Universidade de Lisboa

Faculdade de Farmácia

Generic Medicines Development: a fast and cost-effective reverse engineering method

Beatriz Calado Lourenço da Costa

Dissertation supervised by Professor João Almeida Lopes and co-supervised by Professor Catarina Pinto Reis

Master Course in Pharmaceutical Engineering

2022

Universidade de Lisboa

Faculdade de Farmácia

Generic Medicines Development: a fast and cost-effective reverse engineering method

Beatriz Calado Lourenço da Costa

Dissertation supervised by Professor João Almeida Lopes and co-supervised by Professor Catarina Pinto Reis

Master Course in Pharmaceutical Engineering

2022

Acknowledgements

I would like to thank the following people, without whom it would not be possible to complete this thesis and finish the master's degree.

I thank my supervisor, Professor João Almeida Lopes, for his guidance, feedback throughout this project, and for providing the device that made the development of this thesis possible.

Thank to my college Milene Isabel Rosa Fialho, who completed the 4th year of the Integrated Master’s Degree in Pharmaceutical Sciences, who helped in the developed formulations used in this work.

I also want to thank all my family who have always supported me along this path and have always supported all my choices

Preface

This dissertation is submitted for the Master of Pharmaceutical Engineering at the Faculty of Pharmacy, University of Lisbon. The research described here was perform at Pharmacy Laboratories, under the supervision of Prof. Dr. João Lopes. The thesis was co-supervised by Professor Catarina Pinto Reis, between March – August 2022.

Abstract

This work aimed to develop a straightforward method based on infrared spectroscopy for the estimation of solid pharmaceutical formulations compounds concentration in the context of the development of generic medicines. The proposed method can be extremely advantageous as the resources needed to obtain a first quantitative estimation of the formulation require only the knowledge of the infrared spectra of each formulation component (does not require a calibration procedure). It is based on the use of the pure components infrared spectra, the spectrum of the target pharmaceutical product and an especially designed algorithm based on the assumption of the Lambert-Beer's law. The method was tested with a formulation (powder mixture) containing paracetamol and caffeine (the active substance), starch, talc, microcrystalline cellulose, magnesium stearate and lactose (excipients). The proposed method (calibration-free method) was compared with the multivariate curve resolution method (MCR), a supervised method that requires a calibration with standards. A series of formulations were produced according to an experimental design of the type D-optimal by changing within certain ranges the concentration of each component (only paracetamol was kept constant). Two independent sets of formulations were designed: one for calibration and one for testing. The implementation of the MCR and the calibration-free method were performed according to different scenarios, simulating more or less uncertainty in the initial guess of the component's concentrations. Results for MCR showed that for this method it was fundamental that the pure components spectra were known and used as constrains, thus, estimating only the concentrations. Accuracy and precision of the estimations were highly related with the specific features of the infrared spectrum of each component. The calibration-free method demonstrated that estimations of the formulation components concentrations were similar or even better than those obtained for the MCR method. The results also demonstrated that accuracy was somehow dependent on the allowed range for each component (initial guess). Initial guess of the concentration for each component must not deviate above 50% of its real value for an adequate estimation.

As conclusion, the proposed method demonstrated to be an excellent method to obtain a first estimate of the composition of a solid formulation, that can be fined tuned afterwards using other complementary techniques.

Keywords: Generic Medicines; Reverse Engineering; Design of Experiments; Chemometrics; Fourier Transform Infrared Spectroscopy.

Resumo

Este trabalho visava desenvolver um método simples baseado na espectroscopia infravermelha para a estimativa da concentração de compostos sólidos de formulações farmacêuticas no contexto do desenvolvimento de medicamentos genéricos.

O método proposto pode ser extremamente vantajoso uma vez que os recursos necessários para obter uma primeira estimativa quantitativa da formulação requerem apenas o conhecimento dos espectros infravermelhos de cada componente da formulação (não requer um procedimento de calibração). Baseia-se na utilização dos espectros de infravermelhos de componentes puros, no espectro do produto farmacêutico alvo e num algoritmo especialmente concebido com base no pressuposto da lei de Lambert-Beer. O método foi testado com uma formulação (mistura em pó) contendo paracetamol (a substância ativa), cafeína, amido, talco, celulose microcristalina, estearato de magnésio e lactose (o *filler*). O método proposto (método sem calibração) foi comparado com o método de resolução de curva multivariada (MCR), um método supervisionado que requer uma calibração, e, portanto, a existência de padrões. Portanto, foi produzida uma série de formulações de acordo com um desenho experimental do tipo D-optimal, alterando dentro de certos intervalos a concentração de cada componente (apenas o paracetamol foi mantido constante). Dois conjuntos independentes de formulações foram concebidos: um para calibração e outro para testes. A implementação do MCR e do método sem calibração foi realizada de acordo com diferentes cenários, simulando mais ou menos incerteza no palpite inicial das concentrações dos componentes.

Os resultados para o MCR mostram que para este método é fundamental que os espectros dos componentes puros sejam conhecidos e utilizados como constrangimentos, estimando, portanto, apenas as concentrações. A exatidão e precisão das estimativas estava altamente relacionada com as características específicas do espectro de infravermelhos de cada componente. O método sem calibração demonstrou que as estimativas das concentrações dos componentes da formulação eram semelhantes ou mesmo melhores do que as obtidas para o método MCR. Os resultados também demonstraram que a precisão depende de alguma forma do intervalo permitido para cada componente (palpite inicial). A previsão da concentração para cada componente não deve desviar-se acima de 50% do seu valor real para uma estimativa adequada. Em resumo, o método proposto demonstrou ser um excelente método para obter uma primeira estimativa da composição de uma formulação sólida, que pode ser afinada posteriormente utilizando outras técnicas.

Palavras-chave: Medicamentos Genéricos; Engenharia Inversa; Desenho de Experiências; Quimiometria; Espetroscopia de Infravermelhos por Transformada de Fourier.

Resumo Alargado

Um medicamento genérico é um medicamento que deve ser equivalente ao medicamento de referência já estabelecido no mercado. A quota de mercado dos medicamentos genéricos para Portugal é aproximadamente de 47.35% (1). O aparecimento dos medicamentos genéricos no mercado começa quando a patente dos medicamentos de referência expira. Assim, a pesquisa por parte das empresas que produzem genéricos começa antes da patente expirar. A composição destes medicamentos que estão sobre patente não são completamento conhecidas, geralmente não se tem conhecimentos dos excipientes que são utilizados nem é conhecido os métodos de produção. Como consequência direta, as empresas têm de encontrar formas mais rápidas, eficazes e menos dispendiosas de descobrir as composições e os métodos de produção. Surgindo assim a necessidade de utilizar a engenharia reversa para o desenvolvimento de medicamentos genéricos.

No desenvolvimento e produção de medicamentos genéricos, e para que estes sejam bem-sucedidos, é esperado que estes apresentem a mesma bioequivalência farmacêutica que o medicamento de referência, a mesma quantidade substância ativa, com a expectativa de que o genérico tenha a mesma qualidade, segurança e eficácia (2–4). Isto já não é esperado para os medicamentos híbridos que são medicamentos baseados no medicamento de referência, aqui as diferenças esta na dosagem, via de administração, ou na indicação terapêutica (5,6).

Como os medicamentos de referência estão protegidos sob patentes a informação sobre a fórmula e método de preparação é escassa. O que leva a necessidade de encontrar ferramentas que ultrapassem este problema. Assim, a engenharia reversa que é conhecida como o processo de inversão das etapas de engenharia para replicar um sistema e subsistemas quando há falta de informação (7). A engenharia reversa é então utilizada para descobrir os componentes de um produto desconhecido. Podendo levar a identificação, quantificação e caracterização das substâncias ativas e excipientes. Este é um processo que pode levar algum tempo de que pode ser dispendioso(7,8).

A forma farmacêutica mais utilizada para os medicamentos genéricos é a sólida, assim a fórmula que foi realizada e analisada neste trabalho foi uma misturas de pós. Esta misturas foram realizadas no laboratório da Faculdade de Farmácia. Estas misturas contêm paracetamol, cafeína, amido, talco, celulose microcristalina, estearato de magnésio e lactose, as concentrações destes componentes variaram de acordo com um desenho experimental (DoE) do tipo D-optimal, sendo que o paracetamol tinha um valor fixo e a lactose foi usado como *filler*.

Os pós sólidos são preparações formadas por partículas sólidas secas, livre e finas ou não (9). Pode ter uma ou mais substâncias ativas, que neste caso é o paracetamol e a cafeína, e pode ter também diferentes excipientes. Neste trabalho foram utilizados como excipientes: o amido, talco, celulose, celulose microcristalina, estearato de magnésio e lactose.

A engenharia reversa é utilizada na produção de formulações genérica sólidas, começa por descodificar a fórmula quantitativa no medicamento de referência, de seguida é feita a quantificação e identificação dos excipientes no medicamento de referência. O passo seguinte é a caracterização do estado sólido da substância ativa e, por último, é então a processo de fabrico da forma sólida (10).

O método de espetroscopia utilizado foi espectroscopia de infravermelhos por transformada de refletância total atenuada (ATR-FTIR). Este foi usado para a obtenção dos espetros das amostras e dos compostos. O espectro vibracional é único para cada composto, podendo assim ser utilizado como uma técnica de impressão digital para a identificação (11). A espetroscopia vibracional pode ser implementada para estudar diferentes compostos e misturas e passar por testes simples de identificação até uma análise mais profunda, de espetro total, qualitativa e quantitativa (12,13). O método é usado na engenharia reversa para a análise de produtos farmacêuticos e a análise dos resultados pode ser analisada através dos métodos quimiometria, como o *partial lest-squares* (PLS) e análise de curva multivariadas (MCR).

A quimiometria é descrita como uma nova forma de análise de dados químicos onde os elementos estatísticos e químicos são combinados. Para isso são utilizados métodos matemáticos e estatísticos para a obtenção de informação. Há sempre três elementos que são utilizados na aplicação desta tecnologia: 1) modelação empírica, 2) modelação multivariada, e 3) dados químicos (14,15).

A análise de *quality by design* (QbD) deve ser utilizada para se ter um desenvolvimento, uma otimização robusta e um método analítico rentável. Por isso, o objetivo do QbD é ter especificações significativas que se baseiem no desempenho clínico, aumentar a capacidade e reduzir a variabilidade e falhar do produto, melhorar o desenvolvimento do produto e eficácia de produção. Há algumas limitações que devem ser ultrapassadas como a condução inadequada do desenvolvimento e otimização, assim sendo que o desenho experimental (DoE) é utilizado para ultrapassar estes problemas (16).

O pré-processamento dos dados é realizado para linearizar a resposta da análise das variáveis. Quando se tem dados de espectroscopia é importante que sejam pré-processados.

Os métodos utilizados neste trabalho e que foram aplicados foi a correção da *baseline*. Este método é normalmente utilizado para a espectroscopia. O outro foi o valor absoluto que é utilizado para remover a informação negativa dos dados, permitindo a utilização de restrições não negativas, visto que esta é uma das restrições da análise de curva multivariadas (MCR). Por último, foi também utilizado a normalização. Esta é realizada para corrigir as diferenças de escala, ajuda na forma em que dá a todas as amostras um impacto igual no modelo (17,18).

Este trabalho teve como objetivo avaliar um método que não requeria calibração para a estimativa das concentrações dos componentes, em formulações sólidas, no contexto do desenvolvimento de medicamentos genéricos. A hipótese de trabalho foi de verificar se é possível ter estimativas adequadas da composição quantitativa de uma formulação sólida, uma mistura de pó, recorrendo a equipamento de laboratório e espectroscopia infravermelha.

A análise de curvas multivariadas (MCR) foi o primeiro método realizado. Este necessita de calibração e utiliza um conjunto de formulações de calibração e outro de teste. Este é um método mais comum utilizada na engenharia reversa, que requer múltiplas reproduções do produto através de um DoE, procedimento de calibração e, por vezes, até é utilizado a escala de produção, sendo assim mais demorado e dispendioso. O DoE é utilizado neste método devido a necessidade da calibração e validação. Para este método foram realizados 4 modelos distintos onde se usava pré-processamento ou não e onde se dava os espetros puros ou não. Também os intervalos usados variaram para que fosse possível analisar o melhor intervalo para a estimativa das concentrações. O MCR apresentou dificuldades na reconstrução dos espetros dos compostos das misturas quando não eram dados os espetros compostos puros. O método funciona embora tenha as suas limitações e seja necessário conhecer os espetros dos compostos da formulação analisada.

O método proposto não tem a necessidade de qualquer procedimento de calibração ou reprodução de medicamentos, é um algoritmo simples de reconstrução espectral. Neste método foram fornecidos os espetros puros dos compostos e os espetros das misturas. Foi testado para estimar a composição de algumas das misturas de pós e foi também utilizado para a estimativa de todas as amostras. Para este método forma usados dois intervalos e foi executado considerando diferentes margens (concentrações) para cada composto. Foram testadas variações percentuais em torno da concentração conhecida de cada composto desde a melhor situação até a pior. A reconstrução das concentrações reais foi possível quando utilizo o algoritmo, e concluiu-se que as concentrações inicias para cada composição não devem desviar-se acima de 50% do seu valor real para que se obtenha uma estimativa adequada.

Uma comparação entre o método supervisionado (MCR) e o método calibration-free foi realizada. O *design of experiment*s foi utilizado para avaliar o desempenho do método baseado em engenharia reversa (o algoritmo) com o método MCR. Ambos foram capazes de realizar as estimativas dos componentes, mas também mostram ter as suas limitações. O MCR apresentou dificuldades da estimativa da cafeína e do amido, e durante o período de trabalho destinado à realização desta dissertação não foi encontrada uma solução para ultrapassar o problema. No que respeita ao algoritmo, este não pode ter um grande desvio do valor real para alguns dos compostos, como se observou para o paracetamol e lactose.

Table of Contents

List of Figures

List of Tables

List of Acronyms and Abbreviations

- ALS – Alternating Least Squares
- ANDA – Abbreviated new drug application
- API – Active Pharmaceutical Ingredient
- ATR – Attenuated total reflectance
- DoE – Design of Experiment
- EMA – European Medicines Agency
- FDA – U.S. Food and Drug Administration
- FTIR – Fourier Transform Infrared
- MCC – Microcrystalline cellulose
- MCR – Multivariate Curve Resolution
- MgS – Magnesium Stearate
- NIR – Near-infrared Spectroscopy
- OOS – out-of-specification
- PLS – Partial lest-squares
- QbD – Quality by Design
- RE – Relative Error
- RLD – Refence Listed Drug
- RMSEP – Root mean square error of prediction
- SEP – Standard error of prediction

1. Introduction

The development of generic medicines, in general, is cheaper than the innovator drug product. The sales of generic drugs are a more affordable alternative with the same quality, safety, and effectiveness. The marketing authorization process for these medicines requires several steps. Herein, regulation plays a major part in this process ensuring that the approval of a generic must be subjected to the specific regulatory requirements as the reference drug product. To be approved, an abbreviated new drug application (ANDA) report must be submitted to a regulatory agency. Generics must have the same active ingredient, but it can be with different salts, ensured bioequivalence, therapeutic indications, the same dosage form, and similar properties to the reference drug. The development of a generic usually begins before the innovative product patent expires. To address all requirements imposed by regulatory agencies for approval, pharma companies face some challenges in the development process. Identifying the appropriate formulation and manufacturing process to comply with bioequivalence constraints is very often a challenge and a time-consuming task.

Reverse engineering is described as "the reversal of the engineering, starting with the replication of a prevailing component, subassembly, or product itself without the facilitation of drawings, documentation, or computer modeling" (19,20). Through reverse engineering is possible to uncover the compounds unknown in the reference drug product. The use of reverse engineering in the development of generics came with the necessity of overcoming issues related to the patenting process. Since the reference drug products are under patent the information about the excipients and manufacturing process is very scarce.

To develop a generic medicine, the qualitative and quantitative formula must be similar to the reference drug product. This process can resource to multiple methods. The use of vibrational spectroscopy, for example, Fourier Transform Infrared (FTIR) spectroscopy is a possibility to unveil the compounds present in the drug product.

Chemometrics and statistical methods are used to analyze infrared data and also for the process of identifying the formulation. Multivariate curve resolution is a method used for that purpose. This is a method based on the assumption that Beer's Law stands in infrared and can be used to obtain qualitative and quantitative information about the formulations, through infrared analysis of the reference product and reproduction. In this context, and when there is a need to make reproductions of the reference drug product, experimental design (DoE) is normally a resource to use. It is used in the development process to unveil the formulation and manufacturing process. Applying DoEs can provide better and faster results with the less experimental burden.

1.1. Main Goals

The objective of this work is to evaluate the possibility of using a methodology based on infrared spectroscopy to obtain an estimation of the compositions of a solid pharmaceutical form in the context of reverse engineering in the development of generic. The hypothesis is to verify if it is possible to have adequate estimates of the quantitative composition of a finished drug product resourcing to laboratory equipment and infrared spectroscopy coupled to a simple spectral reconstruction algorithm, without the need for any calibration procedure or drug product reproductions. In this master thesis, this algorithm is designated as a "calibration-free" algorithm. The performance of the proposed methodology will be compared with the more common methodology used in reverse engineering, which requires multiple reproductions of the product through an experimental design, calibration procedure, and sometimes even using the production scale, therefore more time-consuming and expensive. In this thesis, the estimations based on multivariate curve resolution will be compared. Both methods will resource on powder blends produced according to an experimental design.

2. Start of The Art

2.1. Generic Medicines

A generic drug product should be equivalent to a reference drug product already on the market. In the successful development and production of generic drugs, it is expected that they will have the same pharmaceutical bioequivalence, the active ingredient, dosage, strength, and route of administration as the reference drug (2–4). The active substance of a generic drug is considered the same if it consists of different salts, esters, ethers, isomers, mixtures of isomers, and complexes or derivatives of an active substance, only if the properties vary significantly concerning efficacy or safety (20). Generic drugs are also expected to have the same bioequivalence, therapeutic equivalence, and safety and efficacy as reference drugs (2–4). A generic company may manufacture a drug that is based on a reference drug, in this new drug what may be different from the reference drug is strength, route of administration, or indication. These are called hybrid medicines and are authorized medicines that depend in part on the test results of the reference medicine and in part on new clinical trial data (5,6).

The approval of a generic must go through a regulator such as the U.S. Food and Drug Administration (FDA) or European Medicines Agency (EMA), among others. The abbreviated new drug application (ANDA) contains the data that will be submitted to the regulator. When the generic drugs are approved, the companies can produce and market them. After this, the market would have an alternative to the reference drug that is equally effective, safe, and more affordable (21,22). In general, this process usually does not require preclinical (animal) and clinical (human) data to establish safety and effectiveness, because of that is used the term “abbreviated” is for generic drug applications. On the other hand, generic companies must demonstrate that the product performs in the same way as the innovator product (21). The guideline on the investigation of bioequivalence describes the “specific requirements for the design, conduct, and evaluation of bioequivalence studies for immediate release dosage forms with systemic action” (20). The pre-clinical tests and clinical trials that are performed on the reference drugs are not performed for the generic drugs must be established bioequivalence being then demonstrated the equivalence in biopharmaceutical quality between the drugs (20).

Since the generic companies should discover the excipients and the production method of the RLD. Being under patent means that the information’s available about the medicine is very scarce. So, the companies should find methods that are efficient and fast, due to the higher investments that are needed to develop the research (7). For the development of a potential generic, the steps that are involved are: 1) the characterization of reference product; 2) design of the generic product and process; 3) pivotal biobatch, 4) bioequivalence study; and 5) commercial product manufacture (23). Comparing to RLDs, generics have a lower risk of total failure, since the active substance have already been established as safe and efficient. The cost efficiency is more important, because of the lower profit margins and competitions with other generics (23).

2.2. Reverse Engineering

Reverse engineering can be applied in the medical field, pharmaceuticals, therapeutic peptide production, bioinformatics, biosystems, and other fields. This tool can be used or required in the pharmaceutical industry for various reasons such as patent infringement, analytical issues, stability problems, safety issues, and generic design and development. The method is used to reproduce and re-design an existing product. For the generic drug business, it is critical to be the first to profit the most, since several companies will be competing for the creation of a successful generic drug that will be off patent. In generic companies, obtaining bioequivalence is the most critical area and most generics are dosage forms (24). Bioequivalence is a prerequisite for applying for generic approval under ANDA to the FDA or other agencies (19). This means that the levels of the active substance in the blood over time must be the same as in the innovator product, one way to show this is through identical dissolution profiles. The same does not apply to biosimilars. In these, the similarity in composition must be high, and despite this requirement, the FDA does not disclose the innovator formulation (24).

The reference product, as already mentioned, does not have prior information available about its components, manufacturing process, and documentation. Therefore, reverse engineering (RE) is a process known as reversing engineering steps to replicate a system and its subsystems or subassemblies when information is lacking (7). Reverse engineering or deformation is used to discover the components of an unknown product. This process can lead to the identification, quantification, and characterization of the active pharmaceutical ingredients (API) and all excipients in reference drugs in the reference product (7,8). The pharmaceutical deformulation can also be used to develop a reformulated product with greater bioequivalence. The RE development process begins with finding the qualitative formulation, then the quantitative formulation of excipients that may be critical to the stability or performance modification of the drug (7).

According to the EMA, all drugs should be manufactured following the quality standards, therefore generic drugs are also included in this standard (22). To be able to launch the generic on the market, it must have essential properties, such as 1) having the same API (Q1 qualitative), 2) having the same quantities (Q2 quantitative), and 3) having the same physical and chemical properties (Q3) as the RLD. Through the application of reverse engineering, it is possible to investigate and obtain the necessary information about Q1, Q2, and Q3 that is required to obtain a generic (7). Generic drugs usually enter the market when the patent of the reference drug expires. Usually, the exclusivity time is 10 years, after this time the generic drugs start appearing on the market (22).

2.3. Solid Forms

The development of solid pharmaceutical forms requires various technical and regulatory challenges. Some of these include active ingredient properties, ensuring compatibility of excipients with active ingredients over the product shelf life, processing and manufacturing, quality controls, and compliance with regulatory agencies (25). The "guideline on manufacture of the finished dosage form" from EMA it describes guidance for the manufacture of a solid finished form. This guidance serves to clarify the type and level of information that is required (26). The development of solid pharmaceutical forms requires various technical and regulatory challenges. Some of these requirements include active ingredient properties, ensuring compatibility of excipients with active ingredients over the product shelf life, processing and manufacturing, quality controls, and compliance with regulatory agencies. The production of solid oral dosage forms has required some tests to manufacture generics. The following API and finished dosage form tests are required. The API test is performed to be able to select the raw material supplier and characterize the quality of the raw material in each batch. The API test is performed to verify the characteristics that can influence formulation development. Regarding the second test, this is performed to identify the formulation, in vitro dissolution screening for acceptable performance, and release of the dosage form (25).

Solid powders are preparations formed by dry solid particles, free and fine or not. This can have one or more active substances, with different excipients, and if necessary, coloring and flavoring. The route of administration can be in water or with water, or other liquids. The powder can be presented as unit-dose or preparations multidose (9). The production of solid powders requires specific knowledge about appropriate particle size for the intended use (9).

The production of generic solid formulations through reverse engineering starts by decoding the quantitative formula in the RLD. The excipients that affect the quality test should be the first to be identified. The next step is to perform the qualification and identification of the excipients in the RLD that will present a challenge by interfering with other excipients for the separation of excipients HPLC method used, and for the quantification, near-infrared spectroscopy (NIR) is used. Then, it is followed by the characterization of the solid-state of active pharmaceutical ingredient (API). The solid state of API can be categorized by the following molecular, particle, or bulk properties. The manufacturing process corresponds to the next step in the production of solid form. It can be manufactured through wet granulation, dry granulation, or direct compression, depending on the stability profile of the API, the API-total tablet weight ratio, and the physic-chemical properties. Figure 1 depicts the steps that must be followed to obtain a solid generic drug that complies with the regulatory agencies (10).

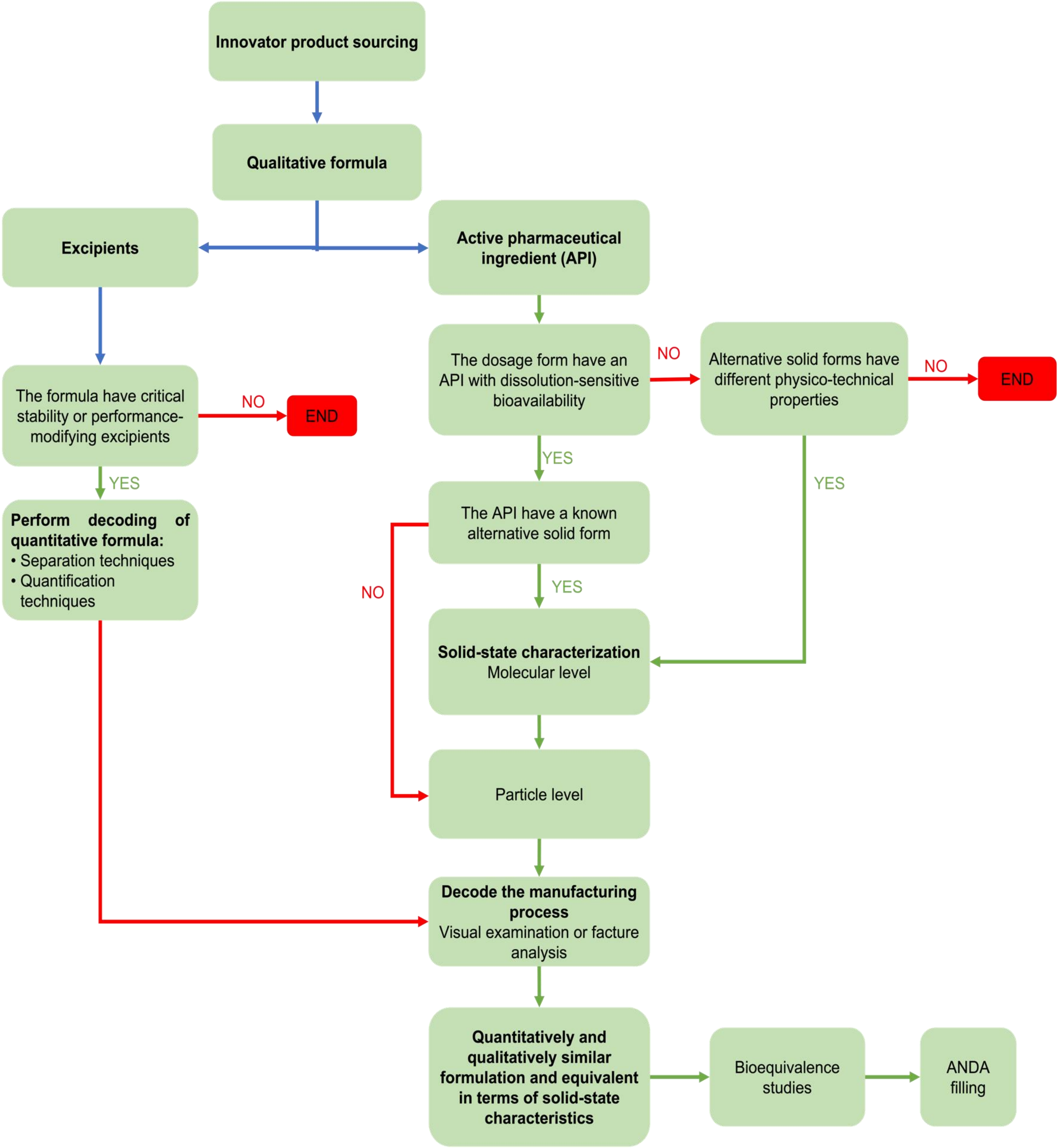
Innovator product sourcing
Qualitative formula
Excipients
Active pharmaceutical ingredient (API)
The dosage form have an API with dissolution-sensitive bioavailability
NO
Alternative solid forms have different physico-technical properties
NO
END
The formula have critical stability or performance-modifying excipients
NO
END
YES
YES
Perform decoding of quantitative formula:
• Separation techniques
• Quantification techniques
YES
The API have a known alternative solid form
NO
YES
Solid-state characterization
Molecular level
Particle level
Decode the manufacturing process
Visual examination or facture analysis
Quantitatively and qualitatively similar formulation and equivalent in terms of solid-state characteristics
Bioequivalence studies
ANDA filling

Figure 1: Protocol that can be applied for reverse engineering when making a solid formulation (10).

