UNIVERSIDADE DE LISBOA

INSTITUTO SUPERIOR TÉCNICO

High-throughput Fourier-transform mid-infrared spectroscopy:
A mechanism-based screening approach to antibiotic discovery

Bernardo José Simões de Almeida Ribeiro da Cunha

Supervisor: Doctor Cecília Ribeiro da Cruz Calado

Co-Supervisor: Doctor Luís Joaquim Pina da Fonseca

Thesis approved in public session to obtain the PhD Degree in Biotechnology and Biosciences

Jury final qualification: Pass with distinction

2021

UNIVERSIDADE DE LISBOA

INSTITUTO SUPERIOR TÉCNICO

High-throughput Fourier-transform mid-infrared spectroscopy:
A mechanism-based screening approach to antibiotic discovery

Bernardo José Simões de Almeida Ribeiro da Cunha

Supervisor: Doctor Cecília Ribeiro da Cruz Calado

Co-Supervisor: Doctor Luís Joaquim Pina da Fonseca

Thesis approved in public session to obtain the PhD Degree in Biotechnology and Biosciences

Jury final qualification: Pass with distinction

Jury

Chairperson: Doctor Duarte Miguel de França Teixeira dos Prazeres, Instituto Superior Técnico, Universidade de Lisboa

Members of the Committee:

Doctor João Paulo Serejo Goulão Crespo, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa

Doctor Jorge Humberto Gomes Leitão, Instituto Superior Técnico, Universidade de Lisboa

Doctor Helena Maria Rodrigues Vasconcelos Pinheiro, Instituto Superior Técnico, Universidade de Lisboa

Doctor Nídia Dana Mariano Lourenço, UCIBIO, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa

Doctor Carla da Conceição Caramujo Rocha de Carvalho, Instituto Superior Técnico, Universidade de Lisboa

Doctor Cecília Ribeiro da Cruz Calado, Instituto Superior de Engenharia de Lisboa – Instituto Politécnico de Lisboa

2021

Printed in Portugal

Acknowledgments

First and foremost, I would like to show my greatest appreciation to my supervisors, Prof. Cecília Calado and Prof. Luís Fonseca, for their open-mindedness to new ideas and their ability to promote critical debates. Throughout these years I always felt I was given a high level of freedom and independence while pursuing this very ambitious project, which I believe was critical towards maximizing my personal growth and scientific productivity. Also, I never lacked the necessary support to bring this project to fruition.

During my stay at the R&D Laboratory in Health and Engineering of Instituto Superior de Engenharia de Lisboa (ISEL), I had the opportunity to share great times with great people. To them, a special thank you for the amazing collaborative environment and all the wonderful learning experiences. In particular, I would like to thank Luís Ramalhete, Paulo Zoio, Pedro Teixeira, Rúben Araujo and Viviana Caldeira. I had equally great times at Instituto Superior Técnico (IST), and within Amir Hassan, Cláudia Godinho, Marisa Silva, Pedro Pais and Rui Pacheco were critical for my integration into the amazing community that is IST.

Throughout this endeavor I also had the opportunity to oversee Ana Russo, Sara Gomes and Ana Correia during their master thesis work. I learned a lot from teaching you, and there is no doubt in my mind it made me a better person and a better researcher, and I hope to have contributed at least as much to your growth as scientists.

At ISEL, I was fortunate to work with Prof. Sandra Aleixo, with who I had the pleasure to collaborate in developing SpecAs' biomarker toolbox and became a great companion through the years. Also within ISEL, I would like to leave my gratefulness to Prof. Manuel Matos, whose institutional support was pivotal for this Ph.D. to even begin. At IST, I had the opportunity to collaborate with Prof. Maria da Conceição Olveira, Gonçalo Justino and Cátia Marques, who introduced me to mass spectrometry and guided me through my brief but intense experience with this amazing technique.

Many people also contributed to SpecA, my custom made spectra analyzer. A lot of work was put into building this amazing tool, which naturally required thousands of decisions. During this process, I always found someone to discuss and share my ideas with. Some of these people were Hugo Correia, Tomas Alegra, Marta Lopes, Pedro Sampaio Filipa Rosa and Kevin Sales.

Also, I would like to thank everyone that organized and participated in the Lab2Market@Tecnico2020 edition, it was an extraordinary experience that allowed me to see my scientific work from a different angle. As part of that experience, I cold-called many people, and I was surprised by so many who agreed to talk to me. No matter how brief, I found all of those conversations of extreme added value and tremendously inspirational. So, thank you Alastair Parkes, Alice Erwin, Glenn Tillotson, John Rex, Karen Shaw, Lynn Silver, Philippe Villain-Guillot and Stefan Miller.

I would like to extend my gratitude to everyone I had the pleasure to meet at Institute for Bioengineering and Biosciences of IST, and at the Departamento de Engenharia Quimica of ISEL. I could not have felt more at home, for which I would like to leave my gratefulness to all the individuals who enriched my personal experience. I would also like to extend my gratitude to Carla Noronha, for making me believe and for her contribution to my scientific foundations.

Finally, to all my loved ones: My wife and sons, for all the unconditional support and comprehension, for all the love every time I come back home; my Mother, such a profound life example; my Father, for watching out for me, even if worlds apart; my parents-in-law, who always made sure I had the best conditions to complete this Ph.D.; and my family and friends, for being there for me and making me smile.

Aos meus filhos e à minha mulher

Abstract

Since its golden age, antibiotic discovery has been slowing to a halt. In thirty years, antibiotic resistant infections are expected to cause the death of more people than cancer does now. A solution should stem from cell-based screening, the most successful first-in-class antibiotic discovery platform, which allows a 'brute force' approach more likely to produce short-term results. In cell-based screening, the mechanism of action (MOA) is identified after antimicrobial activity, so only potent candidates are detected. One option to increase the 'screenable' chemical space is to turn the page on the activity-based paradigm and shift to one based on mechanism. This requires probing the MOA of low-potency antibiotic candidates, which are often good candidates for medicinal chemistry programs, while ensuring sufficient throughput to screen large libraries. However, MOA identification is either slow, low-throughput, difficult to scale, costly, labor-intensive, or a combination thereof. A new discovery strategy requires a different technological approach to MOA identification, for which Fourier-transform infrared spectroscopy (FTIRS) is well-suited, but its full potential was yet to be investigated. In that regard, we have shown that FTIRS can identify the MOA of commercial antibiotics beyond the biosynthetic pathway level, and often in greater detail than the antibiotic class; we have determined the dose-response of MOA identification, which suggested FTIRS can discover hits from the grey chemical matter, effectively expanding the 'screenable' chemical space to low-potency compounds; we have demonstrated that potent compounds are identified, if not by their mechanism then by their inhibitory effect; and validated our protocol for the rapid exclusion of compounds with predominant off-target effects, which is relevant both during hit discovery or lead optimization, with two model organisms, *Escherichia coli* and *Staphylococcus aureus*. Furthermore, our fast and high-throughput protocol, entirely based on 96-well microtiter plates, is highly scalable, which enables screening large compound libraries. As such, FTIRS can fuel a new generation of mechanism-based screening assays that may swing the war on infectious diseases back in our favor.

Keywords: Antibiotic discovery; Chemometrics and machine learning; Fourier-Transform infrared spectroscopy; High-throughput screening; Mechanism of action.

Resumo

Desde a sua época de ouro, a descoberta de antibióticos tem vindo a estagnar. Estima-se que em menos de trinta anos, as infeções resistentes a antibióticos irão causar a morte de mais pessoas do que o cancro atualmente. Uma solução para este problema deve basear-se em ensaios de varrimento com células, a plataforma de maior sucesso na descoberta de antibióticos de primeira linha, e que permite uma abordagem de 'força-bruta' com maior probabilidade de produzir resultados a curto prazo. Nos ensaios baseados em células, o mecanismo de ação (MOA) é identificado depois da atividade antimicrobiana, portanto apenas compostos potentes são detetados. Aumentar o espaço químico detetável requer evoluir do paradigma baseado em atividade para um baseado no mecanismo. Isso requer a capacidade de varrer o MOA de compostos de baixa potência, que costumam ser bons candidatos para programas de química medicinal, com débito suficiente para varrer grandes bibliotecas. Normalmente, a identificação do MOA envolve técnicas lentas, de baixo rendimento, difíceis de escalar, caras, trabalhosas ou uma combinação das mesmas. Uma nova estratégia de descoberta de antibióticos requer uma abordagem tecnológica diferente para a identificação do MOA, para a qual a espectroscopia de infravermelho com transformada de Fourier (FTIRS) tem características adequadas, mas o potencial desta técnica ainda não foi esmiuçado. A esse respeito, mostrámos que a FTIRS pode identificar o MOA de antibióticos comerciais em maior detalhe do que a via biosintética, e por vezes do que a classe do antibiótico; determinámos a relação dose-resposta da previsão do MOA, o que sugere que a FTIRS pode identificar compostos da matéria química cinzenta, expandindo efetivamente o espaço químico detetável até compostos de baixa potência; demonstramos que compostos potentes são identificados, se não pelo mecanismo, então pelo efeito inibitório; e validámos o nosso protocolo para a rápida exclusão de compostos com efeitos predominantemente fora do alvo, seja durante a descoberta ou otimização de candidatos a antibióticos, com duas bactérias-modelo, *Escherichia coli* e *Staphylococcus aureus*. Além disso, o nosso protocolo é rápido e de alto débito, através de placas de placas de microcultura de 96 poços, e facilmente escalável, o que permite o varrimento de grandes bibliotecas de compostos. Como tal, a FTIRS pode estimular uma nova geração de ensaios de varrimento baseados no mecanismo que podem virar a guerra contra as doenças infeciosas de volta para nosso favor.

Palavras-chave: Descoberta de antibióticos; Espectroscopia de infravermelho com transformada de Fourier; Mecanismo de ação; Quimiometria e aprendizagem automática; Varrimento de alto débito.

Thesis Publications

Ribeiro da Cunha, B., Fonseca, L.P. & Calado, C.R.C. (2021) Simultaneous elucidation of antibiotic mechanism of action and potency with high-throughput Fourier-transform infrared (FTIR) spectroscopy and machine learning. *Appl Microbiol Biotechnol* 105. doi.org/10.1007/s00253-021-11102-7.

Ribeiro Da Cunha, B., Fonseca, L. P. and Calado, C. R. C. (2020) Metabolic fingerprinting with Fourier-transform infrared (FTIR) spectroscopy: Towards a high-throughput screening assay for antibiotic discovery and mechanism-of-action elucidation, *Metabolites*, 10(4). doi: 10.3390/metabo10040145.

Ribeiro da Cunha, B., Fonseca, L. P. and Calado, C. R. C. (2019) A phenotypic screening bioassay for *Escherichia coli* stress and antibiotic responses based on Fourier-transform infrared (FTIR) spectroscopy and multivariate analysis, *Journal of Applied Microbiology*, 127(6), pp. 1776–1789. doi: 10.1111/jam.14429.

Ribeiro da Cunha, B., Fonseca, L. P. and Calado, C. R. C. (2019) Antibiotic Discovery: Where Have We Come from, Where Do We Go?, *Antibiotics*, 8(2), pp. 1–21. doi: 10.3390/antibiotics8020045.

Ribeiro da Cunha, B., Aleixo, S. M., Fonseca, L.P. & Calado, C.R.C. Fast identification of off-target liabilities in early antibiotic discovery with Fourier-transform infrared (FTIR) spectroscopy (*Submitted*).

Ribeiro da Cunha, B., Fonseca, L. P. and Calado, C. R. C. (2020) Technologies for high-throughput screening of antibiotic mechanism of action *(In preparation).*

Book Chapter

Ribeiro da Cunha, B., Ramalhete, L., Fonseca, L. P., & Calado, C. R. C. (2020). Fourier-Transform Mid-Infrared (FT-MIR) Spectroscopy in Biomedicine. In Y. Tutar (Ed.), Essential Techniques for Medical and Life Scientists: A Guide to Contemporary Methods and Current Applications- Part II (pp. 1–39). doi: 10.2174/9789811464867120010004.

Additional publications

Marques, V., **Cunha, B.**, Couto, A., Sampaio, P., Fonseca, L. P., Aleixo, S., & Calado, C. R. C. (2019). Characterization of gastric cells infection by diverse *Helicobacter pylori* strains through Fourier-transform infrared spectroscopy. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 210, 193–202. doi.org/10.1016/j.saa.2018.11.001.

Sales, K. C., Rosa, F., **Cunha, B. R.**, Sampaio, P. N., Lopes, M. B., & Calado, C. R. C. (2017). Metabolic profiling of recombinant *Escherichia coli* cultivations based on high-throughput FT-MIR spectroscopic analysis. *Biotechnology Progress*, 33(2), 285–298. doi.org/10.1002/btpr.2378.

Rosa, F., Sales, K. C., **Cunha, B. R.,** Couto, A., Lopes, M. B., & Calado, C. R. C. (2015). A comprehensive high-throughput FTIR spectroscopy-based method for evaluating the transfection event: estimating the transfection efficiency and extracting associated metabolic responses. *Analytical and Bioanalytical Chemistry*, 407(26), 8097–8108. doi.org/10.1007/s00216-015-8983-9.

Conference Proceedings

Ribeiro da Cunha, B., Aleixo, S. M., Fonseca, L. P., & Calado, C. C. (2019). Towards an automated statistical workflow for biomarker screening in Fourier-transform infrared spectroscopy. *6th Portuguese Meeting on Bioengineering (ENBENG)*. 22-23 February 2019, Instituto Superior de Engenharia de Lisboa, Lisboa, Portugal. doi.org/10.1109/ENBENG.2019.8692544.

Ribeiro da Cunha, B., Fonseca, L. P., & Calado, C. R. C. (2017). High-throughput bioassay for mechanism of action determination of antibacterial drugs. *5th Portuguese Meeting on Bioengineering (ENBENG)*. 16-18 February 2017, Faculty of Medicine, University of Coimbra, Coimbra, Portugal. doi.org/10.1109/ENBENG.2017.7889478.

Ribeiro da Cunha, B., Russo, A., Fonseca, L., & Calado, C. (2017). Optimization of a bioassay to evaluate Escherichia coli stress responses. *5th Portuguese Meeting on Bioengineering (ENBENG)*. 16-18 February 2017, Faculty of Medicine, University of Coimbra, Coimbra, Portugal. doi.org/10.1109/ENBENG.2017.7889424.

Rosa, F. O. P.**, Ribeiro da Cunha, B.**, Carmelo, J. G., Fernandes-Platzgummer, A., Da Silva, C. L., & Calado, C. R. C. (2017). Mid-infrared spectroscopy: A groundbreaking tool for monitoring mammalian cells processes. *5th Portuguese Meeting on Bioengineering (ENBENG)*. 16-18 February 2017, Faculty of Medicine, University of Coimbra, Coimbra, Portugal. doi.org/10.1109/ENBENG.2017.7889479.

Sales, K.**, Cunha, B. R.**, & Calado, C. R. C. (2017). Biopharmaceuticals process monitoring based on infrared spectroscopy according to Quality by Design. *5th Portuguese Meeting on Bioengineering (ENBENG)*. 16-18 February 2017, Faculty of Medicine, University of Coimbra, Coimbra, Portugal. doi.org/10.1109/ENBENG.2017.7889460.

Ribeiro da Cunha, B., Sampaio, P. N., & Calado, C. R. C. (2015). Development of a high-throughput monitoring technique of bacteria photodynamic inactivation. *4th Portuguese Meeting on Bioengineering (ENBENG)*. 26-28 February 2015, Faculty of Engineering, University of Porto, Porto, Portugal. 10.1109/ENBENG.2015.7088827

B.R. Cunha, K.C. Sales, F. Rosa, M.B. Lopes, P.N. Sampaio, C.R.C. Calado (2013). Evaluation of anti-bacterial agents using a high-throughput FTIR spectroscopy. *Micro-Biotec13, Portuguese Congress of Microbiology and Biotechnology*. 6-8 December 2013, Aveiro University, Aveiro, Portugal.

Table of Contents

List of abbreviations

6-APA – 6-aminopenicillanic acid

7-ACA – 7-aminocephalosporanic acid

ADP(s) – Antibiotic discovery platform(s)

AEC – Average eigenvalue criterion

AGS – Adenocarcinoma gastric cells

ANN – Artificial neural networks

ATR – Attenuated total reflectance

AUC – Area under the curve

AUROC – Area under the receiver operating characteristic curve

BCP – Bacterial cytological profiling

BIB – Band importance in biomarker

CAEC – Corrected average eigenvalue criterion

CFU – Colony forming unit

CLT – Central limit theorem

CRISPR – Clustered regularly interspaced short palindromic repeats/dCas9

DMSO – Dimethyl sulfoxide

DPT – Data point table

EMSC – Extended multiplicative scatter correction

EMSCrep – Extended multiplicative scatter correction with replicate correction

ERMSPE – Estimated root mean squares prediction error

FPA – Focal plane array

FTIR – Fourier-transform infrared

FTIRS – Fourier-transform infrared spectroscopy

GC – Gas chromatography

GFP – Green fluorescent protein

GPR – Gaussian process regression

GUI – Graphic user interface

HCA – Hierarchical cluster analysis

HCS – High-content screening

HTS – High-throughput screening

IR – Infrared

iTRAQ - Isobaric tags for relative and absolute quantification

KCA – *k*-means clustering

KNN – *k* nearest neighbor

LC – Liquid chromatography

LDA – Linear discriminant analysis

LMSC – Loopy multiplicative scatter correction

LOO-CV – Leave-one-out cross-validation

LV(s) – Latent variable(s)

MAE – Mean absolute error

MIC – Minimum inhibitory concentration

MIR – Mid-infrared

MIRS – Mid-infrared spectroscopy

MOA – Mechanism of action

MS – Mass spectrometry

MSC – Multiplicative scatter correction

NIR – Near-infrared

NIRS – Near-infrared spectroscopy

NMR – Nuclear magnetic resonance

ORF – Open reading frame

PC(s) – Principal component(s)

PCA – Principal component analysis

PCR – Principal component regression

PLS – Partial least squares

PLSDA – Partial least squares discriminant analysis

PLSR – Partial least squares regression

PR – Precision-recall

QDA – Quadratic discriminant analysis

ROC – Receiver operating characteristic

ROS – Reactive oxygen species

RS – Raman scattering

RSS – Raman scattering spectroscopy

SEIRA – Surface enhanced infrared absorption

SG – Savitzky-Golay

SVM – Support vector machine

VIP – Variable importance projection

VS – Vibrational spectroscopy

Chapter I

Introduction

The discovery of antibiotics has been considered a miracle of modern medicine. However, the world is on the verge of not having effective treatments for many common infectious diseases. It is estimated that by 2050, antibiotic resistant bacteria will claim more lives than cancer does now, with an associated cumulative cost of inaction over $100 trillion. These grim estimates, and our inability to invert the situation, have made antibiotic resistant pathogens a top priority for the World Health Organization.

The 'war' on infectious diseases has taken multiple fronts. This involves increasing public awareness, monitoring and minimizing the spread of resistant pathogens, limiting the use of antibiotics outside clinical practice, developing faster diagnostics to reduce the unnecessary use of antimicrobials, establishing novel therapeutic approaches to infectious diseases like vaccines, as well as increasing the arsenal of antimicrobials available. All of these approaches are extremely important, but the treatment of microbial infections will still require compounds capable of killing or halting the proliferation of pathogens, e.g., antibiotics. Therefore, the discovery of new antibiotics is at the very core of the 'war' on infectious diseases, and new advances are required to invert the alarmingly low rates of discovery.

In order to understand where to go, it is important to appreciate where we come from. As such, **Chapter II** presents the discovery timeline of the major antibiotic classes, highlighting the underlying approaches and the context of their application. Starting at the birth of chemotherapy, the importance of the Waksman platform in fueling semi-synthesis and fully synthetic antibiotics is discussed, followed by the technological revolution brought by the genomics era, and the present-day efforts in the post-genomics era. Historically, the evolution of antibiotic resistance has been outpaced with structural modifications of antibiotic scaffolds found in nature, most of which were discovered with the Waksman platform.

High-tech solutions brought by the genomics boom shifted the paradigm to target-based screening, but this proved lackluster and cell-based screening was revived. Although this strategy has been extremely successful in the past, the steady decline of the antibiotic pipeline makes it clear that the 'low-hanging fruit' has been plucked. Nevertheless, various meta-omics studies have shown that the large majority of nature's antibiotic repositories remain untapped, which suggests this strategy is not necessarily exhausted, but a different approach is required. One that is capable of consistently and efficiently delivering new scaffolds.

There is a growing tendency for cell-based assays to evolve from measuring antimicrobial activity towards probing the mechanism of action (MOA) of antibiotics. Despite the plethora of high-tech approaches that have been applied in the context of antibiotic discovery, the MOA of antibiotics remains particularly elusive. In that regard, **Chapter III** reviews technologies that have been used for high-throughput MOA identification. Technologies that can convey molecular information useful for mechanistic determinations, with sufficient throughput for screening large libraries of compounds, could enable shifting antibiotic discovery from an activity-based paradigm to one centered on MOA.

A mechanism-based approach is particularly relevant because it complements traditional activity-based assays, where MOA identification is frequently a bottleneck, thereby increasing the probability of success of antibiotic discovery programs; secondly, a mechanism-based approach can be used to rapidly exclude compounds with predominant off-target effects, i.e., toxic compounds, either during the early stage of hit discovery or later during lead optimization, where off-target liabilities resulting from medicinal chemistry programs can quickly be identified by monitoring MOA profiles; and thirdly, perhaps the most important advantage, it expands the 'screenable' chemical space by probing the grey chemical matter, which are 'not-so-low-hanging' compounds capable of inducing some level of phenotypic modulation, but without sufficient potency to induce cell death or inhibit growth.

One promising technology amendable to a mechanism-based antibiotic discovery approach is Fourier-transform infrared (FTIR) spectroscopy (FTIRS), an established high-throughput technique that reveals the biochemical composition of samples. Hence, **Chapter IV** discusses the fundamentals of FTIRS as a vibrational spectroscopy tool, including the advantages brought by equipment's with Michelson interferometers combined with the Fourier-transform, the types of samples and associated detection modes, as transmission and transflection, high-throughput measurements using micro-plates in transmission mode, and fibre optic probes coupled to attenuated total reflection detection.

Due to the complex biochemical composition of biological samples, mid-infrared spectra are usually very difficult to interpret without the application of complex and sophisticated mathematical and statistical analysis routines, such as: spectra preprocessing, to minimize noise and other non-informative data that compromise subsequent pattern recognition or regression models; deconvolution methods to resolve overlapped spectral bands; dimensionality reduction and feature extraction; supervised and non-supervised pattern

recognition, as support vector machines and artificial neural networks. Therefore, a review of these routines was also presented alongside their application to different datasets.

Another integral part of FTIRS studies is the data analysis workflow, which often has a determining role in the outcome of experiments. In the absence of a fully satisfactory software, SpecA was created. **Chapter V** revolves around SpecA, a graphic user interface built to handle the data analysis pipeline of an infrared spectroscopy experiment while maximizing the biological information retrieved from the data. In particular, SpecA facilitates loading and handling large datasets, allows for an in-depth appraisal of the data quality through quality control routines, enables the comparison and optimization of preprocessing combinations and parametrizations, performs operations such as calculating average or difference spectra, and allows easy visualization of intrinsic patterns by means of various chemometrics procedures such as Principal Component Analysis (PCA) or Hierarchical Cluster Analysis (HCA).

Moreover, SpecA is fully compatible with MATLAB, therefore the transfer of datasets to other tools for more advanced machine learning algorithms is stress-free; SpecA can easily export the dataset at any point in the workflow to a cross-platform compatible Excel file; generates figures suitable for publication, as done for most of the figures in this thesis; includes an automated statistical workflow for univariate biomarker screening; among other essential tasks than can make the difference in the outcome of a scientific experiment. Ultimately, SpecA became the powerhouse that catalyzed the scientific productivity of this Ph.D.

Having researched a problem, suggested a solution, and prepared the tools for its implementation, the next step was its execution. In other words, to validate FTIRS as a mechanism-centered approach for antibiotic discovery. For that, a phenotypic screening assay for stress and antibiotic responses was developed in **Chapter VI**. Here, bacterial cells were stressed at sub-inhibitory concentrations, which can be seen as an extreme case of the grey chemical matter since some level of phenotypic modulation occurred, but without sufficient potency to induce cell death. Because FTIRS bioassays require considerable optimizations to maximize the biological relevance of the molecular information, the effect of nutrient content, bacterial growth phase and stress agent exposure time were evaluated. SpecA was used to apply various preprocessing strategies, and evaluate their effect on unsupervised chemometrics algorithms, namely a combination of PCA-HCA.

This simple data analysis approach was preferred for two reasons: firstly, it is of easier interpretation than more advanced algorithms, such as support vector machines or neural

networks, where the relationships within the biological samples being studied is harder to grasp, while still providing a quantitative comparison of the effect of bioassay parameters on both metabolic resolution and reproducibility; and secondly, using unsupervised algorithms ensures none of the observed data patterns were preferentially highlighted, which consolidated the notion that FTIR spectra have intrinsic patterns that reflect the phenotypic modulation bacteria undergo as part of stress and antibiotic responses. Moreover, using SpecA, spectra were partitioned into biologically relevant regions, which suggested complete spectra provide more informative metabolic signatures for phenotypic screening of both stress and antibiotic responses. Ultimately, this study served as a proof-of-concept of FTIRSs' suitability to explore the grey chemical matter.

In addition to the ability to detect the underlying biochemical alterations of bacterial stress responses, further steps were required towards validating FTIRS as a mechanism-centered approach in antibiotic discovery. Therefore, **Chapter VII** revolved around accurately differentiating the MOA of antibiotics, for which a macro-cultivation protocol coupled with high-throughput spectra acquisition was used. Firstly, a method to quantitatively evaluate the effect of different preprocessing combinations on the successful classification of PLSDA after Leave-One-Out Cross-Validation (LOO-CV) was established. This method was used to objectively select the optimal preprocessing combination that maximized the predictive performance of a PLSDA model of the major biosynthetic pathway targeted by 15 antibiotics.

SpecA enabled a detailed analysis of the score plots of both PLSDA and PCA, which not only revealed the similarities between metabolic fingerprints induced by antibiotics acting on the same biosynthetic pathway, but simultaneously divulged that antibiotics that disturb the same pathway via different mechanisms have sufficiently distinct fingerprints to allow their differentiation, which suggested FTIR's MOA resolution extends beyond the pathway level. Secondly, the coherence observed between PLSDA and PCA indicated that the observed results were not due to the known issues of PLSDA, but rather a reflection of the intrinsic spectral patterns induced by the different antibiotics.

Although the potential of FTIRS as a tool for antibiotic discovery has been unraveled, a deeper examination of its performance for MOA prediction was required. Similarly, it was vital that the bioassay be adapted to a higher throughput design, using micro-cultivation 96-well plates, to validate its application for screening large libraries. As such, **Chapter VIII** firstly focused on the application of machine learning algorithms towards predicting the MOA at the

level of the biosynthetic pathway, class, and individual mechanism of different antibiotics. Importantly, machine learning models were used to predict MOA both in the case of known MOAs, where the models were trained with similar MOA profiles; and, in the case of simulated novel MOAs, where the profiles induced by individual antibiotics were withheld during model training. Prior to machine learning, the previously mentioned method based on the LOO-CV successful classification of a PLSDA was used for preprocessing optimization.

Then, we further consolidated the possibility of using FTIRS for exploring the grey chemical matter by determining the dose-response of MOA prediction. This revealed that FTIR spectra have intrinsic data patterns that are relevant for predicting antibiotics MOA, even at very low antibiotic concentrations that only induce slight growth inhibition. Moreover, we aimed to determine if our assay could predict if a candidate molecule had an antibiotic effect. For that, we regressed microbial growth from the spectra, which revealed the possibility of simultaneously estimating antimicrobial activity from the same samples from which MOA was predicted. Importantly, because samples were normalized in regard to biomass prior to spectra acquisition, it seems FTIRS is not only highly accurate in predicting the MOA of antibiotics, but also to the degree that said MOA is affecting bacterial metabolism and inhibiting bacterial growth, beyond cell density.

As the final step of validating FTIRS as a mechanism-centered approach for antibiotic discovery, **Chapter IX** gaged the possibility of identifying compounds with off-target activity, and the monitorization of MOA profiles. For that, we further improved our bioassay by minimizing the number of steps and reducing their duration, while ensuring accurate MOA identification. This translated to a cycle time within a one-shift timeframe (~8h), a hands-on time of 1h, and a highly scalable protocol. This is particularly relevant because a mechanism-based discovery paradigm requires not only mechanistic information, but for that information to be achieved with sufficient throughput for screening purposes.

We used our protocol to capture the metabolic fingerprints induced by a combination of antibiotics and chemical stressors on *Staphylococcus aureus* and *Escherichia coli*. Firstly, we applied SpecAs' biomarker screening functionality to identify the spectral bands, with known biochemical associations, that more often yielded more significant biomarkers of off-target effects. Then, we used the data analysis pipeline built thus far, consisting of spectral preprocessing optimization with the previously described LOO-CV of a PLSDA, followed by various machine learning algorithms, to derive highly discriminative models to distinguish

compounds with predominant off-target effects from antibiotics with well-defined MOA, and from the latter predict their MOA. In addition to showing that our protocol has sufficient sensitivity for the proposed application, we have demonstrated that it is not overly sensitive to subtle shifts in metabolism. For instance, our assay is not disturbed by the metabolic effect of DMSO, nor does it require additional steps to stabilize the bacterial metabolism.

To close, **Chapter X** contextualizes the work developed in this Ph.D. in regard to the advances made, and future work to be addressed.

Chapter II

Antibiotic discovery: Where do we come from, where do we go?

This chapter is adapted from the manuscript:

Ribeiro da Cunha, B., Fonseca, L. P. and Calado, C. R. C. (2019) 'Antibiotic Discovery: Where Have We Come from, Where Do We Go?', *Antibiotics*, 8(2), pp. 1–21. doi: 10.3390/antibiotics8020045.

Author contribution

Bernardo Ribeiro da Cunha reviewed the literature, prepared the original draft, reviewed and edited its final version.

Abstract

Given the increase in antibiotic-resistant bacteria, alongside the alarmingly low rate of newly approved antibiotics for clinical usage, we are on the verge of not having effective treatments for many common infectious diseases. Historically, antibiotic discovery has been crucial in outpacing resistance and success is closely related to systematic procedures—platforms—that catalyzed the antibiotic golden age, namely the Waksman platform, followed by the platforms of semi-synthesis and fully synthetic antibiotics. Said platforms resulted in the discovery of the major antibiotic classes: aminoglycosides, amphenicols, ansamycins, beta-lactams, lipopeptides, diaminopyrimidines, fosfomycins, imidazoles, macrolides, oxazolidinones, streptogramins, polymyxins, sulfonamides, glycopeptides, quinolones and tetracyclines. During the genomics era came the target-based platform, mostly considered a failure due to limitations in translating drugs to the clinic. Therefore, cell-based platforms were re-instituted, and are still of the utmost importance in the fight against infectious diseases. Although the antibiotic pipeline is still lackluster, especially of new classes and novel mechanisms of action, there is an increasingly large set of information available on microbial metabolism in the post-genomic era. The translation of such knowledge into novel platforms will hopefully result in the discovery of new and better therapeutics, which can sway the war on infectious diseases back in our favor.

Contents

II. 1 INTRODUCTION – THE DESPERATE NEED FOR NEW ANTIBIOTICS

Infectious diseases have been a challenge throughout the ages. From 1347 to 1350, approximately one-third of Europe's population perished to Bubonic plague. Advances in sanitary and hygienic conditions sufficed to control further plague outbreaks. However, these persisted as a recurrent public health issue. Likewise, infectious diseases in general remained the leading cause of death up to the early 1900s, e.g., accounting for 25% of England's mortality. However, by the mid-1900s, the mortality of infectious diseases in England shrunk to under 1% after the commercialization of antibiotics [1]. Given their impact on the fate of mankind, these are regarded as a 'medical miracle'. Moreover, the non-therapeutic application of antibiotics has also greatly affected humanity, for instance those used as livestock growth promoters to increase food production after World War II.

The term 'antibiotic' was introduced by Selman Waksman as any small molecule, produced by a microbe, with antagonistic properties on the growth of other microbes [2]. An antibiotic interferes with bacterial survival via a specific mechanism of action (MOA) but more importantly, it is sufficiently potent at therapeutic concentrations to be effective against infection and simultaneously presents minimal toxicity. Most antibiotic classes in use today were identified in the 1940–1960s, a period referred to as the antibiotic golden age. During this period, it was common belief that, given the antibiotics discovered, and particularly their rate of discovery, infectious diseases would soon be a controlled public health issue [3,4]. In fact, in 1970, the US Surgeon General stated "It's time to close the book on infectious diseases… and shift national resources to such chronic problems as cancer and heart disease" [5].

Currently, more than 2 million North Americans acquire antibiotic resistant infections every year, resulting in 23,000 deaths [6]. In Europe, nearly 700 thousand cases of antibiotic-resistant infections directly develop into over 33,000 deaths yearly [7], with an estimated cost over €1.5 billion [8]. Despite a 36% increase in human use of antibiotics from 2000 to 2010 [9], approximately 20% of deaths worldwide are related to infectious diseases today [10]. This situation deteriorated further as nosocomial infections became a leading cause of morbidity and mortality [11], resulting in lengthier hospital stays and increased health care costs [12]. Furthermore, over 15% of nosocomial infections are already caused by multidrug-resistant pathogens [13] — for some of which, there are no effective antimicrobials [14]. Future perspectives are no brighter, for instance, a government-commissioned study in the United Kingdom estimated 10 million deaths per year from antibiotic resistant infections by 2050 [15].

The increase in drug-resistant pathogens is a consequence of multiple factors, including but not limited to high rates of antimicrobial prescriptions, antibiotic mismanagement in the form of self-medication or interruption of therapy, and large-scale antibiotic use as growth promotors in livestock farming [16]. For example, 60% of the antibiotics sold to the USA food industry are also used as therapeutics in humans [17]. To further complicate matters, it is estimated that $200 million are required for a molecule to reach commercialization [18], with the risk of antimicrobial resistance rapidly developing, crippling its clinical application, or on the opposing end, a new antibiotic might be so effective it is only used as a last-resort therapeutic, thus not widely commercialized. Either way, the bottom line implies similar risks with considerably lower returns on investment compared with other drugs [19], which renders antibiotic discovery as an unattractive business. In an attempt to counter this scenario, the European Federation of Pharmaceutical Industries and Associations consorted with the European Union to establish the largest worldwide life sciences public–private partnership, the Innovative Medicines Initiative. Through funding and a highly ambitious agenda, under the New Drugs for Bad Bugs program, this initiative encourages action in areas ranging from antibiotic discovery, clinical research, through to reshaping the use of antibiotics, in hopes of catalyzing the approval of novel antibiotics [20].

The systematic procedures—Antibiotic Discovery Platforms (ADPs)—behind the discovery of major antibiotic classes that fueled the antibiotic golden age, have become exhausted. Modern ADPs have yielded redundant discoveries and/or failed in translation to the clinic, which dimmed the overly optimistic expectations created with the development of novel technologies throughout the genomics era. From 2004–2009, the overall rate of antibacterial approval was a mindboggling single drug per year [21], which doubled from 2011–2014 when the FDA approved a still impressively scarce eight new antibiotics or combinatorial therapies [22]. According to the antibiotic pipeline surveillance by The Pew Charitable Trusts [23], from 2014 onwards, the situation is slowly improving, if at all. As seen on Figure 1, the total antibiotic pipeline appears to be timidly increasing, although the number of drug candidates close to approval (phase III clinical trials and those that have filed a New Drug Application) or recently approved (phase IV) remains alarmingly low. Despite great efforts, most approved antibiotics only target either the ribosome, cell wall synthesis machinery and DNA gyrase or topoisomerase [24,25]. Beyond conventional antibiotics, some interesting therapeutic alternatives are noteworthy, including bacteriophages, antivirulence strategies, probiotics, vaccines, immune stimulation, antimicrobial peptides, antibiofilm therapies and antibodies, among others. Despite some of these alternatives having reached clinical trials, it is estimated

that across the next 10 years, over £1.5 billion will be needed to further test and develop them before their clinical impact is felt [26].

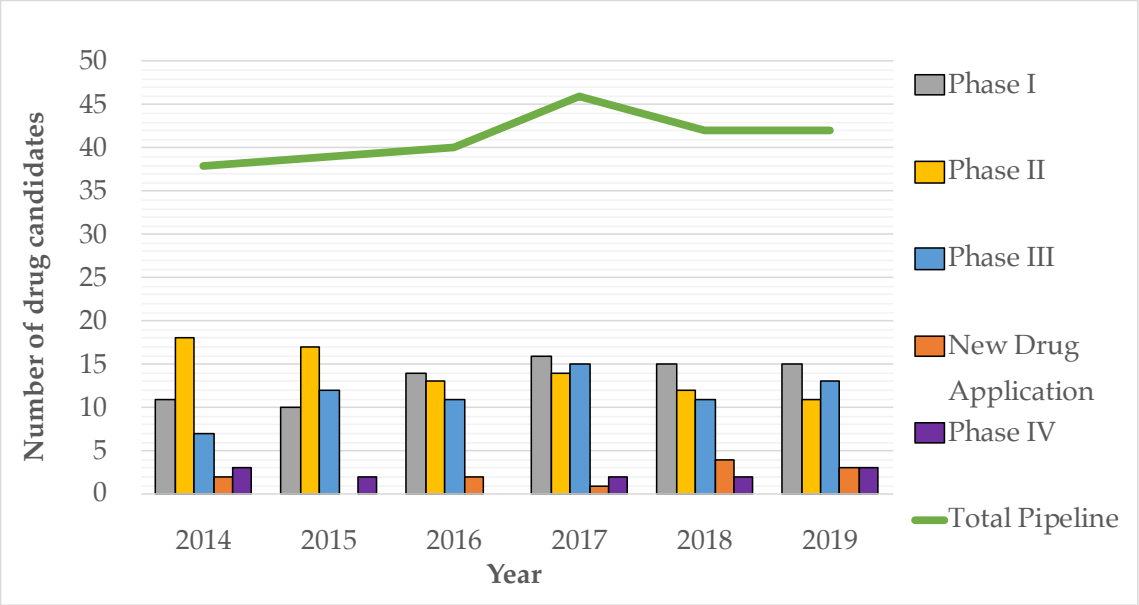


Figure 1. Evolution of the antibiotic pipeline by stage of development, which includes: Clinical trials ranging from Phase I, to evaluate safety; Phase II, to access effectiveness and safety; Phase III, to gather statistically significant data on safety, effectiveness and benefits-versus-risk; submission of a New Drug Application, for marketing approval; and lastly, Phase IV for post-marketing surveillance.

During 2011, the director general of the World Health Organization made the clear forewarning that we are "on the brink of losing these miracle cures… In the absence of urgent corrective and protective actions, the world is heading towards a post-antibiotic era, in which many common infections will no longer have a cure and, once again, kill unabated" [27]. Besides a more efficient management of antibiotic use, there is a pressing need for new platforms capable of consistently and efficiently delivering new lead substances, which should attend to their precursors impressively low rates of success, in today's increasing drug resistance scenario. The present manuscript reviews the discovery timeline of the major antibiotic classes from an ADPs perspective, highlighting their underlying technological basis and the context of their application, beginning with the birth of chemotherapy, the establishment of the Waksman platform, semi-synthesis and fully synthetic antibiotics, followed by the technological revolution during the genomics era, and the present-day efforts in the post-genomics era.

II. 2 THE BIRTH OF ANTIMICROBIAL CHEMOTHERAPY

Traditional behaviors and primitive rituals suggest ancient human use of antibiotics [28,29], although the first scientific record of the therapeutic use of antibiotics dates to 1899 when Emmerich and Löw explored the therapeutic potential of *Pseudomonas aeruginosa* extracts. While their investigation was discontinued due to inconsistent effects, the antimicrobial effect observed was later associated with quorum-sensing molecules [30]. When discussing antimicrobial chemotherapy, highlighting the contributions of Pasteur, Lister and Koch to the foundation of medical microbiology is a tribute of sorts, for which reviews are available [31–34]. The road towards the first modern antimicrobial began in 1854, when Antoine Béchamp produced aniline, via the reduction of nitrobenzene with iron in the presence of hydrochloric acid. In 1859, Béchamp produced atoxyl, by reacting aniline with arsenic acid, in his pursuit of developing aniline derivatives. Simultaneously, Paul Ehrlich noticed that chemical dyes stained specific histological and cellular structures, which inspired his side-chain theory in 1897, where he hypothesized about therapy targeting structures exclusive to pathogens [35]. Ehrlich, together with Alfred Bertheim and Sahachiro Hata, synthetized and screened multiple arsenical derivatives based on Béchamp's discovery of atoxyl and, by 1907, discovered Salvarsan [36], the first antimicrobial that was a more effective and safer therapeutic against syphilis, and became the most prescribed drug until the introduction of penicillin [37].

The systematic application of chemical modifications to expand a library of lead molecules, followed by screening its effect on a disease model, contributed to the discovery of Neosalvarsan, a more water-soluble derivative with reduced side effects, and laid the foundation of modern pharmaceutical research. Given its success, the Friedrich Bayer Company explored synthetic chemicals for therapeutic purposes in the 1920s [38]. The azo compounds, a class of synthetic dyes with antibacterial activity, were the starting point for the synthesis of diverse structural variants. In 1932, Gerhard Domagk recognized the curative potential of Prontosil, synthetized by chemists Josef Klarer and Fritz Mietzsch, from studies on streptococci-infected mice and later, on two dire cases of children in life-threatening situations, including Domagk's daughter. Prontosil became commercially available by 1935, simultaneous to the discovery of its active principle, which was unrelated to the azo functional group or the dye fraction. In fact, Prontosil is a precursor to the active molecule, sulfanilamide, widely used in the dye industry, hence not patentable, and whose synthesis was readily achievable. In the following years, over 5000 derivatives known as the sulfa drugs were synthetized, some of which are still used today, e.g., sulfamethoxazole.

Arguably, it was Alexander's Fleming 'accidental' detection of Staphylococci growth inhibition around mold colonies in petri dishes, forgotten at his lab throughout a holiday period, that mostly impacted the future of antimicrobial discovery [39]. Fleming's observation in 1928 motivated his studies on the mold's product, penicillin, regarding its activity spectrum, potency, leukocyte interaction and toxicity. In fact, it was the first substance noted to present more antibacterial than antileukocytic activity [40]. Fleming's rigorous methods, and their underlying rational, are still hallmarks for antibiotic discovery. Nonetheless, Fleming faced problems associated with the large-scale growth of the penicillin-producing mold and it was not until 1939 that Howard Florey, Norman Heatly and Ernst Chain described a method that made penicillin sufficiently available for clinical testing. This bioprocess was greatly up-scaled when Florey and Heatly moved to the USA and Canada, given the necessity of antibiotics imposed by World War II [41]. Ultimately, their work on bioprocess optimization surpassed the production of penicillin and sparked the fermentation industry, which is highly relevant for the production of diverse antibiotics and other medicines such as insulin, erythropoietin, interferon, and antibodies, among others [42].

Although penicillin's bioprocess scale-up breakthroughs enabled its widespread clinical use during the late period of World War II, efforts pursued an outperforming chemical synthesis protocol. During late 1945, penicillin antimicrobial activity was traced to the β-lactam ring [43]. Ernst Chain believed that fully synthetic penicillin would require new chemical techniques that John Sheehan achieved in 1950, and from which the first synthetic natural penicillin V was produced in 1957. The year after, Sheehan described the production of 6-aminopenicillanic acid (6-APA) via both semi- and fully synthetic methods, and 6-APA became a scaffold for multiple C6 sidechain modifications, further discussed ahead in the context of semi-synthesis.

II. 3 TOWARDS THE GOLDEN ERA: THE WAKSMAN PLATFORM

Impelled by the remarkable successes at the beginning of the 20th century, Selman Waksman adventured into the realm of drug discovery. In 1937, noticing that complex soil bacteria—actinomycetes—inhibited the growth of other bacteria, Waksman acknowledged that these biological mechanisms, which evolved from competitive growth [44], could become the conceptual basis of a screening platform for antibiotic-producing organisms [45]. From 1939 onwards, it is estimated that his systematic agar overlay process, referred to as the Waksman platform, screened well over 10,000 strains of different microbes [46], which exemplifies the scalability of this method—a key characteristic for the coming successes. Equally important,

over 90% of clinical antibiotics derive from actinomycetes [10], making these microbes an antibiotic gold mine of sorts.

The Waksman platform promptly revealed new antimicrobials: actinomycin, streptothricin, fumigacin and clavacin; but it was not until 1944 that a *Streptomyces griseus* strain was found to produce a non-toxic aminoglycoside antibiotic, named streptomycin, which inhibits protein synthesis by binding to the bacterial 30S ribosomal subunit. At the time, it was not possible to patent natural products in the USA, but together with Merck lawyers, Waksman convinced the authorities that purified antibiotics were sufficiently distinct, sparking a new range of business opportunities, a significant stride towards an economic stimulus that bolstered the antibiotic golden age. Merck obtained FDA approval for streptomycin [44] and began its commercialization by 1946 for the treatment of tuberculosis and tuberculous meningitis, and later for pathogens outside penicillin's spectrum of activity [47]. The Waksman platform revealed various antibiotic classes, many of which are the major antibiotic classes currently in clinical use, as described next.

Chloramphenicol was originally isolated in 1947 from the actinomycete *Streptomyces venezuelae*, thereby introducing the amphenicol class. Chloramphenicol's antimicrobial activity derives from its reversible binding to the 50S ribosomal subunit, thereby inhibiting bacterial protein synthesis. It was the first FDA-approved broad-spectrum antibiotic, displaying excellent tissue and fluid permeability. However, in the 1960s, various toxicity issues impaired its administration, and it is currently rarely prescribed [48]. Chlorotetracycline marked the introduction of the tetracycline antibiotic class in 1948, which also disrupts protein synthesis by acting on the 30S subunit of the ribosome. Chlorotetracycline, a product of *Streptomyces aureofaciens*, is unstable at both ends of the pH scale, which hampers its bioavailability [49].

Macrolides are the second most prescribed class of therapeutic antibiotics, introduced in 1949 with erythromycin that is produced by *Saccharopolyspora erythrea*. Erythromycin binds to the 50S bacterial ribosomal subunit, but its therapeutic use was characterized by instability under acidic conditions and overall poor oral bioavailability [50]. Virginiamycin was the first identified streptogramin, originally isolated from *Streptomyces virginiae* in 1952. Streptogramins are a class of antibiotics formed by two chemically unrelated substances, a polyunsaturated macrolactone and a cyclic hexadepsipeptide. Either group binds to the 50S subunit of bacterial ribosomes, presenting mediocre activity, but their synergistic effect empowers its therapeutic application [51].

Unlike the antibiotic classes described thus far, which target bacterial protein synthesis, glycopeptides disrupt cell wall synthesis. The first antibiotic of the glycopeptide class, vancomycin, was discovered in 1956 to be produced by *Amycolatopsis orientalis* and is currently a last-resort antibiotic. Vancomycin interferes with the transpeptidation and transglycosylation steps of the cell wall synthesis, thereby inhibiting cross-linking and cell wall maturation [52]. Similarly, ansamycins differ from protein synthesis inhibitors (e.g., amphenicols, tetracyclines, macrolides and streptogramins) and cell wall synthesis inhibitors (e.g., glycopeptides). For instance, rifamycins inhibit the DNA-dependent RNA polymerase of prokaryotes. Rifamycin B was first isolated in 1959 from *Streptomyces mediterranei* (later classified as *Amycolatopsis mediterranei*), and despite considerably low antimicrobial effect, it introduced a unique metabolic target in bacteria [53]. The discovery of fosfomycin came from the isolation of three Streptomyces strains in 1969. Its antimicrobial effect derives from the inhibition of the initial steps of the cell wall biosynthesis pathway, disrupting the action of phosphoenolpyruvate synthetase. However, fosfomycin presents a broad spectrum of activity, making it an appealing antimicrobial [54].

Although the first report of a lipopeptide antibiotic dates to 1947 with the discovery of polymyxin E, produced by *Paenibacillus polymyxa*, the therapeutic use of this class was limited to experimentations for a mere couple of years, given multiple concerning adverse effects, but was eventually reconsidered [55]. The production of daptomycin by *Streptomyces roseosporus* was revealed in 1980 and although Eli Lilly and Co. attempted its commercialization, clinical trials were discontinued under the belief that there was a small window between therapeutic efficacy and toxicity. As such, this calcium-dependent cyclic lipopeptide is seen as the precursor of the lipopeptide class of antibiotics, with surpassing antimicrobial activity in comparison with polymyxin E, albeit limited to Gram positive pathogens. Interestingly, daptomycin was revived by Cubist Pharmaceuticals and, with dosing adjustments, reached the market by 2003 [56]. Moreover, daptomycin's mode of action is still unclear: permeabilization and depolarization of the cell membrane being the most probable; interference in cell wall synthesis; and/or disruption of cellular division are other suggestions. Although more cyclic lipopeptides have been described, daptomycin remains the only approved therapeutic antibiotic of this class [57].

II. 4 ONTO THE MEDICINAL CHEMISTRY ERA: SEMI-SYNTHESIS

Antibacterial semi-synthesis is the modification of existing scaffolds, or molecular backbones, obtained by a fermentative procedure. Historically, most scaffolds originated from the Waksman platform. Thus, they are the evolutionary outcome of selective pressures, e.g., from the actinomycete-bacteria 'fight', and are therefore extremely well fit to reach and bind to their target. However, this does not translate to therapeutic effectiveness or safety, which can often be improved by means of semi-synthesis, alongside its chemical stability, reduction of undesirable side effects, among other features that are crucial in marketing antibiotics, for instance patenting derivatives, which increases profitability of antibiotic development programs, essential for this generally unattractive business. Semi-synthesis began with the catalytic hydrogenation of streptomycin, which resulted in dihydrostreptomycin by 1946, and was characterized by greater chemical stability along with similar antimicrobial activity. Although both streptomycin and its novel derivative quickly made their way to clinical use, eventually their prescription has been reevaluated due to ototoxicity concerns [58].

Conversely, it took over a decade before a bioproduction method made penicillin a therapeutic possibility. While the identification of penicillin's antimicrobial effect preceded its 'discovery', it was Fleming's willpower that pushed penicillin beyond only being obtainable in small and unstable quantities. This in turn enabled its semi-synthesis, which expanded penicillin from a single drug to a range of semi-synthetic derivatives constituting an entire class of antibacterial drugs, the beta-lactams. These comprise over 60% of antibiotics for human use [59], with a multitude of subclasses and marketed antibiotics within, as seen in Table 1. The rate at which derivatives with improved properties can be synthetized kept the upper hand against infectious diseases, a key characteristic of semi-synthesis. Nonetheless, resistance to these semi-synthetic antimicrobials has been rapidly increasing, which is thought to be related to their high rate of prescription and highlights the importance of continuously developing novel semi-synthetic derivatives [60].

Semi-synthetic penicillins are obtained by producing penicillin G, which is hydrolyzed into 6-APA, purified, and later chemically altered, e.g., at the acyl side chain, to achieve various semi-synthetic penicillins [61]. Another beta-lactam example parallel to penicillins is the semi-synthesis of cephalosporins, which have reduced incidence of both side effects and resistance, alongside an additional site for chemical modification [62]. Cephalosporin C was firstly identified as a metabolite of *Cephalosporium acremonium* in 1948. By 1959, its hydrolysis

Table 1. Beta-lactam subclasses highlighting their diversity with examples of marketed antibiotics.

Subclasses	Examples of Marketed Antibiotics
Penicillins	Penicillin G, Penicillin V, Ampicillin, Amoxicillin, Bacampicillin, Cloxacillinm, Floxacillin, Mezlocillin, Nafcillin, Oxacillin, Methicillin[a], Dicloxacillin[a], Carbenicillin[b], Idanyl[b], Piperacillin[b], Ticarcillin[b]
Cephalosporins	Cefalothin[c], Cephradinea[c], Cefadroxyl[c], Cefazolin[c], Cephalexin[c], Cefuroxine[d], Cefaclor[d], Cefotetam[d], Cefmetazole[d], Cefonicid[d], Cefixime[e], Ceftibuten[e], Cefizoxime[e], Ceftriaxone[e], Cefamandol[e], Cefoperazone[e], Cefotaxime[e], Proxetil[e], Cefprozil[e], Ceftazidime[e], Cefuroxime Axetil[e], Cefpodexime[e], Cefepime[f], Ceftobiprole[g]
Other Minor Subclasses	Flomoxef[h], Latamoxef[h], Cefoxitin[i], Loracarbef[i], Imipenem[j], Meropenem[j], Panipenem[j], Aztreonam[k], Carumonam[k]

[a] Penicillinase-resistant and [b] Anti-pseudomonal penicillins; [c] First, [d] Second, [e] Third, [f] Fourth, and [g] Fifth generation cephalosporins; [h] Oxycepham; [I] Cefam; [j] Carbapenem; [k] Monobactam.

under acidic conditions yielded 7-aminocephalosporanic acid (7-ACA), which was the precursor to a multitude of semi-synthetic cephalosporins [63]. Figure 2 exemplifies the evolution of semi-synthetic cephalosporins, their timeline of introduction and the pros and cons of the succeeding generations marketed so far.

Another key illustration of semi-synthesis comes from the catalytic hydrogenolysis of chlorotetracycline (discovered in 1948), which resulted in the semi-synthesis of tetracycline by 1953, although it was later also found to be a natural product [64]. While semi-synthetic cephalosporins are mostly derivatives of 7-ACA, obtained via the addition of different molecular groups at the pair of modifiable sites, i.e., C7 and C3', semi-synthetic tetracyclines and macrolides result from serial structural modifications. Each iteration requires the chemical manipulation of the previous semi-synthetic derivative, which may preserve its advantages, but proportionally increases the number of chemical modifications and their complexity across a series of semi-synthetic generations. Therefore, less than 10 semi-synthetic tetracyclines were marketed in the last 60 years, in contrast with over 50 commercialized beta-lactams. However, recent advances in fully synthetic routes reignited tetracycline derivative synthesis [49,65], which is crucial given semi-synthesis is one of the major strategies for antibiotic discovery and particularly important in outpacing the evolution of resistance mechanisms.

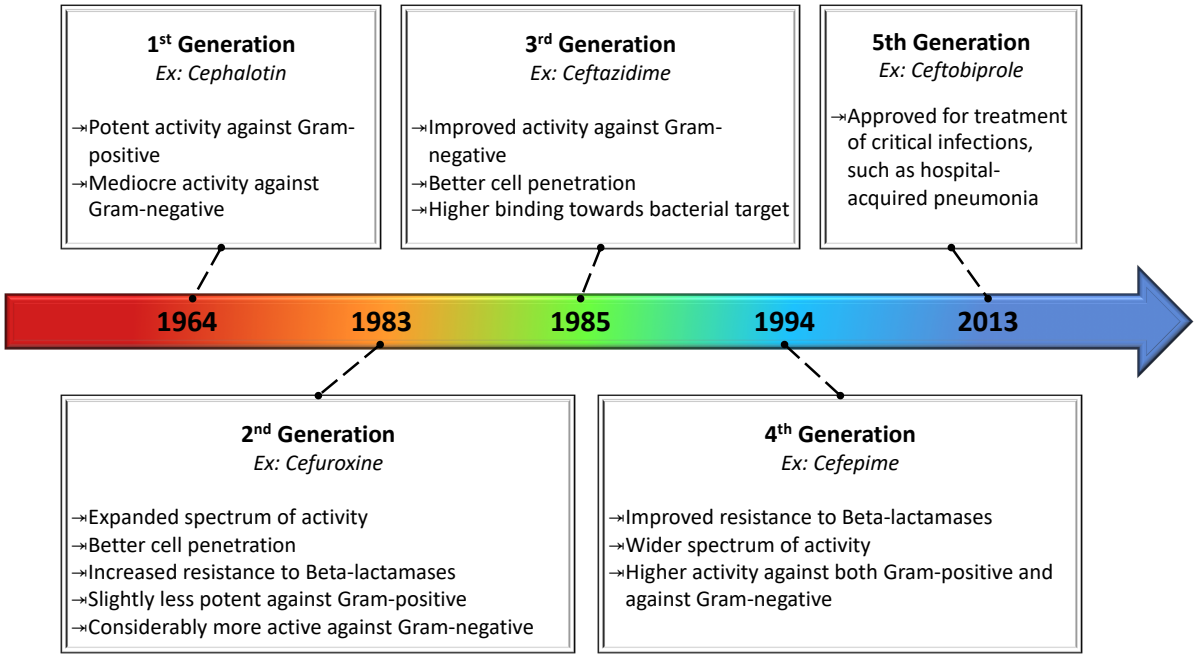


Figure 2. Evolution of cephalosporin characteristics over semi-synthetic generations. Because each generation is the result of substituting different molecular groups to 7-ACA, characteristics are not necessarily inherited by succeeding generations. For instance, second-generation cephalosporins had reduced potency against Gram positive pathogens, despite their otherwise improved properties.

II. 5 FROM THE GROUND UP: FULLY SYNTHETIC ANTIBIOTICS

Fully synthetic antibiotics, beyond introducing novel molecules, enable production at a scale suitable for clinical application. For instance, chloramphenicol became the first fully synthetic antibiotic whose scaffold originated from a natural product to reach the clinic in 1949. Unsurprisingly, the rational of semi-synthesis, that of chemically manipulating a scaffold, applies to a fully synthetic antibiotic like chloramphenicol. In fact, replacing the nitro group with methanesulfonyl resulted in thiamphenicol in 1952, which overcame the most concerning toxicity issues and had greater antimicrobial effect, thereby improving its clinical application [66]. The discovery in 1953 of the natural product azomycin found little clinical application but introduced the nitroimidazole class. In 1962, the search for optimized derivatives revealed metronidazole, currently produced with a fully synthetic protocol, which is active against the trichomoniasis parasite. Curiously, its activity against anaerobic bacteria was a fortuitous discovery, for which it is still in use [67]. Analogously to metronidazole, the natural product fosfomycin only had reasonable clinical application once a racemic synthesis protocol was

developed by Merck, and is still prescribed today [54]. While the advantages of chemically synthetizing natural products are straightforward, fully synthetic antibiotics also resulted in novel scaffolds. Given synthetic analogs of pyrimidine and purine bases inhibit bacterial growth, a diaminopyrimidine derivative named trimethoprim was introduced in 1962 [68], but only commercialized in 1969 together with sulfamethoxazole due to *in vitro* synergies, which are being questioned in light of recent *in vivo* observations [69].

Most of the fully synthetic antibiotics discussed have limited application to uncomplicated infections or as an economic alternative in developing countries. The quinolone class, which was unexpectedly discovered as a by-product of the synthesis of the antimalarial compound chloroquine, despite limited activity, was an important scaffold in the synthesis of nalidixic acid in 1962 [70]. Three more generations, the fluoroquinolones, were later obtained via chemical modification. Quinolones are currently the third most prescribed antibiotic to outpatients, behind macrolides and beta-lactams [71], and their antimicrobial effect is traced to the formation of a DNA gyrase-quinolone-DNA complex, which hampers replication and induces cellular death in both Gram positive and Gram negative pathogens. Another major antibiotic class, macrolides, are produced by semi-synthesis from erythromycin, which may involve simpler routes (e.g., four steps to derive azithromycin) up to more intricate modifications (e.g., 16 steps for the drug candidate solithromycin). The recent report of a fully synthetic protocol that yielded over 300 macrolides [72] brings new hope to this class of antibiotics and portrays the importance of the fully synthetic platform up to this day, not only in facilitating the synthesis, but also increasing the diversity, of the antibiotics available.

The case of fully synthetic beta-lactams is paramount since more intricate antibiotics were synthetized, eventually leading to a panoply of subclasses. Two important examples are the subclasses of carbapenems and monobactams. Carbapenems have a similar core structure to penicillins, differing at the C2–C3 double bond and the replacement of C1 sulfur for carbon, yielding improved potency, spectrum of activity, and better resistance to the action of beta-lactamases. Currently, 10 carbapenems have been marketed, or are under clinical development, since their discovery in 1985. Given that carbapenems have the widest activity spectrum among beta-lactams, including resistant pathogens, they are currently a first-in-line option for treating multidrug-resistant infections [73]. Likewise, monobactams have higher stability regarding beta-lactamases and are a promising way forward. These monocyclic beta-lactams were introduced to the clinic in 1984 with aztreonam and are currently being developed

towards siderophore moiety, a Trojan horse strategy that uses the bacterial iron uptake machinery to facilitate entry into Gram negative bacteria [74].

The class of oxazolidinones is divided into two groups differing in their MOA. The first acts on cell wall biosynthesis and was introduced with the natural product cycloserine in 1952, which is currently produced by synthetic means. Cycloserine is still used as a second-line therapeutic option for tuberculosis, especially in its multidrug-resistant form. The other group of oxazolidinones was found in 1984 to target protein synthesis and, despite reasonable antimicrobial activity, presented limiting toxicity issues [75]. From these, the DuPont group synthetized various derivatives from which resulted the discovery of linezolid, approved in 2000 as the first novel antibiotic class since the discovery of nalidixic acid, with almost half a century discrepancy [76]. Although no major resistance to linezolid has been reported, its limited effectiveness against Gram positive bacteria and toxicity in prolonged treatments limits its therapeutic use as a last-resort alternative against complicated cases of multidrug-resistant pathogens. Over the last decade, there has been substantial interest in developing novel oxazolidinones, given its low resistance profile, thus a handful of companies have been developing novel analogues [77]. Semi-synthesis, along with complete chemical routes, have catalyzed the dawn of the medicinal chemistry era, and together with the Waksman platform, yielded the vast majority of clinically relevant antibiotics. These were characterized by increasing potency and diminishing side effects with succeeding iterations, which gave mankind the upper hand on infectious diseases.

II. 6 Advent of genomics: Target-based screening

After the successes of the antibiotic golden age, the discovery rate of the underlying ADPs gradually decreased while in-class and multidrug-resistance mechanisms flourished. This weakened the therapeutic efficacy of the antibiotic arsenal and revived the issue of infectious diseases. The need for a new strategy coincided with the genomics era, which redefined the scientific paradigm governing antibiotic discovery and shaped new high-tech platforms. During the genomics era (1995–2004), the total number of sequenced microbial genomes increased from 3 to over 200 [78], and in the post-genomics era (2004–2014) reached a staggering 30,000 [79]. In this context, the first platform to arise was based on comparative genomics, where novel targets essential for pathogen survival were identified from repositories of sequenced and annotated genomes. These targets can encode pathogenicity mechanisms, highlighted by comparing genome sequences of pathogenic and non-pathogenic strains.

Furthermore, comparing these genomes to those of the host dismisses targets that are not exclusive to the pathogen, thus minimizing drug–host interactions and therapeutic side effects. Figure 3 resumes the target-based ADP: after target discovery, follows target validation by evaluating if they are essential for bacterial survival, e.g., with knockout analysis and/or mutational studies. After, the target is cloned, overexpressed and incorporated in a high-throughput screening (HTS) assay to search chemical libraries for binding agents.

