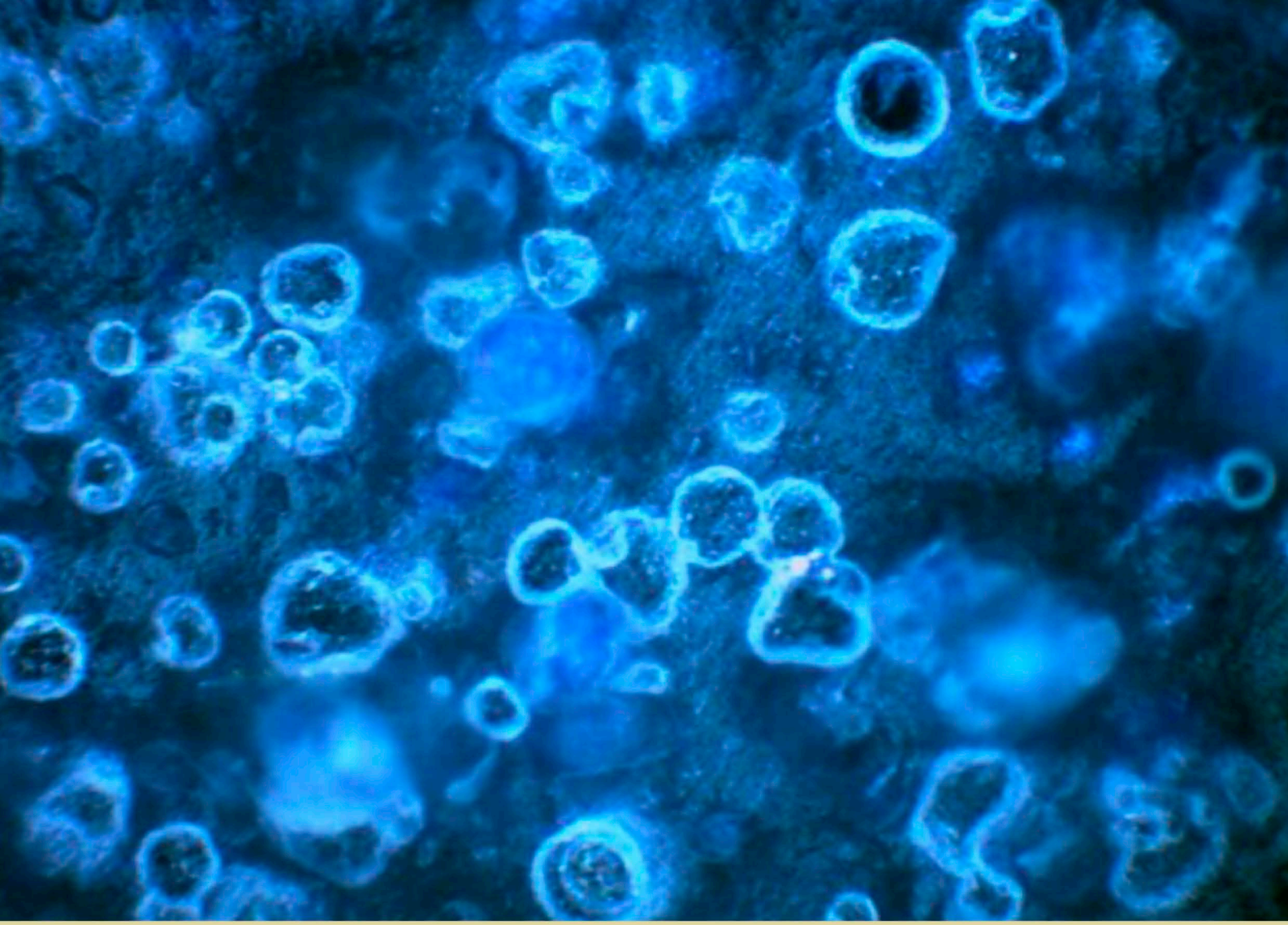

Ricardo Jorge Ferreira Margarido

Data-Driven Process Improvement in the Pharmaceutical Industry

Thesis submitted to the Faculty of Sciences and Technology of the University of Coimbra for the degree of Master in Biomedical Engineering with specialization in Bioinformatics, supervised by Prof. Dr. Marco Seabra dos Reis

July 2018

· U C ·

Universidade de Coimbra

• U C •

FCTUC FACULDADE DE CIÊNCIAS
E TECNOLOGIA
UNIVERSIDADE DE COIMBRA

Ricardo Jorge Ferreira Margarido

Data-driven process improvement in the pharmaceutical industry

Thesis submitted to the
University of Coimbra for the degree of
Master in Biomedical Engineering

Supervisors:
Prof. Dr. Marco Seabra dos Reis
Dr. Cláudia Sousa Silva

Coimbra, 2018

This work was developped in collaboration with:

Bluepharma Indústria Farmacêutica S.A.

Resumo

O uso de abordagens que exploram o potencial de informação contida nos dados para melhoria de processos na indústria farmacêutica é uma prática ainda pouco comum. Dada a disponibilidade de um grande volume de dados por processar provenientes de formulações e condições de fabrico de produtos já existentes, existe hoje em dia a possibilidade de usar toda esta informação para melhorar os processos associados.

Neste trabalho procura-se avançar neste sentido, no âmbito do desenvolvimento de uma melhor formulação para a libertação de fármacos através de uma matriz polimérica, nomeadamente no que diz respeito à estabilidade dessa mesma matriz ao longo do tempo quando armazenada. Todo o trabalho foi desenvolvido na Bluepharma Indústria Farmacêutica S.A.

As atividades começaram pela recolha de dados, sua integração e a limpeza. De seguida foi feita uma análise exploratória dos mesmos para que se averiguasse a sua qualidade, identificassem padrões dominantes e analisasse a influência que certos parâmetros poderiam ter no resultado final.

Identificadas as variáveis mais críticas, procedeu-se à construção de modelos empíricos que potenciem a melhor explicação dos dados utilizados e os seus resultados. Em particular, foram exploradas metodologias baseadas em variáveis latentes, como a regressão dos componentes principais (PCR) e mínimos quadrados parciais (PLS).

Por último, depois de obtido um modelo bem ajustado, procedeu-se á sua inversão de maneira a obter as condições de operação que, quando aplicadas, possam resultar numa melhoria na qualidade do produto a ser desenvolvido.

São também dadas recomendações de como estruturar, integrar e analisar os dados recolhidos para que o uso destas abordagens se torne mais ágil. É também apresentada uma perspectiva relacionada com o futuro destas aplicações.

Palavras chave: Ciência de Dados, Libertação de Fármacos, Qualidade a partir do design, Filmes Orais, Espaço de Desenvolvimento

Abstract

The use of data driven approaches to improve processes in the pharmaceutical industry is still an underexplored opportunity. Given the availability of large amounts of unprocessed raw data from formulations and manufacturing conditions of a product, there is a possibility to improve these processes by taking the most out of it.

In the present work, this endeavour is pursued, by considering the case of improving the formulation for drug release through a polymeric matrix, namely the stability of this matrix throughout its storage time. All the activities were carried out at Bluepharma Indústria Farmacêutica S.A.

The work developed begins by collecting and cleaning all the available data. Once all data is prepared, it is explored, both visually and analytically, to extract any relevant patterns of variation, check their quality and find out which variables mostly influence the outcome of the product.

Once the most important features are identified, the process is modelled using latent variable frameworks, such as Principal Components Regression (PCR) and Partial Least Squares (PLS).

Finally, the model is inverted in order to provide a set of values for the used features that, when applied, can result in an improvement to the product's quality.

Guidelines regarding how to structure, integrate and analyse the collected data are also given, in order to implement this approaches more efficiently in the future. A quick perspective on the evolution of these approaches is also presented.

Keywords: Data Science, Drug Release, Quality by Design, Oral Films, Design Space.

Agradecimentos

Esta dissertação marca o culminar da minha passagem por Coimbra. Em primeiro lugar, e porque sem ele o projeto nunca existiria, expresso o meu profundo e sincero agradecimento ao Prof. Doutor Marco Seabra dos Reis. Quando pela primeira vez reuni com ele encontrava-me bastante perdido no que ao projeto diz respeito. A sua disponibilidade, ajuda e confiança foram sem dúvida um dos grandes fatores para o sucesso deste trabalho.

À Bluepharma Indústria Farmacêutica S.A. por tão rapidamente me abrirem as portas e aceitarem um projeto ambicioso como este. Em especial à Doutora Cláudia Sousa Silva pela celeridade e interesse demonstrados bem como à Doutora Branca Almeia Silva por todas as horas que despendeu comigo em dúvidas e análises de informação.

À minha família no geral por todo o apoio dado desde sempre. Em especial aos meus pais por serem as minhas pedras basilares. Apesar de não o dizer frequentemente sei que sabem a importância que sempre tiveram na minha vida e nas decisões que tomei. O vosso apoio em especial nestes últimos 5 anos, não só financeiro, mas sobretudo anímico foi indespensável. Todo este trabalho é um espelho da pessoa que me moldaram e incentivaram a ser. A vocês o meu maior obrigado.

Aos meus amigos, principalmente os que me conseguiram suportar continuamente nestes últimos 5 anos. Sei do quão difícil consigo ser e encontrei em vocês algo desconhecido até à altura: uma segunda família. A maneira como nos conhecemos e agrupámos não foi de todo convencional mas a vossa atitude descontraída e o peculiar sentido de humor foi uma grande ajuda principalmente nos tempos mais árduos.

À Erica, por todo o apoio e incentivo que me deu nesta reta final da minha vida académica. O fascínio pelas coisas mais simples e o valor que sempre depositou em mim permitiram-me encarar estes últimos tempos de uma outra maneira.

Em último lugar às minhas duas cadelas, a Runa e a Mila, por perceberem todos os momentos em que me sinto mais em baixo e simplesmente se sentarem ao meu lado. Mais que simples animais são para mim família.

A todos estes e a todos os que me possa ter esquecido, o meu sincero obrigado.

A Coimbra, um último FRA.

" *You can't stop the change, any more than you can stop the suns from setting.*"

Shmi Skywalker

Contents

List of Tables **xv**

List of Figures **xvii**

Glossary **xix**

1 Introduction **1**
1.1 Motivation . . . 1
1.2 Data Analysis in the pharmaceutical industry . . . 2
1.3 Goals . . . 3
1.4 Bluepharma Indústria Farmacêutica S.A . . . 4
1.5 Outline . . . 4

2 Background **5**
2.1 Oral Films . . . 5
2.2 Production of Oral films . . . 6
2.3 Applications, advantages and challenges . . . 7
2.4 BlueOS . . . 7

3 Data collection and experimental methods **9**
3.1 Manufacturing Data . . . 10
3.2 Descriptors . . . 11
3.2.1 Release Profile . . . 11
3.2.2 Stability . . . 12

4 Methodology **15**
4.1 Data Cleaning . . . 15
4.2 Exploratory Data Analysis . . . 16
4.3 Quality by Design . . . 18
4.3.1 Model Fitting . . . 19

4.4 Model Inversion 23

5 Dataset, Results and Discussion **25**
5.1 Collected Data 25
5.2 Response Calculation 31
5.2.1 Release Profile 31
5.2.2 Stability 35
5.3 Model Fitting 37

6 Conclusions **45**
6.1 Future Work 46

Appendices **49**
A Unused Data 51

Bibliography **59**

List of Tables

5.1 Spearman's correlation for both APIs and τ. 35
5.2 Spearman's correlation for both APIs and β. 37

List of Figures

3.1 An example of the pipeline involved in manufacturing a film 10
3.2 The quantification of the release profile (red dots) and the obtained curve (blue line) with 3.1 12

4.1 An anecdote about correlation not meaning causation, taken from `http://tylervigen.com/spurious-correlations` 18

5.1 Boxplot of some well explored features (V22, V26 and V25 for example) and some poorly explored ones (V21 and 23 as example) 26
5.2 Pearson's correlation matrix, absolute values 27
5.3 Spearman's correlation matrix, absolute values 28
5.4 Variable trends and constraints part 1, all scaled to range from 0 to 1 30
5.5 Variable trends and constraints part 2, all scaled to range from 0 to 1 30
5.6 A good fit of the release profile to the chosen equation 31
5.7 A poor fit of the release profile to the chosen equation 32
5.8 Boxplot with the τ values of each API and the reference τ as a blue circle 33
5.9 Boxplot with the τ error values of each API 33
5.10 Some features with τ as the y value 34
5.11 A linear regression to get β with 4 time points(months) 35
5.12 Boxplot with the β values of each API and the target β as a blue line 36
5.13 Some features with β as the y value and the red circle identifying the best trial 37
5.14 Variation of RMSE with the number of PLS dimensions for API1 for calibration and cross validation 38
5.15 Variation of RMSE score with the number of PCR dimensions for API1 for calibration and cross validation 38
5.16 Variation of RMSE score with the number of PLS dimensions for API2 for calibration and cross validation 39

5.17 Variation of RMSE score with the number of PCR dimensions for API2 for calibration and cross validation 39
5.18 Percentage of each variable explained for a PCR model with two latent variables 40
5.19 Variable trends part 1, scaled, with the proposed design space for API1 as a black dashed line and for API2 as a blue dashed line . . . 41
5.20 Variable trends part 2, scaled, with the proposed design space for API1 as a black dashed line and for API2 as a blue dashed line . . . 41

A.1 A schematic of how FTIR works, adapted from `www.thermofisher.com` 52
A.2 A schematic of how Raman works, adapted from `www.semrock.com` . 53
A.3 A schematic of how in-depth Raman works, adapted from `www.sigma-epd.com` 53
A.4 An example of microscope used for polarized microscopy adapted from `soft-matter.seas.harvard.edu` 54
A.5 An example of a provided Raman Spectroscopy plot 55
A.6 Guided user interface for the plot of in depth Raman that lets the user select the lower bound, upper bound and spacing between values respectively 56
A.7 An example of a plot of in depth Raman between Z=0 and Z=90 with a spacing of 10 56
A.8 An example of a provided FTIR plot 57
A.9 An example of a provided polarized microscopy image 58

Glossary

API	Active Pharmaceutical Ingredient.
BcF	Buccal Films.
FTIR	Fourier-transform infrared spectroscopy.
LV	Latent Variables.
ODF	Orodispersable Films.
PCR	Principal Components Regression.
PLSR	Partial Least Squares Regression.
QbD	Quality by Design.
R&D	Research and Development.
RMSE	Root Mean Square Error.
SF	Sublingual Film.
Tg	Glass transition temperature.

1

Introduction

1.1 Motivation

For a long time a large amount of data has been collected and stored, in different industries. The pharmaceutical industry is no exception to this rule resulting in the development and large adoption of data driven methods namely data science. With Moore's Law [1] advocating the doubling of the processing power by modern computers every two years (or the halving of the costs) it has became easier and more accessible to process this data.

Data science can be defined as a field dedicated to the extraction of information from existing data. It is an interdisciplinary field that, given raw information, processes it, gives it a defined structure to make it useful.

Data science connects statistics, computer science and knowledge about the domain. The main goal is to be able to discover new insights and apply them to solve a specific problem.

Therefore, and given the amount of data recorded and available, there is a great interest to use it to decrease R&D times, cost and also to achieve higher yields.

On the other hand, it also has the capability to reduce the cost for the final consumer making healthcare more available and affordable to everyone. Decreasing the time to market also ensures that the latest technology can be deployed in the least amount of time necessary, making it possible for novel techniques to be used much earlier than before. As a final note, by improving the control and quality of drugs available, the occurrences of faulty dosage or bad interactions are vastly decreased.