2.4. Vibrational Spectroscopy

Several different vibrational spectroscopy techniques are implemented, the most important of which are mid-infrared (IR), near-IR, and Raman spectroscopy. These techniques use specific vibrations that characterize different molecular structures, and like all techniques, these have advantages and disadvantages concerning instrumentation, sample handling, and applications. Sample information can be obtained through the analysis of the absorbance profiles at different single characteristic wavenumbers (11–13). Each component has a specific intensity that will show in the vibrational spectrum (Figure 2) (12).

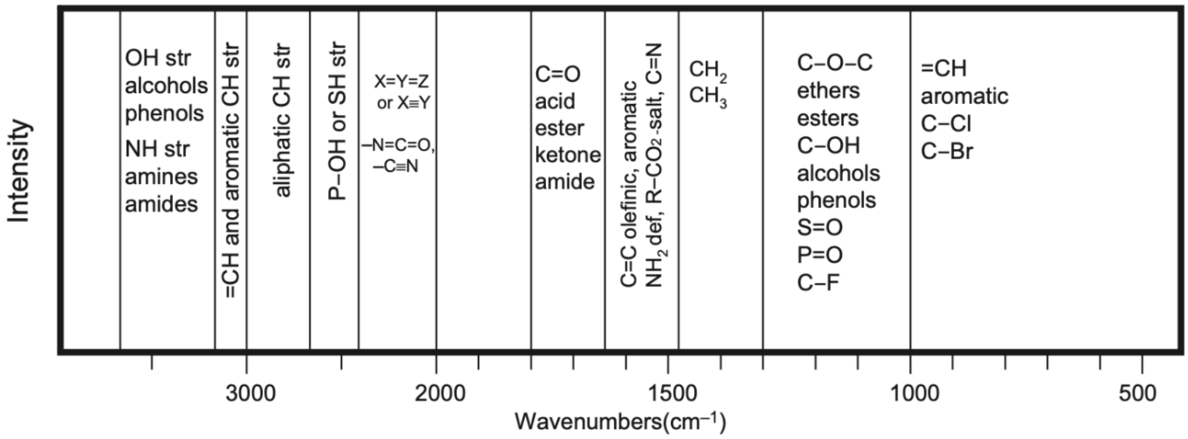

Figure 2: Characteristic group frequencies for the regions of the fundamental spectrum.

The vibrational spectrum is unique for each compound. Infrared (IR) can therefore be used as a fingerprint technique for the identification of compounds. Differences between the compounds can be small and sometimes more sophisticated approaches are needed to develop the analysis of the data (11). Vibrational spectroscopy is implemented to study a vast range of different compounds and mixtures and can be carried out from a simple identification test to an in-depth, full-spectrum, qualitative and quantitative analysis (12,13).

2.4.1. Fourier-Transform Infrared Spectroscopy

The Fourier Transform Infrared Spectroscopy (FTIR) allows the search of an infrared spectrum, either the emission or absorption of liquids, semi-solids, or solids. It detects different functional groups, and the range that can be obtained is between 4000 and 600 cm^{-1} (27). FTIR is a technique that is safer for the environment, reduces time and cost, is non-invasive, has a higher detection capability and is not necessary a prior preparation of the samples (28).

Attenuated total reflectance (ATR) uses the phenomenon of total internal reflection. So, the technique is a method of contact that involves a crystal with a high refractive index that also has good properties of IR transmitting (17). This instrument is an accessory used to quantify changes that happen to an internally reflected infrared beam when it gets in contact with the sample. Therefore, the infrared beam will focus on the crystal with a high refractive index at a set angle (28).

FTIR-ATR spectroscopy has some advantages. The attenuate total reflectance has few optical parts and no slit to attenuate radiation. So, that means that the radiation power that reaches the detector is bigger when compare with other FTIR methods. Another advantage is the higher resolving power and wavelength reproducibility, allowing the analysis of complex spectra. The last one, is the fact that all the elements can reach the detector at the same time, and with this is possible to obtained data for an entire spectrum (29). The use of this technique combined with the multivariate analysis can improve the quality of the results obtained. The most widely used is the partial least squares method (PLS). Therefore, since the PLS is one of the most used techniques, in this work, the multivariate curve resolution (MCR) method is used to analyze the donors and obtain the concentrations of each of the components in the samples (30).

2.5. Chemometrics

Chemometrics can be described as a new way of analyzing chemical data where the statistical and chemical elements are combined. That means mathematical and statistical methods are used to obtain information from physical and chemical phenomena. In this science, these three elements: empirical modeling, multivariate modeling, and chemical data, are always used (14,15).

When a model is being built is likely that prior knowledge or theoretical relations concerning the chemistry of the sample, or the physics of the analyzer will be used. An example of this is the Lambert-Beer Law, Equation 1, which relates the intensity of the spectrum and the concentration (15):

$$A = \varepsilon * b * C \qquad \textit{Equation 1}$$

where A represents absorbance, ε is the molar absorption coefficient in units L/(mol cm), b is the path length of the measurement in units of cm, and C is the concentration in the unit of mol/L (15).

Chemometrics can be used to collect multivariate data and analysis protocols, calibration, process modeling, pattern recognition and classification, signal correction and compression, and statistical control. In the context of this work, chemometrics can help with the determination of the different pharmaceutical properties in powders, granules, and tablets which means that this technique can be ideal for the extraction of quantitative information from the samples.(14).

This technique can be used for various tasks, including experimental design, exploratory data analysis, and for the development of predictive models. In the area of analytical chemistry, the use of chemometrics has proven most effective for two functions: instrument specialization, for the construction of multivariate calibration models that provide selectivity for multivariate analytical instruments, or for information extraction, where the tools of chemometrics are used to obtain the unknown information that is present in information-rich multivariate analytical instruments (15).

2.5.1. Quality-by-Design and Design of Experiments

According to Yu et al., 2014 (16), Quality by Design (QbD) is a “systematic approach to development that begins with predefined objectives and emphasizes product and process understanding and control-based sound science and quality risk management” (16). The objectives of QbD are (16):

1. To have significant specification that are based on clinical performance;
2. To increase capability and reduce product variability and flaws through the improvement of the product and process design, understanding, and control of the product;
3. To enhance product development and production effectiveness;
4. To improve root cause analysis and post-approval change management.

Having a development and robust optimization and cost-effective analytical method is beneficial for the use of the QbD analysis. This implementation allows a better solution to out-of-specification (OOS) product results and reduces the risk of method failure (16).

There are a few limitations that should be overcome. The limitation can be inadequate conduction of the development and optimization. So, the Design of experiments (DoE) is used to overcome these problems. DoE can offer better results with a smaller number of experiments (16).

The design of experiments, DoE, is a structured and organized method to determine the associations between input factors (independent variables) affecting one or more output responses (dependent variables) by establishing mathematical models. So, this method allows the incorporation of quality into the product. It also enables cost reduction, saves time, has reliable quality, and the formulation provided is robust. To have a proper DoE must be considered a few aspects. Therefore, the objectives should be defined, and the number of inputs factors and interactions, the statistical validity and effectiveness of each design should be studied (31).

2.5.2. Pre-Processing Methods

Spectrum manipulation usually is made to improve and help with the qualitative and quantitative interpretation of spectra. Therefore, several techniques are available to the user of infrared spectrometers. After the samples have been scanned with infrared spectroscopy, the data will be manipulated using different methods, such as baseline correction, smoothing, difference spectra, derivatives, deconvolution, and curve-fitting (17).

Consequently, the preprocessing of data is performed to linearize the response of the analysis of the variables. Preprocessing will modify the data and is made before building the model. It has special importance when you have data obtained by spectroscopy, including infrared spectroscopy.

Some pre-processing methods (18) used in this thesis were:

- Baseline (Weighted Least Squares)
  In this method, the baseline offset is automatically removed from the data. The method is usually used in spectroscopy applications when the signal of variables is owed only to the baseline (background).
- Normalization
  The normalization is made to correct the scaling differences that arise from path length effects, scattering effects, sources or detector variations, or other general instrumental sensitivity effects. This helps in the way that gives all the samples an equal impact on the model.
- Absolute value
  This method is used to remove the negative information in the data and can be used after the derivative method or after other methods. The correction allows the use of non-negative constraints or improves the analysis of derivatized spectra. The use of this pre-processing can create a non-linear response and complicate modeling.

2.5.3. Multivariate Curve Resolution

Multivariate Curve Resolution (MCR) is a bilinear model that offers simplified and understandable information of the process data under the principle that the multicomponent Beer's law is valid. So, MCR is one of the most commonly used multivariate calibration models. The application of the method has proven successful in different types of data that came from distinct instruments, like IR, chromatography, hyperspectral imaging, nuclear magnetic resonance, and X-ray fluorescence (32–34). MCR is used to obtain additional information about the concentration profile or the spectra (identification) of the pure components of a mixture. Equation 2 describes, mathematically, the Multivariate Calibration Resolution (MCR) method (32,33):

$$D = CS^T + E \qquad \textit{Equation 2}$$

In Equation 2, the D is the original data matrix (in this work matrix composed of multiple IR spectra), C is the concentration matrix (in this are the concentrations of the compounds), S is the non-augmented matrix (the pure spectra), and the E is the residual matrix. The alternating least squares algorithm is usually applied to estimate C and S matrices from D. Its application depends on the level of knowledge existing for the C and S (concentrations of compounds and pure spectra of compounds, respectively). This method typically does not need a lot of prerequisites or prior information about the chemical identity of the components, but it is convenient to know the number of pure components. There is not any "golden rule" in this method. A trial-and-error approach based on different estimations is recommended, and the residuals must be analyzed carefully. Adding the profiles of the components is an implicit assumption in the method, and since this is the case, this and other restrictions must be studied. Therefore, to have an optimized model that is the most appropriate, a trial-error method must be performed (33). Since this method uses different constraints like adding the profiles of the components this is an implicit assumption (33). Th Consequentially, the constraints that are more important in the multivariate calibration resolution alternating least squares (MCR-ALS) are the non-negative, equality, and closure. Non-negative means that negative values for C and/or S are not accepted when this model is applied. The equality constraint is imposed when there is knowledge (total or partial) about C or S, and the closure constraint limits the total concentration of the constituents (typically assumes that the total mass fraction of each sample is 1) (33).

The validation of the methods can be made by resourcing to the equations provided below (34). As with any chemometric method, an independent validation data set should be employed to appropriately validate the results. The validation is called the prediction (P) dataset.

Root mean square error of prediction (RMSEP):

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{n}(c_i - c1_i)^2}{n}} \quad \text{Equation 3}$$

Bias:

$$bias = \frac{\sum_{i=1}^{n}(c_i - c1_i)}{n} \quad \text{Equation 4}$$

Standard error of prediction (SEP):

$$SEP = \sqrt{\frac{\sum_{i=1}^{n}(c_i - c1_i - bias)^2}{n-1}} \quad \text{Equation 5}$$

Relative percentage error in the concentration predictions (RE, %):

$$RE(\%) = 100\sqrt{\frac{\sum_{i=1}^{n}(c_i - c1_i)^2}{\sum_{i=1}^{n} c_i^2}} \quad \text{Equation 6}$$

The c_i and $c1_i$ are the known and predicted analyte concentration in the sample i, respectively, and the n is the total number of samples used in the validation set.

2.6. Literature Review

In the review of *Fakayode et al.*, 2020 it is said that Raman, NIR, and FTIR spectroscopies have been thoroughly applied to analytical method development, instrumentation calibration, chemical analysis, and quality control and assurance of consumable products especially since 2007 (35). *Capková et al.*, 2022, have applied reverse engineering in tablets using Raman and chemometrics. This paper investigated a manufacturing technology to obtain tablets and the particle size of the incoming API, and also the quantitative composition of each excipient was analyzed. The authors, through the use of chemometric methods, were able to identify the composition of the tablet, and based on their experiments and their analysis it can be applied in reverse engineering (36). Also, *Shafaq* and collaborators used Raman spectroscopy to do a quantitative analysis of solid dosage forms and use chemometrics tools the determination the concentration of the API (37). Different reviews are available for Raman, FTIR, and other spectroscopies. *Deidda* and colleagues published some data about vibrational spectroscopy and the utility of this method in the analysis of pharmaceutical materials (38).

ATR-FTIR spectroscopy that had been used in the pharmaceutical field and chemometrics is used to complement this technique. *Custers et al., 2015* use the ATR-FTIR to describe how this can be useful for the screening of counterfeit medicines, so the author explains how this tool can help the customs to obtain the first evaluation of suspected samples. Custers used as a chemometric tool by the PCA to evaluate if the technique can result in a clustering of samples this can be useful for the creation of classification models (39). Other examples of the application of FTIR in the pharmaceutical fields are in the article of *Mallah et al, 2015*. These authors used the FTIR spectroscopy method for the quantification of paracetamol in a solid pharmaceutical formulation. The proposed of this work is to evaluate a simple model that uses Beer's law calibration and used also a more common method the PLS a popular multivariate model. The two here are used to analyze the solid pharmaceutical samples. The results of this work were accurate according to the permissible limits of the pharmacopeia (40). In the review of *Verma et al, 2021*, FTIR is used in combination with a chemometric method in the quantitative approach. So here this technique is used for the quantitative study of varieties of analytes like API, adulterants, caffeine, cocaine, lipids, fats and oils, sugar, and others. In this study was possible to conclude that FTIR and chemometrics are beneficial methodologies for the quantitative study of the substances analyzed in the work (41).

3. Materials and Methods

3.1. Active Substances

3.1.1. Paracetamol

Paracetamol is used to treat moderate pain, like headaches, menstrual periods, toothaches, backaches, and others, and is also used to reduce fever. This comes as a tablet, chewable tablet, capsule, suspension, or solution (liquid), extended-release tablet, and orally disintegrating tablet (42). It is a non-opioid analgesic and antipyretic agent, this can be used in combination with aspirin and caffeine (43).

3.1.2. Caffeine

Caffeine is a methylxanthine alkaloid found in coffee, tea, cola, cocoa, guarana, yerba mate and other products, can be administrated topically, orally, inhalation, or by injection. This substance is used in beverages, cardiac and respiratory stimulants, diuretics, cosmetics, can also be used pain relief and to combat drowsiness (44,45).

3.2. Excipients

3.2.1. Starch

Starch is odorless and tasteless, fine, white to off-white powder. Consist of very small spherical or ovoid granules or grains. The functional category is tablet and capsule diluent; tablet and capsule disintegrant; tablet binder; thickening agent (46). If used as diluent, ant adherent and lubricant the quantities are between 3-10% and when used as a disintegrant the concentrations can be in the range of 3-25%, and a typical concentration is 15% (46).

3.2.2. Talc

Talc is a fine, white to grayish-white, odorless, impalpable, unctuous, crystalline powder. The functional category of this excipient: anticaking agent, glidant, tablet and capsule diluent, tablet, and capsule lubricant (46). In oral solid formulations is widely used as a lubricant and diluent. If used as a glidant and tablet lubrification the concentration can vary between 1-10% and if used as diluent in a tablet or capsule the variation in concentration is in the range of 5-30% (46).

3.2.3. Microcrystalline Cellulose

Microcrystalline cellulose (MCC) is a purified, partially depolymerized cellulose that occurs as a white, odorless, tasteless, crystalline powder of porous particles. The functional category is absorbent; suspending agent; tablet and capsule diluent; tablet disintegrant (46). In pharmaceuticals it is used as a binder/diluent in oral tablets with a concentration between 20-90% and of capsules between 20-90% and can also be used in tables as a disintegrant with a concentration in the range of 5-15% (46).

3.2.4. Magnesium Stearate

Magnesium Stearate (MgS) is fine, light white, precipitated or milled, impalpable powder of low bulk density, having a faint odor of stearic acid and a characteristic taste. It can be used in tablet and capsule as a lubricant (46). This is used in cosmetics, foods and pharmaceutical formulations and is mainly used as a lubricant in tablets and capsules with a concentration in the range of 0.25-5% (46).

3.2.5. Lactose Monohydrate

Lactose Monohydrate occurs as white to off-white crystalline particles or powder and is odorless. The functional category of lactose is a dry powder inhaler carrier; lyophilization aid; tablet binder; tablet and capsule diluent; and tablet and capsule filler (46).

3.3. Formulations

3.3.1. Experimental Design

It was generated the calibration and validation formulations through an experimental design. For the calibration and testing formulation sets, it was selected a *D-optimal* design in the software that used. For each active/excipient, it was defined a range of potential compositions (typical compositions for each excipient) and used to produce the experimental designs. To build the experiments, MODDE® software (Sartorius Data Analytics) was used. A total of 23 formulations for calibration (Table 1) and 13 formulations for testing the models were designed (Table 2).

The DoE was used to evaluate the performance of the RE IR-based method with the MCR method. MCR requires calibration and validation, which is why it is necessary to produce the formulations using the DoE. On the other hand, the calibration-free method only needs pure spectra and does not require calibration and validation. This method was used to estimate the composition of some of the DoE powder mixtures.

The experimental design considered seven formulation components. Paracetamol was fixed at 20% w/w, and lactose was considered the filler. The remaining components' mass fractions were defined as described below

- Caffeine: 0.5%, 3% or 6%.
- MCC: 1%, 10% or 20%.
- Starch: 1%, 5% or 10%.
- Magnesium Stearate: 0.5%, 3% or 6%.
- Talc: 0.5%, 3% or 6%.

Table 1: Calibration powder mixtures.

Samples	**Paracetamol (mg)**	**Caffeine (mg)**	**MCC (mg)**	**Starch (mg)**	**MgS (mg)**	**Talc (mg)**	**Lactose (mg)**
1	0.20	0.010	0.10	0.050	0.0050	0.0050	0.63
2	0.20	0.010	0.20	0.010	0.030	0.06	0.49
3	0.20	0.0010	0.010	0.10	0.0050	0.060	0.62
4	0.20	0.0010	0.20	0.010	0.060	0.0050	0.52
5	0.20	0.050	0.20	0.10	0.060	0.0050	0.36
6	0.20	0.050	0.20	0.010	0.0050	0.0050	0.53
7	0.20	0.050	0.010	0.010	0.060	0.0050	0.67
8	0.20	0.0010	0.20	0.10	0.0050	0.0050	0.489
9	0.20	0.0010	0.20	0.010	0.0050	0.060	0.52
10	0.20	0.010	0.010	0.10	0.060	0.030	0.59
11	0.20	0.0010	0.010	0.10	0.060	0.0050	0.62
12	0.20	0.050	0.010	0.10	0.0050	0.0050	0.63
13	0.2	0.0010	0.010	0.010	0.0050	0.0050	0.77
14	0.20	0.050	0.20	0.10	0.0050	0.060	0.39
15	0.20	0.010	0.10	0.050	0.030	0.030	0.58
16	0.200	0.050	0.010	0.010	0.0050	0.060	0.67
17	0.20	0.050	0.010	0.10	0.060	0.060	0.52
18	0.20	0.050	0.20	0.010	0.060	0.060	0.42
19	0.20	0.010	0.10	0.050	0.030	0.030	0.58
20	0.20	0.0010	0.010	0.010	0.060	0.060	0.66
21	0.20	0.010	0.10	0.050	0.030	0.030	0.58
22	0.20	0.0010	0.20	0.10	0.060	0.060	0.38
23	0.20	0.050	0.10	0.050	0.030	0.030	0.54

Table 2: Test powder mixtures.

Samples	Paracetamol (mg)	Caffeine (mg)	MCC (mg)	Starch (mg)	MgS (mg)	Talc (mg)	Lactose (mg)
1	0.2	0.002	0.02	0.08	0.05	0.01	0.638
2	0.2	0.002	0.15	0.02	0.05	0.05	0.528
3	0.2	0.021	0.085	0.05	0.03	0.03	0.584
4	0.2	0.04	0.02	0.08	0.01	0.01	0.64
5	0.2	0.04	0.15	0.02	0.01	0.05	0.53
6	0.2	0.021	0.085	0.05	0.03	0.03	0.584
7	0.2	0.002	0.02	0.02	0.01	0.01	0.738
8	0.2	0.04	0.15	0.08	0.05	0.05	0.43
9	0.2	0.002	0.02	0.08	0.01	0.05	0.638
10	0.2	0.021	0.085	0.05	0.03	0.03	0.584
11	0.2	0.002	0.15	0.08	0.01	0.01	0.548
12	0.2	0.04	0.15	0.02	0.05	0.01	0.53
13	0.2	0.04	0.02	0.02	0.05	0.05	0.62

Each formulation (a powder mix) was made using a mixer from Fisher-Kendall (lab scale orbital mixer). Each formulation was prepared considered a total mass of 30g. All components were weighed and mixed in the same container (a plastic bottle of 150 mL). After that, the container was placed in an orbital mixer to mix for 7 minutes at 25 rpm.

3.4. FT-IR Spectral Measurements

The mixtures were analyzed in an FT-IR spectrometer, model Nicolet™ iS™ 5, from ThermoFisher Scientific. The measurement conditions are explained in Table 3. Before each scan, the ATR crystal was clean with isopropanol then, after a few seconds, the sample was placed on the crystal to perform the reading. Because the compounds used were solid, there was necessary to use the punch (to compress the samples against the crystal).

Table 3: Parameters and conditions used in the FT-IR spectral acquisition.

Parameters	Condition applied
Mode	Attenuated Total Reflectance
Accessory	iD5 ATR
Resolution	2 cm^{-1}
Scans	16
Spectral Range	4000 – 600 cm^{-1}
Replicates	Triplicates

3.5. Chemometric Analysis

3.5.1. Supervised Method (MCR)

The Chemometric analysis was done using MATLAB version 9.1.0.441655 (R2016b) (MathWorks, Massachusetts) software using the PLS Toolbox version 8.2.1. The spectra of the mixtures used in the calibration/testing were analyzed with and without pre-processing. Additionally, MCR models were attempted to provide the spectra of the pure compounds or not. The pure components spectra were pre-processed using baseline correction and absolute value. Four attempts were tested to verify the best pre-processing option (Table 4). It also was tested at different wavenumber intervals (Table 4). In Table 4 are the Models that will be analyzed in the results, Model 1, Model 2, Model 3, Model 4-A, Model 4-B, Model 4-C, and Model 4-D, these were the names given to differentiate each of the Models. The pre-processing was applied to the spectra of the samples and the pure spectra. In Model 4-A and 4-B, the same preprocessing was used, for Model 4-C and Model 4-D the order of polynomials was changed from 2nd to 3rd, and in Model 4-D the second pre-processing was changed from absolute value to normalization to unit area. The pre-processing applied in the pure spectra, that was provided to the methods, had the same pre-processing applied to each of the models. In Model 2 was not used any pre-processing, but was provided pure spectra, and for this, was used the correction of the baseline and the absolute value.

Table 4: Conditions applied in the MCR.

Model		Spectra of pure components	Pre-processing	Wavenumber Range (cm^{-1})
1		Not given	None	[3500-2750 1800-800]
2		Given	None	[3500-2750 1800-800]
3		Not Given	Baseline correction (2nd order) and absolute value	[3700-2320 1800-800]
4	A	Given	Baseline correction (2nd order) and absolute value	[3700-2320 1800-800]
	B		Baseline correction (2nd order) and absolute value	[3700-2320 1800-1100]
	C		Baseline correction (3rd order) and absolute value	[3700-2320 1800-800]
	D		Baseline correction (3rd order) and normalization by unit area	[3700-2320 1800-800]

The pre-processing used in the different models was different. The identification of the best pre-processing for the method was done by trial and error. So, first, the baseline (Automatic Weighted Least Square) pre-processing was tried, which is predefined to use the 2nd order polynomial, then the 1st and 3rd order polynomials were also used to see if any significant change in the results occurred. Following this pre-processing, absolute value was used, since the spectra obtained had values below zero, and one of the restrictions of MCR is non-negativity. Finally, the absolute value was replaced by normalization to the unit area, so that all samples had an equal impact on the model. This last pre-processing was used so that a comparison of the two methods with the same pre-processing could be made since the pre-processing used in the second method is the baseline and then normalization to unit area.

After obtaining the estimated mass, scores (C), a linear correction were performed, since the components may not all absorb in the same way. Once the values were corrected, these were used for the analysis of the MCR method.

3.5.2. Calibration-free Method

The calibration free-method algorithm is fully described in Annex 1.