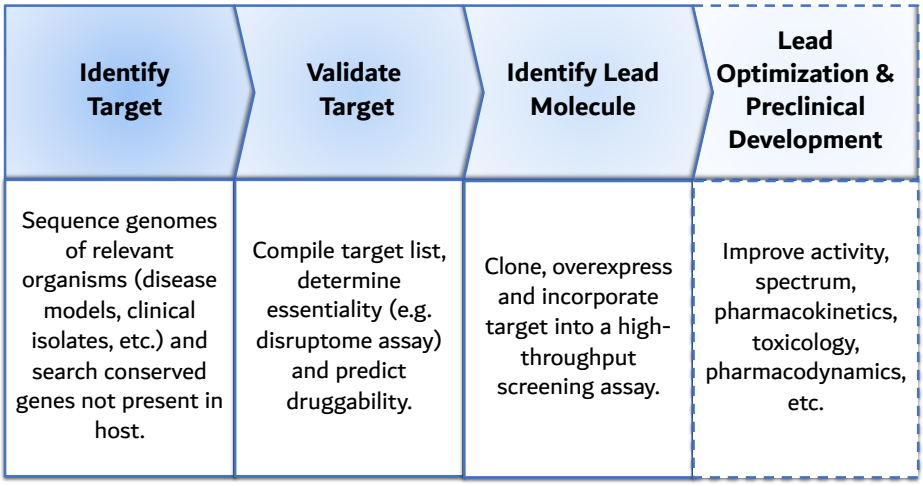


Figure 3. Schematic representation of the target-based antibiotic discovery platform: potential targets are identified from the genome sequence of pathogens and the host, the products of genes exclusive and essential for bacteria are incorporated into high-throughput screening assays, which identify candidates lead optimization and preclinical development. The latter falls outside the scope of this chapter and was not discussed.

Given that a manageable number of proteins are exclusive and conserved in bacteria, new MOAs were expected to surface, so some companies launched pioneering target-based screening programs. GlaxoSmithKline developed a target list of over 300 bacterial genes from 1995 to 2001, of which approximately 160 were considered essential for survival, and deemed 'druggable' in the search for broad-spectrum antibiotics [80]. Elitra pharmaceuticals, one of the top 10 start-up companies of 2001, submitted patents on over 4,000 targets after developing a proprietary strategy that identified essential genes in several pathogens [81]. Although target-based screening is suitable for finding potent inhibitors of said targets, their inability to reach their target, either due to the low permeability of bacterial membranes or the action of efflux pumps, hinders their *in vivo* activity. In a physiological context, the bacterial cell wall is a very efficient barrier against most small molecule drugs. Moreover, said targets may present functional redundancy. Alongside the aforementioned difficulties, the target-based screening

approach also failed because not all targets could be readily cloned, purified and incorporated into *in vitro* screening assays; and in some cases, the oversimplified environment of the assay excludes cofactors and lacks sensitivity for off-target effects. For instance, researchers at Merck found that low guanine–cytosine Gram positive pathogens have increased resistance to fatty acid biosynthesis targets when grown on media mimicking the human host [82], which a target-based assay cannot consider. Also, single gene targets are prone to single point mutations conferring resistance, thus are more likely to select resistant mutants, a major disadvantage of this approach.

Despite the massive bacterial genome sequencing, coupled with the development of bioinformatics tools to analyze said sequences, there are still many genes whose biological function has not been experimentally characterized. Moreover, genetic diversity further complicates target-based screening at the level of model organism selection, e.g., GlaxoSmithKline researchers reported an unrelated copy of genes conferring resistance in 20% of clinical isolates [83]. Ultimately, antibiotic discovery remains a challenging affair unattainable with an exclusively target-based genomics approach, and many consider the comparative genomics platform as rather unsuccessful, since not a single new drug was discovered [84]. Nonetheless, it reignited the quest towards understanding bacterial physiology, which had unquestionable positive implications in the development of antimicrobial chemotherapy. The reductionist approach of target-based screening, e.g., analyzing a single gene/protein (target) outside its biological context, evolved towards a more holistic phenotypic and pathway-based analysis. Subsequent platforms stemmed from taking a step back and reviving whole-cell screening, which was the basis of the Waksman platform, and bears the intrinsic advantage that lead compounds can interact anywhere on the pathway, on multiple constituents of the network or even on different metabolisms, and most importantly, replicating *in vivo* conditions.

II. 7 REVERSE GENOMICS: REVIVAL OF CELL-BASED SCREENING

The case of anti-tuberculosis drug discovery is a good example of this change in strategy: researchers moved away from target-based ADPs and returned to cell-based screening [85], with greater success in the discovery of novel, more diverse, lead molecules for subsequent optimization [86]. In general, cell-based screening results in higher variability and more complex data than the binary hit/no-hit of target-based screening, which is more difficult to relate with biologic phenomena. In cell-based assays, after a positive hit, e.g., an interaction of a drug with a microorganism such that its phenotype becomes altered, counter-screening with human cells allows for cytotoxic evaluation of drug candidates with antimicrobial activity. Cell-based ADPs first identify antimicrobial activity and only later endeavor to characterize MOA, and thus are also named reversed genomics, as represented in Figure 4. This is not necessarily a limitation as the FDA does not require the identification of the molecular target to initiate clinical trials, or to obtain marketing approval [87].

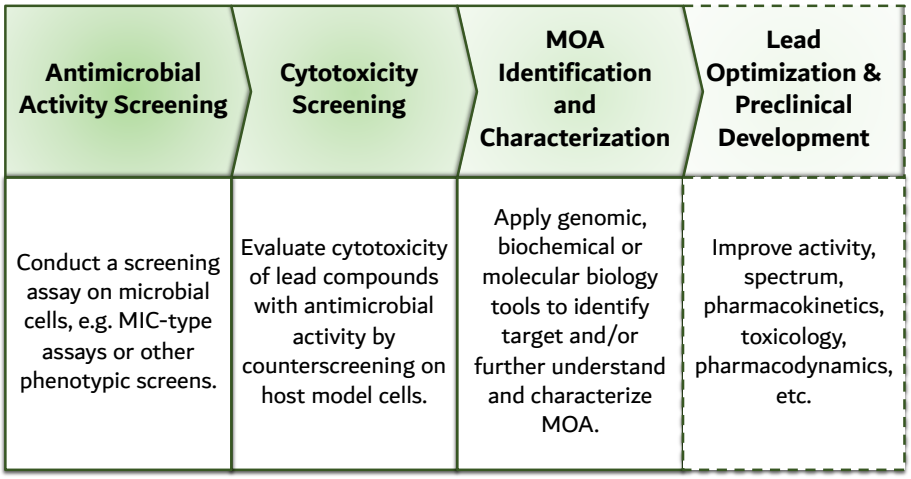


Figure 4. Schematic representation of the cell-based antibiotic discovery platform: drug candidates are identified from cell-based screening assays, a counter-screen excludes cytotoxic compounds, and subsequently other tools are applied to identify MOA. Although MOA is not a requisite, it may facilitate lead optimization and preclinical development, for instance, structural information on the target can enable a rational modification of the drug candidate.

In a broad sense, cell-based assays include screening large libraries in a systems-based mentality in order to evaluate the complex network of responses that antibiotics elicit [88], and are often termed phenotypic screening. Typically, if said screening probes phenotypic changes

free of target hypothesis, the term target-agnostic may be applied. Moreover, cell-based screening may follow a chemocentric approach, e.g., on compounds and its derivatives presenting a known biological effect. The development of cell-based screening methods has been of paramount importance and in its simpler form, these are centered on determining the minimum inhibitory concentration (MIC) to quantify antimicrobial activity. MIC assays are still relevant given that they complement other ADPs, for instance, the initial steps of the fully synthetic protocol to produce macrolide derivatives developed by Seiple et al. [72], included a MIC assay to evaluate antimicrobial activity and prioritize further efforts. Attempts to extend cell-based screens beyond MIC assays, which do not provide insight on the MOA of candidate molecules, have been directed at developing assays that quantify either: mitochondrial activity, by measuring a fluorescent product of a mitochondrial reaction; cellular integrity, evaluating the release of intracellular enzymes or the uptake of dyes that are impermeable when the cell is healthy; or measuring ATP content, etc. The reporter gene technology is still prevalent, where the activation and expression of a gene, which yields a quantifiable signal, e.g., luminescence or fluorescence, 'reports' biomolecular interactions. For instance, Hutter et al. [89,90] developed a HTS assay with a panel of twelve *Bacillus subtilis* strains, modified with luciferase reporter genes, to indicate the MOA of various antibiotics with sensitivity ranging from generic pathways, antibiotic class and the specific MOA of some drugs.

Alternatives to MIC-type assays require genetic manipulation and/or the use of a label, either in the form of a fluorescent or radioactive molecule, or a reporter gene. This is a limiting factor since, on the one hand, genetic manipulation implies *a priori* knowledge and on the other hand, the indication of gene transcription using a reporter gene may not always be coherent with alterations of enzymatic activity, thereby crippling the inherent sensitivity of these assays. Moreover, these signal transduction events can take considerable time to become detectable, thus limiting assay capacity and throughput [91]. In addition to the impact of a reporter gene, some of these labeled assays are limited on miniaturization. Despite said issues, cell-based screening still contributes greatly towards advancing antibiotic discovery, for instance, in neglected diseases such as malaria and human African trypanosomiasis [92]. For the latter, phenotypic screening lead to the discovery of fexinidazole (a nitroimidazole) which has been recently approved as the first oral therapy for human African trypanosomiasis and Chagas disease [93].

Eder et al. [94] reviewed the discovery platforms of first-in-class small molecule drugs, in particular the role of target- versus cell-based screening. First-in-class drugs act on a new

target or biological pathway. Between 1999 to 2008, phenotypic screening was more productive. However, in the period up to 2013, target-based approaches delivered most first-in-class drugs. Given the period ranging from the burdensome process of drug discovery until commercialization, there is a latency between the timeline of said review and the timeline presented in this manuscript. Concerning antibiotics, from 2000 to 2015, only five first-in-class new drugs were marketed: linezolid, daptomycin, retapamulin, fidaxomicin and bedaquiline [95]. Retapamulin binds to the 50S ribosomal subunit, fidaxomicin acts at the "switch" region of the bacteria RNA polymerase and bedaquiline specifically inhibits the ATPase of *Mycobacterium tuberculosis*. From these five new drugs, three are derived from natural products (daptomycin, retapamulin and fidaxomicin), and two were chemically synthesized (linezolid and bedaquiline) [90].

Historically, the success of antibiotic therapy relied on the discovery of natural scaffolds that were chemically optimized or produced. As such, it remains a rational decision to continuously develop screening strategies that probe natures repositories [96], especially using cell-based assays [97]. In fact, the major antibiotic scaffolds currently in use are derived from natural products, except for fluoroquinolones, sulfonamides and trimethoprim [98]. New cell-based ADPs should identify molecules with antimicrobial activity without the limitations of a label. Moreover, these assays would ideally provide insight on MOA whilst being capable of screening very large libraries, as isolated projects have very low rates of success [99].

II. 8 POST-GENOMICS

Biological research tends towards specialization, through increasingly focused and localized research; however, system-wide understanding of the biological constituents and their interactions is gaining importance. It is now possible to extract, handle and interpret information from much higher dimension and diverse origins, such as transcripts (i.e., transcriptomics), proteins (proteomics), and other molecules such as lipids (lipidomics), etc. The influence of these omics' technologies on the field of antibiotic discovery is undisputable, especially in understanding antibiotics MOA, identifying novel targets, and supplying insights to bacterial metabolism and physiology. Given the importance of screening in the ADPs discussed thus far, this mindset should be the backbone of future platforms. However, neither transcriptomics, proteomics nor lipidomics have matured to the throughput capacity of cell-based screening assays, and therefore are not the core technology of any ADP. These technologies convey insight on the biomolecules they probe, and not the holistic dynamics of

bacterial metabolism, thereby serving as complementary, albeit crucial, tools for the antibiotic discovery process.

Initial transcriptomics technologies were hybridization-based, e.g., Northern Blotting and microarrays [100]. Microarrays became the reference by mid-1990s [101] until next-generation sequencing extended to transcriptomics, which copes better with high genetic variation and non-specific hybridizations, as well as being label-free, e.g., unbiased and with a greater upper limit of detection. RNA-seq outperforms microarrays in predicting differentially expressed genes (90% versus 76%), but both technologies had similar performances when estimating the MOA of anti-cancer drugs [102]. However, RNA-seq enables studying non-coding RNA, which has a regulatory role in microbial responses to antibiotics, and therefore can be an alternative for new antimicrobial targets and/or novel combinatorial therapies [103]. Although next-generation sequencing [104], unbiased transcriptomics [105] and non-coding RNA [106] technologies have been applied to drug discovery in general, few studies discuss their application to antibiotic discovery. Whole-genome expression profiling elucidates the molecular and cellular responses to antibiotic stresses, which is particularly helpful for MOA determination, still a major gap in the field of antibiotic discovery. For instance, Salvarsan's MOA remained unclear over a century since its discovery, and so had its chemical structure [107]. In general, antimicrobials of the same class, thus with similar MOA, give rise to analogous transcriptional responses, which provides insight on the MOA of uncharacterized antibiotics [108]. For instance, the cell-based HTS assay developed by Hutter et al. [89,90] used transcriptomics to characterize the effect of various antibiotics, which then guided the genetic manipulation of a bacterial panel that 'reports' lead molecules MOA. Additionally, these signature responses are also being used to elucidate resistance mechanisms [109].

Genome expression technologies expand beyond the transcript level to biological events at the level of proteins. Not only do these occur without transcriptome alterations, but the instability of bacterial RNA raises both conceptual and technological limitations, which stress the need to complement transcriptomics with proteomics. Early proteomics studies relied on 2D gel-based assays or on difference gel electrophoresis [110], which require high-purity protein samples given their little sensitivity for low-abundance proteins, co-migration of proteins, and different modifications on the same protein [111]. Moreover, gel-based techniques are laborious, poorly automatable, and therefore difficult to apply in large-scale studies, so the evolution of mass spectrometry (MS) coupled with chromatographic separation rapidly became an alternative [112]. Proteomics has contributed towards identifying novel antimicrobial targets

[113], understanding resistance mechanisms to therapeutic antibiotics, and to some extent MOA elucidation [114], although unable to fully characterize MOA [115]. Importantly, the application of transcriptomics and proteomics technologies sheds new light on the function of various genes, leading to updates on existing annotations, and to improved understanding of bacterial metabolism and physiology. Although not at the core of any ADP, these technologies complement other ADPs, revealing information that is building the way forward. Since proteins interact with different biomolecules, including nucleic acids and lipids, specialized techniques have been developed to probe said interactions [116]. Moreover, within the realm of proteomics, the field of phosphoproteomics has 'spun-off'. Although this type of post-translational modifications was thought to be exclusive to eukaryotes, it affects bacterial homeostasis, virulence [117], and signal transduction [118]. Virulence mechanisms are interesting since the machinery used by bacteria to cause disease, for instance tyrosine kinases and phosphatases, are structurally different from the hosts and therefore can be exclusively targeted. Further descriptions on phosphoproteomics for drug discovery exist, albeit outside the scope of infectious diseases [119].

Understanding the physiological role of lipids, especially at the molecular level, has been considerably limited due to a technological gap that is being filled with very selective and sensitive lipidome characterization studies using MS, and combining various targeted and non-targeted approaches [120]. To achieve the required lipid separation, various chromatographic methods are routinely applied in combination with MS, for instance hydrophilic interaction liquid chromatography or gas chromatography [121]. Besides their structural function, lipids take part in a panoply of different biological events including signaling, trafficking and even metabolite functions. Regarding infectious diseases, an example of the application of lipidomics is the characterization of pathogenic microbe's cell wall, thereby unveiling its regulation and role in pathogenesis. This has revealed essential enzymes involved in fatty acid synthesis that are conserved across many of the most clinically relevant pathogens, e.g., FabI, FabH, FabF and acetyl-CoA carboxylase. As such, inhibitors of said enzymes are promising targets for future development [122], especially for persistent mycobacteria infections that use fatty acids as a carbon source [123].

II. 9 FROM METABOLOMICS TOWARDS META-OMICS

Gene expression data from transcriptomics and proteomics faces challenges, for instance, increases in RNA levels might not coherently result in changes at the protein level, added to

conceptual and technological limitations associated with the instability of bacterial RNA, and differences in protein levels are often poor estimators of metabolic activity. Consequently, interest in small-molecule metabolites has also emerged [124]. Metabolomics provides a more in-depth view of the biological reality governing microbial metabolism, using complex analytical methods like nuclear magnetic resonance (NMR) and chromatographic techniques associated with MS, alongside advanced data analysis algorithms [125,126]. Since bacterial responses to antibiotics begins rapidly and encompasses a variety of pathways, metabolomics is well suited to elucidate MOA. For instance, Hoerr et al. [127] explored NMR-based metabolomics to differentiate the MOA of nine antibiotics on *Escherichia coli*. Moreover, metabolomics complements other omics, for instance, rhodomyrtone's antimicrobial activity was identified via phenotypic screening, but its MOA was revealed with proteomics and metabolomics. Specifically, rhodomyrtone cripples capsule biosynthesis enzymes and metabolites of *Streptococcus pneumoniae* [128]. Additionally, it is possible to construct metabolic networks that aggregate catalytic activity (i.e., enzymes) alongside its coding and expression (i.e., genes, and their transcriptional and translational control). Over 50 networks of different organisms have been described, which sparked a new approach for antimicrobial target discovery [129].

Metagenomics, and meta-omics in general, reinforced natural product discovery, which has had a central role in antibiotic discovery, and chemotherapy in general, ranging from oncological to immunologic treatments, e.g., approximately 50% of all FDA-approved therapeutics are natural products or their derivatives [130]. Metagenomic studies estimated that only 10% of natural products have been identified, so the suggestion that only 1% of the complete natural products repository has been investigated comes as no surprise [131]. Therefore, the search for new drugs from natural sources is being pursued with renewed hopes [132]. In this regard, sampling new natural product sources, such as plants and marine organisms [133], and endophytes or epiphytes [134], is expected to reveal an even wider range of metabolic pathways with potential therapeutic applications. Moreover, exploring microorganisms unculturable in traditional laboratory conditions, or certain pathways not activated in typical laboratory conditions, requires efforts to develop adequate protocols. Given the meta-omics revelation of natures 'untapped' repositories, these could very well be the next 'gold mine' after actinomycetes, thereby justifying such efforts.

An interesting device, the iChip, allows for the high-throughput cultivation of microbial species in their natural habitat, with a growth recovery of 50% versus 1% of traditional methods, thereby giving access to otherwise 'uncultivable' microorganisms [135]. The iChip was used to

collect extracts from 10,000 isolates, from which a new species of Beta-proteobacteria thought to belong to a new genus related to Aquabacteria, was shown to produce an antibiotic named teixobactin, a peptidoglycan synthesis inhibitor. Teixobactin is mostly active against Gram positive pathogens, some of which are drug-resistant, and its bactericidal activity even surpasses that of vancomycin (a last-resort antibiotic), along with no indication of resistance mechanisms currently existing [136]. Metagenomics enables a different approach, instead of attempting to grow these 'uncultivable' microorganisms, sequences of interest can be identified from metagenomes, which can then be cloned and expressed in laboratory-friendly microbes. This avoids *in situ* cultivations, like with the iChip, or the burdensome tasks of deciphering the conditions required for growth or activation of unexplored pathways, and could provide novel molecules for antibiotic discovery [137].

The interaction of antibiotics with the human microbiome has also been enabled by meta-omics, which has created further opportunities for antibiotic discovery [138]. In fact, human-associated metagenomic studies revealed gene clusters with antibiotic potential. For instance, in the case of *Staphylococcus lugdunensis* nasal colonization, these commensal bacteria inhibit the presence of *Staphylococcus aureus* strains, thereby preventing opportunistic infections. This effect was traced to the production of lugdunin, a novel class antibiotic (macrocyclic thiazolidine peptides) produced by *S. lugdunensis*, which has bactericidal activity on key pathogens and importantly, presents a reduced risk of resistance development [139]. Likewise, lactocillin, a novel thiopeptide antibiotic, was identified from the vaginal microbiota and demonstrated considerable activity against typical pathogens [140]. On a different note, recurrent *Clostridium difficile* infections have been treated with complete microbiome transplantation [141], which is a 'brute-force' alternative in comparison to pinpointing the key molecular agent responsible for the regulation between commensal flora and pathogenic agents. These studies suggest that either introducing healthy microbiota, the targeted manipulation of commensal microbial populations, or even the purified molecular agents that regulate commensal bacteria, can be the source of novel therapeutics, all of which enabled by meta-omics technologies.

As seen, the technologies introduced in the post-genomics era have contributed towards new opportunities in antibiotic research, although these have not been at the core of any ADP per se. The case of teixobactin, for instance, heavily relied on the revelations brought by meta-genomics and the technologies required to build a device such as the iChip. However, identifying which of the molecules recovered with the iChip have antimicrobial activity, along

with insights into their MOA, were revealed with cell-based assays in a reverse genomics platform. Since phenotypic screening has had greater success in revealing first-in-class molecules, it is a great starting point for ADPs. However, the drawback is the reduced mechanistic information derived, for which the omics technologies supply accelerated insight on the MOA, including the molecular target and its regulation. Then, with the required mechanistic information, target-based screening can be applied in order to optimize lead molecules into best-in-class medicines [142]. Although most new antibiotics in late clinical development belong to existing classes [143], the paradigm of combining target- and cell-based screening brings renewed hope moving forward.

To handle the massive amount of information created in the post-genomic, The Pew Charitable Trusts launched SPARK [144], a database focusing on all aspects of Gram negative bacteria permeability, in hopes to facilitate information sharing and ease collaborations among the research community. Similarly, Farrell et al. [145] launched AntibioticDB, a database of antibiotics at all stages of development, including those that were discontinued. While some compounds were legitimately abandoned, e.g., in light of clinical results, due to toxicity issues or inferior effectiveness, the majority were discontinued for unknown reasons, and some were discarded for circumstantial reasons. If re-evaluated with novel chemical synthesis methods, or with post-genomics technologies, many abandoned compounds may prove to be effective therapeutics. If the requirements regarding toxicity profiles of antibiotics are made more in line with those of anticancer drugs, then many compounds may be worth investigating [146]. The case of daptomycin is a good example of how a compound can be revived, nearly 20 years after its abandonment, and still become the most financially successful intravenous antibiotic in the US [56].

II. 10 CONCLUSIONS

Once considered a resolved health issue, infectious diseases have resurfaced as a topic requiring urgent action. Antibiotic discovery has come a long way since the success of the Waksman platform, semi-synthesis and fully synthetic ADPs. As seen, the establishment of systematic procedures—platforms—was crucial for the discovery of the major antibiotic classes in use. Given the limitations of target-based screening, cell-based ADPs were revived during the genomics era. While some consider the genomics era platforms disappointing, the importance of the lessons learned should not be minimized. Considerable technological advances have given researchers unprecedented access to biological events and repositioned

the mind-set for antibiotic research in a systems biology context. Paradoxically, in the field of antibiotic discovery, the more we know the less we can discover. Although not at the core of any ADP, omics technologies have been proven of unquestionable value as auxiliary tools for antibiotic discovery. Importantly, cell-based screening requires MOA characterization, for which omics technologies are indispensable. Despite offering added-value information on biological events, their reduced throughput capacity alongside complementarity, in terms of resourcing to multiple omics simultaneously, implies a limited application in ADPs aiming to screen large libraries, for instance the reservoir of untapped natural products that is likely the next antibiotic 'gold mine'. There is a void between phenotypic screening (high-throughput) and omics-centered assays (high-information), where some mechanistic and molecular information complements antimicrobial activity, without the laborious and extensive application of various omics assays. Given the novelty of the various omics technologies, we are yet to extract their full potential and it seems feasible that these technologies will mature to fulfill this gap. Alternatively, innovative technologies favoring high throughput may be developed, even by sacrificing molecular sensitivity to some extent. In any case, the increasing need for antibiotics drives the relentless and continuous research on the foreground of antibiotic discovery. This is likely to expand our knowledge on the biological events underlying infectious diseases and, hopefully, result in better therapeutics that can swing the war on infectious diseases back in our favor.

II. 11 References

1. Smith, P.W.; Watkins, K.; Hewlett, A. Infection control through the ages. *Am. J. Infect. Control* **2012**, *40*, 35–42.
2. Clardy, J.; Fischbach, M.A.; Currie, C.R. The natural history of antibiotics. *Curr. Biol.* **2009**, *19*, 437–441.
3. Livermore, D.M. Discovery research: The scientific challenge of finding new antibiotics. *J. Antimicrob. Chemother.* **2011**, *66*, 1941–1944.
4. Saga, T.; Yamaguchi, K. History of Antimicrobial Agents and Resistant. *Japan Med. Assoc. J.* **2009**, *137*, 103–108.
5. WHO WHO | WHO Report on Global Surveillance of Epidemic-prone Infectious Diseases - Introduction Available online: http://www.who.int/csr/resources/publications/introduction/en/index1.html (accessed on Oct 12, 2018).
6. Center for Disease Control and Prevention Antibiotic Use in the United States, 2017: Progress and Opportunities. *US Dep. Heal. Hum. Serv.* **2017**, 1–40.
7. Cassini, A.; Högberg, L.D.; Plachouras, D.; Quattrocchi, A.; Hoxha, A.; Simonsen, G.S.; Colomb-Cotinat, M.; Kretzschmar, M.E.; Devleesschauwer, B.; Cecchini, M.; et al. Attributable deaths and disability-adjusted life-years caused by infections with antibiotic-resistant bacteria in the EU and the European Economic Area in 2015: a population-level modelling analysis. *Lancet Infect. Dis.* **2018**, *3099*, 1–11.

8. European Centre for Disease Prevention and Control; European Medicines Agency *The bacterial challenge : time to react*; 2009; Vol. 6 July 201; ISBN 9789291931934.
9. Laxminarayan, R.; Matsoso, P.; Pant, S.; Brower, C.; Røttingen, J.A.; Klugman, K.; Davies, S. Access to effective antimicrobials: A worldwide challenge. *Lancet* **2016**, *387*, 168–175.
10. Martens, E.; Demain, A.L. The antibiotic resistance crisis, with a focus on the United States. *J. Antibiot. (Tokyo).* **2017**, *70*, 520–526.
11. Akova, M. Epidemiology of antimicrobial resistance in bloodstream infections. *Virulence* **2016**, *7*, 252–266.
12. McFee, R.B. Nosocomial or Hospital-acquired Infections: An Overview. *Disease-a-Month* **2009**, *55*, 422–438.
13. Theuretzbacher, U. Accelerating resistance, inadequate antibacterial drug pipelines and international responses. *Int. J. Antimicrob. Agents* **2012**, *39*, 295–299.
14. Frieri, M.; Kumar, K.; Boutin, A. Antibiotic resistance. *J. Infect. Public Heal. « doi 10.1016/j.jiph.2016.08.007. [Epub ahead print]»* **2016**.
15. O'Neill, J. Antimicrobial Resistance: Tackling a crisis for the health and wealth of nations The Review on Antimicrobial Resistance Chaired b by J O'Neil, and supported by the Wellcome Trust and the UK Government. **2014**.
16. Rather, I.A.; Kim, B.C.; Bajpai, V.K.; Park, Y.H. Self-medication and antibiotic resistance: Crisis, current challenges, and prevention. *Saudi J. Biol. Sci.* **2017**, *24*, 808–812.
17. Dodds, D.R. Antibiotic resistance: A current epilogue. *Biochem. Pharmacol.* **2017**, *134*, 139–146.
18. Renwick, M.J.; Brogan, D.M.; Mossialos, E. A systematic review and critical assessment of incentive strategies for discovery and development of novel antibiotics. *J. Antibiot. (Tokyo).* **2016**, *69*, 73–88.
19. Smith, D. Antibiotics Are Money-Losers For Big Pharma. How Can We Incentivize The Development Of New Ones? Available online: https://www.forbes.com/sites/quora/2018/01/02/antibiotics-are-money-losers-for-big-pharma-how-can-we-incentivize-the-development-of-new-ones/#9318630487f3 (accessed on Mar 23, 2019).
20. Kostyanev, T.; Bonten, M.J.M.; O'Brien, S.; Steel, H.; Ross, S.; François, B.; Tacconelli, E.; Winterhalter, M.; Stavenger, R.A.; Karlén, A.; et al. The Innovative Medicines Initiative's New Drugs for Bad Bugs programme: European public-private partnerships for the development of new strategies to tackle antibiotic resistance. *J. Antimicrob. Chemother.* **2016**, *71*, 290–295.
21. Boucher, H.W.; Talbot, G.H.; Bradley, J.S.; Edwards, J.E.; Gilbert, D.; Rice, L.B.; Scheld, M.; Spellberg, B.; Bartlett, J. Bad Bugs, No Drugs: No ESKAPE! An Update from the Infectious Diseases Society of America. *Clin. Infect. Dis.* **2009**, *48*, 1–12.
22. Brown, D.G. New Drugs and Emerging Leads in Antibacterial Drug Discovery. *Ref. Modul. Chem. Mol. Sci. Chem. Eng.* 2016.
23. The Pew Charitable Trusts Antibiotics Currently in Global Clinical Development Available online: https://www.pewtrusts.org/en/research-and-analysis/data-visualizations/2014/antibiotics-currently-in-clinical-development (accessed on Mar 23, 2019).
24. Lewis, K. Platforms for antibiotic discovery. *Nat. Rev. Drug Discov.* **2013**, *12*, 371–387.
25. Tommasi, R.; Brown, D.G.; Walkup, G.K.; Manchester, J.I.; Miller, A.A. ESKAPEing the labyrinth of antibacterial discovery. *Nat. Rev. Drug Discov.* **2015**, *14*, 529–542.
26. Czaplewski, L.; Bax, R.; Clokie, M.; Dawson, M.; Fairhead, H.; Fischetti, V.A.; Foster, S.; Gilmore, B.F.; Hancock, R.E.W.; Harper, D.; et al. Alternatives to antibiotics-a pipeline portfolio review. *Lancet Infect. Dis.* **2016**, *16*, 239–251.
27. Herrmann, M.; Nkuiya, B.; Dussault, A.R. Innovation and antibiotic use within antibiotic classes: Market incentives and economic instruments. *Resour. Energy Econ.* **2013**, *35*, 582–598.

28. Aminov, R.I. A brief history of the antibiotic era: lessons learned and challenges for the future. *Front. Microbiol.* **2010**, *1*, 134.

29. Cui, L.; Su, X. Discovery, mechanisms of action and combination therapy of artemisinin. *Expert Rev. Anti. Infect. Ther.* **2009**, *7*, 999–1013.

30. Fetzner, S.; Drees, S.L. Old molecules, new biochemistry. *Chem. Biol.* **2013**, *20*, 1438–1440.

31. Nicholson, D.J. Biological atomism and cell theory. *Stud. Hist. Philos. Sci. Part C Stud. Hist. Philos. Biol. Biomed. Sci.* **2010**, *41*, 202–211.

32. Levine, R.; Evers, C. The Slow Death of Spontaneous Generation. *North Carolina State Univ.* **2000**, 7–8.

33. Brack, A. Introduction. *Mol. Orig. Life Assem. Pieces Puzzle* **1998**.

34. Newsom, S.W.B. Pioneers in infection control - Joseph Lister. *J. Hosp. Infect.* **2003**, *55*, 246–253.

35. Strebhardt, K.; Ullrich, A. Paul Ehrlich's magic bullet concept: 100 Years of progress. *Nat. Rev. Cancer* **2008**, *8*, 473–480.

36. Williams, K.J. The introduction of "chemotherapy" using arsphenamine - The first magic bullet. *J. R. Soc. Med.* **2009**, *102*, 343–348.

37. Ferrie, J.E. Arsenic, antibiotics and interventions. *Int. J. Epidemiol.* **2014**, *43*, 977–982.

38. Wainwright, M.; Kristiansen, J.E. On the 75th anniversary of Prontosil. *Dye. Pigment.* **2011**, *88*, 231–234.

39. Ligon, B.L. Penicillin: its discovery and early development. *Semin. Pediatr. Infect. Dis.* **2004**, *15*, 52–57.

40. Fleming, A. Sir Alexander Fleming - Nobel Lecture Available online: https://www.nobelprize.org/prizes/medicine/1945/fleming/lecture/ (accessed on Nov 12, 2018).

41. Bynum, B. Rediscovering penicillin. *Lancet* **2018**, *392*, 1108–1109.

42. Sales, K.C.; Rosa, F.; Sampaio, P.N.; Fonseca, L.P.; Lopes, M.B.; Calado, C.R.C. In Situ Near-Infrared (NIR) versus High-Throughput Mid-Infrared (MIR) Spectroscopy to Monitor Biopharmaceutical Production. *Appl. Spectrosc.* **2015**, *69*, 760–772.

43. Bentley, R. Different roads to discovery; Prontosil (hence sulfa drugs) and penicillin (hence β-lactams). *J. Ind. Microbiol. Biotechnol.* **2009**, *36*, 775–786.

44. Woodruff, H.B. Selman A. Waksman, winner of the 1952 nobel prize for physiology or medicine. *Appl. Environ. Microbiol.* **2014**, *80*, 2–8.

45. Lewis, K. Recover the lost art of drug discovery. *Nature* **2012**, *485*, 439–440.

46. Willey, J.; Sherwood, L.; Woolverton, C. Antimicrobial Chemotherapy. In *Prescott, Harley and Klein's Microbiology*; Willey, J., Sherwood, L., Woolverton, C., Eds.; Colin Wheatley/Janice Roerig-Blong: New Yotk, 2008; pp. 835–837 ISBN 978–0–07–299291–5.

47. Wainwright, M. Streptomycin: Discovery and Resultant Controversy. *Hist. Philos. Life Sci.* **1991**, *13*, 97–124.

48. Wiest, D.B.; Cochran, J.B.; Tecklenburg, F.W. Chloramphenicol Toxicity Revisited: A 12-Year-Old Patient With a Brain Abscess. *J Pediatr Pharmacol Ther* **2012**, *17*, 182–188.

49. Liu, F.; Myers, A.G. Development of a platform for the discovery and practical synthesis of new tetracycline antibiotics. *Curr. Opin. Chem. Biol.* **2016**, *32*, 48–57.

50. Cyphert, E.; Wallat, J.; Pokorski, J.; von Recum, H. Erythromycin Modification That Improves Its Acidic Stability while Optimizing It for Local Drug Delivery. *Antibiotics* **2017**, *6*, 11.

51. Mast, Y.; Wohlleben, W. Streptogramins - Two are better than one! *Int. J. Med. Microbiol.* **2014**, *304*, 44–50.

52. James, R.C.; Pierce, J.G.; Okano, A.; Xie, J.; Boger, D.L. Redesign of glycopeptide antibiotics: Back to the future. *ACS Chem. Biol.* **2012**, *7*, 797–804.

53. Floss, H.G.; Yu, T. Rifamycin s Mode of Action , Resistance , and Biosynthesis. *Chem. Rev.* **2005**, *105*, 621–632.

54. Michalopoulos, A.S.; Livaditis, I.G.; Gougoutas, V. The revival of fosfomycin. *Int. J. Infect. Dis.* **2011**, *15*, e732–e739.

55. Levin, A.S.; Barone, A.A.; Penço, J.; Santos, M.V.; Marinho, I.S.; Arruda, E.A.G.; Manrique, E.I.; Costa, S.F. Intravenous Colistin as Therapy for Nosocomial Infections Caused by Multidrug-Resistant Pseudomonas aeruginosa and Acinetobacter baumannii. *Clin. Infect. Dis.* **1999**, *28*, 1008–1011.

56. Eisenstein, B.I.; Oleson, Jr., F.B.; Baltz, R.H. Daptomycin: From the Mountain to the Clinic, with Essential Help from Francis Tally, MD. *Clin. Infect. Dis.* **2010**, *50*, S10–S15.

57. Taylor, S.D.; Palmer, M. The action mechanism of daptomycin. *Bioorganic Med. Chem.* **2016**, *24*, 6253–6268.

58. Adeyemo, A.A.; Oluwatosin, O.; Omotade, O.O. Study of streptomycin-induced ototoxicity: protocol for a longitudinal study. *Springerplus* **2016**, *5*.

59. Townsend, C.A. Convergent biosynthetic pathways to β-lactam antibiotics. *Curr. Opin. Chem. Biol.* **2016**, *35*, 97–108.

60. Güngör, Ö.Ö.N.; Gürkan, P.; Özçelik, B.; Oyardi, Ö. Synthesis and antimicrobial activities of new higher amino acid Schiff base derivatives of 6-aminopenicillanic acid and 7-aminocephalosporanic acid. *J. Mol. Struct.* **2016**, *1106*, 181–191.

61. Deng, S.; Ma, X.; Su, E.; Wei, D. Efficient cascade synthesis of ampicillin from penicillin G potassium salt using wild and mutant penicillin G acylase from Alcaligenes faecalis. *J. Biotechnol.* **2016**, *219*, 142–148.

62. Gröger, H.; Pieper, M.; König, B.; Bayer, T.; Schleich, H. Industrial landmarks in the development of sustainable production processes for the β-lactam antibiotic key intermediate 7-aminocephalosporanic acid (7-ACA). *Sustain. Chem. Pharm.* **2017**, *5*, 72–79.

63. Hu, Y.; Zhu, B. Study on genetic engineering of Acremonium chrysogenum, the cephalosporin C producer. *Synth. Syst. Biotechnol.* **2016**, *1*, 143–149.

64. Nelson, M.L.; Levy, S.B. The history of the tetracyclines. *Ann. N. Y. Acad. Sci.* **2011**, *1241*, 17–32.

65. Charest, M.G.; Lerner, C.D.; Brubaker, J.D.; Siegel, D.R.; Myers, A.G. A convergent enantioselective route to structurally diverse 6-deoxytetracycline antibiotics. *Science (80-.).* **2005**, *308*, 395–398.

66. Perez, M.; Echeverria, P.G.; Martinez-Arripe, E.; Ez Zoubir, M.; Touati, R.; Zhang, Z.; Genet, J.P.; Phansavath, P.; Ayad, T.; Ratovelomanana-Vidal, V. An Efficient Stereoselective Total Synthesis of All Stereoisomers of the Antibiotic Thiamphenicol through Ruthenium-Catalyzed Asymmetric Reduction by Dynamic Kinetic Resolution. *European J. Org. Chem.* **2015**, *2015*, 5949–5958.

67. Alauzet, C.; Lozniewski, A.; Marchandin, H. Metronidazole resistance and nim genes in anaerobes: A review. *Anaerobe* **2019**, *55*, 40–53.

68. Wright, P.M.; Seiple, I.B.; Myers, A.G. The evolving role of chemical synthesis in antibacterial drug discovery. *Angew. Chemie - Int. Ed.* **2014**, *53*, 8840–8869.

69. Caron, F.; Wehrle, V.; Etienne, M. Triméthoprime: un antibiotique en voie de réhabilitation en France. *Med. Mal. Infect.* **2017**, *47*, 253–260.

70. Bisacchi, G.S. Origins of the Quinolone Class of Antibacterials: An Expanded “Discovery Story.” *J. Med. Chem.* **2015**, *58*, 4874–4882.

71. Lee, G.C.; Reveles, K.R.; Attridge, R.T.; Lawson, K.A.; Mansi, I.A.; Lewis, J.S.; Frei, C.R. Outpatient antibiotic prescribing in the United States: 2000 to 2010. *BMC Med.* **2014**, *12*.

72. Seiple, I.B.; Zhang, Z.; Jakubec, P.; Langlois-Mercier, A.; Wright, P.M.; Hog, D.T.; Yabu, K.; Allu, S.R.; Fukuzaki, T.; Carlsen, P.N.; et al. A platform for the discovery of new macrolide antibiotics.

Nature **2016**, *533*, 338–345.

73. El-Gamal, M.I.; Brahim, I.; Hisham, N.; Aladdin, R.; Mohammed, H.; Bahaaeldin, A. Recent updates of carbapenem antibiotics. *Eur. J. Med. Chem.* **2017**, *131*, 185–195.
74. Fu, H.-G.; Hu, X.-X.; Li, C.-R.; Li, Y.-H.; Wang, Y.-X.; Jiang, J.-D.; Bi, C.-W.; Tang, S.; You, X.-F.; Song, D.-Q. Design, synthesis and biological evaluation of monobactams as antibacterial agents against gram-negative bacteria. *Eur. J. Med. Chem.* **2016**, *110*, 151–163.
75. Aminov, R. History of antimicrobial drug discovery: Major classes and health impact. *Biochem. Pharmacol.* **2017**, *133*, 4–19.
76. Leach, K.L.; Brickner, S.J.; Noe, M.C.; Miller, P.F. Linezolid, the first oxazolidinone antibacterial agent. *Ann. N. Y. Acad. Sci.* **2011**, *1222*, 49–54.
77. de Souza Mendes, C.; de Souza Antunes, A. *Pipeline of Known Chemical Classes of Antibiotics*; 2013; Vol. 2; ISBN 5521911243.
78. Mills, S.D. When will the genomics investment pay off for antibacterial discovery? *Biochem. Pharmacol.* **2006**, *71*, 1096–102.
79. Land, M.; Hauser, L.; Jun, S.R.; Nookaew, I.; Leuze, M.R.; Ahn, T.H.; Karpinets, T.; Lund, O.; Kora, G.; Wassenaar, T.; et al. Insights from 20 years of bacterial genome sequencing. *Funct. Integr. Genomics* **2015**, *15*, 141–161.
80. Payne, D.J.; Gwynn, M.N.; Holmes, D.J.; Pompliano, D.L. Drugs for bad bugs: Confronting the challenges of antibacterial discovery. *Nat. Rev. Drug Discov.* **2007**, *6*, 29–40.
81. Foulkes, J. Elitra pharmaceuticals: new paradigms for antimicrobial drug discovery. *Drug Discov. Today* **2002**, *7*, S12–S15.
82. Brinster, S.; Lamberet, G.; Staels, B.; Trieu-Cuot, P.; Gruss, A.; Poyart, C. Type II fatty acid synthesis is not a suitable antibiotic target for Gram-positive pathogens. *Nature* **2009**, *458*, 83–86.
83. Brötz-Oesterhelt, H.; Sass, P. Postgenomic strategies in antibacterial drug discovery. *Future Microbiol.* **2010**, *5*, 1553–1579.
84. Fields, F.R.; Lee, S.W.; McConnell, M.J. Using bacterial genomes and essential genes for the development of new antibiotics. *Biochem. Pharmacol.* **2017**, *134*, 74–86.
85. Yokokawa, F. Recent Progress on the Development of Novel Antitubercular Agents from Whole Cell Screening Hits. **2014**, 1239–1249.
86. Moloney, M.G. Natural Products as a Source for Novel Antibiotics. *Trends Pharmacol. Sci.* **2016**, *37*, 689–701.
87. Lee, J.A.; Berg, E.L. Neoclassic drug discovery: The case for lead generation using phenotypic and functional approaches. *J. Biomol. Screen.* **2013**, *18*, 1143–1155.
88. Brown, E.D.; Wright, G.D. Antibacterial drug discovery in the resistance era. *Nature* **2016**, *529*, 336.
89. Hutter, B.; Schaab, C.; Albrecht, S.; Borgmann, M.; Brunner, N.A.; Freiberg, C.; Ziegelbauer, K.; Rock, C.O.; Ivanov, I.; Loferer, H. Prediction of mechanisms of action of antibacterial compounds by gene expression profiling. *Antimicrob. Agents Chemother.* **2004**, *48*, 2838–2844.
90. Hutter, B.; Fischer, C.; Jacobi, A.; Schaab, C.; Loferer, H. Panel of Bacillus subtilis reporter strains indicative of various modes of action. *Antimicrob. Agents Chemother.* **2004**, *48*, 2588–2594.
91. Michelini, E.; Cevenini, L.; Mezzanotte, L.; Coppa, A.; Roda, A. Cell-based assays: Fuelling drug discovery. *Anal. Bioanal. Chem.* **2010**, *398*, 227–238.
92. Gilbert, I.H. Drug discovery for neglected diseases: Molecular target-based and phenotypic approaches. *J. Med. Chem.* **2013**, *56*, 7719–7726.
93. Deeks, E.D. Fexinidazole: First Global Approval. *Drugs* **2019**, *79*, 215–220.
94. Eder, J.; Sedrani, R.; Wiesmann, C. The discovery of first-in-class drugs: Origins and evolution.

Nat. Rev. Drug Discov. **2014**, *13*, 577–587.

95. Butler, M.S.; Blaskovich, M.A.T.; Cooper, M.A. Antibiotics in the clinical pipeline at the end of 2015. *J. Antibiot. (Tokyo).* **2016**, *70*, 3.

96. Wencewicz, T.A. New antibiotics from Nature's chemical inventory. *Bioorganic Med. Chem.* **2016**, *24*, 6227–6252.

97. Singh, S.B.; Young, K.; Silver, L.L. What is an "ideal" antibiotic? Discovery challenges and path forward. *Biochem. Pharmacol.* **2017**, *133*, 63–73.

98. Cox, G.; Sieron, A.; King, A.M.; De Pascale, G.; Pawlowski, A.C.; Koteva, K.; Wright, G.D. A Common Platform for Antibiotic Dereplication and Adjuvant Discovery. *Cell Chem. Biol.* **2017**, *24*, 98–109.

99. Lewis, K. New approaches to antimicrobial discovery. *Biochem. Pharmacol.* **2017**, *134*, 87–98.

100. Croucher, N.J.; Thomson, N.R. Studying bacterial transcriptomes using RNA-seq. *Curr. Opin. Microbiol.* **2010**, *13*, 619–624.

101. McGettigan, P.A. Transcriptomics in the RNA-seq era. *Curr. Opin. Chem. Biol.* **2013**, *17*, 4–11.

102. Wang, Z.; Deisboeck, T.S. Mathematical modeling in cancer drug discovery. *Drug Discov. Today* **2014**, *19*, 145–150.

103. Dersch, P.; Khan, M.A.; Mühlen, S.; Görke, B. Roles of regulatory RNAs for antibiotic resistance in bacteria and their potential value as novel drug targets. *Front. Microbiol.* **2017**, *8*, 1–12.

104. Woollard, P.M.; Mehta, N.A.L.; Vamathevan, J.J.; Van Horn, S.; Bonde, B.K.; Dow, D.J. The application of next-generation sequencing technologies to drug discovery and development. *Drug Discov. Today* **2011**, *16*, 512–519.

105. Mäder, U.; Nicolas, P.; Richard, H.; Bessières, P.; Aymerich, S. Comprehensive identification and quantification of microbial transcriptomes by genome-wide unbiased methods. *Curr. Opin. Biotechnol.* **2011**, *22*, 32–41.

106. Liu, D.; Mewalal, R.; Hu, R.; Tuskan, G.A.; Yang, X. New technologies accelerate the exploration of non-coding RNAs in horticultural plants. *Hortic. Res.* **2017**, *4*, 1–8.

107. Lloyd, N.C.; Morgan, H.W.; Nicholson, B.K.; Ronimus, R.S. The composition of Ehrlich's Salvarsan: Resolution of a century-old debate. *Angew. Chemie - Int. Ed.* **2005**, *44*, 941–944.

108. Brazas, M.D.; Hancock, R.E.W. Using microarray gene signatures to elucidate mechanisms of antibiotic action and resistance. *Drug Discov. Today* **2005**, *10*, 1245–1252.

109. Suzuki, S.; Horinouchi, T.; Furusawa, C. Prediction of antibiotic resistance by gene expression profiles. *Nat. Commun.* **2014**, *5*, 1–12.

110. Unlu, M.; Morgan, M.E.; Minden, J.S.D.G.E.; A Single Gel Method for. *Detect. Chang. Protein Extr. Electrophor.* **1997**, *18*, 2071–2077.

111. Hannigan, A.; Burchmore, R.; Wilson, J.B. The optimization of protocols for proteome difference gel electrophoresis (DiGE) analysis of preneoplastic skin. *J. Proteome Res.* **2007**, *6*, 3422–3432.

112. Roe, M.R.; Griffin, T.J. Gel-free mass spectrometry-based high throughput proteomics: Tools for studying biological response of proteins and proteomes. *Proteomics* **2006**, *6*, 4678–4687.

113. Pulido, M.R.; García-Quintanilla, M.; Gil-Marqués, M.L.; McConnell, M.J. Identifying targets for antibiotic development using omics technologies. *Drug Discov. Today* **2016**, *21*, 465–472.

114. Freiberg, C.; Brötz-Oesterhelt, H.; Labischinski, H. The impact of transcriptome and proteome analyses on antibiotic drug discovery. *Curr. Opin. Microbiol.* **2004**, *7*, 451–459.

115. Wenzel, M.; Bandow, J.E. Proteomic signatures in antibiotic research. *Proteomics* **2011**, *11*, 3256–3268.

116. Carneiro, D.G.; Clarke, T.; Davies, C.C.; Bailey, D. Identifying novel protein interactions: Proteomic methods, optimisation approaches and data analysis pipelines. *Methods* **2016**, *95*, 46–54.

117. Whitmore, S.E.; Lamont, R.J. Tyrosine phosphorylation and bacterial virulence. *Int. J. Oral Sci.* **2012**, *4*, 1–6.

118. Kobir, A.; Shi, L.; Boskovic, A.; Grangeasse, C.; Franjevic, D.; Mijakovic, I. Protein phosphorylation in bacterial signal transduction. *Biochim. Biophys. Acta - Gen. Subj.* **2011**, *1810*, 989–994.

119. Morris, M.K.; Chi, A.; Melas, I.N.; Alexopoulos, L.G. Phosphoproteomics in drug discovery. *Drug Discov. Today* **2014**, *19*, 425–432.

120. Vihervaara, T.; Suoniemi, M.; Laaksonen, R. Lipidomics in drug discovery. *Drug Discov. Today* **2014**, *19*, 164–170.

121. Sandra, K.; Sandra, P. Lipidomics from an analytical perspective. *Curr. Opin. Chem. Biol.* **2013**, *17*, 847–853.

122. Yao, J.; Rock, C.O. Bacterial fatty acid metabolism in modern antibiotic discovery. *Biochim. Biophys. Acta - Mol. Cell Biol. Lipids* **2017**, *1862*, 1300–1309.

123. Lam, S.M.; Shui, G. Lipidomics as a Principal Tool for Advancing Biomedical Research. *J. Genet. Genomics* **2013**, *40*, 375–390.

124. Lindon, J.C.; Nicholson, J.K. Analytical technologies for metabonomics and metabolomics, and multi-omic information recovery. *TrAC - Trends Anal. Chem.* **2008**, *27*, 194–204.

125. Gao, P.; Xu, G. Mass-spectrometry-based microbial metabolomics: Recent developments and applications. *Anal. Bioanal. Chem.* **2015**, *407*, 669–680.

126. Nagana Gowda, G.A.; Raftery, D. Can NMR solve some significant challenges in metabolomics? *J. Magn. Reson.* **2015**, *260*, 144–160.

127. Hoerr, V.; Duggan, G.E.; Zbytnuik, L.; Poon, K.K.H.; Große, C.; Neugebauer, U.; Methling, K.; Löffler, B.; Vogel, H.J. Characterization and prediction of the mechanism of action of antibiotics through NMR metabolomics. *BMC Microbiol.* **2016**, *16*, 1–14.

128. Mitsuwan, W.; Olaya-Abril, A.; Calderón-Santiago, M.; Jiménez-Munguía, I.; González-Reyes, J.A.; Priego-Capote, F.; Voravuthikunchai, S.P.; Rodríguez-Ortega, M.J. Integrated proteomic and metabolomic analysis reveals that rhodomyrtone reduces the capsule in Streptococcus pneumoniae. *Sci. Rep.* **2017**, *7*, 1–13.

129. Chavali, A.K.; D'Auria, K.M.; Hewlett, E.L.; Pearson, R.D.; Papin, J.A. A metabolic network approach for the identification and prioritization of antimicrobial drug targets. *Trends Microbiol.* **2012**, *20*, 113–123.

130. Milshteyn, A.; Schneider, J.S.; Brady, S.F. Mining the metabiome: identifying novel natural products from microbial communities. *Chem. Biol.* **2014**, *21*, 1211–1223.

131. Fischbach, M.A. Antibiotics from microbes: converging to kill. *Curr. Opin. Microbiol.* **2009**, *12*, 520–527.

132. Wright, G.D. Opportunities for natural products in 21stcentury antibiotic discovery. *Nat. Prod. Rep.* **2017**, *34*, 694–701.

133. Singh, S.B.; Barrett, J.F. Empirical antibacterial drug discovery - Foundation in natural products. *Biochem. Pharmacol.* **2006**, *71*, 1006–1015.

134. Kealey, C.; Creaven, C.A.; Murphy, C.D.; Brady, C.B. New approaches to antibiotic discovery. *Biotechnol. Lett.* **2017**, *39*, 805–817.

135. Nichols, D.; Cahoon, N.; Trakhtenberg, E.M.; Pham, L.; Mehta, A.; Belanger, A.; Kanigan, T.; Lewis, K.; Epstein, S.S. Use of Ichip for High-Throughput In Situ Cultivation of "Uncultivable" Microbial Species. *Appl. Environ. Microbiol.* **2010**, *76*, 2445–2450.

136. Ling, L.L.; Schneider, T.; Peoples, A.J.; Spoering, A.L.; Engels, I.; Conlon, B.P.; Mueller, A.; Schäberle, T.F.; Hughes, D.E.; Epstein, S.; et al. A new antibiotic kills pathogens without detectable resistance. *Nature* **2015**, *517*, 455–459.

137. Kolter, R.; van Wezel, G.P. Goodbye to brute force in antibiotic discovery? *Nat. Microbiol.* **2016**,

1, 15020.

138. Kolmeder, C.A.; de Vos, W.M. Metaproteomics of our microbiome - Developing insight in function and activity in man and model systems. *J. Proteomics* **2014**, *97*, 3–16.

139. Zipperer, A.; Konnerth, M.C.; Laux, C.; Berscheid, A.; Janek, D.; Weidenmaier, C.; Burian, M.; Schilling, N.A.; Slavetinsky, C.; Marschal, M.; et al. Human commensals producing a novel antibiotic impair pathogen colonization. *Nature* **2016**, *535*, 511–516.

140. Donia, M.S.; Cimermancic, P.; Schulze, C.J.; Wieland Brown, L.C.; Martin, J.; Mitreva, M.; Clardy, J.; Linington, R.G.; Fischbach, M.A. A systematic analysis of biosynthetic gene clusters in the human microbiome reveals a common family of antibiotics. *Cell* **2014**, *158*, 1402–1414.

141. Dopazo, J. Genomics and transcriptomics in drug discovery. *Drug Discov. Today* **2014**, *19*, 126–132.

142. Swinney, D.C. Phenotypic vs. Target-based drug discovery for first-in-class medicines. *Clin. Pharmacol. Ther.* **2013**, *93*, 299–301.

143. Fernandes, P.; Martens, E. Antibiotics in late clinical development. *Biochem. Pharmacol.* **2017**, *133*, 152–163.

144. Thomas, J.; Navre, M.; Rubio, A.; Coukell, A. Shared Platform for Antibiotic Research and Knowledge: A Collaborative Tool to SPARK Antibiotic Discovery. *ACS Infect. Dis.* **2018**, *4*, 1536–1539.

145. Farrell, L.J.; Lo, R.; Wanford, J.J.; Jenkins, A.; Maxwell, A.; Piddock, L.J.V. Revitalizing the drug pipeline: AntibioticDB, an open access database to aid antibacterial research and development. *J. Antimicrob. Chemother.* **2018**, *73*, 2284–2297.

146. Rolain, J.M.; Baquero, F. The refusal of the Society to accept antibiotic toxicity: Missing opportunities for therapy of severe infections. *Clin. Microbiol. Infect.* **2016**, *22*, 423–427.

Chapter III

High-throughput screening of antibiotic mechanism of action

This chapter is adapted from the manuscript:

Ribeiro da Cunha, B., Fonseca, L. P. and Calado, C. R. C. (2020) 'Technologies for high-throughput screening of antibiotic mechanism of action' *(In preparation).*

Author contribution

Bernardo Ribeiro da Cunha reviewed the literature, prepared the original draft, reviewed and edited its final version.

Abstract

There are two main strategies for antibiotic discovery: target-based and phenotypic screening. The latter has been much more successful in delivering first-in-class antibiotics, despite the major bottleneck of delayed mechanism of action (MOA) identification. Although finding new antimicrobial compounds is a very challenging task, identifying MOA has proven equally challenging. MOA identification is important because it is a great facilitator of lead optimization and improves the chances of commercialization. Moreover, the ability to rapidly detect MOA could enable a shift from an activity-based discovery paradigm towards a mechanism-based approach. This would allow probing the grey chemical matter, an underexplored source of structural novelty. In this study we review techniques with throughput suitable to screen large libraries and sufficient sensitivity to distinguish MOA beyond the biosynthetic pathway. In particular, the techniques used in chemical genetics (e.g., based on overexpression and knockout/knockdown collections), promoter-reporter libraries (e.g., using fluorescence or luminescence detection), transcriptomics (e.g., using microarrays and RNA sequencing), proteomics (e.g., either gel-based or gel-free techniques), metabolomics (e.g., resourcing to nuclear magnetic resonance or mass spectrometry), bacterial cytological profiling, and vibrational spectroscopy (e.g. Fourier-transform infrared or Raman scattering spectroscopy) were discussed. Ultimately, the new and reinvigorated collection of phenotypic assays brings renewed hope in the discovery of a new generation of antibiotics.

Contents

III. 1 INTRODUCTION

Antibiotics have significantly improved many aspects of society. From their application in medicine resulted an increase of life expectancy and well-being, without which even the simplest of medical interventions would pose life-threatening risks [1]. However, antibiotic discovery has stagnated at alarmingly low rates since its golden age, when most classes in use today were discovered. Infectious disease in general, and multidrug resistant pathogens in particular, are increasingly a worldwide concern, and many calls for action have been issued, especially to reiterate the desperate need for new drugs [2].

There are two main antibiotic discovery strategies, target-based and phenotypic screening. While the target-centric approach begins with a target whose inhibition should result in the desired therapeutic effect, phenotypic screening starts with a cell-based assay that monitors a phenotype, e.g. growth inhibition [3]. While phenotypic screening has a higher likelihood of identifying candidate drugs, along those that target poorly understood biological pathways, their molecular targets are not identified in the process and require subsequent efforts [4]. This results in higher rates of rediscovery, which is a key challenge in natural product antibiotic discovery [5], and an inability to detect low potency candidates, which can be later modified for enhanced therapeutic effect [6].

Identifying the mechanism of action (MOA) is still very challenging [7]. For instance, penicillins MOA is still subject to debate, with recent studies suggesting a more complex mechanism than inhibition of cell wall synthesis [8]. As such, MOA determination is a major bottleneck when screening hundreds of thousands of compounds is a reasonable throughput of a drug discovery program, in part due to the ease in synthetizing bioactive compounds, and in part given the increasing availability of natural product libraries [9]. This is particularly important because knowledge on the exact molecular target, and the pathways it is involved in, facilitates lead optimization by rapidly excluding derivatives with increased activity due to off-target effects [10], thereby guiding medicinal chemistry programs towards improved chances of commercialization [11].

Moreover, the ability to rapidly detect MOA could enable a shift from the activity-based discovery paradigm towards a mechanism-based approach. This would expand the chemical space towards the grey chemical matter, i.e., compounds that induce phenotypic modulation without sufficient potency to be detected in traditional screening assays. The grey chemical

space is an underexplored source of structural novelty that, after structural optimization, could yield much needed new antibiotics [12].

Conventional MOA studies are based on macromolecular synthesis assays, which measure radioactively labeled molecular precursors to ascertain the inhibition of DNA, RNA, protein, lipid or peptidoglycan synthesis. This implies that compounds which act on different steps of the same pathway cannot be distinguished, thereby missing out on potentially novel MOA. Additionally, all pathways are apparently inhibited by compounds which kill bacteria very rapidly, such as disinfectants, even though they affect specific pathways. To make matters worse, these assays are typically slow, laborious, low resolution, low accuracy and low throughput [13]. Alternatively, biochemical approaches, like affinity chromatography, can identify the exact biomolecule to which a compound binds [14,15], but only in the case of a high-affinity small molecule and a fairly abundant protein receptor [16]. Moreover, these require large quantities of test compound, which are not always attainable.

Novel methods capable of probing this complex phenomena are urgently needed to ease the process of antibiotic discovery [17]. In this study, several applications of system-wide profiling techniques for MOA identification were reviewed. Particular attention was given to techniques with sufficient throughput to be employed in screening campaigns of large libraries, whose advantages and limitations were described in light of several examples.

III. 2 CHEMICAL GENETICS

At its core, chemical genetics map the effect of exogenous ligands across genetic variance [18]. When the exogenous ligand is an antibiotic candidate, the effect across a mutant library enlightens its MOA. Interestingly, querying the MOA of antibiotics with chemical genetics contributed to our comprehension of many microbial processes, such as the synthesis of nucleic acids, proteins and the cell wall [19].

III.2.1 OVEREXPRESSION COLLECTIONS

Identifying MOA with overexpression collections involves screening mutants that, when exposed to a compound that targets the product of an overexpressed gene, display a resistant profile. In other words, if the target is overexpressed, a larger antibiotic dose is required compared with a wild-type strain. Additionally, because challenging these resistant collections

with antibiotics generates unique mechanistic fingerprints of the multilevel interactions induced, these assays reveal some, if not all, the participants in the network targeted by the antibiotic.

The first published application of resistant libraries to ascertain antibiotics MOA in a high-throughput phenotypic screening assay was attempted by Li et al. [20], who screened a 8,640 small molecule commercial collection for growth inhibitors of a 20,000 random mutant library of *Escherichia coli* MC1061. The plasmids in these clones were sequenced and two genes were identified, *folA* and *acrB*, which translate to dihydrofolate reductase and the (multidrug) acridine efflux pump. While the first was the target of two similar compounds, the latter surprisingly could efflux the remaining compounds. Given the nature of multicopy suppression assays, one hurdle is filtering out genes that code some drug-resistance mechanism from those that code the molecular target of compounds. Also, despite it being a high-throughput assay, the library size is a clear disadvantage for even higher throughput.

Building on the previous study, the ASKA collection was constructed. ASKA is an ordered high-expression *E. coli* library containing (nearly) all ORFs from the K12 W3110 strain in pCA24N high copy number plasmids [21]. As such, both essential and non-essential genes can be queried in regard to overexpression. Using ASKA, Pathania et al. [22] screened ~50,000 small molecules at a range of concentrations, which allowed a stringency-type analysis, whereby suppression of growth inhibition by a given mutation was evaluated along drug dose, therefore proving a more precise identification of the main molecular target at high drug doses, but also revealing other secondary targets at lower doses.

At high-stringency (16xMIC), the targets of 5 known antibiotics (fosfomycin, fosmidomycin, trimethoprim, sulfamethoxazole and D-cycloserine) were clearly identified, but not of spectinomycin, whose target only became apparent at lower stringencies (8xMIC). More importantly, a novel compound was identified from the high-stringency analysis, MAC13243, whose target is the periplasmatic protein LolA, responsible for lipoprotein transport across the periplasmic region. MAC13243 represented a novel promising antibacterial, whose target and MOA belonged to a (then) novel class and thus warranted further investigation.