Therefore, there is an opportunity to explore the use of data science to boost the performance of a wide variety of industrial activities, including the Pharmaceutical industry. In particular to a sub field that deals with innovation in oral films, which is the scope of this thesis.

1.2 Data Analysis in the pharmaceutical industry

According to the International Pharmaceutical Federation, the pharmaceutical industry aims to guarantee patient safety by delivering a product that is safe and has been tested according to highest standards available. Following this premise it is extremely important to be able to innovate and improve the solutions available.

Innovation can come in two major categories: drug discovery and drug delivery. While the first focus on the research for new active pharmaceutical ingredients, APIs, the second is related to discover and improve the way drugs are administrated.

Drug delivery and controlled release are the main topics of interest nowadays subject to constant innovation and improvements whether by the development of new technologies and fabrication processes or by the given insights provided by data analysis.

The later is still in it's early stages of development and adoption by the community with specific groups dedicated to applying data science in the pharmaceutical industry appearing in the last couple years. The increasing adoption of these approaches are driven by the good results obtained.

The large amount of recorded data provides a good opportunity to improve this field and applying knowledge extracted from previous formulations and experiences to new drugs, processes and release methods. Not only by analysing success cases but most of the times by taking lessons from failed experiments and trials, that help to eliminate some costly expenses given their poor performance for a given problem.

In this particular case, the information used is collected from the production Sublingual films (referred from now on as SFs) that are a specific type of polymeric matrices used for drug delivery [2]. Their properties are a sum of many factors which include the components used in the fabrication process by themselves, the interactions with each other and also with the Active Pharmaceutical Ingredient (API).

The complexity of these interactions, and their hidden nature, result in an impos-

sibility to predict any type of outcome *a priori*. Adding to this problem the fact that different APIs have different interactions and can be more hydrophobic or hydrophilic making the prediction of an outcome even more unreliable because past experiments may not translate fully to new ones.

When observing results from the same formulation one can infer some differences in the final product as being the result of a simple change of one of the concentration of reagents or the conditions during the mixture stages.

1.3 Goals

Overall this case study takes unstructured raw data as an input and processes it in order to provide more information about the different stages of R&D. By being able to provide a clear structure and a pipeline to feed data into, enabling researchers to better understand the problem, the value of data driven approaches is shown.

Another goal is to be able to empower R&D with methods to integrate collected data in a standard format, aggregating it into a system able to analyse it. By nature all the involved systems and machinery during manufacturing provide heterogeneous data both in type and format so it is extremely important to translate into a template easier to understand.

Drug release must obey some specifications to be able to pass certification and clinical trials. The main focus of the analysis done is to find tools and predictors that are able to measure the outcome of a trial providing some insight to the factors involved and, at the same time, helping to achieve a desirable result.

It needs to be guaranteed that the product has a correct drug release but that this property does not change when it is stored waiting to be consumed. This stability is another topic to be analysed in order to increase the safety of the final product.

By analysing the data provided one must be able to see trends and also regions where the values can be adjusted without breaking the set boundaries. This set of conditions, also referred as design space, can be provided whenever a new drug or drug release method is in it's infant R&D phase. It is also important that after the formulation of a film there is a methodology to access it's quality and also how it will evolve over time.

1.4 Bluepharma Indústria Farmacêutica S.A

At this point the importance of Bluepharma for the success of this project has to be highlighted. Bluepharma is one of the biggest innovation groups in the pharmaceutical industry with a remarkable reputation built in the last seventeen years. It comprises eighteen companies with around 580 collaborators with delegations in 7 countries. Around 86% of the products manufactured are exported for more than 40 different regions in the world.

Bluepharma's activity goes through the entire value chain of the drug, from R&D to the market, imposing itself for the excellence of its production unit. Taking advantage of the company's location in Coimbra there is a very close relationship with the University and the investigation centres available. Due to this proximity with academia there was a very positive mindset from the start which propelled the study made.

1.5 Outline

During this document it will be explained all the steps taken to arrive at the end result. The main structure roughly follows this pattern.

1. Introduction - Explaining the problem, the motivation behind it and a brief review of the methodology used
2. Background - Extensive literary review on the problem addressed in this study
3. Data collection and experimental methods - Explanation of processes, data, where it comes from and how it is handled
4. Methodology - In depth explanation of every method used during the work
5. Dataset, Results and Discussion - A presentation of the obtained dataset, achieved results, their importance, how they relate to the problem in hand and their meaning
6. Conclusion - A summary of all the findings and their value and a brief note on what is expected for future Work - new strategies to adopt with the increasing amount of data

2

Background

As the goal of this case study is to bring data science and data driven approaches into the R&D of a new product it is important to have knowledge about the progress already made in this field. The following chapter aims to cover as much as possible the progress achieved in the scope of the oral administration of drugs.

The administration of drugs through the oral pathway is the most predominant one. Given its desirable features as safety, simplicity, cost-effectiveness and, most importantly, its preference by patients and ease of access makes this delivery method an attractive one although some challenges such as enzymatic degradation are present [3–5].

A lot of different products for this administration route have been developed and tested to market which include tablets, sprays and oral films [3, 6]. In the next sections an overview is provided for the latter type of products.

2.1 Oral Films

One of the forms of oral administration is oral films. These stamp-sized films are designed to disintegrate during their stay in the patient's mouth and their formulation is made specifically for this propose. This formulation includes polymers, plasticizers and stabilizers, with the first two being the most critical components of the mixture. [7, 8]

Polymer choice is a key factor since it controls the disintegration time, drug capacity and mechanical properties of the film with a combination of different polymers being able to achieve the desirable end product in this key aspects. [7–9]

Due to water retention and absorption properties hydrophilic polymers like cellulose derivatives, polyvinyl alcohol or polyethylene oxide are the most common in the oral

films currently on the market. [7, 9–11] Although these properties are beneficial for stability and delivery rate proposes they can also be a negative factor due to its loss in flexibility and drug stability. [11–13]

On the other hand plasticizers play a major role in counteracting the hardness or brittleness imposed by the polymers used. By decreasing the Tg (glass transition temperature) of the polymer it gives its chains more mobility and increases both the elasticity and plasticity of the resulting film.

These excipients need to be compatible not only with the polymers used but also with the drug substance used, other excipients while being non toxic [7, 9, 14, 15]. Polymer selection directly influences key characteristics such as disintegration time, drug loading, chemical and mechanical properties [7–9]. Using the terminology present at [8] one can divide oral films in two main branches by the site of absorption:

- Orodispersable Films (ODF), for gastrointestinal absorption
- Sublingual and Buccal Films (SF and BcF respectively), for oral absorption

2.2 Production of Oral films

Oral films are manufactured either by solvent casting or hot-melt extrusion with new techniques being published and surfacing. [7, 16]

Solvent casting is the predominant one and consists in preparing a liquid mixture, using water or organic solvents, containing the polymers that are part of the film, drug substances and needed excipients. After homogenization it is casted and dried in order to form films that are cut to specified dimensions. [9–11]

During this process several parameters can be monitored such as viscosity, drying temperatures, drying times, room humidity and the visual aspect of the mixture (mainly air bubbles and agglomerates). [5, 7, 9, 14]

These monitored variables influence the drug delivery rate and it's stability since a small change in either of them can result in an alteration of the final film properties.

2.3 Applications, advantages and challenges

Oral films are still in its early adoption stage being considered a niche market with only a couple dozen of products available [17, 18]. Oral films have limited drug loading meaning that drugs must be administered in low dosage and high potency, making it difficult to ensure the proper quantity is present. Another factor is related to taste since it influences the patient's willingness to in fact take the medication.

On the other hand oral films possess clear advantages as there is no need for water, have improved bio availability and better dose accuracy causing less side effects for the patient. [7, 8, 14, 17]

2.4 BlueOS

Due to the challenges in developing hydrophilic polymers for oral films [11, 13] a new technology is in need. This challenges are related to the properties of water absorption and retention of these polymers which is essential to provide flexibility but can also affect the stability of the drug substance and result in a sticky film. [11–13].

By using hydrophobic polymers the fast disintegration and release is still achievable while solving the problem of water absorption. To address this necessity BlueOS®, a proprietary technology from Bluepharma Indústria Farmacêutica S.A, was developed, which has the capability of a new drug delivery technology that also includes sweeteners, flavours and colourants, making it more appealing to the patient, besides the usual components found in these types of films (polymers, plasticizers and stabilizers).

3

Data collection and experimental methods

During batch formulation and preparation a good amount of variables are measured and collected. As previously stated the availability of this information was one of the main propellers of this work. Following the standard of the pharmaceutical sector a good collection of variables and results was already collected and stored.

It is of extreme importance to understand where all the data comes from and how it is collected so one can be sure there is no fundamental flaw in this step. It also helps by giving insight of what variables relates to, what should be the reference value for scaling and centering and possible acquisition errors or wrong values.

As all data was collected independently of this project there are some aspects that can be improved such as the variation of some variables in a direction that can better explore the search space.

The amount of missing data is probably the worst offender in this study because data collecting does not follow a standard procedure and some variables may not be collected at a given time point and are subject to the interpretation of who collected it.

To add to this some other features could have been extracted that can be useful to understand and model the behaviour better and the outputs for every formulation should have been measured and quantified.

Even with all these compromises the collected information is well documented and stored. Most variables were scraped from paper were some ambiguity was noticeable . Mixture percentages were passed on in spreadsheets. Release profiles and other *à posteriori* measured properties were also available in spreadsheets.

Overall and given the reduced amount of preparation for setting up this project, data

quality is reasonable and vast which helps with some of the shortcomings mentioned.

3.1 Manufacturing Data

All data collected is segmented in trials and their respective casts. A trial contains all the films made with a certain mixture. A cast is a part of a trial where some process conditions can be changed. So overall the data presented has n trials, each with m casts.

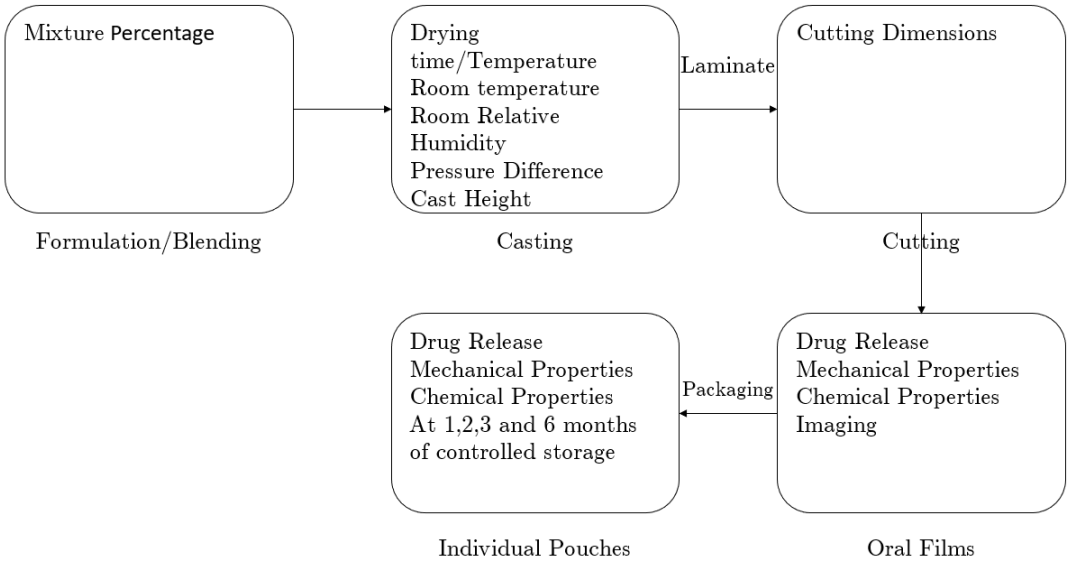

Figure 3.1: An example of the pipeline involved in manufacturing a film

The first step in making a new batch is measuring the amount of each component that goes into a liquid mixture once that the method use is solvent casting based. This can be done in grams/millilitres or as a percentage in mixture. The latter methodology is used to quantify each component individually.

Besides mixture components other features are measured some related to room conditions (temperature and relative humidity) to mixture drying (time and temperature) and dimensions of the film are some examples.

Collected variables can be split into two main categories. In the manufacturing data's case all of the pre-drying features are taken into account. This includes not only mixture related data but also working conditions of machinery and room.

This group is the most important to study and model since it is the only that allows adjusting conditions. Defining a good design space where a good release profile with

an increased stability is achievable is one of the goals of this project.

Manufacturing data is also the only group of data that can be in fact adjusted. Other features are all measured after the film is formed and dried, therefore being a result of mixture and working conditions. Although this results can't be adjusted they also give an important window to look into what may be happening at a molecular level.

To better understand how these different features are measured and collected 3.1 is of great help. Besides providing some variables extracted in each part it also gives a clear view of the overall process. Not included in the image but also important is that all working conditions are collected during the casting phase.

3.2 Descriptors

After the mixture is dried and cut into films another set of features can be collected. This include, but are not limited to, the aspect of said film (measures of homogeneity like bubbles and lumps), pH, purity, chemical and mechanical properties.

The main goals of this study (release profiles and their stability) are also measured at this stage. Along with the features mentioned in the paragraph above these are measured at specific points in time (0, 1, 2, 3 and 6 months after storage at controlled humidity and temperature). Some of these time points have been skipped giving a less than ideal insight of the evolution and ageing suffered by the film.

3.2.1 Release Profile

The release profile is measured by the cumulative percentage of API released at a given moment in time (measured in minutes). It is an indication of a slower or faster release resulting from the permeability of the film. In order to quantify the release profile of each trial a regression was made using 3.1. This formula was already previously used for a similar case study [2] and by looking at the available data it is a good match to the profiles.

$$PR(t) = 100 * (1 - e^{-\frac{t}{\tau}}) \tag{3.1}$$

By providing the values of t, the time points, in minutes, and the percentage released, PR, in them a new parameter (τ) is calculated. This is done to every trial and during

the different months the release is measured.

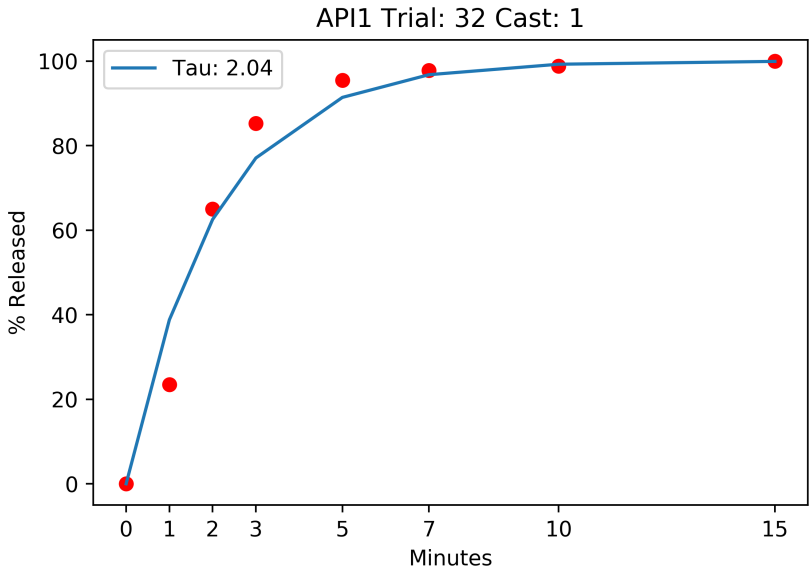

Figure 3.2: The quantification of the release profile (red dots) and the obtained curve (blue line) with 3.1

The obtained τ value is an indicator of the release profile. A smaller value means a faster release rate of the API and the larger the value the slower the release.