In summary, the algorithm picks the FT-IR spectra of the active substances and the excipients and then resources to Lambert Beer's law, similarly to MCR. However, the idea was to use a bootstrapping strategy to try to reach the "real" composition of some powder mix samples. The algorithm received the target powder mix of unknown quantitative composition and tried to reconstruct it using pure spectral profiles. This reconstruction was performed by combining the pure spectral profiles using the Lambert-Beer law. The concentrations were generated randomly for each constituent from a range provided by the user. For some components, that concentration may be fixed (paracetamol for example), and for other components, the user can postulate some possible mass fraction ranges. This is what the formulator in the generic drug product development will do. One starts from previous knowledge. There is normally an admissible range of concentration for each formulation component that can be used.

The algorithm generates the spectrum of different potential mixtures picking values randomly from within the different ranges set for each compound. Typically, the algorithm will perform this process 10 000 times. Each spectrum is then compared to the spectrum of the target product, and a distance is calculated based on some metric (Euclidean distance in this case). The different simulations are then ranked, and the n simulations that best match the reference product are selected. The average and standard deviation calculated for each component considering the n best simulations gives the result of the algorithm. The average value gives the estimation of the concentration (mass fraction) of that component, while the standard deviation is a measure of the uncertainty in the estimation of that quantity. A confidence interval for the predicted mass fractions can be obtained by calculating the average +/- 2 standard deviations (approximately a 95% confidence interval). Before applying the Lamber Beer law, the algorithm uses two spectral pre-processing methods: baseline removal and normalization to unit area.

The script used to run the algorithm is provided in Annex 2. In the Figure 3 is described the algorithm through a scheme.

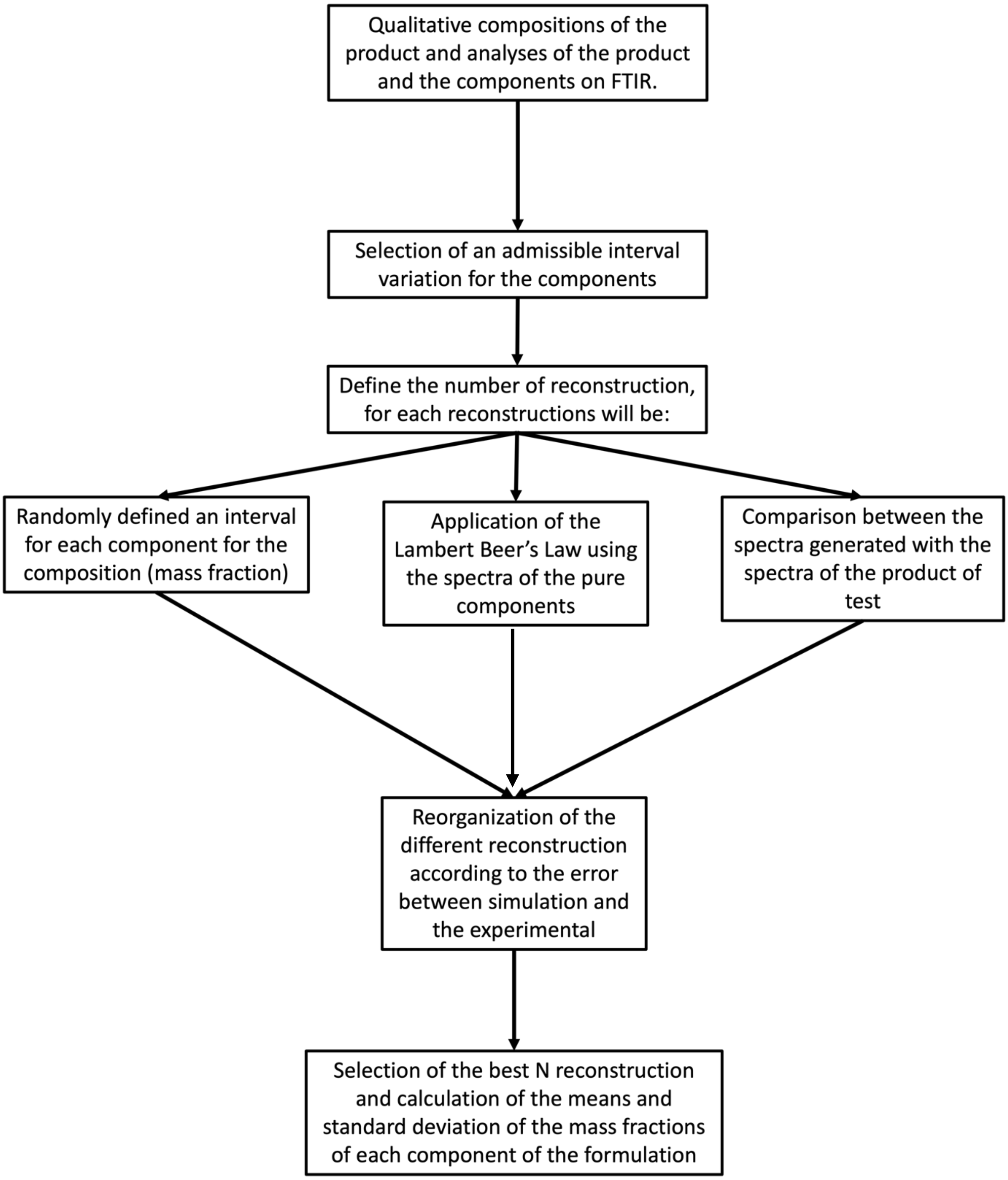
Qualitative compositions of the product and analyses of the product and the components on FTIR.
Selection of an admissible interval variation for the components
Define the number of reconstruction, for each reconstructions will be:
Randomly defined an interval for each component for the composition (mass fraction)
Application of the Lambert Beer's Law using the spectra of the pure components
Comparison between the spectra generated with the spectra of the product of test
Reorganization of the different reconstruction according to the error between simulation and the experimental
Selection of the best N reconstruction and calculation of the means and standard deviation of the mass fractions of each component of the formulation

Figure 3: Algorithm application scheme.

The application of the algorithm went through different stages an initial stage in which the algorithm was run so that it only performed the analysis for one sample at a time, and a second stage in which the algorithm analyzed all samples sequentially. Table 5 shows the wavelength intervals considered in the analyses. The algorithm considering was run with different margins for each compound. We tested percentage variations around the known concentration of each compound from 0 (the best situation, that is fixed amount), 5, 25, 50, 75, 100 up to 1000% (worst scenario).

Table 5: Conditions applied to the algorithm.

Spectra components	Wavelength Range (cm^{-1})	Column Indexes
All components	[3700:2320 1800:800]	[1245:6970 9127:13275]
	[3700:2320 1800:1100]	[1245:6970 9127:12445]

The samples used to test the algorithm are in Table 6. These samples were chosen to have the maximum, minimum, and intermedium values of most compounds. All samples (Table 1) were also used to analyze the precision and accuracy of the algorithm.

Table 6: Concentrations of each component of the samples used to apply the algorithm.

	Sample 6 (mg)	Sample 11 (mg)	Sample 19 (mg)
Starch	0.011	0.100	0.052
Caffeine	0.051	0.001	0.010
MgS	0.005	0.063	0.030
Lactose	0.536	0.626	0.578
Paracetamol	0.202	0.202	0.201
Talc	0.005	0,006	0.030
MCC	0.199	0.010	0.101

4. Results and Discussion

4.1 Pure Components Spectra

Figures 4 to 10 contain the pure spectrum of each component. Through a visual analysis is possible to see that the MCC, Starch, Lactose, and Talc have overlapping peaks in the range between 1300-800 cm^{-1}. This can be a problem when estimating the mass using MCR. But, for the Talc, that is not a problem since this component has another peak that will make it possible to distinguish it from the other components between the 3700-3500 cm^{-1}.

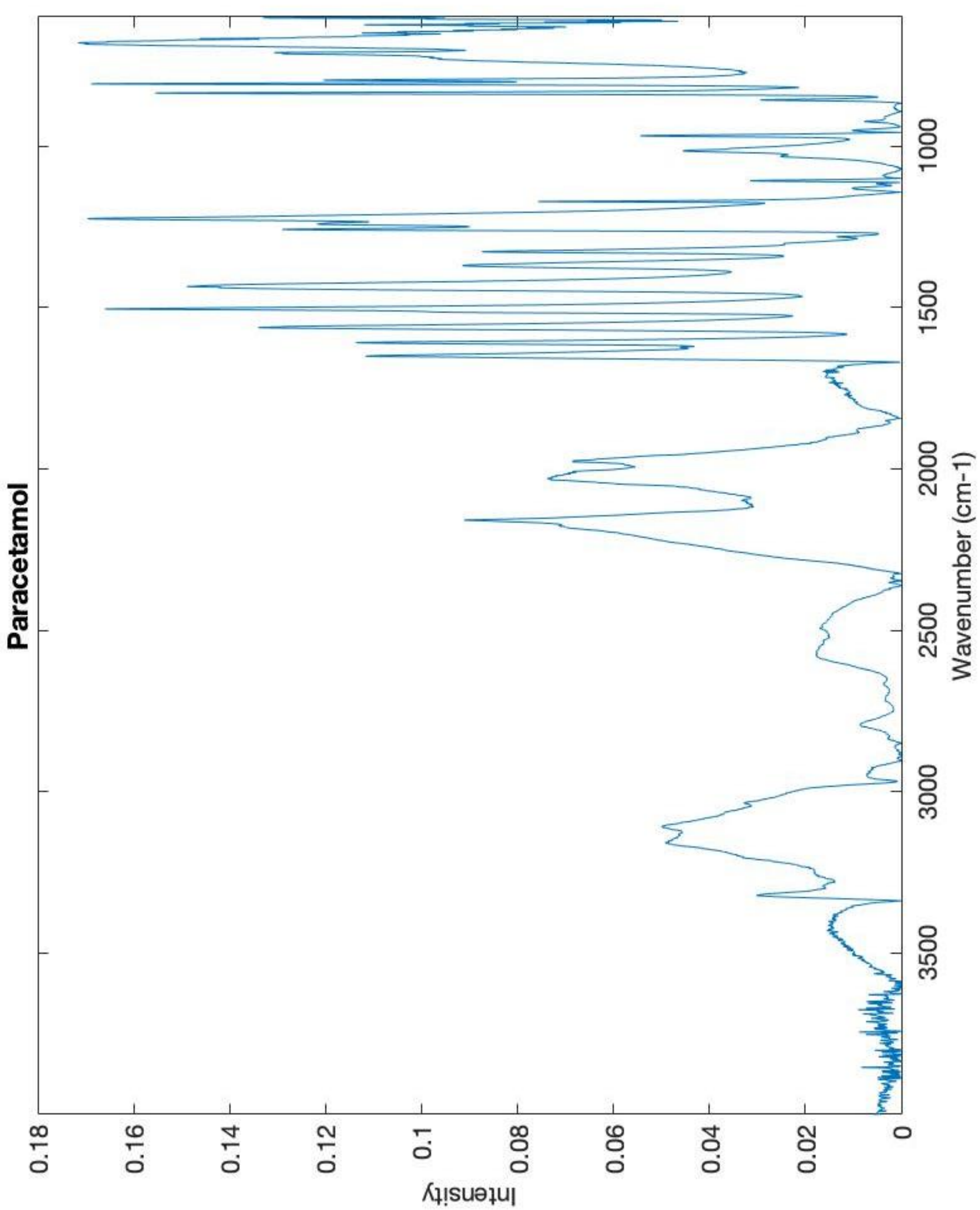

Figure 4: Pure spectrum of paracetamol.

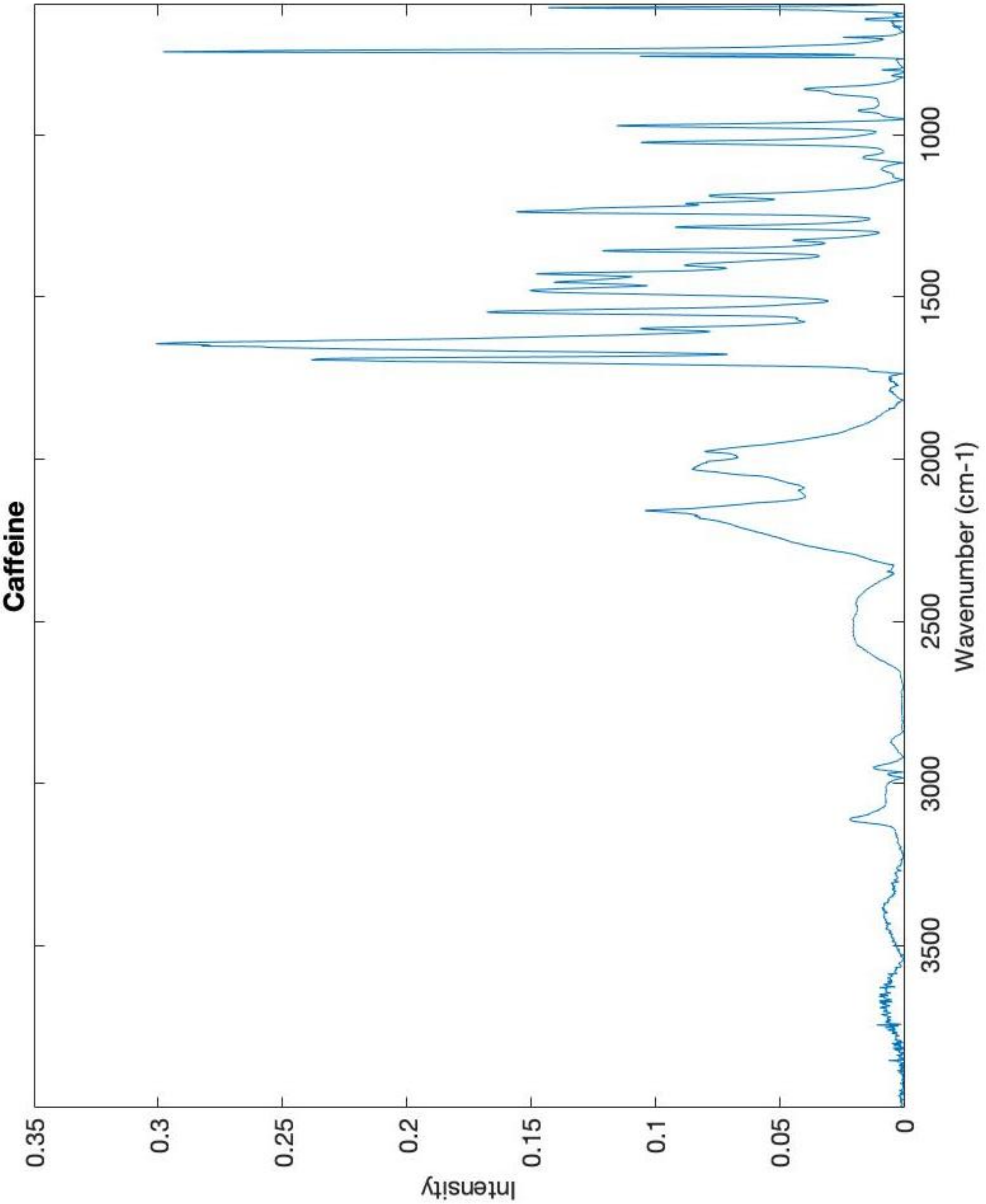

Figure 5: Pure spectrum of caffeine.

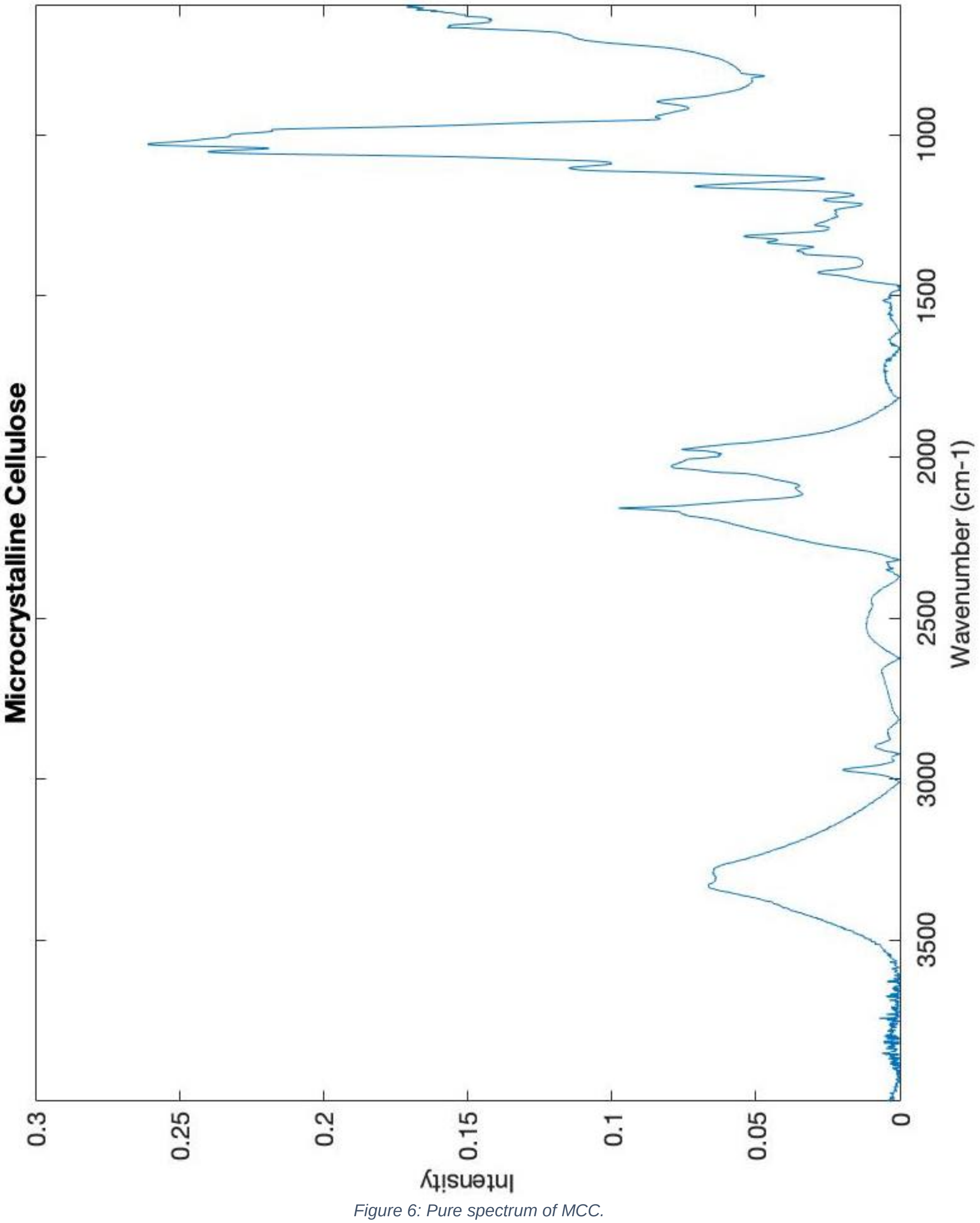

Figure 6: Pure spectrum of MCC.

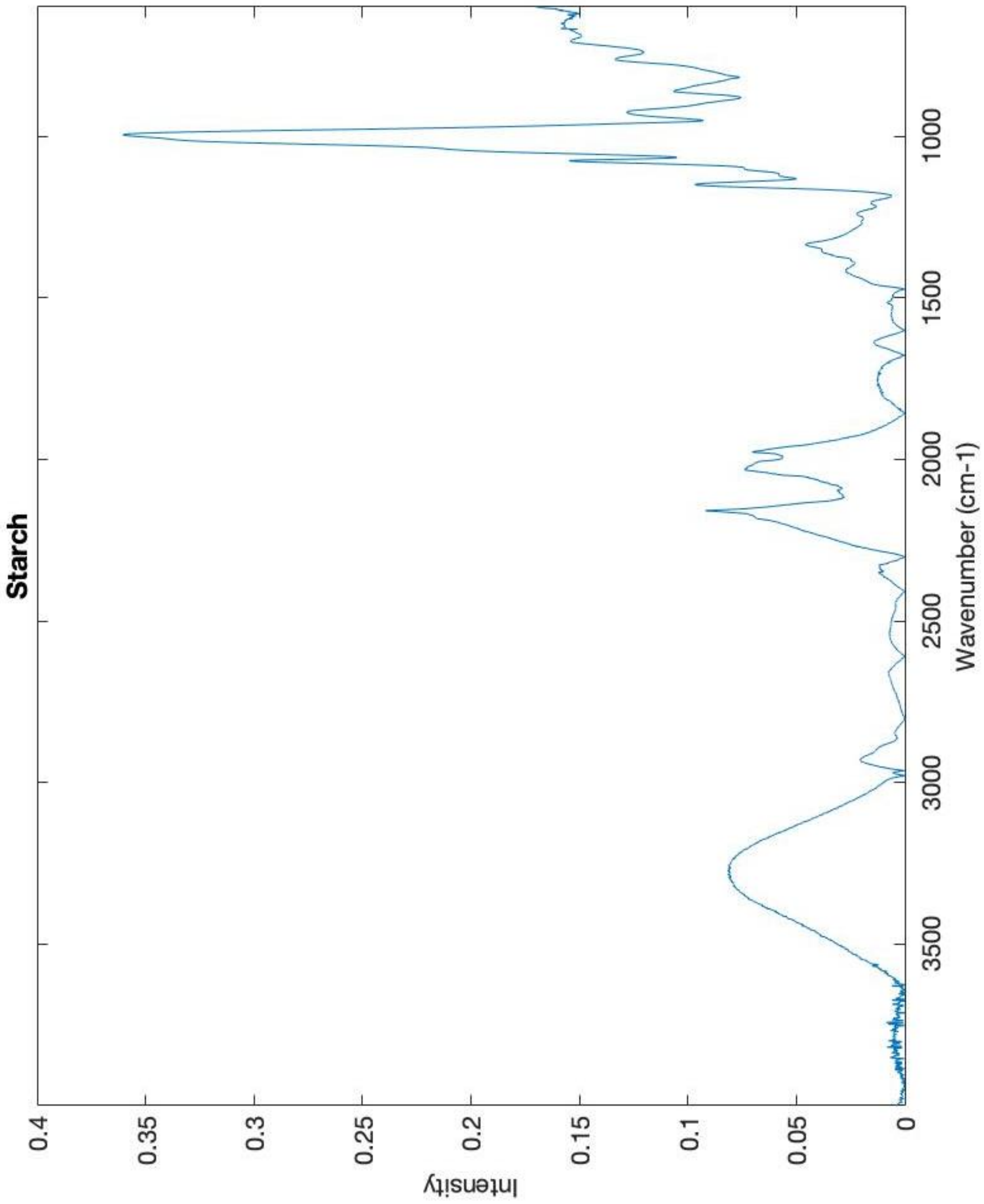

Figure 7: Pure spectrum of starch.

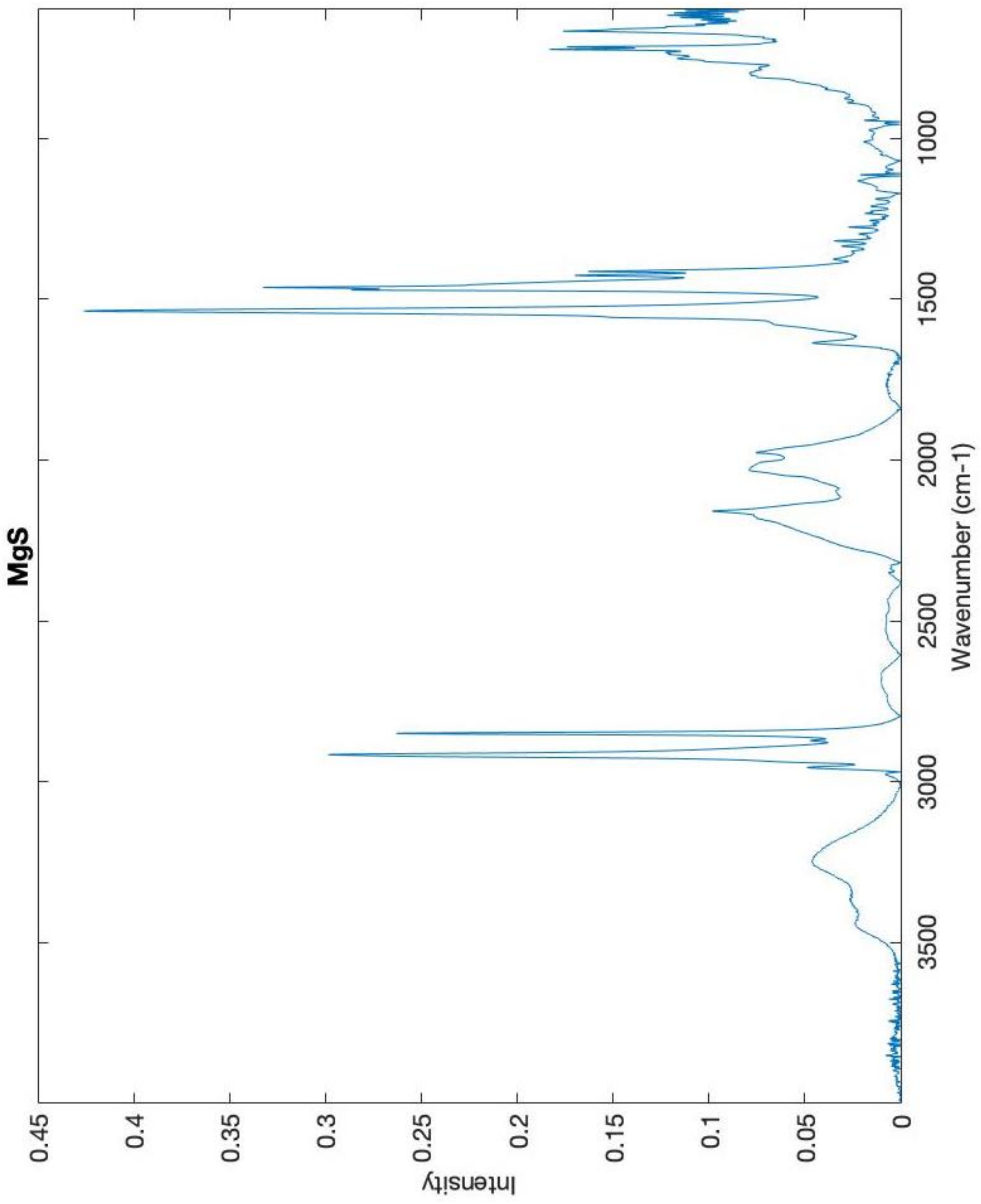

Figure 8: Pure spectrum of MgS.