Later studies into the degradation of MAC13243 revealed that the breakdown product S-(4-chlorobenzyl)isothiourea was responsible for its antibacterial activity, and this compound is in fact a structural analogue of S-(3,4-dichlorobezyl)isothiourea, whose ability to disrupt the actin-like cell shape-determining MreB protein had already been reported [23]. Because MAC13243 breaks down in aqueous medium, its use as a lead molecule was questioned.

However, its alternative use as a permeabilization agent, to potentiate large-scaffold antibiotics, has been suggested, although further structural optimization is likely required [24].

Despite the limitations of MAC13243 as a therapeutic agent, its target, LolA, is part of a five-protein system (LolABCDE) that is an attractive target of Gram negative pathogens. Since the outer membrane of Gram negative bacteria is a permeability barrier; confers greater structural integrity; and participates in a panoply of other roles, including the translocation of proteins and nutrients, adhesion and signal transduction; it is not only essential for survival, but also more easily accessible to drugs in comparison with cytoplasmatic targets [25].

III.2.2 KNOCKOUT AND KNOCKDOWN COLLECTIONS

In contrast with overexpression libraries, the Keio collection is single-gene knockout library of *E. coli* K12 BW25113, where the kanamycin resistance cassette takes the place of the deleted gene [26]. Being a knockout library, only non-essential genes can be probed. Although not as useful for MOA identification as overexpression libraries, the Keio collection highlighted the potential of combinatorial therapies [27]; contributed towards the characterization of gene essentiality and chromosomal organization [28]; revealed the complex interplay of metabolic pathways elicited during nutrient stress, which elucidated gene function and unwrapped new antibiotic targets [29]; and illuminated mechanisms of resistance, including determinants of drug permeability, efflux, degradation as well as stress responses [30].

Stokes et al. [31] used the Keio collection to comprehend how *E. coli* became susceptible to vancomycin, a narrow-spectrum antibiotic active against Gram positive bacteria, under cold stress. While this hydrophilic antibiotic is unable to pass the outer membrane of Gram negative bacteria, transient ‘cracks’ in the outer membrane caused by low temperatures allow its diffusion into the periplasm, allowing it to reach its target.

While controlling essential gene dosage by knockout mutations is technically accessible in diploid eukaryotic organisms, for instance via genome-wide haploinsufficiency profiling, this is more challenging in prokaryotes, and early studies were limited to a low number of genes [32]. Probing essential genes requires a conditional knockdown, e.g., the use of mutants that only display a mutant phenotype in a given restrictive condition [33].

The first hypersensitized microbial collection that allowed the modulation of essential genes used xylose-inducible antisense RNA expression in *Staphylococcus aureus*. Here, 245 target-depleted strains could be tuned to control essential gene expression to obtain moderate growth suppression (~20%) through to the knockout phenotype. As such, comparison of a hypersensitized phenotype with that obtained after exposure to a compound reveals its target [34]. Using this collection, Phillips et al. [5] conducted a natural product screening program that revealed kibdelomycin, a novel type II topoisomerase inhibitor. Given its broad spectrum of activity, especially against Gram positive bacteria, along lack of cross-resistance, there was great expectation for kibdelomycin, which so far has not materialized into new drugs reaching the market.

Gene downregulation with antisense RNA is not without limitations. In cases involving polycistronic mRNA, the entire strand may be degraded when the antisense RNA binds, resulting in the suppression of more genes than desired. In the cases where there are common motifs, undesirable gene suppression may also occur [35]. A more specific and efficient methodology for gene knockdown relies on Clustered Regularly Interspaced Short Palindromic Repeats/dCas9 (CRISPR) transcriptional regulation. Peters et al. [36] employed CRISPR to develop a knockdown library of *Bacillus subtilis* also modulated by xylose, but in this case the plasmids were integrated in the microbial chromosome. With this technique the MOA of MAC0170636 was identified, namely the inhibition of undecaprenyl pyrophosphate synthetase, which is essential for cell wall synthesis.

Interestingly, some gene annotations are being revised with CRISPR. Liu et al. [37] characterized the function of previously 'hypothetical' genes of *Streptococcus pneumoniae*. However, only 73% of genes deemed essential with transposon sequencing were considered essential with CRIPSR knockdown. While this could be due to inadequate annotation using transposon sequencing, it could also be due to different growth conditions, or to insufficient suppression with CRISPR. Regardless, given CRISPR systems are found across a range of bacteria and can be easily transferred to other microorganisms, constructing similar downregulated libraries could be a promising step forward towards novel antibiotics and better therapeutics.

III. 3 PROMOTER-REPORTER LIBRARIES

Promoter-reporter strains are used to build a map of how antibiotics affect the activation of a promoter, which controls the expression of one or more genes. Comparison with baseline transcriptional levels reveals the differential effects induced by antibiotics, and doing so across multiple strains in a library maps out the global response, which can be used as a MOA profile [9], as well as to query cellular pathways and unravel off-target effects [38]. Because a fluorescent or luminescent signal is produced with the transcription of the promotor-reporter, this technique offers adequate temporal resolution to study the effect of antibiotic exposure through a time course [19].

III.3.1 FLUORESCENCE-BASED REPORTERS

Zaslaver et al. [39] developed an *E. coli* K12 MG1655 library were the green fluorescent protein (GFP) gene was fused to ~1,820 different promoter regions, which is over 75% of known promoters, in low copy-number plasmids pUA66. This allowed highly accurate near-genome-wide measurements of promoter activity. Nonetheless, some transcriptional activity outside the individual promoter was occasionally detected, given other post-transcriptional regulation or intergenic regions with multiple promoters. This limitation of promoter-reporter assays is independent of the reporter chosen. Even so, this library revealed the mechanism ruling suppression antagonism of DNA and protein synthesis inhibitors, where DNA stress responses result in nonoptimal regulation of ribosomal genes, distorting the DNA-to-protein ratio, and suppressing protein synthesis inhibition [40].

More recently, promoter-reporter collections have been used to characterize stress responses elicited by bioactive compounds. To that end, Elad et al. [38] constructed a panel of 15 *E. coli* strains with the *lux*CDABE plasmid, to which promoters elicited during particular stresses were inserted. These stresses include DNA damage, protein misfolding, inhibition of fatty acid synthesis, increase of reactive oxygen species and the presence of metals. Of the 420 FDA-approved drugs tested, 89 elicited a response, even though some were not directed at prokaryotes. Interestingly, these responses were clustered in accordance with drug class, and often predicted their toxicity. In line with this, the collection produced by Zaslaver et al. was challenged with 9 antibiotics, which revealed the cellular pathway affected, off-target effects, and to some extent the MOA. However, comparing results obtained with luminescent

and fluorescent reporters, some discrepancies were found for one compound, which highlights the importance of validating these label-based assays with multiple approaches.

III.3.2 LUMINESCENCE-BASED REPORTERS

In comparison with fluorescence-based assays, luminescence is typically preferred for screening purposes given bacteria's considerable autofluorescence, which adds background noise. As such, luciferase reporters as *lux*CDABE are preferred given their low background noise, capacity to screen small colonies on solid media, with high reproducibility, without the need for substrate addition for continuous signal. Nonetheless, luciferase synthesis requires ATP and is affected by the redox potential, as such some false positives occasionally occur.

Baptist et al. [41] transformed the Keio collection with the promoter of the *acs* gene and of the *sdh* gene within *lux*CDABE plasmids, which carry the ampicillin resistance *bla* gene. Moreover, a second fluorescence-based reporter assay was used to exclude non-coherent mutants. A large number of mutants displayed altered transcription of *acs*, which is at the end of a well-described complex regulatory network, indicating that the genetic regulatory network of *E. coli* is densely connected, and that it is strongly linked with metabolism. Although not directly relevant to identifying the target of antibiotics, this study provides insight into the complexity of genetic interactions governing phenotype, which illustrates why the resolution of promoter-reporter assays often impairs their application for MOA identification.

Engelberg et al. [42] developed a *Pseudomonas aeruginosa* reporter strain deprived of 5 efflux pumps and an outer membrane protein, to which a mini-Tn7-based *lux* reporter was integrated in its genome. This strain was used to screen extracts of 529 fungal isolates, of which 12 inhibited bacterial growth by at least 80%, and were further investigated. Using an efflux-deficient strain implies that a higher compound quantity enters and remains within the bacterial cell, which unveils more compounds with inhibitory effect by postponing the efflux issue to a later stage of the drug discovery process. Second, the assay is considerably faster (~3h) than typical absorbance-based assays (~18-24h), which improves throughput.

III.3.3 ALTERNATIVE ASSAYS

Given that the β-lactamase gene of *Citrobacter freundii* is induced by cell wall inhibitors in general, and not exclusively β-lactams, Sun et al. [43] developed an indirect induction

system for pathway specific screening by cloning its *ampR-ampC* locus into *E. coli*. Since the β-lactamase concentration determines the rate of nitrocefin hydrolysis, the resulting product can be quantified by measuring the absorbance at 490nm, which reveals the extent of cell wall inhibition. Using said reporter, Nayar et al. [44] identified a sulfonyl piperazine and a pyrazole that target LpxH, a lipopolysaccharide anchor, as well as LolC and LolE, part of the LolCDE transporter.

Another drug that targets the LolCDE transporter, compound G0507, was recently identified by Nickerson et al. [25] using an *E. coli* strain transformed with a plasmid containing the σ^{E} stress response promoter rpoHP3 followed by the β-galactosidase gene, lacZ. This reporter is activated not only by peptidoglycan damage, but also inhibitors of the outer membrane, lipoprotein and lipopolysaccharide biosynthesis, along with inhibitors of transport mechanisms as AmpC, of cell wall biogenesis, polycationic compounds and aminoglycosides. Although G0507 has been used to probe lipoprotein trafficking mechanisms, it had reduced activity, for which structural modifications were applied to convert it into a promising anti-infective drug.

Lastly, bacteriophages have been used as reporters. While still in the proof-of-concept stage, Mido et al. [45] used the MS2 phage-MAGPIX reporter to detect the presence of live *E. coli* cells. With a burst time of 30 minutes and 5,000-10,000 phages released per infected cell, a sensitivity of approximately 100 cells/mL was achieved in as little as 3h. Because the fluorescent signal comes from the SAPE probe, which binds to the antibody that recognizes the phage, this sandwich type immunoassay detection is considered label-free and can be adapted to a variety of bacteria.

III. 4 TRANSCRIPTOMICS

Transcripts were initially characterized with northern blotting or polymerase chain reaction, but these assays were mostly limited to a few transcripts. The development of microarrays in the 1990s enabled the quantitative detection of most known transcripts of a strain in a single assay [46]. In practice, microarrays reveal the hybridization between a oligonucleotides of reference strain and its labeled complementary DNA obtained from an experimental strain, e.g., one that has been exposed to an antibiotic. As such, microarrays are limited to strains whose genome is known and available when developing the microarray.

Alternative hybridization-based techniques include high-density bead arrays, electronic microarrays or suspension bead arrays, which have been revised elsewhere [47].

III.4.1 HYBRIDIZATION ASSAYS

Boshoff et al. [48] used whole-genome microarrays to measure the effect of various drugs, along with growth-inhibitory conditions, on *Mycobacterium tuberculosis*. Transcripts with at least a 3-fold change were analyzed and clustered according to their known regulatory network. These were coherent with the known MOA of inhibitors of cell-wall synthesis, protein synthesis, transcription and DNA gyrase. Moreover, these clusters also suggested the MOA of uncharacterized compounds, including a natural product extract in either its crude or purified form, whose MOA was later identified by reporter strains and biochemical assays.

Similarly, Liang et al. [49] explored the transcriptional response of *M. tuberculosis* to linezolid, which binds to the 50S ribosomal subunit and inhibits protein synthesis. In total, 729 genes were differentially expressed, including genes involved in protein synthesis, sulfite metabolism, cell-wall synthesis, among others. Surprisingly, genes closely related to linezolids' target were down-regulated. This reveals the complexity of transcriptional responses to antibiotics, as well as the challenge of pinpointing the biomolecular target of the drug. For example, Bonn et al. [50] identified a similar number of genes with differential expression, however the genes closely related to linezolids' target were found to be up-regulated.

Although hybridization assays provide reasonable throughput at a relatively low cost, issues regarding reliability and reproducibility are well known. For instance, only high copy transcripts are easily detected, outputs often have high background noise due to cross-hybridization, signal saturation is common, and several issues arise given that probes are based on predicted open reading frames of sequenced genomes [51].

III.4.2 THE UPRISING OF NEXT-GENERATION SEQUENCING

Many of the disadvantages of microarrays can be overcome with sequencing strategies. Initially reliant on the Sanger method, and later variants thereof, sequencing techniques were limited by high cost, inability to map short reads, and incomplete transcript sequencing [52]. Next-generation sequencing allowed the analysis of larger sequence numbers, with higher reproducibility, from strains whose genome need not be sequenced. Moreover, RNA-seq has

little to no background noise and virtually no saturation [52]. Recently the cost of RNA-seq has been steadily decreasing into reasonable levels, which explains its increasingly widespread application [53].

Many transcriptomic studies in the RNA-seq era focus on the particular transcriptional response of a single compound [54]. By design, these studies were descriptive, rather than predictive, of MOA. Nevertheless, these have consolidated the correlation between MOA and transcriptional response, thereby established transcriptomics as an essential tool in antibiotic discovery. Because many of said studies have been revised elsewhere [55,56], the focus of this section shifts to RNA-seq studies that compare different MOA with a predictive approach.

One of the advantages of RNA-seq is the ability to probe non-coding RNA. For instance, Howden et al. [57] explored the role of both the mRNA and sRNA of a multidrug-resistant *S. aureus* after exposure to four last-resort antibiotics. Interestingly, mRNA profiles more closely reflected growth conditions, and the strain analyzed, than antibiotic exposure, except for the linezolid-induced transcriptional response, where a clear profile emerged. On the other hand, 39 differently expressed sRNAs confirmed with northern blotting provided better MOA profiles. Moreover, it is interesting to note that many antisense sRNAs associated with protein synthesis genes were down-regulated independently of the drug tested.

Similarly, Molina-Santiago et al. [58] explored the transcriptional response induced by 8 antibiotics on *Pseudomonas putida*, which was chosen as a model organism given its resistance to high concentrations of various antibiotics. In total, 5756 mRNAs, 58 tRNAs and 154 sRNAs were identified. Two-fold changes in mRNA levels resulted in two clusters, one with kanamycin, ampicillin and chloramphenicol, all of which had similar profiles to the control samples. The second cluster presented more distinct transcriptional responses regarding the control. Using sRNA profiles, only ampicillin and chloramphenicol clustered together with the controls, suggesting sRNA profiles are better suited for MOA classification.

Sequencing-based transcriptomics also allows to study the pathogen within the host, which reveals microbial gene expression during infection. For instance, the adaptation of *Mycobacterium* inside macrophages as disease progresses, but also the efficacy of treatment in both animal models and human patients, has been evaluated [56]. Pathogens *in vivo* and *in situ* present lower variation of gene expression, as well as up-regulated genes regarding SOS stress response, alginate biosynthesis and efflux pumps, among others [59]. In that regard, *in vitro* transcriptional profiles might not have a direct relationship to *in situ* and *in vivo* profiles,

thus this technique could open the door to new research that might eventually yield better therapeutics more suited to the *in vivo* biological phenomena.

III. 5 PROTEOMICS

The downstream products of transcripts offer an alternative view on the dynamics of the effect of antibiotics. In particular, there are post-transcriptional mechanisms whose role is often central to understanding which genes actually yield proteins. Moreover, proteins are often the primary targets of antibiotics, so the application of proteomics for MOA identification is well-justified, and as a technique, proteomics is increasingly more prominent in drug discovery [60].

III.5.1 GEL-BASED ASSAYS

Early proteomics studies began with gel-based methods, e.g., 2D gel electrophoresis, which profiled the protein constituents of a sample. Comparison of these profiles with the use of fluorophores allowed multiplexing, and as discussed so far, this revealed differential expression [61]. To identify proteins within a certain spot on the gel, digestion followed by mass spectrometry (MS) allowed database querying. In line with this, Wang et al. [62] explored the MOA of juglone, a plant-derived 1,4-naphthoquinone, against *S. aureus*. Of the 21 differentially expressed spots, 13 were identified by matrix assisted laser desorption ionization-time of flight/time of flight MS. These included proteins that participate in the tricarboxylic acid cycle, DNA and RNA synthesis and protein synthesis. However, other studies into the MOA of juglone against other microorganisms have revealed considerably different proteomic responses, and juglones' biomolecular target remains elusive.

Likewise, Bandow et al. [63] compiled a proteomic profile database of *B. subtilis* response to 30 antibiotics, most of which have been well characterized regarding their MOA. While each antibiotic presented a complex proteomic profile, with some overlap across similar compounds, a sufficiently unique profile was obtained for each antibiotic. For example, most disrupters of translation accuracy resulted in up-regulation of heat shock proteins, which are induced by the accumulation of misfolded proteins. In the end, 122 proteins with at least a 2-fold change in regard to the control samples provided sufficiently distinct profiles to identify the MOA of BAY 50-2369, a novel compound that acts at the peptidyltransferase step of protein synthesis.

The identification of the MOA of BAY 50-2369 as a translation inhibitor using proteomics was straightforward in the sense that its proteomic response was similar to other known antibiotics. However, in the case of acyldepsipeptides, their novel MOA could not be extrapolated. Nonetheless, its proteomic profile identified the up-regulation of ClpP and ClpC, two protease subunits, along with down-regulation of the GroEL chaperon and the TU elongation factor, which staged an initial MOA hypothesis. From here, biochemical and chemical genetics assays identified ClpP as the target, which was confirmed with crystallography. Interestingly, acyldepsipeptides don't inhibit ClpP, but release it from ATPase regulation so intact proteins enter its proteolytic chamber, resulting in indiscriminate degradation and cell death [64].

Similarly, Wenzel et al. [65] explored the proteomic response of *B. subtilis* to lantibiotics. More specifically, mersacidin, gallidermin and nisin were investigated given these bind to lipid II, thereby inhibiting cell wall synthesis. However, mersacidin does not integrate with the cytoplasmatic membrane, while nisin fully integrates and induces the formation of large pores, which impairs membrane potential and leads to nutrient and ion leakage. In between these two ends of the spectrum, gallidermin integrates the membrane, but only induces pore formation in some bacteria. These MOA correlated with proteomic profiles, and when compared with the profiles of other antibiotics that target the bacterial envelope compiled by Bandow et al., YtrE, PspA and NadE, along with, YceC were revealed as marker proteins of cell wall biosynthesis inhibition, membrane stress and general cell envelope stress respectively. Since these marker proteins correlate to the specific steps inhibited, they are indicative of a narrow range of possible molecular targets, and thus enlighten the MOA.

More recently, Maaß et al. [66] used proteomic profiling of *Clostridium difficile* to characterize the MOA of metronidazole, vancomycin and fidaxomicin. Here, 425 protein markers constructed profiles specific to the individual response to each antibiotic, with very little overlap across different antibiotics. Even so, metronidazole affected proteins involved in protein biosynthesis and degradation, DNA replication, recombination and repair; fidaxomicin altered the expression of proteins with cell envelope functions, cell motility, transcription and amino acid synthesis; while vancomycin affected a greater diversity of pathways. Although these antibiotics act on different pathways, which validates proteomics as a high-value tool for MOA identification, a higher number of antibiotics should be studied to consolidate the uniqueness of these proteomic profiles. Moreover, testing antibiotics with very similar MOA would allow to test the sensitivity of these profiles in regard to the biomolecular target.

III.5.2 GEL-FREE METHODS

Gel-based methods are limited regarding throughput and require visual comparison prior to peptide mass fingerprinting. This is a severe limitation when constructing large libraries of proteomic profiles of MOA [67]. Alternatively, one of the preferred gel-free proteomics assays, isobaric tags for relative and absolute quantification (iTRAQ), relies on isobaric peptide labelling for chromatographic separation and their quantification using MS. Other gel-free techniques include stable isotope labelling or selected reaction monitoring [68]. Moreover, enrichment techniques such as antibodies, ionic interaction or specific enzymes allow the evaluation of post-translational modifications relevant in the infectious process [69].

One of the advantages of iTRAQ is its multiplexing capabilities, whereby commercially available reagents allows testing of 2-8 samples in a single liquid chromatography separation and MS analysis [70]. In that sense, Ma et al. [71] quantified the proteomic expression of *S. aureus* to daptomycin. In total, 34 proteins were found to be up-regulated, while 17 proteins were down-regulated, of the 872 differentially expressed proteins. Ultimately, the MOA of daptomycin is different from other known antibiotics, and although its exact target remains elusive, two proteins associated with the metabolism of nucleotide acid were validated as a universal response to daptomycin across *S. aureus* strains. Moreover, evidence is building towards cell membrane disruption with sufficient damage to impair its integrity, along with chromosomal aggregation, which results in DNA release and cell death.

Although gel-free proteomics has not been extensively applied in the context of MOA characterization of large antibiotic libraries, it has elucidated protein function and cellular interactions. Larger-scale studies, with comparable procedures, are still required before the felt impact of proteomics matches that of previously discussed techniques. In addition to MOA characterization, proteomics has been used to investigate many different determinants of resistance and infection [72], as well as the formation and dynamics of biofilms [73]. Here, the biological processes are evaluated *in vivo*, as discussed for transcriptomics. One example of such is the iTRAQ study on the cell membrane protein expression of *P. aeruginosa* in cystic fibrosis isolates, which have revealed the host-specific microevolution of this pathogen [74].

III. 6 METABOLOMICS

The last stage of the 'omics' cascade focusses on the metabolites present in a biological system. Metabolites are small molecules that belong to different biomolecular classes, including organic acids, amino acids, fatty acids, sugars, sugar alcohols, steroids, nucleic acids, among others [61]. As such, metabolomics is closer a closer representation of phenotype, thus has enormous potential to enlighten the complete dynamics of bacterial physiology in response to antibiotic exposure [75]. Although the metabolome complements the transcriptome and proteome, it reflects cellular activities that are regulated by a wider range of mechanisms. In other words, the metabolome reflects physiological states in a amplified state in comparison with transcriptomics and proteomics [76].

There are two dominant techniques in metabolomics: Nuclear Magnetic Resonance (NMR) and MS coupled with a chromatographic step. While at first both had a similar scientific output, MS has recently become the preferred approach. Even though the last couple of years have seen an increase in NMR metabolomics databases, both in quantity and quality, when uncatalogued or unknown metabolites need identification, MS is still the go-to technique [77]. Nonetheless, NMR is able to quantify abundant metabolites, and does not require laborious sample preparation, fractioning procedures, nor derivatization, and can analyze difficult to ionize compounds. Moreover, stable isotope labeled NMR can probe the dynamics of metabolite transformation *in vivo*, as the technique is non-destructive [78].

Regardless of the technique, metabolomics can be targeted or untargeted, the difference being that targeted metabolomics aims to quantify and identify a predetermined subset of metabolites. On the other hand, untargeted metabolomics describes the complete metabolome, then identifies the key metabolites regarding the observed phenotype. In theory, untargeted metabolomics is more suited for MOA identification, mostly because no *a. priori* knowledge is required. In practice, a targeted metabolomics protocol can equally predict MOA, as such this distinction was not discussed in the following study reviews.

III.6.1 NUCLEAR MAGNETIC RESONANCE

Kozlowska et al. [79] employed ^{1}H high-resolution magic angle spinning NMR spectroscopy to identify the lowest concentration of antimicrobial peptides that induced a detectable metabolic response in *E. coli*. Although this study did not aim to discriminate

different MOA, its rational was to detect antibiotic-specific responses but also avoid large scale death cell and its associated non-specific response. The authors tested four structurally similar, and physically related, cationic amphipathic antimicrobial peptides with different degrees of activity. Unique metabolic responses for each compound at sub-inhibitory concentrations were found, and these were not coherent with the observed minimum inhibitory concentration (MIC) and recovery assays, which suggest some sort of phenomena related to the ratio between antibiotic and bacteria and goes to show the gap between observable phenotype and underlying biological mechanisms.

Hoerr et al. [80] used NMR spectroscopy to study both the intracellular fingerprint and extracellular footprint of *E. coli* when exposed to 9 antibiotics from 5 classes, which inhibited protein synthesis, nucleic acid synthesis or cell wall biosynthesis. While antibiotics acting on intracellular targets consistently presented fingerprints coherent with class-action, only antibiotics acting on the cell wall had a distinct metabolic footprint, which has been associated to the loss of membrane integrity and subsequent metabolite leakage. As such, metabolic fingerprinting is a more relevant approach to explore antibiotic MOA, despite the fact that it is more laborious. More importantly, a descriptive model built from the metabolic profiling data was used to successfully predict the MOA of antibiotics not used in the training dataset. Even so, further studies with more MOA diversity, as well as larger libraries, were deemed necessary.

Interestingly, Birkenstock et al. [7] conducted a exometabolome analysis using ^{1}H NMR spectroscopy, which determined the target of triphenulbismuthdichloride as pyruvate dehydrogenase in *S. aureus*. This analysis revealed that pyruvate concentrations strongly increase with exposure to triphenulbismuthdichloride, along with detectable suppressive effects on glucose and amino acid consumption, as well as reduced accumulation of acetate. Additionally, acetolactate, acetoine, butanediole, lactate, formate and ethanol also accumulated, which indicated that accumulated pyruvate was directed to alternative pathways. As such, this antibiotic was suspected of interfering with pyruvate catabolism, which was the basis for further enzymatic analysis to pinpoint the exact target.

III.6.2 Mass spectrometry

Regarding MS-based metabolomics, two separation techniques are typically employed prior to mass spectra acquisition, Gas Chromatography (GC) and Liquid Chromatography (LC), although others have been described. Small molecules in general, and metabolites in particular,

are subject to high temperatures in the preparation and analysis with GC-MS. In the case of blood plasma, this alters the molecular peak pattern by up to 40%, including the formation of degradation and transformation products [81]. Because of this, and given LC achieves better separation, the focus of this section was placed on LC-MS.

Schelli, Zhong & Zhu [82] compared the metabolic response of a methicillin-resistant and susceptible strain of *S. aureus* to sub-inhibitory concentrations of ampicillin, kanamycin and norfloxacin, which belong to the classes of β-lactams, aminoglycosides and quinolones, respectively. Interestingly, minor differences were induced in the sensitive strain by kanamycin and norfloxacin, but marked differences were induced by ampicillin. A similar pattern was observed on the resistant strain, however these differences in metabolic profile were more prominent across the three antibiotics with the susceptible strain. In total, 109 and 107 metabolites were significantly altered by antibiotic exposure for the sensitive and resistant strains, respectively. Moreover, Principal Component Analysis (PCA) score plots of the metabolic response showed moderately good separation according to class. On the downside, dispersion of untreated samples with either strain highlighted a critical limitation of MS-based metabolomics, namely its high sensitivity often masks the MOA-related signal.

Zampieri et al. [17] probed the initial response (1 minute to 1 hour) of *E. coli* to 9 antibiotics and hydrogen peroxide. Here, 324 putatively annotated metabolites were significant altered in comparison with untreated samples, out of 784 detected. In general, antibiotic exposure had an extensive effect on metabolism. Not only did antibiotic exposure alter a large number of metabolites, but 37 of these displayed similar patterns after exposure to at least two compounds with considerably different MOA. Even so, because not all significantly altered metabolites displayed the same pattern across all antibiotics, as well as the fact that all samples exposed to antibiotic with similar MOA had at least one exclusive responsive metabolite, except for amoxicillin and ampicillin, the authors highlight the suitability of MS-based metabolomics to unravel the MOA of antibiotics.

One limitation of metabolomics techniques is its high sensitivity to subtle between-sample variations. For instance, some antibiotics are better dissolved in particular solvents, and this often masks the MOA-related signal that is the objective of these studies [80]. Another limitation regards peak annotation using databases of known metabolites [83]. Also, metabolomics is still limited regarding throughput, although a 10-100x gain can be achieved regarding proteomics, which requires peptide fragmentation and more demanding acquisition

conditions [17]. Although chromatography-free MS-based metabolomics can achieve a throughput of 10,000 samples/day with a single flow injection electrospray equipment, cycle time was expected to be reduced to anywhere between 4-8 seconds in the near future [84].

In that regard, Zampieri et al. [85] optimized a chromatography-free protocol for rapid metabolome profiling, relying on microtiter plates, thus suitable for automation and increased throughput. Using direct flow injection high-resolution MS, MOA profiles of 62 reference compounds, where gathered with *Mycobacterium smegmatis*. These served to extrapolate the MOA of 212 compounds of a GlaxoSmithKline antibacterial library, a few of which were experimentally validated on *M. tuberculosis*. Despite the scale of this study, only 8% of said library targeted unconventional cellular processes, e.g., those involving the trehalose and lipid metabolism, an indication of the challenge that is to discover novel MOA. Importantly, MS-metabolomics identified a large range of MOA, some of which putative, even in the absence of growth inhibition. However, discerning drug-target effects from indirect metabolic adaptations is difficult and requires further investigation, for instance, with transcriptomics, chemical genetics, or biochemical assays.

III. 7 BACTERIAL CYTOLOGICAL PROFILING

The evolution of technologies such as electron and fluorescence microscopy has been critical towards understanding various microbial processes, including those between pathogens and their host. Importantly, the simultaneous acquisition of multiple parameters from microscopy images enabled profiling the MOA of antibiotics towards their prediction [86]. First introduced by Giuliano et al. [87], High-Content Screening (HCS) was proposed as a technique for drug discovery that combined high-throughput screening with a tabletop instrument capable of reading up to four channels of fluorescence at sub-cellular resolution. HCS was originally demonstrated with the drug-induced transport of GFP from the cytoplasm to the nucleus of human tumor cells. Several hurdles had to be cleared for HCS to be applied to prokaryotes. For instance, bacterial cells lack the organelles for which fluorescent labels were developed, and the high-throughput image-based screening technologies were not suitable at the magnifications required to analyze bacterial cells. In addition to this, because of their smaller size, fluorescent signals were generally weaker, which complicated subsequent analysis. Nonetheless, the rational was that antibiotic-induced morphological changes would be indicative of the MOA of antibiotics [6].

One of the first attempts at HCS of bacteria was described by Peach et al. [88], who used epifluorescence microscopy to build cytological profiles of intertwined monolayers of *Vibrio cholerae* challenged with a set of 58 FDA-approved antibiotics with known MOA. Using feature segmentation and extraction, key structural metrics composed cytological profiles, from which the MOA of a natural product library was estimated. These profiles identified the three major pathways affected by antibiotics, namely protein, DNA and cell wall synthesis. Although cytological profiles were unsuitable to pinpoint the exact molecular target, these were beneficial in large-scale screening campaigns as a triage step applicable at early stages of discovery programs, thereby minimizing and guiding subsequent testing [89].

Similarly, and nearly simultaneously, Nonejuie et al. coined this technique as Bacterial Cytological Profiling (BCP), and used it to ascertain the MOA of 41 antibiotics, of 26 structural classes, on *E. coli* [13]. Firstly, the profiles induced by inhibitors of transcription, translation, DNA replication, lipid and peptidoglycan synthesis were gathered, and these presented considerable differences that served to distinguish among them. Then, cytological profiles induced by antibiotic of different classes, but inhibiting the same pathway, were clustered. Namely, three clusters of antibiotics were found for protein synthesis inhibitors: those that completely block elongation, those that promote mistranslation and alter membrane permeability, and those which result in premature termination. Therefore, cytological profiling seemed suitable to discriminate the effect of compounds beyond protein synthesis inhibition. A similar result was found for the remaining major pathways targeted, albeit MOA could not be resolved in finer detail.

Although there had been previous studies applying fluorescence microscopy to reveal the biological effects of antibiotics [90], the notion of comparing cytological profiles to provide insight into the MOA of antibiotics was first reported by Lamsa et al. [91]. But it was the work of Peach et al. and Nonejuie et al. who consolidated BCP as a primary screening technique for antibiotic discovery from large libraries, and for tangential objectives likewise, such as susceptibility testing [92].

Most importantly, BCP has been adopted by the ‘big pharma’. For instance, McLoed et al. [93] used BCP to further validate the suspected MOA of a promising compound identified after a high-throughput phenotypic screen of ~1.2 million compounds of the AstraZeneca compound collection against *E. coli*. An alternative use of BCP was described by Zoffman et al. [12], who aimed to identify the lowest concentration of antibiotics that induced significant

changes to bacterial phenotype. Because phenotypic changes were detected under the MIC, this approach effectively expanded the 'screenable' space of the compound library to include the grey chemical matter, which is characterized by limited activity that could be improved with medicinal chemistry efforts, and eventually reach the clinic.

III. 8 VIBRATIONAL SPECTROSCOPY

Vibrational spectroscopy (VS) is based on chemical bonds having a unique vibrational energy. Thus, a samples spectrum reflects the vibrational modes of major cellular biomolecules, such as proteins, carbohydrates, lipids or nucleic acids. This provides a sensitive and complete metabolic fingerprint, which can be obtained using simple, rapid, reagent-less and label-free procedures [94]. Within VS, two techniques hold great potential: Raman Scattering (RS) spectroscopy (RSS) and Fourier-Transform Infrared (FTIR) spectroscopy (FTIRS). FTIRS measures the vibrational modes of molecular bonds that result from dipole moment changes, i.e., charge differences in the electric field of atoms. RSS probes electric polarizability changes, so it is often complementary to FTIRS. While these are far from generating comprehensive metabolite-level data on the metabolome, the information provided is sufficiently revealing of the metabolic networks involved to be applied for metabolic fingerprinting [76], and ultimately MOA-centric studies.

III.8.1 RAMAN SCATTERING SPECTROSCOPY

López-Díez et al. [95] investigated the effect of amikacin, an aminoglycoside antibiotic, on *P. aeruginosa* using UV resonance RSS, which is particularly suitable to probe nucleic acids and aromatic amino acids. Qualitative and quantitative multivariate analysis on the concentration-dependent effect of amikacin revealed that, as the concentration increased, there was a shift from protein-associated bands towards nucleic acid peaks. This finding is coherent with the MOA of amikacin, which binds to ribosomal RNA, resulting in the misincorporation of amino acids and therefore inhibition of translation. As protein synthesis is repressed, there is an accumulation of nucleic acids and reduction of proteins in the cell.

Similarly, Athamneh et al. [96] employed RSS to evaluate the antibiotic response of *E. coli* to 15 antibiotics of 5 classes. Firstly, the objective was to discriminate each antibiotics' effect using a PCA followed by a linear discriminant analysis, which achieved an accuracy of 83.6%. Then, some antibiotics were held out from the model building and the analysis was

repeated, yielding a less impressive accuracy of 48%. While this study is a more robust proof-of-concept than the previous, given a more comprehensive set of antibiotics and their MOA were considered, the question whether a truly novel MOA could be predicted remains.

On a slightly different note, Liu et al. [97] explored surface-enhanced RSS, which is particularly adequate to study low-abundance molecules, towards the susceptibility testing of both *E. coli* and *S. aureus*. While this study was not aimed at ascertaining the MOA, several RS spectra alterations were coherent with the MOA of antibiotics. Although this issue was not fully considered, it suggested the possibility of MOA identification with this technique.

A similar conclusion was reached by Teng et al. [98], who explored single-cell RSS to differentiate the *E. coli* stress response to ethanol, ampicillin, kanamycin, n-butanol and heavy metals. Here, RS bands were related with distinct stress responses, and more importantly with the mechanism of the stressor. Using a combination of RSS and transcriptomics, Germond et al. [99] predicted the acquired resistance mechanisms of 10 laboratory-evolved strains *E. coli*, even in the absence of antibiotics. Interestingly, a linear relationship was found between resistance mechanism-associated bands and the expression levels of the genes known to grant resistance to the action of antibiotics.

III.8.2 FOURIER-TRANSFORM INFRARED SPECTROSCOPY

Nguyen et al. [100] explored FTIRS in combination with RSS to individually compare *E. coli* control samples with those exposed to ampicillin, cefotaxime, tetracycline and ciprofloxacin. Interestingly, both FTIRS and RSS distinguished the control samples at different time-points (3, 6, 8 and 24h), with RSS revealing more bands indicative of growth phase. Regarding the antibiotic-exposed samples, altered bands detected with FTIRS include those associated with carbohydrates and proteins, while those detected with RSS were associated with nucleic acids and phenylalanine. Unfortunately, a simultaneous comparison of all antibiotics was not presented, so evaluating the ability of either FTIRS or RSS to ascertain MOA is challenging. Even so, it is interesting to note that the standard deviation of technical replicates was higher with RSS than with FTIRS.

Huleihel, Pavlov & Erukhimovitch [101] used FTIRS to differentiate the effect of caffeic acid phenethyl ester, a natural honeybee product with potent antimicrobial activity, and ampicillin, on 9 Gram negative and 8 Gram positive bacteria. The effect of this compound was

remarkably different on Gram negative and Gram positive bacteria, which could be due to its biological effect. An increase of the bands associated with proteins and sugars, along with a major reduction of the band associated with nucleic acids, was observed in Gram positive bacteria. Similarly, an increase of the band associated with sugar content, along a decrease of the protein-associated band, was detected in Gram negative microbes.

FTIRS has been underexplored for MOA classification, but a few examples with tangential objectives can be brought to the discussion. Moen et al. [102] compared the global transcriptomic profile and FTIR spectra of *E. coli* when challenged with 10 adverse conditions. Although 40% of the 4279 genes investigated had differential expression, no correlation between the transcriptional profile and the biomolecular profile obtained with FTIRS was found. Nonetheless, a PCA on the spectral regions associated with fatty acids, proteins and carbohydrates revealed stress induced sample separation. Similarly, Corte et al. [103] developed a FTIRS toxicity assay with *Saccharomyces cerevisiae*, for which various spectral stress indexes were developed. While their study had an ecotoxicology focus, the stress response to ethanol, sodium hypochlorite, sodium chloride and sulfur dioxide at low concentrations and after a short exposure was clearly captured on the FTIR spectra.

Given FTIR spectroscopy has been underexplored but holds great potential, we dwelled into its application towards MOA identification. Firstly, we applied a macro-cultivation assay, from which it became clear that metabolic fingerprints reflect the MOA of antibiotics [104]. We then refined our assay into a high-throughput micro-cultivation protocol, from which we successfully predicted the MOA of antibiotics at the level of the major biosynthetic pathway, class and individual antibiotics. Moreover, MOA was accurately predicted, at all levels, when models were trained with similar samples, to simulate cases of rediscovery, and when models were trained without similar samples, to simulate novelty. Our assay seems to be suitable to probe the grey chemical matter, as when the dose-response of MOA prediction was determined, an average growth inhibition of 15% yielded 70% accuracy at MOA prediction. Lastly, using spectra that were obtained from normalized samples regarding biomass, we were able to predict growth inhibition, which suggested that metabolic fingerprints obtained with FTIR spectra have intrinsic patterns that reflect growth inhibition beyond cell density, and opened the door to a single-step assay that simultaneously predicts MOA and potency [105].

III. 9 CONCLUDING REMARKS

Throughout this review, the application of various techniques to MOA identification and characterization have been described. Given MOA identification is a major bottleneck of the phenotypic screening approach, and that phenotypic screening has been the most successful in delivering first-in-class antibiotics, techniques for MOA identification are increasingly important. The focus of this review was on techniques capable of both sufficient throughput to be employed in screening campaigns of large libraries, and sufficient sensitivity to accurately distinguish antibiotics MOA. Additionally, given an increasing appreciation that antibiotics elicit system-wide responses, emphasis was put on techniques that output a holistic MOA profile.

While mutants of overexpression or knockdown/knockout libraries evaluate a gene at a time, a single reporter strain probes the genes under the regulation of a promoter. Promoter-reporter systems share some advantages with mutant libraries, but also some disadvantages. While the readout is also simple and can be associated with a cluster of genes, collections comprise a large number of reporter strains (e.g., >1,500), crippling the throughput of this technique. Some strategies attempted to reduce the number of strains to a minimum while obtaining relevant profiles. However, additional disadvantages include the background noise of fluorescent labels, or an inability to probe post-transcriptional events, which results in a biological gap between gene transcription and phenotype that this technique does not cover.

An alternative to reporters of gene transcription is to probe the transcriptome. The advantage is that transcriptional profiles provide a system-level readout, but this made the biological interpretation of the results more challenging. Regarding hybridization-based transcriptomics, these are generally lower-cost and their multiplex capabilities imply a good degree of throughput. However, only the genes used to build the microarray can be detected, and issues arise due to cross-hybridization, as well as with low-abundance transcripts. Beyond hybridization techniques, those based on sequencing have seen their cost steadily decrease into reasonable levels, which enables a more commonplace usage. In addition to detecting all transcripts of a sample, RNA-seq allows querying post-transcriptional events, which is a step forward towards closing the gap between gene transcription and phenotype.

Most antibiotics target proteins. As such, proteomics is well-suited to investigate MOA. Additionally, because proteins are the end-product of genes, the effect of post-transcriptional and post-translational regulatory mechanisms can be queried, which is a step closer to

understanding the biological mechanisms ruling antibiotic-induced phenotype. While proteomics does not require manipulated strains, it is generally a laborious, low-throughput and challenging technique. Despite their critical role in understanding bacterial physiology, and the MOA of antibiotics, proteomic profiles are less distinctive of MOA than transcriptomics. Also, the throughput of proteomics has not reached that of metabolomics. As such, there have been fewer studies applying proteomics to illuminate the MOA of large antibiotic libraries.

The last component of the Omics cascade are metabolites, which have higher variability in terms of structure and biological function, but are a closer reflection of phenotype than proteins, transcripts and genes. As such, metabolomics can not only pinpoint an antibiotics' target, but also identify the MOA of antibiotics that don't target the metabolism. Despite a few promising protocols having 10-100x the throughput of proteomics, high-throughput metabolomics-based MOA identification is still in its infancy. For NMR, low metabolome coverage and fewer annotation resources are common issues. Regarding MS, the optimization of chromatographic separation, peak annotation, and masking of MOA-related signal by subtle undesirable sample variations are still major bottlenecks. Nonetheless, MS has been successfully applied to large compound libraries, which validates its use for screening of large antibiotic libraries.

An alternative to the techniques discussed thus far lies in BCP, which was originally developed as HCS for eukaryotic cells. Although its application to prokaryotes only recently gained traction, the ultrastructural and morphological alterations induced by antibiotics yield a cytological profile with good predictive ability of the MOA of antibiotics. The issue with this approach, which has a different perspective on the issue of MOA, is one of sensitivity. While the aptitude to detect the major pathway affected has been well established, contradictory findings have been reported regarding the ability to separate profiles induced by drugs that act on the same pathway, and it is still unclear if BCP can be used as a standalone technique or requires complementation with biomolecular techniques. Nonetheless, its high- to very high-throughput has made BCP an attractive technique, rightfully justifying further efforts towards its consolidation in the discovery pipeline.

Another alternative is VS, which has also been increasingly gaining traction for metabolic fingerprinting. Similarly to BCP, this technique is promising because it enables adequate throughput for screening purposes with sufficient biological information for MOA identification. However, because it reflects the biochemical composition of the sample, it has

the potential to be more informative than BCP, without requiring complementary assays. Also, because these techniques are high-throughput, reagent-less, label-free and involve reduced or no sample preparation, they hold tremendous potential. Even so, few studies have explored large antibiotic libraries, which by itself justifies a deeper evaluation on the application of VS towards antibiotic discovery.

Due to the lack of novelty in the antibiotic pipeline, and the desperate need for new antibiotics, it has become increasingly evident that a new approach is needed. The ability of the various Omics techniques discussed, as well as BCP and VS, to rapidly and robustly detect MOA could enable a shift from the activity-based antibiotic discovery paradigm towards a mechanism-based approach. In turn, a mechanism-based discovery approach increases the 'screenable' space by probing the grey chemical matter, an underexplored source that is extremely promising because medicinal chemistry can improve the properties of these compounds, in particular their potency. Moreover, a low-cost high-throughput technique for MOA identification can also aid during this process by rapidly excluding off-target liabilities. With a higher number of hits, and their better guided structural optimization, it is expected that new compounds can eventually reach the clinic and spark a new generation of antibiotics.

III. 10 References

1. De Mol, M.L.; Snoeck, N.; De Maeseneire, S.L.; Soetaert, W.K. Hidden antibiotics: Where to uncover? *Biotechnol. Adv.* **2018**, *36*, 2201–2218.
2. Ribeiro da Cunha; Fonseca; Calado Antibiotic Discovery: Where Have We Come from, Where Do We Go? *Antibiotics* **2019**, *8*, 45.
3. Kubota, K.; Funabashi, M.; Ogura, Y. Target deconvolution from phenotype-based drug discovery by using chemical proteomics approaches. *Biochim. Biophys. Acta - Proteins Proteomics* **2019**, *1867*, 22–27.
4. Ohki, Y.; Sakurai, H.; Hoshino, M.; Terashima, H.; Shimizu, H.; Ishikawa, T.; Ogiyama, T.; Muramatsu, Y.; Nakanishi, T.; Miyazaki, S.; et al. Perturbation-Based Proteomic Correlation Profiling as a Target Deconvolution Methodology. *Cell Chem. Biol.* **2019**, *26*, 137–143.
5. Phillips, J.W.; Goetz, M.A.; Smith, S.K.; Zink, D.L.; Polishook, J.; Onishi, R.; Salowe, S.; Wiltsie, J.; Allocco, J.; Sigmund, J.; et al. Discovery of kibdelomycin, a potent new class of bacterial type II topoisomerase inhibitor by chemical-genetic profiling in Staphylococcus aureus. *Chem. Biol.* **2011**, *18*, 955–965.
6. Kurita, K.L.; Glassey, E.; Linington, R.G. Integration of high-content screening and untargeted metabolomics for comprehensive functional annotation of natural product libraries. *Proc. Natl. Acad. Sci.* **2015**, *112*, 11999–12004.
7. Birkenstock, T.; Liebeke, M.; Winstel, V.; Krismer, B.; Gekeler, C.; Niemiec, M.J.; Bisswanger, H.; Lalk, M.; Peschel, A. Exometabolome analysis identifies pyruvate dehydrogenase as a target for the antibiotic triphenylbismuthdichloride in multiresistant bacterial pathogens. *J. Biol. Chem.* **2012**, *287*, 2887–2895.
8. Cho, H.; Uehara, T.; Bernhardt, T.G. Beta-lactam antibiotics induce a lethal malfunctioning of the

bacterial cell wall synthesis machinery. *Cell* **2014**, *159*, 1300–1311.

9. French, S.; Ellis, M.J.; Coutts, B.E.; Brown, E.D. Chemical genomics reveals mechanistic hypotheses for uncharacterized bioactive molecules in bacteria. *Curr. Opin. Microbiol.* **2017**, *39*, 42–47.
10. Cunningham, M.L.; Kwan, B.P.; Nelson, K.J.; Bensen, D.C.; Shaw, K.J. Distinguishing on-target versus off-target activity in early antibacterial drug discovery using a macromolecular synthesis assay. *J. Biomol. Screen.* **2013**, *18*, 1018–1026.
11. Bantscheff, M.; Drewes, G. Chemoproteomic approaches to drug target identification and drug profiling. *Bioorganic Med. Chem.* **2012**, *20*, 1973–1978.
12. Zoffmann, S.; Vercruysse, M.; Benmansour, F.; Maunz, A.; Wolf, L.; Blum Marti, R.; Heckel, T.; Ding, H.; Truong, H.H.; Prummer, M.; et al. Machine learning-powered antibiotics phenotypic drug discovery. *Sci. Rep.* **2019**, *9*, 1–14.
13. Nonejuie, P.; Burkart, M.; Pogliano, K.; Pogliano, J. Bacterial cytological profiling rapidly identifies the cellular pathways targeted by antibacterial molecules. *Proc. Natl. Acad. Sci.* **2013**, *110*, 16169–16174.
14. Sato, S. ichi; Murata, A.; Shirakawa, T.; Uesugi, M. Biochemical Target Isolation for Novices: Affinity-Based Strategies. *Chem. Biol.* **2010**, *17*, 616–623.
15. Nishiya, Y.; Hamada, T.; Abe, M.; Takashima, M.; Tsutsumi, K.; Okawa, K. A new efficient method of generating photoaffinity beads for drug target identification. *Bioorganic Med. Chem. Lett.* **2017**, *27*, 834–840.
16. Burdine, L.; Thomas, K. Target Identification in Chemical Genetics: The (Often) Missing Link. *Chem. Biol.* **2004**, *11*, 593–597.
17. Zampieri, M.; Sekar, K.; Zamboni, N.; Sauer, U. Frontiers of high-throughput metabolomics. *Curr. Opin. Chem. Biol.* **2017**, *36*, 15–23.
18. Cacace, E.; Kritikos, G.; Typas, A. Chemical genetics in drug discovery. *Curr. Opin. Syst. Biol.* **2017**, *4*, 35–42.
19. Barker, C.A.; Farha, M.A.; Brown, E.D. Chemical Genomic Approaches to Study Model Microbes. *Chem. Biol.* **2010**, *17*, 624–632.
20. Li, X.; Zolli-Juran, M.; Cechetto, J.D.; Daigle, D.M.; Wright, G.D.; Brown, E.D. Multicopy Suppressors for Novel Antibacterial Compounds Reveal Targets and Drug Efflux Susceptibility. *Chem. Biol.* **2004**, *11*, 1423–1430.
21. Kitagawa, M.; Ara, T.; Arifuzzaman, M.; Ioka-Nakamichi, T.; Inamoto, E.; Toyonaga, H.; Mori, H. Complete set of ORF clones of Escherichia coli ASKA library (A complete set of E. coli K-12 ORF archive): unique resources for biological research. *DNA Res.* **2005**, *12*, 291–299.
22. Pathania, R.; Zlitni, S.; Barker, C.; Das, R.; Gerritsma, D.A.; Lebert, J.; Awuah, E.; Melacini, G.; Capretta, F.A.; Brown, E.D. Chemical genomics in Escherichia coli identifies an inhibitor of bacterial lipoprotein targeting. *Nat. Chem. Biol.* **2009**, *5*, 849.
23. Barker, C.A.; Allison, S.E.; Zlitni, S.; Nguyen, N.D.; Das, R.; Melacini, G.; Capretta, A.A.; Brown, E.D. Degradation of MAC13243 and studies of the interaction of resulting thiourea compounds with the lipoprotein targeting chaperone LolA. *Bioorganic Med. Chem. Lett.* **2013**, *23*, 2426–2431.
24. Muheim, C.; Götzke, H.; Eriksson, A.U.; Lindberg, S.; Lauritsen, I.; Nørholm, M.H.H.; Daley, D.O. Increasing the permeability of Escherichia coli using MAC13243. *Sci. Rep.* **2017**, *7*, 1–11.
25. Nickerson, N.N.; Jao, C.C.; Xu, Y.; Quinn, J.; Skippington, E.; Alexander, M.K.; Miu, A.; Skelton, N.; Hankins, J. V.; Lopez, M.S.; et al. A Novel Inhibitor of the LolCDE ABC Transporter Essential for Lipoprotein Trafficking in Gram-Negative Bacteria. *Antimicrob. Agents Chemother.* **2018**, *62*, 1–16.
26. Baba, T.; Ara, T.; Hasegawa, M.; Takai, Y.; Okumura, Y.; Baba, M.; Datsenko, K.A.; Tomita, M.; Wanner, B.L.; Mori, H. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Syst. Biol.* **2006**, *2*, 2006.0008.

27. Liu, A.; Tran, L.; Becket, E.; Lee, K.; Chinn, L.; Park, E.; Tran, K.; Miller, J.H. Antibiotic sensitivity profiles determined with an Escherichia coli gene knockout collection: Generating an antibiotic bar code. *Antimicrob. Agents Chemother.* **2010**, *54*, 1393–1403.

28. Nichols, R.J.; Sen, S.; Choo, Y.J.; Beltrao, P.; Zietek, M.; Chaba, R.; Lee, S.; Kazmierczak, K.M.; Lee, K.J.; Wong, A.; et al. Phenotypic landscape of a bacterial cell. *Cell* **2011**, *144*, 143–156.

29. Côté, J.-P.; French, S.; Gehrke, S.S.; MacNair, C.R.; Mangat, C.S.; Bharat, A.; Brown, E.D. The Genome-Wide Interaction Network of Nutrient Stress Genes in Escherichia coli . *MBio* **2016**, *7*, 1–12.

30. Shiver, A.L.; Osadnik, H.; Kritikos, G.; Li, B.; Krogan, N.; Typas, A.; Gross, C.A. A Chemical-Genomic Screen of Neglected Antibiotics Reveals Illicit Transport of Kasugamycin and Blasticidin S. *PLoS Genet.* **2016**, *12*, 1–19.

31. Stokes, J.M.; French, S.; Ovchinnikova, O.G.; Bouwman, C.; Whitfield, C.; Brown, E.D. Cold Stress Makes Escherichia coli Susceptible to Glycopeptide Antibiotics by Altering Outer Membrane Integrity. *Cell Chem. Biol.* **2016**, *23*, 267–277.

32. DeVito, J.A.; Mills, J.A.; Liu, V.G.; Agarwal, A.; Sizemore, C.F.; Yao, Z.; Stoughton, D.M.; Cappiello, M.G.; Barbosa, M.D.F.S.; Foster, L.A.; et al. An array of target-specific screening strains for antibacterial discovery. *Nat. Biotechnol.* **2002**, *20*, 478–483.

33. Griffiths, A.J.; Miller, J.H.; Suzuki, D.T.; Lewontin, R.C.; Gelbart, W.M. *An Introduction to Genetic Analysis*; 7th Edition.; W. H. Freeman: New York, 2000;

34. Donald, R.G.K.; Skwish, S.; Forsyth, R.A.; Anderson, J.W.; Zhong, T.; Burns, C.; Lee, S.; Meng, X.; LoCastro, L.; Jarantow, L.W.; et al. A Staphylococcus aureus Fitness Test Platform for Mechanism-Based Profiling of Antibacterial Compounds. *Chem. Biol.* **2009**, *16*, 826–836.

35. Forsyth, R.A.; Haselbeck, R.J.; Ohlsen, K.L.; Yamamoto, R.T.; Xu, H.; Trawick, J.D.; Wall, D.; Wang, L.; Brown-driver, V.; Froelich, J.M.; et al. A genome-wide strategy for the identification of S. aureus essential genes.pdf. **2002**, *43*, 1387–1400.

36. Peters, J.M.; Colavin, A.; Shi, H.; Czarny, T.L.; Larson, M.H.; Wong, S.; Hawkins, J.S.; Lu, C.H.S.; Koo, B.M.; Marta, E.; et al. A comprehensive, CRISPR-based functional analysis of essential genes in bacteria. *Cell* **2016**, *165*, 1493–1506.

37. Liu, X.; Gallay, C.; Kjos, M.; Domenech, A.; Slager, J.; van Kessel, S.P.; Knoops, K.; Sorg, R.A.; Zhang, J.; Veening, J. High-throughput CRISPRi phenotyping identifies new essential genes in *Streptococcus pneumoniae*. *Mol. Syst. Biol.* **2017**, *13*, 931.

38. Elad, T.; Seo, H. Bin; Belkin, S.; Gu, M.B. High-throughput prescreening of pharmaceuticals using a genome-wide bacterial bioreporter array. *Biosens. Bioelectron.* **2015**, *68*, 699–704.

39. Zaslaver, A.; Bren, A.; Ronen, M.; Itzkovitz, S.; Kikoin, I.; Shavit, S.; Liebermeister, W.; Surette, M.G.; Alon, U. A comprehensive library of fluorescent transcriptional reporters for Escherichia coli. *Nat. Methods* **2006**, *3*, 623–628.

40. Bollenbach, T.; Quan, S.; Chait, R.; Kishony, R. Nonoptimal Microbial Response to Antibiotics Underlies Suppressive Drug Interactions. *Cell* **2009**, *139*, 707–718.

41. Baptist, G.; Pinel, C.; Ranquet, C.; Izard, J.; Ropers, D.; De Jong, H.; Geiselmann, J. A genome-wide screen for identifying all regulators of a target gene. *Nucleic Acids Res.* **2013**, *41*.

42. Engelberg, R.; Danielson, A.; Wang, S.; Singh, M.; Wai, A.; Sorensen, J.; Duan, K.; Hausner, G.; Kumar, A. Creation of a drug-sensitive reporter strain of Pseudomonas aeruginosa as a tool for the rapid screening of antimicrobial products. *J. Microbiol. Methods* **2018**, *152*, 1–6.

43. Sun, D.; Cohen, S.; Mani, N.; Murphy, C.; Rothstein, D.M. A Pathway-specific Cell Based Screening System to Detect Bacterial Cell Wall Inhibitors. *J. Antibiot. (Tokyo).* **2002**, *55*, 279–287.

44. Nayar, A.S.; Dougherty, T.J.; Ferguson, K.E.; Granger, B.A.; McWilliams, L.; Stacey, C.; Leach, L.J.; Narita, S.; Tokuda, H.; Miller, A.A.; et al. Novel Antibacterial Targets and Compounds Revealed by a High-Throughput Cell Wall Reporter Assay. *J. Bacteriol.* **2015**, *197*, 1726–1734.

45. Mido, T.; Schaffer, E.M.; Dorsey, R.W.; Sozhamannan, S.; Hofmann, E.R. Sensitive detection of

live Escherichia coli by bacteriophage amplification-coupled immunoassay on the Luminex® MAGPIX instrument. *J. Microbiol. Methods* **2018**, *152*, 143–147.

46. Nagaraj, N.S.; Singh, O. V. Using genomics to develop novel antibacterial therapeutics. *Crit. Rev. Microbiol.* **2010**, *36*, 340–348.

47. Miller, M.B.; Tang, Y.W. Basic concepts of microarrays and potential applications in clinical microbiology. *Clin. Microbiol. Rev.* **2009**, *22*, 611–633.

48. Boshoff, H.I.M.; Myers, T.G.; Copp, B.R.; McNeil, M.R.; Wilson, M.A.; Barry, C.E. The Transcriptional Responses of Mycobacterium tuberculosis to Inhibitors of Metabolism . *J. Biol. Chem.* **2004**, *279*, 40174–40184.

49. Liang, J.; Tang, X.; Guo, N.; Zhang, K.; Guo, A.; Wu, X.; Wang, X.; Guan, Z.; Liu, L.; Shen, F.; et al. Genome-wide expression profiling of the response to linezolid in mycobacterium tuberculosis. *Curr. Microbiol.* **2012**, *64*, 530–538.

50. Bonn, F.; Pané-Farré, J.; Schlüter, R.; Schaffer, M.; Fuchs, S.; Bernhardt, J.; Riedel, K.; Otto, A.; Völker, U.; van Dijl, J.M.; et al. Global analysis of the impact of linezolid onto virulence factor production in S. aureus USA300. *Int. J. Med. Microbiol.* **2016**, *306*, 131–140.

51. Draghici, S.; Khatri, P.; Eklund, A.C.; Szallasi, Z. Reliability and reproducibility issues in DNA microarray measurements. *Trends Genet.* **2006**, *22*, 101–109.

52. Wang, Z.; Gerstein, M.; Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **2009**, *10*, 57–63.

53. Goodwin, S.; McPherson, J.D.; McCombie, W.R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **2016**, *17*, 333.

54. Hua, X.; Chen, Q.; Li, X.; Yu, Y. Global transcriptional response of Acinetobacter baumannii to a subinhibitory concentration of tigecycline. *Int. J. Antimicrob. Agents* **2014**, *44*, 337–344.

55. Wecke, T.; Mascher, T. Antibiotic research in the age of omics: From expression profiles to interspecies communication. *J. Antimicrob. Chemother.* **2011**, *66*, 2689–2704.

56. Briffotaux, J.; Liu, S.; Gicquel, B. Genome-wide transcriptional responses of Mycobacterium to antibiotics. *Front. Microbiol.* **2019**, *10*, 1–14.

57. Howden, B.P.; Beaume, M.; Harrison, P.F.; Hernandez, D.; Schrenzel, J.; Seemann, T.; Francois, P.; Stinear, T.P. Analysis of the Small RNA Transcriptional Response in Multidrug-Resistant Staphylococcus aureus after Antimicrobial Exposure. *Antimicrob. Agents Chemother.* **2013**, *57*, 3864–3874.

58. Molina-Santiago, C.; Daddaoua, A.; Gómez-Lozano, M.; Udaondo, Z.; Molin, S.; Ramos, J.L. Differential transcriptional response to antibiotics by Pseudomonas putidaDOT-T1E. *Environ. Microbiol.* **2015**, *17*, 3251–3262.

59. Hébert, F.O.; Boyle, B.; Levesque, R.C. Direct In Vivo Microbial Transcriptomics During Infection. *Trends Microbiol.* **2018**, *26*, 732–735.

60. Frantzi, M.; Latosinska, A.; Mischak, H. Proteomics in Drug Development: The Dawn of a New Era? *Proteomics - Clin. Appl.* **2019**, *13*, 1–13.

61. dos Santos, B.S.; da Silva, L.C.N.; da Silva, T.D.; Rodrigues, J.F.S.; Grisotto, M.A.G.; Correia, M.T. do. S.; Napoleão, T.H.; da Silva, M. V.; Paiva, P.M.G. Application of omics technologies for evaluation of antibacterial mechanisms of action of plant-derived products. *Front. Microbiol.* **2016**, *7*, 1–13.

62. Wang, J.; Wang, Z.; Wu, R.; Jiang, D.; Bai, B.; Tan, D.; Yan, T.; Sun, X.; Zhang, Q.; Wu, Z. Proteomic Analysis of the Antibacterial Mechanism of Action of Juglone against Staphylococcus aureus. *Nat. Prod. Commun.* **2016**, *11*, 825–827.

63. Bandow, J.E.; Brötz, H.; Leichert, L.I.O.; Labischinski, H.; Hecker, M. Proteomic approach to understanding antibiotic action. *Antimicrob. Agents Chemother.* **2003**, *47*, 948–955.

64. Kim, W.; Hendricks, G.L.; Tori, K.; Fuchs, B.B.; Mylonakis, E. Strategies against methicillin-

resistant Staphylococcus aureus persisters. *Future Med. Chem.* **2018**, *10*, 779–794.

65. Wenzel, M.; Kohl, B.; Münch, D.; Raatschen, N.; Albada, H.B.; Hamoen, L.; Metzler-Nolte, N.; Sahl, H.-G.; Bandow, J.E. Proteomic Response of Bacillus subtilis to Lantibiotics Reflects Differences in Interaction with the Cytoplasmic Membrane. *Antimicrob. Agents Chemother.* **2012**, *56*, 5749–5757.

66. Maaß, S.; Otto, A.; Albrecht, D.; Riedel, K.; Trautwein-Schult, A.; Becher, D. Proteomic Signatures of Clostridium difficile Stressed with Metronidazole, Vancomycin, or Fidaxomicin. *Cells* **2018**, *7*, 213.

67. Brötz-Oesterhelt, H.; Bandow, J.E.; Labischinski, H. Bacterial proteomics and its role in antibacterial drug discovery. *Mass Spectrom. Rev.* **2005**, *24*, 549–565.