It is also important to notice that as is common in every regression there is an associated error that can be bigger if the measurements don't follow the formulated curve. An effort must be made to find a good function that can adjust well to a large number of instances to minimize errors.

Given an already approved and on the market drug of the same type its' release profile is used as reference and the τ value calculated. Converging to a result similar to this one, at the initial time point, is the aim of this part of the work.

3.2.2 Stability

Stability assessment is the other goal of this case study. A lot more importance was given to it since release profile at the initial time point is easier to adjust by increasing/decreasing film dimensions and height, if needed.

Stability translates the ageing and deterioration of films during its storage at controlled conditions (25°C and 60 % relative humidity). It is a measurement of the evolution of the release profile over time (measured in months).

The stability parameter evaluated, β, is the slope obtained with the values of τ at different months (t), using 3.2.

$$\tau(t) = \beta * t + k \tag{3.2}$$

Given the error while calculating τ one must be careful because it might skew β results. To add to this error there is also the inner error from making a linear regression and also the factor of some trials only having two months (0 and another one) as time points therefore prone to portraying a wrong stability. With more data these cases should be excluded but given the small amount of β available they were included in this study.

The best case scenario would be a β of 0 indicating perfect stability and that the release profile is the same throughout product's storage time. Getting as close as possible to this perfect stability is the most important aim of this analysis.

Descriptors are the second group of features. These range from bubbles and lumps to purities and mechanical properties. As said above none of these can be changed directly and are a result of the already made films.

However besides providing more information this group of variables can also be important for an pre diagnosis of film's evolution and can be helpful to discard or closely monitor some trials instead of others. These provide a different perspective to what might be happening at a micro and macro level in the matrix and second order interactions.

In this group not only numerical process data is available but also data surfacing from other sources such as imaging. Although all can be analysed together it makes the most sense to evaluate them individually given the different natures.

All release profiles analysed were collected beforehand and passed on in spreadsheets.

4

Methodology

All the work presented was developed with the help of Excel for building the dataset and then Python with several packages (numpy, pandas, sklearn and matplotlib) to analyse it as well as building the shown visualizations.

As stated all used data was previously collected and stored with its' origin discussed earlier. Due to the lack of structure displayed there is a need to clean it, explore it and see what conclusions can be derived from it.

The aim of this work is to be able to suggest a good design space comprising two different goals: on one hand a acceptable release profile and on the other a good stability of said release profile. Therefore caution is needed to see if one doesn't exclude the other.

4.1 Data Cleaning

As previously stated the large amount of data collected had some issues. The main one is indeed the presence of missing data (consequence of not having a clear method to register data in action). Some variables are not measured in some trials and time points making it harder to use them as features.

Many of this missing values follow a systematic pattern so it is not recommend or even a good practice to impute any values. Therefore most of the variables with a large amount of missing values are readily discarded, shrinking the number of variables available.

Ambiguous data is also another problem to tackle. Being in the form of a written report or conflicting values for the same variable in different spreadsheets these occurrences were quite common and were sorted out by looking at the raw data and talking with researchers.

Another issue is some variables are only collected in the moment right after the film is made and in no other time point after. Therefore although they might be complete they can not be used to see how they evolve over time.

Once all variables were collected, cleaned and processed an Excel file was built. This file can be seen as a fixed structure to be employed in the future as a standard for what types of data to collect and how to aggregate them.

4.2 Exploratory Data Analysis

It is important to notice that in the given context there is a sub set of problems to address and solve derived from having to deal with two APIs and two different goals, the release profile in itself and the stability of said profile throughout the time once the film is packed and stored.

There was a special attention given to data visualization during all the stages whether to inspect the behaviour of some variables or to interpret the results obtained.

As visual information is pre conscious it frees up the conscious part to problem solving while identifying patterns more quickly and understanding more complex systems [34].

Therefore in order to communicate easily with anyone exterior to the problem it is advised to provide the most attractive visualizations possible since they have a greater chance of being looked more closely and stimulating problem thinking. [35].

In this visualization some key aspects must be noted such as the utility of boxplots. Boxplots are extremely useful as a quick tool to inspect the distribution, skewness and presence of outliers of a feature.

Boxplots are also useful to compare two different groups and infer if there is any statistical difference between then to be tested after. Overall a lot of information can be condensed in a simple graphic.

Overall boxplots should be used for any form of scientific inquiry since the simplicity and elegance makes critical information to be quickly exposed. This makes comparisons easier and is vastly superior to histograms. [36, 37]

Summarizing, boxplot analysis allows a quick inspection of:

- Outliers

- Maximum
- Quartiles
- Median
- Minimum
- Skewness
- Group Difference

Plotting the release profile and its' regression helps to understand what kind of errors can be happening and if the curve provides a good overall fit to the data. It also shows if the release is faster or slower when compared side by side.

Analogously plotting the values of τ of a trial over the months they were measured and the regression made gives a good understanding not only of the error associated when estimating β but also of how does τ evolves with time.

Trends in variables can be observed in two ways. The first one is through correlation, both Pearson's and Spearman's. While Pearson's focuses on true values and linear relationships, Spearman's focuses on ranks and monotonic relationships (increasing or decreasing at a non constant rate). Therefore each variable has two correlation values for each of the others.

One of the main problems to tackle while dealing with this kind of data is the existence of a lot of correlation between the X variables. May it be in the form of some of the components of the liquid mixture being added as an already made composite or by influence of one variable in other. Therefore some measurement of correlation between variables is needed.

In [38] it is concluded that overall Spearman's is better than other methods of finding correlation in continuous non normal data and less sensitive to outliers [39]. Pearson's should only be used for normal data [39].

As a final note both [39, 40] advise for the fact that most of the time these statistics, and statistics overall, are poorly applied and interpreted and suggest some caution. Not over interpreting these values and be aware of the fact they give a generalized trend and not a dependency measure are two main takes.

As an anecdote in 4.1 one can see that caution is indeed advised when trying to infer some meaning from correlation. Two highly correlated variables may or not be influenced by each other or another hidden process.

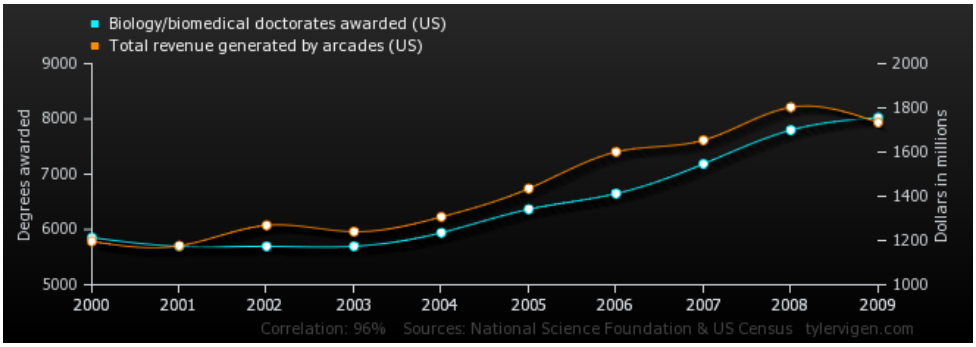

Figure 4.1: An anecdote about correlation not meaning causation, taken from `http://tylervigen.com/spurious-correlations`

This is especially true when trying to calculate the correlation between a variable and a response, being it τ or β, since it is not safe to assume that because a correlation is shown a causation effect is present.

Plotting each feature with the output as the y value (it being either τ or β) gives visual information of any correlation between a feature and the desired output. This correlation must then be accessed with the proper statistical methods.

As a way to evaluate the quality and how good these release profiles are, the obtained data is benchmarked against the reference product on the market to see if it is well adjusted or not.

4.3 Quality by Design

Analysing data, and, in this case, such an heterogeneous amount of data needs to be by different methods related to traditional and non traditional machine learning and statistics.

It is also absolutely necessary to research on the topic at hand first so one can inject prior knowledge into this exploration and know in which direction to move.

Another important aspect to research and retain is what has already been done in the context of this problem in order to solve it or, at least, improve it.

Data science, and in particular modelling, has been used before on drug release. Although not many studies can be found, proving the lack of adoption, their contributions are important [41–44].

These approaches can be included in a wider view: Quality by Design. Quality by

design (concept introduced by Joseph M. Juran in [45]) is, by definition, an approach to product development with predefined objectives [10].

The main aspects of this methodology are process understanding and controlling process conditions to improve performance [10]. By identifying critical variables and process phases to the target quality these can be closely monitored and adjusted. [46–48].

An important step in QbD is the design of experiments. These need to be done in a manner that yields the maximum amount of information. [49,50]. The interaction between variables and the influence of variables in the outputs should be investigated during this phase. [48–50]

Quality by design also allows one to establish a design space where the process can occur without flaws and result in a desired product. [46, 48, 51]. A design space is the combination and interaction of different input variables to provide a successful product quality wise. [52].

The methods involving Quality by Design are still in early stages of adoption and quickly being developed and improved [48,51,53]. They provide insight about product design and process [46–48] helping in defining a better design space [46, 51] to provide more robust methods with less failure rate, increasing efficiency coupled with a smaller development time. [48, 51, 53]

One particular case study is very similar [2] and serves as a guideline to the laid work. It is shown how by applied quality by design principles to the manufacturing process a new design space can be created within a margin of error for a specific output.

4.3.1 Model Fitting

After all feature engineering and selection is over and all variables have been inspected (visually and statistically) the next step is to fit a model to the selected data.

This model needs to be able not only to generalize the behaviour and influence of features into a final output but also to be inverted in a way that, given a desirable output, can provide a set of variables to achieve it (the correct design space).

No Free Lunch principle [54] advises against choosing a method blindly because given all possible problems the performance of one algorithm/method is equal to

any other. This can also be seen as the performance of a algorithm or method being case dependent.

Prior knowledge about the problem, what has been done in the topic and what results were presented is needed to make a good choice.

Given the nature of the data gathered and how strongly correlated some of the variables are, it is not wise to perform simple traditional regressions in order to obtain a model. Therefore a more robust method that is able to deal with many, noisy and collinear variables is needed [55].

These correlation and some constraints being imposed by the nature of the process requires the use of a model that can be fit not only to the variables but also their inner structure.

PLSR and PCR fill all this requisites, they are the simplest methods and also the most used ones in pharmaceutical industry . One can support this decision on the principle of Occam's Razor, stating that often the best solution for a given problem is also the simplest one.

These types of regressions are also proven to improve with the increasing number of variables relevant to the process and with observations, both a property of an industrial processes.

When using PLSR or PCR models it is assumed that the models are influenced by just a few variables, not apparent, called latent variables. Given this assumption it is clear that neither the measured variables nor the outputs are independent because they are the expression of these latent variables.

One does not know *à priori* how many LV will be present but these can be interpreted and it is extremely important to do so. These variables can be related to microscopic concepts (molecules and reactions) in chemistry and biology [55].

These can be seen as new variables resulting from combinations of the original ones given their linearity, correlation or similar structure. After fitting the model to the data, the inspection of explanation percentage conveys the information about variables that are important and variables that do not give crucial information to the model (large and small percentages, respectively).

This small explanation percentage may also be a result of a poor fit of the model to the data so no immediate conclusions should be taken. By analysing this percentage with what variables have a stronger correlation with the output one can see how good the performance of the model is.

PLSR can also be interpreted as a projection method similar to PCR, selecting the best projections of the variables that retain most of the information but are also more related to the output given.

This projection method can handle missing data in both the variables and the output. In [55] it is explained that for 20 variables in 20 observations PLSR can handle between 10 to 20 % missing data, if they are not systematic.

PLSR also provides an option to fit a model for an output at the time or for more. The general rule of thumb is that if the outputs are correlated one must analyse them together. If they measure different properties not dependent then a model should be made for each case.

This last case provides a model much easier to interpret and understand where the other can give more insight to the influence of variables in several outputs.

Deciding the number of dimensions to use is also a parameter to optimize in this type of regression. In general a model with a strong fit to the data is wanted but the trade off comes in the form of over fitting resulting in the loss of predictive power.

Over fitting can be prevented by cross validation and measured by evaluating R^2, Q^2 or RMSE that also gives an idea of the overall fit of the model. [55, 56]

Before introducing data into a regression model it's advised to pre-process it because these methods are sensitive to the scale of variables. If no knowledge exists on variables that can be more relevant to the problem it is advised to scale variables to unit variance (dividing by standard deviation) and center them (subtracting averages) in a process known as auto-scaling in machine learning.

Once there is a need to have a model that besides making use of latent variables is also able to be inverted both PLSR and PCR should be tested and the best model selected to invert.

For a good grasp and fitting of a model one needs:

1. Knowledge about the problem/domain, important variables and outputs
2. Analyse good data
3. Choose an appropriate amount of dimensions (too few - information is lost, too many - added complexity for no gain that can lead to over fitting)

4.4 Model Inversion

Finding windows of process operating conditions in which the design product exhibits certain, and desirable, characteristics is not an easy process and often can not be made just by guessing or by trial and error methodologies.

Therefore when data is available it is advised to use it as a diagnosis tool and if possible as a predictor for wanted outputs. This approach has revealed promising results in polymerization processes [56].

Inversion of latent variable models takes advantage of historical data from the manufacturing process (formulation, operating conditions such as temperatures, and also outputs) to be able to model and then suggest a design space where the product exhibits desirable characteristics and quality.

These models, although sometimes worse in performance, take into account the already existing constraints in the process and suggest a new design space with these in mind so it doesn't violate them and is still in a working environment.

It is also easier to explain and show results to someone outside the machine learning environment since the final result is a set of values for a number of variables. These results can also be interpreted by a specialist in the field to make sure they are indeed correct and some explanation can be obtained from them.

These constraints are available due to the property of LV models modelling not only the correlation of X with Y but also the structure of X and the inversion imposing the covariance structure of the past operating conditions [56].

For batch processes the recipe is analysed (the percentage of each of the products used in a mixture) and the operating profiles during the process such as temperatures, times and humidity. All the variables must be explored in a meaningful way to ensure the results are consistent. The model must then be inspected, taking into account R^2, Q^2 or RMSE, and see if the inversion is consistent with the constraints known.

Although the design space is supposed to be good for the output desired it must be considered as a starting point and not a finished set of conditions. New experiments should be designed around this design space in order to access the quality of the resulting outputs.

The obtained new set of conditions also need to be inspected in order to see what trends and values are presented. Trends conflicting with the pre observed ones might

indicate that earlier formulations were not guided in the right direction.

Given the nature of collected variables it is clear that none of them can admit negative values. If they are present in the new design space, once that these conditions are not explicitly imposed, serves as an indicator that they should be taken to their lowest value possible.

5

Dataset, Results and Discussion

5.1 Collected Data

As stated, after collecting all important data the first logic step is to visualize it. As stated before it is important to keep in mind that mixture variables range until V18 (included) and all the others are from manufacturing conditions.

From Fig. 5.1 we can see that some variables can contain interesting information to the problem while others clearly are not well explored (either by their nature or a less than appropriate trial design), meaning the collected values should be more spread in the available search space. Therefore these should be discarded for their lack of information.