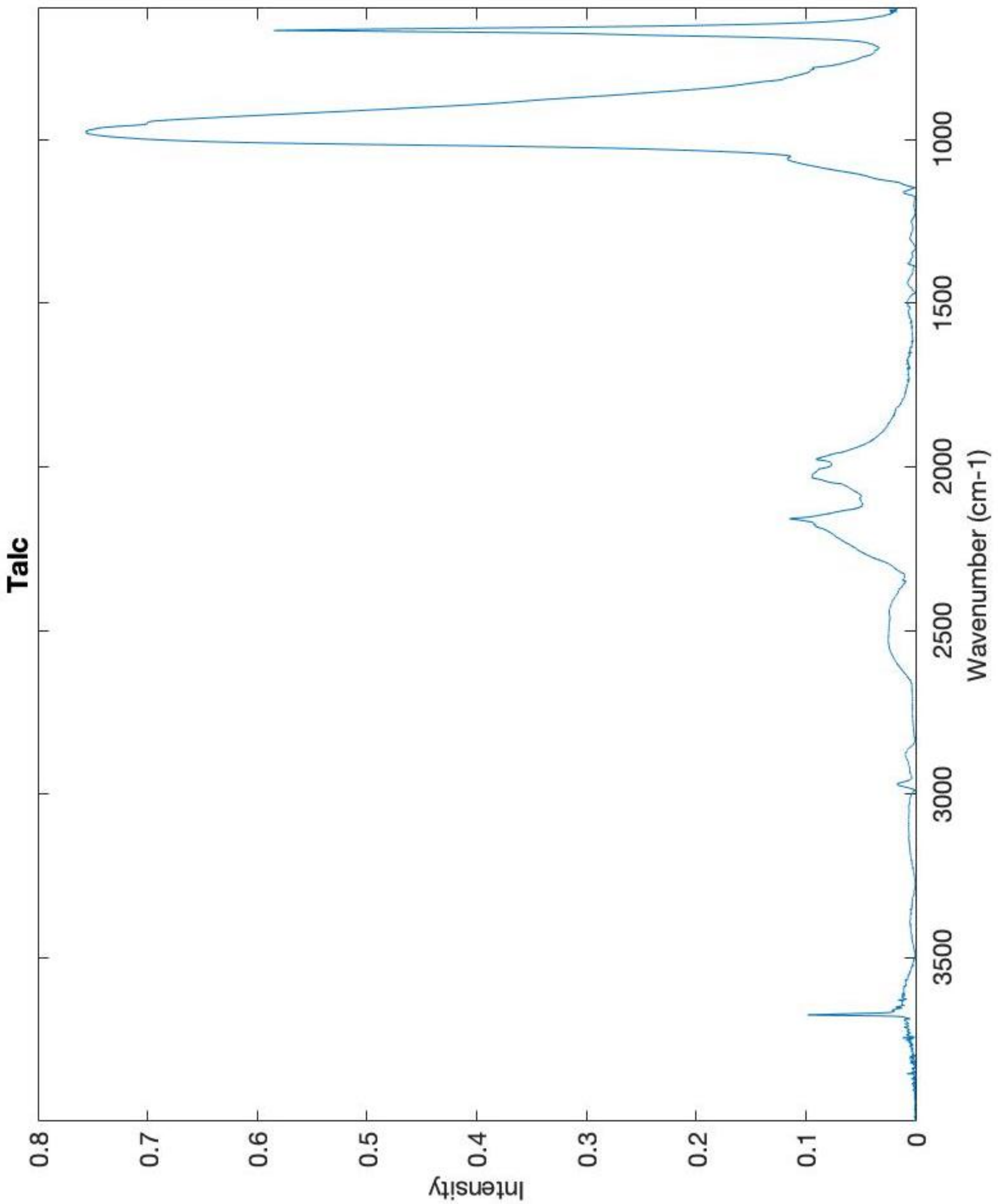

Figure 9: Pure spectrum of talc.

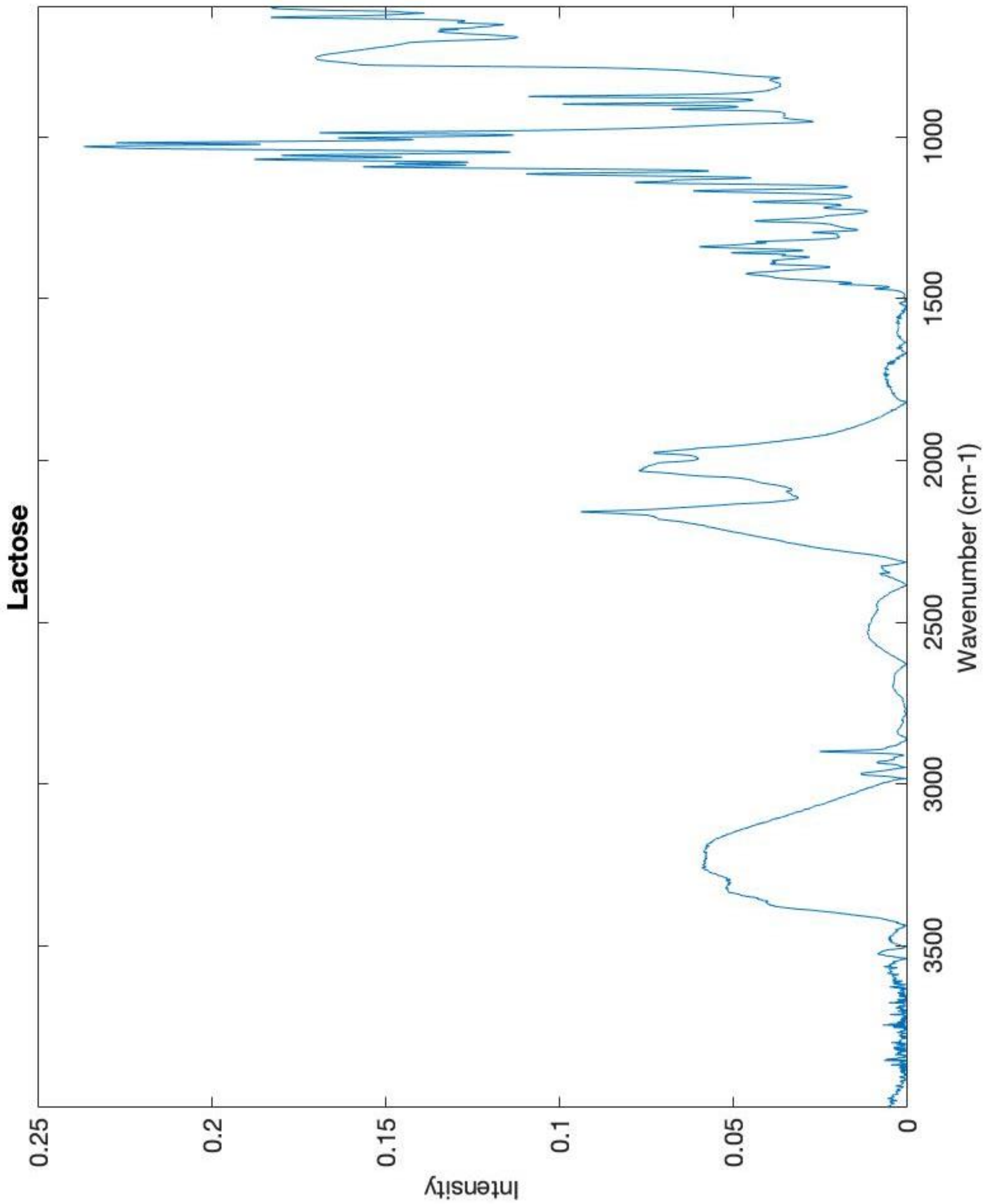

Figure 10: Pure spectrum of lactose.

4.2 Supervised Method (MCR)

The application of the MCR method started with an analysis of different models with conditions. For this supervised method, 7 different scenarios were analyzed. The differences in the models are presented in Table 4. The pre-processing was different, and the interval and pure spectra were given. For models 1 and 2, not all the conditions that were performed for model 4 were applied since the results of these were not acceptable, since these have higher residuals, as will be seen in Section 4.2.1.

Figure 10 shows all samples, and raw data, without pre-processing. Figure 12 shows the data already pre-processed with baseline correction (order 2nd) and absolute values, and Figure 13 shows the data pre-processed with baseline correction (order 3rd) and normalization by unit area.

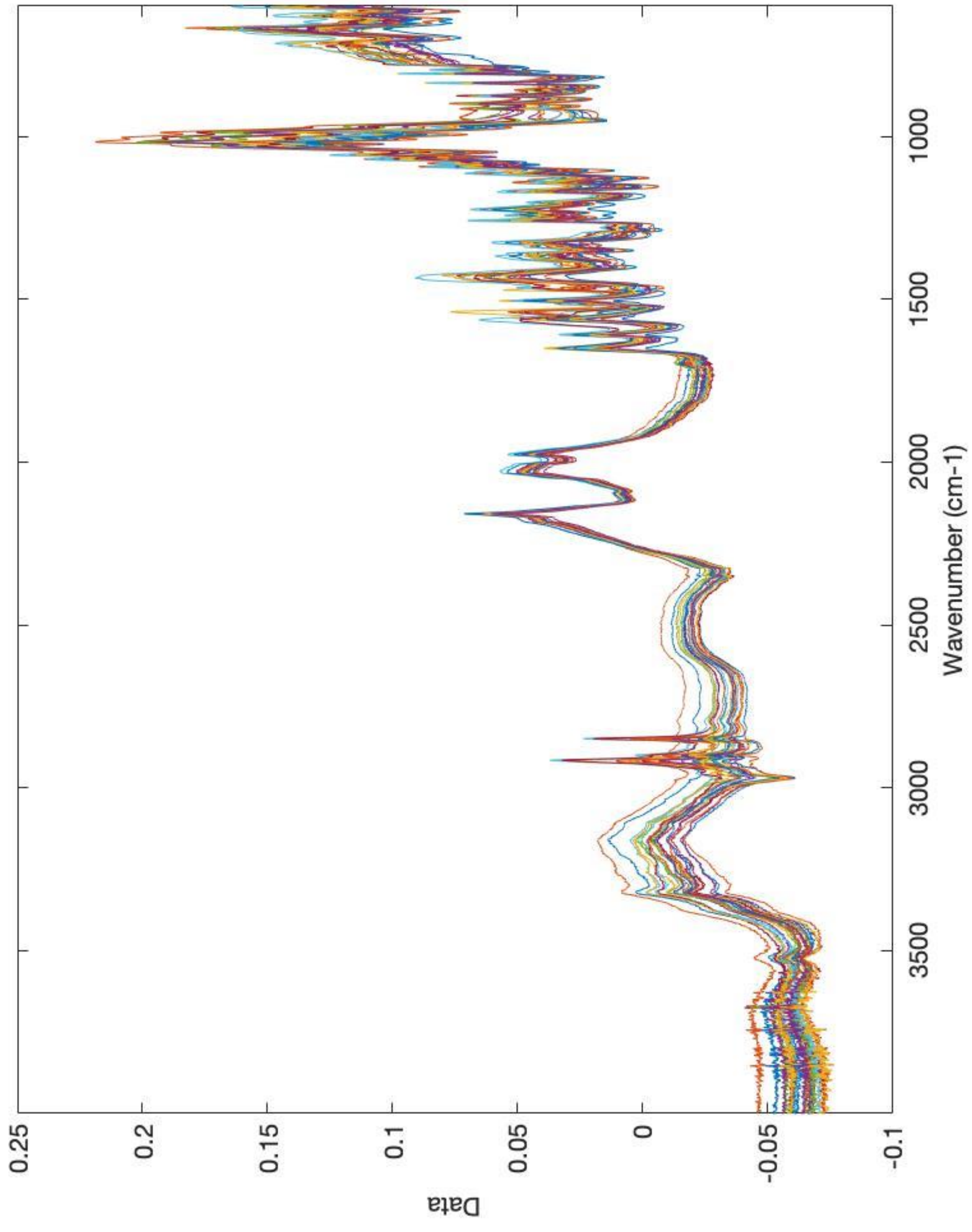

Figure 11: Calibration data without pre-processing.

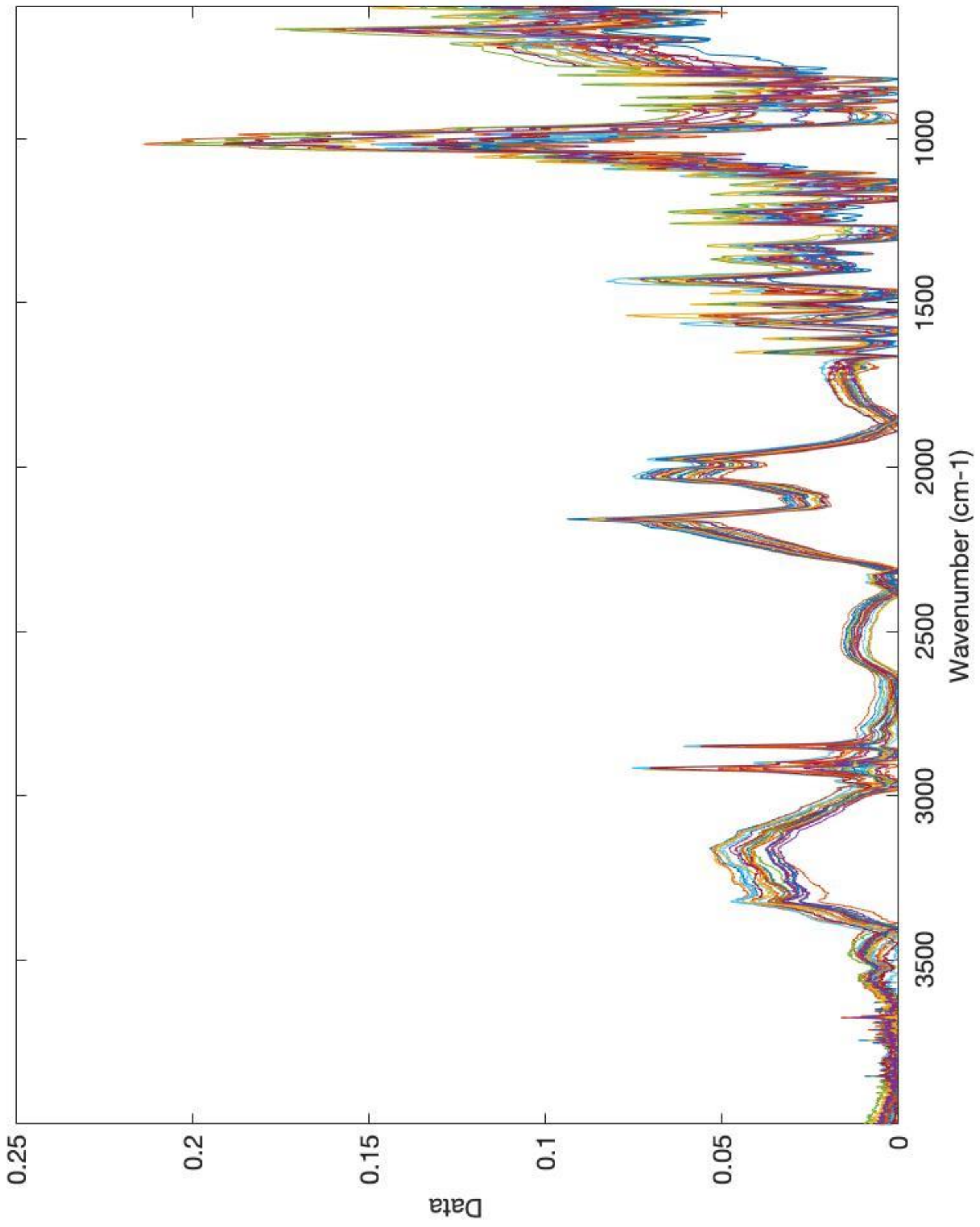

Figure 12: Calibration data with pre-processing (baseline correction and absolute value).

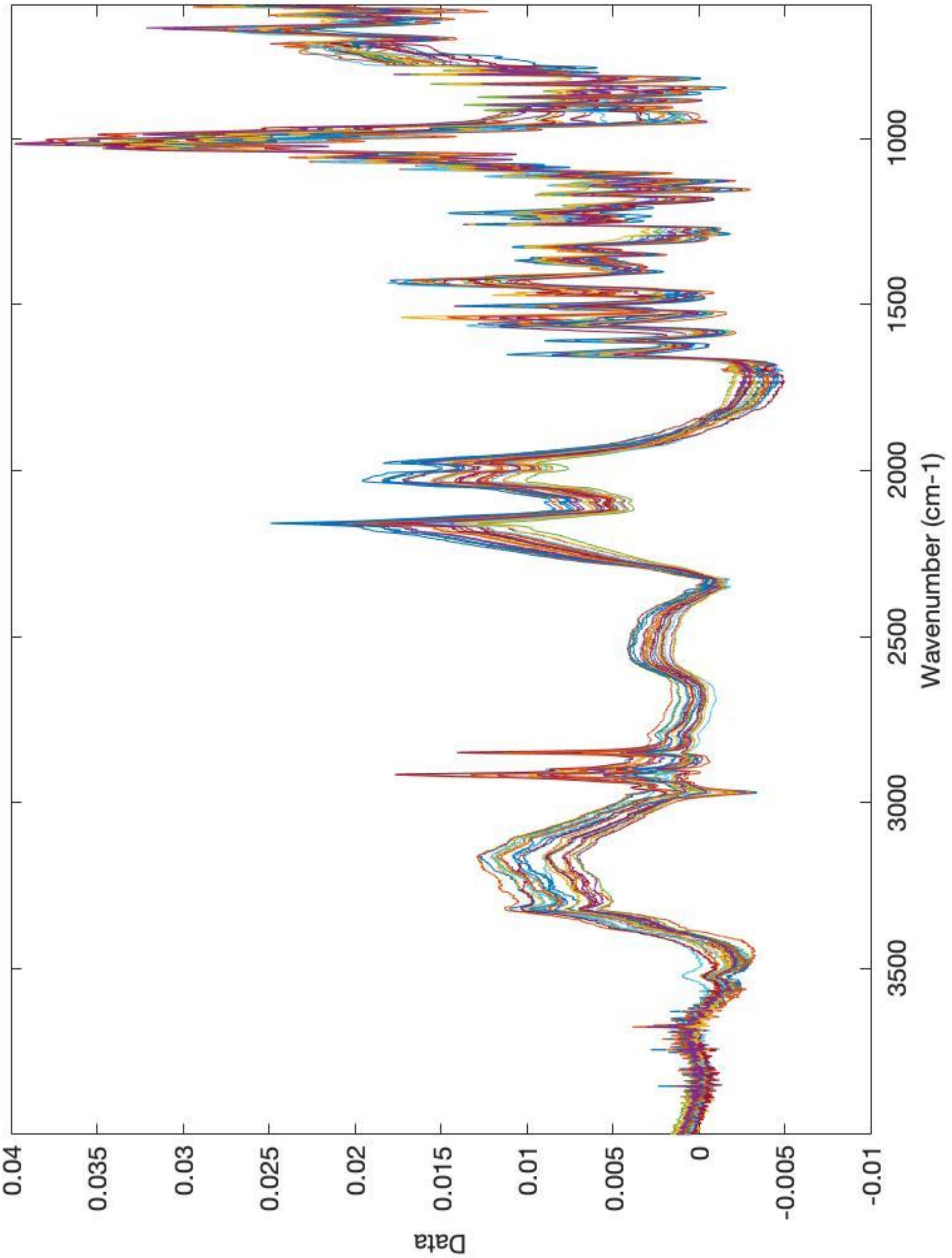

Figure 13: Calibration data with pre-processing (baseline correction and normalization to unit area).

4.2.1 Models analysis

Table 7 shows the residuals of all samples for each of the models. The residuals can be used to evaluate the accuracy of the models. The residual will assess the error between the sample reconstructions and the experimental spectra. So, if the values are high, it means that the error between these two is higher and that the value of these reconstructions is far away from the value of the experimental spectra. Of all the models, Models 1 and 2 are the ones with high residuals. Since that these models have not been used again for other conditions with other pre-processes or intervals.

Table 7: Q Residual (%) for each MCR Model.

Q Residuals (%)							
Samples	Model 1	Model 2	Model 3	Model 4-A	Model 4-B	Model 4-C	Model 4-D
1	3.3	11	0.042	1.9	0.37	0.42	0.15
1'	1.3	13	0.023	0.76	1.3	0.42	0.17
1''	1.9	13	0.02	0.72	1.8	0.33	0.17
2	3.0	16	0.29	0.52	0.92	1.0	0.21
3	3.1	14	0.063	1.3	1.8	0.63	0.17
4	2.4	9.1	0.21	1.3	3.8	0.77	0.33
5	2.3	8,2	0.035	2.7	1.5	0.64	0.45
6	3.9	13	0.095	2.6	0.52	0.92	0.18
7	3.3	11	0.18	1.8	0.6	0.51	0.21
8	2.5	14	0.035	0.42	1.1	0.43	0.16
9	4.4	21	0.013	0.32	0.44	1.0	0.15
10	3.1	8,4	0.028	1.4	1.9	0.30	0.22
11	4.1	11	0.026	1.4	1.2	0.37	0.22
12	3.1	17	0.070	0.54	0.30	0.97	0.17
13	4.0	17	0.047	0.13	0.45	0.37	0.16
14	4.7	20	0.040	1.4	0.21	1.4	0.15
15	5.9	33	0.067	4.4	1.9	2.0	0.17
16	6.2	37	0.011	3.4	1.3	1.9	0.18
17	4.1	13	0,027	0.58	0.82	0.71	0.19
18	5.1	22	0.10	2.4	0.27	1.2	0.20
19	4.3	17	0.055	0.21	0.75	0.64	0.16
20	4.9	15	0.10	0.35	1.3	0.54	0.19
21	4.0	21	0.013	2.9	0.52	1.0	0.18
22	4.6	18	0.047	0.68	0.78	0.92	0.20
23	5.1	12	0.15	1.3	0.83	0.32	0.18

The figures of merit were used to analyze the results obtained using the MCR method. For paracetamol, R2 was not considered since this compound always had the same composition. Thus, starting with Model 1, which does not have any type of processing, only the spectra of the samples, it is observed in Table 8 that the conditions used were not the most effective for the reconstruction or estimation of concentrations. Since the pure spectra were not given, the order of the scores provided by the MCR may not have been the one that corresponded to the actual compounds. A script was proposed to research the correspondence between the estimated loadings (pure spectra profiles) and the real compounds in the blend (Annex 3). The idea was to match one compound to each estimated loading.

In Table 8 are the figures of merit corresponding to Model 1. It is possible to see that the R2 is low for all components. This can also happen since the order used may have been the wrong order in the scores. The Q-Residual for this model was higher than 3 for all samples (Table 7). These high values mean that the model cannot accurately estimate the real concentrations.

Table 8: Figures of merit for the Model 1.

Calibration Set							
	Paracetamol	Caffeine	MCC	Starch	MgS	Talc	Lactose
RMSEC (mg)	0.0010	0.022	0.081	0.038	0.016	0.024	0.090
bias (mg)	-4.3E-05	3.5E-05	-3.1E-05	1.7E-05	-4.2E-06	-7.1E-06	-3.9E-05
SEP (mg)	0.00022	0.0023	0.034	0.0074	0.0013	0.0030	0.042
RE (%)	0.10	20	16	14	10	16	3.2
R^2	----	0.0019	0.000020	0.018	0.58	0.058	0.11
Validation Set							
	Paracetamol	Caffeine	MCC	Starch	MgS	Talc	Lactose
RMSEP (mg)	0.00091	0.017	0.060	0.029	0.0060	0.016	0.091
bias (mg)	-4.1E-05	5.6E-06	-2.0E-02	-5.7E-04	4.6E-04	4.3E-03	3.5E-02
SEP (mg)	0.00016	0.0011	0.088	0.0052	0.0016	0.015	0.10
RE (%)	0.12	22	20	15	5.5	14	4.3
R^2	-----	0.000	0.0081	0.088	0.90	0.16	0.057

Like Model 1, Model 2 (Table 9) also presents unsatisfactory results. In this model, unlike the first one, it was provided the component's pure spectra. So, the order of the scores was known and it can be evaluated through Table 9 that the coefficient of determinations did not improve. The residuals (Table 7) were also quite high for this model, with values above 8 for all samples, meaning that the conditions used for this model were not ideal. So next was performed different model to see if the residuals had improved, to have better residuals this has to be lower (closest to zero).

Table 9: Figures of merit for the Model 2.

Calibration Set							
	Paracetamol	Caffeine	MCC	Starch	MgS	Talc	Lactose
RMSEC (mg)	0.0010	0.017	0.052	0.039	0.022	0.017	0.081
bias (mg)	7.8E-06	-1.2E-05	3.5E-05	-4.6E-05	3.2E-05	-5.0E-06	4.3E-05
SEP (mg)	0.000035	0.0016	0.014	0.0079	0.0024	0.0014	0.033
RE (%)	0.099	16	10	14	15	11	2.9
R^2	----	0.40	0.58	0.00090	0.18	0.56	0.28
Validation Set							
	Paracetamol	Caffeine	MCC	Starch	MgS	Talc	Lactose
RMSEP (mg)	0.00084	0.016	0.064	0.027	0.018	0.021	0.071
bias (mg)	1.4E-04	-9.5E-04	-2.8E-02	-1.9E-03	3.7E-03	-1.5E-02	-2.9E-02
SEP (mg)	0.00053	0.0046	0.12	0.010	0.013	0.056	0.13
RE (%)	0.12	21	21	15	16	18	3.3
R^2	----	0.21	0.24	2.0E-16	4.0E-16	0.33	0.27

In the construction of Model 3, the pure spectra of the compounds were not provided, but the calibration and validation samples were pre-processed as described in Table 4. The procedure done in Model 1 was performed for this model. The script in Annex 2 was also used since the pure spectra were also not provided.

Table 10 shows the figures of merit for the Model 3 and what can be seen is that the R^2 is high for the MgS, but for the remaining the R2 is low. When comparing with Models 1 and 2, a better Q-Residual for all the samples was obtained. The problem with this model was not knowing the correct order of the compounds (scores), which led to an imprecise and inaccurate estimation of the concentration of the components. Due to this problem, no further scenarios were repeated with other pre-processing methods.