68. Evans, C.; Noirel, J.; Ow, S.Y.; Salim, M.; Pereira-Medrano, A.G.; Couto, N.; Pandhal, J.; Smith, D.; Pham, T.K.; Karunakaran, E.; et al. An insight into iTRAQ: Where do we stand now? *Anal. Bioanal. Chem.* **2012**, *404*, 1011–1027.

69. Olsen, J. V; Mann, M. Status of large-scale analysis of post-translational modifications by mass spectrometry. *Mol. Cell. Proteomics* **2013**, *12*, 3444–52.

70. Chen, B.; Zhang, D.; Wang, X.; Ma, W.; Deng, S.; Zhang, P.; Zhu, H.; Xu, N.; Liang, S. Proteomics progresses in microbial physiology and clinical antimicrobial therapy. *Eur. J. Clin. Microbiol. Infect. Dis.* **2017**, *36*, 403–413.

71. Ma, W.; Zhang, D.; Li, G.; Liu, J.; He, G.; Zhang, P.; Yang, L.; Zhu, H.; Xu, N.; Liang, S. Antibacterial mechanism of daptomycin antibiotic against Staphylococcus aureus based on a quantitative bacterial proteome analysis. *J. Proteomics* **2017**, *150*, 242–251.

72. Fields, F.R.; Lee, S.W.; McConnell, M.J. Using bacterial genomes and essential genes for the development of new antibiotics. *Biochem. Pharmacol.* **2017**, *134*, 74–86.

73. Sun, L.; Chen, H.; Lin, W.; Lin, X. Quantitative proteomic analysis of Edwardsiella tarda in response to oxytetracycline stress in biofilm. *J. Proteomics* **2017**, *150*, 141–148.

74. Kamath, K.S.; Pascovici, D.; Penesyan, A.; Goel, A.; Venkatakrishnan, V.; Paulsen, I.T.; Packer, N.H.; Molloy, M.P. Pseudomonas aeruginosa Cell Membrane Protein Expression from Phenotypically Diverse Cystic Fibrosis Isolates Demonstrates Host-Specific Adaptations. *J. Proteome Res.* **2016**, *15*, 2152–2163.

75. Pulido, M.R.; García-Quintanilla, M.; Gil-Marqués, M.L.; McConnell, M.J. Identifying targets for antibiotic development using omics technologies. *Drug Discov. Today* **2016**, *21*, 465–472.

76. Goodacre, R.; Vaidyanathan, S.; Dunn, W.B.; Harrigan, G.G.; Kell, D.B. Metabolomics by numbers: Acquiring and understanding global metabolite data. *Trends Biotechnol.* **2004**, *22*, 245–252.

77. Bingol, K. Recent Advances in Targeted and Untargeted Metabolomics by NMR and MS/NMR Methods. *High-Throughput* **2018**, *7*, 9.

78. Markley, J.L.; Brüschweiler, R.; Edison, A.S.; Eghbalnia, H.R.; Powers, R.; Raftery, D.; Wishart, D.S. The future of NMR-based metabolomics. *Curr. Opin. Biotechnol.* **2017**, *43*, 34–40.

79. Kozlowska, J.; Vermeer, L.S.; Rogers, G.B.; Rehnnuma, N.; Amos, S.B.T.A.; Koller, G.; McArthur, M.; Bruce, K.D.; Mason, A.J. Combined Systems Approaches Reveal Highly Plastic Responses to Antimicrobial Peptide Challenge in Escherichia coli. *PLoS Pathog.* **2014**, *10*.

80. Hoerr, V.; Duggan, G.E.; Zbytnuik, L.; Poon, K.K.H.; Große, C.; Neugebauer, U.; Methling, K.; Löffler, B.; Vogel, H.J. Characterization and prediction of the mechanism of action of antibiotics through NMR metabolomics. *BMC Microbiol.* **2016**, *16*, 1–14.

81. Fang, M.; Ivanisevic, J.; Benton, H.P.; Johnson, C.H.; Patti, G.J.; Hoang, L.T.; Uritboonthai, W.; Kurczy, M.E.; Siuzdak, G. Thermal Degradation of Small Molecules: A Global Metabolomic Investigation. *Anal. Chem.* **2015**, *87*, 10935–10941.

82. Schelli, K.; Zhong, F.; Zhu, J. Comparative metabolomics revealing Staphylococcus aureus metabolic response to different antibiotics. *Microb. Biotechnol.* **2017**, *10*, 1764–1774.

83. Chaleckis, R.; Meister, I.; Zhang, P.; Wheelock, C.E. Challenges, progress and promises of metabolite annotation for LC–MS-based metabolomics. *Curr. Opin. Biotechnol.* **2019**, *55*, 44–50.

84. Fuhrer, T.; Zamboni, N. High-throughput discovery metabolomics. *Curr. Opin. Biotechnol.* **2015**, *31*, 73–78.

85. Zampieri, M.; Szappanos, B.; Buchieri, M.V.; Trauner, A.; Piazza, I.; Picotti, P.; Gagneux, S.; Borrell, S.; Gicquel, B.; Lelievre, J.; et al. High-throughput metabolomic analysis predicts mode of action of uncharacterized antimicrobial compounds. *Sci. Transl. Med.* **2018**, *10*, 1–12.

86. Ang, M.L.T.; Pethe, K. Contribution of high-content imaging technologies to the development of anti-infective drugs. *Cytom. Part A* **2016**, *89*, 755–760.

87. Giuliano, K.A.; DeBiasio, R.L.; Dunlay, R.T.; Gough, A.; Volosky, J.M.; Zock, J.; Pavlakis, G.N.; Taylor, D.L. High-Content Screening: A New Approach to Easing Key Bottlenecks in the Drug Discovery Process. *J. Biomol. Screen.* **1997**, *2*, 249–259.

88. Peach, K.C.; Bray, W.M.; Winslow, D.; Linington, P.F.; Linington, R.G. Mechanism of action-based classification of antibiotics using high-content bacterial image analysis. *Mol. Biosyst.* **2013**, *9*, 1837–1848.

89. Schulze, C.J.; Bray, W.M.; Woerhmann, M.H.; Stuart, J.; Lokey, R.S.; Linington, R.G. "function-first" lead discovery: Mode of action profiling of natural product libraries using image-based screening. *Chem. Biol.* **2013**, *20*, 285–295.

90. Liu, W.T.; Yang, Y.L.; Xu, Y.; Lamsa, A.; Haste, N.M.; Yang, J.Y.; Ng, J.; Gonzalez, D.; Ellermeier, C.D.; Straight, P.D.; et al. Imaging mass spectrometry of intraspecies metabolic exchange revealed the cannibalistic factors of Bacillus subtilis. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 16286–16290.

91. Lamsa, A.; Liu, W.T.; Dorrestein, P.C.; Pogliano, K. The Bacillus subtilis cannibalism toxin SDP collapses the proton motive force and induces autolysis. *Mol. Microbiol.* **2012**, *84*, 486–500.

92. Quach, D.T.; Sakoulas, G.; Nizet, V.; Pogliano, J.; Pogliano, K. Bacterial Cytological Profiling (BCP) as a Rapid and Accurate Antimicrobial Susceptibility Testing Method for Staphylococcus aureus. *EBioMedicine* **2016**, *4*, 95–103.

93. McLeod, S.M.; Fleming, P.R.; MacCormack, K.; McLaughlin, R.E.; Whiteaker, J.D.; Narita, S.; Mori, M.; Tokuda, H.; Miller, A.A. Small-Molecule Inhibitors of Gram-Negative Lipoprotein Trafficking Discovered by Phenotypic Screening. *J. Bacteriol.* **2015**, *197*, 1075–1082.

94. Marques, V.; Cunha, B.; Couto, A.; Sampaio, P.; Fonseca, L.P.; Aleixo, S.; Calado, C.R.C. Characterization of gastric cells infection by diverse Helicobacter pylori strains through Fourier-transform infrared spectroscopy. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2019**, *210*, 193–202.

95. López-Díez, E.C.; Winder, C.L.; Ashton, L.; Currie, F.; Goodacre, R. Monitoring the mode of action of antibiotics using raman spectroscopy: Investigating subinhibitory effects of amikacin on Pseudomonas aeruginosa. *Anal. Chem.* **2005**, *77*, 2901–2906.

96. Athamneh, A.I.M.; Alajlouni, R.A.; Wallace, R.S.; Seleem, M.N.; Sengera, R.S. Phenotypic profiling of antibiotic response signatures in Escherichia coli using raman spectroscopy. *Antimicrob. Agents Chemother.* **2014**, *58*, 1302–1314.

97. Liu, C.Y.; Han, Y.Y.; Shih, P.H.; Lian, W.N.; Wang, H.H.; Lin, C.H.; Hsueh, P.R.; Wang, J.K.; Wang, Y.L. Rapid bacterial antibiotic susceptibility test based on simple surface-enhanced Raman spectroscopic biomarkers. *Sci. Rep.* **2016**, *6*, 1–15.

98. Teng, L.; Wang, X.; Wang, X.; Gou, H.; Ren, L.; Wang, T.; Wang, Y.; Ji, Y.; Huang, W.E.; Xu, J. Label-free, rapid and quantitative phenotyping of stress response in E. coli via ramanome. *Sci. Rep.* **2016**, *6*.

99. Germond, A.; Ichimura, T.; Horinouchi, T.; Fujita, H.; Furusawa, C.; Watanabe, T.M. Raman spectral signature reflects transcriptomic features of antibiotic resistance in Escherichia coli. *Commun. Biol.* **2018**, *1*.

100. Xuan Nguyen, N.T.; Sarter, S.; Hai Nguyen, N.; Daniel, P. Detection of molecular changes induced by antibiotics in Escherichia coli using vibrational spectroscopy. *Spectrochim. Acta - Part A Mol. Biomol. Spectrosc.* **2017**, *183*, 395–401.

101. Huleihel, M.; Pavlov, V.; Erukhimovitch, V. The use of FTIR microscopy for the evaluation of antibacterial agents activity. *J. Photochem. Photobiol. B.* **2009**, *96*, 17–23.

102. Moen, B.; Janbu, A.O.; Langsrud, S.; Langsrud, Ø.; Hobman, J.L.; Constantinidou, C.; Kohler, A.; Rudi, K. Global responses of *Escherichia coli* to adverse conditions determined by microarrays and FT-IR spectroscopy. *Can. J. Microbiol.* **2009**, *55*, 714–728.

103. Corte, L.; Rellini, P.; Roscini, L.; Fatichenti, F.; Cardinali, G. Development of a novel, FTIR (Fourier transform infrared spectroscopy) based, yeast bioassay for toxicity testing and stress response study. *Anal. Chim. Acta* **2010**, *659*, 258–265.

104. Ribeiro da Cunha, B.; Fonseca, L.P.; Calado, C.R.C. Metabolic fingerprinting with fourier-transform infrared (FTIR) spectroscopy: Towards a high-throughput screening assay for antibiotic discovery and mechanism-of-action elucidation. *Metabolites* **2020**, *10*.

105. Ribeiro da Cunha, B.; Fonseca, L.P.; Calado, C.R.C. Simultaneous elucidation of antibiotic mechanism of action and potency with high-throughput Fourier-transform infrared (FTIR) spectroscopy and machine learning. *Appl. Microbiol. Biotechnol.* **2021**, 1–18.

Chapter IV

Fundamentals of Fourier-transform mid-infrared spectroscopy

This chapter is adapted from the book chapter:

Ribeiro da Cunha, B., Ramalhete, L., Fonseca, L. P., & Calado, C. R. C. (2020). Fourier-Transform Mid-Infrared (FT-MIR) Spectroscopy in Biomedicine. In Y. Tutar (Ed.), Essential Techniques for Medical and Life Scientists: A Guide to Contemporary Methods and Current Applications- Part II (pp. 1–39). doi: 10.2174/9789811464867120010004.

Author contribution

Bernardo Ribeiro da Cunha contributed to the literature review, to the preparation of the original draft, reviewed, and edited its final version.

Abstract

Fourier-transform infrared (FTIR) spectroscopy (FTIRS) is a powerful technique that probes the intramolecular vibrations of most molecules, thereby enabling the acquisition of metabolic fingerprint of cells, tissues and biofluids (e.g., serum, urine and saliva, etc.), in a rapid (minutes), simple (without or with minimal sample processing), economic (without reagents consumption), label-free and highly sensitive and specific mode. FTIRS is an analytically flexible technique with diverse sampling techniques and detection modes. These include classical transmission and transflection, high-throughput measurements using micro-plates in transmission mode, and fiber optic probes coupled to attenuated total reflection detection. As such, its applications range from *in situ* analysis trough to micro-spectroscopy with spatial resolution at the sub-cellular level. Due to the complex biochemical composition of biological samples, mid-infrared spectra are usually very difficult to interpret without the application of complex and sophisticated mathematical and statistical analysis routines, such as: spectra preprocessing, to minimize noise and other non-informative data that compromise subsequent pattern recognition models; deconvolution methods to resolve overlapped spectral bands; dimensionality reduction and feature extraction; along with supervised and non-supervised pattern recognition strategies, as support vector machines and artificial neural networks. The present work revises the fundamentals of FTIRS, its main acquisition modes, preprocessing and the most common multivariate spectral analysis.

Contents

IV.1 INTRODUCTION

Spectroscopy is the study of the interaction between matter and radiated energy. In the case said energy is within the infrared (IR) region of the electromagnetic spectrum, the term IR spectroscopy applies. When dealing with solid and liquid samples, IR spectroscopy measures the vibrational modes of molecular bonds that result from dipole moment changes, i.e., charge differences in the electric field of atoms. With exception of monoatomic (e.g., He, Ne) or homopolar diatomic molecules (e.g., H_2, N_2, O_2), almost all molecules present a unique IR spectrum. Given the complex molecular composition of cells, biofluids (e.g., serum, blood, saliva, urine, spinal fluid), and other materials (e.g., tissue, calculi, feces, cartilage, bone), IR spectra of complex biological samples are also distinctive. As such, IR spectroscopy, especially at the mid-infrared (MIR) region of the electromagnetic spectrum, acquires a holistic molecular fingerprint that can be associated with the metabolic state in a highly sensitive and specific mode.

IR spectrometers are either based on dispersive equipment, which yields a continuous wavelength, or interferometers, which result in a time domain signal called an interferogram. Dispersive spectrometers rely on a prism, or a more sophisticated grating, that separates the individual frequencies of energy emitted by the infrared source. On the other hand, most interferometric spectrometers use a Michelson interferometer, which includes a beam splitter, a fixed and a moving mirror. The split IR beam is brought out of phase by increasing the length of one beam path using the moving mirror, and then recombined by carrying the interference pattern. Thus, interferometric spectroscopy is a multiplex technology as all frequencies are observed simultaneously over the scanning period. By applying a Fourier transform, it is possible to switch the signal from the time domain to the frequency domain, thereby producing a single beam spectrum.

Fourier-transform infrared (FTIR) spectroscopy (FTIRS), or interferometric spectroscopy, has intrinsic properties in relation to dispersive spectroscopy [1,2], in particular:

- All frequencies are measured simultaneously, and the same environment is used to acquire all frequencies. So, noise is reduced along with the time required for spectra acquisition, which is in the range of seconds rather than minutes. This is known as the multiplex or Felgett advantage;

- Fast scanning enables the co-addition of several scans, further reducing noise. Thus, FTIR spectra are more sensitive, which is known as Jacquinot advantage.
- An internal HeNe laser is used to calibrate the interference information, which results in high wavenumber accuracy, high reproducibility, and high precision. This is known as the Connes advantage.

Since the first Michelson interferometer was assembled in 1881, various key events took place which consolidated FTIRS as a mature analytical technique. Namely the ability to digitalize the interferogram, the capacity to rapidly apply the Fourier-transform, along with various computational tools required to interpret the data, are just a few examples.

IV.2 General characteristics of mid-infrared spectroscopy

Biological samples present a highly complex molecular composition, ranging from small inorganic and organic molecules, through to macromolecules, as nucleic acids, polysaccharides, and lipids. Since the MIR region of the electromagnetic spectrum (2.5-25 µm or 400-4000 cm^{-1}) reflects fundamental vibrations, and the near-infrared (NIR) region (780-2500 nm or 4000-12821 cm^{-1}) reflects overtones and combinations of vibrations, acquisition in the MIR region gives rise to stronger and better-defined absorbance bands compared to the NIR region, which typically results in weaker and wider spectra. Furthermore, MIR spectra of biological samples are sensitive to biomolecules with functional groups such as C-C, C=C, C-O, C-N, C- H, O-O, O-P, N-H and O-H bonds, while NIR covers groups exclusively containing the hydrogen atom as C-H, N-H, O-H, and S-H bonds.

Consequently, MIR spectra are more informative about biomolecular composition, particularly in the following spectral regions: 3600-2000 cm^{-1}, reflecting mainly stretching vibrations between X-H (where X is C, O or N) present in amide A and amide B (~3300 and 3100 cm^{-1}, respectively), and CH_3 (~2960 and ~2872 cm^{-1}) and CH_2 (~2920 and 2850 cm^{-1}) groups of lipids; 1800–1500 cm^{-1}, reflecting mainly stretching vibrations of double bonds (e.g. C=O, C=C and C=N), present in amide I (~1655 cm^{-1}) and amide II (~1545 cm^{-1}) of proteins and some secondary structure of proteins, and COOR in phospholipids esters (~1740 cm^{-1}); and 1500-400 cm^{-1}, reflecting overlapped vibrations due to proteins, lipids, and nucleic acids, designated as fingerprint region [3,4]. Figure 1 exemplifies MIR spectra obtained from human serum and gastric cells, and Table 1 points out the biochemical assignment of major spectral bands.

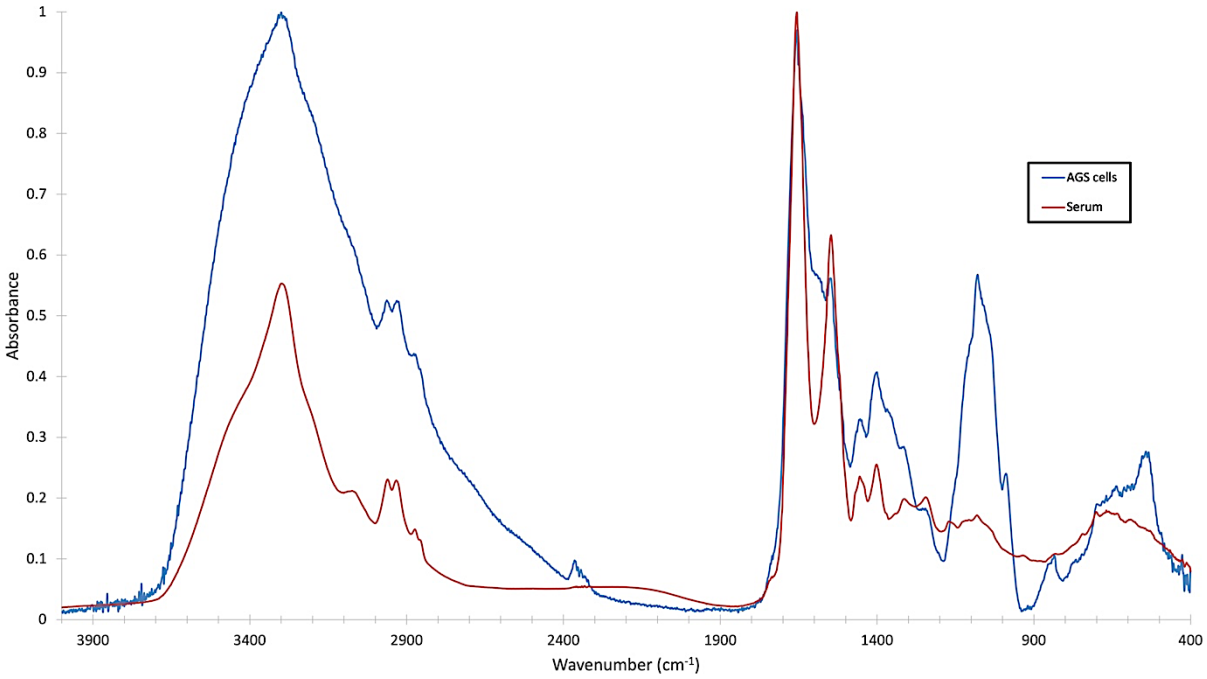


Figure 1. Mid-infrared spectrum of human serum and adenocarcinoma gastric (AGS) cells.

IV.2.1 SENSITIVITY BEYOND MOLECULAR COMPOSITION

MIR spectra, in addition to being informative about molecular composition, also reflect conformation changes of biomolecules, as protein folding, via vibrational resonance originating from polypeptide backbones or side chains that depend on the protein structure and local interactions such as hydrogen bridges [5,6]. In addition, MIR spectra are also informative on the conformation of nucleic acids [7–9], biomembrane organization, fluidity and even other biomolecular interactions [10–12]. Therefore, MIR spectra provide metabolic fingerprints of biological samples with high sensitivity and specificity.

Since MIR spectroscopy (MIRS) is sensitive to a variety of biological features, it has been used to monitor biological processes as cell division, differentiation, apoptosis, necrosis, disease progression, prognosis, diagnosis, and even treatment follow-up. An example of the versatility of this spectroscopic technique is illustrated by its ability to not only distinguish adenocarcinoma gastric cells infected with *Helicobacter pylori* from non-infected cells (Figure 2 – A), but also cluster infected cells according to the specific strains that infect them (Figure 2 – B). These *H. pylori* strains are major etiological agents of different gastric diseases, such as ulcers and cancer.

Table 1. Vibrational frequencies of some functional groups commonly assigned to biomolecules (adapted from Bellisola et al. [13] and Sales et al. [4]).

Wavenumber (cm^{-1})	Vibrational mode on the functional group	Commonly assigned biochemical component
3300	υ(N-H)	Amide A: peptide, proteins
3100	υ(N-H)	Amide B: peptide, proteins
2960	$\upsilon_{as}(CH_3)$	Lipids
2920	$\upsilon_{as}(CH_2)$	Lipids
2870	$\upsilon_{s}(CH_3)$	Lipids
2850	$\upsilon_{s}(CH_2)$	Lipids
1740	υ(C=O))	Phospholipid esters
1655	[80% υ(CO) and 20% υ(CN)] in O=C-NH	Amide I in peptides and proteins
1550	υ(C-C) in O=C-NH	Amide II in peptides and proteins
1450	$\upsilon_{as}(CH_3)$	Lipid, proteins
1395	$\upsilon_{s}(CH_3)$	Lipid, proteins
1380	$\upsilon_{s}(CH_3)$	Phospholipid, fatty acid, triglyceride
1240	$\upsilon_{s}(PO_2^-)$	Phosphodiesters in nucleic acids, phospholipids and phosphorylated proteins
1160 and 1111	υ_{s}(C-O)	Ribose in RNA
1078	$\upsilon_{s}(PO_2^-)$	Phosphodiesters in phosphate energetic level
1055	υ_{s}(C-O-P)	Phosphate esters
1032	def(C-OH)	Glycogen
965	$\upsilon(PO_3^{2-})$	DNA and RNA
950	$\upsilon(PO_3^{2-})$	Phosphorylated proteins
920	υ(COP)	Phosphorylated proteins

*, vibrations type: υ_s, symmetric stretching; υ_{as}, asymmetric stretching; def, deformations.

IV.2.2 TYPES OF SAMPLES AND ACQUISITION MODES

A great advantage of MIRS is its high versatility. MIRS allows to investigate samples in either solid, liquid, or gaseous phase, therefore samples can take a diversity of forms such as: biofluids, solutions, cells, fixed cells or tissues, biopolymers, pastes, powders, tablets, films, fibers, coatings and surfaces. Accordingly, there are numerous acquisition modes of FTIR spectra, including but not limited to: transflectance, a combination between transmission and

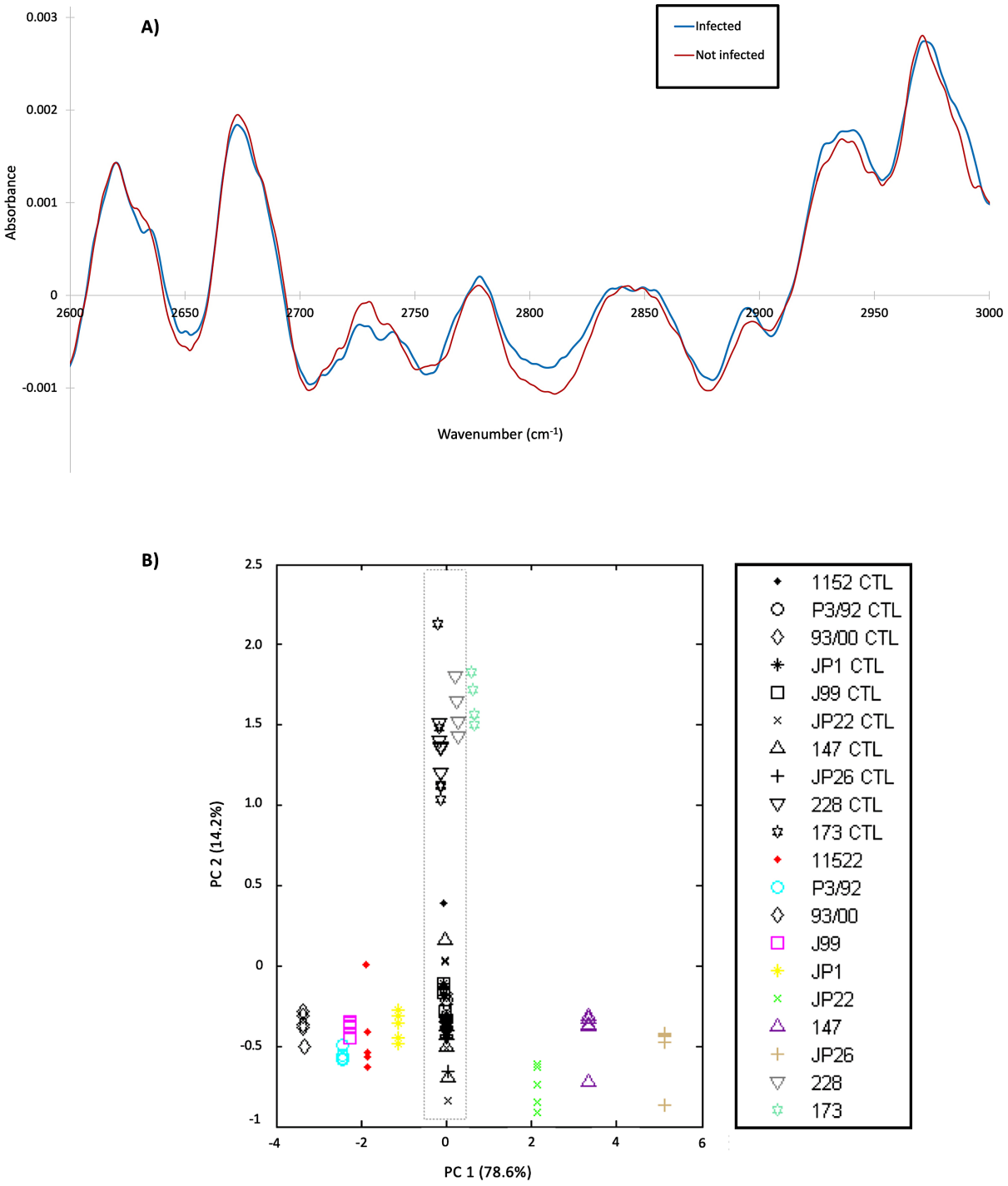


Figure 2. Second derivative of mid-infrared (MIR) spectra of adenocarcinoma gastric human (AGS) cells infected, or not, with *Helicobacter pylori* (A). A spectral region that highlights the differences between infected and non-infected gastric cells was chosen for demonstrative purposes. A principal component analysis of MIR spectra of AGS cells infected with different strains of *H. pylori* distinguished not only non-infected cells (represented inside the doted square) from infected cells, but also separated those infected by more virulent bacterial strains (represented by different symbols), which are associated with more severe gastric pathologies (B).

reflectance, using e.g., calcium or barium fluoride slides; transmission, using e.g., microplates of zinc selenium that enable high-throughput readings; attenuated total reflectance (ATR), which can be coupled to fiber optic probes for *in situ* analysis; and microscopic analysis, using focal plane array (FPA) detectors, which yield 2D spectra in a defined localization of the sample. Independently of the detection mode, it is possible to acquire FTIR spectra in the MIR or NIR region in a rapid (minutes), economic (without reagents), label-free, with minimal or no sample handling, and in the case of MIR, in a highly specific and sensitive mode.

Transmission spectroscopy is the oldest and most straightforward method [14]. To obtain high-quality spectra of biological specimens, samples need to be dehydrated e.g., directly on a multi-well micro-plate as implemented by Sales et al. [4], Rosa et al. [15], and throughout this Ph.D. thesis. To enhance sensitivity, signal amplification can be achieved with the surface enhanced infrared absorption technique, which is analogous to surface enhanced Raman scattering in Raman scattering spectroscopy (RSS). Plasmonic chip-based technology allows *in situ* analysis of aqueous media (reviewed by Baker et al. [16]). An alternative detection mode involves placing the sample on a reflective surface, such as a substrate coated with aluminum/Teflon or on a glass slide coated with reflective tin oxide-based silver. In these cases, a transflectance mode is obtained as the MIR beam passes through the sample: it is reflected off the slide and passes again through the sample (reviewed by Ataka et al. [17]).

ATR is another detection mode noteworthy for biological samples. Here, the MIR beam is propagated through a high refractive index crystal surface, which produces an evanescent wave that penetrates a few microns into the sample. The MIR light is transmitted through the sample, reflects on the top layer, and passes again through the sample, increasing path length and consequently increasing sensitivity. The ATR detection mode can also be coupled to fiber optic cables, thereby enabling *in situ* analysis [18]. An important disadvantage of this system stems from the characteristics of MIR light, which limits the fiber optic cable length to a few meters, whereas NIR light can propagate fiber optic cables across larger distances [11].

The use of micro-spectroscopy facilitates analysis of micro samples and low concentration compounds (e.g., residual substances). Microscopic analysis in transmission mode is usually conducted on 4-10 µm thick dehydrated histological samples, which are placed on an IR transparent material such as barium or calcium fluoride slides. Another method of detection in transflectance mode uses e.g., slides based on Ag/SnO_2. An alternative ATR detection mode can also be applied to analyze aqueous biological samples.

It is also possible to conduct MIRS imaging based on micro-spectroscopy. In this case, the high spatial resolution obtained enables analysis at a sub-cellular level, which is highly relevant in histopathological diagnosis, as represented on Figure 3. MIR micro-spectroscopy imaging can be achieved in transmission or transflectance mode and is based on scanning with either a single element detector, linear arrays, or FPA detectors. Linear and 2D FPA detectors (e.g., arrays of 64x64 or 128x128 detectors) enable a faster analysis, yielding tens to thousands of spectra in just a few minutes. In any case, a few areas within a sample, e.g., tissue slides from biopsies, are usually selected for analysis, as one analysis results in 16,384 spectra. FPA technology is currently the state-of-the-art for MIR micro-spectroscopy imaging, and these enable the analysis of complete histological samples, thus minimizing sampling bias and therefore minimizing the issue of missing diagnostically important areas. Bassan et al. [18] acquired 66 million spectra in 14 hours, each representing 5.5x5.5μm^2, of a prostate cross section using an FPA with 128x128 detectors.

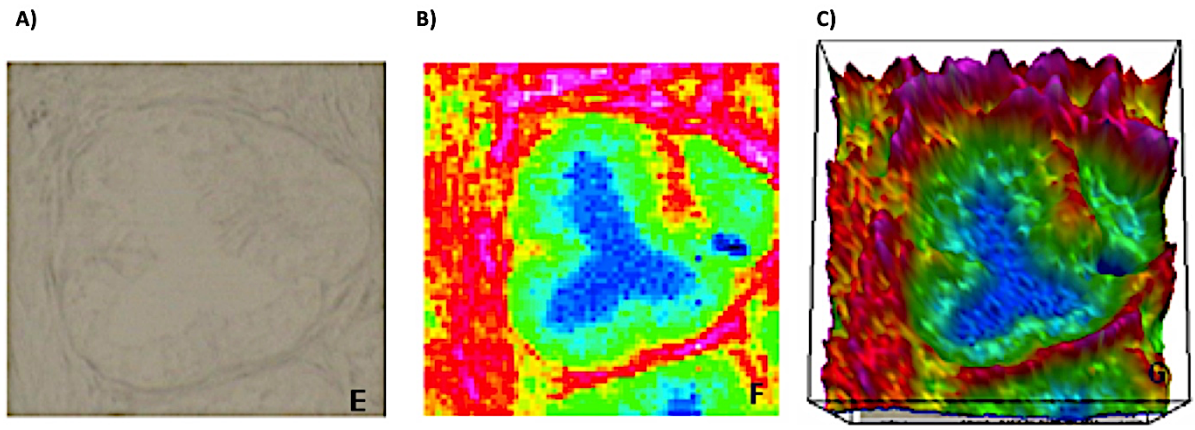


Figure 3. Histological analysis of a prostate sample tissue by optical microscopy (A) and by Fourier-transform micro-spectroscopy using a 15x objective and a 64x64 FPA detector (B), which reveals the chemical composition upon integration of the spectral region between 3000 to 3600cm^{-1} (C).

A conceivable disadvantage of transflectance over transmission mode in FTIR micro-spectroscopy imaging is the so-called electric field standing wave effect. Pilling et al. [19] observed non-linear variations in absorption band strength across spectra of prostate samples, obtained with a 15x objective and a 128x128 FPA in transflectance mode. This resulted in lower discrimination power between benign and cancerous tissue, when compared with a transmission detection mode. Therefore, when performing medical diagnosis with MIR micro-spectroscopy of histologic samples, it is advisable to use a transmission rather than a transflectance detection mode. This translates to using more expensive and fragile IR

transparent slides, as barium fluoride, instead of cheaper transflection substrates such as Ag/SnO_2.

More information concerning the recent technological evolutions of MIR spectra acquisition can be found in the following articles: Ataka et al. [17], Bunaciu et al. [20,21], Baker et al. [22], Tiwari et al. [23] and Mitchell et al. [24]. For example, it is now possible to achieve lateral resolutions of 50 nm, temporal monitoring of biological reactions at the femtosecond scale, and even evaluate the effect of membrane potential on a single proton transfer [17].

Considering the characteristics of MIRS described so far, some of which are especially advantageous in the context of this Ph.D. thesis, this section focused on IR spectroscopy in the MIR region. Even so, a brief discussion of other vibrational spectroscopy techniques is presented given these are not only relevant for many biomedical applications, but often complementary to MIRS.

IV.2.3 Near-Infrared spectroscopy

Since NIR spectra are characterized by overtones and combinations of fundamental vibrations, they are usually less intense and have broader bands in relation to MIR spectra. Furthermore, only bonds with hydrogen are detected, but this does not imply that NIR spectroscopy (NIRS) lacks specific advantages regarding MIRS. For example, support materials used in NIRS, such as glass and quartz, are much cheaper than those used in MIRS. Water is also mostly transparent in the NIR region, whereas MIRS requires sample dehydration in transmission mode or ATR detection mode.

Due to the high energetic frequencies of NIR radiation, and its water-like transparency, it can penetrate several millimeters into human tissue, while MIR light can only penetrate around 100 μm. Moreover, NIRS is sensitive to the differences that arise in defined biomolecules, such as hemoglobin and lipids, between normal and pathogenic tissues, which enables its application as a non-invasive or minimally invasive monitoring technique. For instance, given its sensitivity to the oxygenation state of hemoglobin, NIRS has been investigated for the non-invasive monitoring of tissue oxygenation. Similarly, using fiber optic probes, NIRS can be used on body surfaces, or other minimally invasive examinations, for real-time diagnosis of diverse pathologies like breast, skin, prostate, pancreas and colorectal cancer, metabolic myopathy, valvular heart diseases, peripheral artery diseases,

neuromuscular diseases, chronic obstructive pulmonary disease and type 1 or type 2 diabetes, among others (reviewed by Kondepati et al. [25] and Sakudo [26]). NIRS has also been used during surgeries to identify cancerous tissue in real-time [27].

Molecular probes have been developed for NIRS, to serve as enhancing agents in contrast imaging applications. This further enables the use of non-invasive NIR imaging, i.e., using an acquisition mode external to the organism, to acquire internal signals (reviewed by Frangioni [28] and Hilderbrand et al. [29]). For example, Frangioni [30] developed NIR fluorescent quantum dots to localize sentinel lymph nodes during cancer surgery. Other recent advances of NIRS probes guiding imaging have been reviewed by Namikawa et al. [31].

Similarly, NIRS has been developed for non-invasive and functional neuroimaging studies since the 1990s. In this case, dynamic changes in cerebral blood flow are monitored, which increase at sites with activated neural function. The NIR radiation detected at the surface of the head results from scattered radiation of an incident NIR beam, which passes through skin and bone, reaching a depth of approximately 20 mm, thereby probing the activity of the cerebral cortex. This technique not only enables high temporal resolution (less than 1s), but also allows patients to move during measurements using portable and wireless equipment (reviewed by Hoshi [32] and Sakudo [33]).

In addition, NIRS is relatively inexpensive compared to other functional imaging techniques such as positron emission tomography, single-photon emission computerized tomography or magnetic resonance imaging. Similarly, functional NIR neuroimaging provides insight into the neurobiological factors of hearing that are relevant after cochlear implantation, with clear advantages over traditional methods as it is non-invasive and not subject to electrical artifacts [34].

The non-invasive analysis of clinically relevant analytes in diabetes with NIRS still presents considerable constraints, e.g., measuring glucose depends on a calibration that is specific to the patient's tissue characteristics. Therefore, for this application, most studies could only measure relative analyte variations, but that is not enough when quantitative analysis are needed for clinical purposes (reviewed by Sakudo [33]). Indeed, despite intensive research, a reliable non-invasive glucose sensor does not exist in clinical practice [35]. Kottmann et al. [36] proposed a method for the non-invasive detection of glucose, which applies MIRS to the interstitial fluid of epidermis, which had a detection limit within the physiological range.

NIR analysis of defined metabolites can be conducted directly from body biofluids like urine and serum. In these cases, a multivariate regression method, e.g., Partial Least Squares Regression (PLSR), is usually applied and good predictive models are achieved with low errors and high correlation coefficients (as reviewed by Sakudo [33]). However, it is also possible to conduct MIRS in real-time, which presents the advantage of having more informative spectra in terms of molecular composition. This does not mean specific analytes, e.g., glucose in serum, cannot be adequately predicted with NIRS. For example, NIR analysis of urine samples resulted in predictive models of urea and protein with an equivalent accuracy in relation to the conventional assays, in contrast creatinine presented a lower accuracy [37].

Nevertheless, there are situations where NIRS cannot achieve good predictive models or simply MIRS produces more robust models. For example, Liu et al. [38] developed NIR models from serum samples that were only capable of predicting low-density lipoprotein cholesterol. In turn, Liu et al. [39] developed similar MIR models that could predict both low-density and high-density lipoprotein cholesterol. Likewise, Sandor et al. [40] observed a 50% error reduction of PLSR models to predict glucose, lactate, glutamine, glutamate and ammonia in animal cell cultures, when these were derived from MIRS in comparison to those derived from NIRS.

Even in cases where a NIR set-up resulted in slightly better predictive models of analytes in cell cultures, MIR spectra conveyed further biochemical and metabolic information, thus promoting cumulative knowledge of the biological process towards improved prognostics, diagnostics and follow-up treatments. As an example, Sales et al. [3,4] developed similar models to predict extracellular glucose and intracellular plasmid quantities both with NIR and MIR spectra, but obtained supplemental information that enabled the characterization of the cell's biochemical and metabolic activity with MIRS.

Therefore, if the main goal is to build predictive models for a defined metabolite using biofluids (as urine, saliva, serum, blood, sputum, semen, etc.), the method of choice should be MIRS over NIRS, as MIR spectra yield more robust models that are typically not achievable with a NIR setup. Additionally, MIRS simultaneously and concomitantly provides a metabolic fingerprint of the cell, tissue or organism, which reveals supplemental information that of could be of added value. Furthermore, the high diversity of MIRS set-ups overcomes, to some extent, the advantages of NIRS. For instance, there are alternatives regarding sample preparation that

require no or minimal processing, as well as detection modes that include real-time spectral acquisition, high-throughput assays, and microscopic analysis.

IV.2.4 RAMAN SCATTERING SPECTROSCOPY

RSS is a vibrational spectroscopy technique complementary to FTIRS, since it probes electric polarizability changes. In other words, it probes changes in the electron density distribution of a molecule. Because of minimal interference by water, RSS has great potential in biomedical sciences. RS corresponds to the inelastic light scattering process of photons following interaction with a monochromatic radiation source. A high proportion of photons that interact with matter are elastically scattered with no change in their energy. A minor percentage of the photons, however, transfer their energy as vibrational energy, resulting in scattered photons with lower energy levels – known as Raman-Stokes scattering. Anti-Stokes scattering occurs when incident photons receive energy from vibrating molecules, therefore presenting higher energetic levels.

Given the low probability of a molecule undergoing Raman state transition, a higher concentration of the analyte is required to produce a detectable RS signal in relation to a MIR band. Furthermore, many biological samples undergo fluorescence when submitted to wavelengths used in RSS, which decreases the intensity and sensitivity of the RS signal [41]. For example, prediction models of glucose, lactate and urea mixtures from human cell cultures based on NIR spectra outperformed equivalent regression models derived from RS spectra [42]. There are diverse techniques to overcome these limitations, which work by enhancing the RS signal, such as surface enhanced Raman scattering [43], which enable the detection of residual molecules. Similar work has been conducted to enhance MIR signals [44].

Both RSS and MIRS can be applied in a microscopy detection set-up. This enables not only the detection of residual compounds, but also spectral acquisition in a defined space that can be used for sub-cellular and tissue imaging. The advantage of RS over MIR micro-spectroscopy is its low water interference, along its high spatial resolution, which is because the excitation wavelengths fall in the visible and near-infrared range. On the other hand, the use of an intense laser results in local thermal decomposition of the sample [45]. Examples of RSS reviews for biomedical sciences were carried out by Bunaciu et al. [20], Butler et al. [46], Eberhardt et al. [47], Kong et al. [48] and, Baker et al. [22].

Another key limitation of RSS is its poor repeatability, which hinders reliable quantitative analysis [49]. Although efforts are ongoing to overcome said limitations, as well as to increase the throughput of this technique [50,51], FTIRS is a more mature technique, considered by most the gold standard of molecular spectroscopy, for which high-throughput detections protocols are well-established, well-understood, and considerably cheaper. For these reasons, the focus of this chapter, and this Ph.D. in general, has been on MIRS specifically. From herein, the data analysis discussion is directed at the MIRS subset, although overlaps with NIRS are commonplace. Although the great potential of RSS spectroscopy, and its complementary nature to FTIRS justified its discussion, it did not warrant its application in the context of this research. Even so, it is noteworthy that reviews of preprocessing and analysis algorithms commonly applied in RSS have been discussed elsewhere [52].

IV.3 CHEMOMETRICS AND DATA ANALYSIS

A FTIR spectra of a single sample, acquired at a 2 cm^{-1} resolution, from 400 to 4000 cm^{-1}, includes over 1800 absorbance values. Due to the relatively high dimension of FTIR spectra, alongside the intrinsic variability of biological systems, multivariate analysis is a key aspect of FTIRS. There are also various studies based on univariate analysis, i.e., looking at spectral bands or ratios to retrieve information concerning biomolecules such as lipids, proteins, DNA, glycogen, phosphate levels. In these cases, the deconvolution of overlapped spectral bands along spectral derivatives plays a key role. Also, the metabolic status, such as apparent translational levels or turn-over metabolism, have also been probed with univariate analysis. However, most studies, due to the complexity of FTIR spectra, apply sophisticated multivariate processing methods.

Most of the techniques to be discussed originated from chemometrics, a field that ‘spun-off’ from the branch of analytical chemistry in the mid-1970s greatly due to work of Kowalski and Wold [53]. More recently, machine learning techniques have been increasingly applied in order to tackle more complex biological phenomena; as such, these advanced algorithms are now an integral component of infrared spectroscopy. In conjunction with increasing computational power, these techniques enable the extraction of information from much larger volumes of denser data, i.e., data with more samples and more variables [54]. In order to achieve robust models, a FTIRS workflow should include: quality control routines, to ensure high-quality spectral libraries; preprocessing procedures, to highlight sample-specific

spectral components; dimensionality reduction methods, to facilitate data visualization and improve subsequent analysis; and, chemometrics and/or machine learning algorithms.

IV.3.1 QUALITY TESTS AND SPECTRAL PREPROCESSING

An important first step in FTIRS is to appreciate the quality of the raw spectra, for which three parameters are typically referred: sample thickness, in transmission mode the integrated absorption values can be considered as an estimate for sample thickness that should be within a given range; signal-to-noise ratio, which measures both the signal at Amide I region (1620-1690 cm^{-1}) and the noise in the signal-free region (1800-1900 cm^{-1}), further excluding spectra below a minimum value; and, water content, which is detrimental in FTIRS, and can be probed at the atmospheric water vapor region (1750-1900 cm^{-1}) of the second derivative spectra [55].

To maximize data extraction from the FTIR spectra, preprocessing methods should be explored with the goal of e.g., removing baseline offsets and minimizing physical effects as light scattering scatter. Examples of preprocessing methods are baseline-correction, normalization, multiplicative scatter correction, derivative filtering and combinations thereof. For better understanding and visualization of the workflow, the following discussion is accompanied by examples of an actual FTIRS dataset, obtained from a study based on *Escherichia coli* exposed to rifampicin and nanoencapsulated rifampicin.

The objective behind spectral preprocessing is to achieve numerically what could not be accomplished physically during data acquisition, in other words to ‘enforce’ intra-replica homogeneity and reduce the non-discriminatory sample-specific component of the spectral signal (for simplicity named noise), and simultaneously highlight inter-replica variation, which contains information related to biological variations under study [56].

The most commonly used preprocessing methods include: Baseline correction, which aims to reduce baseline and slope distortions inherent to complex spectra acquisition; Multiplicative scattering correction, which minimizes scatter effects due to particles of different sizes and shapes obtained during the dehydration process; Normalization procedures, to scale spectra within a similar range to compensate differences in the sample’s total biomass and varying optical pathlengths, and so reduce differences in spectral intensity while highlighting differences in spectral patterns; Derivatives, so to enhance the informative component of the data by increasing the resolution of overlapping bands, while minimizing physical interferences.

Determining spectral derivatives has the downside of reducing signal scale and amplifying the noise, however the latter can be minimized if the Savitzky-Golay filter is used [15]. The effects of some of these preprocessing techniques with raw spectra are compared in Figure 4. Other commonly used preprocessing algorithms have been described elsewhere [56].

IV.3.2 Univariate analysis of FTIR spectra

An alternative to multivariate analysis of FTIR spectra is univariate analysis of specific spectral bands and/or ratios. In this type of analysis, one approach is to apply various tests to evaluate the statistical relationship of spectral bands, or ratios, to a given metabolic or phenotypic characteristic, which may reveal that the former can be used as biomarkers of the latter. Independently from the statistical test applied, and the type of spectral feature, it is desirable to apply preprocessing techniques beforehand.

An example of a statistical workflow for biomarker screening has by described by Marques et al. [57], and more recently we have developed an automated workflow that applies the most suitable hypothesis tests to distinguish two or more populations, in mean terms [58]. This enables an extremely efficient investigation of a large number of putative biomarkers; ensures the most powerful test is always applied, if its assumptions are not violated; and, accepts a range of samples and number of observations within.

Alternatively, univariate analysis can be used to associate spectral features with biochemical or metabolic characteristics, such as the levels of RNA, glycogen, phosphate level, CH_3/CH_2 ratios, amide I β-sheet and α-helix turnover metabolism ratios. In these cases, as described by Sales et al. [4], it is often possible and beneficial to deconvolute overlapped spectral bands. Here, overlapping bands composing a single spectral band can be identified with the peaks revealed by the second derivative spectrum, and these can then be resolved into their underlying biochemical contributions.

For example, the second derivative of the spectral band between 1000 and 1150 cm^{-1} reveals five different peaks (Figure 5 – A), which in turn suggests that this apparent single band of the original spectrum (Figure 5 – B, highlighted in bold), is actually composed of five overlapping bands. After spectral deconvolution using Lorentzian shape curve fitting, considering the number of subjacent peaks identified on the second derivative, five bands whose peak is at the location identified on the second derivative, were fitted to the original

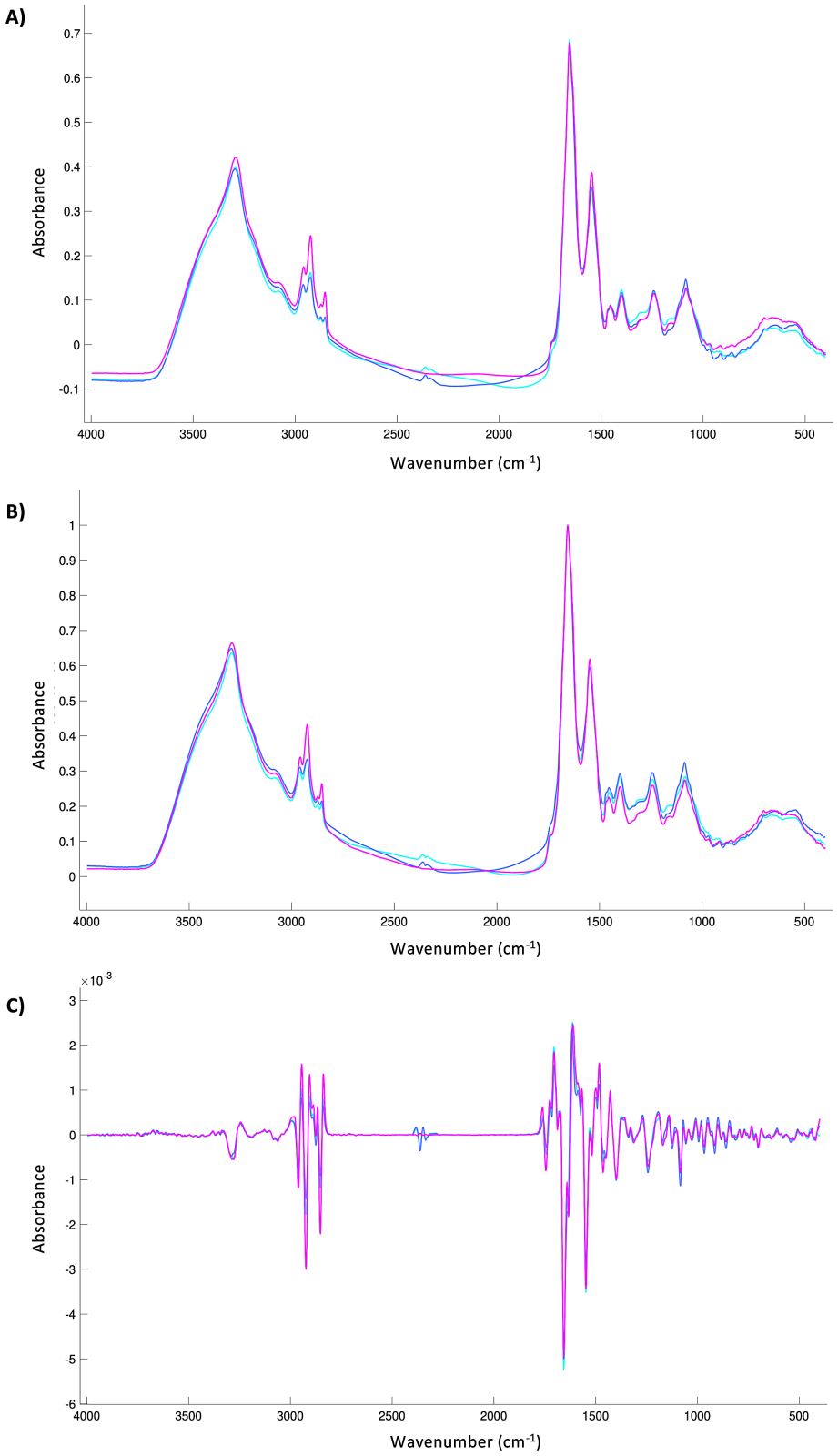


Figure 4. Fourier-transform infrared spectra acquired from *Escherichia coli* cells prior to exposure (—), and after exposure to rifampicin (—) and nanoencapsulated rifampicin (—) without preprocessing (A), after baseline correction followed by normalization at the Amide I band (B), and after the Savitzky-Golay filter was used to determine the second derivative spectra (C).

spectral region (Figure 5 – B). These bands, with peaks at 1032 cm^{-1}, 1055 cm^{-1}, 1078 cm^{-1} and 1111cm^{-1} correspond to C-OH in glycogen molecules, to C-O-P in phosphate esters, to PO^- in phosphodiesters, and to C-O in RNA molecules, respectively. As such, if the objective was to monitor RNA levels, the resolved band at 1111cm^{-1} would be better suited than the original band, whose peak was around 1080cm^{-1}, and its intensity influenced by various biomolecules.

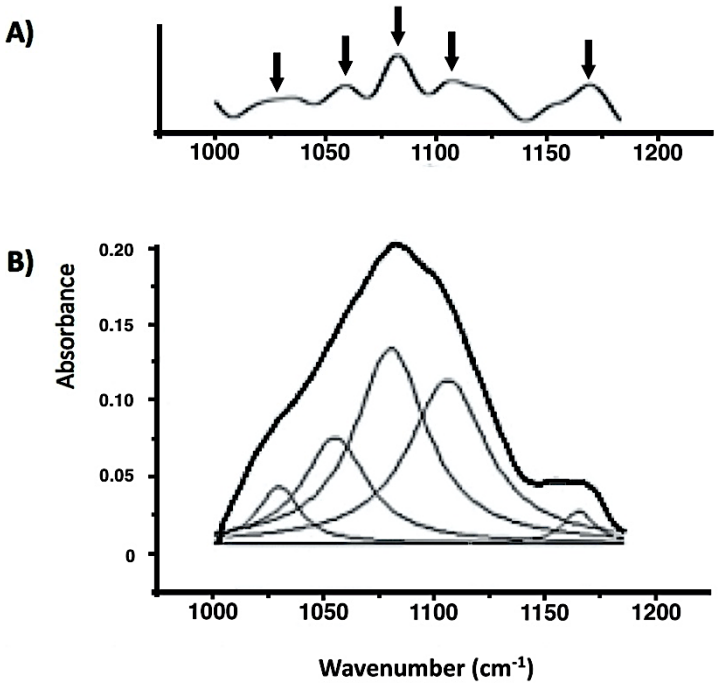


Figure 5. Second derivative of Fourier-transform infrared spectra of *Escherichia coli* cells, obtained in transmission mode in the region between 1000 to 1150 cm^{-1}, where arrows highlight spectral peaks (A). Deconvolution of the original band, highlighted in bold, with Lorentzian shape curve fitting into its five underlying bands whose peak align with those identified on the second derivative spectra (B).

IV.3.3 PRINCIPAL COMPONENT ANALYSIS

Principal Component Analysis (PCA) is the swiss army knife of chemometrics. One of its most common applications is to reduce a set of multivariate measurements to a smaller set of components with less redundancy, which are expected to maintain the original data structure [59]. PCA converts a set of observations and (possibly) correlated variables from the original dataset into a new set of 'projected' observations (scores), on new uncorrelated variables called Principal Components (PCs). Because the number of PCs is less than or equal to the

number of original variables, and few PCs can be used to reconstruct the original signal, PCA reduces data dimension.

By definition, the first PC models as much data variability as possible and succeeding PCs are ranked by decreasing variance, given the constraint that they are orthogonal (i.e., uncorrelated) to preceding PCs. PCs are linear combinations of the original variables, thus the various scores on the PCs space can be analyzed as would the original spectra [60]. PCA interpretation commonly relies on score plots, exemplified on Figure 6, which serves to visualize the relationship between the projected samples on a reduced dimensionality plot.

Projected samples, or PCA scores, can be used as a precursor of further analyses. A straightforward example of which is performing an unsupervised Hierarchical Cluster Analysis (HCA) on PCA scores to identify intrinsic clusters and relate these to a biological phenomenon. Alternatively, it is possible to perform a Linear or Quadratic Discriminant Analysis (LDA and QDA, respectively). Either method determines linear expressions that model descriptors against variables, while maximizing between-class variability in relation to the pooled within-group variability [59]. In other words, these are parametric methods that select directions which maximize class separation using a distance metric [60]. To do so, it is necessary to ‘introduce’ external data on original dataset, thus the term ‘supervised’. In practice, either LDA or QDA quantitatively describe ‘boundaries’ between classes, as seen in Figure 7.

Therefore, it is possible to predict class-membership for unknown samples and inclusively determine the robustness of the predictive model with a multitude of statistical validation tools, which is an integral part of the multivariate model building workflow. It is important to note, however, that PCA must be applied to the complete dataset, and new predictions require a re-calibration of the projected samples on new set of PCs.

In a broader scope of dimensionality reduction, PCA belongs to a class of methods named feature construction. Within these methods, examples include Partial Least Squares (PLS), band fitting or peak picking, among others. In addition to building new features of reduced dimension, it is also possible to select the most relevant features of the original dataset. In this case, PCA can also be applied, but other methods include forward feature selection, genetic algorithm or random forests [61].

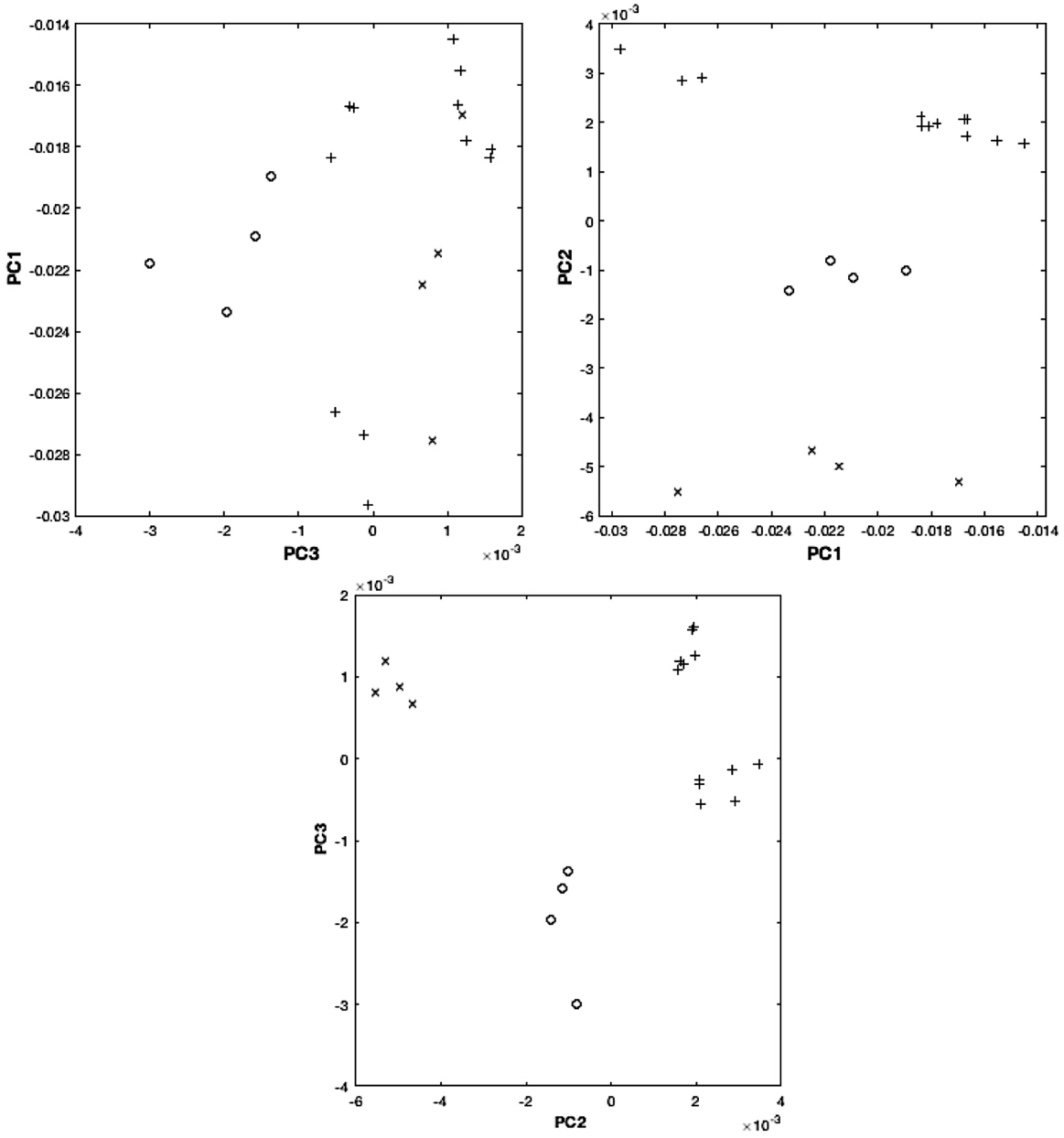


Figure 6. Score plots of the first 3 principal components of a PCA, which was applied to second derivative Fourier-transform infrared spectra preprocessed with the Savitzky-Golay algorithm. FTIR spectra were acquired from *E. coli* cells prior to exposure (+), and after exposure to both rifampicin (o) and nanoencapsulated rifampicin (x). Individual replicates are shown to evaluate reproducibility.

A different output of a PCA are its coefficients, or loadings, which show the correlation between PCs and the variables, e.g., wavenumbers. Loading plots are a visual representation of the correlation between each variable to each PC, which is denoted by a vector pinned at the origin. As such, their length reflects the relative weight of the variable on the PC, while the

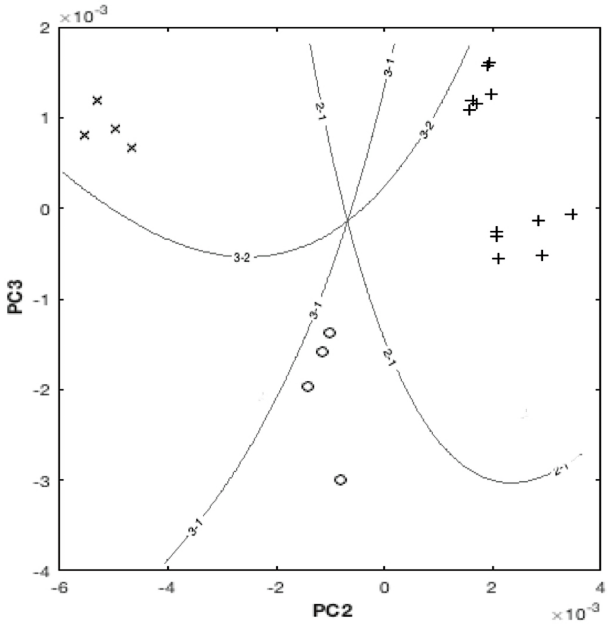


Figure 7. Quadratic discriminant analysis boundaries superimposed on the score plot of second and third principal components of a PCA, which was applied to second derivative Fourier-transform infrared spectra after the Savitzky-Golay algorithm. FTIR spectra were acquired from *E. coli* cells prior to exposure (+), and after exposure to both rifampicin (o) and nanoencapsulated rifampicin (x). Individual replicates are shown to evaluate reproducibility.

angle between vectors reveals the correlation between variables. Given the loading matrix normalization constrain, where the sum of squared elements for a variable is set to one, a correlation circle can be plotted. The variables close to the center of the circle have a smaller contribution to the PCs and are therefore generally less informative [62]. However, the high dimension, and band overlap, of FTIR spectra makes loading plots impractical.