All analysed variables come either from mixture percentages or manufacturing and equipment conditions. Therefore in Fig. 5.1 the values presented are the distribution of each variable in the several trials and respective casts.

As previously stated it is clear from Fig. 5.1 that some variables like V23, V24 or V29 only assume 2 or 3 different values and therefore can't fully express the variations in the output.

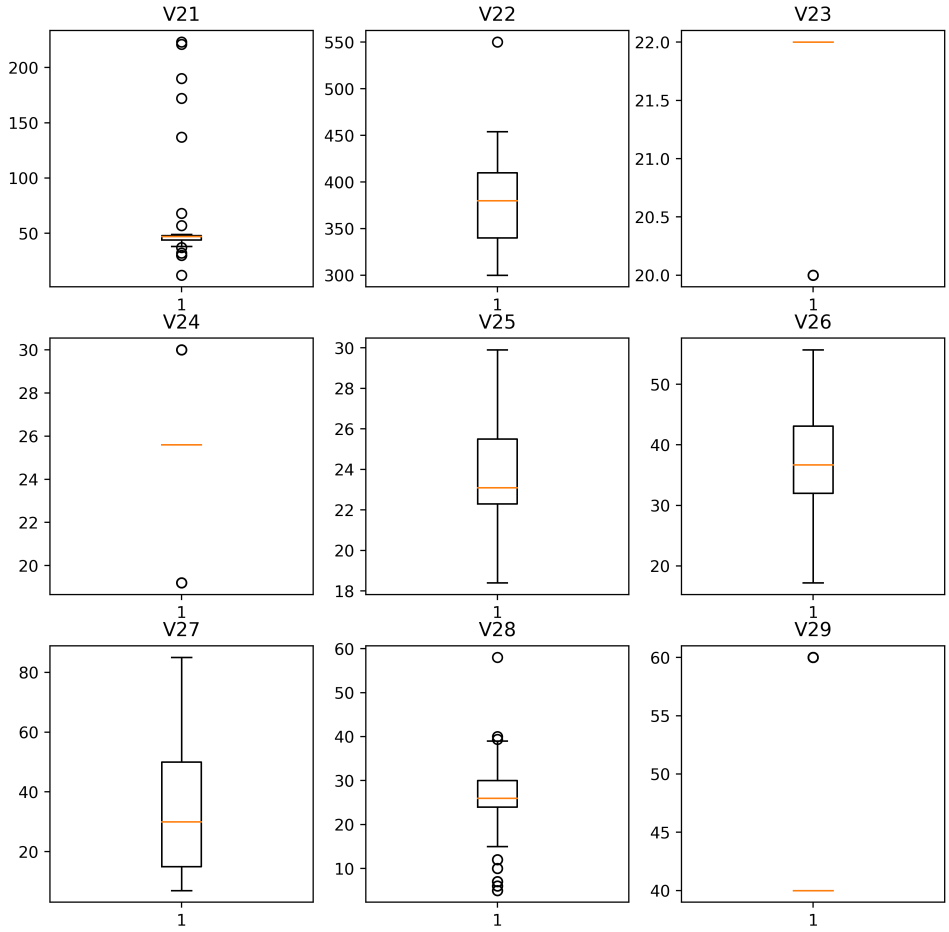

Figure 5.1: Boxplot of some well explored features (V22, V26 and V25 for example) and some poorly explored ones (V21 and 23 as example)

To see how closely correlated the dataset is a graphical representation of both Pearson's and Spearman's correlation matrix is presented making it easier to see which features can be grouped and are influenced by each other. New features can also be derived from these ones.

This correlations can be shown on Fig. 5.2 and Fig. 5.3 with the colour heat map also present to ease the visualization contrasting to the normal case where numbers are displayed.

From these matrices one can see that V3, V4 and V5 form a block and are very strongly correlated. Since these are features from mixture composition and the three of them come already pre-mixed (in a bundle) together, so this correlation is expected and obligatory (the absence would be indicative of problems while collecting data).

The values tend to decrease as the index of the variable increases since these vari-

ables are collected from manufacturing conditions and not mixture formulation and therefore are less rigid and more prone to be changed. Mixture formulation follows much stricter rules compared to casting and drying conditions. The latter are easier to adjust and should not be dependent of mixture percentages.

Overall and setting a threshold of 0.4 (moderate and strong correlations, yellow to red spots on the matrices) there are some other strongly correlated features that impose some constraints to the problem.

The absolute values are shown in the correlation matrices because the important information is the strength of said correlation and not it's direction.

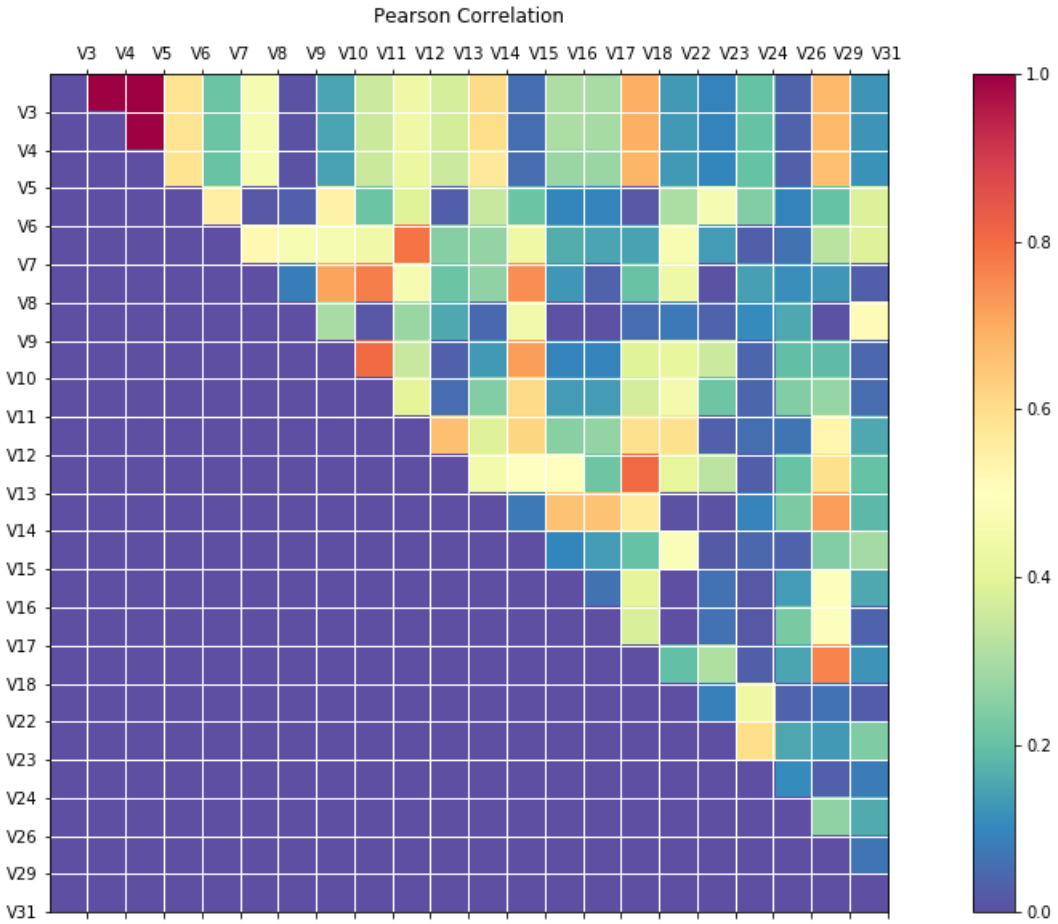

Figure 5.2: Pearson's correlation matrix, absolute values

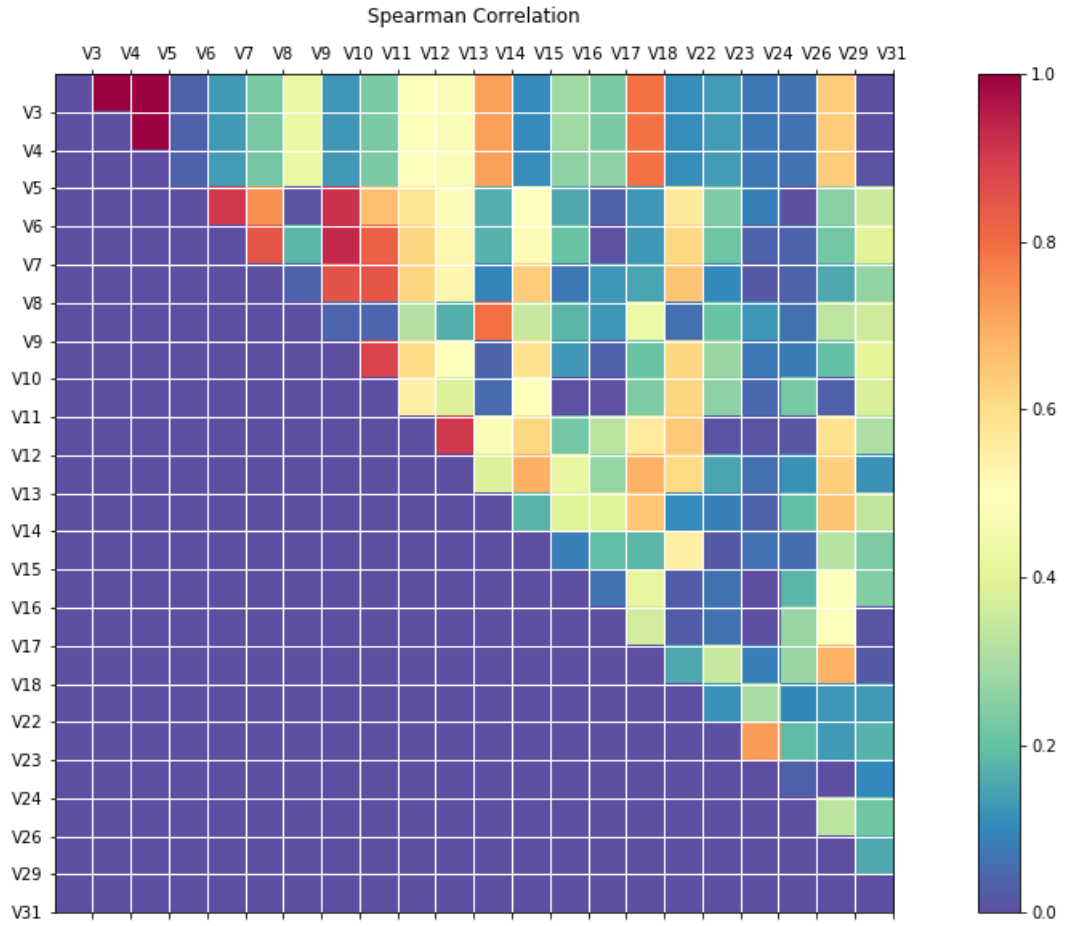

Figure 5.3: Spearman's correlation matrix, absolute values

In Fig. 5.4 and Fig. 5.5 one can see the trends each variable follows and that should serve as a constraint for the fitted models. V3, V4 and V5 show the correlation obtained in both Fig. 5.3 and Fig. 5.2.

As explained earlier the available data can be split into trials and casts (T and C in the graph, respectively) with the measured variables coming from mixture formulation and manufacturing conditions. Until V18 all variables are from mixture composition and all other are related to film manufacturing. All these variables are plotted with values between 0 and 1 (with 1 being the maximum value explored so far).

Although some cast can bee seen as outliers because they trend upwards from one feature to another when the majority transitions downwards and vice versa, both these figures make it clear that mixture formulation is stricter and more correlated than manufacturing conditions as previously seen in Fig. 5.3 and Fig. 5.2.

Overall manufacturing conditions have been better explored as seen with the heterogeneity starting from V24 onwards.

Once a new design space is generated it should be plotted to ensure that no constraint is being violated. These constraints are not only in regards to trend and correlation but also their maximum and minimum values. The models chosen will not include this minimum and maximum constraints so their results should be explained and analysed from a qualitative standpoint.

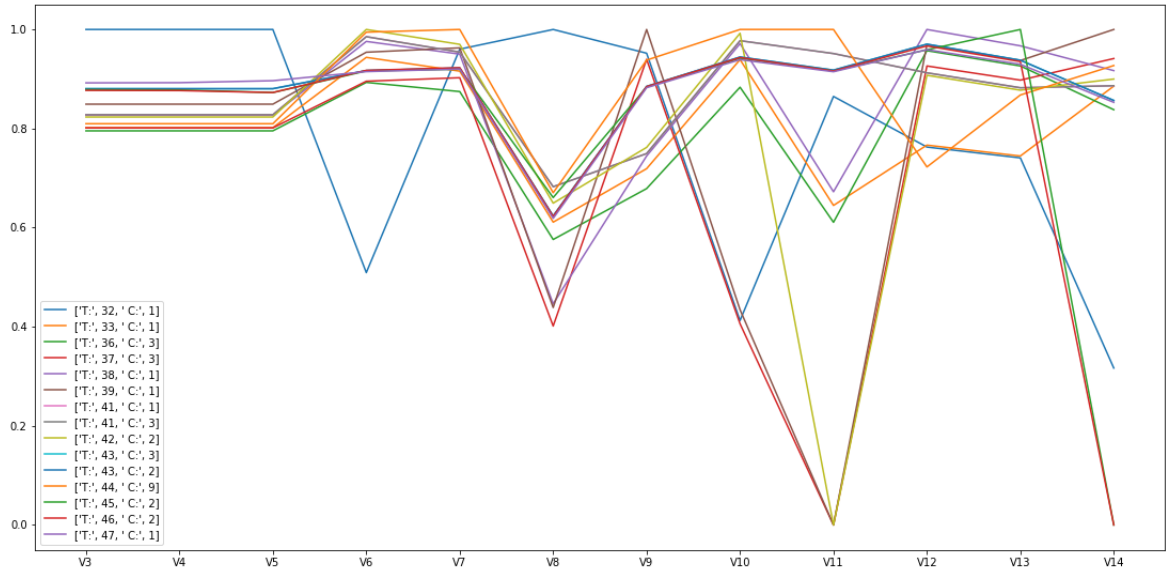

Figure 5.4: Variable trends and constraints part 1, all scaled to range from 0 to 1

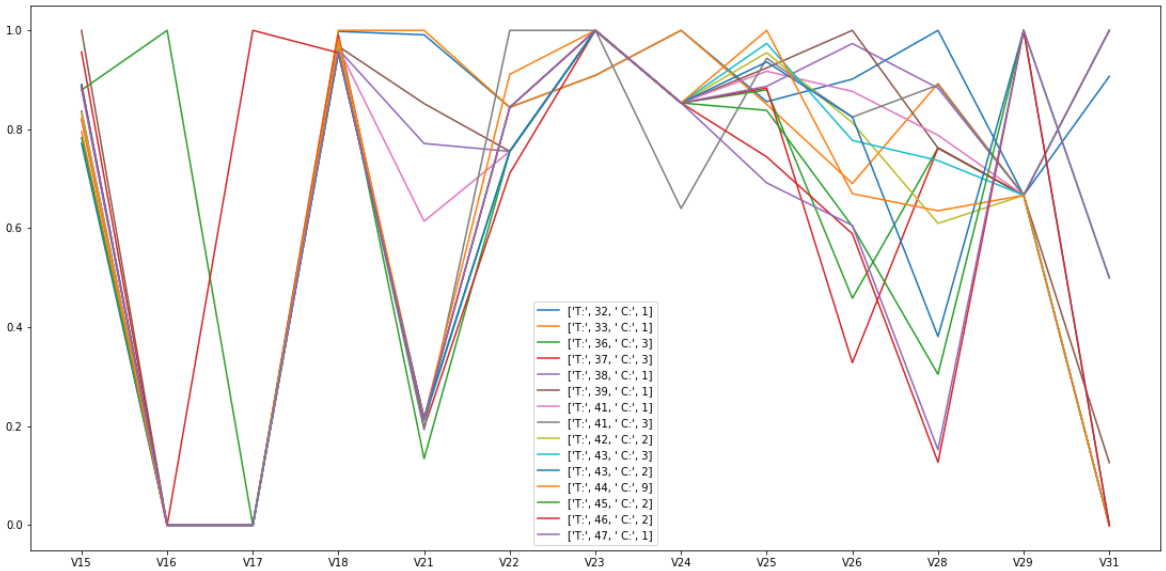

Figure 5.5: Variable trends and constraints part 2, all scaled to range from 0 to 1

5.2 Response Calculation

5.2.1 Release Profile

In Fig. 5.6 the good fit of the chosen equation to the points recorded when handling release profiles is shown. However in Fig. 5.7 a poor fit is displayed. Although some of the release profiles suffer from this poor fit this happens when the trial is already non viable and should be considered as an outlier.