Table 10: Figures of merit for the Model 3.

Calibration Set							
	Paracetamol	Caffeine	MCC	Starch	MgS	Talc	Lactose
RMSEC (mg)	0.0010	0.020	0.076	0.038	0.010	0.022	0.095
bias (mg)	-4.8E-05	-4.9E-05	3.6E-05	2.5E-05	3.3E-05	3.1E-05	-1.6E-05
SEP (mg)	0.00025	0.0022	0.029	0.0073	0.00034	0.0023	0.046
RE (%)	0.099	18	15	14	6.5	14	3.4
R^2	----	0.21	0.12	0.030	0.84	0.23	0.010
Validation Set							
	Paracetamol	Caffeine	MCC	Starch	MgS	Talc	Lactose
RMSEP (mg)	0.00085	0.020	0.075	0.028	0.0085	0.014	0.074
bias (mg)	-2.3E-05	-6.9E-03	-3.8E-02	-5.1E-03	6.5E-03	8.1E-03	2.1E-02
SEP (mg)	0.000090	0.027	0.16	0.022	0.024	0.029	0.059
RE (%)	0.12	25	24	15	7.8	13	3.5
R^2	----	0.041	0.10	0.021	0.92	0.59	0.21

Finally, Model 4 was carried out, where pre-processing was applied to the samples and pure spectra, in Table 4 the types of pre-processing applied can be seen.

In Model 4-A, the baseline correction with the polynomial of the 2nd order was applied, and then the absolute value. Through the R^2 (Table 11) and the Q-Residual (Table 7) was possible to evaluate that this was the path to a better model. Although not all components had an acceptable R-square but were possible to observe an improvement when compared to the models previously performed. In this Model, the Q-Residual was lower than 3 for all the samples, and when compared with Model 3 the values in Model 4 were much higher. Which means that Model 4 can improve. That does not mean that Model 3 has a better estimation for the mass of the components since in the case of Model 4 the RMSEC is lower in almost all the components and the R2 is also higher. Also, the RE is lower for Model 4 when compared with Model 3, meaning that Model 4 is more precisel.

Table 11: Figures of merit for the Model 4-A

Calibration Set							
	Paracetamol	Caffeine	MCC	Starch	MgS	Talc	Lactose
RMSEC (mg)	0.0010	0.022	0.064	0.038	0.011	0.018	0.068
bias (mg)	2.9E-06	-2.2E-05	9.4E-06	-4.8E-05	2.0E-05	1.1E-05	9.9E-06
SEP (mg)	0.000010	0.0025	0.021	0.0076	0.00047	0.0015	0.023
RE (%)	0.10	20	12	14	6.9	12	2.4
R^2	----	0.030	0.38	0.033	0.81	0.50	0.50
Validation Set							
	Paracetamol	Caffeine	MCC	Starch	MgS	Talc	Lactose
RMSEC (mg)	0.00084	0.017	0.054	0.030	0.012	0.010	0.082
bias (mg)	1.1E-04	-6.9E-03	-2.1E-02	-7.1E-03	9.0E-03	2.1E-03	4.4E-02
SEP (mg)	0.00042	0.027	0.091	0.030	0.033	0.0075	0.14
RE (%)	0.12	22	18	16	11	8.4	3.9
R^2	----	0.33	0.25	0.035	0.90	0.69	0,19

For the next Model 4-B, Table 12, the interval used was changed to assess whether the estimation of caffeine and starch improved. The interval used was [3700-2320 1800-1100] (cm^{-1}). Since each component has a pure spectrum that distinguishes them, it should be possible to get good estimates for each, but when the peaks of different components overlap, getting good estimates is compromised. The most distinctive peak for starch is in the region of 1000 cm^{-1}. In this region, other components also have peaks which makes it difficult to estimate the mass of starch in the samples. For this reason, this Model 4-B was performed to evaluate whether there was any improvement in the estimation. Concerning the residuals (Table 4), when comparing the values of Model 4-B with the values of Model 4-A, an improvement in the values for each of the samples can be seen, since the values of Model 4-B in general are lower. So, in Table 12 are the figures of merit for this model. Analyzing Table 12, R^2 remains low for caffeine and starch. No improvement was observed for caffeine (Figure 3) either, the peaks of which were similar to those of paracetamol (Figure 2), which may make it difficult to estimate these quantities. For the remaining components, the estimation was better when compared to the previous models.

Table 12: Figures of merit of the Model 4-B

Calibration Set							
	Paracetamol	Caffeine	MCC	Starch	MgS	Talc	Lactose
RMSEC (mg)	0.0010	0.021	0.049	0.039	0.0072	0.012	0.065
bias (mg)	-1.5E-05	3.7E-05	2.8E-05	1.4E-05	4.7E-05	-6.4E-03	1.7E-05
SEP (mg)	0.000084	0.0021	0.012	0.0076	0.000028	0.033	0.021
RE (%)	0.10	19	9.4	14	4.7	8.0	2.3
R^2	----	0.084	0.63	0.00060	0.91	0.83	0.54
Validation Set							
	Paracetamol	Caffeine	MCC	Starch	MgS	Talc	Lactose
RMSEP (mg)	0.00089	0.017	0.039	0.029	0.0071	0.011	0.060
bias (mg)	6.8E-05	2.5E-03	2.0E-02	-8.8E-04	1.8E-03	-6.6E-03	-2.9E-02
SEP (mg)	0.00025	0.0084	0.069	0.0064	0.0066	0.025	0.12
RE (%)	0.12	22	13	15	6.4	9.7	2.8
R^2	----	0.098	0.71	0.15	0.87	0.74	0.59

The next two Models were Model 4-C and Model 4-D (Tables 13 and 14). In these two models, the differences were in the pre-processing (Table 4). Comparing these two models with the others (Model 4-A and Model 4-B), it was possible to observe the improvements, but not very significant. So, was not possible to solve the problem encountered in estimating the caffeine and starch concentrations.

This can mean that when using the MCR for estimating a similar composition of a drug, with paracetamol, caffeine, and starch, it is difficult to get a good estimate of all the components. Meaning that the MCR is not able to predict the right amounts for these components in the samples. Regarding the precision of the MCR, it can be accurate or precise when the pure spectra of the components are given, and when the pure spectra are not given, it becomes more complicated to estimate the concentrations, making it less precise.

Table 13: Figures of merit for the Model 4-C

Calibration Set							
	Paracetamol	Caffeine	MCC	Starch	MgS	Talc	Lactose
RMSEC (mg)	0.0010	0.019	0.035	0.037	0.0085	0.012	0.057
bias (mg)	-1.3E-05	2.1E-05	-3.7E-05	7.3E-06	1.6E-06	-8.6E-06	2.5E-05
SEP (mg)	0.000071	0.0018	0.0064	0.0070	0.00036	0.00072	0.016
RE (%)	0.099	18	6.7	14	5.6	7.6	2.0
R^2	----	0.24	0.82	0.079	0.88	0.79	0.65
Validation Set							
	Paracetamol	Caffeine	MCC	Starch	MgS	Talc	Lactose
RMSEP (mg)	0.00089	0.014	0.033	0.024	0.0054	0.013	0.047
bias (mg)	9.9E-05	-1.9E-03	4.1E-03	-5.3E-03	2.9E-03	5.8E-03	-9.7E-03
SEP (mg)	0.00037	0.0080	0.011	0.022	0.011	0.021	0.044
RE (%)	0.12	18	11	13	4.9	11	2.2
R^2	----	0.43	0.69	0.26	0.93	0.59	0.63

Table 14: Figures of merit for the Model 4-D

Calibration Set							
	Paracetamol	Caffeine	MCC	Starch	MgS	Talc	Lactose
RMSEC (mg)	0.0010	0.022	0.047	0.036	0.012	0.014	0.054
bias (mg)	1.0E-05	-4.6E-05	4.5E-05	-1.5E-05	-3.7E-05	3.0E-05	-2.4E-06
SEP (mg)	0.000048	0.0027	0.011	0.0067	0,00087	0.00086	0.015
RE (%)	0.10	20	9.0	13	7.5	9.3	1.9
R^2	----	0.0034	0.66	0.14	0.78	0.68	0.68
Validation Set							
	Paracetamol	Caffeine	MCC	Starch	MgS	Talc	Lactose
RMSEP (mg)	0.00090	0.017	0.043	0.034	0.0058	0.012	0.048
bias (mg)	3.5E-05	5.6E-06	7.1E-03	-2.2E-02	1.2E-03	7.9E-03	-4.2E-03
SEP (mg)	0.00013	0.0011	0.020	0.088	0.0045	0.029	0.024
RE (%)	0.12	22	14	18	5.3	11	2.3
R^2	----	0	0.45	0.14	0.90	0.69	0.58

4.3. Calibration-free Method

This method resources only on the pure spectra of the components. In the Section 4.1. are the spectra of the pure components (Figure 4-10). The intervals used for the analysis with the calibration-free algorithm (Section 4.3.1) are in Table 5. For the Section 4.3.2. the interval used was [3700-2320 1800-800] (cm^{-1}).

4.3.1. Individual Samples

Tables 15-17 are the Residuals obtained for the three samples used to test this algorithm. The Residuals represent the error between the experimental data and the estimated data. The three colors, red, yellow, and green mean if the values obtained were reasonable or not. If the error is less than 0.02, in green, that value is reasonable or acceptable; in yellow are the values between 0.02 and 0.05, and in red are the values higher than 0.05.

Different scenarios were carried out with different initial uncertainties for the compounds' mass fraction. As the uncertainty increases, the prediction error should normally increase. Since the interval that is being given to the concentrations is getting larger, makes it more difficult to estimate concentrations and these are further away from the experimental concentrations. Tables 15-17 show that in some cases the error increase, but in others, the error decreases with increased uncertainty. Figures 14 and 15 are the estimated masses and the experimental masses. Figure 14 is the estimation for the 5% scenario for sample 6 and it can be seen that the estimate is very close to the experimental concentrations. Figure 15 represents the 50% scenario, the estimated masses are already farther from the experimental concentrations.

Having a lower error for all or most all the scenarios means that the algorithm can estimate with some accuracy the concentrations of the compounds even when the uncertainty range for the concentrations is large.

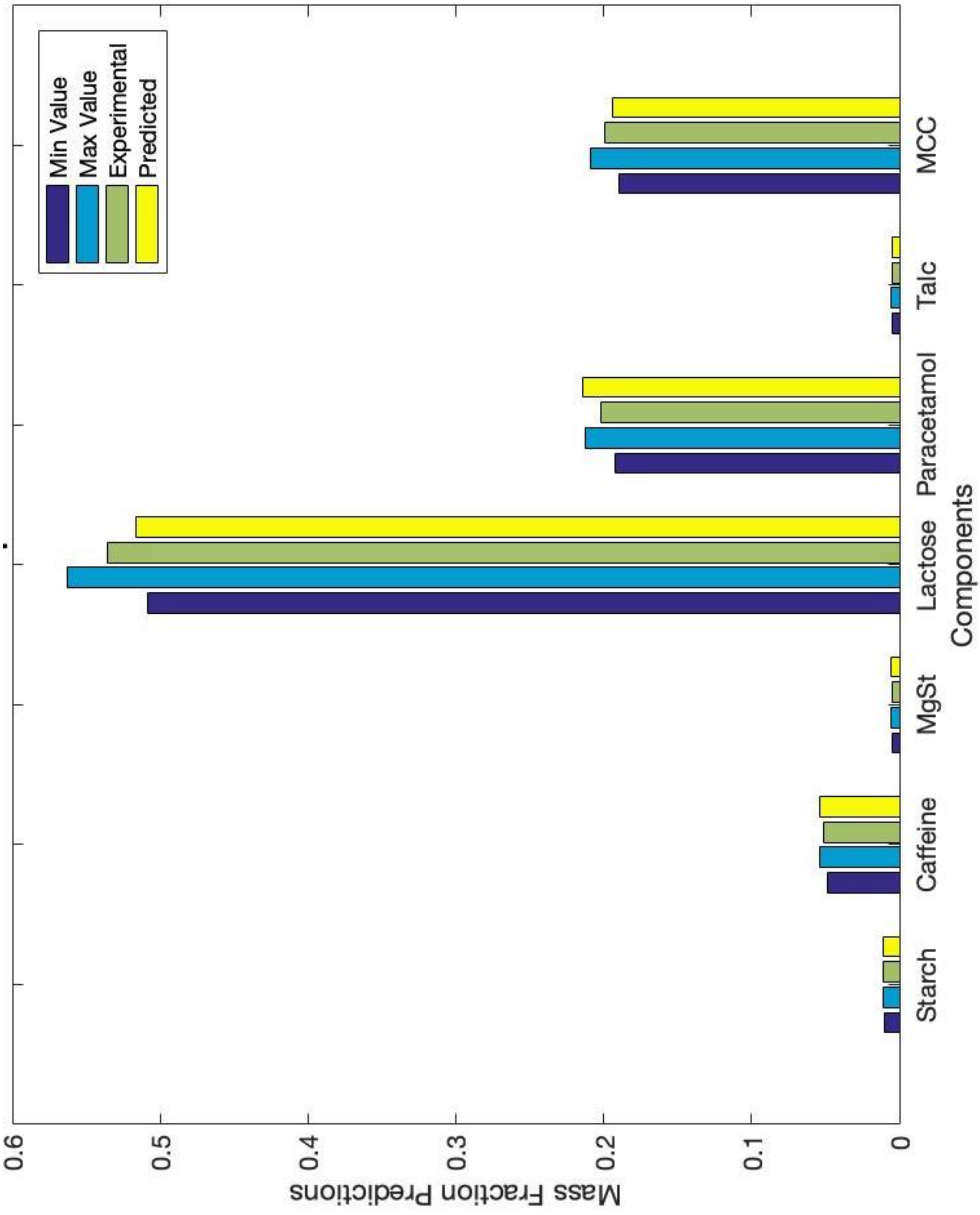

Figure 14: Comparation between the estimated and experimental masses for sample 6 (5% uncertainty scenario).

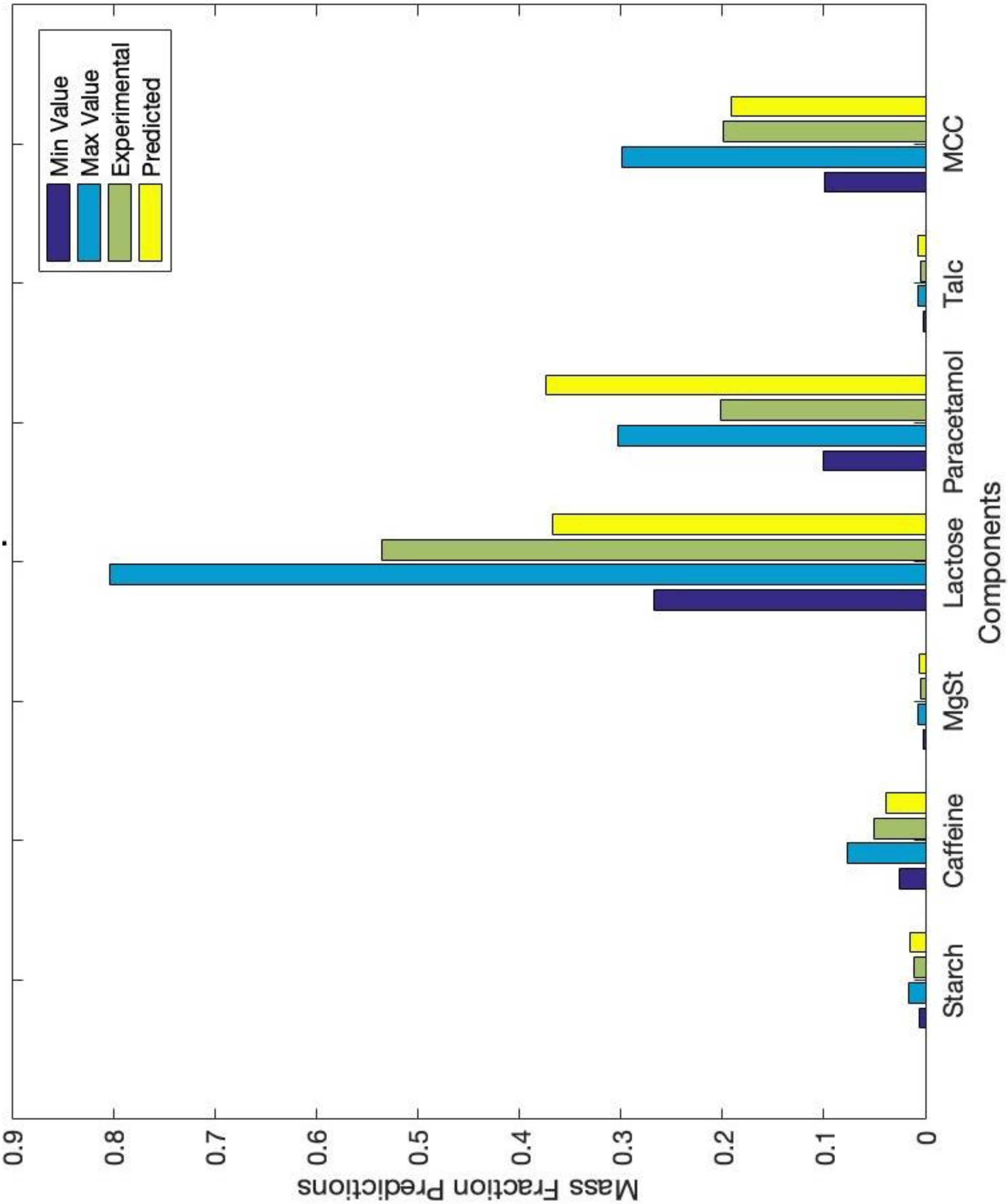

Figure 15: Comparation between the estimated and experimental masses for sample 6 (50% uncertainty scenario).

Table 15: Residuals of sample 6.

Interval 1

Percentage	Starch	Caffeine	MgS	Lactose	Paracetamol	Talc	MCC
0%	0	6.9E-18	0.00	0	0	0	0
5%	0.000080	0.0026	0.00015	0.019	0.012	0.000011	0.0052
25%	0.0000043	0.0029	0.00032	0.072	0.079	0.00092	0.021
50%	0.0045	0.012	0.00045	0.17	0.17	0.0027	0.0079
75%	0.00049	0.029	0.00029	0.23	0.22	0.00065	0.038
100%	0.0023	0.044	0.0030	0.24	0.22	0.0013	0.056
1000%	0.031	0.043	0.0046	0.27	0.21	0.00093	0.050
R²	0.98	0.38	0.73	0.33	0.21	0.0024	0.30

Interval 2

Percentage	Starch	Caffeine	MgS	Lactose	Paracetamol	Talc	MCC
0%	0	6.94E-18	0	0	0	0	0
5%	0.00015	0.0023	0.000035	0.018	0.012	0.00015	0.0065
25%	0.00073	0.0076	0.0012	0.079	0.081	0.00042	0.021
50%	0.0018	0.010	0.0014	0.16	0.18	0.0014	0.030
75%	0.0082	0.028	0.00013	0.20	0.22	0.00037	0.0073
100%	0.0054	0.047	0.0019	0.18	0.23	0.00066	0.013
1000%	0.014	0.040	0.00020	0.19	0.22	0.0030	0.010
R²	0.74	0.32	0.048	0.23	0.22	0.85	0.0053

Table 16: Residuals of sample 11.

Interval 1							
Percentage	Starch	Caffeine	MgS	Lactose	Paracetamol	Talc	MCC
0%	1.4E-17	0	0	0	2.8E-17	0	0
5%	0.0027	5.55E-07	0.0031	0.021	0.012	0.000010	0.000041
25%	0.0093	0.00014	0.018	0.093	0.075	0.00053	0.00094
50%	0.019	0.00052	0.052	0.21	0.17	0.00068	0.0038
75%	0.040	0.00059	0.062	0.24	0.21	0.00092	0.0043
100%	0.015	0.00019	0.055	0.25	0.21	0.0039	0.0066
1000%	0.054	0.00090	0.055	0.23	0.21	0.0016	0.017
R^2	0.63	0.59	0.17	0.18	0.22	0.069	0.90

Interval 2							
Percentage	Starch	Caffeine	MgS	Lactose	Paracetamol	Talc	MCC
0%	1.4E-17	0	0	0	2.8E-17	0	0
5%	0.0029	0.0000050	0.0033	0.020	0.012	0.00012	0.00026
25%	0.010	0.00019	0.020	0.09	0.074	0.00033	0.00069
50%	0.028	0.00034	0.048	0.20	0.17	0.00074	0.00062
75%	0.051	0.00042	0.060	0.23	0.21	0.00073	0.00039
100%	0.085	0.00027	0.060	0.19	0.22	0.0045	0.0036
1000%	0.068	0.0016	0.068	0.22	0.21	0.0021	0.00027
R^2	0.27	0.95	0.32	0.21	0.22	0.10	0.015

Table 17: Residuals of sample 19.

Interval 1

Percentage	Starch	Caffeine	MgS	Lactose	Paracetamol	Talc	MCC
0%	6.9E-18	0	0	0	2.8E-17	0	1.4E-17
5%	0.0032	0.00033	0.00014	0.012	0.0034	0.0017	0.0072
25%	0.021	0.0016	0.0016	0.066	0.0071	0.011	0.039
50%	0.049	0.0015	0.00057	0.14	0.0010	0.026	0.055
75%	0.085	0.0030	0.0070	0.21	0.012	0.016	0.10
100%	0.10	0.0039	0.022	0.21	0.042	0.015	0.072
1000%	0.16	0.0047	0.014	0.23	0.037	0.0080	0.053
R^2	0.68	0.50	0.20	0.33	0.40	0.0074	0.026

Interval 2

Percentage	Starch	Caffeine	MgS	Lactose	Paracetamol	Talc	MCC
0%	6.90E-18	0	0	0	2.8E-17	0	1.4E-17
5%	0.0025	0.00023	0.00062	0.018	0.012	0.00015	0.0012
25%	0.012	0.0016	0.0058	0.044	0.030	0.0038	0.010
50%	0.026	0.0042	0.010	0.062	0.043	0.0034	0.0086
75%	0.048	0.0042	0.011	0.044	0.040	0.0093	0.022
100%	0.062	0.0079	0.013	0.023	0.036	0.0045	0.051
1000%	0.12	0.0028	0.013	0.17	0.047	0.0058	0.018
R^2	0.80	0.0055	0.25	0.88	0.26	0.12	0.020

4.3.2. All Samples

In Section 4.3.1., only one sample was used to estimate the mass. So, in this section, all samples were used for mass estimation. The same analysis that was performed for just one sample (Tables 15-17) was done in this section with all samples. The analysis of the scenarios was performed through the average of the residuals (Table 18) and the average of the standard deviation (Table 19) obtained, corresponding to the average of the residuals obtained for each of the samples. In this case, it was used just the interval 1 (3700-2320 1800-800 cm^{-1}).

In Table 18, the algorithm was able to have a good estimate for all the compounds for the scenario of 0 and 5 %, and for the 25% was not able to have a good estimation for the Lactose, for the rest was made a reasonable estimation.

Table 18: Average of the residuals using all the samples.

				Interval 1			
Percentage	Starch	Caffeine	MgS	Lactose	Paracetamol	Talc	MCC
0%	3.6E-18	1.1E-18	1.1E-18	4.4E-18	8.9E-18	1.9E-18	4.4E-18
5%	0.0018	0.00048	0.0011	0.017	0.010	0.0013	0.0035
25%	0.013	0.0025	0.0062	0.070	0.041	0.0088	0.018
50%	0.028	0.0054	0.014	0.12	0.058	0.019	0.032
75%	0.047	0.011	0.020	0.15	0.063	0.021	0.045
100%	0.063	0.017	0.023	0.16	0.065	0.022	0.051
1000%	0.083	0.017	0.026	0.19	0.065	0.018	0.047
R²	0.57	0.41	0.38	0.38	0.18	0.12	0.23

Table 19: Stander deviation of the residuals for all the samples.

				Interval 1			
Percentage	Starch	Caffeine	MgS	Lactose	Paracetamol	Talc	MCC
0%	6.9E-18	2.8E-18	2.5E-18	1.5E-17	1.6E-17	3.5E-18	9.5E-18
5%	0.0023	0.00080	0.0014	0.0082	0.0087	0.0015	0.0050
25%	0.014	0.0036	0.0086	0.029	0.041	0.011	0.026
50%	0.033	0.0070	0.021	0.057	0.065	0.024	0.050
75%	0.060	0.013	0.034	0.084	0.071	0.032	0.066
100%	0.077	0.020	0.035	0.099	0.069	0.035	0.072
1000%	0.081	0.020	0.039	0.081	0.066	0.033	0.068
R²	0.39	0.43	0.35	0.18	0.15	0.22	0.22

The algorithm was run for the 50% scenario to do a more detailed evaluation since it can be observed that point is where the algorithm starts to have some difficulties in the estimation of the mass, and where the estimations begin to be more distant from the experimental masses. In the 50% scenario, it was assessing the mass estimation and the variation in the residuals. Figures 16-22 display the experimental mass and the estimated mass for each compound. It can be observed that the difference in the mass is higher in some samples. An example of that is samples 3 and 17 and others in Figure 16 (Starch), the samples 4, 7, and 10 in Figure 18 (MgS). Figures 19 and 20, Lactose and Paracetamol, have an estimation similar to the experimental mass. What can be seen when analyzing the Figures is that the lowest experimental concentrations were the easiest to estimate for the algorithm, while the highest concentrations were difficult. This may be happening since what is varying is the experimental concentration and if that concentration was greater the variation will be greater, for lower concentrations although the variation range is high the variation in concentrations will not be as high which makes it easier to estimate concentrations for some components. This can be seen for example in Figure 16 (Starch) for samples 7, 13, and 18, and also in Figure 17 (Caffeine) for samples 3, 4, 8, and others. The same can be observed for the remaining Figures 18-22 as well.