As an alternative, it is possible to directly identify the variables with greater contribution to the sample distribution pattern observed on the score plots, which is a more common approach when dealing with FTIR spectra. In this case, the loading of every variable is plotted for each PC used to build the final projection. Alternatively, for a more holistic perspective, the sum of the variables used to build the final projection can also be used. This results in a plot with identical abscissa to the original spectra, however the ordinate corresponds to the component loading [60]. Here, the presence of signal-rich regions can be directly traced back to the original spectra, in particular certain spectral bands that have been associated with biomolecules.

As seen, choosing the number of PCs to build the final projection of the original data onto the newly created space is a critical step in PCA. For this the scree plot is typically used, which shows the successive, and decreasing, eigenvalues of PCs. This reveals how much variability of the original data is being retained in the projected space and facilitates the definition of a cut-off point. Even so, there is no assurance that the data variability being discarded is not a direct consequence of the biological phenomena being studied, so this approach requires great care in its implementation. On a slightly different note, PCA can be used for outlier detection by plotting the Q residuals along the Hotelling T^2 [63].

As stated, a PCA can be applied to a multitude of problems, and the case of multivariate regression is equally relevant. Principal Components Regression (PCR) applies a multi-linear regression to estimate the outcome (i.e., the response or dependent variable) on a set of covariates (i.e., predictors, explanatory variables, or independent variables) based on a standard linear regression model. In PCR, instead of regressing dependent variable on explanatory variables directly, the PCs of the explanatory variables are used as regressors [60]. In other words, samples are represented on a new space – the score plots – which maintains the original data structure, and those are used for calibration purposes. As before, this application of PCA can always be related back to the original spectra for biological interpretation.

IV.3.4 PARTIAL LEAST SQUARES

The PLS method was first introduced with the non-linear iterative partial least squares' algorithm, aimed at deriving linear models from nonlinear parameters. Later this was adapted for overdetermined regression problems, which were typically solved with PCR, and the method named PLS. Once a method per se, alternative implementations of PLS were determined, particularly regarding the necessary constrains, to avoid trivial solutions, of the successive PLS directions. These are traditionally either orthogonal or uncorrelated scores constrains [64].

Currently, PLSR is one of the most commonly used chemometric methods for calibration. PLSR is particularly suitable when facing a dataset with more variables than observations, and there is multicollinearity across variables [65]. This is precisely the case of FTIR spectra, so it is not surprising that a PLSR model outperforms an equivalent model (i.e., same number of PCs and Latent Variables (LVs)) obtained with PCR, as seen on Figure 8.

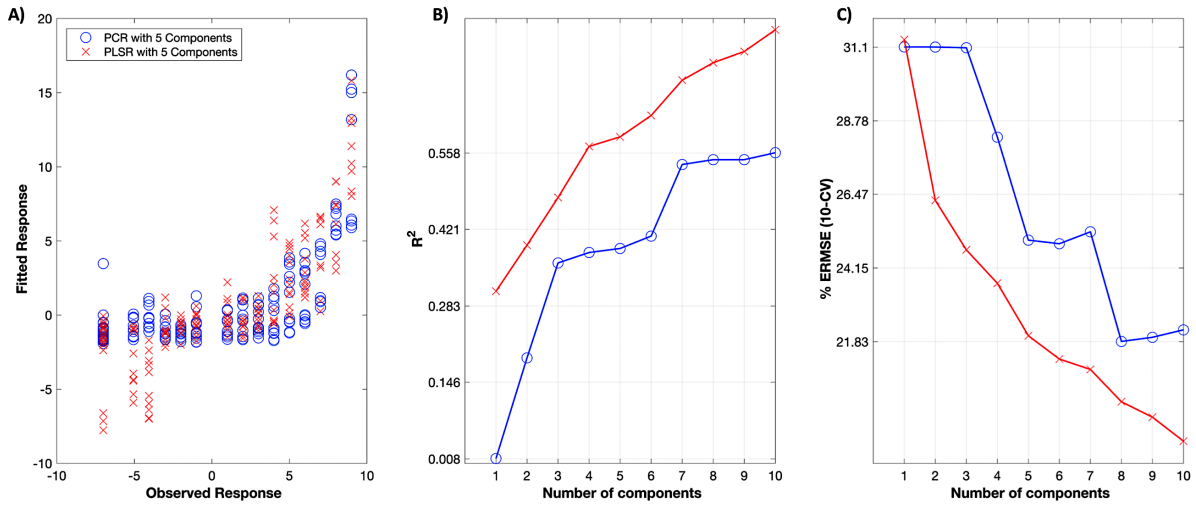


Figure 8. Comparison of observed versus predicted response (A), estimated R^2 (B) and cross-validation root mean squares prediction error (ERMSPE) (C) of a PCR and PLSR. Second derivative spectra of *E. coli* cells exposed to both rifampicin and nanoencapsulated rifampicin were regressed against the $\log_2$ transformed twofold antibiotic concentrations that *E. coli* was exposed to.

Although not originally intended for pattern recognition, Partial Least Squares Discriminant Analysis (PLSDA) has become increasingly applied in a heuristic context. In this case, PLSDA works by fitting a linear PLS regression model to a set of dummy variables that reflect class membership. Thereby, PLSDA enables the projection of observed samples to a new space defined by LVs. In comparison with PCA, PLSDA is advantageous since it not only attempts to maximize the explained variance of the original data, but LVs are constructed so to maximize the correlation between the original data and the dummy variables [66].

In theory, PLSDA is preferred to PCA when the goal is sample separation, which is due to the relationship between PLSDA and canonical correlation analysis, and the subsequent link between the latter and LDA. In fact, PLSDA can be seen as a penalized canonical correlation analysis, with PCAs to determine such penalizations [67]. In practice, the projected samples can be visualized on the newly defined LVs space, which produces a graphical presentation identical to the score plots obtained with a PCA.

However, because PLS maximizes the correlation between the original variables and the classifier, falsely positive predictions occur in a few situations, which is more frequent when the number of variables greatly exceeds the number of samples. Despite this controversial

behavior that contrasts with the original objective of the PLS algorithm, PLSDA is still widely applied, but care should be taken [68] .

Another application of the PLS algorithm is for dimensionality reduction. Similarly to PCA, the loading weights of PLS reflect the importance of the original variables on the LVs and are an easily accessible output. Moreover, because the PLS model can be cross validated, the number of LVs in the fitted model can be optimized. Likewise, the regression coefficients are also typical PLS outputs, and these reflect the weight that each variable has on the fitted response, i.e., the predicted class membership. In either case, setting a minimum threshold above which variables are deemed 'informative' serves to exclude non-informative variables.

However, these thresholds rely on absolute values that may be challenging to define. As such, the variable importance in projection method was developed, which provides a relative measure, and can therefore be expressed as a percentage. In essence, this method tallies the contribution of each original variable to the variance modelled by each LV, which is weighted by the covariance between the original variables and the predicted class membership [66]. Other dimensionality reduction techniques based on the PLS algorithm have been developed and have been reviewed elsewhere [69].

IV.3.5 Unsupervised clustering methods

Unsupervised pattern recognition has been a main objective of chemometrics since it's early days, when researchers faced questions as what intrinsic patterns does the data have? How do these relate with known variables? To answer these, different types of cluster analysis are typically applied, most of which have been adapted from numerical taxonomy towards their application in chemometrics [70].

HCA is an unsupervised pattern recognition algorithm, which assumes that the dissimilarity between samples is directly reflected by the samples distance in the descriptor-defined space (the original dataset or after dimensionality reduction) [59]. After the proximity between samples has been determined, the distances between groups of samples, or linkages, are determined to create a multi-level hierarchy, or tree, that is represented on a dendrogram [65], as shown in Figure 9A.

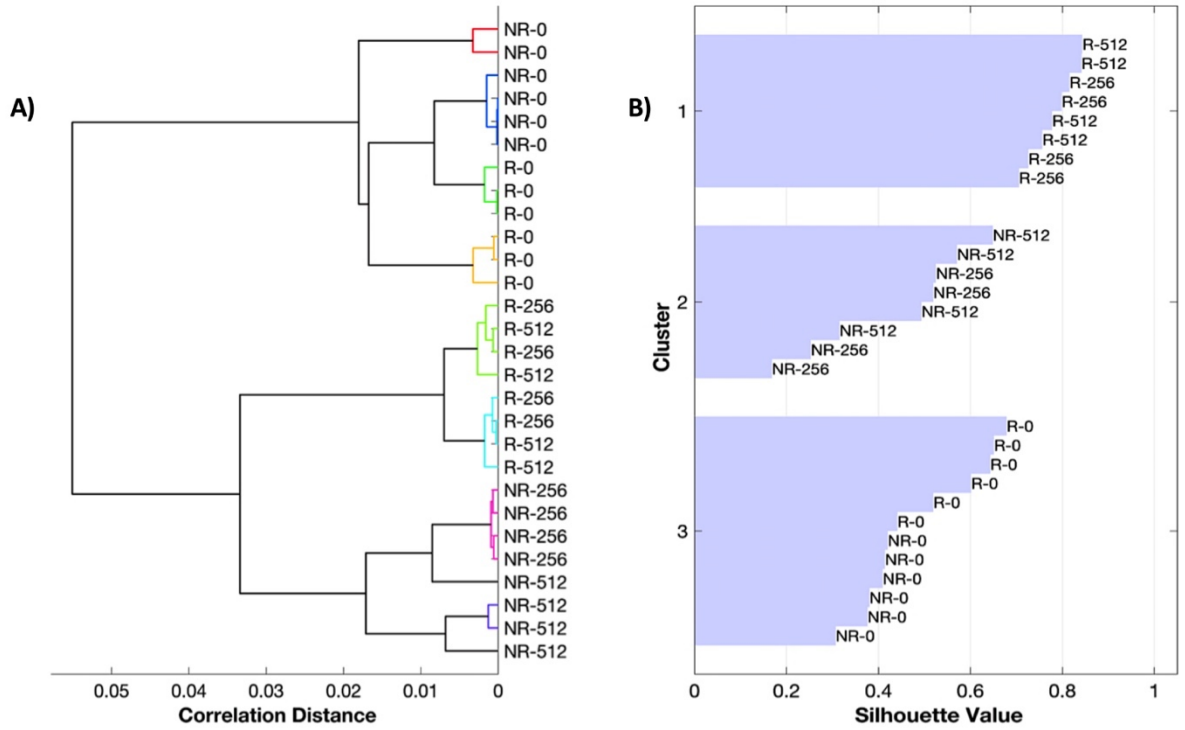


Figure 9. Hierarchical cluster analysis A) dendrogram and B) k-means silhouette plot. Fourier-transform infrared spectra have been acquired from *E. coli* cells prior to exposure (R0 and NR0), and after exposure to both rifampicin (R512 and R256) and nanoencapsulated rifampicin (NR512 and NR256), and individual replicates are shown to evaluate reproducibility and possible misclassifications.

To partition the data into clusters, it is necessary to 'cut' the tree, which can be done arbitrarily or given the natural divisions of the data. In the former, a vertical cut-off is applied on the dendrogram so that a specified number of clusters are formed, despite the fact that these clusters may bear no relationship with the biological phenomenon being studied. In the latter, a cut-off is applied to the inconsistency coefficients of the linkages, which allows to cluster the data regarding how abruptly the samples similarities change among groups of samples.

Unlike HCA, *k*-means operates on observations, rather than their dissimilarity, which results in a single level of clusters. As such, *k*-means is considered a partitioning or non-hierarchical clustering method. The mechanics behind *k*-means are rather simple, which translates to an efficient implementation. In essence, *k*-means assigns class membership of a samples to the cluster whose centroid is closer, and the number of clusters (*k*) is defined by the user [71]. Per definition, various distance metrics can be used, as in HCA.

The resulting clusters are then observed via a silhouette plot, as shown in Figure 9 – B. Here, the value associated with each sample reveals to how close that sample is grouped within its cluster, in relation to the samples in other clusters. Note that silhouette values are contained between -1 and +1. Therefore, smaller negative values suggest that a given sample was poorly clustered, while larger positive values reveal better clustering. Moreover, silhouette values close to zero imply that a given sample is very close to the decision border for a different cluster [72].

IV.3.6 ADVANCED ALGORITHMS

Despite the similar name, *k*-nearest neighbor (KNN) is not related to *k*-means. While the latter is a clustering method, KNN is a supervised classification algorithm that can also be used for regression. As such, KNN is capable of classifying new, unlabeled samples by probing its *k* nearest neighbors. Firstly, the distances within known classes are determined for values of *k* between 1 and the number of samples, which identifies the value of *k* that produces the lowest error. Then, the distance of new samples to those use in training is determined, and the new samples are classified as the majority of its *k* neighbors, which is why typically *k* is an odd number [73].

Soft independent modelling of class analogy is a particularly interesting, supervised pattern recognition method. For each class, a separate PCA is built. For that, either a cross-validation method is used to determine the number of components that must be retained so that sufficient variability of the original data is modelled by the PCA of a given class, or each class is delimited by a region in space that ensures a given samples belongs to its known class, usually within a confidence level [59]. New samples are classified to the class whose PCA model produces the smallest residual. As such, intra-class similarity is favored rather than inter-class dissimilarity [60].

Support Vector Machines (SVM) is considered a generalized linear method that can be applied in both regression and supervised pattern recognition [74]. Typically, SVM is used to build binary classifiers, which work in a similar fashion to the LDA/QDA example previously discussed. In other words, SVM establishes pair-wise boundaries between samples of different classes. However, SVM outperforms LDA and QDA since the support vectors, i.e., the samples at the edge of each class, are used to determine a hyperplane that maximizes inter-class distance [75]. Although SVM is labelled as a linear method, if samples are not linearly

separable on the original space, a kernel transformation can be applied to transform the original samples and represent them on a higher-dimension space where separation can be achieved linearly [75]. Therefore, SVM tends to present better generalization capacity and performs better with higher-dimension data, and in particular, noisy data [76].

A non-parametric alternative are classification and regression trees. Irrespectively of the end goal, this type of analysis is based on a decision tree that starts with an observations' descriptors (root) and evolves towards a response (leaves). Building said tree is typically a three step process where a maximal (overfitted) tree is grown, then pruned into sub-trees, and lastly the optimal tree size is selected [77]. In addition, it is possible to use ensemble methods to combine more than one decision tree in a single classifier. The rational is that, even if the decision tree is just above chance level, i.e., barely outperforms random choices, a combination of several tress of equally poor performances yields a classifier with better performance that a single learner that is theoretically stronger. Examples of these are AdaBoost, bagged trees, or random forests. Moreover, ensemble methods are not exclusive to decision trees. In fact, these can be applied to more than decision trees, for instance to KNN, LDA, QDA or SVM [78].

Artificial Neural Networks (ANNs) can generalize and cope well with nonlinear problems. As such, ANNs are suited for pattern recognition and multivariate regression in degraded, missing or noisy data. In general, large numbers of processing units (neurons) are organized in layers, interconnected by synapses with variable strengths, and an activation function (usually a nonlinear sigmoid function) determines the output, given the weighted sum of inputs [79]. Two widely applied algorithms are probabilistic neural networks and self-organizing maps. While the former are great at rapid supervised classification with relatively low computational requirements [80], the latter are very reliable at unsupervised categorization [81]. The extensive range of existing variations of ANN and SVM implies that their revision, and subsequent application, with the rigor and detail needed to present meaningful examples would require an unreasonably extensive document. Ultimately, as the complexity of algorithms increases, not only does parameterization and optimization become more demanding, but also interpreting their results becomes increasingly challenging.

IV.4 FINAL REMARKS

Advances in the field of FTIRS regarding instrumentation, sampling and detection modes currently permit the analysis of a large variety of samples, including biological fluids, isolated cells, whole tissues and tissue sections. In addition, the multitude of chemometrics methods available, and the informatic tools for their application, maximize the data that can be retrieved from FTIR spectra, which in turn facilitates their interpretation into biologically relevant information. As such, FTIRS is a simple, rapid, economic and mature technology capable of capturing highly sensitive and specific biochemical and metabolic information, whose application potentiates knowledge concerning various biological phenomena. Therefore, the application of FTIRS to mechanism-based high-throughput screening antibiotic discovery is well-justified and holds great potential.

IV.5 REFERENCES

1. Schultz, C.P. Precision infrared spectroscopic imaging: The future of FT-IR spectroscopy. *October* **2001**, *16*.
2. Singh, P.; Andola, H.C.; Rawat, M.S.M.; Pant, G.J.N.; Purohit, V.K. Fourier Transform Infrared (FT-IR) Spectroscopy in An-Overview. *Res. J. Med. Plant* **2011**, *5*, 127–135.
3. Sales, K.C.; Rosa, F.; Sampaio, P.N.; Fonseca, L.P.; Lopes, M.B.; Calado, C.R.C. In Situ Near-Infrared (NIR) versus High-Throughput Mid-Infrared (MIR) Spectroscopy to Monitor Biopharmaceutical Production. *Appl. Spectrosc.* **2015**, *69*, 760–772.
4. Sales, K.C.; Rosa, F.; Cunha, B.R.; Sampaio, P.N.; Lopes, M.B.; Calado, C.R.C. Metabolic profiling of recombinant Escherichia coli cultivations based on high-throughput FT-MIR spectroscopic analysis. *Biotechnol. Prog.* **2017**, *33*, 285–298.
5. López-Lorente, Á.I.; Mizaikoff, B. Mid-infrared spectroscopy for protein analysis: Potential and challenges. *Anal. Bioanal. Chem.* **2016**, *408*, 2875–2889.
6. Yang, H.; Yang, S.; Kong, J.; Dong, A.; Yu, S. Obtaining information about protein secondary structures in aqueous solution using Fourier transform IR spectroscopy. *Nat. Protoc.* **2015**, *10*, 382–396.
7. Wood, B.R. The importance of hydration and DNA conformation in interpreting infrared spectra of cells and tissues. *Chem. Soc. Rev.* **2016**, *45*, 1980–1998.
8. Whelan, D.R.; Bambery, K.R.; Heraud, P.; Tobin, M.J.; Diem, M.; McNaughton, D.; Wood, B.R. Monitoring the reversible B to A-like transition of DNA in eukaryotic cells using Fourier transform infrared spectroscopy. *Nucleic Acids Res.* **2011**, *39*, 5439–5448.
9. Whelan, D.R.; Bambery, K.R.; Puskar, L.; Mcnaughton, D.; Wood, B.R. Quantification of DNA in simple eukaryotic cells using Fourier transform infrared spectroscopy. *J. Biophotonics* **2013**, *6*, 775–784.
10. Ataka, K.; Kottke, T.; Heberle, J. Thinner, smaller, faster: IR techniques to probe the functionality of biological and biomimetic systems. *Angew. Chemie - Int. Ed.* **2010**, *49*, 5416–5424.
11. Pistorius, A.M.A. Biochemical applications of FT-IR spectroscopy. *Spectrosc. Eur.* **1995**, *4*, 8–15.
12. Shai, Y. ATR-FTIR studies in pore forming and membrane induced fusion peptides. *Biochim. Biophys. Acta - Biomembr.* **2013**, *1828*, 2306–2313.

13. Bellisola, G.; Sorio, C. Infrared spectroscopy and microscopy in cancer research and diagnosis. *Am. J. Cancer Res.* **2012**, *2*, 1–21.

14. Stuart, B.H. *Infrared Spectroscopy: Fundamentals and Applications*; 2004;

15. Rosa, F.; Sales, K.C.; Carmelo, J.G.; Fernandes-Platzgummer, A.; da Silva, C.L.; Lopes, M.B.; Calado, C.R.C. Monitoring the *ex-vivo* expansion of human mesenchymal stem/stromal cells in xeno-free microcarrier-based reactor systems by MIR spectroscopy. *Biotechnol. Prog.* **2016**, *32*, 447–455.

16. Baker, M.J.; Hussain, S.R.; Lovergne, L.; Untereiner, V.; Hughes, C.; Lukaszewski, R.A.; Thiéfin, G.; Sockalingum, G.D. Developing and understanding biofluid vibrational spectroscopy: A critical review. *Chem. Soc. Rev.* **2016**, *45*, 1803–1818.

17. Ataka, K.; Kottke, T.; Heberle, J. Thinner, smaller, faster: IR techniques to probe the functionality of biological and biomimetic systems. *Angew. Chemie - Int. Ed.* **2010**, *49*, 5416–5424.

18. Bassan, P.; Sachdeva, A.; Shanks, J.H.; Brown, M.D.; Clarke, N.W.; Gardner, P. Whole organ cross-section chemical imaging using label-free mega-mosaic FTIR microscopy. *Analyst* **2013**, *138*, 7066–7069.

19. Pilling, M.J.; Bassan, P.; Gardner, P. Comparison of transmission and transflectance mode FTIR imaging of biological tissue. *Analyst* **2015**, *140*, 2383–2392.

20. Bunaciu, A.A.; Fleschin, Ş.; Hoang, V.D.; Aboul-Enein, H.Y. Vibrational Spectroscopy in Body Fluids Analysis. *Crit. Rev. Anal. Chem.* **2017**, *47*, 67–75.

21. Bunaciu, A.A.; Fleschin, S.; Aboul-Enein, H.Y. Biomedical Investigations Using Fourier Transform-Infrared Microspectroscopy. *Crit. Rev. Anal. Chem.* **2014**, *44*, 270–276.

22. Baker, M.J.; Faulds, K. Fundamental developments in clinical infrared and Raman spectroscopy. *Chem. Soc. Rev.* **2016**, *45*, 1792–1793.

23. Tiwari, S.; Bhargava, R. Extracting knowledge from chemical imaging data using computational algorithms for digital cancer diagnosis. *Yale J. Biol. Med.* **2015**, *88*, 131–143.

24. Mitchell, A.L.; Gajjar, K.B.; Theophilou, G.; Martin, F.L.; Martin-Hirsch, P.L. Vibrational spectroscopy of biofluids for disease screening or diagnosis: Translation from the laboratory to a clinical setting. *J. Biophotonics* **2014**, *7*, 153–165.

25. Kondepati, V.R.; Heise, H.M.; Backhaus, J. Recent applications of near-infrared spectroscopy in cancer diagnosis and therapy. *Anal. Bioanal. Chem.* **2008**, *390*, 125–139.

26. Sakudo, A. Near-infrared spectroscopy for medical applications: Current status and future perspectives. *Clin. Chim. Acta* **2016**, *455*, 181–188.

27. Holt, D.; Okusanya, O.; Judy, R.; Venegas, O.; Jiang, J.; DeJesus, E.; Eruslanov, E.; Quatromoni, J.; Bhojnagarwala, P.; Deshpande, C.; et al. Intraoperative near-infrared imaging can distinguish cancer from normal tissue but not inflammation. *PLoS One* **2014**, *9*.

28. Frangioni, J. V. In vivo near-infrared fluorescence imaging. *Curr. Opin. Chem. Biol.* **2003**, *7*, 626–634.

29. Hilderbrand, S.A.; Weissleder, R. Near-infrared fluorescence: application to in vivo molecular imaging. *Curr. Opin. Chem. Biol.* **2010**, *14*, 71–79.

30. Frangioni, J. V; Kim, S.-W.; Ohnishi, S.; Kim, S.; Bawendi, M.G. Sentinel lymph node mapping with type-II quantum dots. *Methods Mol. Biol.* **2007**, *374*, 147–159.

31. Namikawa, T.; Sato, T.; Hanazaki, K. Recent advances in near-infrared fluorescence-guided imaging surgery using indocyanine green. *Surg. Today* **2015**, *45*, 1467–1474.

32. Hoshi, Y. Towards the next generation of near-infrared spectroscopy. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **2011**, *369*, 4425–4439.

33. Sakudo, A. Near-infrared spectroscopy for medical applications: Current status and future perspectives. *Clin. Chim. Acta* **2016**, *455*, 181–188.

34. Saliba, J.; Bortfeld, H.; Levitin, D.J.; Oghalai, J.S. Functional near-infrared spectroscopy for neuroimaging in cochlear implant recipients. *Hear. Res.* **2016**, *338*, 64–75.

35. Ferrante do Amaral, C.E.; Wolf, B. Current development in non-invasive glucose monitoring. *Med. Eng. Phys.* **2008**, *30*, 541–549.

36. Kottmann, J.; Rey, J.M.; Luginbühl, J.; Reichmann, E.; Sigrist, M.W. Glucose sensing in human epidermis using mid-infrared photoacoustic detection. *Biomed. Opt. Express* **2012**, *3*, 667.

37. Shaw, R.A.; Kotowich, S.; Mantsch, H.H.; Leroux, M. Quantitation of protein, creatinine, and urea in urine by near-infrared spectroscopy. *Clin. Biochem.* **1996**, *29*, 11–19.

38. Liu, K.Z.; Shi, M.; Man, A.; Dembinski, T.C.; Shaw, R.A. Quantitative determination of serum LDL cholesterol by near-infrared spectroscopy. *Vib. Spectrosc.* **2005**, *38*, 203–208.

39. Liu, K.Z.; Shaw, R.A.; Man, A.; Dembinski, T.C.; Mantsh, H.H. Reagent-free, simultaneous determination of serum cholesterol in HDL and LDL by infrared spectroscopy. *Clin. Chem.* **2002**, *48*, 499–506.

40. Sandor, M.; Rüdinger, F.; Bienert, R.; Grimm, C.; Solle, D.; Scheper, T. Comparative study of non-invasive monitoring via infrared spectroscopy for mammalian cell cultivations. *J. Biotechnol.* **2013**, *168*, 636–645.

41. Teixeira, A.P.; Oliveira, R.; Alves, P.M.; Carrondo, M.J.T. Advances in on-line monitoring and control of mammalian cell cultures: Supporting the PAT initiative. *Biotechnol. Adv.* **2009**, *27*, 726–732.

42. Ren, M.; Arnold, M.A. Comparison of multivariate calibration models for glucose, urea, and lactate from near-infrared and Raman spectra. *Anal. Bioanal. Chem.* **2007**, *387*, 879–888.

43. Moskovits, M. Surface-enhanced spectroscopy. *Rev. Mod. Phys.* **1985**, *57*, 783–826.

44. Osawa, M.; Ken-Ichi, A.; Yoshii, K.; Yuji, N. Surface-Enhanced Infrared Spectroscopy: The Origin of the Absorption Enhancement and Band Selection Rule in the Infrared Spectra of Molecules Adsorbed on Fine Metal Particles. *Appl. Spectrosc.* **1993**, *47*, 1497–1502.

45. Butler, H.J.; Ashton, L.; Bird, B.; Cinque, G.; Curtis, K.; Dorney, J.; Esmonde-White, K.; Fullwood, N.J.; Gardner, B.; Martin-Hirsch, P.L.; et al. Using Raman spectroscopy to characterize biological materials. *Nat. Protoc.* **2016**, *11*, 664–687.

46. Butler, H.J.; Ashton, L.; Bird, B.; Cinque, G.; Curtis, K.; Dorney, J.; Esmonde-White, K.; Fullwood, N.J.; Gardner, B.; Martin-Hirsch, P.L.; et al. Using Raman spectroscopy to characterize biological materials. *Nat. Protoc.* **2016**, *11*, 664–687.

47. Eberhardt, K.; Stiebing, C.; Matthäus, C.; Schmitt, M.; Popp, J. Advantages and limitations of Raman spectroscopy for molecular diagnostics: an update. *Expert Rev. Mol. Diagn.* **2015**, *15*, 773–787.

48. Kong, K.; Kendall, C.; Stone, N.; Notingher, I. Raman spectroscopy for medical diagnostics - From in-vitro biofluid assays to in-vivo cancer detection. *Adv. Drug Deliv. Rev.* **2015**, *89*, 121–134.

49. Dong, D.; Zhao, C. Limitations and challenges of using raman spectroscopy to detect the abiotic plant stress response. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, E5486–E5487.

50. Mondol, A.S.; Patel, M.D.; Rüger, J.; Stiebing, C.; Kleiber, A.; Henkel, T.; Popp, J.; Schie, I.W. Application of High-Throughput Screening Raman Spectroscopy (HTS-RS) for Label-Free Identification and Molecular Characterization of Pollen. *Sensors* **2019**, *19*, 1–16.

51. Schie, I.W.; Rüger, J.; Mondol, A.S.; Ramoji, A.; Neugebauer, U.; Krafft, C.; Popp, J. High-Throughput Screening Raman Spectroscopy Platform for Label-Free Cellomics. *Anal. Chem.* **2018**, *90*, 2023–2030.

52. Gautam, R.; Vanga, S.; Ariese, F.; Umapathy, S. Review of multidimensional data processing approaches for Raman and infrared spectroscopy. *EPJ Tech. Instrum.* **2015**, *2*, 8.

53. Wold, S. Chemometrics; what do we mean with it, and what do we want from it? *Chemom. Intell. Lab. Syst.* **1995**, *30*, 109–115.

54. Rodionova, O.Y.; Pomerantsev, A.L. Chemometrics: achievements and prospects. *Russ. Chem. Rev.* **2006**, *75*, 271–287.

55. Lasch, P. Spectral pre-processing for biomedical vibrational spectroscopy and microspectroscopic imaging. *Chemom. Intell. Lab. Syst.* **2012**, *117*, 100–114.

56. Randolph, T.W. Scale-based normalization of spectral data. *Cancer Biomarkers* **2006**, *2*, 135–144.

57. Marques, V.; Cunha, B.; Couto, A.; Sampaio, P.; Fonseca, L.P.; Aleixo, S.; Calado, C.R.C. Characterization of gastric cells infection by diverse Helicobacter pylori strains through Fourier-transform infrared spectroscopy. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2019**, *210*, 193–202.

58. Ribeiro da Cunha, B.; Aleixo, S.M.; Fonseca, L.P.; Calado, C.C. Towards an automated statistical workflow for biomarker screening in Fourier-transform infrared spectroscopy. In Proceedings of the 2019 IEEE 6th Portuguese Meeting on Bioengineering (ENBENG); 2019; pp. 1–4.

59. Zielinski, A.A.F.; Haminiuk, C.W.I.; Nunes, C.A.; Schnitzler, E.; van Ruth, S.M.; Granato, D. Chemical Composition, Sensory Properties, Provenance, and Bioactivity of Fruit Juices as Assessed by Chemometrics: A Critical Review and Guideline. *Compr. Rev. Food Sci. Food Saf.* **2014**, *13*, 300–316.

60. Roggo, Y.; Chalus, P.; Maurer, L.; Lema-Martinez, C.; Edmond, A.; Jent, N. A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies. *J. Pharm. Biomed. Anal.* **2007**, *44*, 683–700.

61. Trevisan, J.; Angelov, P.P.; Carmichael, P.L.; Scott, A.D.; Martin, F.L. Extracting biological information with computational analysis of Fourier-transform infrared (FTIR) biospectroscopy datasets: Current practices to future perspectives. *Analyst* **2012**, *137*, 3202–3215.

62. Abdi, H.; Williams, L.J. Principal component analysis. *WIREs Comput. Stat.* **2010**, *2*, 433–459.

63. Römer, M.; Heinämäki, J.; Strachan, C.; Sandler, N.; Yliruusi, J. Prediction of tablet film-coating thickness using a rotating plate coating system and NIR spectroscopy. *AAPS PharmSciTech* **2008**, *9*, 1047–1053.

64. Nie, F.; Xiang, S.; Liu, Y.; Hou, C.; Zhang, C. Orthogonal vs. uncorrelated least squares discriminant analysis for feature extraction. *Pattern Recognit. Lett.* **2012**, *33*, 485–491.

65. Lourenço, N.D.; Lopes, J.A.; Almeida, C.F.; Sarraguça, M.C.; Pinheiro, H.M. Bioreactor monitoring with spectroscopy and chemometrics: A review. *Anal. Bioanal. Chem.* **2012**, *404*, 1211–1237.

66. Tran, T.N.; Afanador, N.L.; Buydens, L.M.C.; Blanchet, L. Interpretation of variable importance in Partial Least Squares with Significance Multivariate Correlation (sMC). *Chemom. Intell. Lab. Syst.* **2014**, *138*, 153–160.

67. Barker, M.; Rayens, W. Partial least squares for discrimination. *J. Chemom.* **2003**, *17*, 166–173.

68. Brereton, R.G.; Lloyd, G.R. Partial least squares discriminant analysis: Taking the magic away. *J. Chemom.* **2014**, *28*, 213–225.

69. Mehmood, T.; Liland, K.H.; Snipen, L.; Sæbø, S. A review of variable selection methods in Partial Least Squares Regression. *Chemom. Intell. Lab. Syst.* **2012**, *118*, 62–69.

70. Brereton, R.G. Statistical Concepts. In *Applied Chemometrics for Scientists*; John Wiley & Sons, Ltd, 2007; pp. 63–109 ISBN 9780470057780.

71. Anzanello, M.J.; Fogliatto, F.S.; Ortiz, R.S.; Limberger, R.; Mariotti, K. Selecting relevant Fourier transform infrared spectroscopy wavenumbers for clustering authentic and counterfeit drug samples. *Sci. Justice* **2014**, *54*, 363–368.

72. Taboada, H.A.; Coit, D.W. Data Clustering of Solutions for Multiple Objective System Reliability Optimization Problems. *Qual. Technol. Quant. Manag.* **2007**, *4*, 191–210.

73. Castillo, R.; Contreras, D.; Freer, J.; Ruiz, J.; Valenzuela, S. Supervised pattern recognition

techniques for classification of Eucalyptus species from leaves NIR spectra. *J. Chil. Chem. Soc.* **2008**, *53*, 1709–1713.

74. Hu, N.; Wang, G.; Wu, Y.H.; Chen, S.F.; Liu, G.D.; Chen, C.; Wang, D.; He, Z.S.; Yang, X.Q.; He, Y.; et al. LDA-SVM-based EGFR mutation model for NSCLC brain metastases: An observational study. *Med. (United States)* **2015**, *94*, e375.

75. Hajiloo, M.; Rabiee, H.R.; Anooshahpour, M. Fuzzy support vector machine: An efficient rule-based classification technique for microarrays. *BMC Bioinformatics* **2013**, *14*, S4.

76. Verma, R.; Melcher, U. A Support Vector Machine based method to distinguish proteobacterial proteins from eukaryotic plant proteins. *BMC Bioinformatics* **2012**, *13*, S9.

77. Deconinck, E.; Sacré, P.Y.; Coomans, D.; De Beer, J. Classification trees based on infrared spectroscopic data to discriminate between genuine and counterfeit medicines. *J. Pharm. Biomed. Anal.* **2012**, *57*, 68–75.

78. Shin, J.; Im, C.H. Performance Improvement of Near-Infrared Spectroscopy-Based Brain-Computer Interface Using Regularized Linear Discriminant Analysis Ensemble Classifier Based on Bootstrap Aggregating. *Front. Neurosci.* **2020**, *14*, 1–11.

79. Mouwen, D.J.M.; Capita, R.; Alonso-Calleja, C.; Prieto-Gómez, J.; Prieto, M. Artificial neural network based identification of Campylobacter species by Fourier transform infrared spectroscopy. *J. Microbiol. Methods* **2006**, *67*, 131–140.

80. Jia, J.; Liang, C.; Cao, J.; Li, Z. Application of probabilistic neural network in bacterial identification by biochemical profiles. *J. Microbiol. Methods* **2013**, *94*, 86–87.

81. Ballabio, D.; Vasighi, M. A MATLAB toolbox for Self Organizing Maps and supervised neural network learning strategies. *Chemom. Intell. Lab. Syst.* **2012**, *118*, 24–32.

Chapter V

Author contribution

Bernardo Ribeiro da Cunha designed, implemented, tested and used SpecA.

Abstract

What started as a simple graphical user interface with a few buttons to automate the application of preprocessing routines eventually became SpecA. SpecA is a custom-made spectra analyzer to streamline most of the data analysis pipeline of an infrared spectroscopy experiment. This includes loading and handling large datasets, quality control routines, preprocessing, dimensionality reduction, quantitative and qualitative analyses, output generation, among others. Importantly, SpecA makes any of these tasks effortless, and allows the comparison of different parameters with objective performance indicators. In turn, this allows researchers to swiftly evaluate their data and ensures an easier identification of data patterns related with the biological phenomena being investigated. Throughout this chapter, several figures of SpecA 'in action' are accompanied by brief descriptions of how SpecA works from a user perspective. As such, this chapter provides a clear picture of SpecAs' functionalities and showcases how SpecA can be used, thus highlighting the advantages of using SpecA, which in turn validates the reasoning behind creating SpecA. Without SpecA, many of the experiments conducted throughout this Ph.D. would not have been as successful, if at all. SpecA became the powerhouse that catalyzed the scientific productivity of this Ph.D.

Contents

V. 1 INTRODUCTION

SpecA is a custom-made spectra analyzer that streamlines the typical data analysis pipeline of an infrared spectroscopy experiment. SpecA was built using MATLABs' graphics user interface (GUI) design environment, which is a hybrid platform for designing GUIs or apps that combines a drag and drop approach with an integrated editor, so that the behavior of visual components can be programmed.

At its essence, SpecA is a GUI that gives access to a series of algorithms without the need to implement any code. Importantly, SpecA ensures that each analysis can be swiftly executed without sacrificing the functionality or granular control over the algorithms available. Typically, resourcing to code, either in the form of a command line or a script, allows the user to fine-tune several parameters, but the downsides of this approach are that a higher level of expertise is required, the data analysis pipeline becomes considerably less fluid and is more error prone. SpecA was constructed while aiming for an equilibrium between functionality and simplicity, between control and ease of use.

Before dwelling into the data analysis pipeline enabled by SpecA, a few concepts regarding the data structure and nomenclature are worth visiting. Figure 1 shows the data structure employed when working with a SpecA experiment. At its essence, a dataset is a matrix. If nothing else is done, then a dataset is simply a matrix whose entries are the absorbance registered on different samples (rows) over a range of wavenumbers (columns). Within a dataset there are lower hierarchical levels, but these only serve to organize the preprocessing combinations applied by the user. Each dataset has parameters that describe what it is, e.g., a spectra matrix, a principal component analysis (PCA) model, etc., and how it was obtained, e.g., number of principal components (PCs), PCA algorithm, etc. Often, these parameters were named 'Internal variables', since they are part of the dataset, in contrast with the 'external variables' that complement and describe the dataset.

A dataset can be analyzed, in which case a new dataset is generated. For instance, a new dataset is generated if a PCA is applied to a spectral matrix, which shares the same variables, or a subset thereof, with the original dataset. As such, an assay is a group of datasets that share the same variables, or a subset thereof, with the original dataset. An assay can be created by loading a new dataset, by concatenating other assays, or applying a set of rules, which will be discussed ahead.

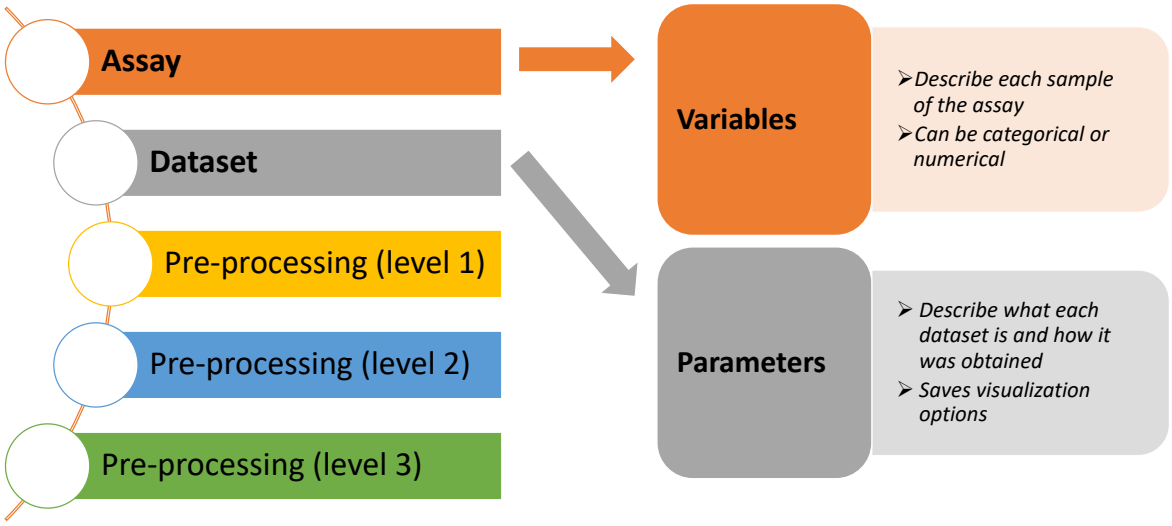


Figure 1. Representation of a SpecA experiment data structure. Major levels are represented in bold and their properties are outlined. Minor levels are merely for organizational purposes.

A collection of assays is called an experiment. This choice of data structure was paramount in creating the listbox selection layout that is characteristic of using SpecA, which allows to swiftly scroll through various levels of preprocessing, datasets and assays.

Rather than describing how SpecA works from a programming or algorithmic perspective, this chapter aims to showcase how SpecA can be used, thus provides a clear notion of its functionalities and highlights the advantages of using SpecA, which in turn validates the reasoning behind creating SpecA. As such, the remainder of this chapter resources to several figures, accompanied by brief descriptions, which capture the user interaction that can be expected when using SpecA.

Two thigs are noteworthy: firstly, the previous chapter has covered many of the fundamentals of the data analysis algorithms that are showcased in the following sections; secondly, the application of these algorithms was described in the context of the research conducted throughout this Ph.D. thesis. Therefore, this chapter is mostly focused on the data analysis workflow with a brief mention to its implementation. Whenever available, MATLABs' built-in functions were preferably used. Otherwise, these were downloaded from File Exchange, a MathWorks community repository, and explicitly mentioned.

V. 2 HOME SCREEN

SpecA is divided into two panels. A smaller control panel on the left, and a larger visualization panel on the right (Figure 2). SpecAs' home page shows the control panel contents on the visualization panel overlaid by brief explanations that remind the user of its functionalities and operation. Importantly, SpecAs' control panel is where nearly all user interactions occur. SpecA takes advantage of the tabbed panel feature, which allows to organize the available functionalities according to their role, namely those regarding the Dataset, Variable, Visualization, or Others, such as exporting. From herein, whenever a word directly refers to a SpecA button or component, it has been highlighted in bold, this way the user knowns what to look for on the respective figures. Moreover, many of SpecAs' functionalities are accessible by clicking on a popup-menu, so whenever relevant, available functionalities were highlighted with an open popup-menu.

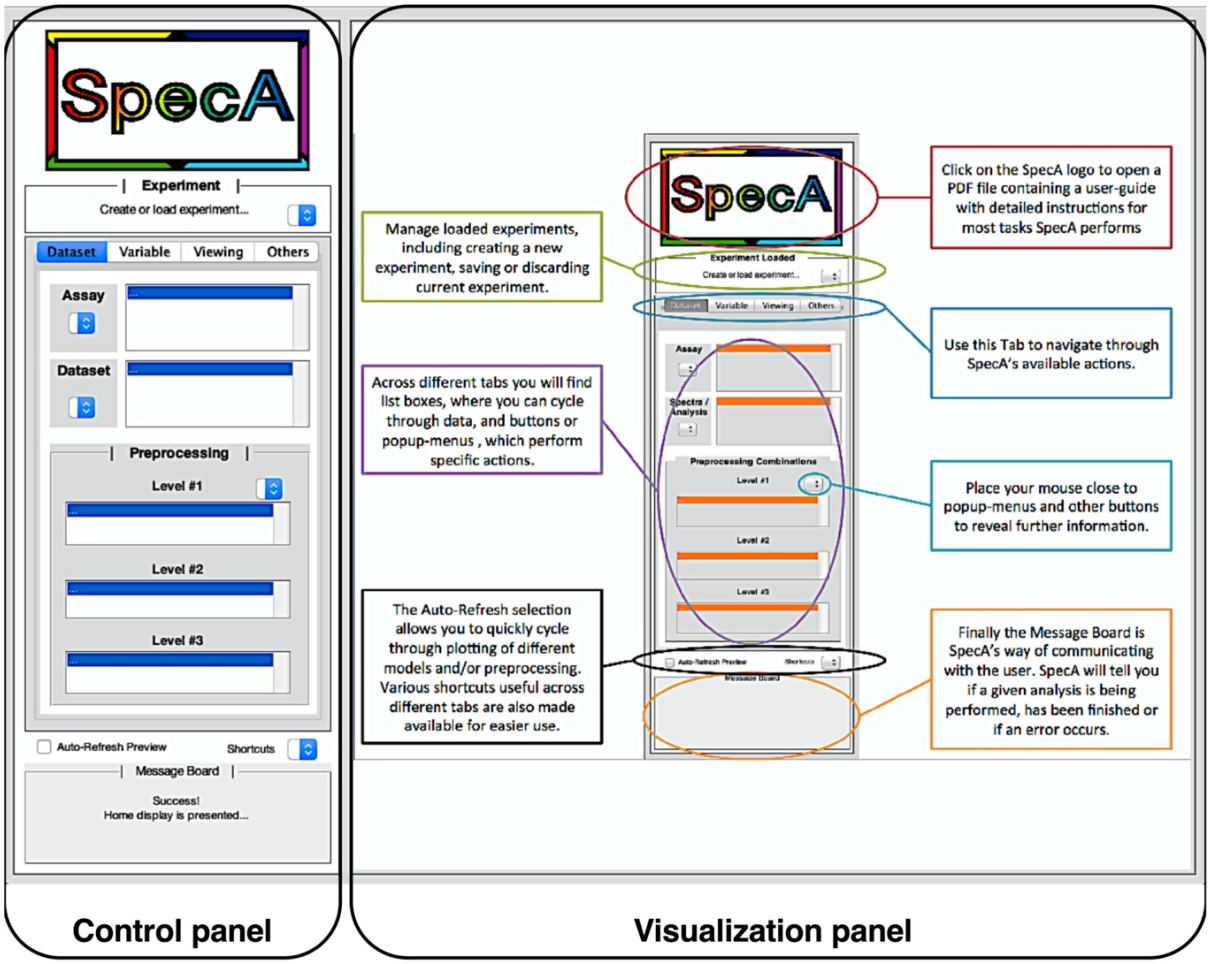


Figure 2. SpecAs' home page. SpecA is composed of a control and a visualization panel. On the latter a brief description of how to use the control panel is presented.

V. 3 CREATING AN EXPERIMENT

To start an experiment, click on the respective popup-menu and choose **Create new experiment** (Figure 3). A sub-GUI is shown with various fields that need to be filled in order to load a dataset within a new assay under the new experiment. This method of overlaying sub-GUIs over the visualization panel will be coherently used herein to showcase the required inputs for each action. SpecA accepts manually selecting data point table (DPT) files (**DPT Manual**), introducing the name of serially numbered DPT files that were placed in SpecAs' input folder (**DPT Series**), or a range within an Excel file (**Microsoft Excel**).

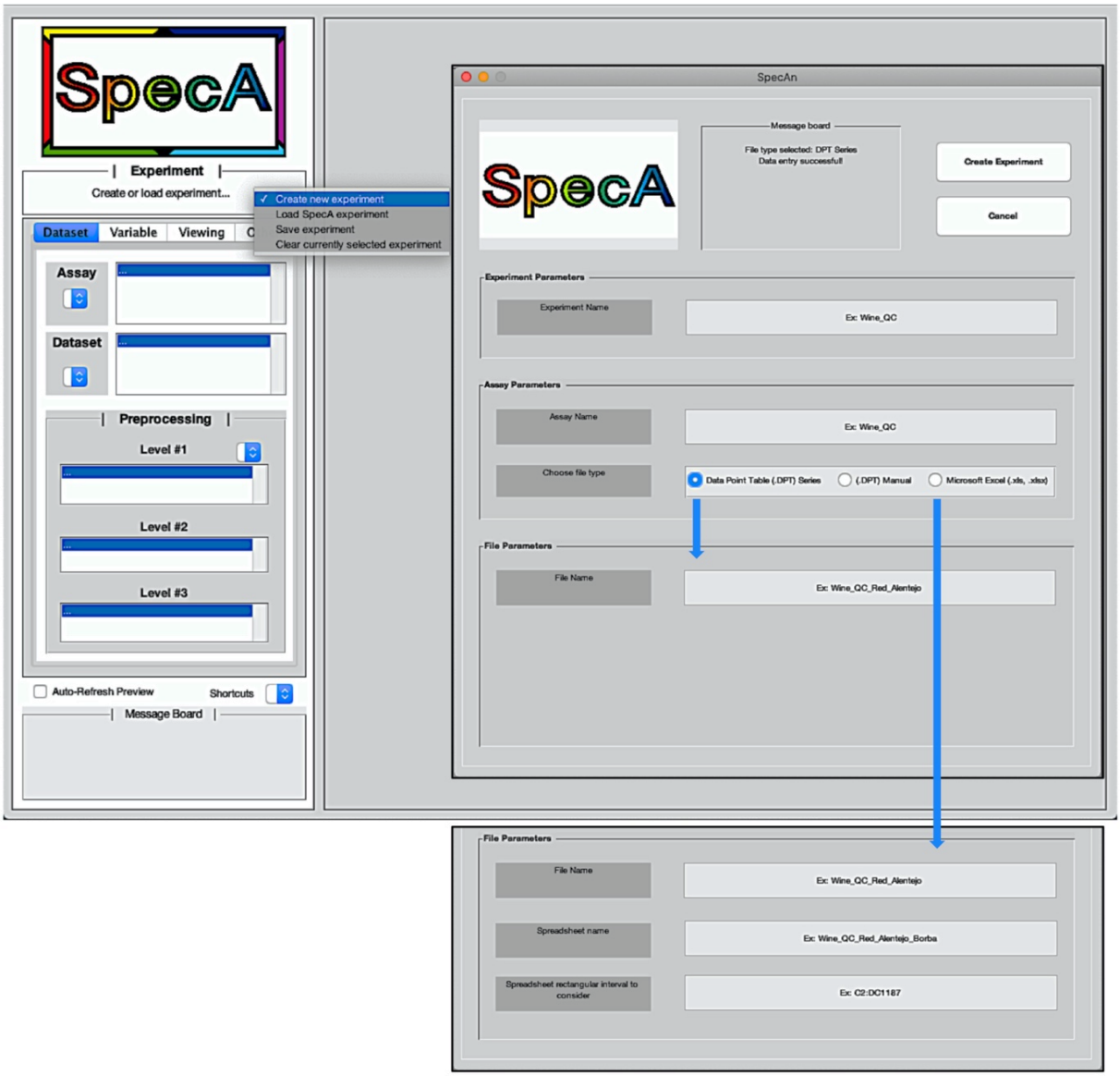


Figure 3. Starting a new SpecA experiment. Highlighted are the required fields when the file type is a DPT file series or an Excel document.

V.3.1 LOADING A NEW DATASET

An assay can be loaded to an existing experiment by choosing **Load new assay** from the same popup-menu shown in Figure 3. In this case, the sub-GUI that is shown lacks the field to enter the experiment name, but other than that is the same.

V.3.2 MANIPULATING EXISTING ASSAYS

Assays can be concatenated (**Group into new assay**), divided (**Create assay from variable rule set**), deleted (**Delete assay**), renamed (**Rename assay**), or duplicated (**Duplicate assay**) (Figure 4). While most of these functionalities are self-explanatory, the ability to divide an assay works by creating an assay whose samples adhere to user-defined rules and discarding the remainder. SpecA allows cumulative rules to be applied and can also carry-over existing preprocessing to the newly created assay.

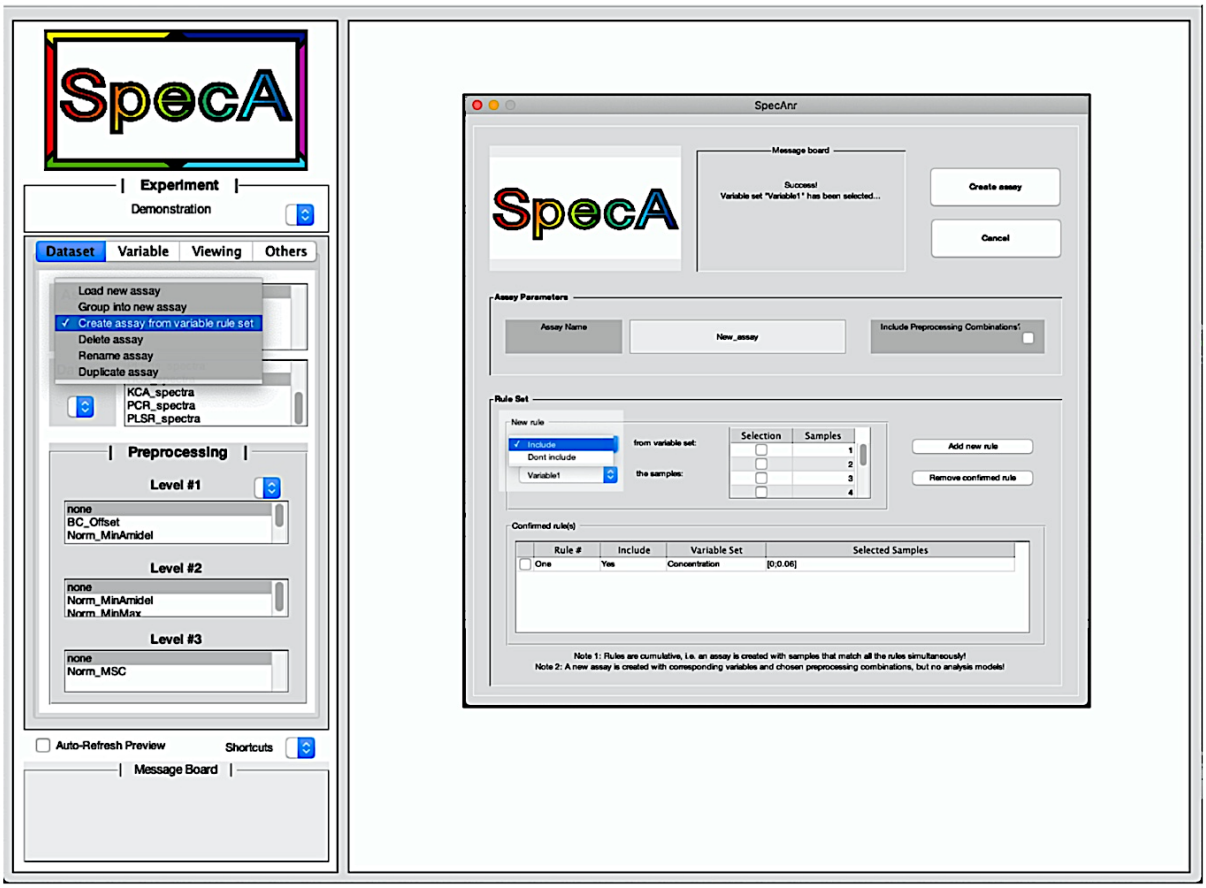


Figure 4. Functionalities available to manipulate assays. The sub-GUI shown allows the user to divide an assay using user-defined rules.

V. 4 ADDING VARIABLES

Before analyzing a dataset, SpecA requires the user to introduce at least one variable (i.e., a numeric response for regression or a label). There are four ways to achieve this.

V.4.1 REPETITIVE VARIABLES

The first is by navigating to the **Variable** tab in the control panel, clicking on the **External Variable** popup-menu, and choosing **Create variable dataset** (Figure 5). Then a sub-GUI is shown where the user can enter the values assumed by the variable, and the number of replicas of each value, given these occur in succession. This is most often useful when variables follow a repetitive pattern, as is the case if replicas are acquired. If other variables exist, and some of which are constant among replicas, choosing **Create replica variable set** automatically creates a new variable with a unique identifier for each replica. For that, the user only needs to identify which variables are constant among replicas.

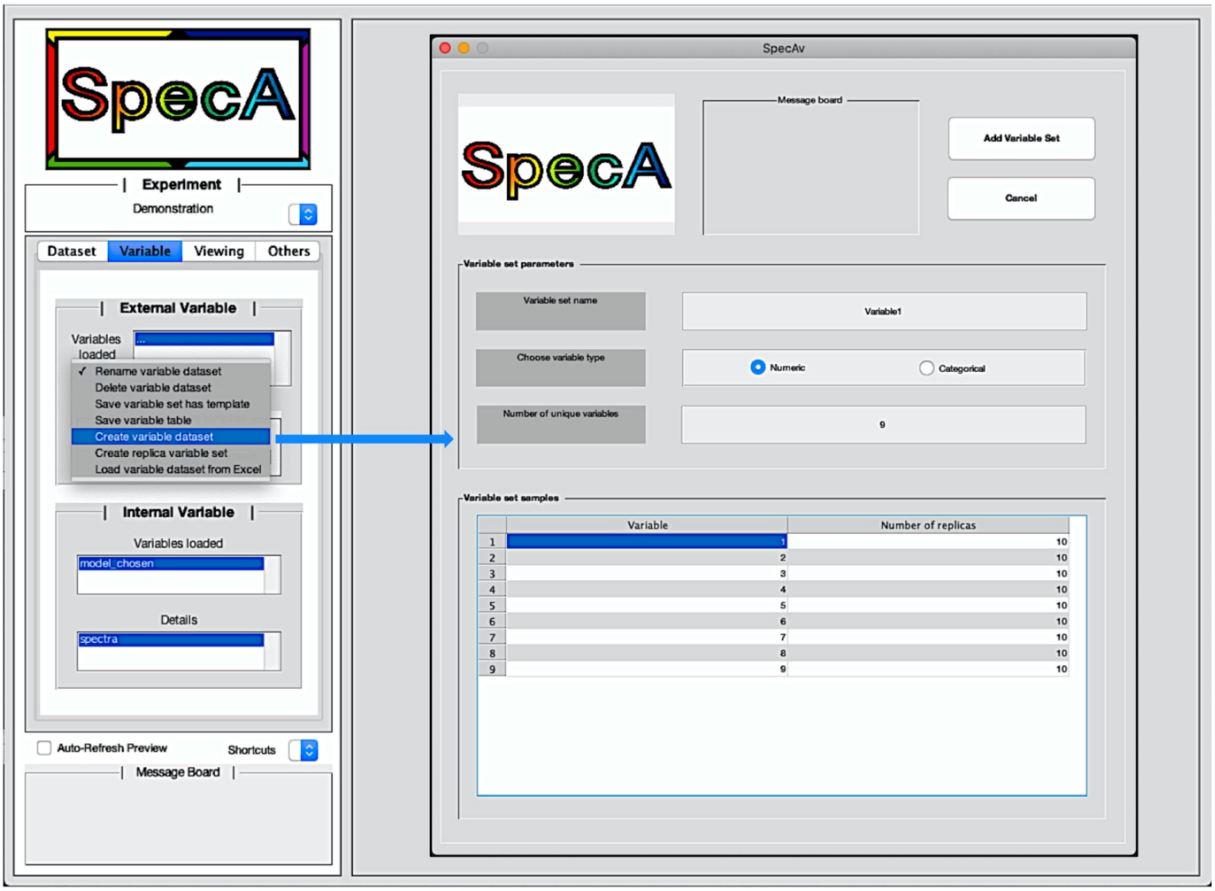


Figure 5. Adding a variable to a dataset using SpecAs' easy variable insertion sub-GUI.

V.4.2 FROM AN EXCEL FILE

Often, it is the case that the variables don't have the replicas in succession, don't follow a clear pattern, or are already recorded on an Excel file. In that case, SpecA enables a smooth importation of data from a spreadsheet, so long as the file is in MATLABs' directory. For that, choose **Load variable dataset from Excel** and fill in the fields of the dialog box that is shown (Figure 6).

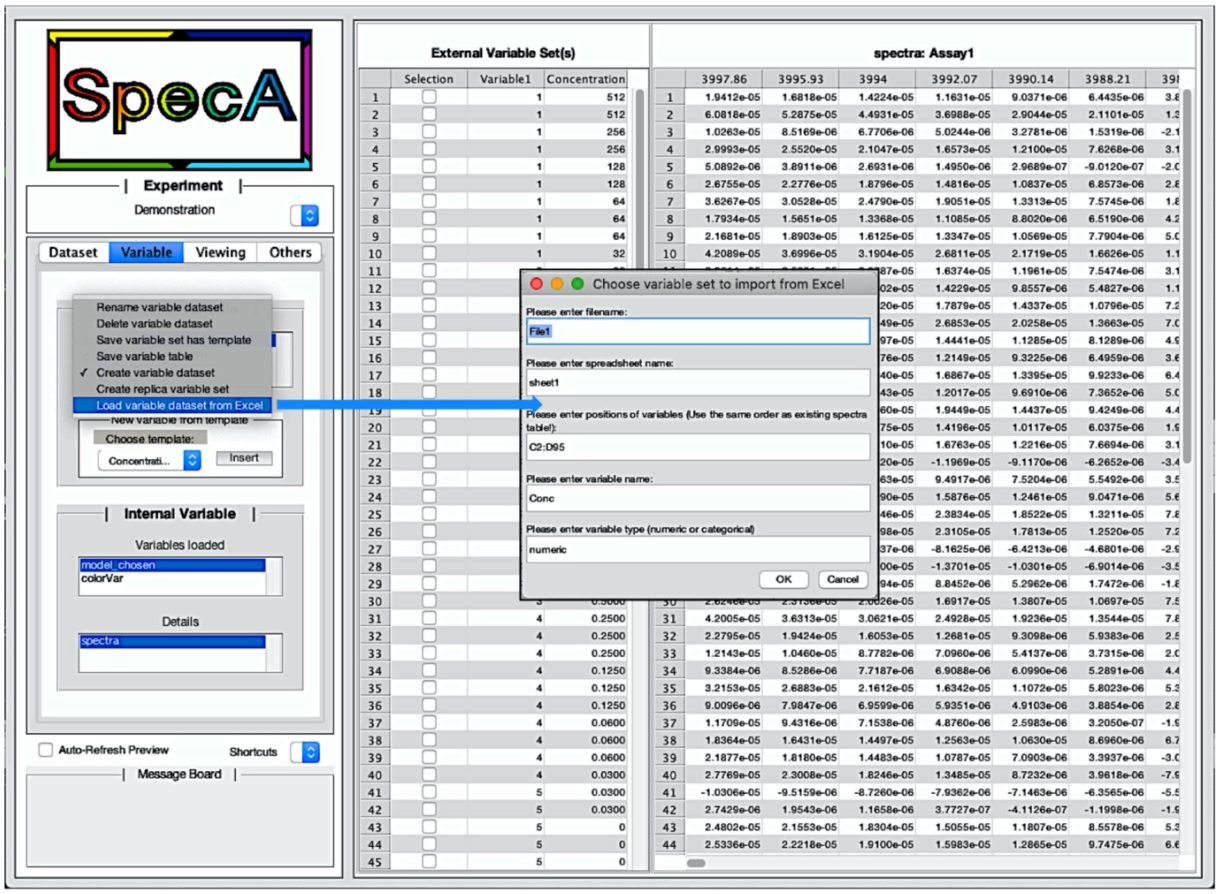


Figure 6. Importing a variable from an Excel file. The process is as simple as inserting the file and spreadsheet name, the cell range to be imported, a name for the new variable, and defining it as numeric or categorical.