Although the fit is not the best for these cases, given the good fit to the expected results and the fact that these trials have indeed a higher value of τ and a slower release the used equation is appropriate when looking at the associated error in Fig 5.9. This can be verified for both APIs.

A bad fit of τ to this line is also a good diagnostic tool of some problem in either the formulation or the process conditions resulting in a much less than adequate film.

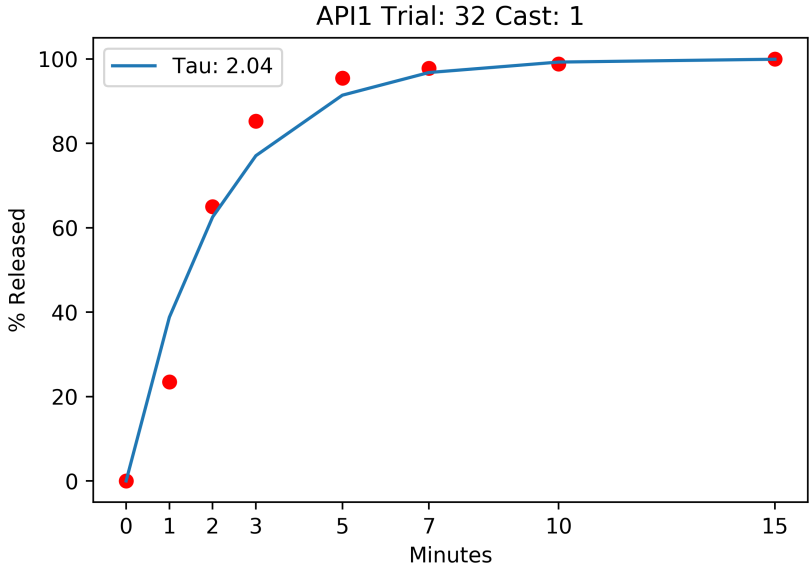

Figure 5.6: A good fit of the release profile to the chosen equation

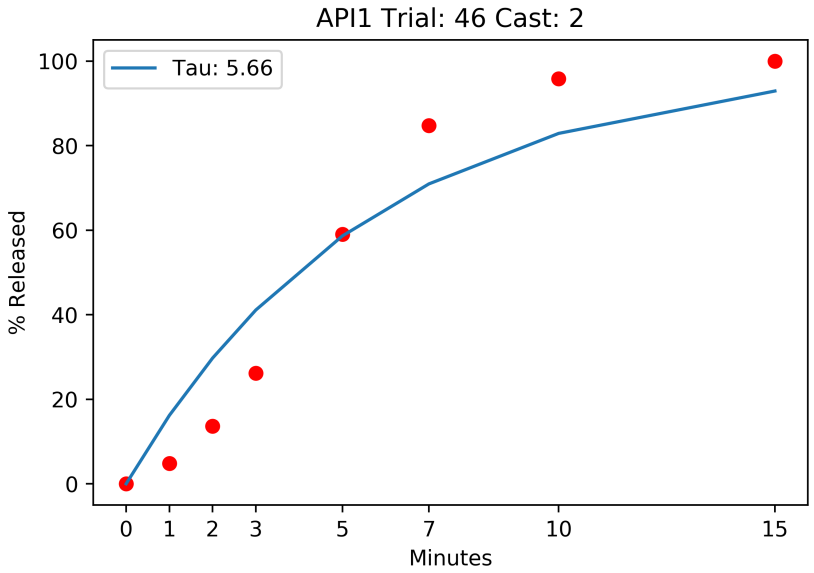

Figure 5.7: A poor fit of the release profile to the chosen equation

After all values of τ are calculated Fig. 5.8 is generated with the reference τ as a blue point. It is clear that throughout some of the trials the intended value of τ was either achieved or a close value was obtained. This means that some of the conditions experimented were able to achieve a release profile close to the reference one.

This was already known since τ adjustment is considered to be easy at least when compared to the stability problem. With the spread of the boxplot one can also see that the case of API1 seems harder to control then API2 which displays a much smaller spread of values and is more stable.

This can also be seen in Fig. 5.9 where the error of API2 when finding τ is much smaller than API1. This figure also complements what was said earlier about the errors being overall small (about 10%) indicating a good fit.

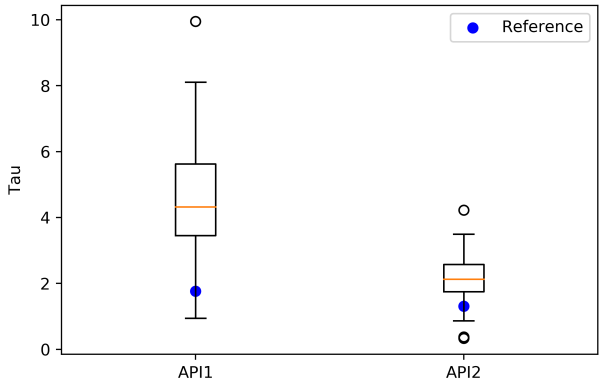

Figure 5.8: Boxplot with the τ values of each API and the reference τ as a blue circle

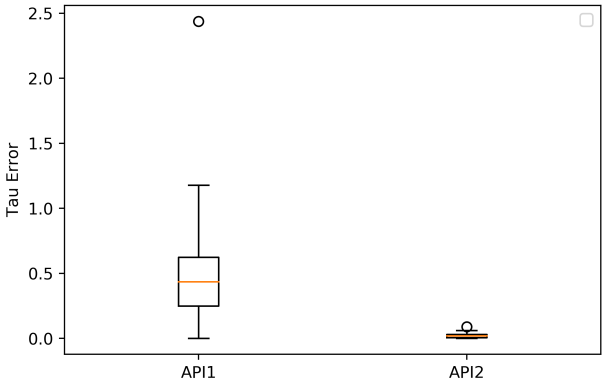

Figure 5.9: Boxplot with the τ error values of each API

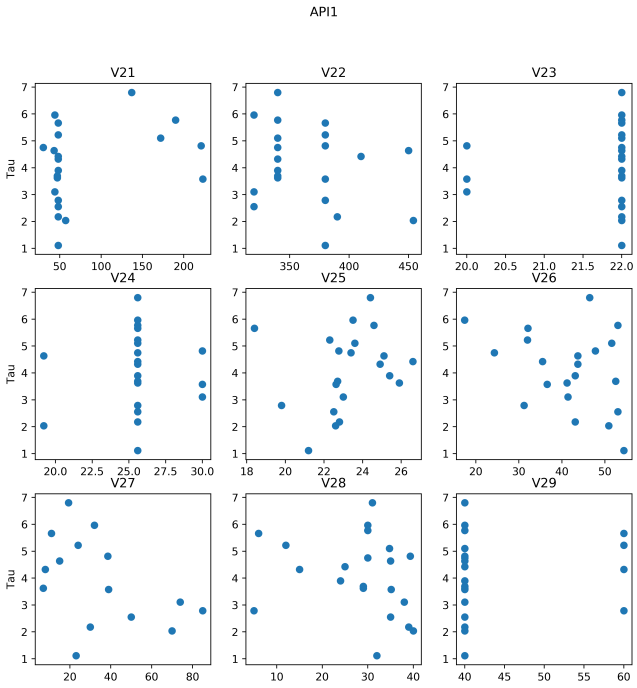

Figure 5.10: Some features with τ as the y value

Plotting variables with τ as y value and calculating Spearman's correlation is useful to see if some variable is more important to the problem. In table 5.1 a cut off of 0.3 was applied and only variables present in both APIs are displayed since the same film must contain API1 and API2 and the formulation is the same for both. This way the presented features are capable of adjusting both τ values at the same time. Although some weak to moderate correlations can be obtained none of them seem strong enough to fully explain the variation in τ values by itself.

The higher the value of correlation the greater the impact of a variable in the final output. Very weak correlations (from 0 to 0.3) should not be considered. Weak correlations (from 0.3 to 0.5) should be seen as small contributions to the final outcome. Moderate correlations (from 0.5 to 0.7) indicate bigger influence on the output. Strong correlations (from 0.7 to 1) mean that the variable has a big impact on the final outcome.

Table 5.1: Spearman's correlation for both APIs and τ.

τ	API1	API2
V3	0.512	0.514
V4	0.512	0.514
V5	0.504	0.509
V11	0.306	0.413
V18	-0.443	-0.392
V25	0.306	0.332
V28	-0.323	-0.349

5.2.2 Stability

After obtaining all τ values stability also needs to be calculated. As mentioned before, this is done using a simple linear regression where the slope (β) is taken. There are some cases where only two time points are present for the regression therefore the error is 0.

However this regression is worse then the one with 4 time points in Fig. 5.11 because even though the latter has a much bigger associated error it is more representative of the evolution of the film over time. With the increase of data one must move to exclude trials with only 2 months of τ to develop a more robust analysis.

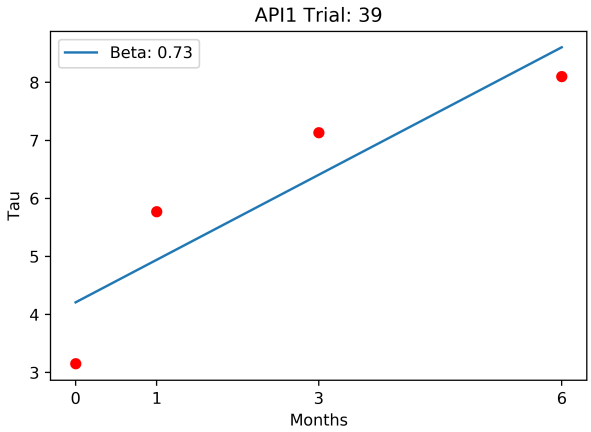

Figure 5.11: A linear regression to get β with 4 time points(months)

Similar to τ, once all β values are obtained the respective boxplots are created to provide an easier inspection. Here neither the reference value of 0 nor an approximated value is achieved. This was expected since the main issue was the lack of

stability of the films when stored.

However API2 presents much better β values overall because the API in itself is hydrophilic where API1 is hydrophobic. As storage conditions involve around 60% relative humidity API1's release profile might suffer more degradation due to its' hydrophobic nature resulting in matrix ageing having a bigger impact in this case.

By analysing the spreads of both boxplots one can see, once more, that API2 has a much smaller spread of values indicating a more stable problem than API1.

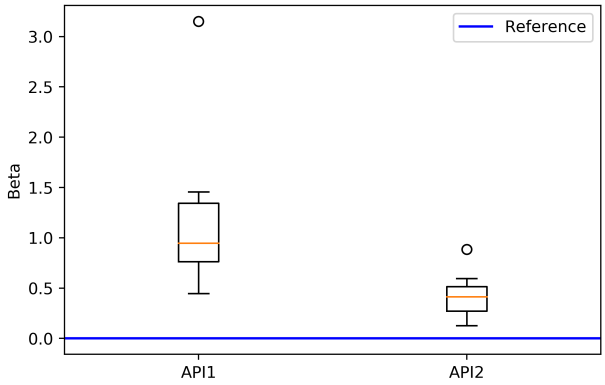

Figure 5.12: Boxplot with the β values of each API and the target β as a blue line

Following the same approach of τ once again all variables are plotted with β as the y value and Spearman's correlation calculated. in Fig. 5.13 a good trial is highlighted. This trial is the best value of β achieved for both APIs. The figure also helps identifying some correlation between any variable and β.

In table 5.2 a cut off of 0.3 was established and only variables present in both APIs are displayed.

Once again some weak correlations are found but not strong enough to be a solution to the problem at hand by themselves. These correlations are even lower when compared to τ values indicating an ever harder to solve problem with more hidden interactions happening.

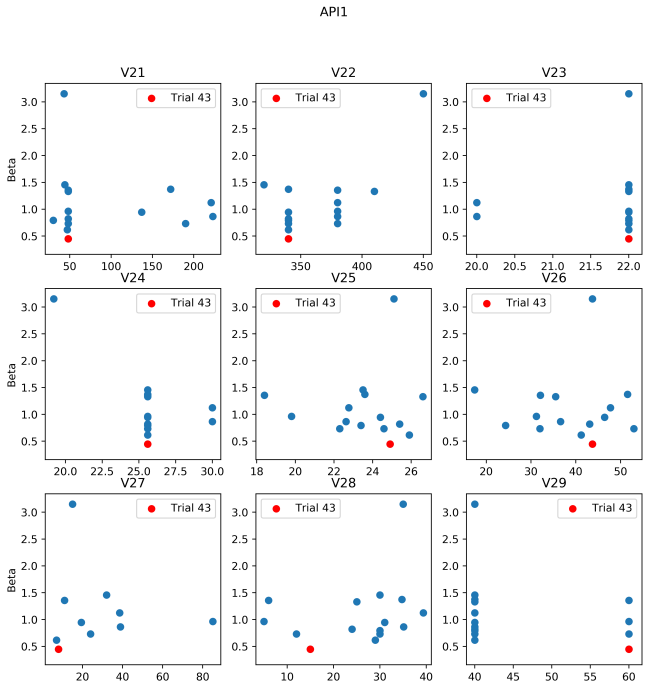

Figure 5.13: Some features with β as the y value and the red circle identifying the best trial

Table 5.2: Spearman's correlation for both APIs and β.

β	API1	API2
V12	-0.332	-0.406
V13	-0.372	-0.542
V18	0.300	0.513
V22	0.326	0.345
V28	0.323	0.588

5.3 Model Fitting

Choosing between PLSR or PCR models involves not only choosing the best performer but also the best type of model to the problem being handled. Both of them can be inverted, a property extremely important to be able to obtain a new design space.

Fitting PLSR and PCR models with just a few instances warrants some caution.

Although a good model is one without over fitting therefore having a good generalization this parameters are not easily measured when occurrences are scarce.

The typical approach of a 30-70% test-train split doesn't yield enough instances for either a good model or a good score calculation. Another choice must be made in the number of dimensions desired. Cross validation can be used but with caution.

For this second choice the best approach is to see how the percentage explained by the model or the RMSE evolves with the increasing number of dimensions and define a cut-off when adding a dimension doesn't translate in a meaningful increase in percentage explained or decrease in RMSE (similar to PCA).