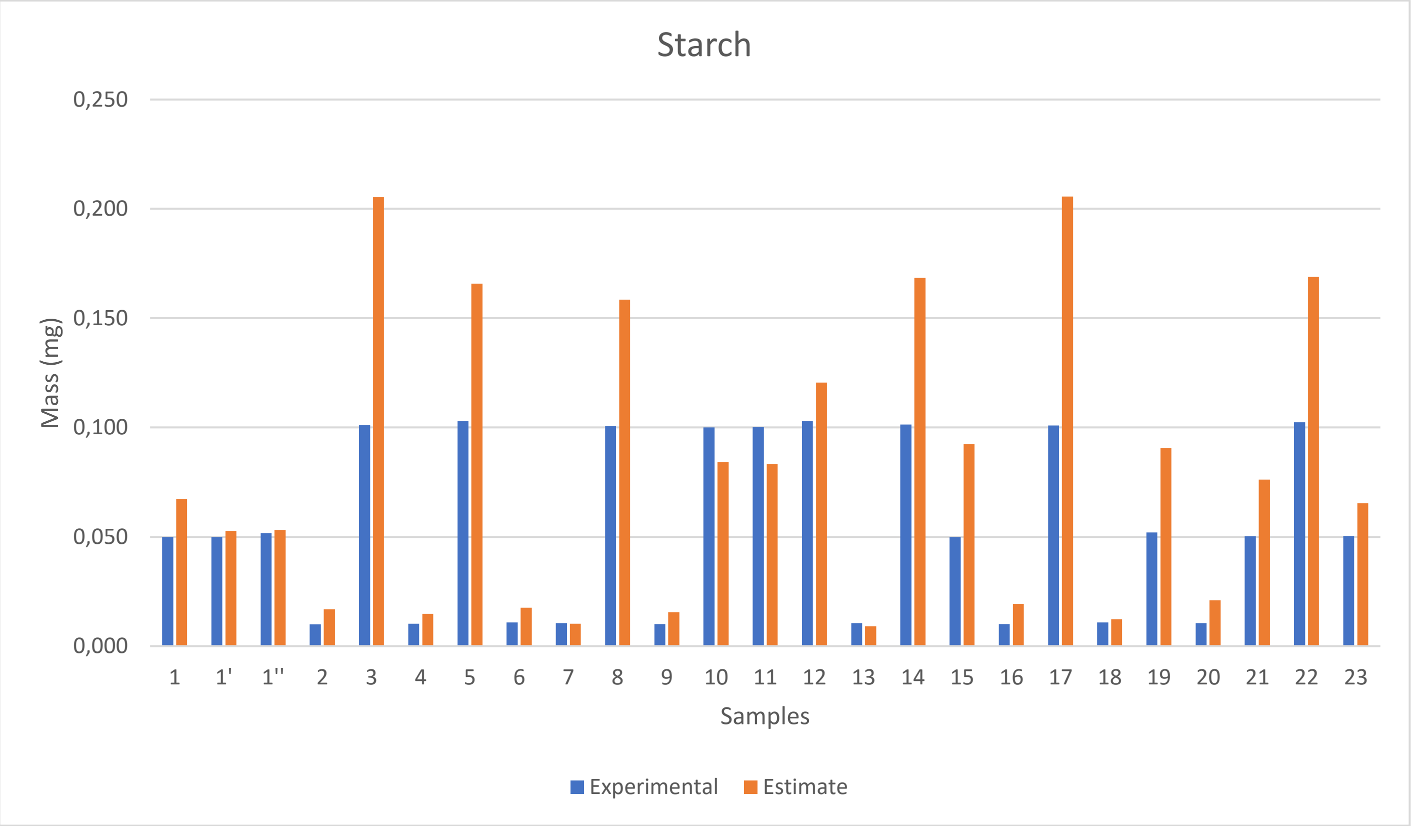

Figure 16: Comparison between the experimental and estimated mass for Starch.

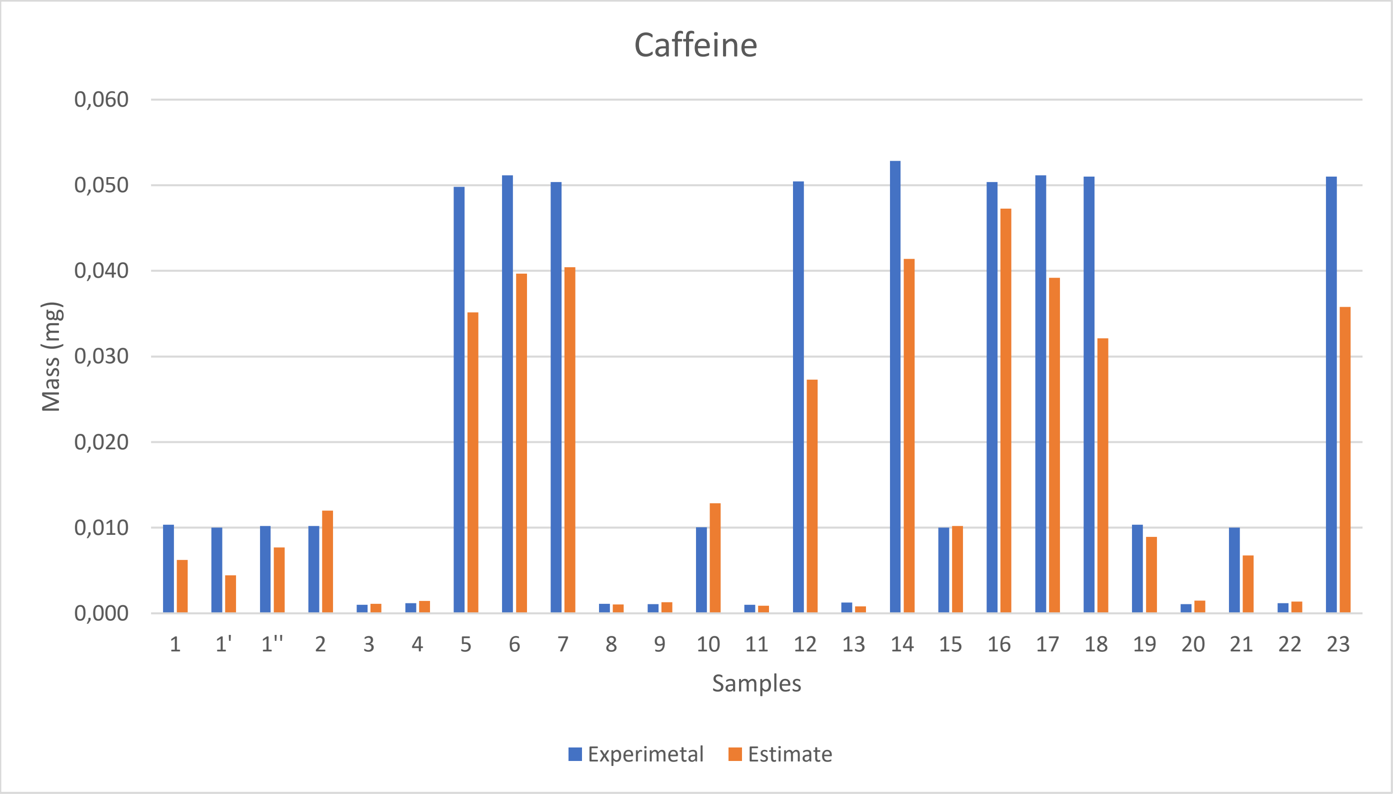

Figure 17: Comparison between the experimental and estimated mass for Caffeine.

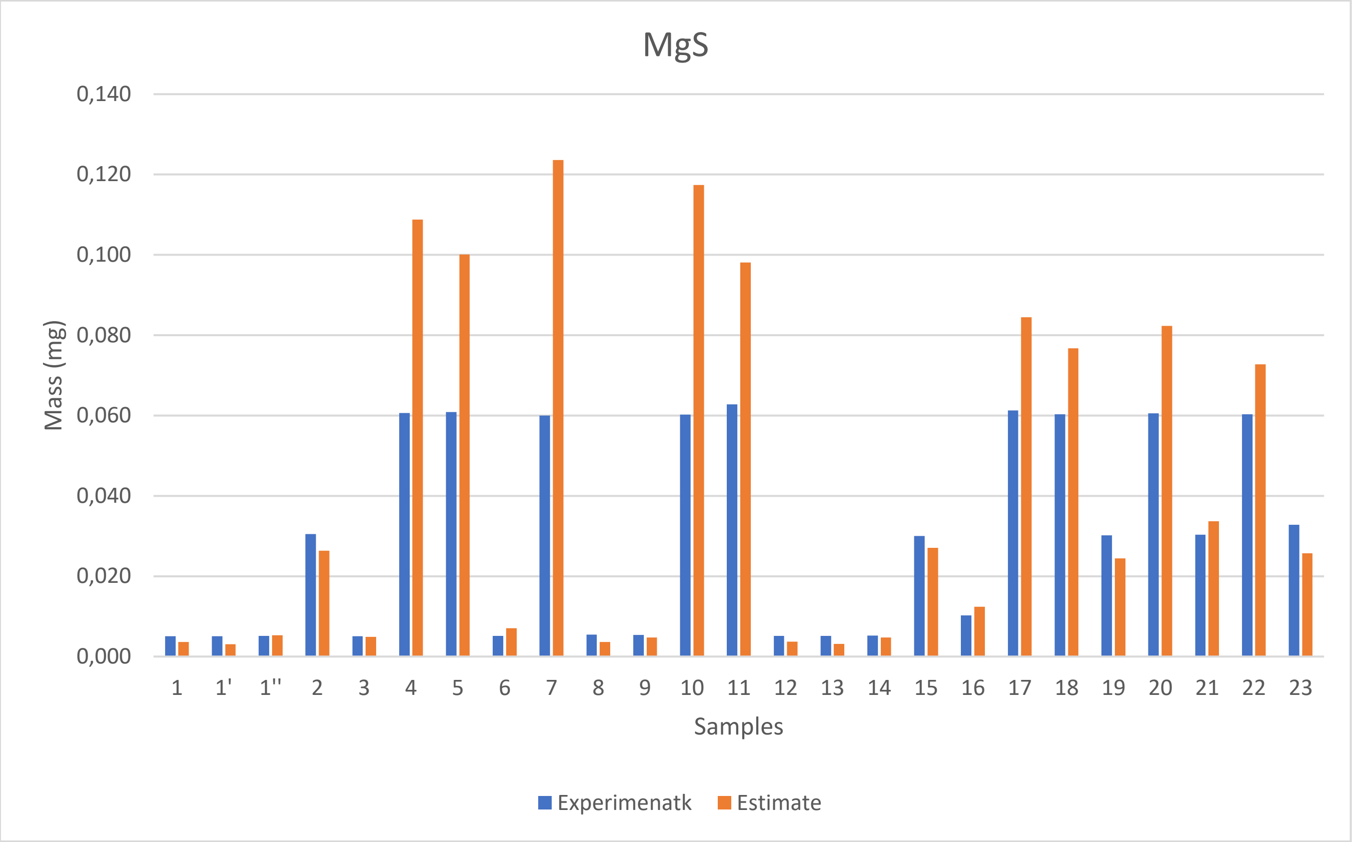

Figure 18: Comparison between the experimental and estimated mass for MgS.

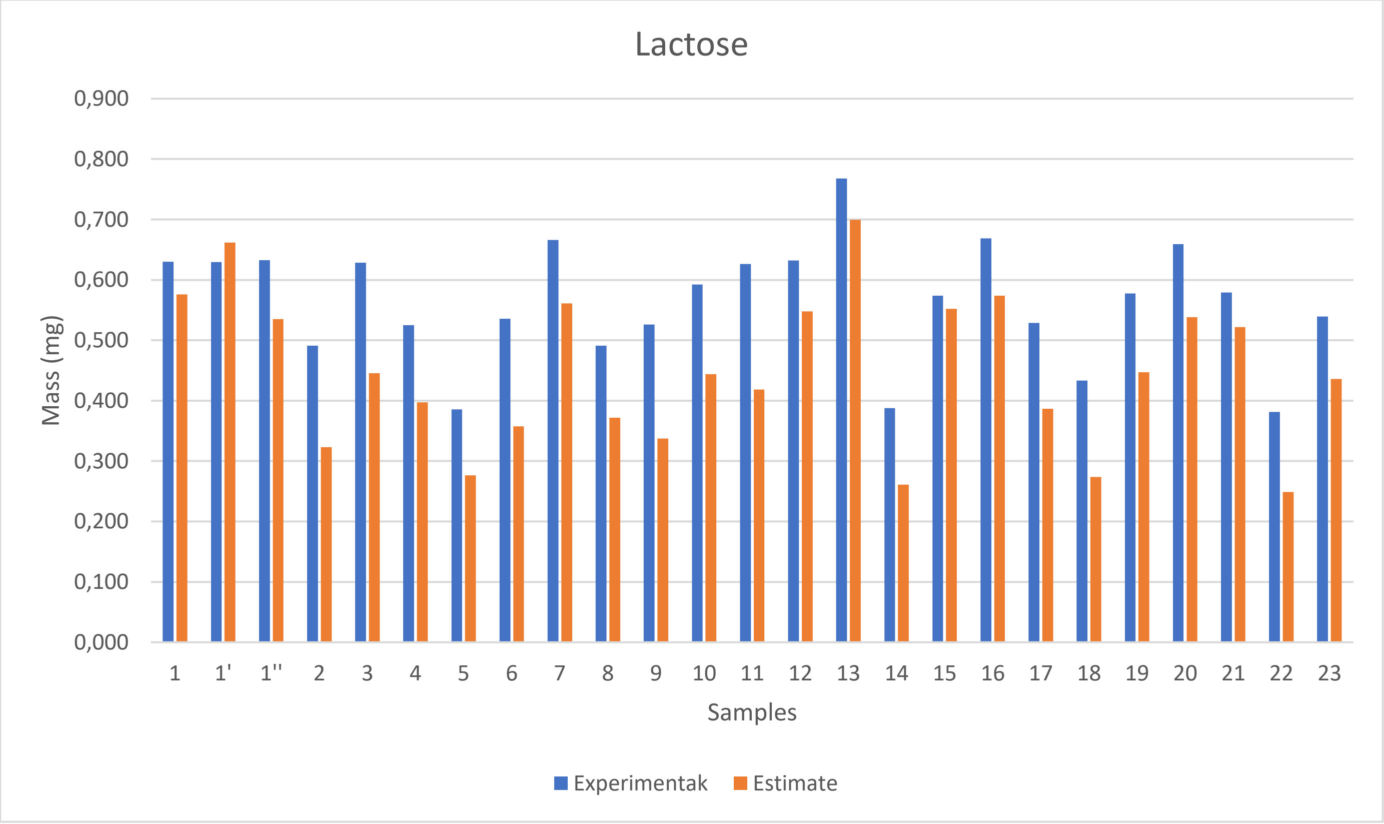

Figure 19: Comparison between the experimental and estimated mass for Lactose.

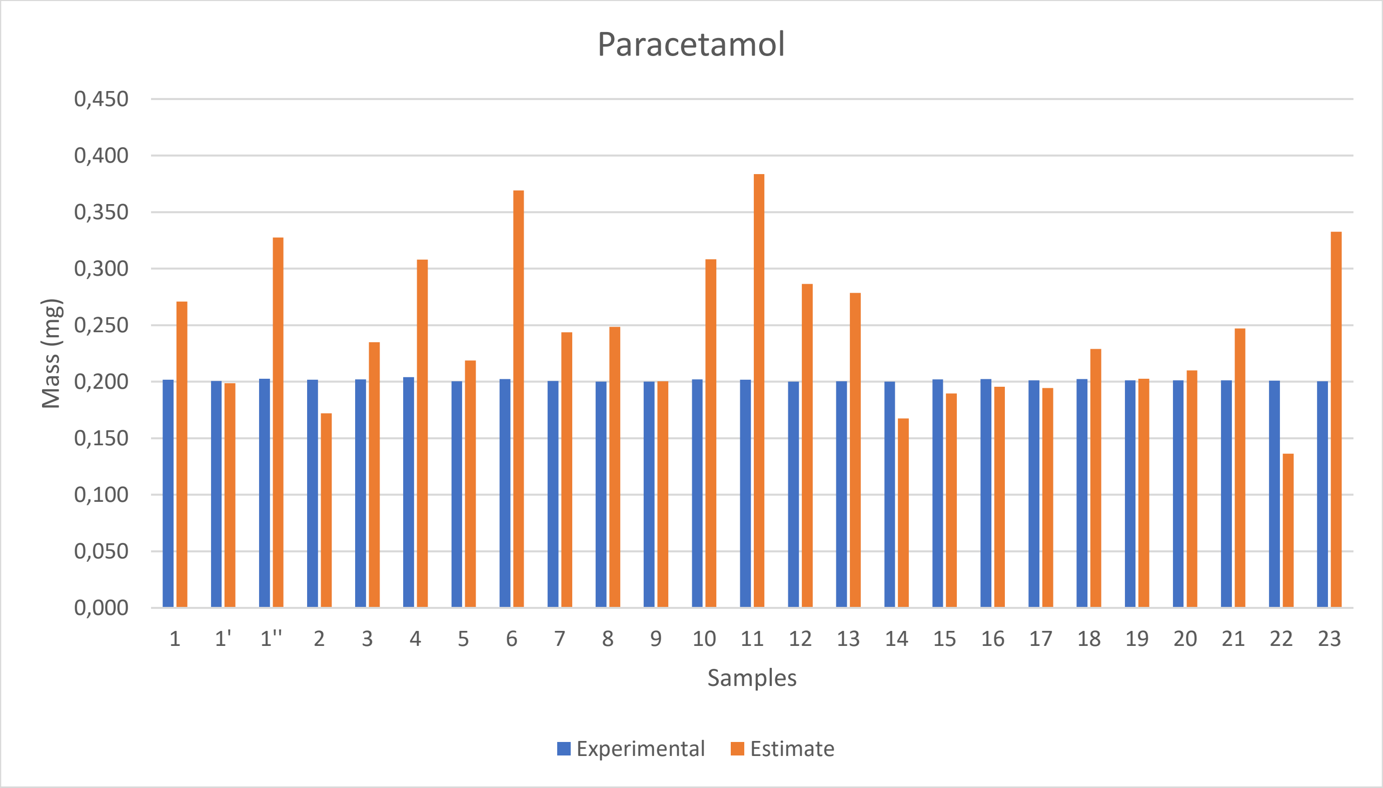

Figure 20: Comparison between the experimental and estimated mass for Paracetamol.

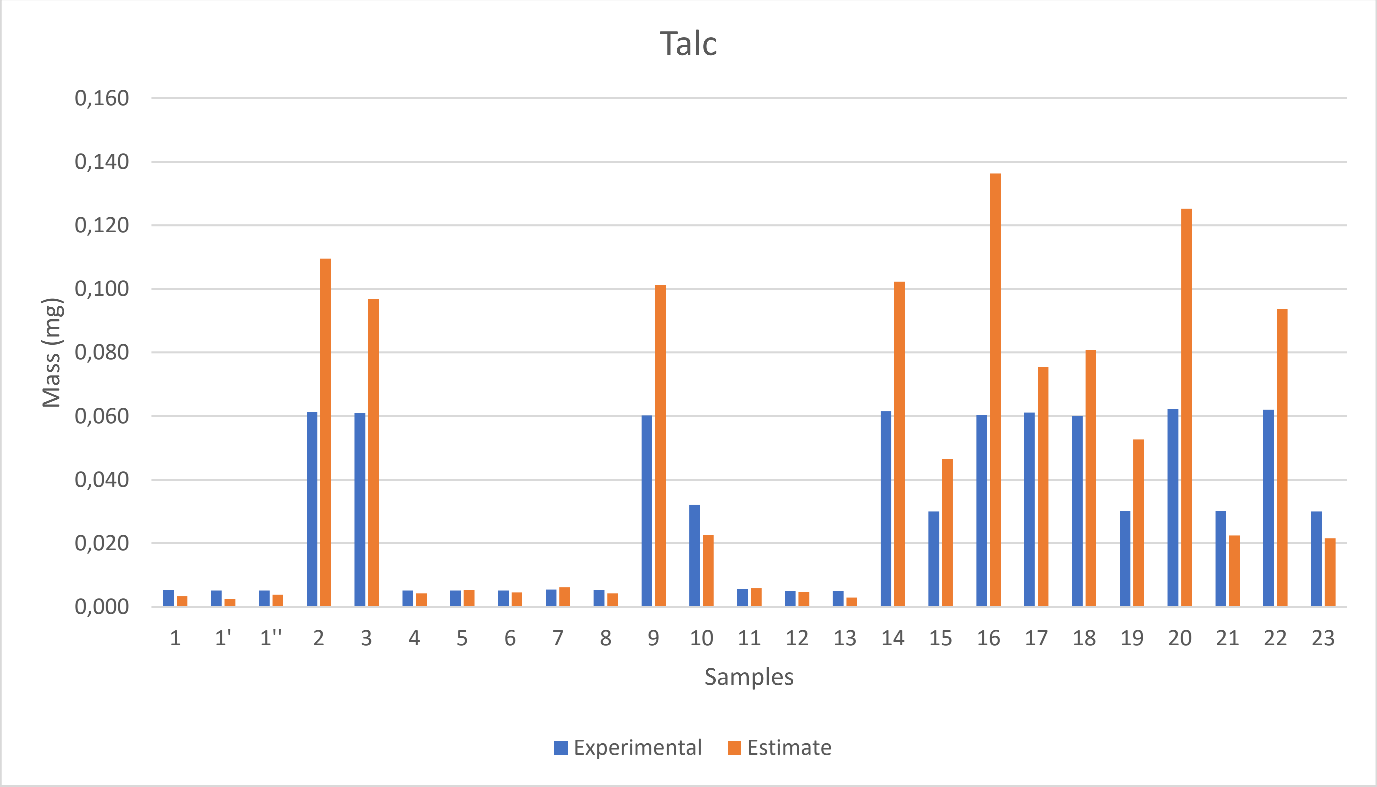

Figure 21: Comparison between the experimental and estimated mass for Talc.

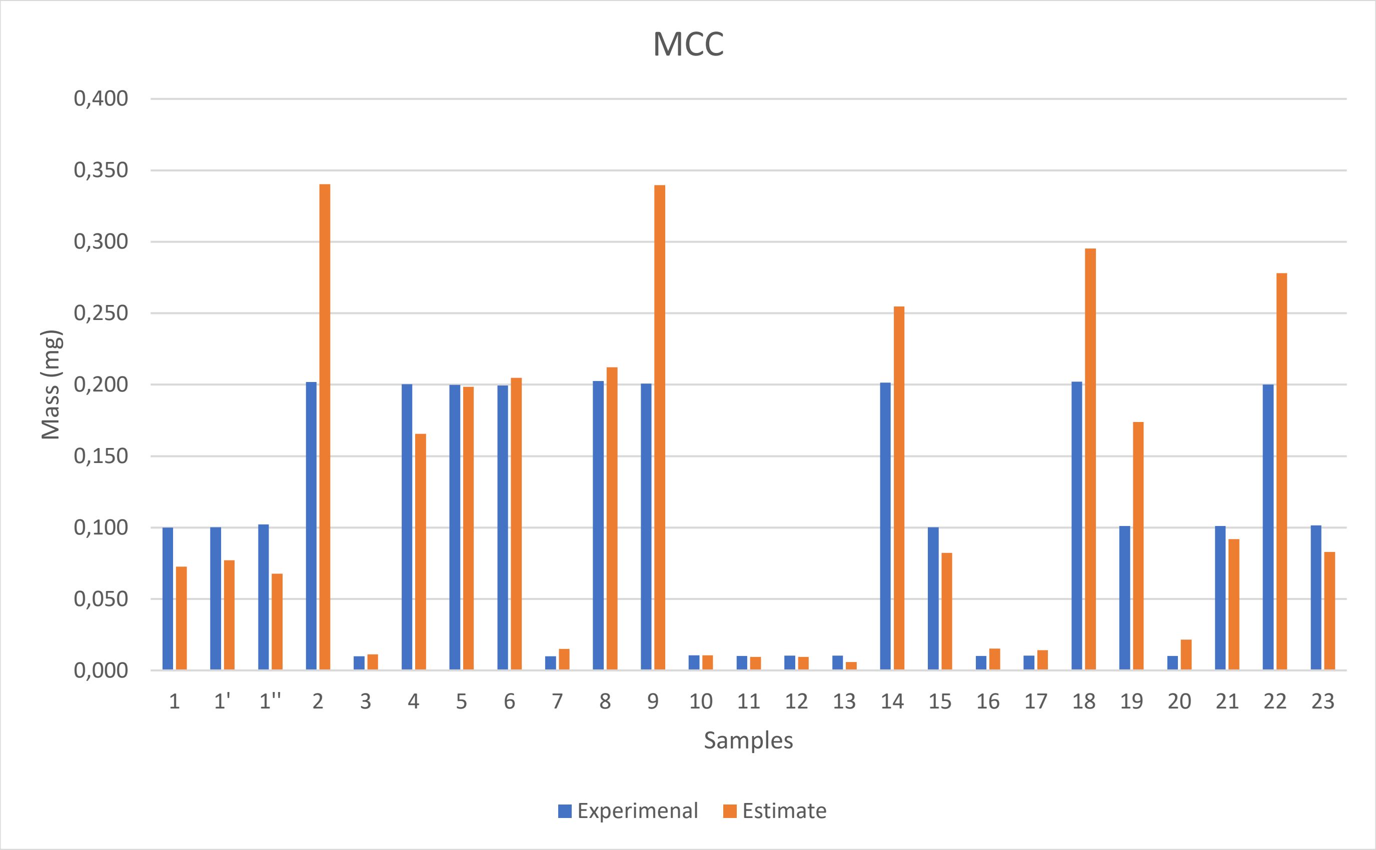

Figure 22: Comparison between the experimental and estimated mass for MCC.

For the same scenario, it was obtained the residuals for each sample. Figures 23-29 it is represented the variation of the residual for each sample and separated by compounds. If the Residual is less than 0.02, the values obtained in the estimation were reasonable, between 0.02 and 0.05 being estimation slightly less good, and higher than 0.05 are not acceptable.

In Figure 23, corresponding to Starch, there are 16 samples below 0.02, 3 samples between 0.02 and 0.05, and the rest of them higher than 0.05. For Figure 24, Caffeine, almost all the values are below 0.02, and just one is between 0.02 and 0.05. Figure 25, which corresponds to MgS, depicts 18 samples with values below 0.02, 5 samples between 0.02 and 0.05, and 2 samples higher than 0.05. Lactose, in Figure 26, has 2 samples between 0.02 and 0.05, and the rest is higher than 0.05. For Figure 27, corresponding to Paracetamol, 8 samples are lower than 0.02, 7 samples have values between 0.02 and 0.05, and 10 samples are higher than 0.05. For Talc, Figure 28, 16 samples are lower than 0.02, 7 samples are between 0.02 and 0.05, and 2 samples are higher than 0.05. The last Figure 29 corresponds to MCC, for this, there are 15 samples lower than 0.02, 4 samples between 0.02 and 0.05, and 6 samples higher than 0.05.