V.4.3 USING TEMPLATES

It could also be the case that another assay has an identical variable that the user wishes to import. In this case, variables can be copied to a new assay by first selecting the original variable in the list box under **External variable**, for which the user must navigate to the **Variable** tab of the control panel. Then choose the **Save variable has template** from the popup-menu. SpecA saves that variable, under its original name, as a new template. When creating a new variable from a template, so long as the dimensions of the datasets match, the user only has to choose a name for the variable (Figure 7). As a fourth option, SpecA has pre-recorded templates that either have been useful in previous work (e.g., **Concentrations** and **Replicas**), or can be used to manually enter the data. For the latter, choose either the **Zeros**, **Ones**, or **Name** template and create a new variable. Then use the table on the visualization panel, under **External Variable Set(s)**, to edit the new variable. After, choose **Save variable table** from the variable popup-menu. When using this option, ensure that the template matches the variable type, e.g., **Name** for categorical, and **Zeros** or **Ones** for numeric variables.

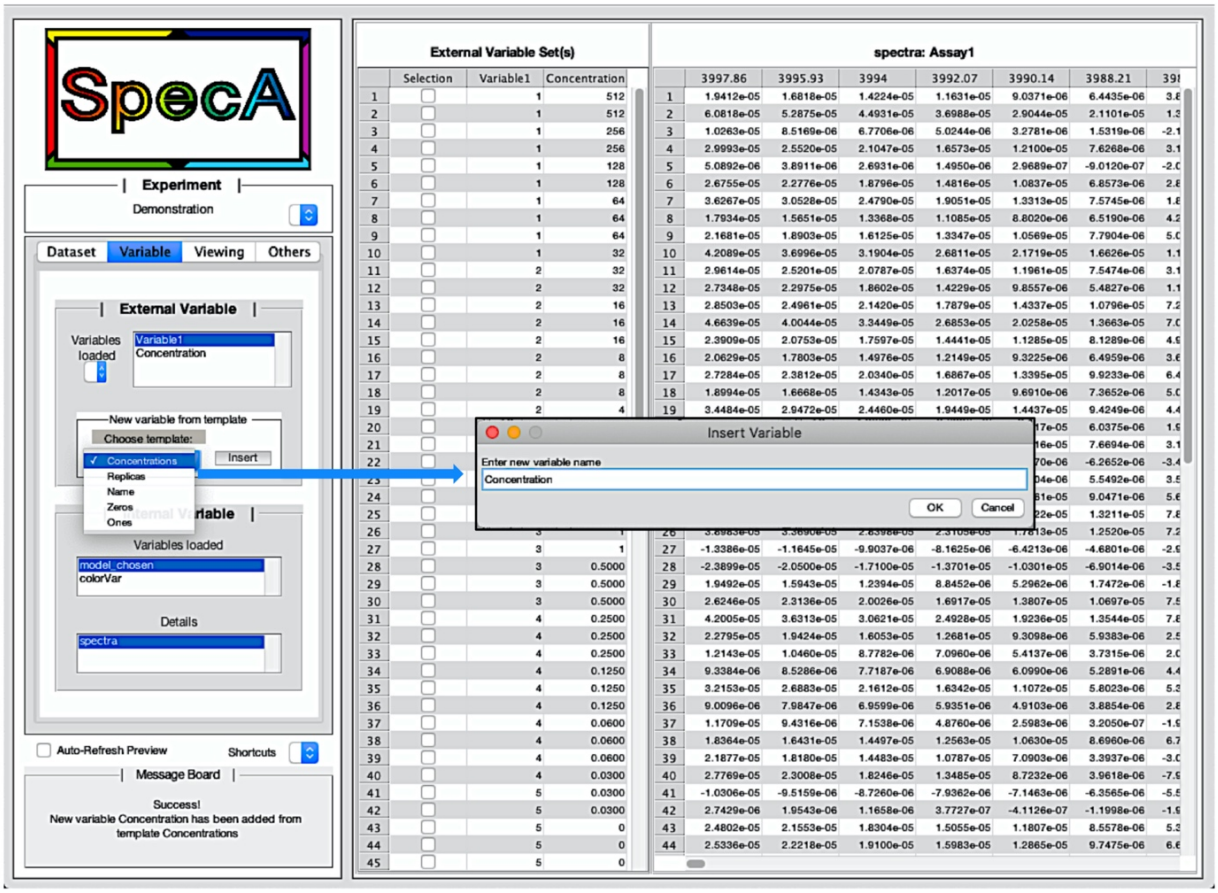


Figure 7. Importing a variable from a template. The user only needs to select the template and a name for the new variable.

V. 5 VISUALIZING SPECTRA

Although SpecA allows visualization of spectra immediately after an experiment has been created, a variable is required for the data to be plotted in a meaningful fashion. When visualizing spectra, it is possible to select (or alter) the coloring variable, in other words the samples assigned to each color, by navigating to the **Viewing** tab, under **Temporary Parameters**, **Coloring variable** (Figure 8). Moreover, informative feedback is presented to the user upon selecting a given point on the plot, which can be navigated using the mouse or the keyboard arrows.

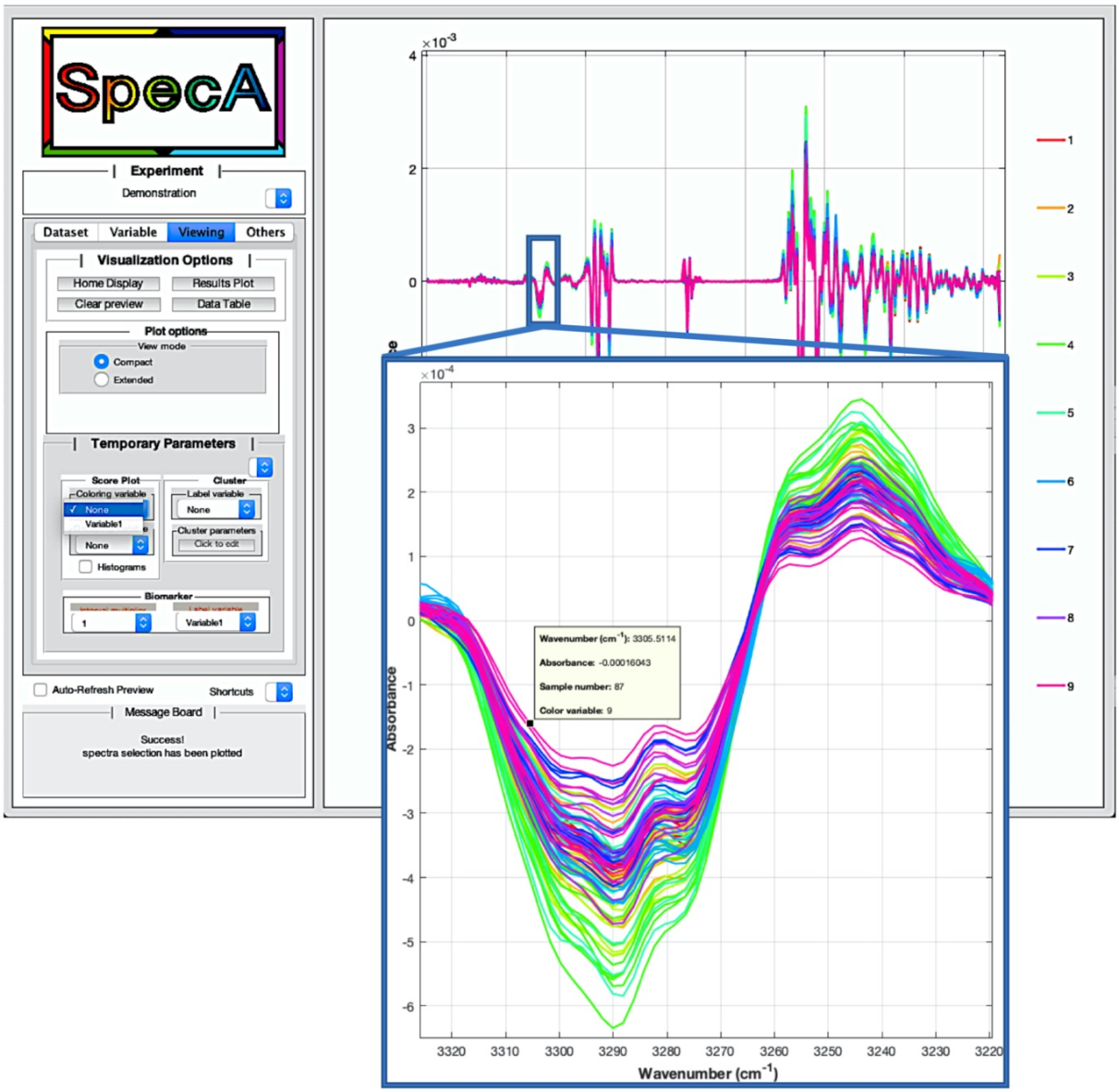


Figure 8. Visualization of second derivative spectra colored with a dummy variable. Blue box shows the visualization panel should the user select to zoom-in on that spectral region. A data point was chosen to demonstrate the information displayed.

V. 6 SPECTRA PREPROCESSING

The last step before analyzing a dataset is to preprocess the spectra. For that, navigate to the **Dataset** tab and open the **Preprocessing** popup-menu (Figure 9). The resulting sub-GUI allows the user to choose the **Type of Pre-processing Combination** to be loaded, either the **Default**, **Template**, or **User-defined**. For the latter, click on the **Allow customization of default preprocessing parameters** tick box, which reveals the **Customized preprocessing combination** panel. Here, after selecting one preprocessing strategy for each level using the popup-menus available, and entering any parameters required for a specific preprocessing algorithm, click **Add custom combination.** This adds the preprocessing combination to those under **Detailed preprocessing combinations to apply**. Note that below the popup-menus to select the type of preprocessing, other popup-menu shown reveal the list of available preprocessing algorithms.

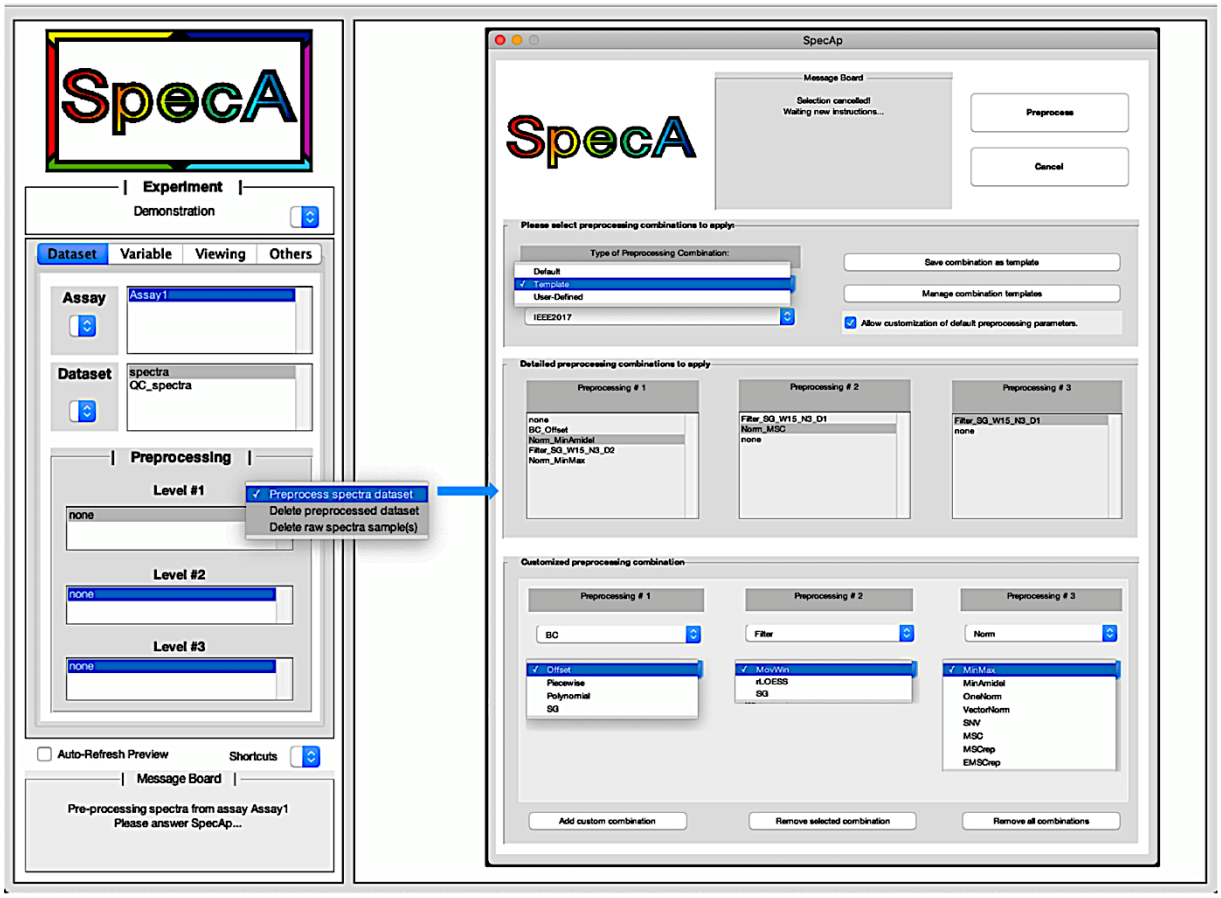


Figure 9. Preprocessing a spectra dataset using a template and complementing it with a customized combination. There are no limits to how many preprocessing combinations, and their parametrization, can be included.

Additionally, a combination of preprocessing algorithms can be loaded from memory by choosing **Template** from the popup-menu below **Type of Pre-processing Combination**, then choosing the desired template from the popup-menu below that one. Clicking **Save combination as template** opens a dialog to record a name for the chosen preprocessing combination, while **Manage combinations templates** allows editing or deleting templates.

V. 7 GENERAL CHARACTERISTICS OF DATASET ANALYSIS

For an equilibrium to be reached between control and ease of use, considerable effort was put into the ability to parametrize the analysis algorithms, while keeping the workflow intuitive, and limiting error prone behavior. In the dataset analysis sub-GUI (Figure 10) the user can select the **Input dataset** to analyze, which defines which **Analysis Algorithm** can be applied for that particular input dataset. Although the user is given lots of freedom to choose successive analysis algorithms, some combinations are not possible. Also, each analysis requires a **Model Name**. A nifty feature is that clicking on the **Suggestion** below the model name actually adds the current suggestion as the model name.

If the user ticks the **Customize Model** box, then not only does the option to **load customization from a template** is presented, but also the **Preprocessing**, **Dataset Restriction** and **Notes** tabs are created (Figure 10). All the selections made across the different tabs can be saved to a template by clicking on **Save as template**. Close to this button, there are other self-explanatory buttons that expedite the burdensome parametrization of a dataset, in addition to reducing error prone repetitive tasks.

Regardless of the type of analysis the user might want to conduct, there are a few parameters that are constant, namely those presented by the **Preprocessing**, **Dataset Restriction** and **Notes** tabs.

In the **Preprocessing** tab the user can quickly choose to include **All** the preprocessing combinations, just the raw spectra (**None**), or a **Custom** selection. In the latter, three listboxes appear below **Confirmed Preprocessing Combination**, which describe the combinations currently selected for analysis.

In the **Dataset Restriction** tab, the user can **Restrict Samples** to choose which values assumed by a given variable are included for analysis. Likewise, the dataset matrix can be

restricted regarding its columns (e.g., **Limit wavenumbers numbers** as depicted on Figure 10, or the number of PCs if the input dataset is a PCA, for instance), in which case the process is equally simple, just fill the required fields.

Moreover, the user can record some metadata along with the model. For example, the **Author Name** can be added to the model, as well as a timestamp (**Include creation date**) and any **descriptive notes** the user finds helpful to record along with the dataset. If the user elects to publish an automatic report, as discussed ahead, this information is then presented alongside the model results within the report.

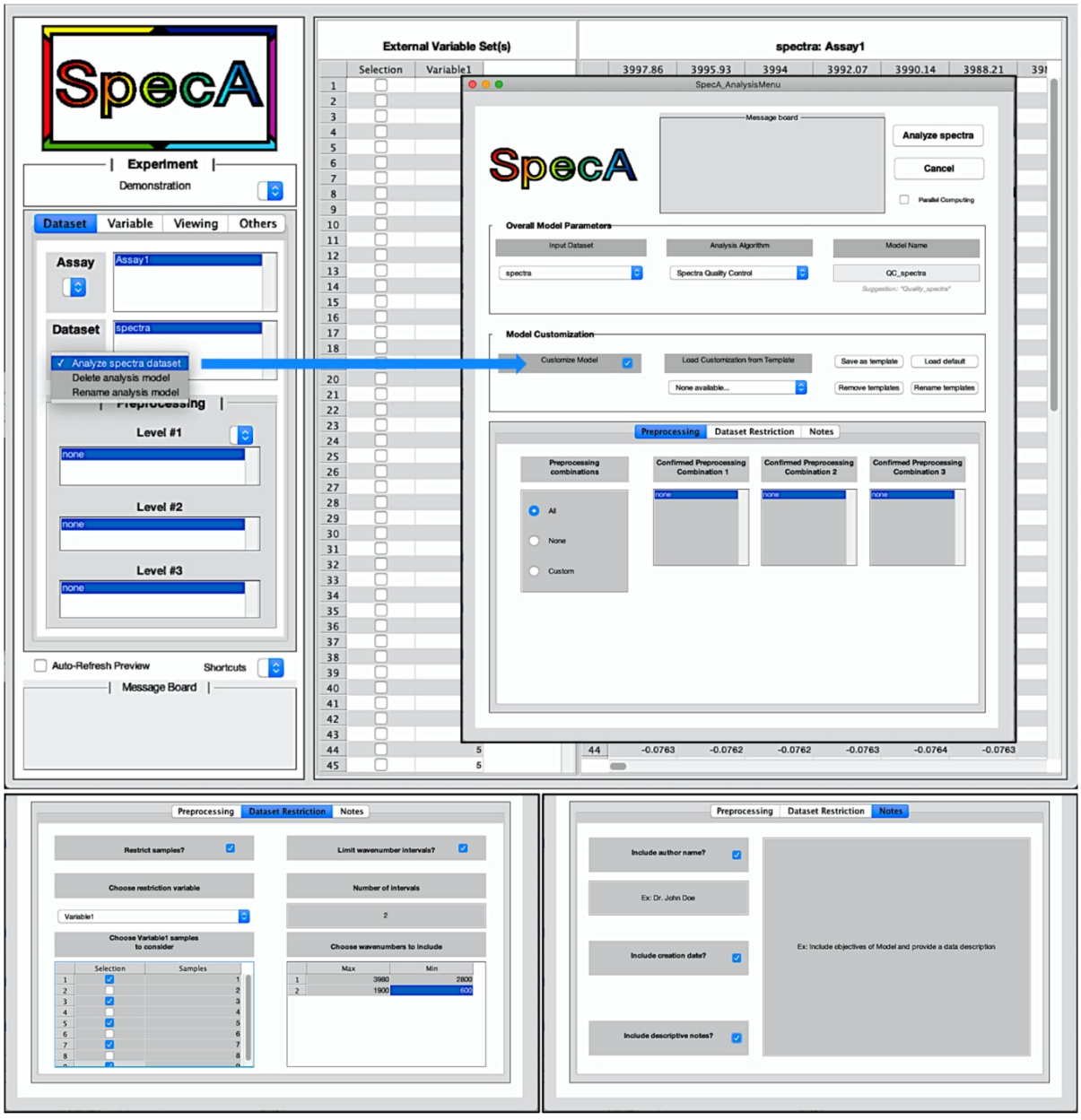


Figure 10. General dataset analysis inputs that are available for all analysis algorithms.

V. 8 QUALITY CONTROL ROUTINES

Within SpecA there are three quality indicators, namely the **signal-to-noise ratio**, **sample thickness** and **water content** [1] (Figure 11), for which three visualization options are available on the visualization panel. These are **Grouping**, which allows the visualization of each indicator per each value assumed by a variable. This is useful for instance when analyzing spectra of different bacteria, where visualizing possible differences in sample thickness that might be associated with that particular grouping could be useful. Alternatively, if a variable is created with the layout of a 96-well plate, then it is also easy to visualize if any particular wells have higher water content, e.g., wells in the middle of the plate tend to require longer for complete dehydration prior to spectra acquisition. Also, the user can choose to view **box-and-whisker** plots, as shown, or the same data represented in **histograms**. Lastly, the **compact** view is shown on Figure 11, while an **extended** view mode expands each quality control indicator over the entire visualization panel and is useful when there are many grouping variables. As for spectra visualization, choosing a datapoint highlights certain information.

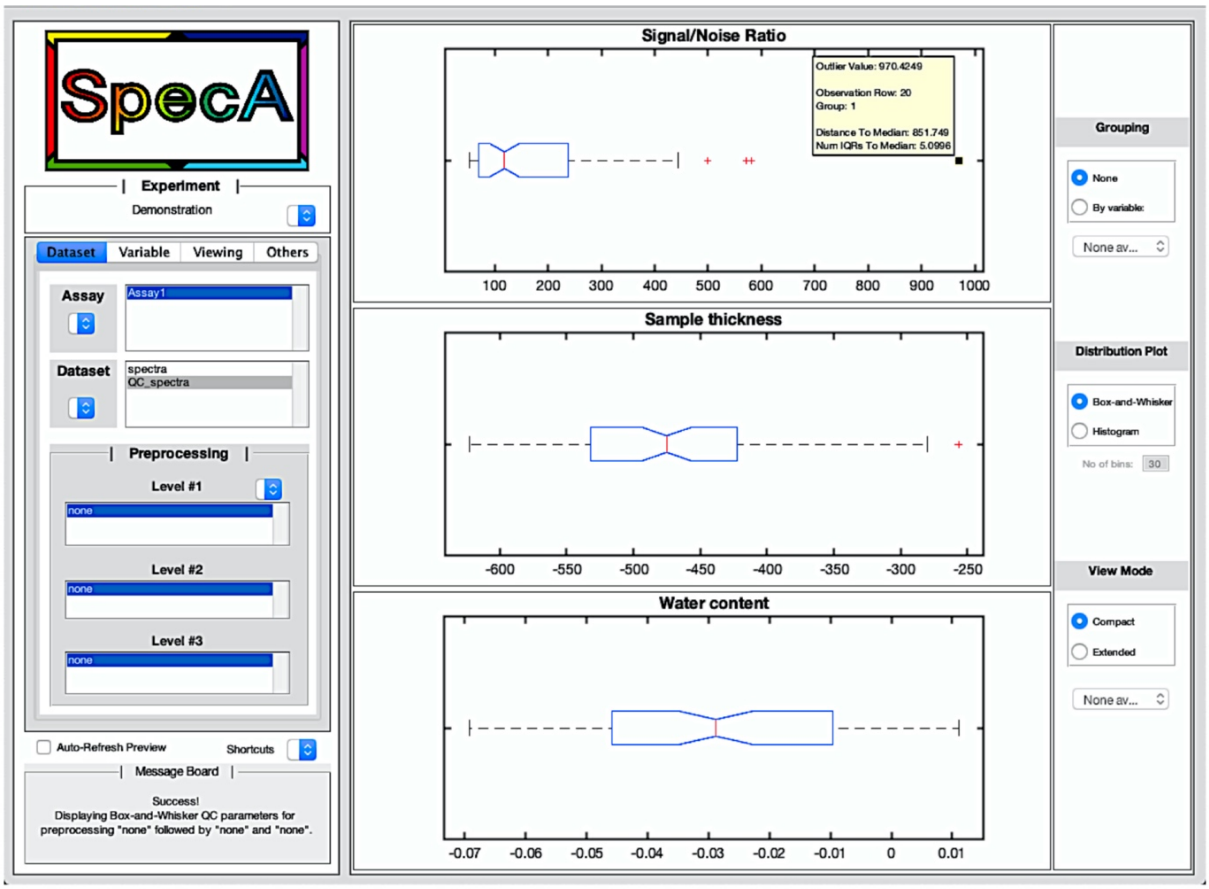


Figure 11. Signal-to-noise ratio, sample thickness and water content estimated with a quality control analysis. Also shown are the available visualization options.

V. 9 DIFFERENCE, AVERAGE AND MEDIAN SPECTRA

The simplest dataset analysis operations that SpecA can apply are determining the **average spectra** and **median spectra**. For that, a **supervision variable** is required to identify which samples are to be considered (Figure 12). This type of analysis is particularly valuable when the acquisition protocol includes mechanical replicates, and the user wants to average these to reduce operator variability. Additionally, SpecA can also determine the **difference spectra**, in which case further selections are required to identify which samples are to be subtracted. Following the traditional application of difference spectra, where the **control** samples are subtracted from **experimental** samples, SpecA presents a popup-menu to quickly select control samples. When the same sample is selected as experimental and control, it is effectively removed from the dataset.

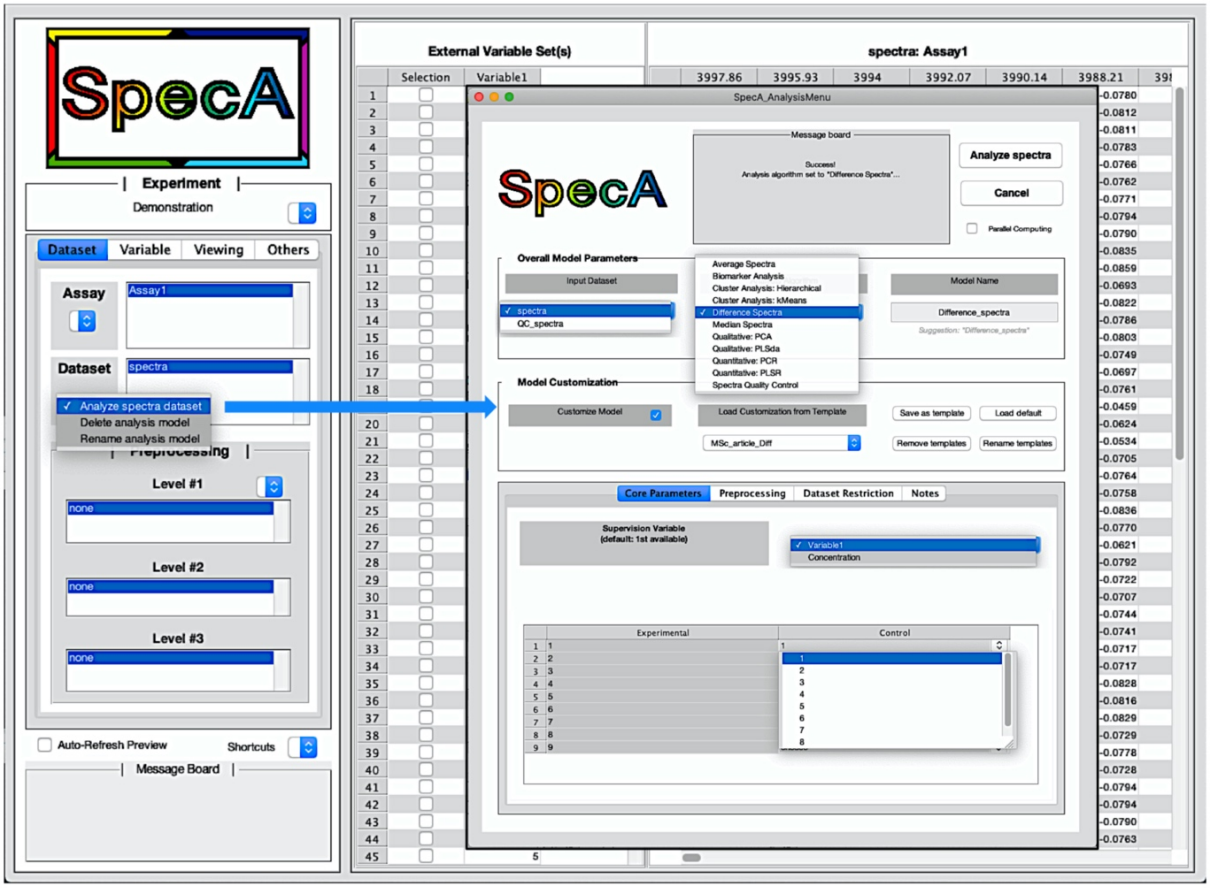


Figure 12. SpecAs' inputs for difference spectra. Average and median spectra have a similar layout regarding the supervision variable but lack the selection table for experimental and control samples.

V. 10 BIOMARKER SCREENING

Biomarker screening with SpecA works by first choosing spectral features, then analyzing them with a custom-made statistical workflow that applies a series of hypothesis tests to determine if samples come from different populations, in mean terms (Figure 13). The rationale behind this decision tree is provided in the supplementary materials.

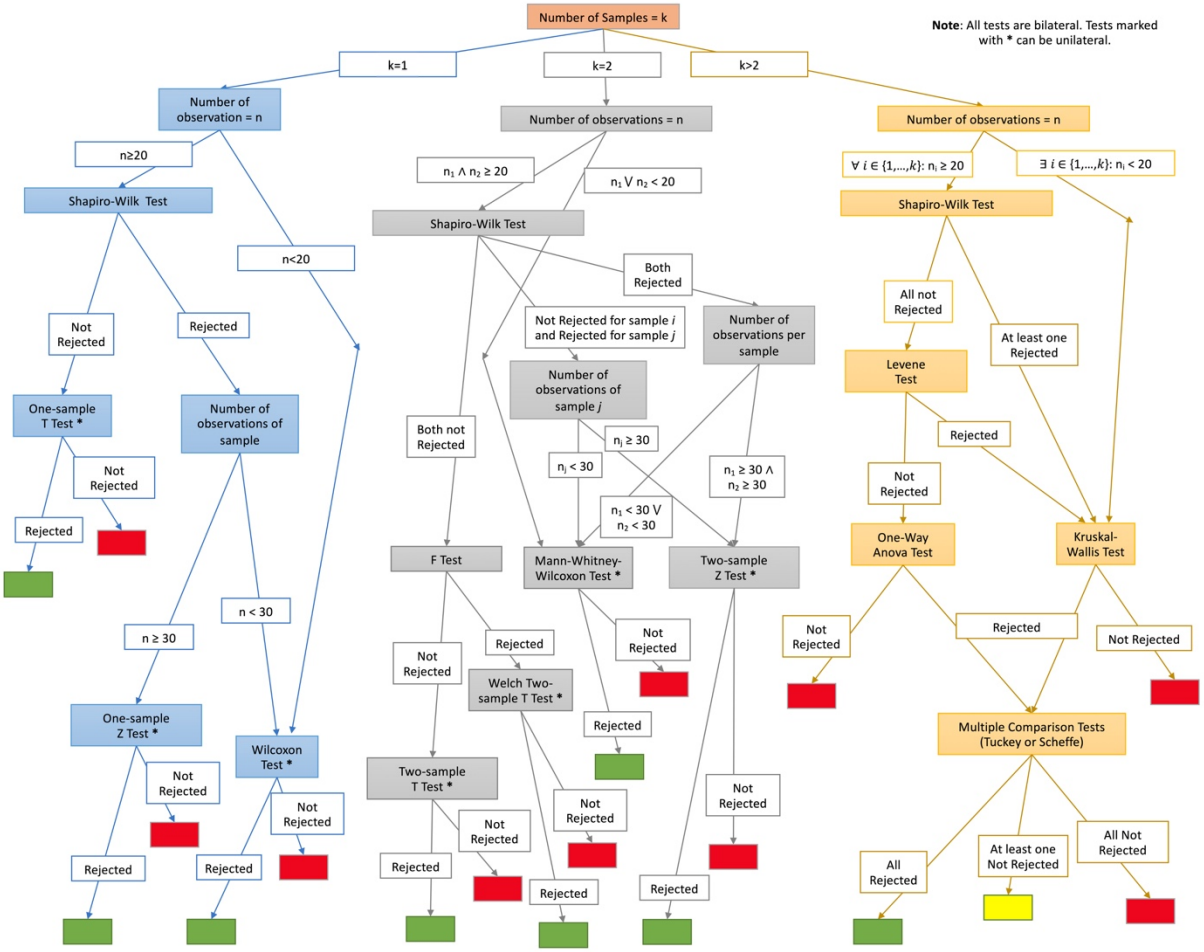


Figure 13. Decision tree of SpecAs' statistical workflow to identify spectral biomarkers.

For that, the user must navigate to the **Core Parameters** tab and select a **supervision variable**, which identifies samples according to their grouping; a **significance level**, to reject or not reject the hypothesis tested; a **band width**, for SpecA to search the datasets' closest wavenumber to the suggested band localization; and a **Monte-Carlo tolerance** for the computation of statistical tests (Figure 14). It is important to highlight that biomarkers can be spectral bands or ratios. Under the **Biomarker Management** tab, a series of **known spectral bands** with biological assignment are included in SpecA by default, and the user can quickly

select them for analysis, for which the **available notes** shown below the list boxes, are often helpful. In addition, the user can manually enter a new band for analysis. SpecA also includes **known spectral ratios** that are of biological interest, which the user can confirm for analysis as done with spectral bands. In addition to manual entry of spectral bands for analysis, SpecA also computes all combinations of confirmed spectral bands into ratios and adds these for analysis by clicking **Add ratios from bands**.

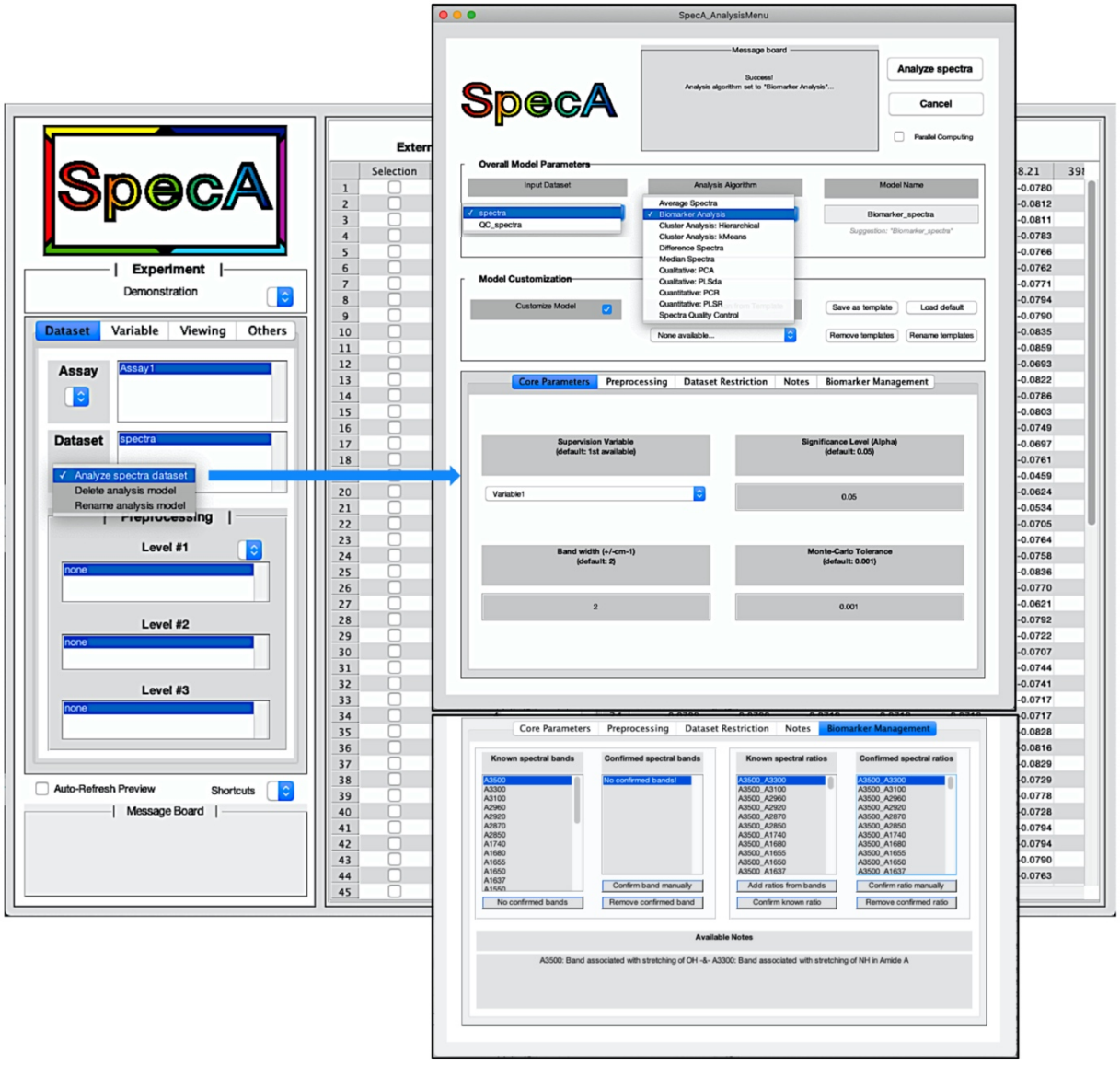


Figure 14. SpecAs' inputs for biomarker analysis.

In terms of results, at the center of a biomarker analysis is a table that briefly summarizes the p-values of the statistical tests applied (Figure 15). SpecA labels each putative biomarker (rows) as: **red**, if the hypothesis that samples come from different populations, in mean terms, is rejected; **yellow**, in the case of at least one hypothesis not being rejected in a multiple comparison test; or **green**, if the hypothesis is not rejected, in which case the band or

ratio is considered a biomarker, within the significance level chosen by the user. Below the table there is a tabbed panel that shows: the ratio or band value for each value of the supervision variable with a **box-and-whisker plot,** and a **bar plot**. Follows the usual **descriptive statistics**, decomposed into the constituents of the supervision variable; detailed information on the **hypothesis tests** applied, in particular the test statistic is often useful to scrutinize; and the **Biomarker Location**, where the original dataset reveals the patterns of the original dataset around the suggested location of the band or ratio.

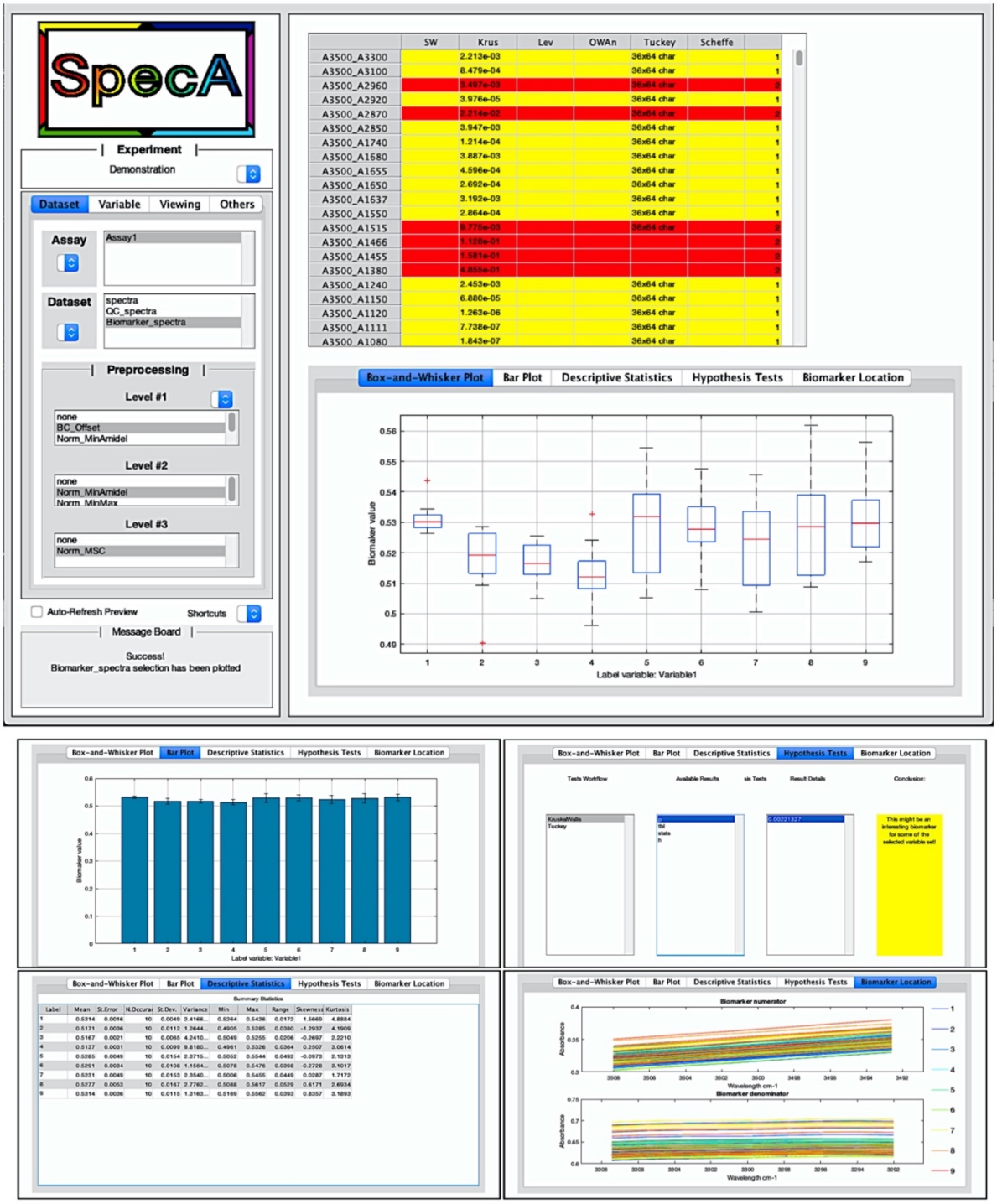


Figure 15. Set of outputs of a biomarker analysis.

V. 11 PRINCIPAL COMPONENT ANALYSIS

SpecA has three **PCA algorithms** (Figure 16), according to MATLABs' built-in function. In addition, the user is also queried for the **maximum number of principal components** (PCs) of the model, as well as a **grouping variable set** and **coloring variable set**. These variable sets are used to plot the data with different markers and colors, respectively.

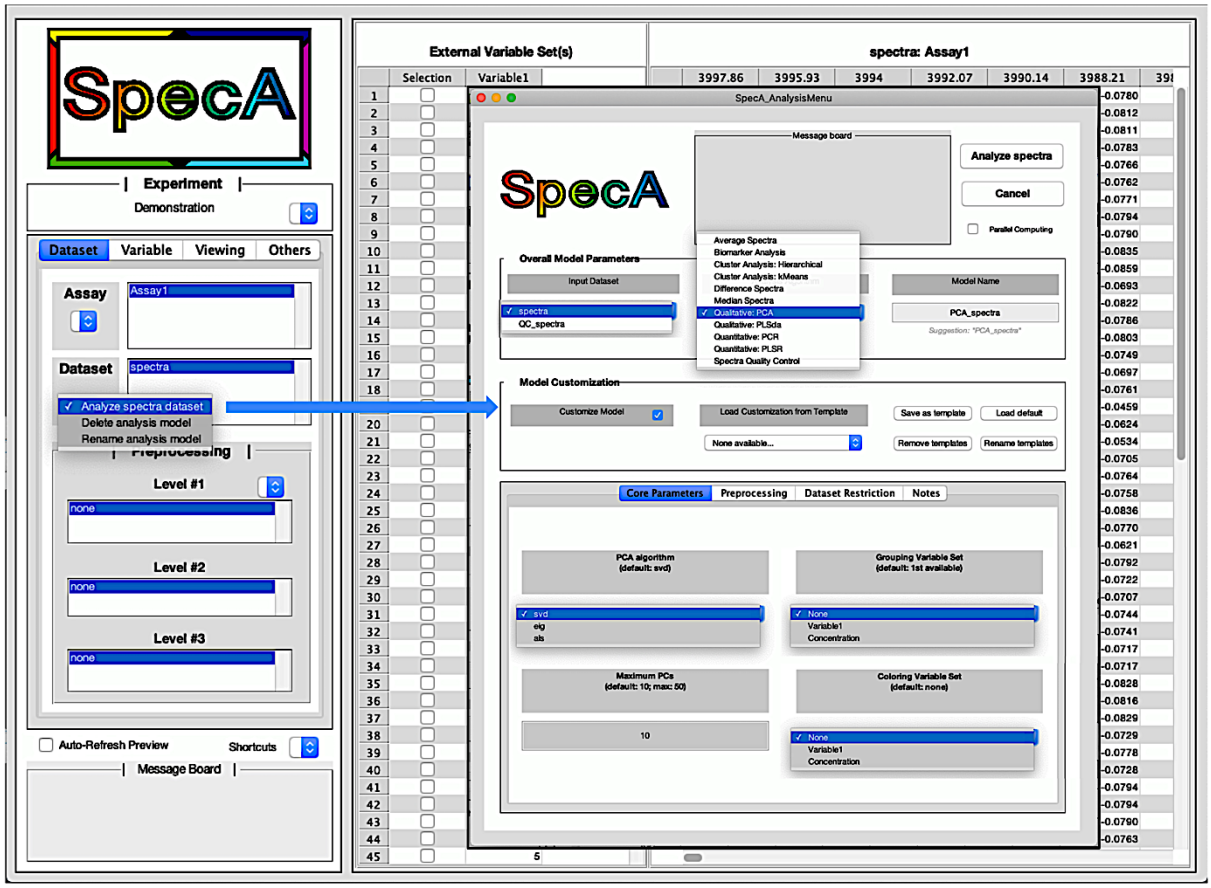


Figure 16. SpecAs' inputs for principal component analysis.

Typically, **3-D score plots** are difficult to visualize. However, because SpecA takes full advantage of MATLABs' figure handling functions, it allows the user to rotate and pan the axis to better visualize any underlying patterns (Figure 17). In addition, the more traditional **2D score plots** are also available, which can be configurable to show pair of PCs the user deems necessary (Figure 18). The **model coefficients** are an easy way to visualize the contribution of spectral regions to the sample separation on the PCA-defined space, which can be further decomposed for individual coefficients as shown. Both the **variance explained** and the **eigenvalues** grant the user with an easy appreciation of how increasing the number of components also increases how much variability in the original data is captured by the PCA model. Moreover, the average eigenvalue criterion (**AEC**) and corrected average eigenvalue criterion (**CAEC**) are non-arbitrary cutoffs to objectively select the optimal number of PCs to retain for a given dataset. For that, only PCs above said cutoffs should be included [2]. Lastly, the **influence plot** identifies outliers according to the Hotelling T^2 statistic and Q residuals, which in this case are shown for a 99% confidence interval. These parameters are often relevant for outlier detection [3].

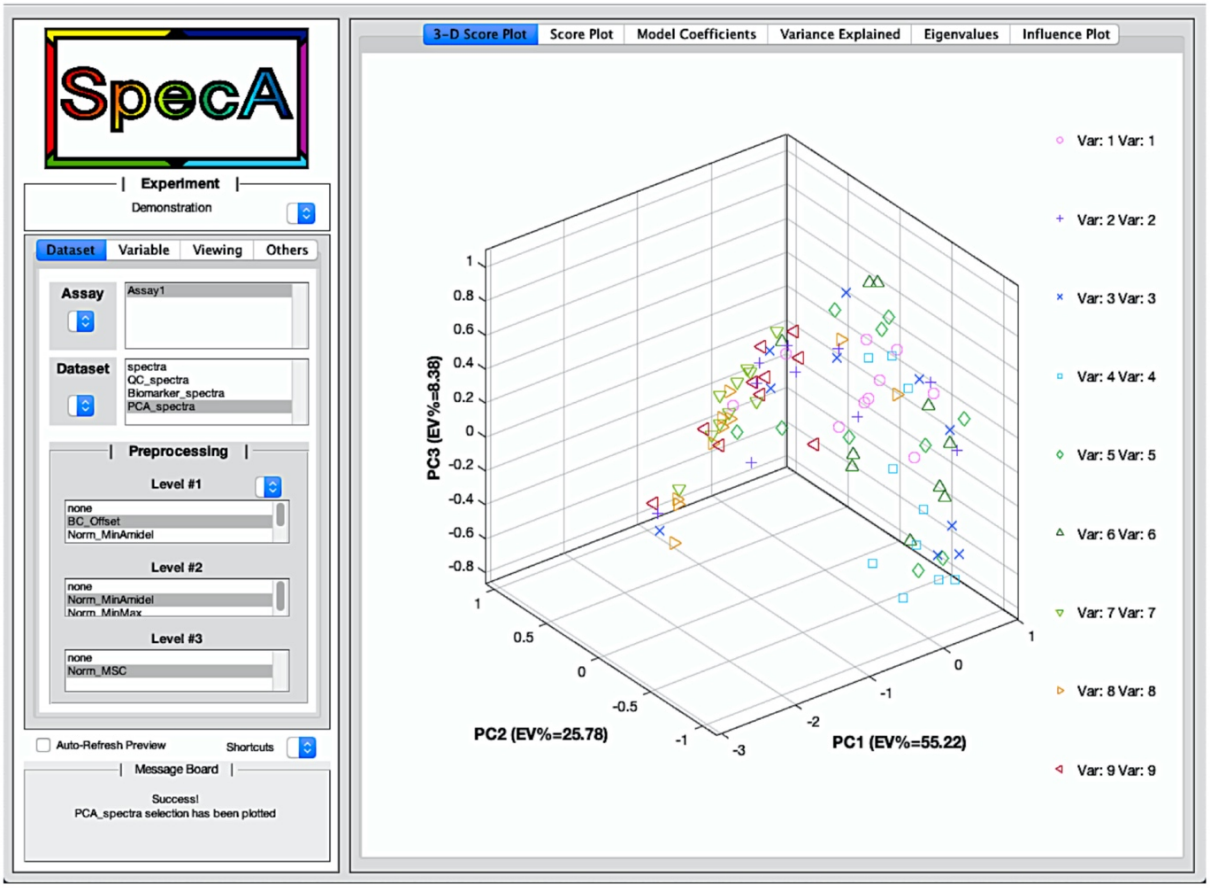


Figure 17. Set of outputs of a principal component analysis highlighting 3D score plots.

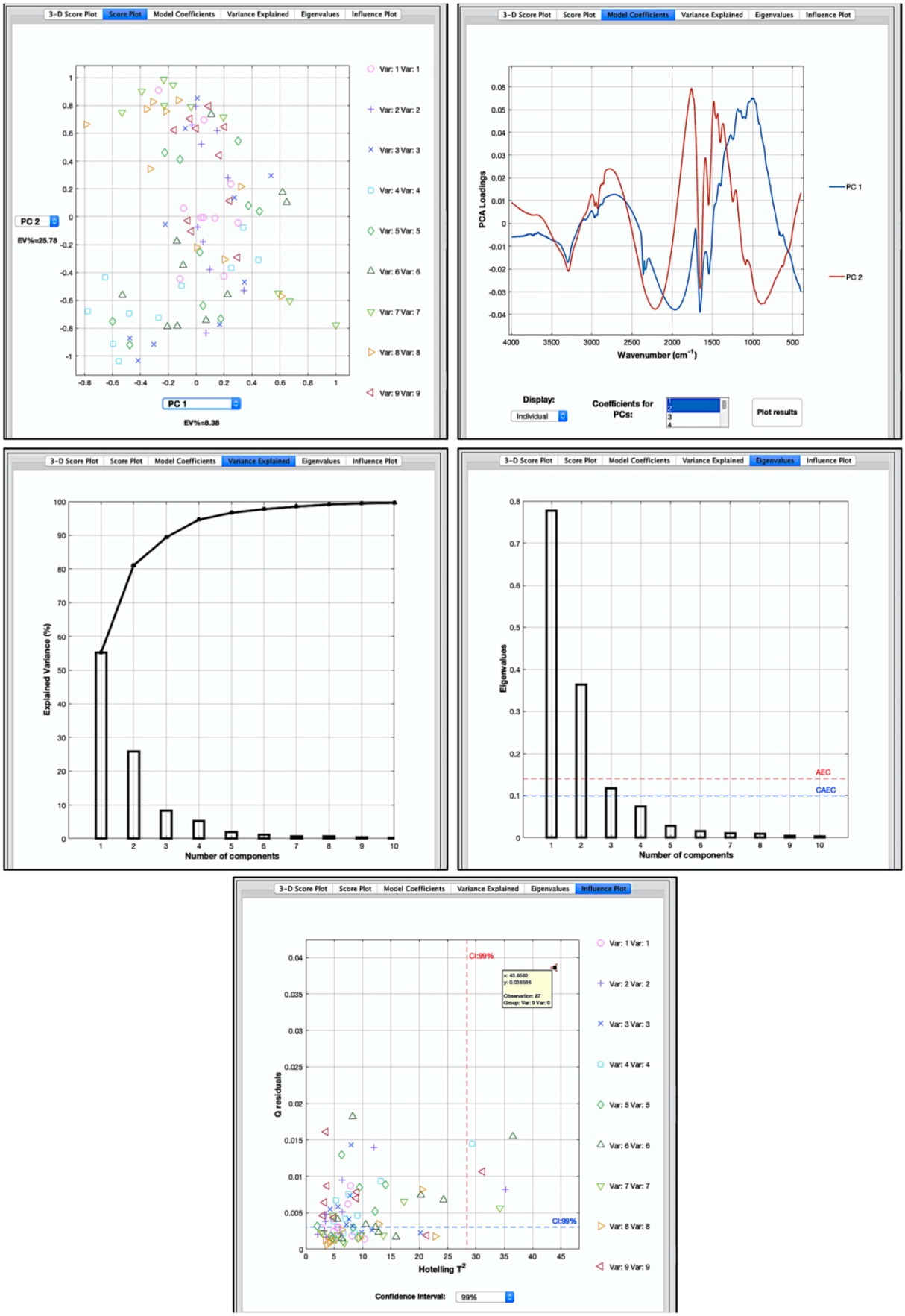


Figure 18. Set of outputs of a principal component analysis highlighting 2D score plots, model coefficients, variance explained, eigenvalues and influence plot.

V. 12 PRINCIPAL COMPONENT REGRESSION

Regarding principal component regression (PCR), SpecAs' inputs include the **PCA algorithm**, the **maximum number of PCs** and the **regression variable**, which is the numeric response variable to be modelled from the dataset. Because of its resemblance with PCA, no figure was shown exemplifying the available inputs for PCR. In terms of outputs, the **Observed Vs. Fitted** tab provides an easy overview of the PCR performance, for instance the goodness of fit is indicated by how close the samples are in relation to the $x = y$ diagonal, and heteroskedasticity is made evident by samples diverging more from the diagonal on one region of the plot more than others (Figure 19). The **regression vector** and the **variance explained** are identical to the **model coefficients** and **variance explained** of a PCA. The **prediction error** and R^2 presented are calculated after a leave-one-out cross-validation and are presented over a range of PCs (Figure 20), which helps in choosing an adequate number of PCs to avoid overfitting but ensure an adequate modelling of the data.

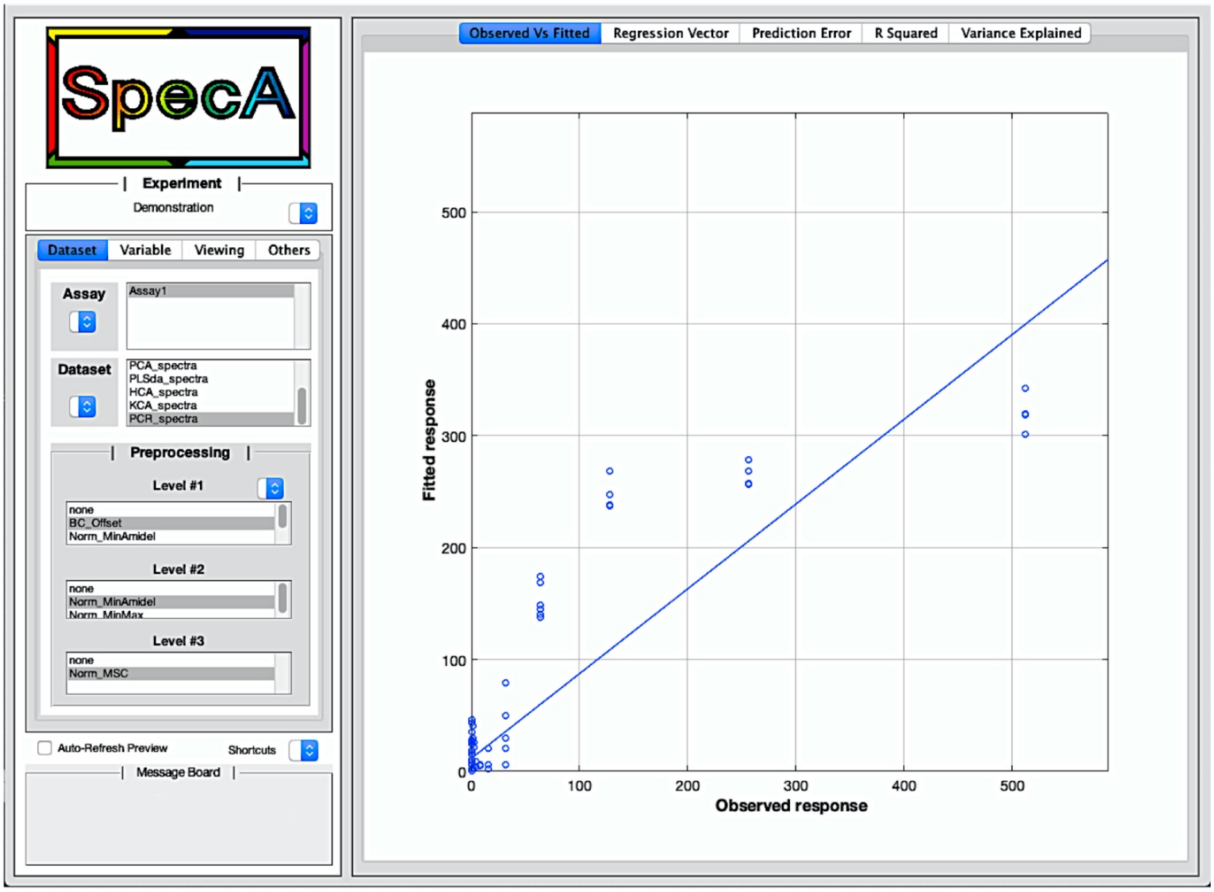


Figure 19. Set of outputs of a principal component regression highlighting the observed versus fitted response plot.

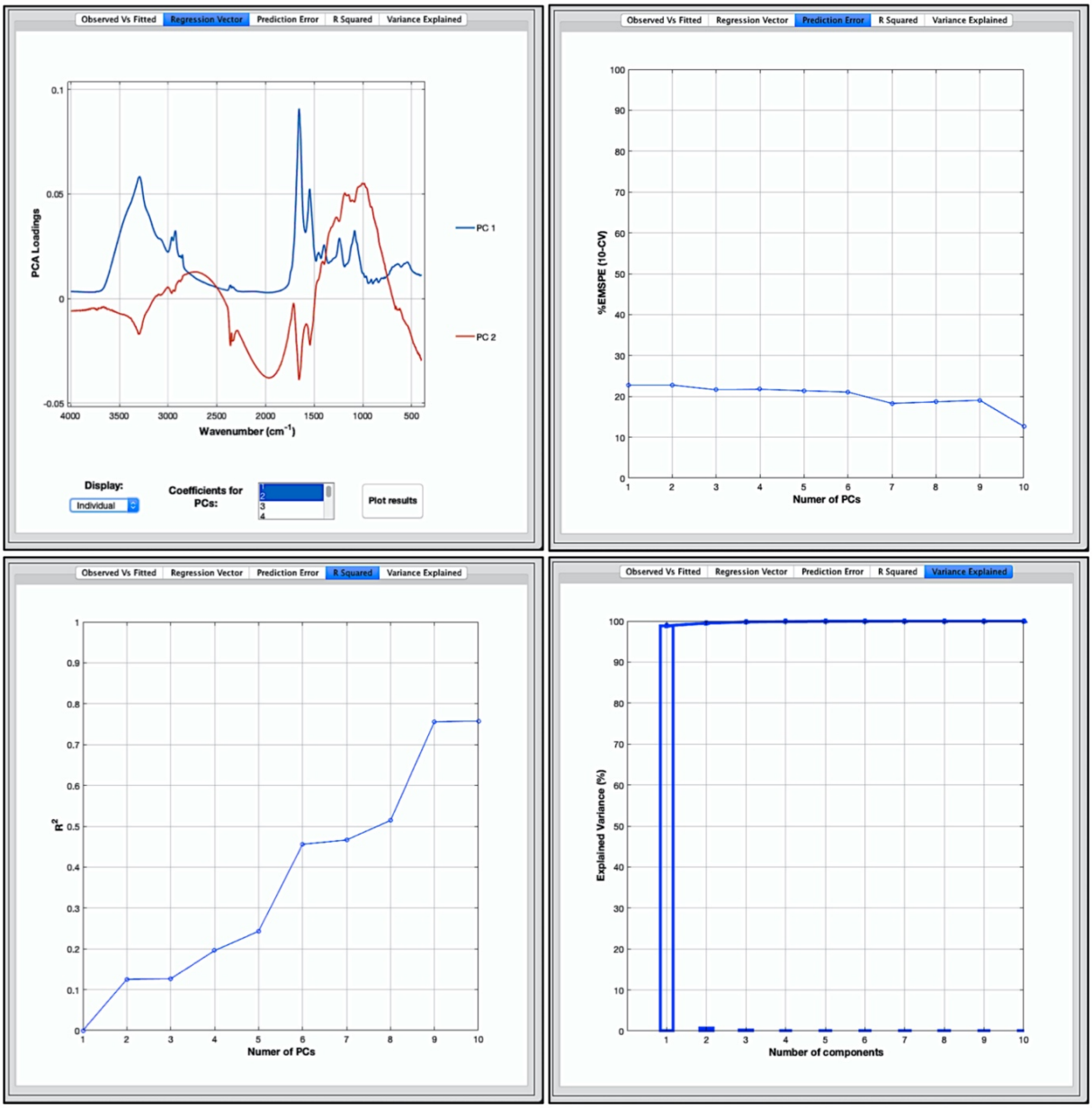


Figure 20. Set of outputs of a principal component regression highlighting the regression vector, which is equivalent to the model coefficients of a principal component analysis, cross-validated prediction error and R^2, along variance explained.

V. 13 PARTIAL LEAST SQUARES DISCRIMINANT ANALYSIS AND REGRESSION

To calculate a partial least squares (PLS) discriminant analysis (PLSDA), SpecA uses a community file uploaded to File Exchange by Cleiton A. Nunes from UFLA, Brazil. Unfortunately, this particular function is no longer publicly available, but it was kept in SpecA given it is still fully functional, and more importantly integrates cross-validation both in PLSDA and PLS regression (PLSR). Eventually, the work by Nunes et al. evolved into a standalone GUI called Chemoface [4], which is publicly available.

PLSDA has the same input requirements as PCA, except PLSDA requires a **supervision variable** that reflects class membership for prediction. This input appears where the algorithm selection was available for PCA. Additionally, in PLS the PCs are called latent variables (**LVs**), which is another required input for PLSDA.

In terms of outputs, PLSDA has the same number of outputs and these are mostly identical to PCA. Figure 21 highlights PLSDA that are different from those of a PCA. Particularly, PLSDA does not have an eigenvalue plot, but includes a **variable importance projection plot**; PLSDA substitutes the influence plot with one of **leverages versus residuals**; importantly, while the former can be used to remove uninformative spectral regions, the latter is helpful for outlier detection [5,6].