The RMSE should be accessed in both calibration and cross validation (RMSEC and RMSECV) since the first evaluates the error in the training set and the other the error in cross validation. The error in calibration is expected to decrease with the number of dimensions used. In cross validation this can lead to a bigger error since the added dimensions can be related to noise or simply not contain any information relevant to the problem.

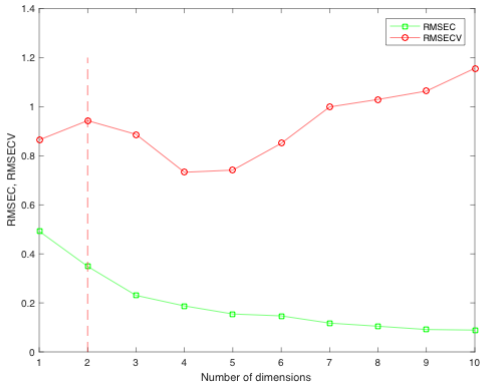

Figure 5.14: Variation of RMSE with the number of PLS dimensions for API1 for calibration and cross validation

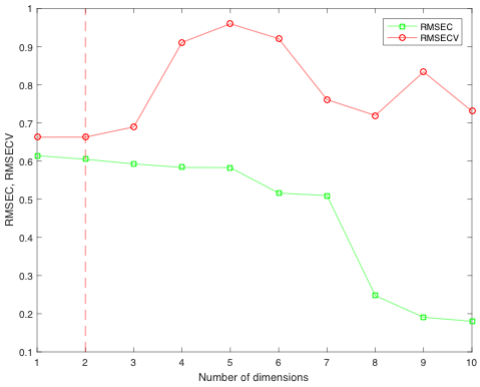

Figure 5.15: Variation of RMSE score with the number of PCR dimensions for API1 for calibration and cross validation

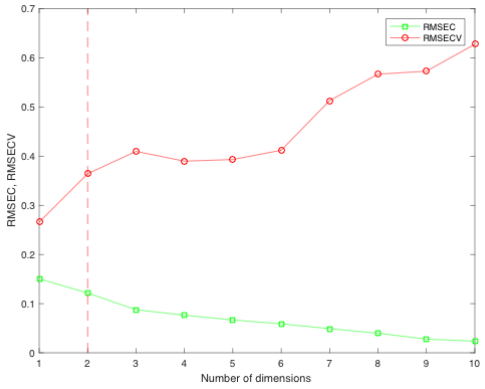

Figure 5.16: Variation of RMSE score with the number of PLS dimensions for API2 for calibration and cross validation

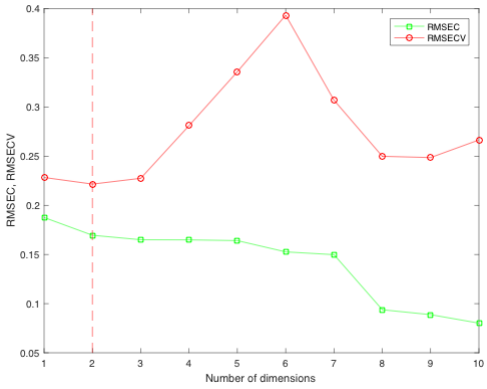

Figure 5.17: Variation of RMSE score with the number of PCR dimensions for API2 for calibration and cross validation

From Fig. 5.15 and Fig. 5.14 the better performance of PCR is evident. By having a lower value of RMSE in cross validation (the y scales are different) and needing only 2 dimensions to achieve this result as opposite to the 4 or 5 needed in PLSR for the least RMSE in cross validation.

The same result can be seen by looking at Fig. 5.17 and Fig. 5.16 (once again, the y scales are different) showing the strength of PCR for correlated and process derived data due to the existence of latent variables as was shown in literature. PCR is the best model to use in order to model the problem and find a new suitable design space.

By looking at the values of RMSE for both the models it is also noticeable the difference in API1 and API2 with the later having a much lower value of RMSE therefore having a better fit to the model, corroborating the assumption of being an easier problem to tackle than API1.

To choose how many dimensions to use the rule of thumb is to keep including dimensions until the increase in complexity doesn't yield an improvement in RMSE. By examining Fig. 5.15 and Fig. 5.17 a cut off at 2 dimensions seems appropriate and yields a fairly simple model. PCR was chosen as the model to use and invert not only because of this factors but also because of its intrinsic simplicity (making it easier to explain to anyone not so familiar with machine learning).

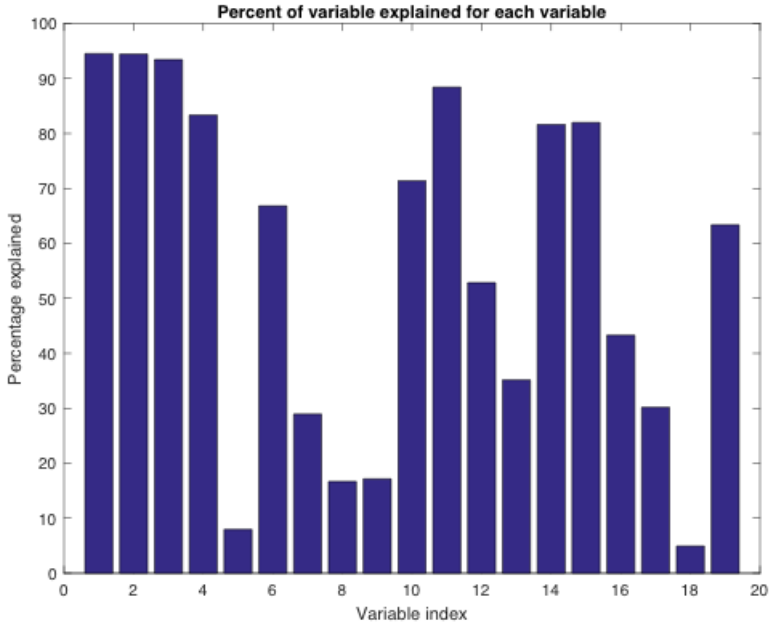

Figure 5.18: Percentage of each variable explained for a PCR model with two latent variables

Another output of the functions used to fit this model is Fig. 5.18, a graph bar showing how much of a variable is explained in the used model. Although overall most variables have a good percentage others are clearly under explained.

This can be a result of a variable not being important for the model fitting and therefore not being related to the output or a bad fit of the model, ignoring a otherwise important feature. This results need to benchmark to the previous obtained ones to see which variables are being excluded and if this exclusion is expected.

So the indexes of variables explained less than 50% are 5,7,8,9,13,16,17 and 18 translating into V7,V9,V10,V11,V15,V24,V25 and V26. By looking up at table 5.2 none of these variables appear, being a strong indicator of the lack of importance of these features to the model output.

On the other hand by analysing which features are over 80% their indexes are 1,2,3,4,11,14,15 corresponding to V3,V4,V5,V6,V13,V18 and V23. Once again and with the help of 5.2 V13 and V18 appear in both.

This was expected given their strong correlation mainly in API2. The other explained features might indicate a more complex interaction between them and the output and therefore not being detected when checking for correlation one by one.

After a good model, one that successfully explains the data and proves to not be over

fitted, is obtained the last step is to invert it so new values for all included features are generated. These values represent the design space for the desired output.

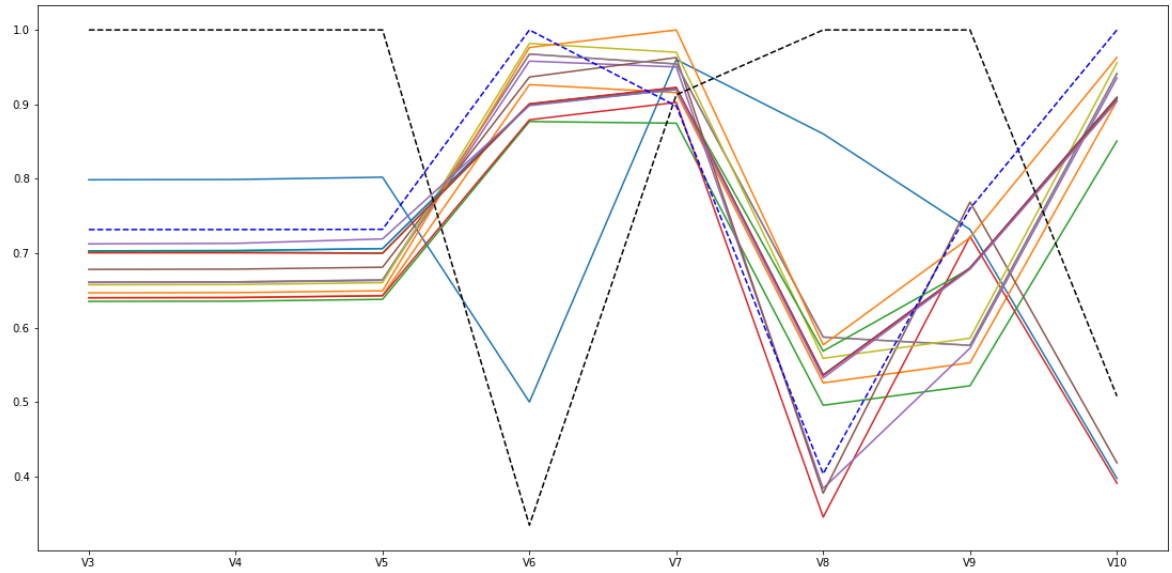

Figure 5.19: Variable trends part 1, scaled, with the proposed design space for API1 as a black dashed line and for API2 as a blue dashed line

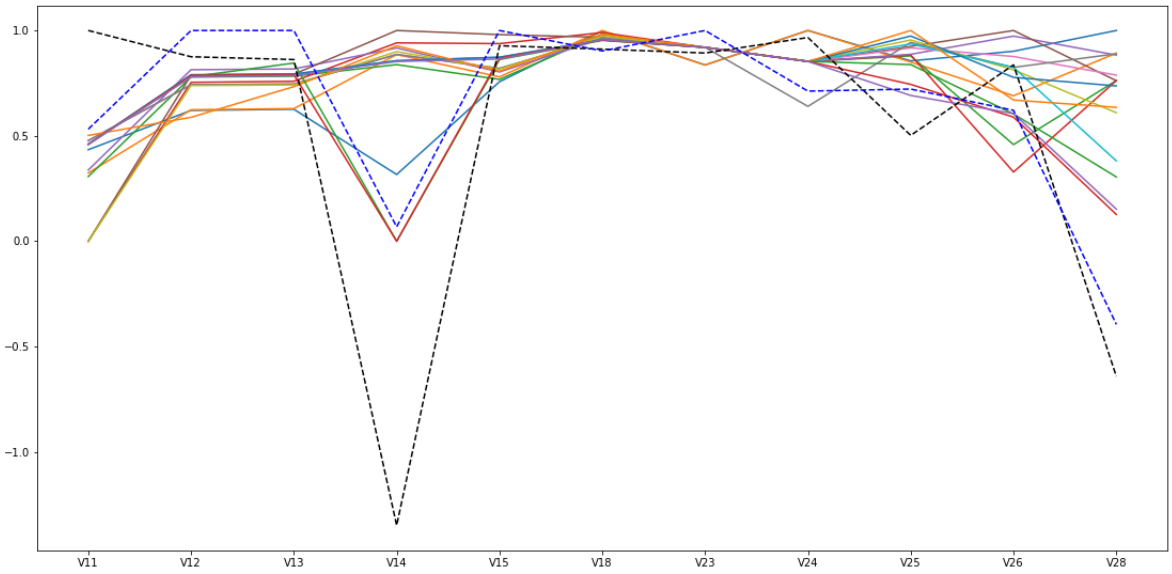

Figure 5.20: Variable trends part 2, scaled, with the proposed design space for API1 as a black dashed line and for API2 as a blue dashed line

As previously, plotting each cast with all variables between 0 and 1 and after plotting the 2 new design spaces obtained can be helpful in visualizing the goodness of these inversions. If the new sets of variables are within the bounds of previously explored values then they are feasible. If not the problems should be closely analysed.

Similar to the previously showed figures, the solid lines represent casts from collected data. The new dashed lines are the result of inverting the fitted models for API1 and API2.

At a first glance there are some problems with the proposed design space. The most obvious one is the presence of negative values. Given the nature of the variables collected (percentage in formulation, drying times, temperatures and so on) none of these admits a negative value.

These are a result of a worse than perfect fit of the model and must be interpreted as trends. Fixing this problem would require imposing limits to each variable (both lower and upper bound) to ensure the design space formulated was inside this limits.

However and given the distance from the desired β it was decided to not impose this limits since the model should only be constricted when near a acceptable value of β.

Therefore in variables that assume a negative value (V14 for API1 and V28 for both APIs) the more conservative approach is to take their minimum possible values (evidenced by V14 in API2).

The variables found to be important by analysing correlations with β and the percentage explained on the model (V13 and V18) are proposed in the new design space with values according to their previous correlation found in table 5.2. Although the correlations were weak there is still some correlation that is also present in the fitted models.

For V13 the negative value of correlation with β is a result of the increase in this specific variable being observed in a decrease of β value. This assumption is corroborated by Fig. 5.20 as V13 appears to have a new value, higher than any previously observed value for both of the cases.

As for V18 the value of correlation is positive. This means that an increase in this variable also translates into an increase in β values. As both the new design spaces appear to have a much lower value than any previously observed this evidence is also corroborated.

Once again, taking into account not only trends but also values, API2 seems to be much easier to model and adjust then API1. Values for API2 fall in line with already explored ones with some deviation but point to the fact that as seen in Fig. 5.12 some trials are near the ideal one and so is the output.

As for API1 it is noticeable the need to stretch to the limit (and sometimes passing

it) some of the variables. In some areas (V6,V9,V10 and V11) the trend is opposite to the one followed so far (and suggested in API2). In these regions (both the limits and counter trending) caution is advised when formulating a new trial.

The best approach is indeed to experiment around these areas, taking the obtained results as a starting point and seeing if greatly increasing/decreasing some variables yields better results while also trying to approximate both design spaces to reach a good β for both APIs.

Given the lack of direction to guide the development of these oral films, these two new design spaces provide new guidance in order to which variables should be altered. Given the complexity of the problem and limitations in mixture and manufacturing the new values need to be taken as advice.

6

Conclusions

Given the availability of raw, untreated data already collected by the pharmaceutical industry the work proposed was to use this untapped resource to cut cost and times in the research, development and improvement of drugs and delivery methods.

To achieve this goal data collected by Bluepharma Indústria Farmacêutica S.A was paramount. This data contained not only mixture properties but also manufacturing conditions with the corresponding output. Cleaning the data was the first major challenge faced given some irregularities and contradicting reports.

The first main outcome of this thesis is applied here. A structured Excel file was provided with indications on how to collect and aggregate all collected data. Some recommendations on which variables to register and how to better explore them is also given. Overall this internal documentation aims to improve the research method to boost the presented data driven approaches.

Also provided was a tool to help with data visualization, mainly in depth Raman, allowing this type of data to be inspected in a much quicker manner.

To tackle the main issue faced, stability of a film while stored, variables were inspected one by one to check not only their importance but also how they were explored. This process showed that most variables could be better explored but also provided insight to their importance.

By analysing trends in the data collected, fitting models to data, inverting the model and benchmarking the final result with the trends already observed a set of operating conditions were carried out in which the stability point of a film is supposed to be much closer to the desired one.

This design space was discussed suggested for a new trial although given the time it takes to measure the release profile in at least 3 different time points (at least 3 months) made it impossible to see the results of the suggested formulation before

the conclusion of this project. This new formulation also shows that the stability of API2 is much easier to tackle than API1 as it was shown during this work.