.

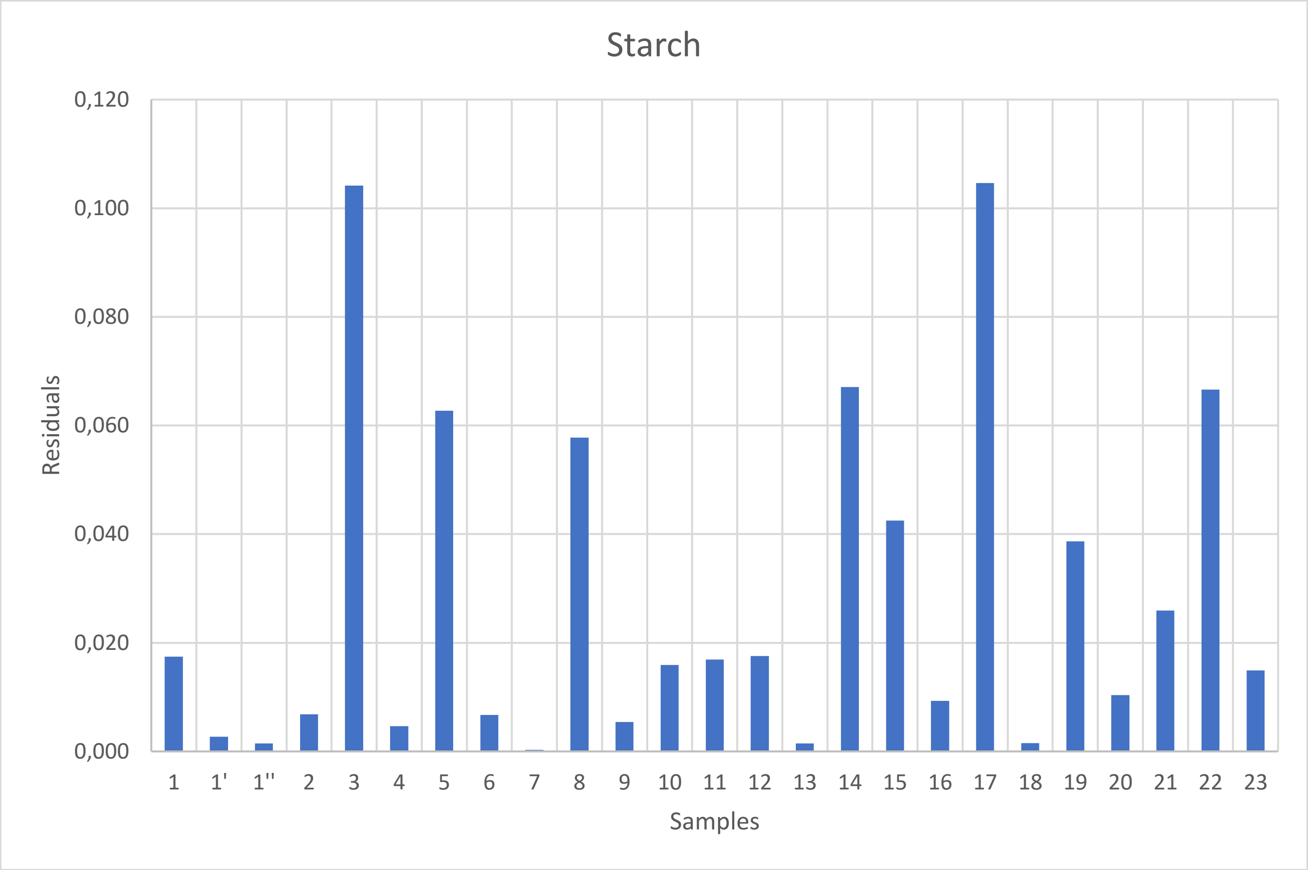

Figure 23: Residuals of each sample for Starch.

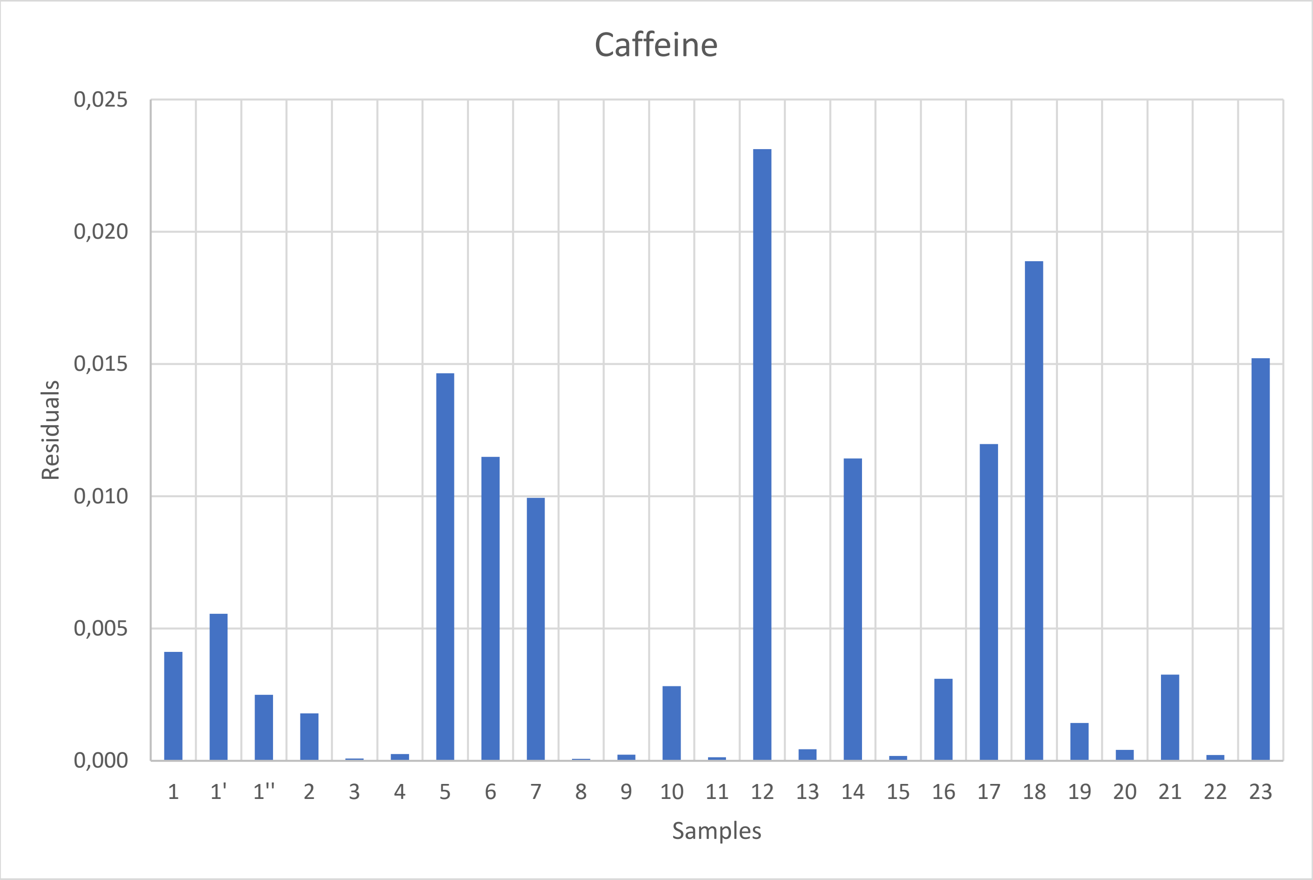

Figure 24: Residuals of each sample for Caffeine.

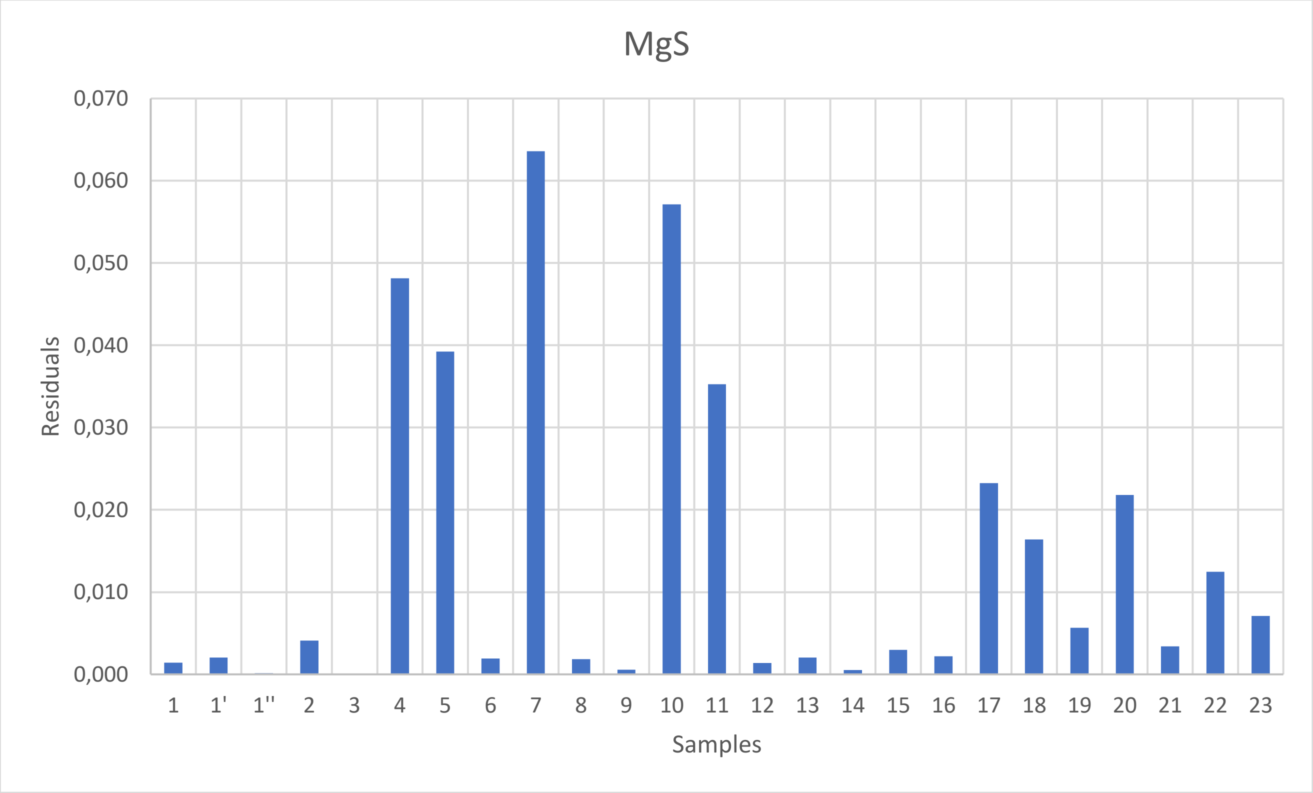

Figure 25: Residuals of each sample for MgS.

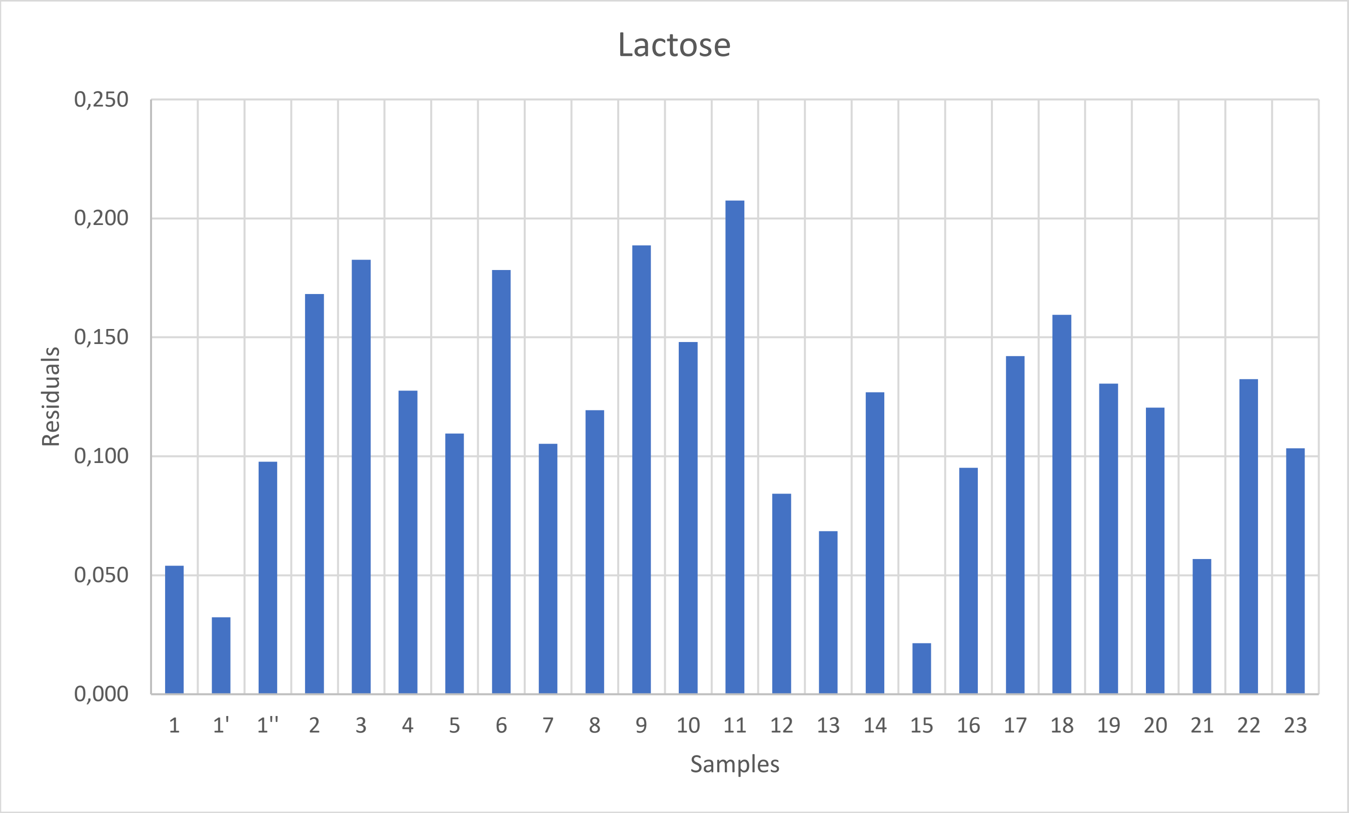

Figure 26: Residuals of each sample for Lactose.

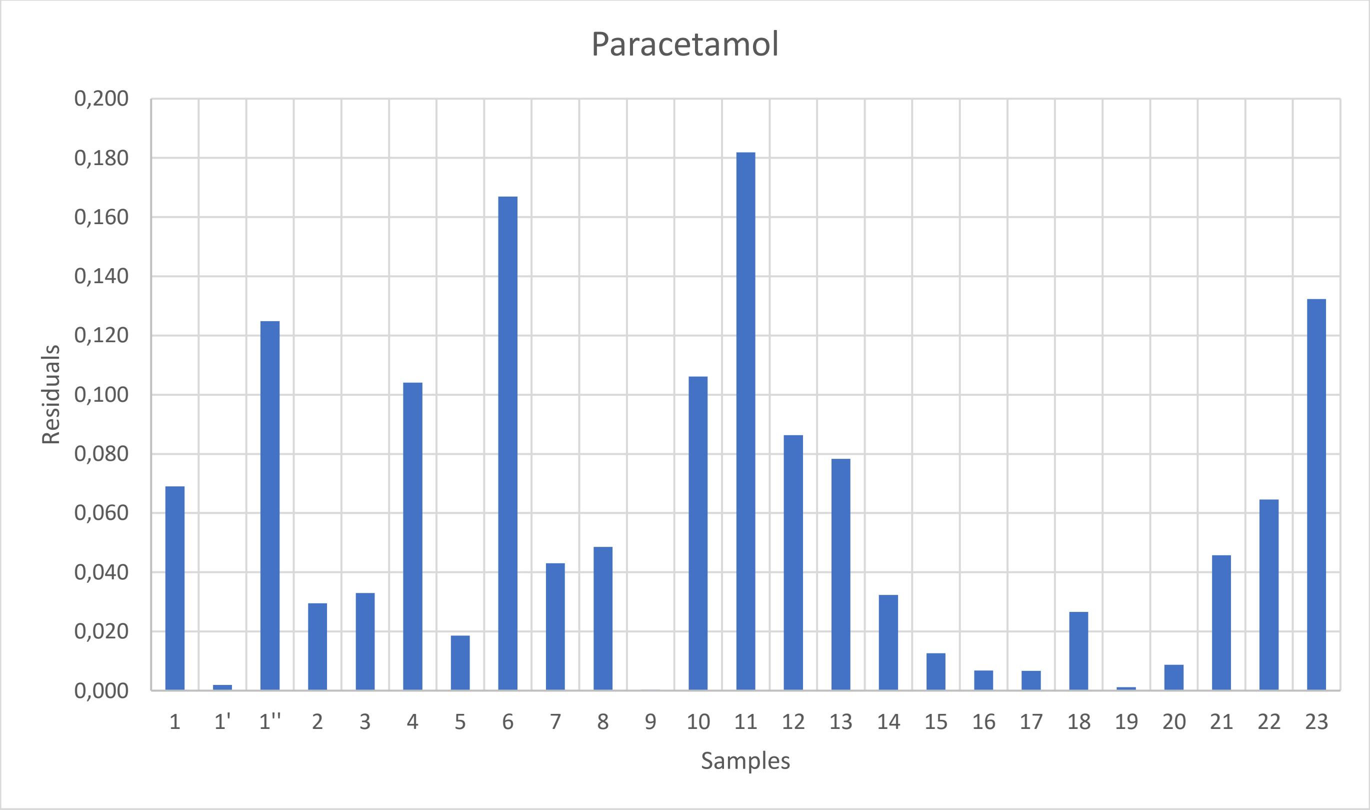

Figure 27: Residuals of each sample for Paracetamol.

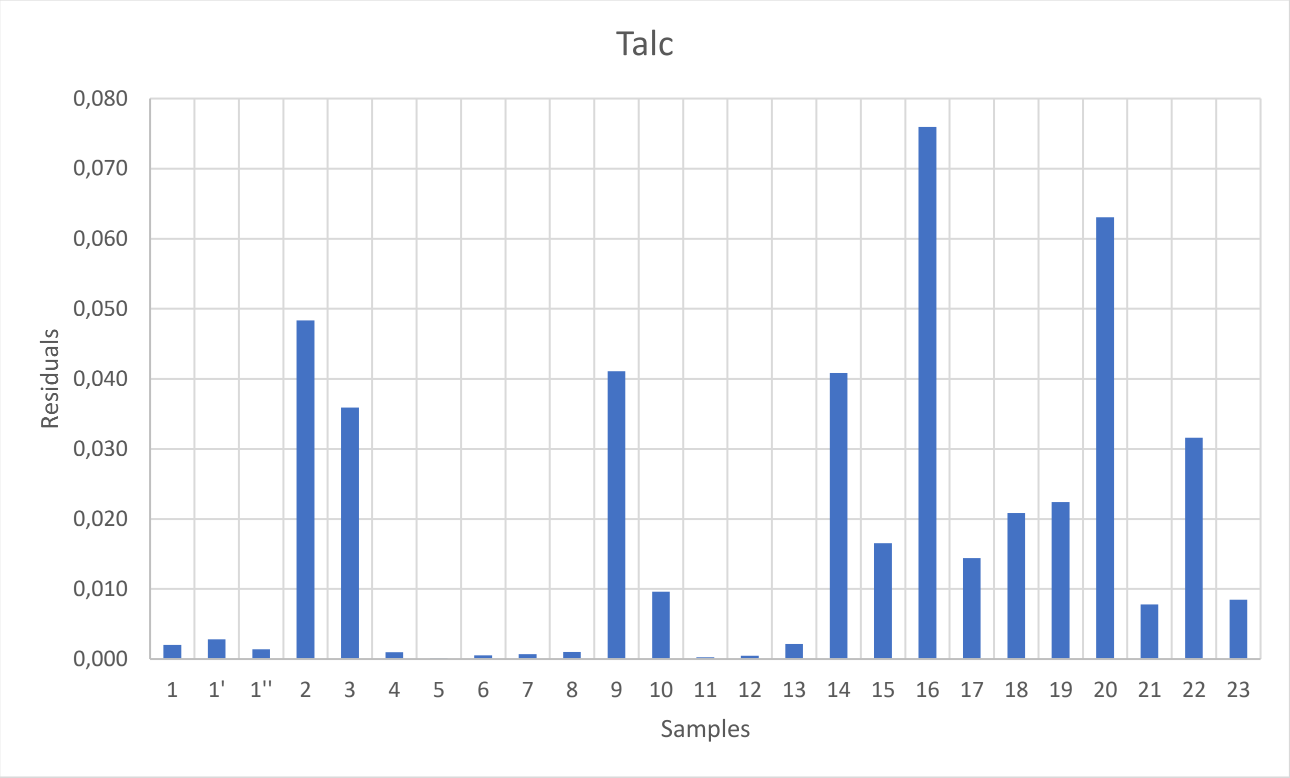

Figure 28: Residuals of each sample for Talc.

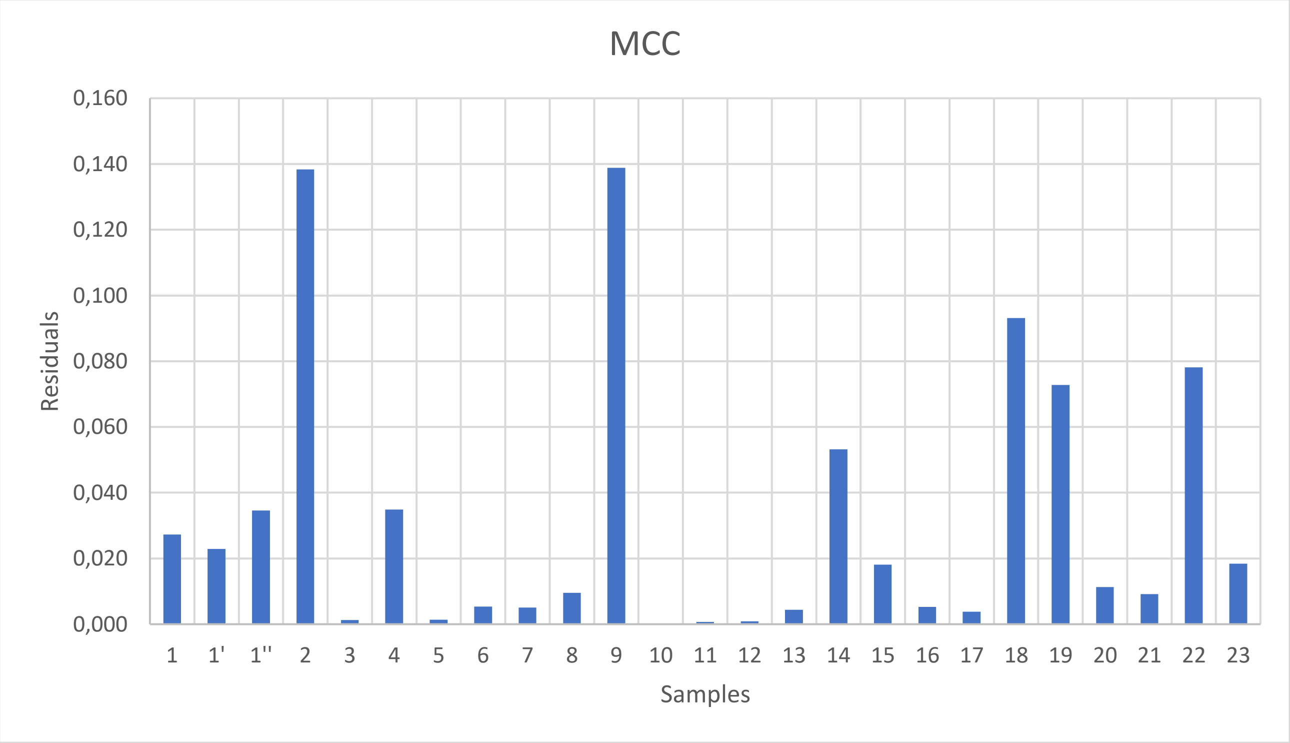

Figure 29: Residuals of each sample for MCC.

In Table 20, the average of the residuals of the samples of Figures 23-29 is represented. Table 20 shows that the results obtained were acceptable for caffeine, MgS, and Talc, for starch and MCC the values of the residuals are at an intermediate level, neither good nor bad, whereas for lactose and paracetamol the values are not acceptable, since it has a residual above 0.05. Therefore, to have a good estimate or an acceptable estimate, it is not recommended to have an uncertainty range above 50%.

Table 20: Average residuals and standard deviation of the residuals for the 50% scenario.

Average of the Residual

Starch	Caffeine	MgS	Lactose	Paracetamol	Talc	MCC
0.028	0.006	0.014	0.118	0.057	0.018	0.032

Standard Deviation of Residual

Starch	Caffeine	MgS	Lactose	Paracetamol	Talc	MCC
0.034	0.007	0.021	0.055	0.064	0.024	0.049

4.4. Comparison between the Two Methods

The comparison of both methods was performed through the analysis of the residuals. In Table 21 are the results of the average of the residuals for each component and each model for the Supervised Method and in Table 22 are the results of the average for the Calibration-free Method. The same color scheme was used, with green corresponding to results with reasonable errors, yellow for intermediate values, and red for the not acceptable.

If each component is compared individually, paracetamol had better residuals in the MCR than had on the Calibration-free method. Caffeine had better residuals in the Calibration-free (Table 21). In both methods, lactose has poor results but has two reasonable results for the 0 and 5% uncertainty intervals (Table 22). In the MCR, it has two intermediate values, which means that for paracetamol it is better to use the MCR to estimate the concentration, and for caffeine and lactose, the Calibration-free method is better.