Regarding PLSR, SpecAs' inputs include the **maximum number of LVs** and the **regression variable**, which are identical to those of PCA; as are the outputs, so these were not presented nor discussed.

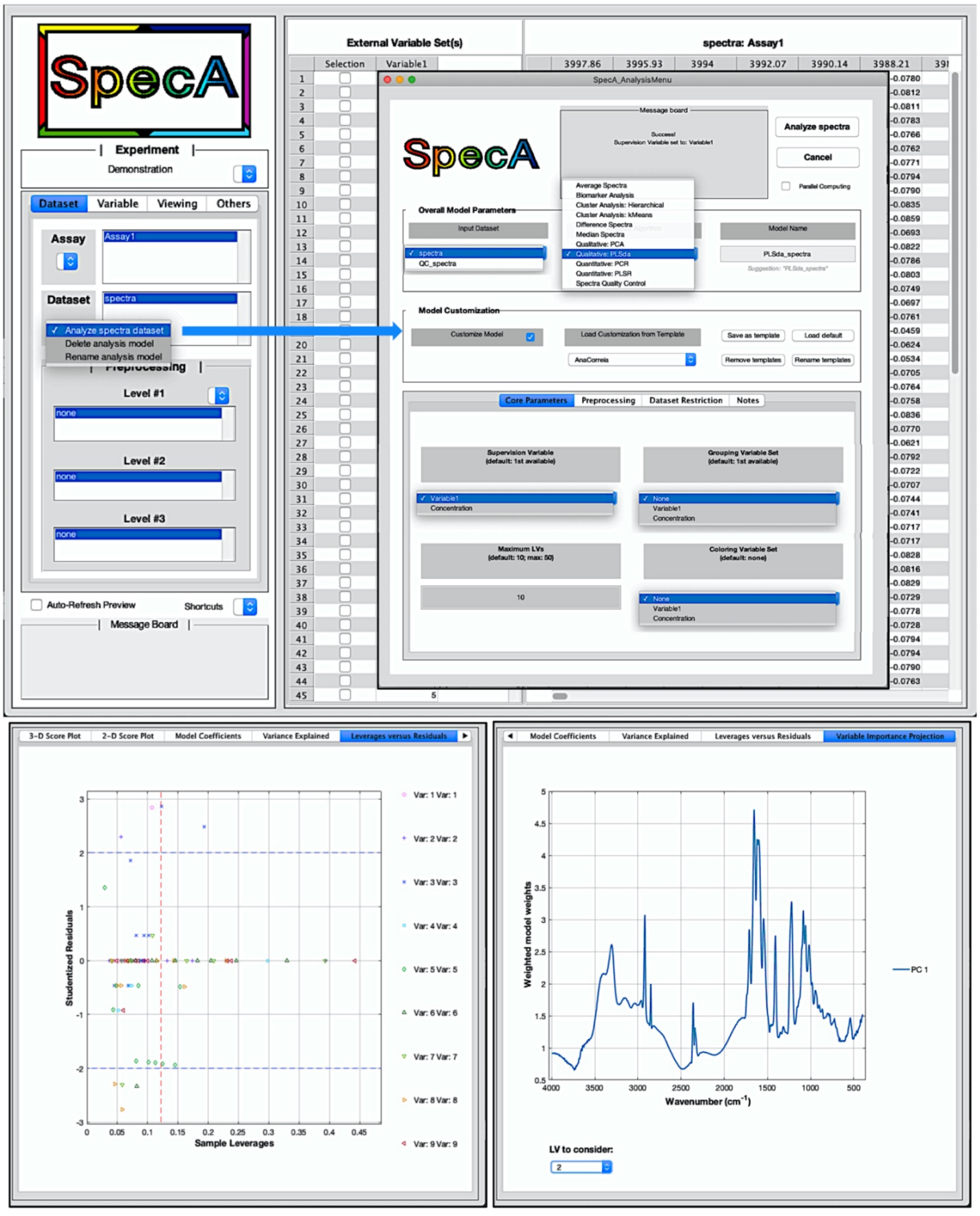


Figure 21. SpecAs' inputs and outputs exclusive of partial least squares discriminant analysis.

V. 14 HIERARCHICAL CLUSTER ANALYSIS

Hierarchical cluster analysis (HCA) requires a different set of inputs (Figure 22). Firstly, a **label variable** is required to identify individual samples on both the dendrogram and the silhouette plots. HCA also involves a **distance algorithm** and a **linkage algorithm**. These can be singular, e.g., **Average**; or **Multiple**, e.g., all available. The user is also asked for the **maximum number of clusters**, which is useful to limit the maximum number of clusters from which the optimal number of clusters is identified. Choosing a number of clusters forces a horizontal slice on the dendrogram that prunes the tree into a user-defined number of clusters, which may or may not be related to the natural divisions of the data. Opposingly, to consider the natural divisions of the data, the user can specify cluster boundaries that are not a horizontal slice on the dendrogram. For that, the user chooses an **inconsistency cut-off**, where higher cutoff values effectively allow clusters to contain more dissimilar samples within, resulting in a smaller number of clusters. The **dendrogram color threshold** controls the linkage below which groups of nodes are colored in unique colors and can help in visualization.

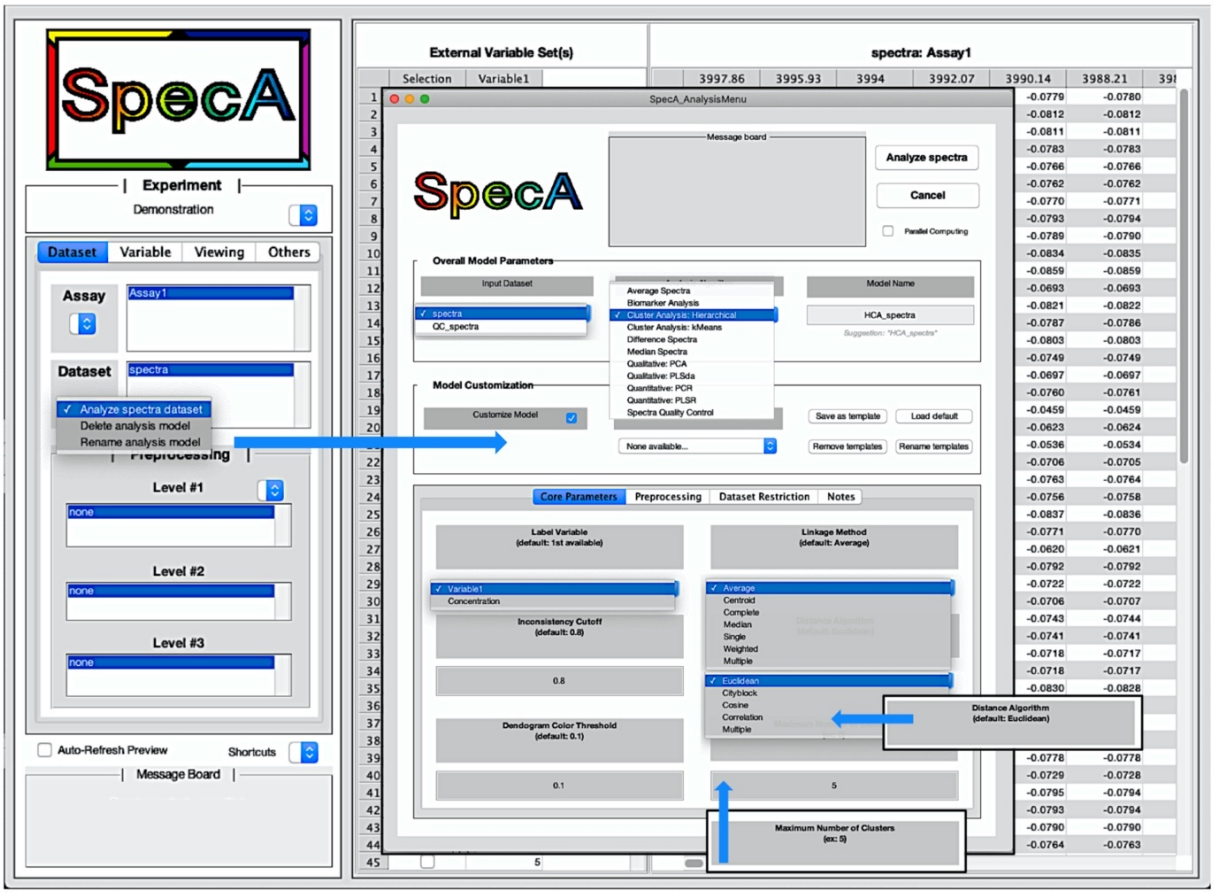


Figure 22. SpecAs' inputs for hierarchical cluster analysis.

If multiple **distances** and/or **linkages** are chosen, it is possible to scroll through those selected for analysis as can be done with preprocessing combinations, datasets or assays. For that, the user should navigate to the **Variable** tab of the control panel, and under **External Variables**, choose either **distance** or **linkage** under the list box named **Variable loaded,** and the particular metric desired for visualization on the list box below under **Details**.

Regarding the outputs of a HCA, the most common is a **Dendrogram** (Figure 23). Here, sample label font size is adjusted according to the number of samples being plotted and can be interactively zoomed in with the mouse. It is also possible to select a data point to rapidly view the height of a given node. The **Cophenet** tab (Figure 24) provides the user with a quick comparison of the cophenetic correlation coefficient, which is higher when the correlation between the linking of objects in the tree is strongly correlated with the distances between objects. Because this parameter is calculated for all combinations of **linkage** and **distance** chosen for analysis, this translates to an objective method to choose the optimal combination of these parameters.

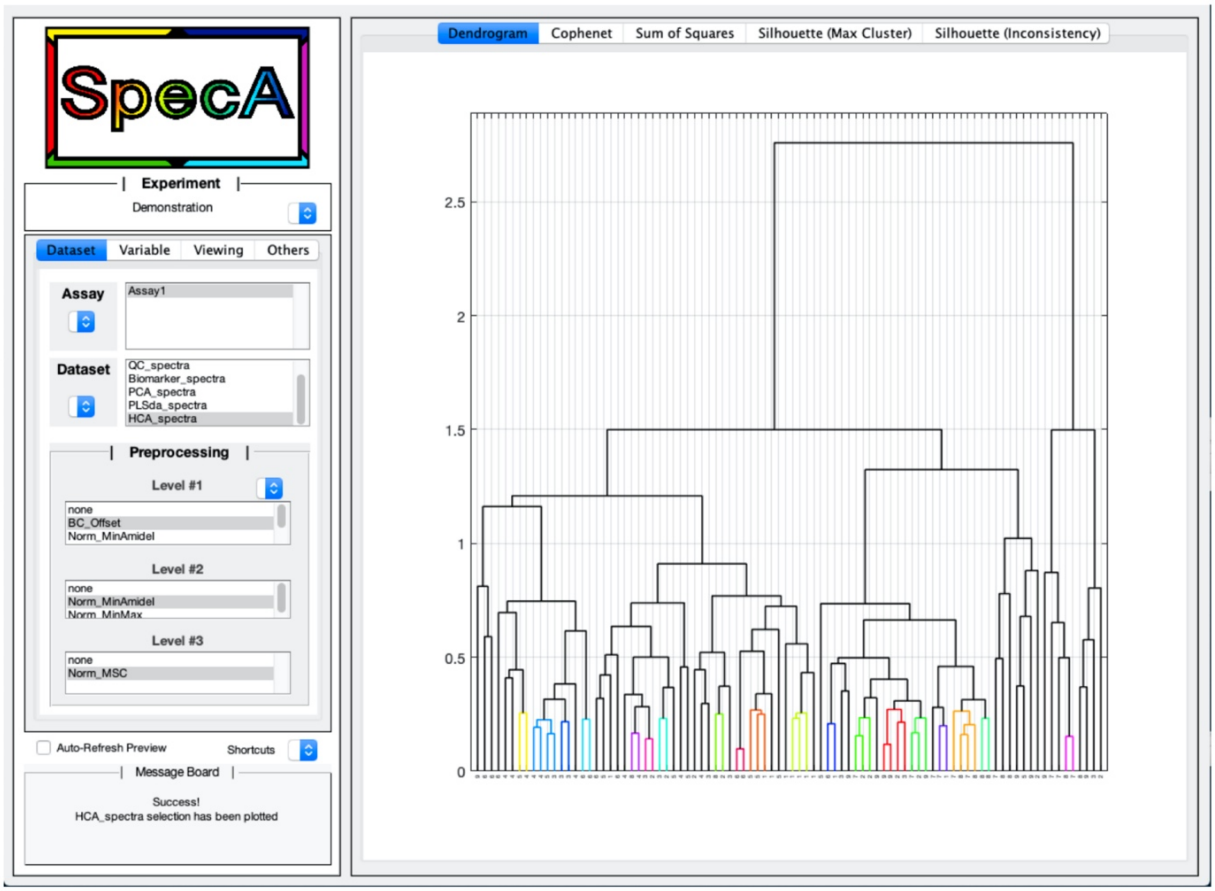


Figure 23. Set of outputs of a hierarchical cluster analysis highlighting a dendrogram.

The **Sum of Squares** tab shows how the within-cluster sum-of-squares performs over a range of clusters, broken down into the combinations of **linkage** and **distance**. When this value is lower, there is lower variability of the samples within each cluster, so the clusters are more 'compact', and the underlying **linkage** and **distance** is seen as 'better'.

To evaluate the number of clusters obtained by pruning the tree with a user-defined number of clusters, or those resulting from an inconsistency cut-off, SpecA presenta a **Silhouette (Max Cluster)** and a **Silhouette (Inconsistency)** respectively.

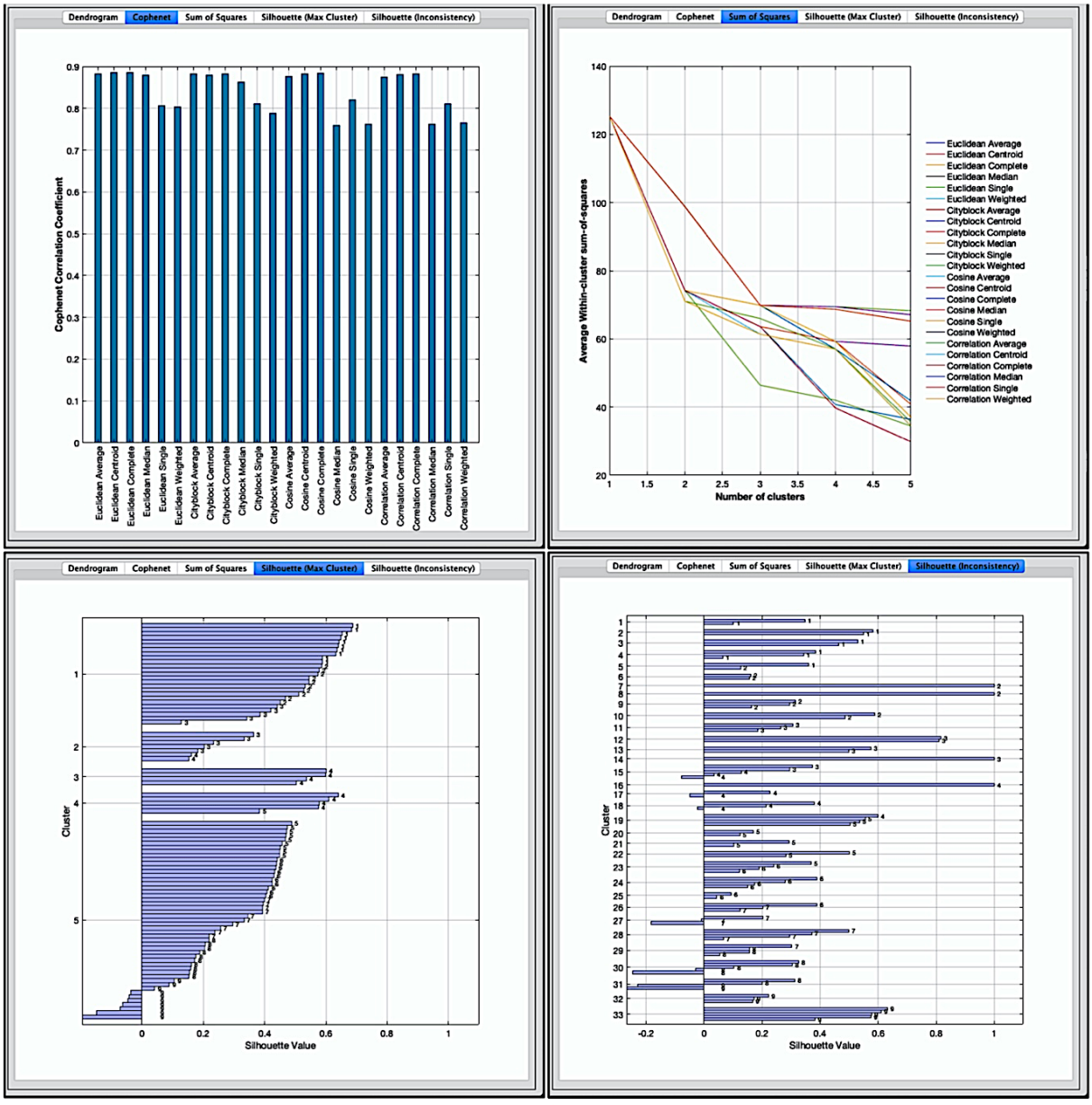


Figure 24. Set of outputs of a hierarchical cluster analysis highlighting the cophenetic correlation coefficient, the within-cluster sum-of-squares, and silhouette plots calculated with a user-defined number of clusters and an inconsistency cut-off.

V. 15 *K*-MEANS CLUSTERING

K-means cluster analysis, referred to as KCA within SpecA, requires the user to define a **label variable,** to identify individual samples on the silhouette plot; a **maximum number of clusters**, which is used to not only build the *K*-means model, but up to which the performance indicators are calculated; and a **distance** metric (Figure 25). Additionally, KCA asks for a number of **iterations**, which is typically the highest that can be computationally afforded so the algorithm converges; and a number of **replicates**, which is the number of times the calculations are repeated with different initial positions, to avoid local minima. SpecA presents the KCA model with lowest within-cluster sums of point-to-centroid distances.

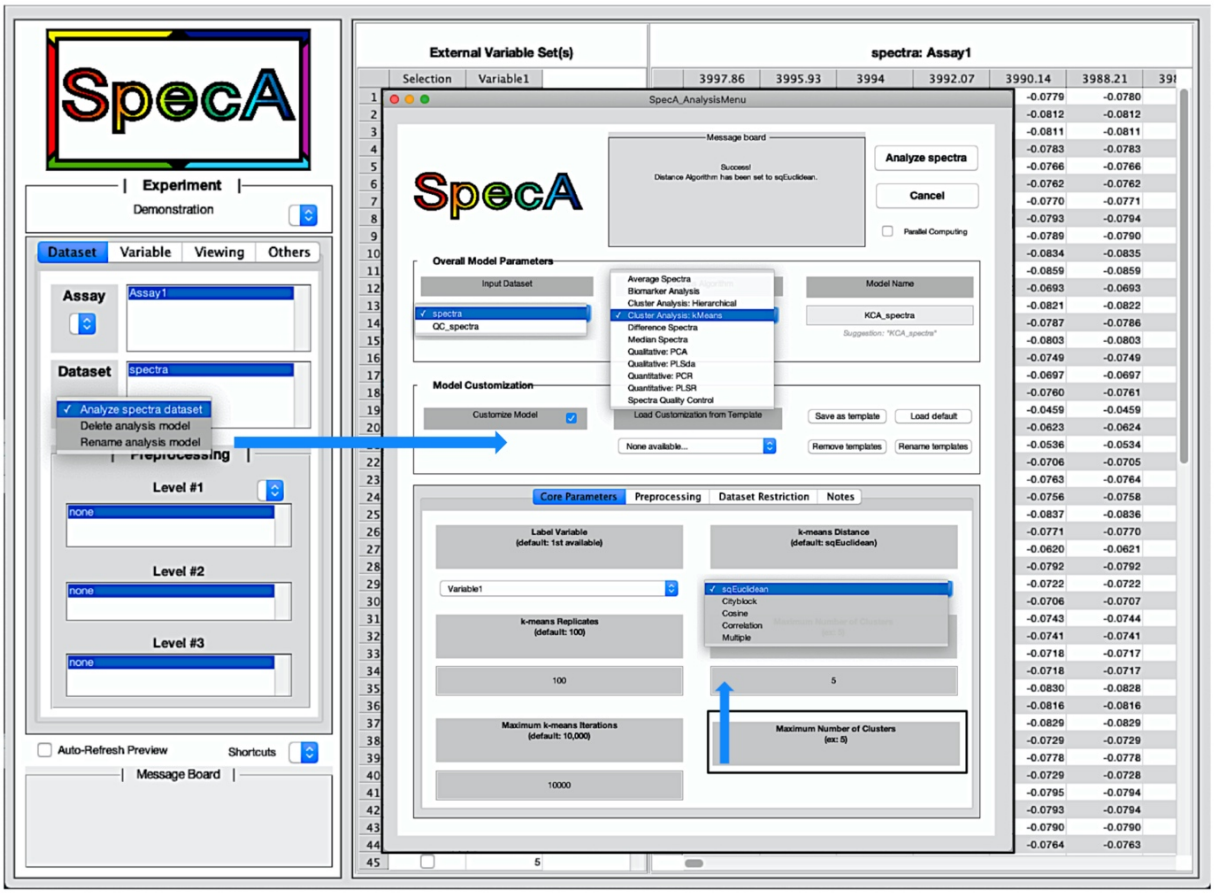


Figure 25. SpecAs' inputs for a *K*-means cluster analysis.

KCA has two outputs, a silhouette plot and performance indicators (Figure 26). By default, the silhouette is plotted with the maximum number of clusters, but this is customizable in the **Variable** tab, under **External Variables**, similarly to the distance metric. The performance indicators presented are the **Silhouette width values**, the **Davies-Bouldin index**, and the **Calinski-Harabasz index**, which have been labeled as to whether these should be maximized or minimized for those who are less familiar. Moreover, these performance indicators are shown for all the distance metrics, and for up to the maximum number of clusters, chosen by the user. As such, these performance indicators reveal not only the optimal distance metric, but also the number of clusters that most likely reflects the natural divisions of the data

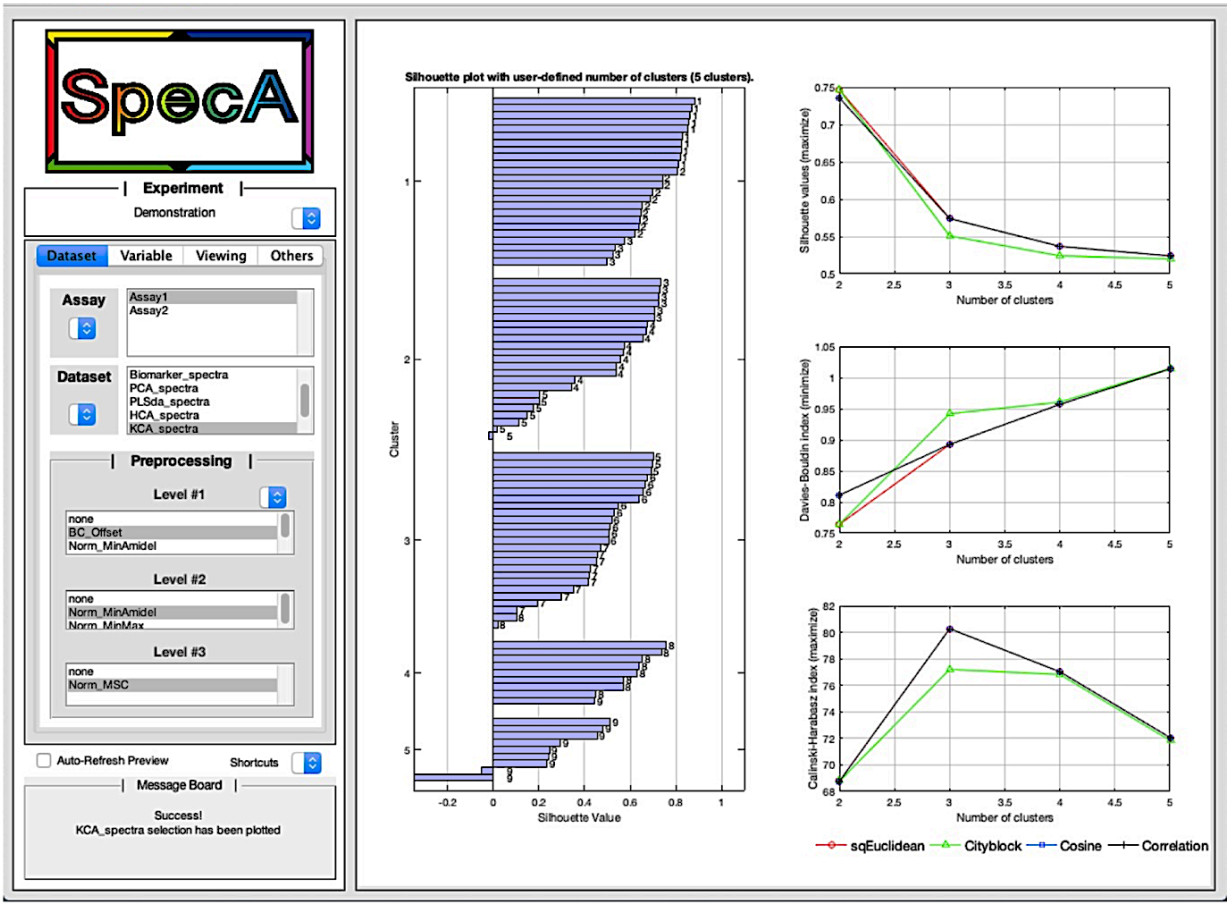


Figure 26. SpecAs outputs for a *K*-means cluster analysis.

V. 16 GENERATING OUTPUTS

Upon completion of an analysis, further actions are often necessary to finalize an experiment, for which SpecA has implemented a series of outputs generation actions that can be conveniently accessed under the **Others** tab of the control panel (Figure 26).

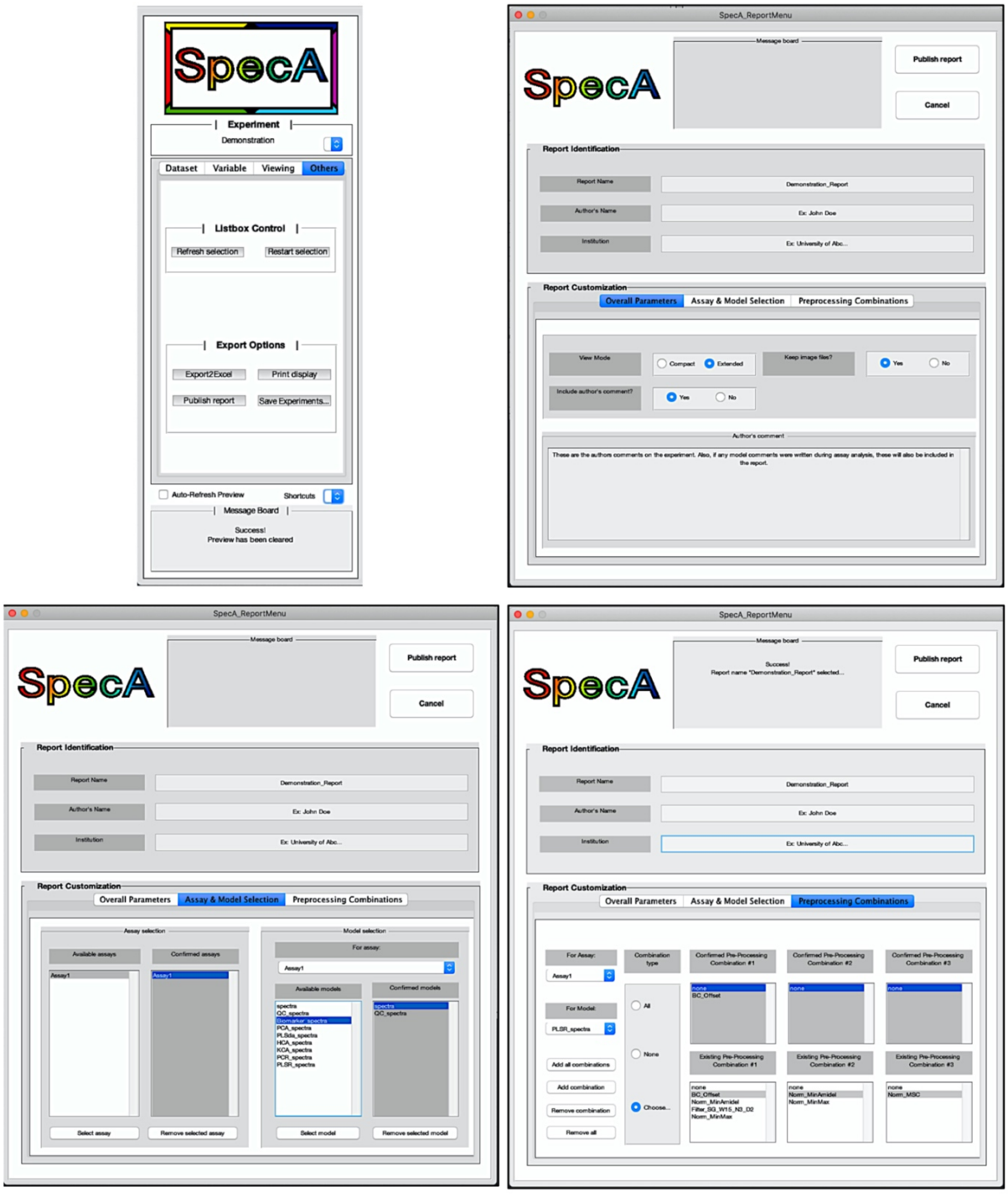


Figure 27. SpecAs' output tab of the control panel and sub-GUI for generating reports.

Firstly, SpecA generates high-resolution images, which researchers often require for publication, by clicking on **Print display**. This calls upon a File Exchange submission by Yair Altman [7], which can be highly customized to achieve the desired specifications. Additionally, it could be the case that a set, or sub-set, of an experiment is to be compiled into a single report. This might be useful to keep a record of what analyses have been done, or to share your findings with a colleague, for instance. For that, SpecA has a sub-GUI that appears by clicking on **Publish Report**, which automates this process (Figure 27). Here the user can identify the report by adding a **Report Name**, an **Author Name** and an **Institution**. Then, the report can be customized by choosing the **View Mode**, which controls if the images appear in an extended or compact format. The user can choose to **keep the image files** after the report is published or delete them. The user can **include author comments** on the report, which can be a description of the experiment conducted, for instance. It is noteworthy that any comments added when a dataset is analyzed also appear on the report. Lastly, several list boxes, popup-menus and buttons were included so the user can select which assays are to be published, and within the datasets and preprocessing combinations. The report is generated as a word file using a template that can be easily customized by the user.

V. 17 CONCLUSION

What started as a simple graphical user interface, with a few buttons that automated the application of preprocessing routines, eventually became a tool for most of the data analysis pipeline of an infrared spectroscopy experiment. Since early in the development of SpecA, the goal was to facilitate the data analysis component of the research being conducted, and more specifically, to automate the most repetitive tasks so that what seemed impossible became standard practice. For example, applying different combinations of preprocessing was challenging at first; evaluating their effect on a model more so; and currently this is a mundane occurrence. In fact, it is now normal to apply a multitude of preprocessing combinations and parametrizations and immediately gauge their effect on the following datasets: PCA of preprocessed spectra; PCA of difference preprocessed spectra; HCA after PCA of preprocessed spectra; HCA after PCA of difference preprocessed spectra; and for the last two datasets multiple distance and linkage metrics are compared. What initially was simply inconceivable given the sheer volume of data, it is now standard practice just to evaluate the quality of an experiment, and sometimes of a day's work. This not only provides near real-time feedback on the quality of the laboratory work, but also ensures the data analysis component of the research is thoroughly optimized.

An extremely positive side-effect of developing SpecA was that my proficiency with MATLAB, and coding in general, greatly improved, which very strongly reflected on all the data analysis that has been conducted and is presented on the following chapters of this thesis. There is no doubt that without SpecA, many of the experiments conducted throughout this Ph.D. would not have been as successful, if at all. SpecA has not been a secondary objective of this Ph.D., but rather it has been the powerhouse that catalyzed the scientific quality of this Ph.D.

V. 18 Acknowledgements

I would like to acknowledge the contribution of Tomas Alegra, to the design and logo of SpecA; of Hugo Correia, to the lengthy discussions regarding many decisions undertaken during SpecAs' creation process; of Sandra M. Aleixo, for her guidance while implementing the biomarker toolbox; and also, of many others who participated during the creation of SpecA, namely Pedro Sampaio, Marta Lopes, Filipa Rosa, Kevin Sales, among others.

V. 19 Supplementary Material

There is a general agreement among most researchers that quantitative data should be evaluated with the appropriate parametric tests when the respective assumptions are verified, because generally these tests are more powerful than its nonparametric counterparts. This means that they are more likely to detect true differences or relationships. However, if at least one of the assumptions are violated, this power advantage may be negated [8,9].

Some researchers argue that the necessary transformation of quantitative data into an ordinal rank order or categorical or nominal format, in order to use a nonparametric test, leads to a loss of information of the data. Because of this, they still think that is more prudent to employ the appropriate parametric test even one or more of its assumption have been violated, arguing that most parametric tests are robust. Others researchers claim that many times, parametric and nonparametric tests employed to evaluate the same set of data lead to identical or similar conclusions [8]. Several researchers of different areas have revealed their concerns about the incorrect use of parametric tests, claiming that if the assumptions are not met properly then the parametric tests may provide results that are not statistically valid and consequently they can lead to inaccurate conclusions [10–14]. We believe that a correct use of suitable tests will provide more accurate conclusions. Thus, automatic workflow to

distinguish two or more populations, in mean terms, by applying suitable hypothesis tests to the corresponding independent groups of observations was created for biomarker screening. The differences between populations are considered statistically significant if the p-value is less than or equal to the significant level of 5%.

In order to construct a credible and feasible automatic workflow, we first had to make two decisions: first, the minimum sample size to be used for normality tests had to be defined so that they have an acceptable reasonable power, to actually make sense to use them to assess whether the sample comes from a population with a normal distribution; and, which normality test should be used, in order to be more powerful considering different alternative distributions.

The literature is not consensual nor for the choice of the minimum sample size to be used for normality tests so that they can be effective, nor which test to use. Some argue that for small samples ($n < 20$) the normality tests are unlikely to detect non-normality. Also, several simulations studies had been carried out to compare different normality tests for some sample sizes and considering different types of alternative symmetric and asymmetric distributions, where test powers were very low for small samples, and even for $n = 20$, but this was the turning point where tests start to be reasonable for several situations. The Shapiro Wilk normality test performs well in many cases, having a good power in several situations and becoming the most powerful in some situations [15–17].

Ghamesi and Zahediasl [13] refer some normality tests, and discuss their importance to the validity of parametric tests. In their view, one of the most common normality tests, the Kolmogorov-Smirnov after Lilliefors correction, should no longer be used owing to it low power. They further argue that to assess normality both visually and through a normality test are preferable. For the latter, a Shapiro-Wilk test is highly recommended for several researchers [18]. Originally this test was restricted to $n < 50$, but after Roystons' modifications, an improved approximation to the weights result in an algorithm which can be used for $3 < n < 5000$ [19].

Of course, checking normality visually using graphics like histograms or steam-and-leaf plot (for large samples), boxplot, probability-probability plot and quantile-quantile plot (the last two more so for small samples), would be helpful to aid to take the right conclusion about the distribution of a variable [10,13,17] but this is not compatible with an automatic scheme. Attending to the mentioned studies, we opted to apply the Shapiro-Wilk test (adjusted by Royston) to evaluate normality for samples where $n \geq 20$, considering that at this size, the

powers obtained by simulation, although weak became reasonable. Indeed, for small samples ($n < 20$) the normality tests are unlikely to detect non-normality, and consequently this assumption for parametric tests cannot be verified, so, in these cases, nonparametric tests will be preferable.

Regarding the automated workflow for biomarker screening: all parametric and non-parametric tests involved in this workflow have been described elsewhere [8]. First, we are going to explain the statistical methodology applied to distinguish two independent samples, in mean terms. Here, f at least one sample is small ($n < 20$), the non-parametric Mann-Whitney-Wilcoxon test was applied. Otherwise, the Shapiro-Wilk normality test was used to verify if each one of the two samples came from a population with normal distribution. Then several situations may occur:

If both samples come from populations with normal distribution, parametric T tests were used. In order to choose the suitable T test to compare the two populational mean values, a F test was first applied to evaluate the populational equality of variances, which determines the subsequent type of T-test to use: a two sample T-test if the population variances are equal; or the Welsh T test otherwise.

If both samples do not come from populations with normal distribution, but both have size at least 30, then the parametric Z test was applied, because according to the Central Limit Theorem (CLT) the distribution of each of the sample mean estimator is approximately normal and consequently the difference between the two sample means estimators is a random variable which follows an approximately normal distribution.

If both samples do not come from populations with normal distribution, but both have size less than 30, then we should apply the non-parametric Mann-Whitney-Wilcoxon test.

If one of the samples comes from a population with normal distribution (sample i) and the other does not come from a population with normal distribution (sample j), two situations may occur, depending on the dimension of the sample that does not come from a population with normal distribution (n_j): If $n_j < 30$, the non-parametric Mann-Whitney-Wilcoxon test was used; otherwise, according to the CLT we can apply the parametric Z test.

In relation to the statistical methodology applied to distinguish more than two independent samples, in mean terms, if at least one of the samples is a small sample ($n < 20$)

then the non-parametric Kruskal-Wallis test was applied because normality tests may not properly assess the assumption of normality in this case. Then, if all the samples have size at least equal to 20 the Shapiro-Wilk test was used to evaluate if each sample came from a population with normal distribution. Two situations may occur:

If at least one of the samples did not come from a population with normal distribution, the non-parametric Kruskal-Wallis test was used; or if all samples came from populations with normal distributions, the normality assumption of the parametric one-way ANOVA test is satisfied. In this case, a Levene's test was used to evaluate the other assumption of the one-way ANOVA called homoscedasticity, where two situations may happen:

If the assumption of equality of population variances is satisfied, then we can apply the parametric one-way ANOVA test to evaluate equality of the population means values. If the homoscedasticity assumption is violated, then the non-parametric Kruskal-Wallis test was applied to evaluate the equality of the population central locations.

As a side note, some parametric procedures were developed to deal with heterogeneity of variance and could be applied instead of the non-parametric Kruskal-Wallis test, for instance the Welch One-way ANOVA. However, Keppel in 1991 noted the Welch One-way ANOVA is not acceptable for more than four samples ($k > 4$). Similarly, Moder [20] argued for the unsuitability of this test for more than two or three samples. These observations render this test incompatible with our workflow, given it was built to be as generalist as possible.

In any of the parametric or non-parametric tests used to assess the equality of more than two populations in mean terms, if this hypothesis is rejected, multiple comparison tests were applied to evaluate which populations differ from the others.

V. 20 REFERENCES

1. Lasch, P. Spectral pre-processing for biomedical vibrational spectroscopy and microspectroscopic imaging. *Chemom. Intell. Lab. Syst.* **2012**, *117*, 100–114.
2. Ballabio, D. A MATLAB toolbox for Principal Component Analysis and unsupervised exploration of data structure. *Chemom. Intell. Lab. Syst.* **2015**, *149*, 1–9.
3. Römer, M.; Heinämäki, J.; Strachan, C.; Sandler, N.; Yliruusi, J. Prediction of tablet film-coating thickness using a rotating plate coating system and NIR spectroscopy. *AAPS PharmSciTech* **2008**, *9*, 1047–1053.
4. Nunes, C.A.; Freitas, M.P.; Pinheiro, A.C.M.; Bastos, S.C. Chemoface: A novel free user-friendly interface for chemometrics. *J. Braz. Chem. Soc.* **2012**, *23*, 2003–2010.
5. Tran, T.N.; Afanador, N.L.; Buydens, L.M.C.; Blanchet, L. Interpretation of variable importance in Partial Least Squares with Significance Multivariate Correlation (sMC). *Chemom. Intell. Lab. Syst.* **2014**, *138*, 153–160.
6. Ballabio, D.; Consonni, V. Classification tools in chemistry. Part 1: Linear models. PLS-DA. *Anal. Methods* **2013**, *5*, 3790–3798.
7. Altman, Y. export_fig Available online: https://github.com/altmany/export_fig/releases/tag/v3.14.
8. Kafadar, K.; Sheskin, D.J. Handbook of Parametric and Nonparametric Statistical Procedures. *Am. Stat.* **2006**, *51*, 374.
9. Marshall, E. The Statistics Tutor's Quick Guide to Commonly Used Statistical Tests. *Statstutor Community Proj.* **2016**.
10. Chan, Y.H. Biostatistics 102 : Quantitative Data – Parametric. *Significance* **2003**, *44*, 391–396.
11. Lang, T. Twenty statistical errors even you can find in biomedical research articles. *Croat. Med. J.* **2004**, *45*, 361–370.
12. Neideen, T.; Brasel, K. Understanding Statistical Tests. *J. Surg. Educ.* **2007**, *64*, 93–96.
13. Ghasemi, A.; Zahediasl, S. Normality tests for statistical analysis: A guide for non-statisticians. *Int. J. Endocrinol. Metab.* **2012**, *10*, 486–489.
14. Kaur, A.; Kumar, R. Comparative Analysis of Parametric and Non-Parametric Tests. *J. Comput. Math. Sci. | www.compmath-journal.org* **2015**, *6*, 336–342.
15. Romão, X.; Delgado, R.; Costa, A. An empirical power comparison of univariate goodness-of-fit tests for normality. *J. Stat. Comput. Simul.* **2010**, *80*, 545–591.
16. Noughabi, H.A.; Arghami, N.R. Monte carlo comparison of seven normality tests. *J. Stat. Comput. Simul.* **2011**, *81*, 965–972.
17. Yap, B.W.; Sim, C.H. Comparisons of various types of normality tests. *J. Stat. Comput. Simul.* **2011**, *81*, 2141–2155.
18. Thode, H.C. *Testing For Normality*; CRC Press, 2002; ISBN 9780429213250.
19. Royston, P. A Remark on Algorithm as 181: The W-Test for Normality. *J. R. Stat. Soc. Ser. C (Applied Stat.* **1995**, *44*, 547–551.
20. Moder, K. Alternatives to F-Test in One Way ANOVA in case of heterogeneity of variances (a simulation study). *Psychol. Test Assess. Model.* **2010**, *52*, 343–353.

Chapter VI

Phenotypic screening of stress and antibiotic responses

This chapter is adapted from the manuscript:

Ribeiro da Cunha, B., Fonseca, L. P. and Calado, C. R. C. (2019) 'A phenotypic screening bioassay for *Escherichia coli* stress and antibiotic responses based on Fourier-transform infrared (FTIR) spectroscopy and multivariate analysis', *Journal of Applied Microbiology*, 127(6), pp. 1776–1789. doi: 10.1111/jam.14429.

Author contribution

Bernardo Ribeiro da Cunha conceptualized the study methodology, conducted the investigation and formal analysis, prepared the original draft, reviewed and edited its final version.

Abstract

We aimed to develop and optimize a Fourier-transform infrared spectroscopy (FTIRS) phenotypic screening bioassay for stress responses, regarding the effect of nutrient content, bacterial growth phase and stress agent exposure time. For that, a high-throughput FTIRS bioassay was developed to distinguish the Escherichia coli stress responses to sodium hydroxide, hydrochloric acid, sodium chloride, sodium hypochlorite, and ethanol. Principal component analysis and hierarchical clustering were used to quantify the effect of each parameter on bioassay performance, namely its reproducibility and metabolic resolution. Bioassay performance varied greatly, ranging from poor to very good. Spectra were partitioned into biologically relevant regions to evaluate their contributions to bioassay performance, but further improvements were not observed. Bioassay optimization was validated against empirical parameters, which confirmed a closer representation of known mechanisms on the antibiotic-induced stress responses. The optimized bioassay used standard nutrient content, cells in the late-stationary growth phase, and an 8h exposure. Only the optimized bioassay adequately and reproducibly distinguished the E. coli stress and antibiotic responses. The absence of performance improvements using partitioned spectra indicated that stress responses are imprinted on the whole-spectra metabolic signature. Ultimately, highly optimized FTIRS bioassay parameters are vital in capturing whole-spectra metabolic signatures that can be used for satisfactory and reproducible phenotypic screening of stress and antibiotic responses.

Contents

VI. 1 INTRODUCTION

The aptitude of bacteria to thrive in challenging environments depends on their adaptation to a panoply of stimuli, especially stress-inducing ones [1]. In the case of foodborne pathogens, these stimuli include industry procedures to neutralize biological contaminants, along with those encountered inside the host. Stress responses to such stimuli play a vital role on host–pathogen interactions and consequently on pathogenesis [2]. Given the panoply of stimuli that bacteria face, mechanisms that are non-specifically activated, e.g., the sigma factors RpoS and SigB in *Escherichia coli* [3], evolved alongside very precise responses whose stimuli may differ in a single unpaired electron [4].

Beyond single-cell responses, there are dynamic behaviors granted by population-wise stress responses, as in the case of altering nutrient sources or sublethal exposure to a stressing agent [5]. These play a key role in antibiotic resistant pathogens, where the loss of fitness due to resistance-conferring mutations is balanced by other mutants and/or the human microbiota [6,7]. Parallel to antibiotic resistance, stress response mechanisms are important in antibiotic tolerance [8], namely the increase of efflux pumps [9], reduction of membrane fluidity and permeability [10], alteration of antibiotic targets, induction of a general state of dormancy, increased rate of mutations, and the formation of biofilms [11].

Antibiotic-induced stresses also contribute to increased protection against environmental stressing factors, for example when *E. coli* is exposed to sublethal doses of trimethoprim, its tolerance to acidic environments increases [12], or when exposed to sublethal concentrations of ampicillin, its survivability in later exposure to other stressing factors and to lethal ampicillin doses is improved [13]. Understanding cross-protection mechanisms provides further insight into the nonspecific mechanisms elicited by bacteria upon antibiotic challenges, alongside illuminating the relationship between metabolome alterations and the mechanism of action (MOA) of antimicrobials [14]. Interestingly, many bactericidal antibiotics elicited similar metabolic alterations in *E. coli*, which contrasts with the traditional premise of target-specific inhibition resulting in the antibiotic effect [15].

In the paradigm of systems biology, the advent of genomics, transcriptomics and proteomics has contributed to undisputable advances. However, researchers often have limited insight on phenotype, given the complexity of biological systems, e.g. at the metabolite level [16]. For example, stress agents alter gene expression, however, many differentially

expressed genes are only occasionally relevant at the phenotype level [17–24]. Often, simpler indicators of stress surpass transcriptomics in probing stress responses [25]. The incoordination between transcriptional responses and phenotypic observations has been proposed as a survival mechanism, in cases where a stress response has not been evolutionarily imprinted on the micro-organism, thus preventing unreliable adaptations.

On the other hand, metabolic stresses, such as nutrient depletion, are a common challenge to which micro-organisms have a well-defined and coordinated response [26]. Therefore there are intrinsic advantages in metabolomics, given it is a closer representation of phenotype, as it relies on downstream cellular products [27]. Metabolomics techniques have very high sensitivity for individual metabolites, and protocols are well established [28]. However, in a context of phenotypic screening, focus is put on distinguishing different phenotypes, rather than on characterizing a given phenotype, thus features such as throughput, automation and cost of analysis are highly important [29].

In that regard, Fourier-transform infrared (FTIR) spectroscopy (FTIRS) acquires phenotypic profiles in a high-throughput, rapid, label-free, automatable, considerably inexpensive and reasonably simple mode [30]. FTIRS may be applied to discriminate and quantify diverse molecules, from proteins, nucleic acids, lipids through to diverse metabolites [31–33], with the advantage of representing the whole omics of a cell [34]. However, this implies *a priori* laborious calibration procedures for molecule identification and quantification.

In untargeted analysis, whole FTIR spectra have been associated with specific phenotypes, e.g., to elucidate the antimicrobial effect of novel extracts [35], to distinguish the effect of various surfactants on *E. coli* cells [36], to compare the global responses of *E. coli* to diverse stress conditions with transcriptomics and FTIRS [37], to develop a bioassay for toxicity testing in yeasts [38], among others. In comparison with the various omics techniques, FTIRS is not as data-intensive and provides valuable information without considerable detriment of metabolic sensitivity, as seen by previous examples.

In the present work, we have optimized a bioassay coupled with high-throughput FTIR spectra acquisition to discriminate highly diverse *E. coli* stress responses and began to pave the way for its applicability for antibiotic-related studies. Given its similarity in key metabolic pathways with other micro-organisms, *E. coli* is a preferred bacterial model [39,40]. Since FTIRS bioassays require considerable optimizations to maximize the biological relevance of the molecular information, most importantly regarding reproducibility and metabolic resolution

(e.g., specificity towards individual stress responses), the growth phase of the *E. coli* cells, nutrient media content and the duration of exposure to stress agents were optimized. The described approach aimed to identify the impact of each parameter on the ability of the bioassay to distinguish between the metabolic impact of a diverse set of stress agents. Additionally, partitioning FTIR spectra into key regions served to provide insight into their contribution towards the stress response clusters obtained. Lastly, the bioassay optimization was validated in the context of its application in antibiotic studies, by comparing the optimized parameters against our empirical experimental conditions, regarding the performance in distinguishing stress responses induced by antibiotics with known MOA.

VI. 2 MATERIALS AND METHODS

VI.2.1 BACTERIAL CULTURES

The model organism *E. coli* strain JM 101 (ATCC 33876) was used as it is a biosafety class 1 organism along with no described resistance mechanisms in its genome: supE thi-1 Δ(lac-proAB) [F´ traD36 proAB lacIqZΔM15]. Bacterial growth was performed by adding 1 ml aliquots, from a working cell bank at –80°C, to 60 ml of 1% yeast extract (Difco Laboratories, Inc. Detroit, MI) and 2% Bacto-Tryptone (BD Biosciences, San Jose, CA) with PBS at 0·02 mol l^{-1}, pH 7·0, in a 250 ml Erlenmeyer.

The bacteria were grown at 37°C, in an orbital incubator (TH30 and SM30; Edmund Buhler GmbH, Bodelshausen, Germany), at 250 rev min^{-1} until one of the following growth phases: mid-exponential, early- and late-stationary, which corresponded approximately to an OD_{600} of 7, 14 and 13 after 8, 16 and 24 h, respectively. To ensure consistency, the growth curve was estimated from three independent cultures, from which the approximate time and corresponding OD_{600} of each growth phase were defined. The resulting bacterial biomass was centrifuged at 4000 rev min^{-1} (3935 g) for 15 min (Rotanta 460R; Hettich Zentrifugen, Tuttlingen, Germany) and resuspended in 0·9% sodium chloride (NaCl) (Merck KGaA, Darmstadt, Germany) to obtain an OD_{600} of 50 for subsequent exposure to stressing agents.

VI.2.2 STRESS AGENT EXPOSURE

Exposure to stress-inducing agents was based on an incubation mixture of: 500 μl of a stock solution containing the stress agent at twice the final concentration, 100 μl of bacterial suspension at OD_{600} of 50 and 400 μl of nutrient media at a concentration to obtain the final

concentrations of half, standard or double that used for the bacterial cultures. Exposure to stress agents occurred at 37°C on an incubator (TH30 and SM30; Edmund Buhler GmbH) for either 1 h (immediate), 8 h (one-shift) or 24 h (overnight).

The following stress agents were evaluated, since these challenge bacteria with a wide range of stresses, thereby providing a broad variety of responses: 10% (v/v) ethanol (Sigma-Aldrich, St. Louis, MO); 8% (w/v) NaCl (Merck); 80 ppm sodium hypochlorite (VWR, Lisboa, Portugal); 2 mol l^{-1} hydrochloric acid solution (Sigma-Aldrich) corresponding to a final pH of 2·00; and a 0·5 mol l^{-1} sodium hydroxide solution (Fluka, Diegem, Belgium) corresponding to a final pH of 10·66. These concentrations are adequate to study *E. coli* stress responses since they are insufficient to induce extensive cell death [41–45]. Moreover, each stress agent had a paired control sample, where the stress agent was replaced with distilled sterile water.

To minimize the number of experiments, the following parameter combinations were evaluated: the effect of the nutrient content was studied using cells grown until the late-exponential phase and exposed to stress-inducing agents for 1 h; the effect of the *E. coli* growth phase was studied using a standard nutrient content and exposed to stress-inducing agents for 1 h; the effect of the exposure time was studied using the standard nutrients content and cells cultured until a late-exponential growth phase. These parameter combinations were selected since, in previous exploratory work, the cells grown until the late-stationary phase resulted in more reproducible assays, a low exposure period provided faster bioassays and the standard nutrient content is generally recommended. Triplicate cultures, for each defined study, were conducted over different days.

VI.2.3 Antibiotic exposure

Escherichia coli cells were exposed to six antibiotics: amoxicillin (Sigma-Aldrich), ampicillin (Sigma-Aldrich), ciprofloxacin (Bayer Healthcare, Leverkusen, Germany), metronidazole (Sigma-Aldrich), neomycin (Sigma-Aldrich) and erythromycin (Sigma-Aldrich). Amoxicillin and ampicillin target cell wall synthesis, ciprofloxacin and metronidazole disrupt nucleic acid synthesis, while neomycin and erythromycin cripple protein synthesis. Choosing antibiotics that act via different MOA allowed for an evaluation of the metabolic resolution of the bioassay, both at the MOA level and for the response to specific antibiotics.

Antibiotic stock solutions were prepared at twice the final concentration of 125 mg l^{-1}, in the recommended solvent, and adjusted to the respective potency. Moreover, antibiotic exposure also evaluated our empirical experimental conditions, namely standard nutrient content with immediate exposure (i.e., 1 h) and using cells grown until late-stationary phase; versus the optimized conditions, namely the same as above but using 8 h exposure.

VI.2.4 DATA ACQUISITION

Following exposure to stress-inducing agents, 1 ml of the incubation mixture was centrifuged for 3 min at 13 000 rev min^{-1} (13 793 g) (Z160M; Hermle Labortechnik, Wehingen, Germany), and the pellet was resuspended in 200 μl of 0·9% NaCl to obtain an OD_{600} of 25 for FTIR readings. Quintuplicates of 20 μl of these cell suspensions were plated on a ZnSe 96-well micro-plate, which was dehydrated for 3 h in a vacuum desiccator with silica. Spectra were acquired in high-throughput transmission mode using a HTS-XT module coupled to a Vertex-70 spectrometer (Bruker Optics, Ettlingen, Germany), and consisted of 64 co-added scans at a resolution of 4 cm^{-1}. Spectra were exported from the opus software (OPUS; Bruker) to Data Point Table files and analyzed with Matlab R2012b (MathWorks, Natick, MA).

VI.2.5 SPECTRAL PRE-PROCESSING

The following spectral pre-processing combinations were employed: normalization at the Amide I band (1690–1620 cm^{-1}), to compensate for intensity variations due to differences in sample biomass by forcing spectra maximum to one on that region; multiplicative scatter correction, to counteract phenomena with additive effects; and the Savitzky-Golay algorithm was applied to calculate the first derivative of spectra, using a 2^{nd} degree polynomial and a 15-point window.

To reduce spectra heterogeneity originating from operator handling, the spectra of quintuplicate mechanical replicates were averaged. Difference spectra were then calculated by subtracting the averaged control spectra from these averaged exposure spectra (i.e., both exposure and control quintuplicates were averaged, then each pair was subtracted). Difference spectra have the advantage of minimizing the signal component that is not associated to the specific stress agent response, e.g., slight variations between independent cultures.

Principal component analysis (PCA) was applied on preprocessed, averaged, difference spectra, which adds the advantage of reducing data dimensionality and subsequently facilitate its visualization, thereby helping to identify major trends.

Hierarchical cluster analysis (HCA) was used to evaluate intrinsic data patterns and to determine pairwise distances that were visualized in hierarchical trees (i.e., dendrograms). These served to evaluate the bioassay reproducibility regarding the independent cultures (by determining the maximum distance within a cluster containing three independent cultures exposed to a specific stress-inducing agent), and metabolic resolution, in terms of the ability to distinguish between the effect of different stress agents (by determining the distance between a cluster of independent cultures exposed to a given stress-inducing agents and the remaining clusters). This approach was also used to evaluate the effect of partitioning spectra in order to identify specific spectral regions associated with the specific phenotypic response.

VI. 3 RESULTS

Under the premise that the specific *E. coli* responses to stress agents were imprinted on FTIR spectra, these were used as metabolic signatures of the individual stress responses. In order to optimize the bioassay performance in distinguishing diverse stress responses, especially considering reproducibility (e.g., coherence between three independent culture replicates), and metabolic resolution (e.g., specificity towards varying stress responses), the following bioassay parameters were evaluated: the growth phase from which *E. coli* cells were obtained (from mid-exponential, early- and late-stationary phase), the nutrient media concentration used (half, standard and double), and the duration of exposure to the stress agent (1, 8 and 24 h).

VI.3.1 NUTRIENT CONTENT

Rather than using different media for the exposure mixture, which would imply different nutrient compositions, we focused strictly on the nutrient content. As expected, the concentration of nutrients had a clear effect on the cellular metabolism, which affected the bioassay sensitivity towards the stress responses (Figure 1). The following nutrient concentrations were evaluated: standard, double and half concentration in relation to that used for the bacterial cultures. In general, optimized conditions ensured lower intra-culture variability,

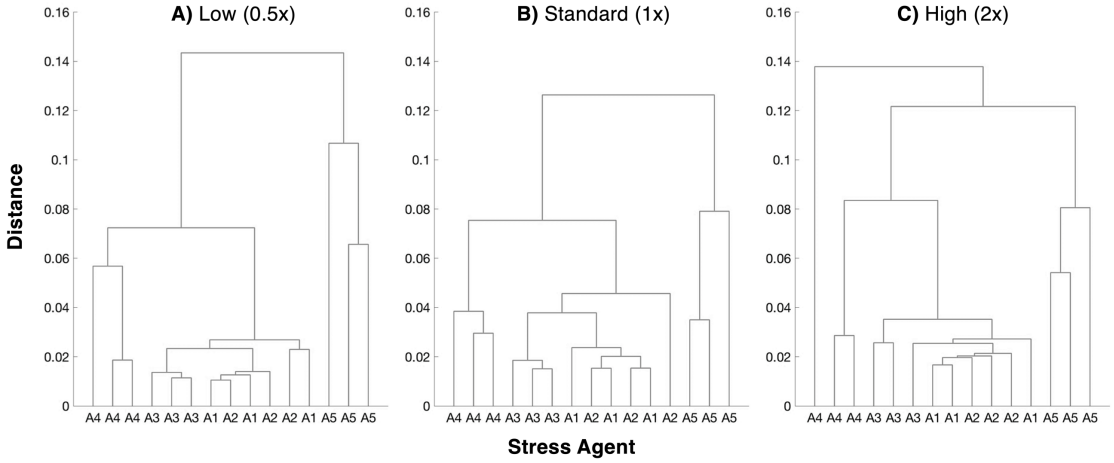


Figure 1. Effect of the nutrient content using A) Half, B) Standard, and C) Double media concentrations on bioassay performance for *E. coli* cells exposed to Ethanol (A1), Sodium hypochlorite (A2), Sodium chloride (A3), Hydrochloric acid (A4) and Sodium Hydroxide (A5).

thus a more coherent bioassay, as well as increased the difference between the metabolic signature of different stress agents, which suggested greater bioassay sensitivity.

Another critical parameter for evaluating the said bioassay's performance was the number of misclassifications. As such, the high nutrient content negatively affected the performance of the proposed bioassay, given that only the sodium hydroxide samples were adequately grouped. Consequently, the high nutrient content condition was not further discussed. Additionally, the ethanol and sodium hypochlorite metabolic signatures were not adequately classified in any of the parameters investigated, thus were also not subject to further discussion.

Increasing nutrient content from half to standard resulted in smaller intra-culture variability of samples stressed with sodium hydroxide (–25·88%), in other words the maximum distance that separated the three independent cultures exposed to sodium hydroxide decreased by 25·88% when the nutrient content increased, likewise for hydrochloric acid (–32·28%). In contrast, this also led to a higher variability of the sodium chloride metabolic signature (+35·92%). Regarding the bioassay sensitivity, the variability between the sodium hydroxide, hydrochloric acid and sodium chloride stress-induced cultures and the remaining decreased (–11·85%), slightly increased (+4·17%) and considerably increased (+61·96%), respectively. In other words, the distance between the cluster of cultures exposed to each of

the stress agents from the remaining, decreased, slightly increased and considerably increased, which indicated poorer, slightly better and considerably better performance, since a greater distance between different clusters implied greater bioassay sensitivity.

In resume, the increase of nutrient content from half to standard concentration resulted in more coherent metabolic signatures of the sodium hydroxide and hydrochloric acid stress agents among the three independent cultures, although the opposite was observed for the sodium chloride exposed samples. Moreover, a more noticeable metabolic signature of hydrochloric acid assay was obtained, although at the cost of a less prominent sodium hydroxide and sodium chloride stress signatures in comparison with the other stress responses.

VI.3.2 ESCHERICHIA COLI GROWTH PHASE

Another key parameter of the suggested bioassay was the starting metabolic state of the *E. coli* cells, which influenced the sensitivity towards the spectral signature induced by the stress agents, as seen by the effect of the *E. coli* cells growth phase in the fermentation industry [46]. As such, the metabolic state of the *E. coli* cells, obtained at different growth phases prior to stress agent exposure, clearly affected the bioassay performance (Figure 2). Three key points of the *E. coli* growth curve were investigated: mid-exponential, early-stationary and late-exponential.

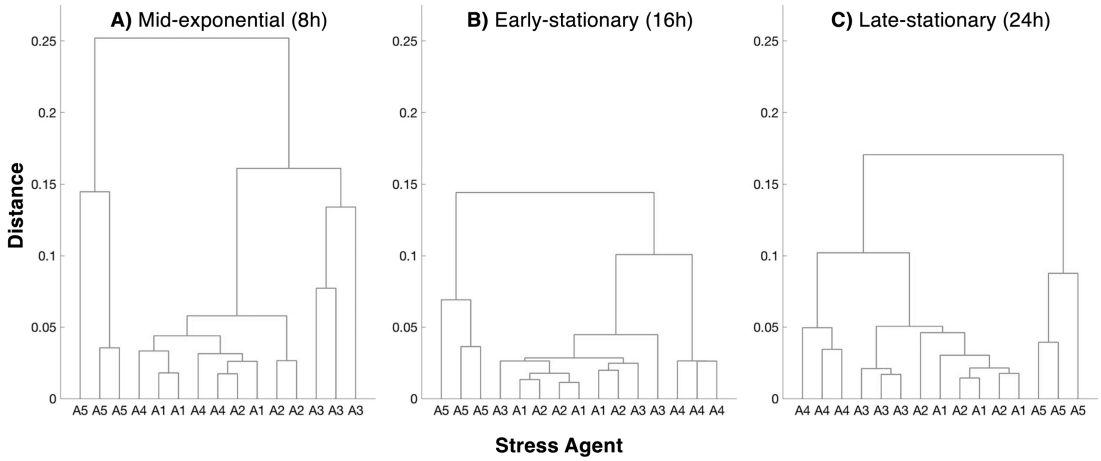


Figure 2. Effect of the *E. coli* metabolic state on the bioassay performance. *E. coli* cells grown until A) Mid-exponential, B) Early-stationary, and C) Late-stationary growth phases were exposed to Ethanol (A1), Sodium hypochlorite (A2), Sodium chloride (A3), Hydrochloric acid (A4) and Sodium Hydroxide (A5).

