen, L.A.M., van de Sandt, E.J.A.X., Eppink, M.H.M, Hubbuch, J., Ottens, M. (**2013**), "High-throughput comprehensive multi-dimensional fractionation and characterization of CHO host cell proteins", 9th International PhD Seminar on Chromatographic Separation Science, Weggis, Switzerland

Hanke, A.T., Ramirez Vazquez, M.d.P., Verhaert, P.D.E.M., van der Wielen, L.A.M., van de Sandt, E.J.A.X., Eppink, M.H.M, Hubbuch, J., Ottens, M. (**2013**), "High-throughput crude

feedstock profiling: the characterization of CHO host cell proteins", 245th ACS National Meeting, New Orleans, LA, USA

Hanke, A.T., Ramirez Vazquez, M.d.P., Verhaert, P.D.E.M., van der Wielen, L.A.M., van de Sandt, E.J.A.X., Eppink, M.H.M, Hubbuch, J., Ottens, M. (**2013**), "High-throughput comprehensive multi-dimensional fractionation and characterization of CHO host cell protein", 2nd European Congress of Applied Biotechnology, Den Haag, The Netherlands

Hanke, A.T., Verhaert, P.D.E.M., van de Sandt, E.J.A.X., Eppink, M.H.M, Ottens, M. (**2014**), "Multiplexed high-throughput characterization of cell culture supernatants", 10th International PhD Seminar on Chromatographic Separation Science, Egmond and Zee, The Netherlands

Hanke, A.T., Verhaert, P.D.E.M., van de Sandt, E.J.A.X., Eppink, M.H.M, Ottens, M. (**2014**), "Multiplexed high-throughput characterization of cell culture supernatants", 247th ACS National Meeting, Dallas, TX, USA

Hanke, A.T., Verhaert, P.D.E.M., van der Wielen, L.A.M., van de Sandt, E.J.A.X., Eppink, M.H.M, Ottens, M. (**2014**), "Comprehensive high-throughput multi-dimensional liquid chromatography for the characterization of complex biological mixtures", 15th Netherlands Biotechnology Congress, Ede, The Netherlands

Hanke, A.T., Verhaert, P.D.E.M., van der Wielen, L.A.M., van de Sandt, E.J.A.X., Eppink, M.H.M, Ottens, M. (**2014**), "Comprehensive high-throughput multi-dimensional liquid chromatography for the characterization of complex biological mixtures", 10th European Symposium on Biochemical Engineering Sciences, Lille, France

Hanke, A.T., Verhaert, P.D.E.M., van der Wielen, L.A.M., van de Sandt, E.J.A.X., Eppink, M.H.M, Ottens, M. (**2014**), "Hybridized experimental and algorithmic analysis of complex chromatographic separations", 3rd International Conference on HTPD, Sienna, Italy

Hanke, A.T., Verhaert, P.D.E.M., van der Wielen, L.A.M., van de Sandt, E.J.A.X., Eppink, M.H.M, Ottens, M. (**2015**), "Robust characterization of the chromatographic behavior of complex biological feedstocks", 11th International PhD Seminar on Chromatographic Separation Science, Sorpesee, Germany

Pirrung, S.M., **Hanke, A.T.**, van der Wielen, L.A.M., Verhaert, P.D.E.M., van de Sandt, E.J.A.X., Eppink, M.H.M., Ottens, M. (**2015**), "A new Paradigm in Bio Purification Process Development", 11th International PhD Seminar on Chromatographic Separation Science, Sorpesee, Germany (Oral presentation)

Hanke, A.T., Verhaert, P.D.E.M., van der Wielen, L.A.M., van de Sandt, E.J.A.X., Eppink, M.H.M, Ottens, M. (**2015**), "Robust characterization of the chromatographic behavior of complex biological feedstocks", 249th ACS National Meeting, Denver, CO, USA

Poster Presentations

Hanke, A.T., Ramirez Vazquez, M.d.P., Koppejan, V.W., Verhaert, P.D.E.M., van der Wielen, L.A.M., van de Sandt, E.J.A.X., Eppink, M.H.M, Hubbuch, J., Ottens, M. (**2013**), "High-throughput chromatography: Rapid characterization of protein binding dynamics", 2nd European Congress of Applied Biotechnology, Den Haag, The Netherlands (Poster presentation)

Pirrung, S.M., **Hanke, A.T.**, van der Wielen, L.A.M., Verhaert, P.D.E.M., van de Sandt, E.J.A.X., Eppink, M.H.M.,

Ottens, M. (**2014**), "A new Paradigm in Bio Purification Process Development", NBC 15, Ede, The Netherlands (Poster presentation)

Pirrung, S.M., **Hanke, A.T.**, van der Wielen, L.A.M., Verhaert, P.D.E.M., van de Sandt, E.J.A.X., Eppink, M.H.M., Ottens, M. (**2014**), "A new Paradigm in Bio Purification Process Development", 10th European Symposium on Biochemical Engineering Sciences and 6th International Forum on Industrial Bioprocesses, Lille, France (Poster presentation)

Pirrung, S.M., **Hanke, A.T.**, van der Wielen, L.A.M., Verhaert, P.D.E.M., van de Sandt, E.J.A.X., Eppink, M.H.M., Ottens, M. (**2015**), "Model-based biopurification process development", 249th ACS National Meeting, Denver, USA (Poster presentation)

Curriculum vitae

Alexander Thomas Hanke was born on May 28th, 1986 in Heidelberg, Germany. He graduated with honours and with special recognition for his achievements in chemistry from the Gymnasium in Walldorf, Germany in 2005. That same year he started studying Life Science Engineering at the Karlsruhe Institute of Technology, where he majored in Mechanical Process Engineering and Separation of Biomolecules.

He first came to the Netherlands in 2008, to optimize the fermentation conditions for the production of single domain antibody fragments towards more favourable downstream processing properties of the resulting feed stream, in the context of an industrial internship at BAC, now part of Thermo Fisher Scientific. Back in Germany he studied analytical technologies for the characterization of PEGylated proteins in the lab of Prof. J. Hubbuch.

As part of collaboration between the Technical University of Delft and the Karlsruhe Institute of Technology, he came to Delft to work on the 'Characterization of Sf9/Baculovirus feedstocks on purification relevant parameters'. This work led to a PhD position in the Bioprocess Engineering Group under Dr. ir. Marcel Ottens and Prof. Dr. ir. Luuk van der Wielen, from September 2011a to September 2015.

In October 2015 he started working as a Principal Scientist in Downstream Processing at Novartis Biologics in Basel, Switzerland.

ISBN: 978-94-6186-544-1