Although a lot of work and thought was put into defining a formulation suitable for both APIs the fact they are not fully compatible may lead to a region where the output is not as good as it should be.

It is also important to point out that some variables that can have a greater influence on the end result may not have been discovered yet and therefore are not being collected. This situation is normal given the difficulty involved.

Overall and given the complexity of the problem addressed the results are very satisfactory, not only by providing a new direction to make new formulations but also giving some novel insights about the interaction of some features with the output that were not apparent until now.

The new suggested formulations can drive the research and development process to a better final quality stability wise. The inclusion of data that was not used or analysed so far can change the perspective of researchers and provide new insights on the problem.

6.1 Future Work

Given the amount of information collected and it's heterogeneity some directions are proposed in what regards further studying this problematic.

If new trials and a new mindset in registering all data available it is expected that the dataset can grow fast in quantity and quality.

With more complete information on variables that are, at the moment, severely lacking values on numerous instances the models can have a better fit and therefore be more robust in predictions.

Another good pathway to access is related to the others types of data mentioned that were not analysed given its sparse collection. Almost all of these are related to methods of diagnostic and not predicting and as such can be seen as a tool to quickly discard unfeasible trials.

The best case scenario for predicting the outcome of a mixture would be to measure even more properties and check if any of them has a good correlation with the desired output.

Looking at FTIR, Raman or even polarized microscopy besides accessing the state of the film can provide even more knowledge about the problem and it's domain.

Coupled with the extrapolation of what can be happening to produce said graphs/results can be used to once more strengthen the models and provide a better design space.

Appendices

A

Unused Data

A.1 FTIR

Fourier-transform infra-red spectroscopy gets its' name from the need of using Fourier transform to transform the acquired data to an interpretable spectrum. It works by using a light beam with a large amount of frequencies of light and measuring how much is absorbed by the sample. The composition of the light beam is changed and absorbance is measured again.

This process is repeated until all the configurations are sampled and the resulting data (interferogram, deriving from Michelson interferometer) is processed by a computer using Fourier transforms and computing the values for each wavelength A.1. This results in a typical FTIR graphic in which the obtained values can be mapped to known molecules.

Although FTIR raw data is provided, the FTIR graphic is also given. The raw data structure is easy to read and analyse since it is a double column with the first being wavelength and the second the transmittance percentage.

FTIR usage yields good results in compound identification in a wide range of different areas. [19–23]

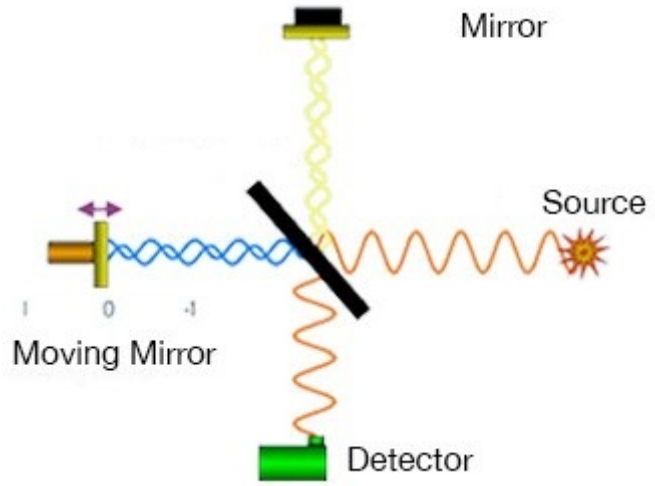

Figure A.1: A schematic of how FTIR works, adapted from `www.thermofisher.com`

A.2 Raman

Raman works on the principle of inelastic scattering of monochromatic light giving information about molecule vibrations as seen in Fig. A.2. This process produces mainly Rayleigh scattering with a very small amount $10^{-5}\%$ of the beam having increased or decreased energy. Raman scattering is the latter part of the scattering and similar to FTIR after being collected is plotted in a graphic for ease of analysis and matching molecules to Raman peaks.

Each Raman file provided follows the same structure with a double column, the first one being the wavelengths used and the second one the counts for each of the wavelengths making the visualization rather easy. (Same as FTIR)

Similar to FTIR, Raman has been widely used to identify compounds in different fields, proving it's worth in this task. [24–30]

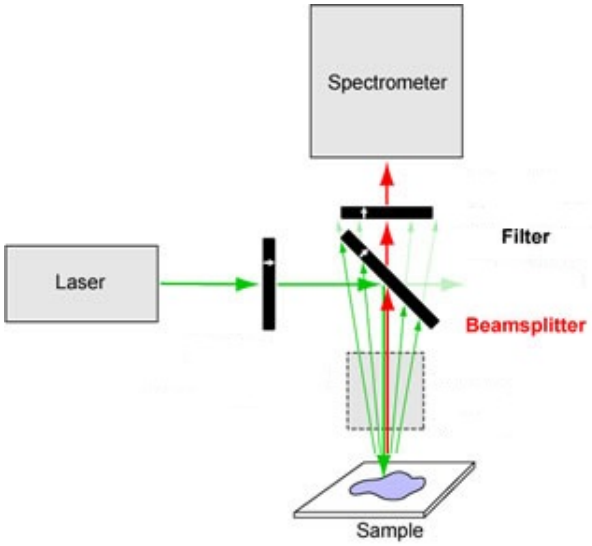

Figure A.2: A schematic of how Raman works, adapted from `www.semrock.com`

A.2.1 In-depth Raman

A particular case of Raman that was obtained is 3D Raman. Also known as confocal Raman it takes advantage of the film being transparent to the laser and is able to produce several Raman spectra at different height levels. This technique gives insight not only about the matrix surface but also the interior as it can be seen on Fig. A.3.

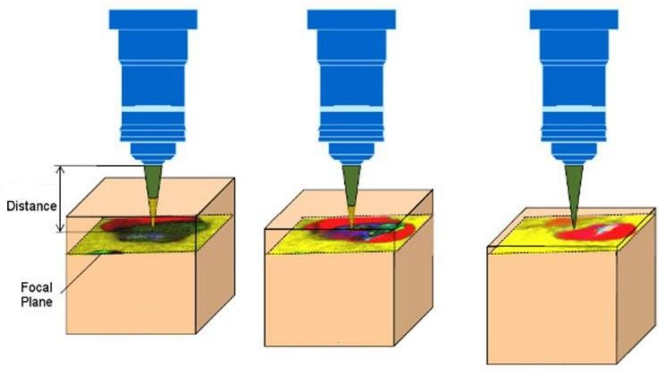

Figure A.3: A schematic of how in-depth Raman works, adapted from `www.sigma-epd.com`

The files handled corresponding to this Raman type all have the same layout with the first line being the wavelength used and each of the following lines starting with the depth value followed by the counts for each of the wavelengths. In order to ease visualization a small tool was made to plot each Raman at its corresponding depth.

A.3 FTIR vs Raman

FTIR and Raman should be viewed as complementary to each other allowing for a more precise identification and quantification of the compounds present in the film. In FTIR the molecule must undergo a change in dipole moment and in Raman the electron cloud must suffer positional change.

Therefore molecules that can't be subjected to dipole moment change (for example by being symmetrical) rely on Raman to be identified. FTIR is useful to evaluate functional groups and Raman is able to distinguish different levels of covalent bonding.

Numerous cases of success in identifying components and monitoring polymerization processes are documented. Shifts in Raman and FTIR and the presence of new peaks are a good indicator of changes either in process or film content [31, 32]. Some of these changes can also be indicative of changes in miscibility [33].

A.4 Polarized Microscopy

Polarized Microscopy uses polarized light to enhance contrast between light and dark areas of the films with a good resolution. By polarizing light these differences can be accentuated and more pronounced. It is mainly used to check the surface of a film and see if there is any crystalline structures that can be related to a change in the film's stability.

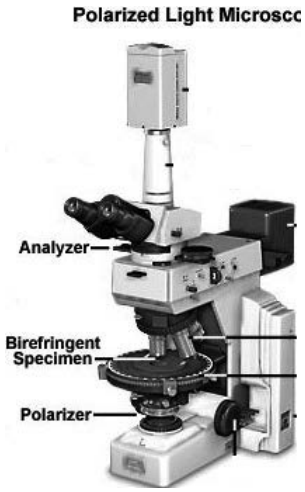

Figure A.4: An example of microscope used for polarized microscopy adapted from `soft-matter.seas.harvard.edu`

A.5 Dataset

As previously stated there are four main types of data that were not used because they were not present in enough trials in order to infer any conclusions from them. One of them is the measured properties of a film after its fabrication, being them chemical, mechanical or visual. The other three are presented below. With the increased collection of this data some new perspectives are open.

A.5.1 Raman

The first type is Raman spectroscopy. The files provided were either normal Raman or in depth Raman and an example of both is presented next (A.5 and A.7). In the case of in depth Raman a basic visualization tool was developed for internal use to ease up this process with the corresponding guided user interface to abstract from coding anything (A.6). In both images the wavelength values were omitted since they are what makes possible to identify the compounds present.

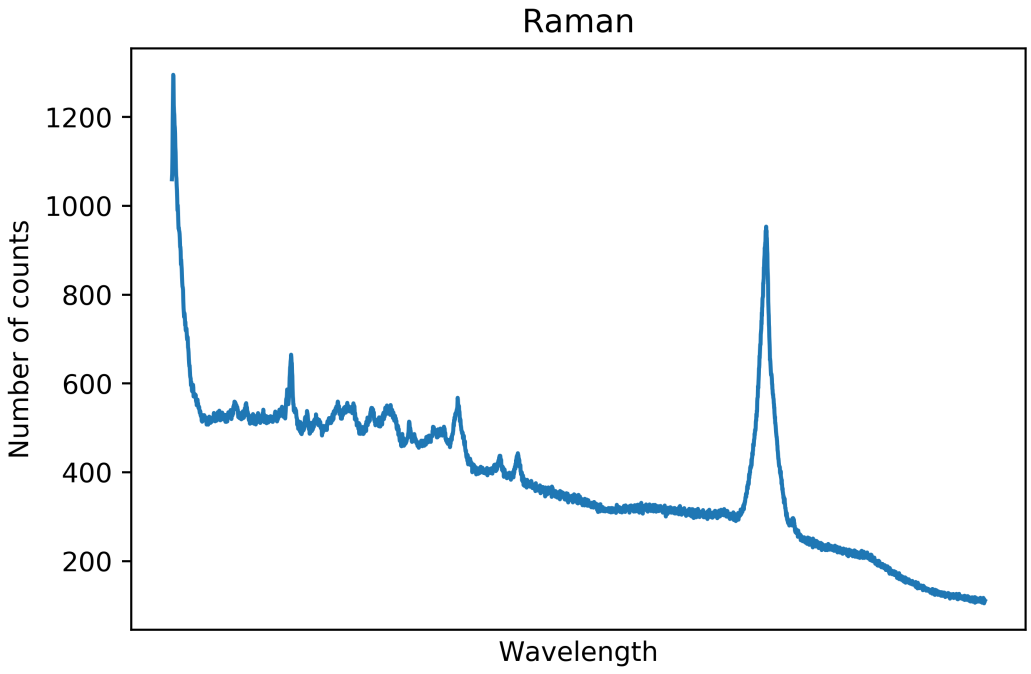

Figure A.5: An example of a provided Raman Spectroscopy plot

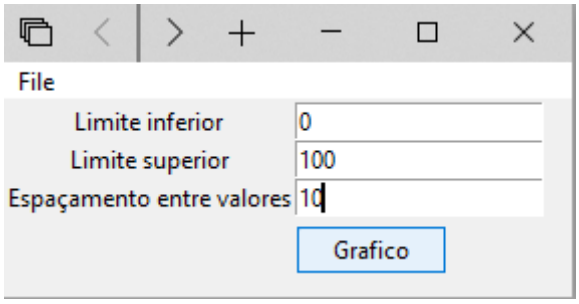

Figure A.6: Guided user interface for the plot of in depth Raman that lets the user select the lower bound, upper bound and spacing between values respectively

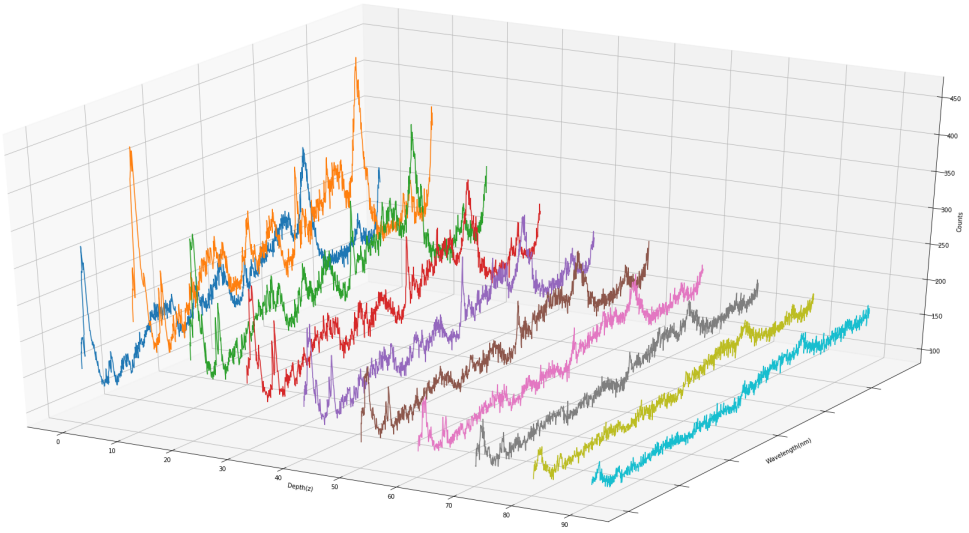

Figure A.7: An example of a plot of in depth Raman between Z=0 and Z=90 with a spacing of 10

A.5.2 FTIR

FTIR as a complementary method to Raman was only provided in its simplest form, equal to normal Raman and a plot is presented in A.8. Once again wavelength values are omitted.

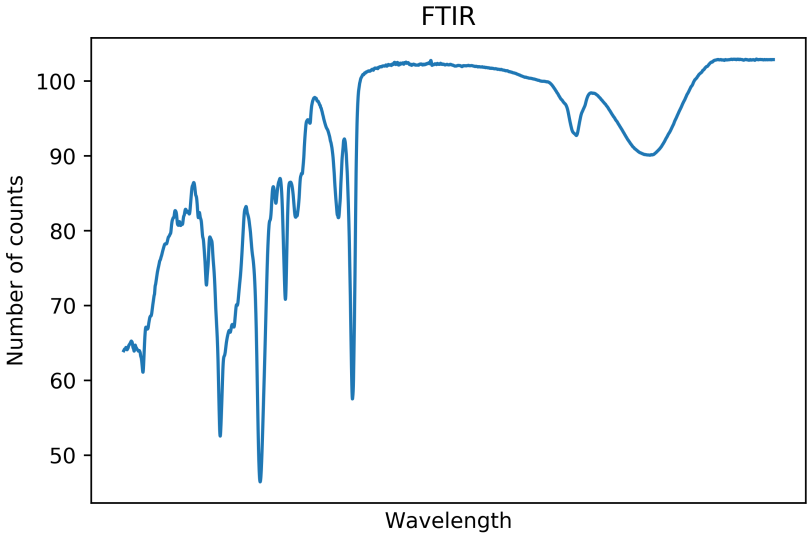

Figure A.8: An example of a provided FTIR plot

A.5.3 Polarized Microscopy

Polarized microscopy is the last type of data received. This comes in image format so traditional and cutting edge image processing (wavelets and deep learning respectively for example) can be used to check if there is more information to be extracted from this dataset. This methods have been proven to be very effective with large amounts of data.