When comparing all components for the MCR, the best models were Model 4-C and Model 4-D, as there were no residuals greater than 0.05. For the Calibration-free method, the best ones were 0 and 5%, where all components are below 0.02. Also, the 25% could be considered reasonable since only one component presents a value above 0.05 and an intermediate value.

In general, neither method is 100% effective in estimating the concentration of all components, but for certain components, it is possible to have a better estimate of the concentration using one of the two methods. Caffeine, MCC, lactose, and starch are best estimated using the calibration-free method, with starch with up to 25% uncertainty. On the other hand, paracetamol has a better estimation using the MCR. Finally, both talc and MgS had better estimates using either method.

A limitation of the calibration-free method is that it is more convenient to have an estimate of the actual concentrations up to an uncertainty of 25% but is it possible to make a proper estimate of up to 50%. After 50%, it is no longer possible to have reliable estimates for all compounds. So, after that, the method becomes less precise and accurate in estimating the concentrations.

Table 21: Average of residual for the Models used on MCR.

Average of the Residual							
	Starch	Caffeine	MgS	Lactose	Paracetamol	Talc	MCC
Model 1	0.034	0.021	0.012	0.076	0.00085	0.022	0.070
Model 2	0.034	0.013	0.020	0.064	0.00082	0.013	0.048
Model 3	0.033	0.018	0.0086	0.077	0.00082	0.020	0.066
Model 4-A	0.034	0.020	0.0092	0.053	0.00084	0.015	0.051
Model 4-B	0.034	0.019	0.0060	0.050	0.00083	0.010	0.038
Model 4-C	0.032	0.016	0.0073	0.046	0.00081	0.010	0.028
Model 4-D	0.032	0.021	0.010	0.045	0.00084	0.011	0.042
R²	0.53	0.034	0.32	0.78	---	0.61	0.61

Table 22: Average of the residual for the Calibration-free method.

Interval 1							
Percentage	Starch	Caffeine	MgS	Lactose	Paracetamol	Talc	MCC
0%	3.6E-18	1.1E-18	1.1E-18	4.4E-18	8.9E-18	1.9E-18	4.4E-18
5%	0.0018	0.00048	0.0011	0.017	0.010	0.0013	0.0035
25%	0.013	0.0025	0.0062	0.070	0.041	0.0088	0.018
50%	0.028	0.0054	0.014	0.12	0.058	0.019	0.032
75%	0.047	0.011	0.020	0.15	0.063	0.021	0.045
100%	0.063	0.017	0.023	0.16	0.065	0.022	0.051
1000%	0.083	0.017	0.026	0.19	0.065	0.018	0.047
R²	0.57	0.41	0.38	0.38	0.18	0.12	0.23

5. Conclusions

The purpose of this work was to evaluate the best or most efficient method to estimate the concentration of components of a solid formulation, using two distinct methods: one using calibration (MCR) and other calibration-free (in-house developed algorithm). In this work, methods of reverse engineering were performed to analyze a solid formulation that was intended to simulate a pharmaceutical formulation.

To sum up:

The supervised method (Multivariate Curve Resolution) needs a calibration of the method and then validation. So, two sets of formulations were needed for the use of this method, the calibration and validation formulations. The best MCR Model was built imposing the knowledge of the pure spectra and adequate pre-processing. This also presented limitations and was not possible to have a good estimation of the concentration for all the components. The limitations were with the estimation of the Starch, Caffeine, and MCC.

The calibration-free method is an algorithm developed and optimized to be a faster and more cost-effective method. For this, no calibration was required, only the knowledge of the pure components' spectra. In addition, for this method, different initial concentration uncertainty percentages were set to observe to what extent it can be used and still be precise and accurate. The conclusion was that the initial concentrations must not deviate by more than 50% to obtain adequate estimates for all components. Making this a limitation of the method, meaning that the range of mass given to this method cannot be very different from the exact mass of the individual components of a target pharmaceutical product.

Concluding, with this work, it was possible to evaluate two different methods and compare both to understand which method would be better to be a faster and more cost-effective reverse engineering method. Both can be used to support the development of generic drugs, but both have their limitations as discussed along this dissertation.

6. Future Perspectives

This work focused on the development of solid formulations through reverse engineering using Chemometrics for the evaluation of the data obtained by FTIR. Following are some ideas for possible future work:

- Improve the estimation of the concentration of some excipients using the Multivariate Curve Analysis.
- Calibration-free method may require some improvements.
- Collect the spectra using other infrared spectroscopy methods.
- Validate using commercial products (solid).
- Perform the analysis with different formulations: liquid and semi-solid.

7. References

1. Medicamentos Genéricos [Internet]. [cited 2022 Dec 11]. Available from: https://app.infarmed.pt/cotasgenericos/acessibilidade.asp#Lisboa
2. Product-Specific Guidances for Generic Drug Development [Internet]. [cited 2022 Aug 15]. Available from: https://www.accessdata.fda.gov/scripts/cder/psg/index.cfm
3. Generic Drug FAQs: What is a Generic Drug? [Internet]. [cited 2022 Jun 14]. Available from: https://www.drugs.com/article/generic_drugs.html
4. Generic Drugs: Questions & Answers | FDA [Internet]. [cited 2022 Jun 14]. Available from: https://www.fda.gov/drugs/frequently-asked-questions-popular-topics/generic-drugs-questions-answers
5. Generic and hybrid medicines [Internet]. [cited 2022 Sep 15]. Available from: https://www.ema.europa.eu/en/human-regulatory/marketing-authorisation/generic-hybrid-medicines
6. Hybrid medicine [Internet]. [cited 2022 Sep 15]. Available from: https://www.ema.europa.eu/en/glossary/hybrid-medicine
7. Paliwal R, Mamgain A, Kenwat R, Paliwal SR. Reverse Engineering in Pharmaceutical Product Development. In: Mehra N, Gulbake A, editors. Micro- and Nanotechnologies-Based Product Development [Internet]. First ediction. Boca Raton: CRC Press; 2021 [cited 2022 Jun 15]. p. 235–42. Available from: https://www.taylorfrancis.com/books/9781003043164/chapters/10.1201/9781003043164-15
8. ANDA Regulatory Pathway: Q1/Q2(Q3) Deformulation & Equivalence [Internet]. [cited 2022 Jun 14]. Available from: https://www.element.com/nucleus/2022/q1-q2-q3-deformulation-equivalence
9. Monografia das Formas Farmacêuticas. Lisbon; 2008.
10. Bansai AK, Koradia V. The Role of Reverse Engineering in the Development of Generic Formulations [Internet]. [cited 2022 Aug 4]. Available from: www.pharmtech.com
11. Erxleben A. Application of Vibrational Spectroscopy to Study Solid-state Transformations of Pharmaceuticals. Curr Pharm Des. 2016 Oct 26;22(32):4883–911.
12. Larkin P. Infrared and Raman Spectroscopy Principles and Spectral Interpretation. First. Elsevier; 2011. 1–5, 39–41 p.
13. Rytwo G, Zakai R, Wicklein B. The use of ATR-FTIR spectroscopy for quantification of adsorbed compounds. Journal of Spectroscopy. 2015;
14. Singh I, Juneja P, Kaur B, Kumar P, Haji Shabani M, Klodzinska E, et al. Pharmaceutical Applications of Chemometric Techniques. ISRN Analytical Chemistry

[Internet]. 2013 [cited 2022 Aug 4];2013:13. Available from: http://dx.doi.org/10.1155/2013/795178

15. Bakeev KA. Process Analytical Technology. Bakeev K, editor. Vol. 2. Greant Britain: Wiley; 2010. 82,83,354,355.

16. Yu LX, Amidon G, Khan MA, Hoag SW, Polli J, Raju GK, et al. Understanding pharmaceutical quality by design. Vol. 16, AAPS Journal. Springer New York LLC; 2014. p. 771–83.

17. Stuart BH. Infrared Spectroscopy: Fundamentals and Applications [Internet]. John Wiley & Sons, Ltd; 2004 [cited 2022 Jun 27]. 33–35, 51–57 p. Available from: http://www.pharmaresearchlibrary.com/wp-content/uploads/2013/04/Infrared-Spectroscopy-Fundamentals-and-Applications-Barbara-Stuart.pdf

18. Wise B, Gallagher N, Bro R, Shaver J, Windig W, Koch RS. Chemometrics Tutorial. Wenatchee: Eigenvector Research, Inc.; 173–182 p.

19. Bhatti A, Syed NA, John P. Reverse engineering and its applications. In: Omics Technologies and Bio-engineering: Towards Improving Quality of Life. Elsevier Inc.; 2018. p. 95–110.

20. Committee for medicinal products for human use (CHMP)) [Internet]. London; 2010 Jan [cited 2022 Sep 12]. Available from: https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-investigation-bioequivalence-rev1_en.pdf

21. Abbreviated New Drug Application (ANDA) | FDA [Internet]. [cited 2022 Aug 17]. Available from: https://www.fda.gov/drugs/types-applications/abbreviated-new-drug-application-anda

22. Medicines Agency E. Questions and answers on generic medicines. 2012 [cited 2022 Aug 16]; Available from: www.ema.europa.eu

23. Polli JE, Abrahamsson BS, Yu LX, Lionberger RA. FDA Critical Path Initiatives: Opportunities for Generic Drug Development. 2008;

24. Warad TA, Karbhari VN, Dhanasure SG. Recent trends in pharmaceutical reverse engineering. Indo American Journal of Pharmaceutical Research [Internet]. 2018;8(09):1670–5. Available from: www.iajpr.com

25. Swarbrick J, Augsburger LL, Dressman JB, Hughes JA, Jones TM, Lee VHL, et al. Generic Drug Product Development [Internet]. Shargel L, Kanfer I, editors. New York: Marcel Dekker; 2005 [cited 2022 Sep 16]. 31–50 p. Available from: https://pharmachitchat.files.wordpress.com/2015/05/generic-drug-product-development-solid-oral-dosage-forms.pdf

26. Medicines Agency E. Committee for Human Medicinal Products (CHMP) Guideline on manufacture of the finished dosage form. 2017 [cited 2022 Sep 15]; Available from: www.ema.europa.eu/contact

27. Fourier Transform Infrared Spectroscopy - an overview [Internet]. [cited 2022 Jun 29]. Available from: https://www.sciencedirect.com/topics/engineering/fourier-transform-infrared-spectroscopy

28. Hinton-Sheley P. ATR-FTIR: An Overview [Internet]. 2021 [cited 2022 Jun 28]. Available from: https://www.azolifesciences.com/article/What-is-ATR-FTIR.aspx

29. Skoog D, Holler FJ, Crouch S, editors. Principles of Instrumental Analysis [Internet]. 17th ed. Bostom: Cengage Learning; 2016 [cited 2022 Jun 27]. 186–187 p. Available from: https://www.chemcome.com/wp-content/uploads/2020/11/Principles-of-Instrumental-Analysis-7th-edition-Skoog-by-Douglas-A.-Skoog-F.-James-Holler-Stanley-R.-Crouch-z-lib.org_.pdf

30. Müller ALH, Flores ÉMM, Müller EI, Silva FEB, Ferrão MF. Attenuated Total Reflectance with Fourier Transform Infrared Spectroscopy (ATR/FTIR) and Different PLS Algorithms for Simultaneous Determination of Clavulanic Acid and Amoxicillin in Powder Pharmaceutical Formulation. Article J Braz Chem Soc [Internet]. 2011 [cited 2022 Aug 6];22(10). Available from: http://www.infometrix.com

31. Kolekar YM. Understanding of DoE and its advantages in Pharmaceutical development as per QbD Approach. Asian Journal of Pharmacy and Technology. 2019;9(4):271.

32. Santana IM, Breitkreitz MC, Pinto L. Multivariate curve resolution alternating least squares applied to chromatographic data: From the basics to the recent advances. Vol. 8, Brazilian Journal of Analytical Chemistry. DKK Comunicacao; 2021.

33. Ruckebusch C, Blanchet L. Multivariate curve resolution: A review of advanced and tailored applications and challenges. Vol. 765, Analytica Chimica Acta. 2013. p. 28–36.

34. Mostafa A, Shaaban H. Quantitative analysis and resolution of pharmaceuticals in the environment using multivariate curve resolution-alternating least squares (MCR-ALS). Acta Pharmaceutica. 2019 Jun 1;69(2):217–31.

35. Fakayode SO, Baker GA, Bwambok DK, Bhawawet N, Elzey B, Siraj N, et al. Molecular (Raman, NIR, and FTIR) spectroscopy and multivariate analysis in consumable products analysis. Vol. 55, Applied Spectroscopy Reviews. Bellwether Publishing, Ltd.; 2020. p. 647–723.

36. Čapková T, Pekárek T, Hanulíková B, Matějka P. Application of reverse engineering in the field of pharmaceutical tablets using Raman mapping and chemometrics. J Pharm Biomed Anal. 2022 Feb 5;209.

37. Shafaq S, Irfan Majeed M, Nawaz H, Rashid N, Akram M, Yaqoob N, et al. Quantitative analysis of solid dosage forms of Losartan potassium by Raman spectroscopy. Spectrochim Acta A Mol Biomol Spectrosc. 2022 May 5;272.

38. Deidda R, Sacre PY, Clavaud M, Coïc L, Avohou H, Hubert P, et al. Vibrational spectroscopy in analysis of pharmaceuticals: Critical review of innovative portable and handheld NIR and Raman spectrophotometers. TrAC - Trends in Analytical Chemistry. 2019 May 1;114:251–9.

39. Custers D, Cauwenbergh T, Bothy JL, Courselle P, de Beer JO, Apers S, et al. ATR-FTIR spectroscopy and chemometrics: An interesting tool to discriminate and characterize counterfeit medicines. J Pharm Biomed Anal. 2015 Aug 1;112:181–9.

40. Mallah MA, Sherazi STH, Bhanger MI, Mahesar SA, Bajeer MA. A rapid Fourier-transform infrared (FTIR) spectroscopic method for direct quantification of paracetamol content in solid pharmaceutical formulations. Spectrochim Acta A Mol Biomol Spectrosc. 2015 Apr 15;141:64–70.

41. Verma K, Akhtar M, Anchliya A. Combination of FTIR Spectroscopy and Chemometric Method on Quantitative Approach - A Review. Austin J Anal Pharm Chem [Internet]. 2021 [cited 2022 Dec 12];8(1):01–015. Available from: www.austinpublishinggroup.com

42. Acetaminophen [Internet]. [cited 2022 Aug 4]. Available from: https://medlineplus.gov/druginfo/meds/a681004.html

43. Acetaminofeno [Internet]. [cited 2022 Aug 4]. Available from: https://www.lecturio.com/pt/concepts/acetaminofeno/

44. Caffeine [Internet]. [cited 2022 Aug 4]. Available from: https://pubchem.ncbi.nlm.nih.gov/compound/Caffeine

45. Caffeine: Uses, Interactions, Mechanism of Action [Internet]. [cited 2022 Aug 4]. Available from: https://go.drugbank.com/drugs/DB00201

46. Rowe R, Sheskey P, Quinn M, editors. Handbook of Pharmaceutical Excipients. 6th ed. London: Pharmaceutical Press; 2009. 129–130, 364, 404, 686, 782 p.

Attachments

Annex 1

```
function Result=RE_CompositionEstimation_Alg1(Samples,Pure,Sample2Test,PureLabels,WRange,CRange,flag)
%
%Result=RE_CompositionEstimation_Alg1(Samples,Pure,Sample2Test,PureLabels,WRange,CRange,flag)
%
%Algorithm to estimate mass fractions from a formulation based on IR
%spectra
%
%Inputs:
%Samples: the IR spectra of the formulations
%Pure: the IR spectra of the pure compounds of the formulation
%Sample2Test: an integer corresponding for the sample to be tested
%PureLabels (optional): Labels of the pure compounds
%WRange (optional): Indexes of the columns to use in the analysis
%CRange (optional): A two row matrix with minimum (first row) and maximum (second row) mass fractions for the components
%Flag: If set to 1 then results are displayed
%
%Outputs:
%Result: a matrix with the mass fractions corresponding to the best estimation and confidence limits (95%)
%
%JAL, MEFARM, 2022
%
%


%Change WRange and CRange according to the needs
% WRange=[1245:6970 9527:12445];
% CRange=[
%  0 0.5    %Starch
%  0 1    %Caffeine
%  0 1    %MgSt
%  0 1   %Lactose
%  0 1   %Paracetamol
%  0 1    %Talc
%  0 1    %MCC
%  ]';


N=10000;
PureLabels={'Starch','Caffeine','MgSt','Lactose','Paracetamol','Talc','MCC'};

if flag==1
disp(['Sample to test: ' Samples.label{1}(Sample2Test,:)]);
end

Sexc=wlsbaseline(Pure.data,3);
```

```
Sref=wlsbaseline(Samples.data(Sample2Test,:),3);

Sexc=normaliz(Sexc,0,2);
Sref=normaliz(Sref,0,2);

Sexc=Sexc(:,WRange);
Sref=Sref(:,WRange);

for k=1:size(CRange,2)
    R(:,k)=(CRange(2,k)-CRange(1,k)).*rand(N,1)+CRange(1,k);
end
Rold=R;

ufix=find((CRange(2,:)-CRange(1,:))==0);
uvar=find((CRange(2,:)-CRange(1,:))>0);

for k=1:size(R,1)
    Stot=1-sum(R(k,ufix));
    R(k,uvar)=R(k,uvar)./sum(R(k,uvar)).*Stot;
end

Srec=R*Sexc;
E=repmat(Sref,N,1)-Srec;
E=E.*E;
E=sqrt(sum(E'));

[u1,u2]=min(E);[u3,u4]=sort(E);


if flag==1
%figure;plot(sort(E))
figure
subplot(2,1,1)
plot(Samples.axisscale{2}(WRange),Srec(u2,:),'r');hold
on;plot(Samples.axisscale{2}(WRange),Sref,'b');legend('Best Estimation','Test Sample');
title(['Sample ' Samples.label{1}(Sample2Test,:) ', RMSEP=' num2str(u1)])
xlabel('Wavenumbers (cm-1)');ylabel('Absorbance')
end

b=5;s=1;
BestE=R(u4(1:b),:);
AvgBestEst=mean(BestE);
StdBestEst=std(BestE);

if flag==1
subplot(2,1,2)
bar(AvgBestEst,'b');hold
on;plot(AvgBestEst+s.*StdBestEst,'ok','markerfacecolor','r');plot(AvgBestEst-
s.*StdBestEst,'ok','markerfacecolor','r');
end

u=find((CRange(2,:)-CRange(1,:))==0);

if flag==1
bar(u,AvgBestEst(u),'g')
```

```
for k=1:size(Pure,1)
   line([k k],[AvgBestEst(k)+s.*StdBestEst(k) AvgBestEst(k)-s.*StdBestEst(k)],'color',[1 0 0]);
end
set(gca,'xticklabel',PureLabels)
xtickangle(45)
xlabel('Components');ylabel('Mass Fractions %m/m')
disp(['Fitting Error (RMSE) = ' num2str(u1)])
disp('Estimation of mass fractions (%)')
for k=1:size(Pure,1)
   if StdBestEst(k)==0
      u='FIX';
   else
      u='VAR';
   end
   disp([num2str(k) ' - ' PureLabels{k} ' ' u ' - ' num2str(AvgBestEst(k)*100) '% ('
num2str((AvgBestEst(k)-s.*StdBestEst(k))*100) ' - '
num2str((AvgBestEst(k)+s.*StdBestEst(k))*100) '%)']);
end

end

Estimations=[AvgBestEst ; AvgBestEst+s.*StdBestEst ; AvgBestEst-s.*StdBestEst]';
Result=array2table(Estimations,'VariableNames',{'Average_Mass_Fraction','Upper_95_Limit'
,'Lower_95_Limit'},'RowNames',PureLabels);
```

Annex 2

```
%% Script 1
clear Erro ErrroR ConcEst

WRange=[1245:6970 9127:12445];

CRange=[
 0.01 0.1    %Starch
 0.001 0.05    %Caffeine
 0.005 0.06    %MgSt
 0.38 0.76    %Lactose
 0.2 0.2    %Paracetamol
 0.005 0.06    %Talc
 0.01 0.2    %MCC
 ]';

%% Script 1 - for analysing just just one sample
    k=12;
    disp(['Checking DoE Sample ' num2str(k)]);
    C=RE_CompositionEstimation_Alg1(CalSet1,MPSet1,k,[],WRange,CRange,1);
    ConcEst{k}=C;
    Erro=table2array(C(:,1))'-table2array(Massa(k,:));
    ErroR=(table2array(C(:,1))'-
table2array(Massa(k,:)))./table2array(Massa(k,:)).*100;

%% Script 2 - for running all 25 samples

for k=1:25
    disp(['Checking DoE Sample ' num2str(k)]);
    C=RE_CompositionEstimation_Alg1(CalSet1,MPSet1,k,[],WRange,CRange,1);
    Erro(k,:)=table2array(C(:,1))'-table2array(Massa(k,:));
    ErroR(k,:)=(table2array(C(:,1))'-
table2array(Massa(k,:)))./table2array(Massa(k,:)).*100;

end

%% Script 3 - to browse through all samples based on a % of initial
uncertainty

%Click on <Run Section> to run this script

clear CRange CPred Residual AvResidual StdResidual

WRange=[1245:6970 9127:13275];
ExpMass=table2array(Massa);

VarPercent=[0 5 25 50 75 100 1000];
VarPercent=[50];

for u=1:length(VarPercent)
    disp(['Trying variance of ' num2str(VarPercent(u)) '% around
average.']);
for k=1:25
    %Defines the CRange array based on the percentage of variation
    for j=1:7
        CRange(1,j)=max(0,ExpMass(k,j)-VarPercent(u)/100*ExpMass(k,j));
        CRange(2,j)=min(1,ExpMass(k,j)+VarPercent(u)/100*ExpMass(k,j));


    end

CPred{k}=RE_CompositionEstimation_Alg1(CalSet1,MPSet1,k,[],WRange,CRange,0)
;
    Residual(k,:)=table2array(CPred{k}(:,1))'-table2array(Massa(k,:));

end

%Calculate average and standard deviation of residuals
AvResidual(u,:)=mean(abs(Residual));StdResidual(u,:)=std(Residual);


end

%Plot Results
PureLabels={'Starch','Caffeine','MgSt','Lactose','Paracetamol','Talc','MCC'
};
figure;imagesc(AvResidual);colorbar;set(gca,'yticklabel',num2str(VarPercent
'),'xticklabel',PureLabels);title('Average Absolute
Residuals');xlabel('Components');ylabel('% of Initial Uncertainty')
figure;plot(AvResidual,'o-');legend(PureLabels);ylabel('Average Absolute
Residuals');xlabel('% of Initial Uncertainty')
set(gca,'xtick',[1:length(VarPercent)],'xticklabel',num2str(VarPercent'));
figure;plot(StdResidual,'o-');legend(PureLabels);ylabel('Residuals Standard
Deviation');xlabel('% of Initial Uncertainty')
set(gca,'xtick',[1:length(VarPercent)],'xticklabel',num2str(VarPercent'));


% for k=1:25
% figure;bar([ table2array(Massa(k,:))' table2array(CPred{k}(:,1)) ])
% set(gca,'xticklabel',PureLabels);xlabel('Components');ylabel('Mass
Fraction Predictions');legend('Experimental','Predicted');title(['For
sample all']);
% end
```

```
%% Script 4 - Compare predictions for one sample considering one scenario

%Click on <Run Section> to run this script

clear CRange CPred Residual
WRange=[1245:6970 9127:12445];ExpMass=table2array(Massa);

%Set the scenario and the sample to test
VarPercent=1000; %Set the scenario (% of variation around the real mass
fraction)
Sample2Test=4;  %Set sample to test the predictions on this scenario

    disp(['Trying variance of ' num2str(VarPercent(u)) '% around
average.']);

    %Defines the CRange array based on the percentage of variation
    for j=1:7
        CRange(1,j)=max(0,ExpMass(Sample2Test,j)-
VarPercent(u)/100*ExpMass(Sample2Test,j));

CRange(2,j)=min(1,ExpMass(Sample2Test,j)+VarPercent(u)/100*ExpMass(Sample2T
est,j));
    end

CPred=RE_CompositionEstimation_Alg1(CalSet1,MPSet1,Sample2Test,[],WRange,CR
ange,0);
    Residual=table2array(CPred(:,1))'-table2array(Massa(Sample2Test,:));


figure;bar([CRange(1,:)' CRange(2,:)' table2array(Massa(Sample2Test,:))'
table2array(CPred(:,1)) ])
set(gca,'xticklabel',PureLabels);xlabel('Components');ylabel('Mass Fraction
Predictions');legend('Min Value','Max
Value','Experimental','Predicted');title(['For sample '
num2str(Sample2Test)]);
```

Annex 3

```
function [R,BestMatch]=BC_CompareLoadings(Pure,Model)
%Function to compare loadings estimated by MCR with pure spectra
%JAL 2022
ModelLoadings=Model.loads{2}';
ind=Model.detail.includ{2};

for k=1:size(ModelLoadings,1)
```

```
    S2=ModelLoadings(k,:)./max(ModelLoadings(k,:));
    for j=1:size(Pure)
        S1=Pure.data(j,ind)./max(Pure.data(j,ind));
        R(j,k)=mean(sqrt((S2-S1).^2));
    end
end
R(8,:)=0;R(:,8)=0;

 figure;
 subplot(2,1,1);
 pcolor(R);xlabel('Predicted');ylabel('Experimental');colorbar
 set(gca,'xtick',[0.5:1:7.5])
set(gca,'ytick',[0.5:1:7.5])
set(gca,'xticklabel',{'','Pred1','Pred2','Pred3','Pred4','Pred5','Pred6','Pred7'})
set(gca,'yticklabel',{'','Exp1','Exp2','Exp3','Exp4','Exp5','Exp6','Exp7'})
title('RMSE Distance Matrix')

R(8,:)=[];R(:,8)=[];
[u1,u2]=min(R);
for k=1:7
    disp(['Loading ' num2str(k) ' best matches pure compound ' num2str(u2(k))])
    BestMatch(k)=u2(k);
end
subplot(2,1,2)
plot (R','s','linewidth',5)
legend('Exp1','Exp2','Exp3','Exp4','Exp5','Exp6','Exp7');
set(gca,'xtick',[0:8],'xlim',[0 8])
set(gca,'xticklabel',{'','Pred1','Pred2','Pred3','Pred4','Pred5','Pred6','Pred7',''})
ylabel('Distance (RMSE)');xlabel('Predicted')
```