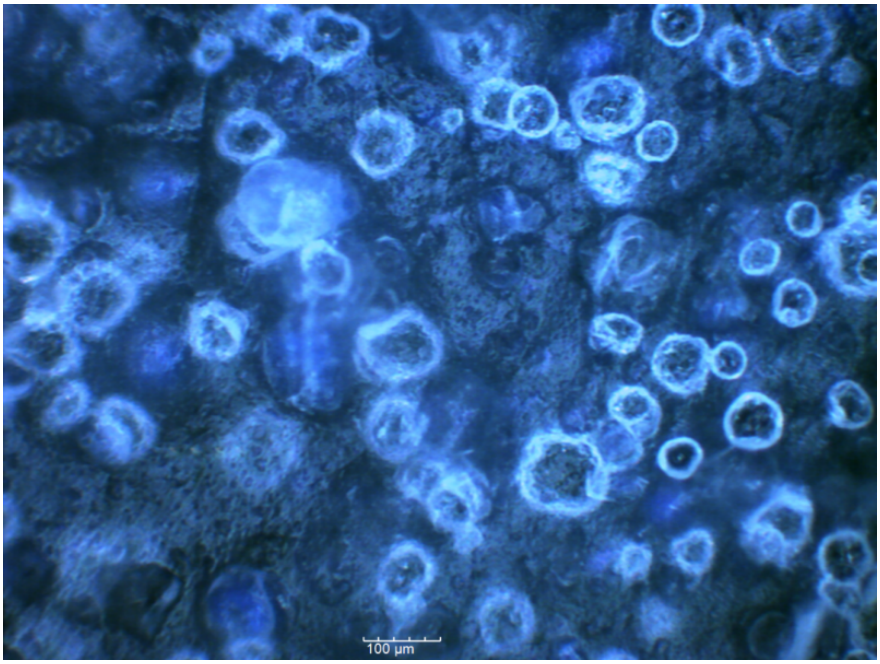

Figure A.9: An example of a provided polarized microscopy image

Bibliography

[1] R. R. Schaller, "Moore's law: past, present and future," *IEEE Spectrum*, vol. 34, pp. 52–59, June 1997.

[2] B. M. Silva, S. Vicente, S. Cunha, J. F. Coelho, C. Silva, M. S. Reis, and S. Simões, "Retrospective quality by design (rqbd) applied to the optimization of orodispersible films," *International Journal of Pharmaceutics*, vol. 528, no. 1, pp. 655 – 663, 2017.

[3] J. K. Lam, Y. Xu, A. Worsley, and I. C. Wong, "Oral transmucosal drug delivery for pediatric use," *Adv. Drug Deliv. Rev.*, vol. 73, pp. 50–62, Jun 2014.

[4] P. Shakya, N. V. Madhav, A. K. Shakya, and K. Singh, "Palatal mucosa as a route for systemic drug delivery: A review," *J Control Release*, vol. 151, pp. 2–9, Apr 2011.

[5] J. O. Morales and J. T. McConville, "Manufacture and characterization of mucoadhesive buccal films," *Eur J Pharm Biopharm*, vol. 77, pp. 187–199, Feb 2011.

[6] S. Şenel, M. J. Rathbone, M. Cans?z, and I. Pather, "Recent developments in buccal and sublingual delivery systems," *Expert Opin Drug Deliv*, vol. 9, pp. 615–628, Jun 2012.

[7] R. P. Dixit and S. P. Puthli, "Oral strip technology: overview and future potential," *J Control Release*, vol. 139, pp. 94–107, Oct 2009.

[8] A. F. Borges, C. Silva, J. F. Coelho, and S. Simoes, "Oral films: Current status and future perspectives: I - Galenical development and quality attributes," *J Control Release*, vol. 206, pp. 1–19, May 2015.

[9] S. Mazumder, N. Pavurala, P. Manda, X. Xu, C. N. Cruz, and Y. S. R. Krishnaiah, "Quality by Design approach for studying the impact of formulation and

process variables on product quality of oral disintegrating films," *Int J Pharm*, vol. 527, pp. 151–160, Jul 2017.

[10] J. C. Visser, W. M. Dohmen, W. L. Hinrichs, J. Breitkreutz, H. W. Frijlink, and H. J. Woerdenbag, "Quality by design approach for optimizing the formulation and physical properties of extemporaneously prepared orodispersible films," *International Journal of Pharmaceutics*, vol. 485, no. 4, pp. 70 – 76, 2015.

[11] A. F. Borges, B. M. Silva, C. Silva, J. F. Coelho, and S. Simoes, "Hydrophobic polymers for orodispersible films: a quality by design approach," *Expert Opin Drug Deliv*, vol. 13, pp. 1357–1374, 10 2016.

[12] G. Szakonyi and R. Zelko, "The effect of water on the solid state characteristics of pharmaceutical excipients: Molecular mechanisms, measurement techniques, and quality aspects of final dosage form," *Int J Pharm Investig*, vol. 2, pp. 18–25, Jan 2012.

[13] S. D. Saoji, S. C. Atram, P. W. Dhore, P. S. Deole, N. A. Raut, and V. S. Dave, "Influence of the Component Excipients on the Quality and Functionality of a Transdermal Film Formulation," *AAPS PharmSciTech*, vol. 16, pp. 1344–1356, Dec 2015.

[14] E. M. Hoffmann, A. Breitenbach, and J. Breitkreutz, "Advances in orodispersible films for drug delivery," *Expert Opin Drug Deliv*, vol. 8, pp. 299–316, Mar 2011.

[15] S. M. Krull, H. V. Patel, M. Li, E. Bilgili, and R. N. Davé, "Critical material attributes (cmas) of strip films loaded with poorly water-soluble drug nanoparticles: I. impact of plasticizer on film properties and dissolution," *European journal of pharmaceutical sciences : official journal of the European Federation for Pharmaceutical Sciences*, vol. 92, p. 146—155, September 2016.

[16] M. Preis, J. Breitkreutz, and N. Sandler, "Perspective: Concepts of printing technologies for oral film formulations," *Int J Pharm*, vol. 494, pp. 578–584, Oct 2015.

[17] B. M. Silva, A. F. Borges, C. Silva, J. F. Coelho, and S. Simoes, "Mucoadhesive oral films: The potential for unmet needs," *Int J Pharm*, vol. 494, pp. 537–551, Oct 2015.

[18] A. F. Borges, C. Silva, J. F. Coelho, and S. Simoes, "Oral films: Current status and future perspectives II - Intellectual property, technologies and market needs," *J Control Release*, vol. 206, pp. 108–121, May 2015.

[19] C. N. P., W. Paul, T. P. A., and M. Richard, "Ftir microscopic imaging of collagen and proteoglycan in bovine cartilage," *Biopolymers*, vol. 62, no. 1, pp. 1–8.

[20] Z. Movasaghi, S. Rehman, and D. I. ur Rehman, "Fourier transform infrared (ftir) spectroscopy of biological tissues," *Applied Spectroscopy Reviews*, vol. 43, no. 2, pp. 134–179, 2008.

[21] J. Muyonga, C. Cole, and K. Duodu, "Fourier transform infrared (ftir) spectroscopic study of acid soluble collagen and gelatin from skins and bones of young and adult nile perch (lates niloticus)," *Food Chemistry*, vol. 86, no. 3, pp. 325 – 332, 2004.

[22] C. Patacz, B. Defoort, and X. Coqueret, "Electron-beam initiated polymerization of acrylate compositions 1 : Ftir monitoring of incremental irradiation," *Radiation Physics and Chemistry*, vol. 59, no. 3, pp. 329 – 337, 2000.

[23] A. Pawlak and M. Mucha, "Thermogravimetric and ftir studies of chitosan blends," *Thermochimica Acta*, vol. 396, no. 1, pp. 153 – 166, 2003. Natas 2001 Si.

[24] M. Jain and S. Annapoorni, "Raman study of polyaniline nanofibers prepared by interfacial polymerization," *Synthetic Metals*, vol. 160, no. 15, pp. 1727 – 1732, 2010.

[25] S. E. Barnes, Z. T. Cygan, J. K. Yates, K. L. Beers, and E. J. Amis, "Raman spectroscopic monitoring of droplet polymerization in a microfluidic device," *Analyst*, vol. 131, pp. 1027–1033, 2006.

[26] M. A. J. and B. R. H., "Raman spectral changes during the solid-state polymerization of diacetylenes," *Journal of Polymer Science: Polymer Physics Edition*, vol. 11, no. 4, pp. 603–619.

[27] A. G. Kalampounias, K. S. Andrikopoulos, and S. N. Yannopoulos, "Probing the sulfur polymerization transition in situ with raman spectroscopy," *The Journal of Chemical Physics*, vol. 118, no. 18, pp. 8460–8467, 2003.

[28] C. Murli and Y. Song, "Pressure-induced polymerization of acrylic acid: A raman spectroscopic study," *The Journal of Physical Chemistry B*, vol. 114, no. 30, pp. 9744–9750, 2010. PMID: 20617845.

[29] K. Aoki, S. Usuba, M. Yoshida, Y. Kakudate, K. Tanaka, and S. Fujiwara, "Raman study of the solid-state polymerization of acetylene at high pressure," *The Journal of Chemical Physics*, vol. 89, no. 1, pp. 529–534, 1988.

[30] S. Parnell, K. Min, and M. Cakmak, "Kinetic studies of polyurethane polymerization with raman spectroscopy," *Polymer*, vol. 44, no. 18, pp. 5137 – 5144, 2003.

[31] M. Talu, E. U. Demiroğlu, Şenay Yurdakul, and S. Badoğlu, "Ftir, raman and nmr spectroscopic and dft theoretical studies on poly(n-vinylimidazole)," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 134, pp. 267 – 275, 2015.

[32] E. Gulari, K. McKeigue, and K. Y. S. Ng, "Raman and ftir spectroscopy of polymerization: bulk polymerization of methyl methacrylate and styrene," *Macromolecules*, vol. 17, no. 9, pp. 1822–1825, 1984.

[33] S. Ramesh, K. H. Leen, K. Kumutha, and A. Arof, "Ftir studies of pvc/pmma blend based polymer electrolytes," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 66, no. 4, pp. 1237 – 1242, 2007.

[34] M. W. Rohrer, "Seeing is believing: the importance of visualization in manufacturing simulation," in *2000 Winter Simulation Conference Proceedings (Cat. No.00CH37165)*, vol. 2, pp. 1211–1216 vol.2, 2000.

[35] N. Cawthon and A. V. Moere, "The effect of aesthetic on the usability of data visualization," in *Information Visualization, 2007. IV '07. 11th International Conference*, pp. 637–648, July 2007.

[36] P. K., K. J., R. R., and J. C.R., "Visualizing summary statistics and uncertainty," *Computer Graphics Forum*, vol. 29, no. 3, pp. 823–832.

[37] K. Potter, "Methods for presenting statistical information: The box plot," 01 2006.

[38] N. S. Chok, "Pearson's versus spearman's and kendall's correlation coefficients for continuous data." September 2010.

[39] M. M. Mukaka, "A guide to appropriate use of correlation coefficient in medical research," *Malawi Med J*, vol. 24, pp. 69–71, Sep 2012. jMMJ.v24.i3.pg69[PII].

[40] J. Hauke and T. Kossowski, "Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data," *Quaestiones Geographicae*, vol. 30, no. 2, pp. 87–93, 2011.

[41] M. H. Shoaib, J. Tazeen, H. A. Merchant, and R. I. Yousuf, "Evaluation of drug release kinetics from ibuprofen matrix tablets using hpmc.," *Pakistan Journal of Pharmaceutical Sciences*, vol. 19, pp. 119–124, April 2006.

[42] S. Dash, P. N. Murthy, L. Nath, P. Chowdhury, *et al.*, "Kinetic modeling on drug release from controlled drug delivery systems," *Acta Pol Pharm*, vol. 67, no. 3, pp. 217–23, 2010.

[43] S. Jamzad, L. Tutunji, and R. Fassihi, "Analysis of macromolecular changes and drug release from hydrophilic matrix systems," *International Journal of Pharmaceutics*, vol. 292, no. 1, pp. 75 – 85, 2005.

[44] J. Siepmann and N. Peppas, "Modeling of drug release from delivery systems based on hydroxypropyl methylcellulose (hpmc)," *Advanced Drug Delivery Reviews*, vol. 48, no. 2, pp. 139 – 157, 2001. Mathematical Modeling of Controlled Drug Delivery.

[45] J. M. Juran *et al.*, *Juran on quality by design: the new steps for planning quality into goods and services*. Simon and Schuster, 1992.

[46] V. Lourenço, D. Lochmann, G. Reich, J. C. Menezes, T. Herdling, and J. Schewitz, "A quality by design study applied to an industrial pharmaceutical fluid bed granulation," *European Journal of Pharmaceutics and Biopharmaceutics*, vol. 81, no. 2, pp. 438 – 447, 2012.

[47] H. Wu, M. White, and M. A. Khan, "Quality-by-design (qbd): An integrated process analytical technology (pat) approach for a dynamic pharmaceutical co-precipitation process characterization and process design space development," *International Journal of Pharmaceutics*, vol. 405, no. 1, pp. 63 – 78, 2011.

[48] L. X. Yu, G. Amidon, M. A. Khan, S. W. Hoag, J. Polli, G. K. Raju, and J. Woodcock, "Understanding pharmaceutical quality by design," *The AAPS Journal*, vol. 16, pp. 771–783, Jul 2014.

[49] L. Eriksson, E. Johansson, N. Kettaneh-Wold, C. Wikström, and S. Wold, "Design of experiments," *Principles and Applications, Learn ways AB, Stockholm*, 2000.

[50] S. N. Politis, P. Colombo, G. Colombo, and D. M. Rekkas, "Design of experiments (doe) in pharmaceutical development," *Drug development and industrial pharmacy*, vol. 43, no. 6, pp. 889–901, 2017.

[51] A. H. Schmidt and I. Molnár, "Using an innovative quality-by-design approach for development of a stability indicating uhplc method for ebastine in the api and pharmaceutical formulations," *Journal of Pharmaceutical and Biomedical Analysis*, vol. 78-79, pp. 65 – 74, 2013.

[52] J. J. Peterson, "A bayesian approach to the ich q8 definition of design space," *Journal of Biopharmaceutical Statistics*, vol. 18, no. 5, pp. 959–975, 2008. PMID: 18781528.

[53] V. F. G. and K. A. S., "Development of quality-by-design analytical methods," *Journal of Pharmaceutical Sciences*, vol. 100, no. 3, pp. 797–812.

[54] D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization," *IEEE transactions on evolutionary computation*, vol. 1, no. 1, pp. 67–82, 1997.

[55] S. Wold, M. Sjöström, and L. Eriksson, "Pls-regression: a basic tool of chemometrics," *Chemometrics and Intelligent Laboratory Systems*, vol. 58, no. 2, pp. 109 – 130, 2001. PLS Methods.

[56] C. M. Jaeckle and J. F. MacGregor, "Industrial applications of product design through the inversion of latent variable models," *Chemometrics and Intelligent Laboratory Systems*, vol. 50, no. 3, pp. 199 – 210, 2000.