Interestingly, as the growth phase of the cells used changed, so did the ability of the bioassay to effectively distinguish stress response signatures. While a mid-exponential growth phase suitably classified both sodium hydroxide- and sodium chloride-exposed samples, the remaining stress agents were all misclassified. Using cells from the early-stationary growth phase no longer adequately classified samples exposed to sodium chloride, however, bacteria stressed with hydrochloric acid and sodium hydroxide became well classified. Lastly, using cells in the late-stationary phase, the highest number of adequate classifications of stress responses was observed, where only the sodium hypochlorite and ethanol stress response signatures were not adequately classified. These misclassifications were consistent with those observed under the tested nutrient concentrations.

Regardless, using cells from a later growth phase, e.g., from mid-exponential to early-stationary, the intra-culture variability for sodium hydroxide was reduced (–52·20%). However, using cells from the late-stationary phase increased the intra-culture variability (+26·69%). A similar pattern was observed for the difference between the sodium hydroxide signature and the other stress responses, namely cells from later growth phases decreased (–42·78%) but subsequently increased (+18·31%) the intra-culture variability.

The remaining stress-inducing agents were not consistently classified across the range of growth phases tested, however, hydrochloric acid's intra-culture variability increased (+87·77%) from the early-stationary to late-stationary, while its difference from the remaining stress agents also increased (+10·21%). Moreover, sodium hydroxide's intra-culture variability decreased (–84·28%) from mid-exponential to late-stationary, and its distinguishability from other stress responses also diminished (–68·58%).

VI.3.3 Stress agent exposure time

The third bioassay parameter evaluated was the duration that the *E. coli* cells were exposed to the various stress agents. The trade-off revolved between a short exposure, which resulted in a more rapid assay, and issues with reproducibility and/or sensitivity associated with a longer exposure. In fact, of the bioassay parameters discussed thus far, the stress agent exposure duration had the most notorious impact on the performance of the bioassay. A maximum of two misclassifications were observed (Figure 3 – A and C), which is on a par with the best performances observed thus far.

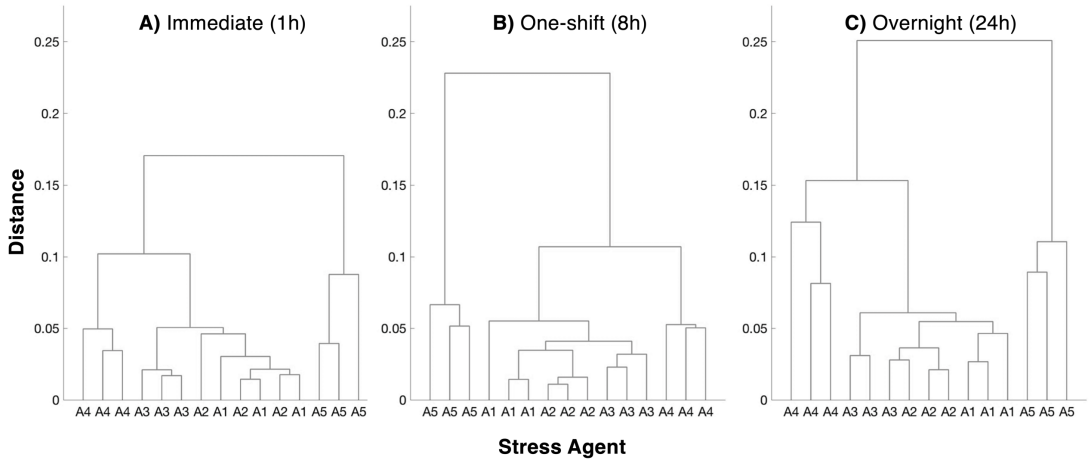


Figure 3. Effect of the stress agent exposure time on bioassay performance, using A) Immediate, B) One-shift, and C) Overnight timeframes for *E. coli* cells exposed to Ethanol (A1), Sodium hypochlorite (A2), Sodium chloride (A3), Hydrochloric acid (A4) and Sodium Hydroxide (A5).

More importantly, the one-shift exposure (Figure 3 – B) adequately classified all five stress agents, although one culture of ethanol was deemed as an outlier. This was because it was clustered as considerably different from the other ethanol-exposed cultures, but also those exposed to sodium hypochlorite and sodium chloride, which suggested a marked difference in that culture alone. Actually, the one-shift exposure duration was the only one in which proper classification of the five stress agents tested was observed.

Regarding the intra-culture variability, longer exposure times resulted in a decrease from immediate to one-shift for sodium hydroxide (–25·06%) but an increase for hydrochloric acid (+6·25%) and sodium chloride (+51·66%). Longer exposures, from one-shift to overnight, had the opposite effect for the cultures exposed to sodium hydroxide, hydrochloric acid and sodium chloride, where intra-culture variability increased by 66·06, 135·81 and 222·52% respectively. On the other hand, the difference between stress response signatures of either sodium hydroxide or hydrochloric acid in regard to the other stress agents increased from immediate to one-shift (+33·70 and +4·80%), while it decreased for sodium chloride (–18·94%).

Regarding metabolic sensitivity, the effect of a longer exposure time, from one-shift to overnight, had a similar effect where the stress response became less prominent for sodium hydroxide (–9·95%), hydrochloric acid (–43·18%) and ethanol (–58·00%). Therefore, the bioassay performance was superior with a one-shift exposure (i.e., 8 h) to stress agents, where

fewer misclassifications were observed, and accurate classifications were largely more coherent and more accurate.

VI.3.4 SPECTRA PARTITIONING TOWARDS PHENOTYPIC RESPONSE-INTERPRETATION

This section aimed to evaluate the spectral regions that most affected the PCA-HCA clustering. Using the bioassay parameters that granted the best performance when optimizing the exposure duration, i.e. the spectra obtained with an exposure duration of one-shift (i.e. 8 h), four key spectral regions (as described by other authors [37,47]) were investigated: region 1 (3400–3200 cm^{-1}), region 2 (3010–2800 cm^{-1}), region 3 (1800–1200 cm^{-1}) and region 4 (1200–700 cm^{-1}).

The differences observed at the level of preprocessed and averaged difference spectra, with the corresponding HCA after PCA metric system previously described, are highlighted in Figure 4. Although the same analysis was conducted for all parameters tested, i.e., exposure duration, growth phase and nutrient content (Figs S1–S6), there was no clear benefit of partitioning the spectra into regions for the PCA-HCA models that originally had weak to moderate performances.

From spectral region 1 (Figure 4 – A) a satisfactory classification was obtained for sodium hydroxide, hydrochloric acid and sodium hypochlorite. Sodium chloride and ethanol were adequately classified considering the whole spectra (Figure 3 – B) but became misclassified resourcing only to region 1, which suggested that the metabolic effect of these stress agents is not sufficiently imprinted at this region for adequate classification.

Moreover, spectral region 2 (Figure 4 – B) resulted in a noticeably poorer classification of stress responses, as only the hydrochloric acid exposed samples, and to some extent those exposed to sodium hydroxide, presented adequate clustering. It was also observed that resourcing to either region 1 (Supplementary Figure 2 – A) or to region 2 (Supplementary Figure 2 – B), the PCA-HCA presented worse results when using immediate and overnight exposure. As such, the various stress responses could not be distinguished with the proposed bioassay resourcing exclusively to these regions.

Unlike the formerly described examples, the use of spectral region 3 apparently resulted in better clustering than observed when based on the whole spectra (Figs 3b and 4c, respectively), as the distances between the stress responses of different stress agents

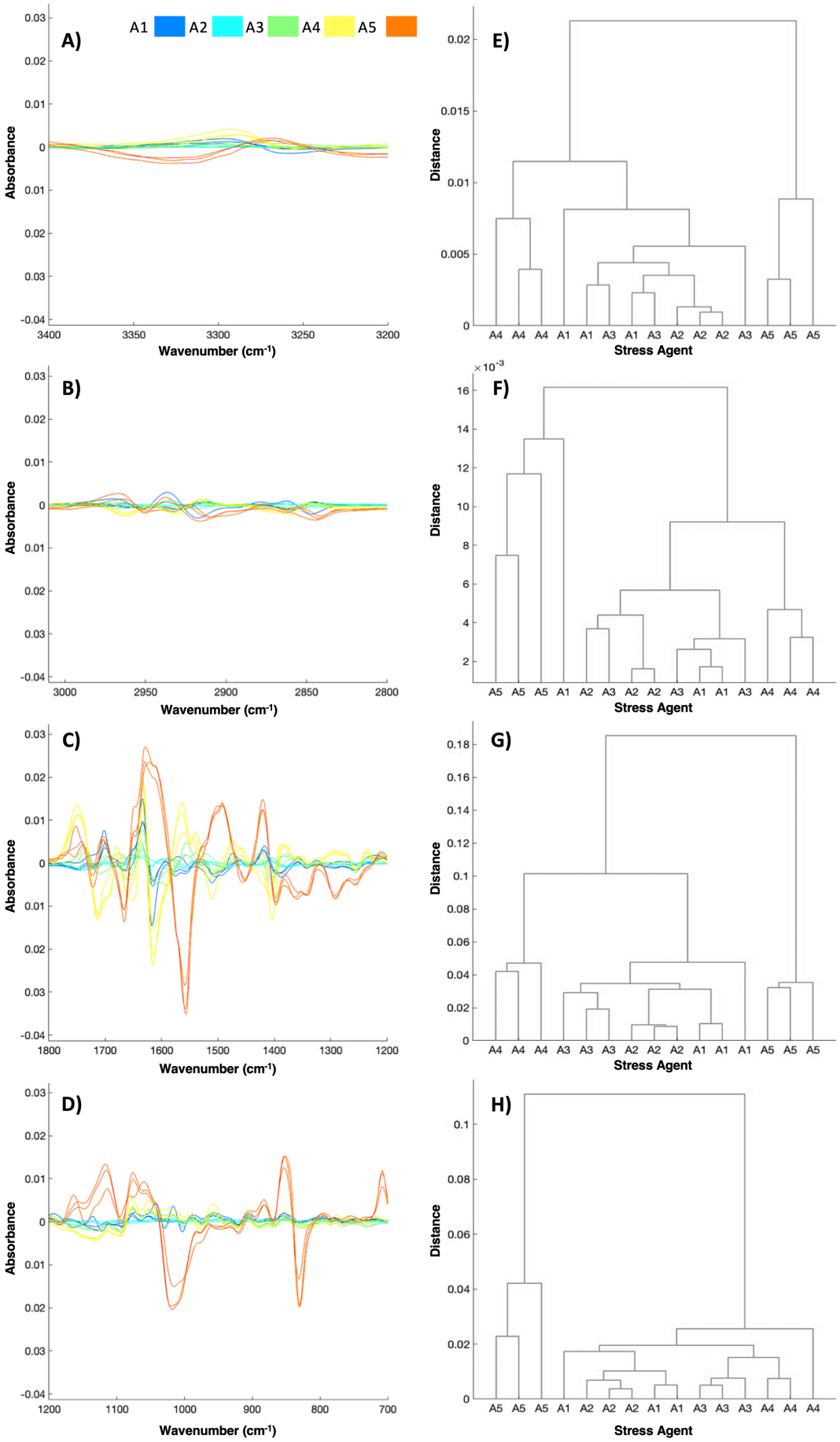

A)
A1 A2 A3 A4 A5
Absorbance
Wavenumber (cm-1)
B)
C)
D)
E)
Distance
Stress Agent
F)
×10-3
G)
H)

Figure 4. Stress response signatures of first derivative spectra normalized at the Amide I and with MSC (left column) and corresponding PCA-HCA metric system (right column) for the regions between A) 3400-3200 cm^{-1}, B) 3010-2800 cm^{-1}, C) 1800-1200 cm^{-1} and D) 1200-700 cm^{-1} obtained with standard nutrient content for *E. coli* cultures exposed to Ethanol (A1), Sodium hypochlorite (A2), Sodium chloride(A3), Hydrochloric acid (A4) and Sodium Hydroxide (A5).

increased, which indicated a higher resolution towards different stress responses. Likewise for whole spectra, a single ethanol-exposed culture was misclassified, and the difference between cells exposed to sodium hydroxide, hydrochloric acid, sodium chloride, sodium hypochlorite and ethanol diminished by 22·85, 5·42, 18·19, 10·84 and 10·84% in regard to the other stress response signatures, respectively. However, this higher resolution came at the cost of increased intra-culture variability, namely of 87·98, 11·68, 9·69, 66·81 and 39·97% for sodium hydroxide, hydrochloric acid, sodium chloride, sodium hypochlorite and ethanol, respectively.

Ultimately, region 3 seemed to provide comparable results to whole spectra, however, the increase in intra-culture variability, and therefore reduced reproducibility of the bioassay, was not sufficiently beneficial to metabolic resolution. Moreover, with an exposure duration of 1 h, sodium hypochlorite exposed cultures were no longer clustered appropriately, which was a clear downgrade of the bioassay performance; and after 24 h exposure, the three ethanol-exposed cultures were grouped, however, a sodium chloride-exposed culture became misclassified along with the sodium hypochlorite-exposed cultures (Supplementary Figure 2 – C).

Comparable results were obtained when the PCA-HCA metric was applied to region 4 (Figure 4 – D). In this case, noticeable differences at the spectra level resulted in a clustering pattern analogous to that obtained exclusively with region 2 or with the whole spectra. However, an additional misclassification of a hydrochloric acid exposed culture suggested a weaker performance. Interestingly, the hydrochloric acid exposed samples were clustered closer to sodium chloride exposed samples, unlike for the complete spectra and region 3.

Additionally, the immediate response PCA-HCA presented better clustering for this particular region when compared with a one-shift exposure (Supplementary Figure 2 – D), although ethanol and sodium hypochlorite-exposed cultures were still misclassified. This revealed that this particular region did not outperform the classification obtained with the whole

spectra, which was also observed with an overnight exposure, where only hydrochloric acid and sodium hydroxide stress responses were accurately clustered.

VI.3.5 EXPOSURE TO ANTIBIOTIC AGENTS

The applicability of the developed bioassay in distinguishing the response to various antibiotics was evaluated, albeit to a more modest degree. In that sense, the bioassay parametrization that resulted in optimal stress response classification was validated by comparing it with our empirical parametrization, e.g., which we typically applied prior to this study. Figure 5 - A shows the clustering obtained with the bioassay parameters before optimization, and Figure 5 – B shows the clustering obtained with the optimized parameters.

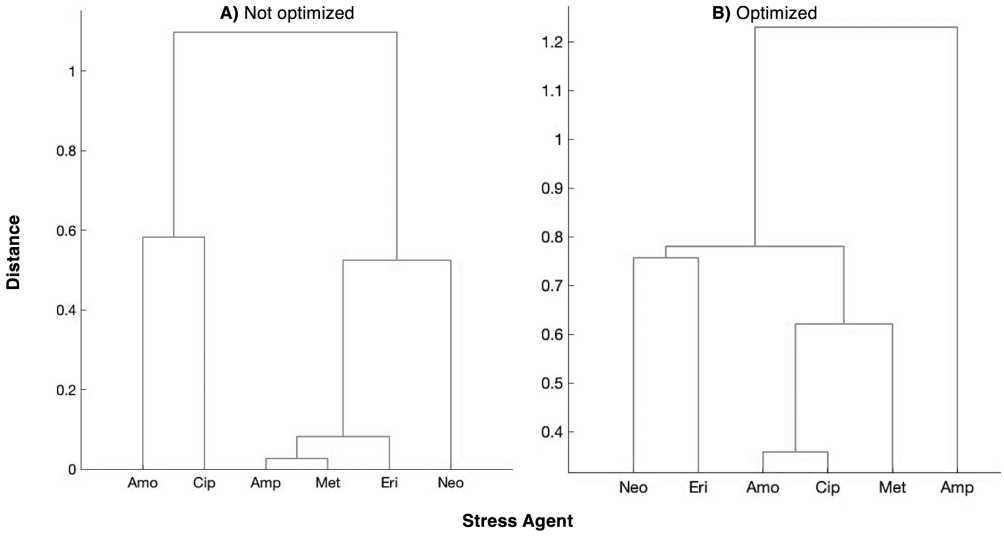


Figure 5. Comparison of bioassay performance with A) not optimized and B) optimized parameters for *E. coli* exposed to Neomycin (Neo), Erythromycin (Eri), Amoxicillin (Amo), Ciprofloxacin (Cip), Metronidazole (Met) and Ampicillin (Amp).

The optimized bioassay provided a closer representation of the known metabolic alterations induced by the antibiotics. In other words, the *E. coli* response to antibiotics was considerably dissimilar, except for amoxicillin and ciprofloxacin, which were grouped as very similar although these have different MOA, at the cell wall biosynthesis and DNA level, respectively. The clustering of metronidazole with ciprofloxacin was a positive indication, given both target nucleic acid synthesis, which suggested a good ability of the bioassay to classify antibiotic stress responses, despite the amoxicillin misclassification. Moreover, neomycin and

erythromycin were grouped closely, which was another promising indication, given that both act at the protein synthesis level.

Additionally, the difference between the neomycin- and erythromycin-exposed samples could be due to the fact that, while the first binds to the 30S-ribosomal subunit, the later inhibits the 50S-ribosomal subunit. This said difference between similar MOA seems within the scale of the difference between the comparable case of ciprofloxacin and metronidazole, which both act on nucleic acid synthesis. Lastly, the ampicillin stress response was not clustered with the amoxicillin stress response, although these act at the cell wall biosynthesis level. Ultimately, this was a more realistic clustering than without optimizing the bioassay, since a more reasonable correspondence with the antibiotics MOA was observed.

Moreover, models built with specific spectral regions, as previously conducted, lead to marginal local improvements, if at all, at the cost of overall weaker performances (data not shown), which indicated that the bioassay more closely captured the unique metabolic signature of antibiotic responses when the whole FTIR spectra was considered.

VI. 4 Discussion

Bacterial responses to stressful events encompass a wide variety of adaptations, many of which have been described for *E. coli*, including those following exposure to: sodium hypochlorite, a strong oxidizer that up-regulates over 380 genes and induces faster post-exposure recovery cell growth [43,48]; ethanol, which disrupts the membrane and cell wall integrity, reducing O_2 levels and ATP production, as well as fostering DNA damage and a six orders of magnitude increase of protein expression [45,49]; sodium chloride, which increases water efflux, thereby affecting morphology and reducing maximum growth, ultimately resulting in extensive cell damage and loss of integrity [44,50–53] and acidic environments that activate mechanisms to counteract the pH imbalance [41,54–56].

Fewer studies have discussed alkaline stress responses in *E. coli*, nonetheless two heat shock proteins and the *rpoS* gene have been shown to take part in the alkaline stress response [42,57], and other propositions have been put forward for other lactic acid bacteria [58]. Ultimately, exposure to these particular stressing agents elicits a variety of responses, which validates their use to estimate the proposed bioassay's metabolic resolution to different stress responses. Moreover, to the best of our knowledge, this is the first time such bioassay

optimization was undertaken with a systematic quantitative approach resulting in considerably more adequate stress response discrimination.

The type and availability of nutrients effect on the growth behavior has been reported for yeast cells [59], however this is an issue with little to no discussion for prokaryotes. Said FTIR study established that higher levels of glucose reduce the bioassay variability, thus improve its reproducibility, however the same concentration of glucose complementing different nutrient sources does not affect bioassay variability to the same extent. This suggests that bioassay reproducibility is mostly affected by the availability of the major energy source, but given differences in study design and model organism, a straightforward comparison cannot be pursued.

However, our study suggests that the effect of nutrient availability has a limit to improving bioassay reproducibility, as indicated by the maximum number of classifications obtained with the high nutrient content, as well as the higher distance between clusters of different stress-inducing agents, compared to the standard and half nutrient concentration. Even so, the study with yeast cells is somewhat on the same level with that reported herein, since the proposed bioassay has increased coherence (or homogeneity) and increased specificity when the nutrient content is increased to the standard concentration, despite not being advantageous across the different stress agents tested.

In general, using bacterial cells obtained at a late growth-phase increases the performance of the bioassay, and ensures a more coherent bioassay, to a scale that outweighs the reduced differences between different stress response signatures. Actually, mass spectrometry analysis of *E. coli* during various points of its growth pattern revealed an increased expression of multiple stress response proteins during the stationary phase [49]. These observations seem coherent with the present study, where more advanced growth phases seem to positively affect the ability of the bioassay to identify spectral signatures unique to each stressing agent, possibly because some of the general stress response mechanisms are also present in the control (not exposed) samples, simply because they are elicited during later growth phases. These can therefore be easily ‘removed’ from the spectral signatures, as part of the pre-processing and multivariate analysis workflow.

Remarkably, metabolome profiling of bacterial responses to antibiotics has revealed that the immediate responses are identical, likely given the activation of general stress response mechanisms, however, these responses then become characteristic of the particular

stress induced [15]. As such, it appears that the longer-than-immediate response becomes more representative of the unique stress response, however, after too long an exposure, such a response is no longer characteristic of the specific effect of the stress agent, possibly because the stress agent becomes depleted and therefore the spectral signature no longer reflects a stress response, but rather the metabolic composition of the culture after recovery, which likewise to the identical initial responses, is not necessarily unique to a given stress.

Several FTIR spectral regions have been associated with different biomolecules [37,47]. However, the locations of these spectral regions are not always coherent across different studies. Therefore, the spectral regions investigated in this study were defined to be as inclusive as possible regarding those found in the literature, i.e., for region 2, Maity et al. suggested 3000–2800 cm^{-1} and Moen et al. suggested 3010–2800 cm^{-1}, so region 2 was defined as 3010–2800 cm^{-1}. Additionally, the association of these regions with underlying predominant biomolecular bonds contributing to the spectra are also not always consistent. For instance, what Maity et al. define as a wider fingerprint region, is actually regarded by Moen et al. as two regions, one fingerprint region and one with predominant contributions from carbohydrates.

Moreover, we deliberately omitted the associations between spectral windows and predominant biomolecules to keep the discussion focused on the performance of the bioassay. Regarding spectra partitioning, although region 3 and 4 seem to provide similar stress response signatures towards the optimal bioassay performance observed for the one-shift analysis, it appears that said optimal clustering cannot be obtained exclusively from a particular region, and therefore the distinctive stress responses are not completely imprinted at the level of the associated spectral regions. This indicates that the successful classification of the five stress-inducing agents results from the whole FTIR spectra, which therefore constitutes a complex metabolic profile of each stress response.

Studies on the effect of sodium hypochlorite on *E. coli* suggest a minor effect compared with other bacteria [60], which is coherent with the observed results. Alongside ethanol, and to some extent, sodium chloride, these were the stress response signatures more sensitive to the bioassay parameters, and therefore for which the optimizations had greater impact. Interestingly, in a study on the antimicrobial activity of disinfectants against *Staphylococcus aureus*, the FTIRS signature of ethanol and sodium hypochlorite were markedly different [61]. Although this said study investigated maximum bacterial inactivation rather than stress

response, their results differ from our findings, where these two stress responses were consistently un-distinguishable, which suggests a similar stress response to these two agents.

However, given that the ethanol and sodium hypochlorite spectra signatures clearly separate into two different stress responses with optimized bioassay parameters, this indicates that there is a narrow window that maximizes the bioassay ability to distinguish stress responses, which ultimately justifies the need to strictly optimize bioassay parameters in stress response studies. Additionally, sodium hydroxide had the most distinct stress response signature, followed by hydrochloric acid stressed samples, across the various optimized parameters, as they were most often clustered very differently from the others and therefore were less affected by the bioassay optimization.

Thus far, the stress agent exposure bioassay optimization not only further validated FTIRS as an adequate tool for phenotypic screening, but also shed light on its application in stress response studies. Beyond said application, some studies have dwelt on the applicability of FTIRS towards antibiotic-related studies, namely the early evaluation of anti-bacterial effect of known and novel molecules, in terms of MOA and activity [35,62–64]. However, to the best of our knowledge, no studies focus on the importance of bioassay optimization towards more sensitive and robust classification and predictive models. Since the antibiotic-induced stress responses were analyzed to a lesser scale than other stress agents, it becomes difficult to evaluate if the patterns described are coherent when considering a greater biological variability. In other words, the amoxicillin response cannot be deemed an outlier, as an ethanol-exposed sample was in the stress exposure duration optimization of one-shift.

Nonetheless, this reiterates the importance of optimizing FTIRS-based bioassays in order to ensure that spectral data accurately and coherently distinguish between different stress responses, including those induced by antibiotics. Moreover, spectra partitioning indicates that the metabolic signature derives from the whole FTIR spectra. Nonetheless, considering more antibiotics acting with similar MOA, for instance several protein synthesis inhibitors, a PCA-HCA analysis of a specific region might be of added-value in discriminating very similar stress responses.

In this study, we proposed a bioassay to discriminate between the spectral signatures of the different *E. coli* stress responses to sodium hydroxide, hydrochloric acid, sodium chloride, sodium hypochlorite and ethanol. More importantly, the pre-processing and multivariate analysis applied, namely PCA-HCA, provided a quantitative evaluation of the effect of three

key parameters on the performance of the proposed bioassay regarding its reproducibility and specificity.

From the parameters tested: the highest nutrient content led to more misclassifications, while the lower nutrient content had a slightly poorer metabolic resolution; the use of *E. coli* cells grown until later growth phases reduced the number of misclassifications and improved overall performance; and lastly for the duration of the stress agent exposure, a one-shift exposure duration resulted in an almost complete correct classification of the three replicated assays for each of the stress agents. Moreover, four specific spectral regions were explored regarding their contribution towards the optimal classification, where importantly, the overall performance of the bioassay did not surpass that of using the whole spectra.

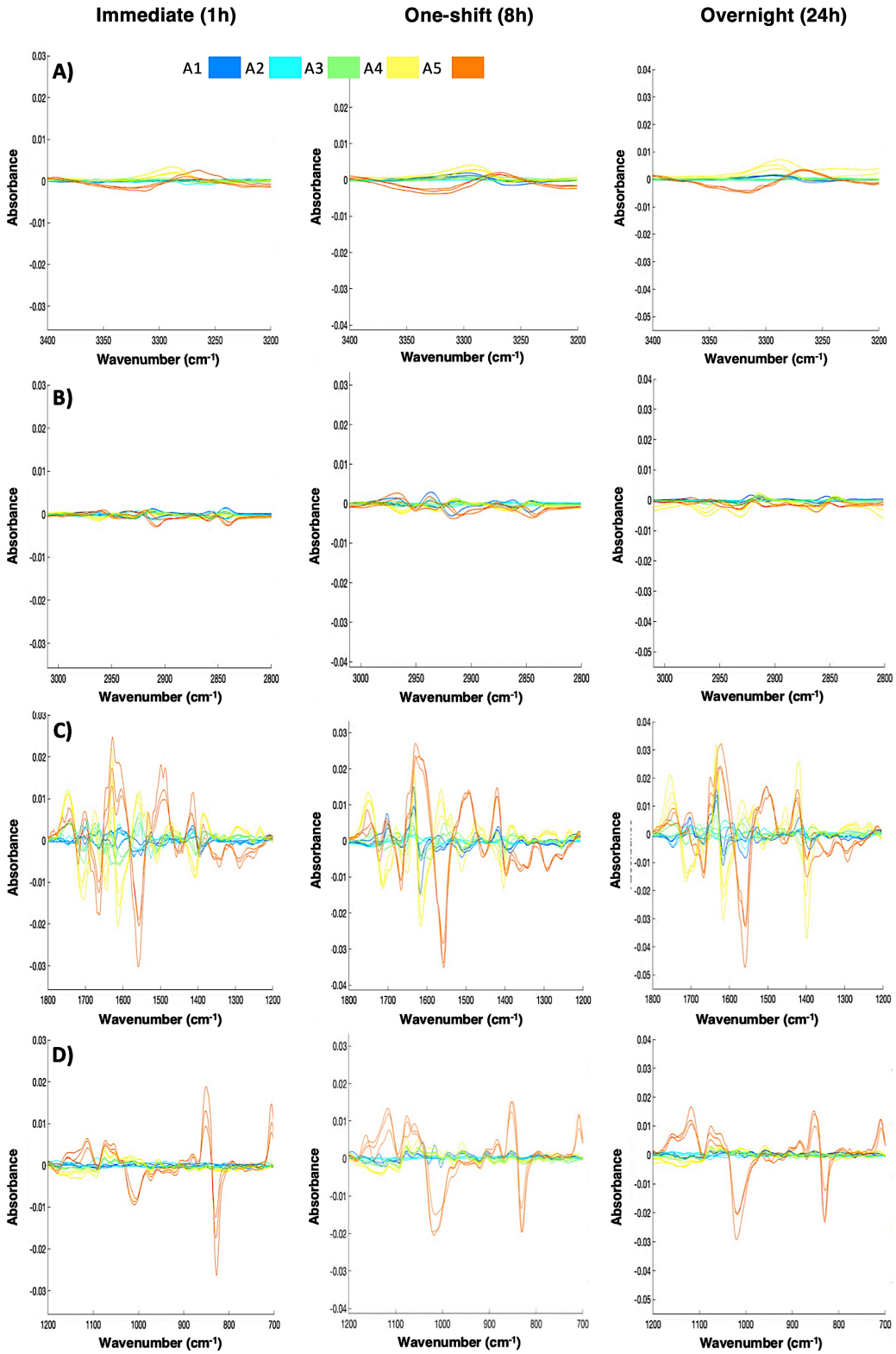
Ultimately, the bioassay optimization ensured a closer representation between the spectral signatures and the MOA of clinical antibiotics. As such, the optimal classification of stress response signatures, including those imposed by exposure to antibiotics, requires highly optimized bioassay parameters and seems to be consequence of the metabolic profile imprinted on the whole FTIR spectra, which reinforces the application of FTIRS to acquire a metabolic signature of general stress responses and antibiotic-related responses.

VI. 5 ACKNOWLEDGMENTS

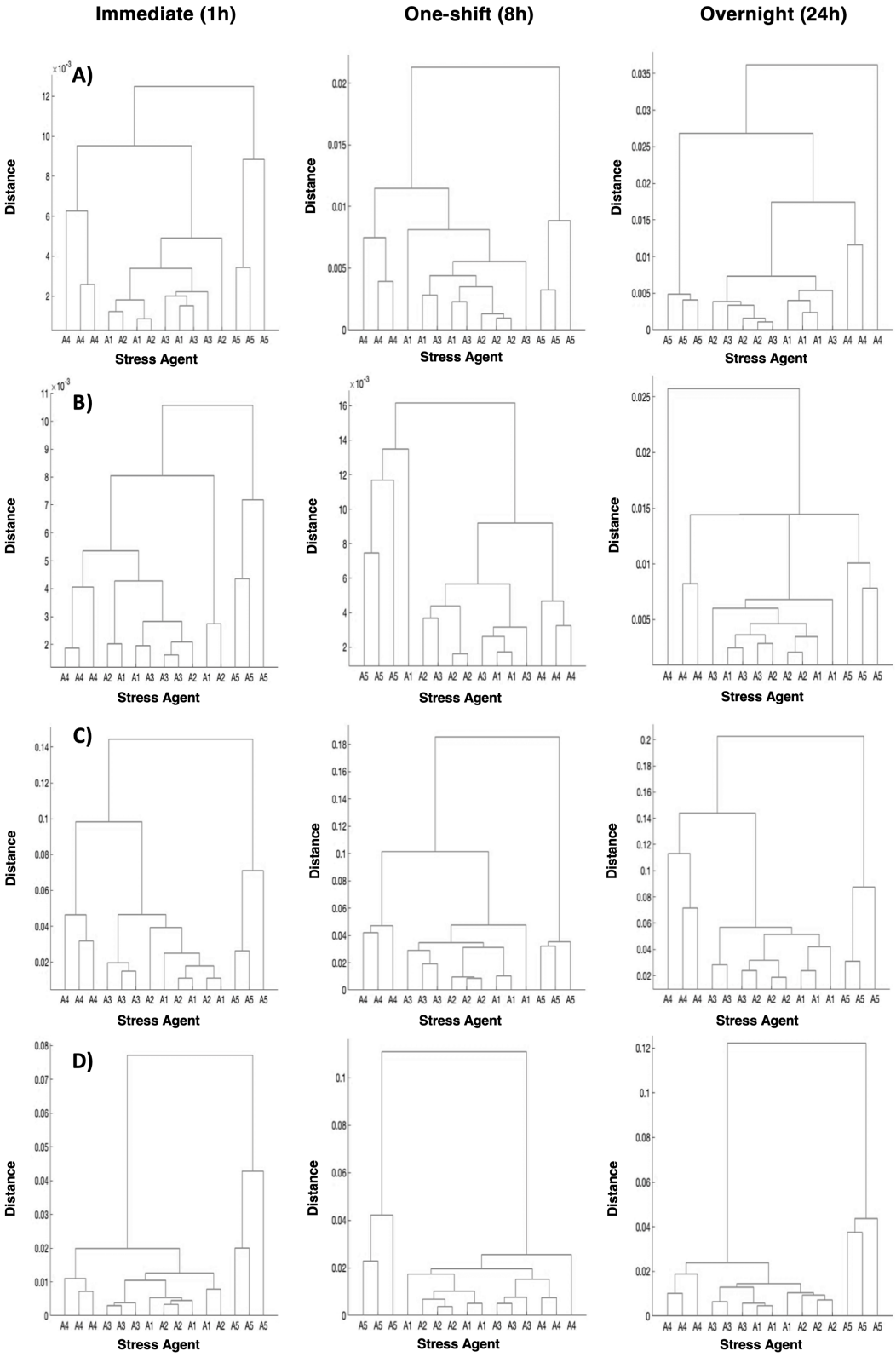
The present work was supported by Instituto Politécnico de Lisboa, grant number IDI&CA/IPL/2017/DrugsPlatf/ISEL and IDI&CA/IPL/2018/RenalProg/ISEL and was partially conducted in the Engineering & Health Laboratory, which results of a collaboration protocol established between Universidade Católica Portuguesa and Instituto Politécnico de Lisboa.

The authors would like to acknowledge the contribution of Ana Lúcia Evaristo Russo to the data collection process.

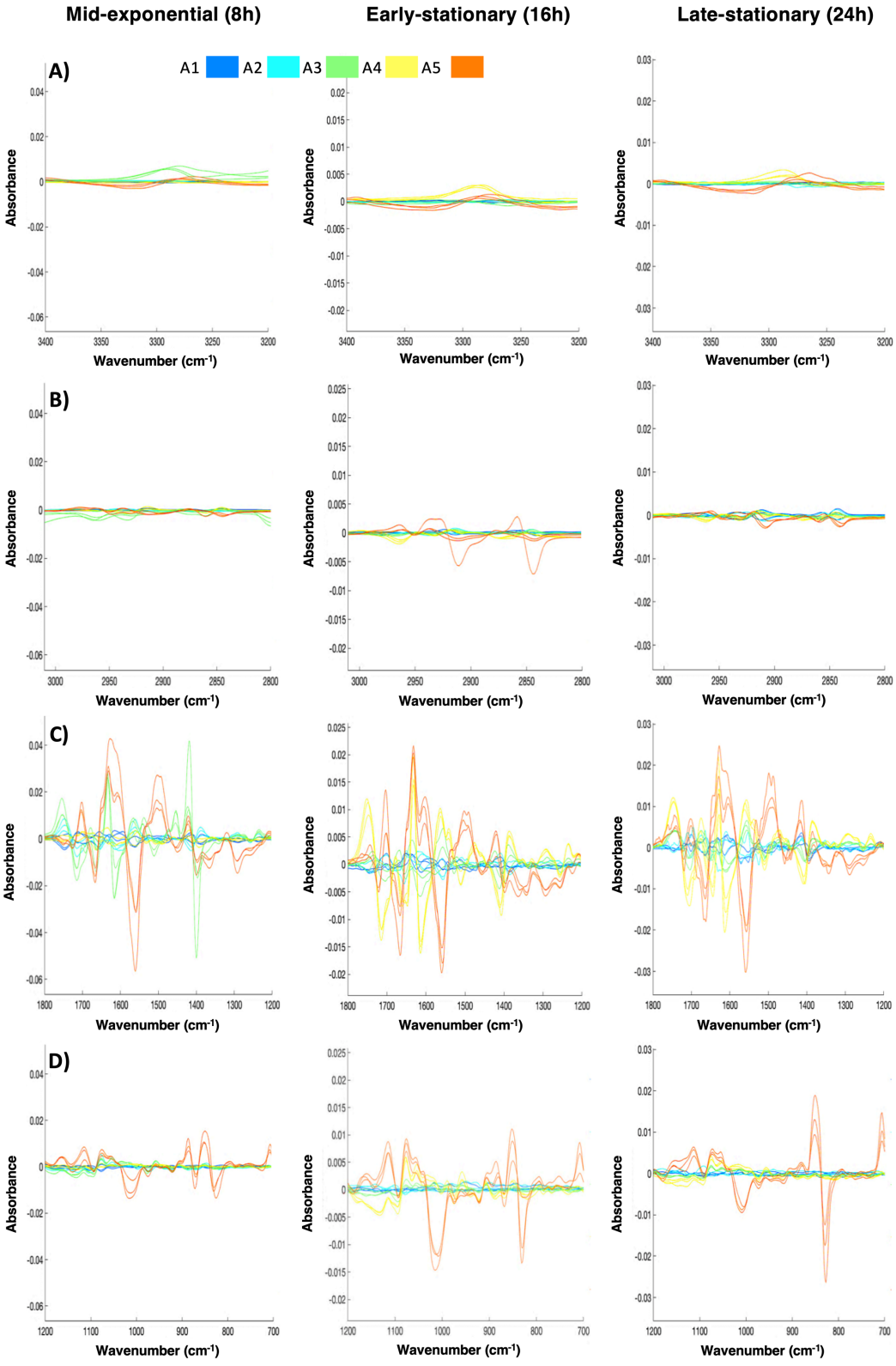
VI. 6 SUPPLEMENTARY DATA



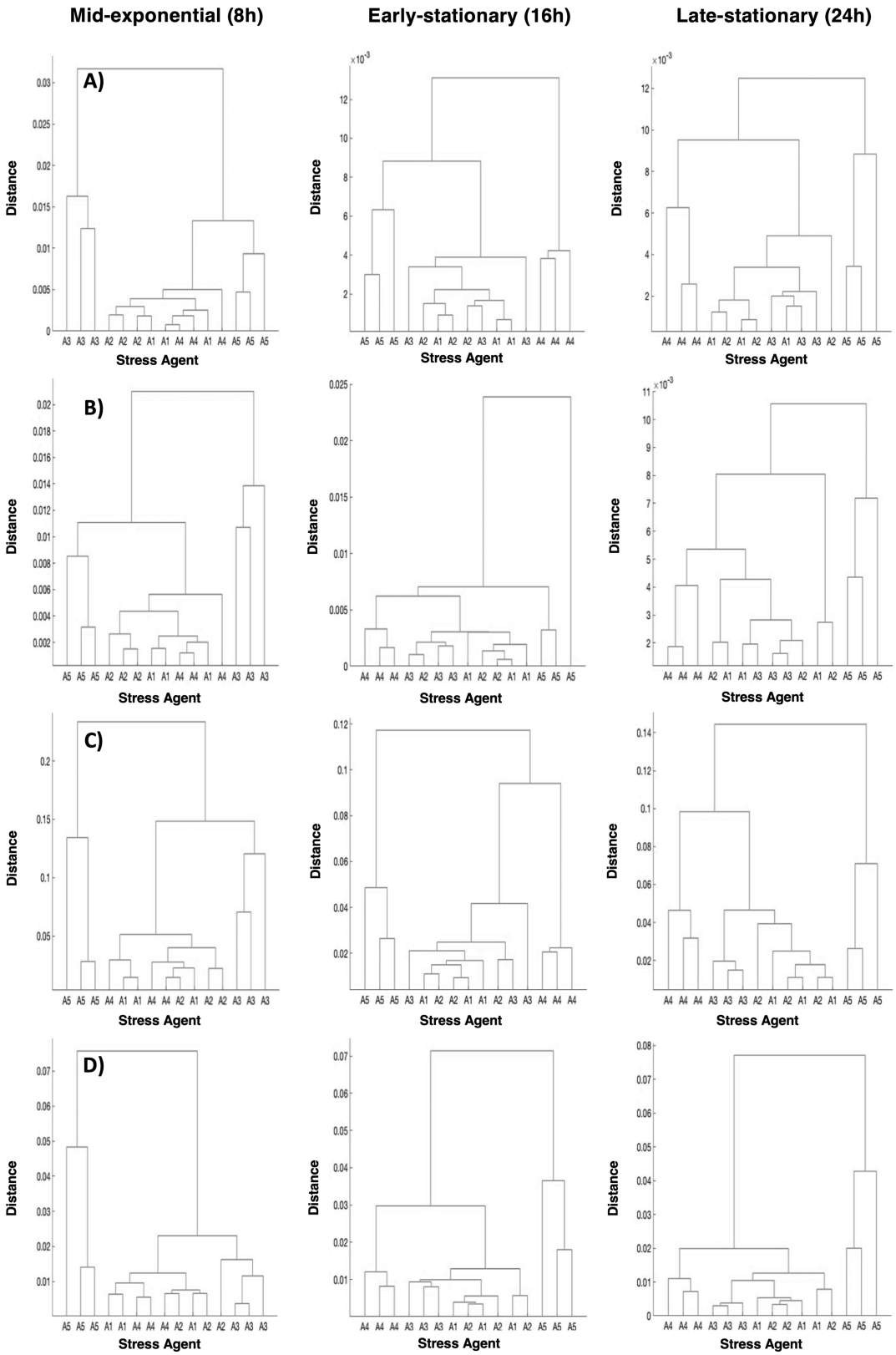
Supplementary Figure 1. Stress response spectral signatures for the regions between A) 3400-3200 cm^{-1}, B) 3010-2800 cm^{-1}, C) 1800-1200 cm^{-1} and D) 1200-700 cm^{-1} obtained after immediate, one-shift and overnight exposure to Ethanol (A1), Sodium hypochlorite (A2), Sodium chloride (A3), Hydrochloric acid (A4) and Sodium Hydroxide (A5).



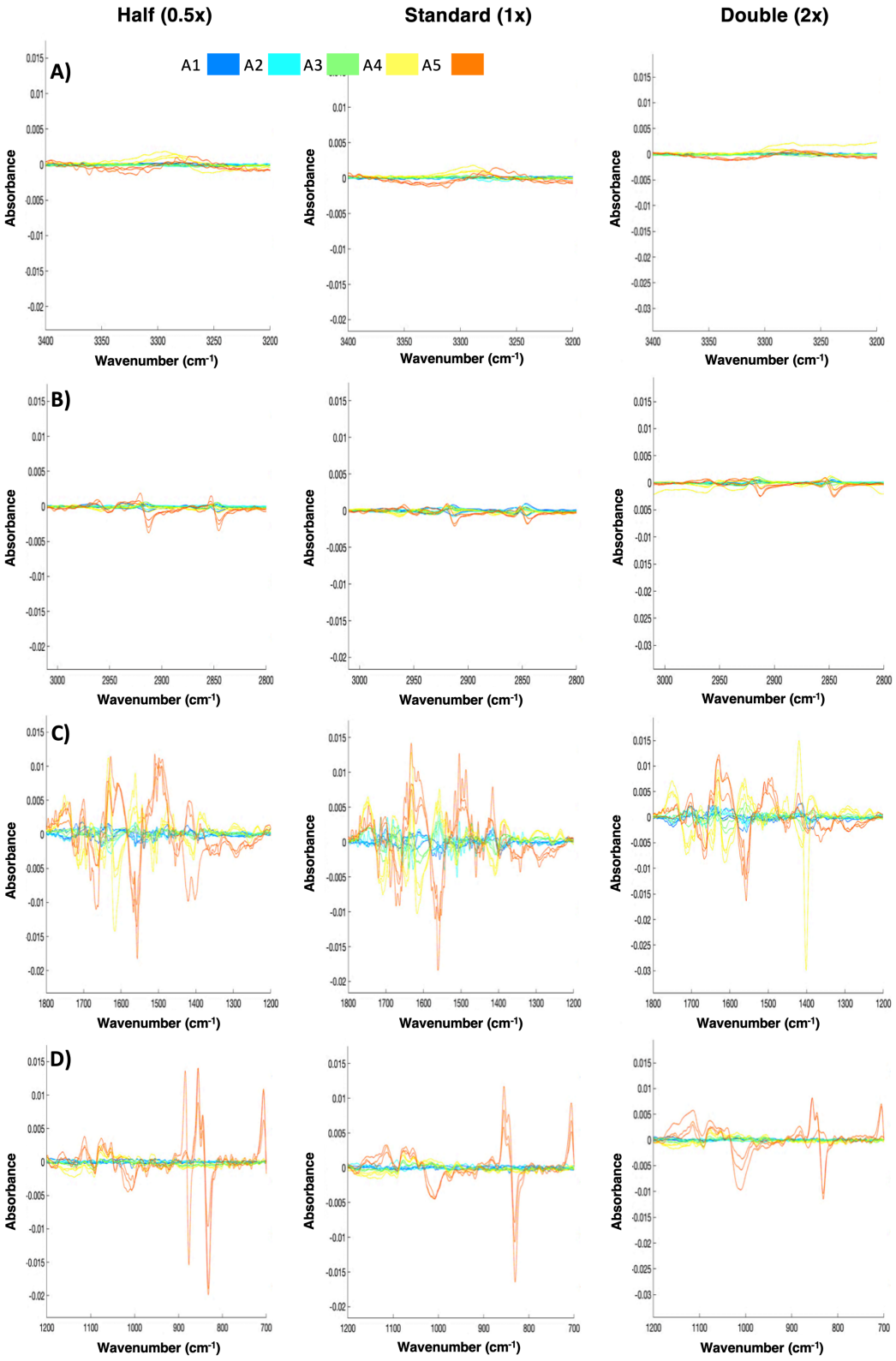
Supplementary Figure 2. Stress response signatures after PCA-HCA analysis for the regions between A) 3400-3200 cm^{-1}, B) 3010-2800 cm^{-1}, C) 1800-1200 cm^{-1} and D) 1200-700 cm^{-1} obtained after immediate, one-shift and overnight exposure to Ethanol (A1), Sodium hypochlorite (A2), Sodium chloride (A3), Hydrochloric acid (A4) and Sodium Hydroxide (A5).



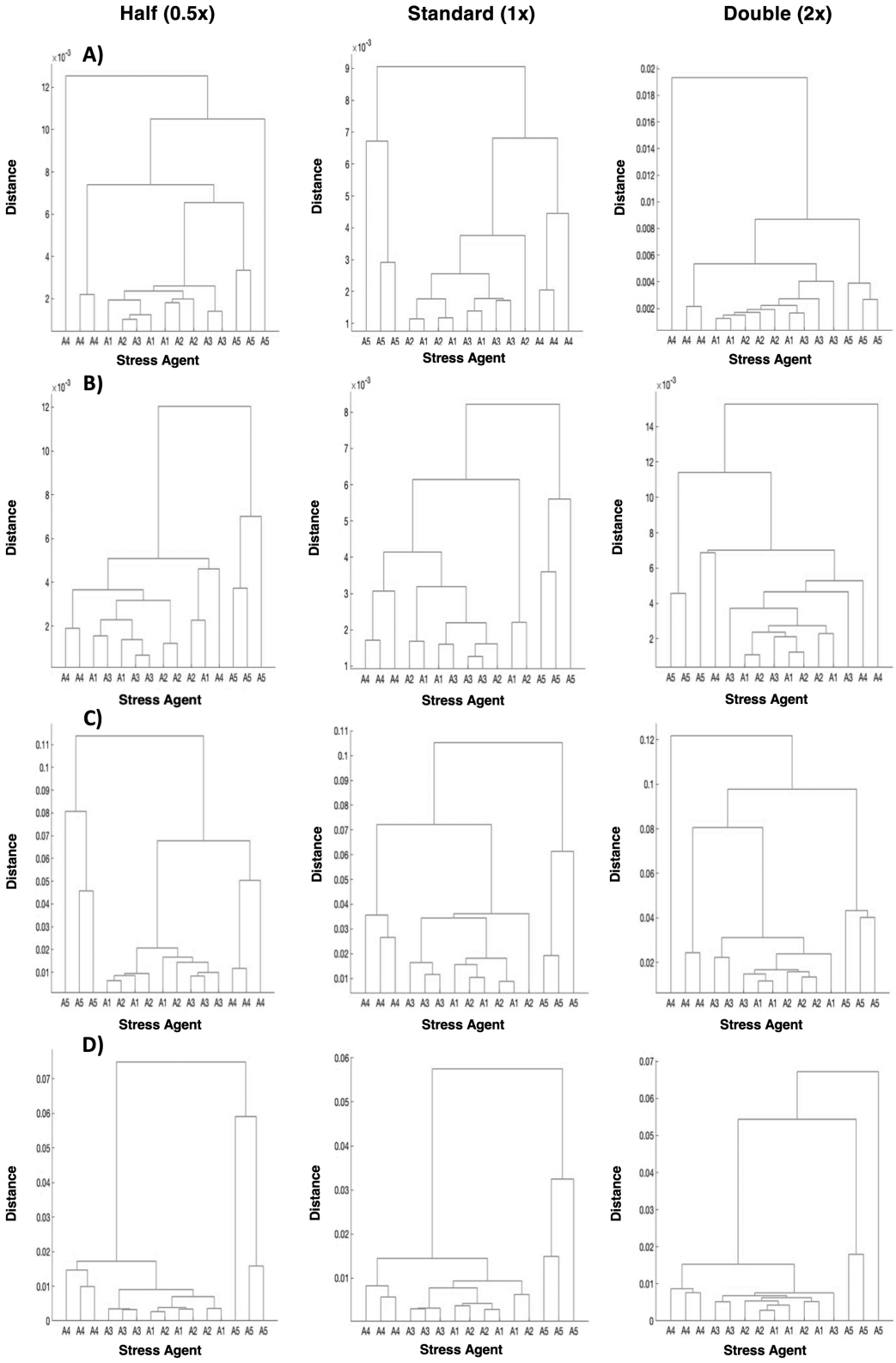
Supplementary Figure 3. Stress response spectral signatures for the regions between A) 3400-3200 cm^{-1}, B) 3010-2800 cm^{-1}, C) 1800-1200 cm^{-1} and D) 1200-700 cm^{-1} obtained after mid-exponential, early- and late-stationary *E. coli* growth phase exposure to Ethanol (A1), Sodium hypochlorite (A2), Sodium chloride (A3), Hydrochloric acid (A4) and Sodium Hydroxide (A5).



Supplementary Figure 4. Stress response signatures after PCA-HCA analysis for the regions between A) 3400-3200 cm^{-1}, B) 3010-2800 cm^{-1}, C) 1800-1200 cm^{-1} and D) 1200-700 cm^{-1} obtained after mid-exponential, early- and late-stationary *E. coli* growth phase exposure to Ethanol (A1), Sodium hypochlorite (A2), Sodium chloride (A3), Hydrochloric acid (A4) and Sodium Hydroxide (A5).



Supplementary Figure 5. Stress response spectral signatures for the regions between A) 3400-3200 cm^{-1}, B) 3010-2800 cm^{-1}, C) 1800-1200 cm^{-1} and D) 1200-700 cm^{-1} obtained using half, standard and double nutrient concentration during exposure to Ethanol (A1), Sodium hypochlorite (A2), Sodium chloride (A3), Hydrochloric acid (A4) and Sodium Hydroxide (A5).



Supplementary Figure 6. Stress response signatures after PCA-HCA analysis for the regions between A) 3400-3200 cm^{-1}, B) 3010-2800 cm^{-1}, C) 1800-1200 cm^{-1} and D) 1200-700 cm^{-1} obtained using half, standard and double nutrient concentration during exposure to Ethanol (A1), Sodium hypochlorite (A2), Sodium chloride (A3), Hydrochloric acid (A4) and Sodium Hydroxide (A5).

VI. 7 REFERENCES

1. Boor, K.J. Bacterial stress responses: What doesn't kill them can make them stronger. *PLoS Biol.* **2006**, *4*, 0018–0020.
2. Fang, F.C.; Frawley, E.R.; Tapscott, T.; Vázquez-Torres, A. Bacterial Stress Responses during Host Infection. *Cell Host Microbe* **2016**, *20*, 133–143.
3. Gomes Neto, N.J.; Magnani, M.; Chueca, B.; García-Gonzalo, D.; Pagán, R.; de Souza, E.L. Influence of general stress-response alternative sigma factors σS (RpoS) and σB (SigB) on bacterial tolerance to the essential oils from Origanum vulgare L. and Rosmarinus officinalis L. and pulsed electric fields. *Int. J. Food Microbiol.* **2015**, *211*, 32–37.
4. Fang, F.C.; Frawley, E.R.; Tapscott, T.; Vázquez-Torres, A. Discrimination and Integration of Stress Signals by Pathogenic Bacteria. *Cell Host Microbe* **2016**, *20*, 144–153.
5. Mathis, R.; Ackermann, M. Response of single bacterial cells to stress gives rise to complex history dependence at the population level. *Proc. Natl. Acad. Sci.* **2016**, *113*, 4224–4229.
6. Hammer, N.D.; Cassat, J.E.; Noto, M.J.; Lojek, L.J.; Chadha, A.D.; Schmitz, J.E.; Creech, C.B.; Skaar, E.P. Inter- and intraspecies metabolite exchange promotes virulence of antibiotic-resistant staphylococcus aureus. *Cell Host Microbe* **2014**, *16*, 531–537.
7. Wolter, D.J.; Hoffman, L.R. Strength in diversity. *Cell Host Microbe* **2014**, *16*, 427–429.
8. Stewart, P.S.; Franklin, M.J.; Williamson, K.S.; Folsom, J.P.; Boegli, L.; James, G.A. Contribution of stress responses to antibiotic tolerance in Pseudomonas aeruginosa biofilms. *Antimicrob. Agents Chemother.* **2015**, *59*, 3838–3847.
9. Schindler, B.D.; Kaatz, G.W. Multidrug efflux pumps of Gram-positive bacteria. *Drug Resist. Updat.* **2016**, *27*, 1–13.
10. Yoon, Y.; Lee, H.; Lee, S.; Kim, S.; Choi, K.H. Membrane fluidity-related adaptive response mechanisms of foodborne bacterial pathogens under environmental stresses. *Food Res. Int.* **2015**, *72*, 25–36.
11. Fruci, M.; Poole, K. Bacterial Stress Responses as Determinants of Antimicrobial Resistance. *Stress Environ. Regul. Gene Expr. Adapt. Bact.* **2016**, *1*, 115–136.
12. Mitosch, K.; Rieckh, G.; Bollenbach, T. Noisy Response to Antibiotic Stress Predicts Subsequent Single-Cell Survival in an Acidic Environment. *Cell Syst.* **2017**, *4*, 393-403.e5.
13. Mathieu, A.; Fleurier, S.; Frénoy, A.; Dairou, J.; Bredeche, M.F.; Sanchez-Vizuete, P.; Song, X.; Matic, I. Discovery and Function of a General Core Hormetic Stress Response in E. coli Induced by Sublethal Concentrations of Antibiotics. *Cell Rep.* **2016**, *17*, 46–57.
14. Zampieri, M.; Zimmermann, M.; Claassen, M.; Sauer, U. Nontargeted Metabolomics Reveals the Multilevel Response to Antibiotic Perturbations. *Cell Rep.* **2017**, *19*, 1214–1228.
15. Belenky, P.; Ye, J.D.; Porter, C.B.M.; Cohen, N.R.; Lobritz, M.A.; Ferrante, T.; Jain, S.; Korry, B.J.; Schwarz, E.G.; Walker, G.C.; et al. Bactericidal Antibiotics Induce Toxic Metabolic Perturbations that Lead to Cellular Damage. *Cell Rep.* **2015**, *13*, 968–980.
16. Kosmides, A.K.; Kamisoglu, K.; Calvano, S.E.; Corbett, S.A.; Androulakis, I.P. Metabolomic Fingerprinting: Challenges and Opportunities. *Crit. Rev. Biomed. Eng.* **2013**, *41*, 205–221.
17. Birrell, G.W.; Brown, J.A.; Wu, H.I.; Giaever, G.; Chu, A.M.; Davis, R.W.; Brown, J.M. Transcriptional response of Saccharomyces cerevisiae to DNA-damaging agents does not identify the genes that protect against these agents. *Proc. Natl. Acad. Sci.* **2002**, *99*, 8778–8783.
18. Giaever, G.; Chu, A.M.; Ni, L.; Connelly, C.; Riles, L.; Véronneau, S.; Dow, S.; Lucau-Danila, A.; Anderson, K.; André, B.; et al. Functional profiling of the Saccharomyces cerevisiae genome. *Nature* **2002**, *418*, 387.
19. Deutschbauer, A.; Price, M.N.; Wetmore, K.M.; Shao, W.; Baumohl, J.K.; Xu, Z.; Nguyen, M.; Tamse, R.; Davis, R.W.; Arkin, A.P. Evidence-based annotation of gene function in Shewanella

oneidensis MR-1 using genome-wide fitness profiling across 121 conditions. *PLoS Genet.* **2011**, *7*.

20. Price, M.N.; Deutschbauer, A.M.; Skerker, J.M.; Wetmore, K.M.; Ruths, T.; Mar, J.S.; Kuehl, J. V.; Shao, W.; Arkin, A.P. Indirect and suboptimal control of gene expression is widespread in bacteria. *Mol. Syst. Biol.* **2013**, *9*, 1–18.
21. Guimaraes, J.C.; Rocha, M.; Arkin, A.P. Transcript level and sequence determinants of protein abundance and noise in Escherichia coli. *Nucleic Acids Res.* **2014**, *42*, 4791–4799.
22. Turner, K.H.; Everett, J.; Trivedi, U.; Rumbaugh, K.P.; Whiteley, M. Requirements for Pseudomonas aeruginosa Acute Burn and Chronic Surgical Wound Infection. *PLoS Genet.* **2014**, *10*.
23. Xie, L.; Liu, W.; Li, Q.; Chen, S.; Xu, M.; Huang, Q.; Zeng, J.; Zhou, M.; Xie, J. First succinyl-proteome profiling of extensively drug-resistant Mycobacterium tuberculosis revealed involvement of succinylation in cellular physiology. *J. Proteome Res.* **2015**, *14*, 107–119.
24. Keren, L.; Hausser, J.; Lotan-Pompan, M.; Vainberg Slutskin, I.; Alisar, H.; Kaminski, S.; Weinberger, A.; Alon, U.; Milo, R.; Segal, E. Massively Parallel Interrogation of the Effects of Gene Expression Levels on Fitness. *Cell* **2016**, *166*, 1282-1294.e18.
25. Feder, M.E.; Walser, J.C. The biological limitations of transcriptomics in elucidating stress and stress responses. *J. Evol. Biol.* **2005**, *18*, 901–910.
26. Jensen, P.A.; Zhu, Z.; van Opijnen, T. Antibiotics Disrupt Coordination between Transcriptional and Phenotypic Stress Responses in Pathogenic Bacteria. *Cell Rep.* **2017**, *20*, 1705–1716.
27. Liang, Q.; Wang, C.; Li, B.; Zhang, A. Metabolic fingerprinting to understand therapeutic effects and mechanisms of silybin on acute liver damage in rat. *Pharmacogn. Mag.* **2015**, *11*, 586.
28. Krishnan, P.; Kruger, N.J.; Ratcliffe, R.G. Metabolite fingerprinting and profiling in plants using NMR. *J. Exp. Bot.* **2005**, *56*, 255–265.
29. Athamneh, A.I.M.; Alajlouni, R.A.; Wallace, R.S.; Seleem, M.N.; Sengera, R.S. Phenotypic profiling of antibiotic response signatures in Escherichia coli using raman spectroscopy. *Antimicrob. Agents Chemother.* **2014**, *58*, 1302–1314.
30. Sales, K.C.; Rosa, F.; Cunha, B.R.; Sampaio, P.N.; Lopes, M.B.; Calado, C.R.C. Metabolic profiling of recombinant Escherichia coli cultivations based on high-throughput FT-MIR spectroscopic analysis. *Biotechnol. Prog.* **2017**, *33*, 285–298.
31. Lopes, M.B.; Gonçalves, G.A.L.; Felício-Silva, D.; Prather, K.L.J.; Monteiro, G.A.; Prazeres, D.M.F.; Calado, C.R.C. In situ NIR spectroscopy monitoring of plasmid production processes: Effect of producing strain, medium composition and the cultivation strategy. *J. Chem. Technol. Biotechnol.* **2015**, *90*, 255–261.
32. Rosa, F.; Sales, K.C.; Carmelo, J.G.; Fernandes-Platzgummer, A.; da Silva, C.L.; Lopes, M.B.; Calado, C.R.C. Monitoring the *ex-vivo* expansion of human mesenchymal stem/stromal cells in xeno-free microcarrier-based reactor systems by MIR spectroscopy. *Biotechnol. Prog.* **2016**, *32*, 447–455.
33. Sampaio, P.N.; Sales, K.C.; Rosa, F.O.; Lopes, M.B.; Calado, C.R.C. High-throughput FTIR-based bioprocess analysis of recombinant cyprosin production. *J. Ind. Microbiol. Biotechnol.* **2017**, *44*, 49–61.
34. Bellisola, G.; Sorio, C. Infrared spectroscopy and microscopy in cancer research and diagnosis. *Am. J. Cancer Res.* **2012**, *2*, 1–21.
35. Álvarez-Ordóñez, A.; Carvajal, A.; Arguello, H.; Martínez-Lobo, F.J.; Naharro, G.; Rubio, P. Antibacterial activity and mode of action of a commercial citrus fruit extract. *J. Appl. Microbiol.* **2013**, *115*, 50–60.
36. Corte, L.; Tiecco, M.; Roscini, L.; De Vincenzi, S.; Colabella, C.; Germani, R.; Tascini, C.; Cardinali, G. FTIR metabolomic fingerprint reveals different modes of action exerted by structural variants of N-alkyltropinium bromide surfactants on Escherichia coli and Listeria innocua cells.

PLoS One **2015**, *10*, 1–15.

37. Moen, B.; Janbu, A.O.; Langsrud, S.; Langsrud, Ø.; Hobman, J.L.; Constantinidou, C.; Kohler, A.; Rudi, K. Global responses of *Escherichia coli* to adverse conditions determined by microarrays and FT-IR spectroscopy. *Can. J. Microbiol.* **2009**, *55*, 714–728.
38. Corte, L.; Rellini, P.; Roscini, L.; Fatichenti, F.; Cardinali, G. Development of a novel, FTIR (Fourier transform infrared spectroscopy) based, yeast bioassay for toxicity testing and stress response study. *Anal. Chim. Acta* **2010**, *659*, 258–265.
39. Bianchi, A.A.; Baneyx, F. Stress responses as a tool to detect and characterize the mode of action of antibacterial agents. *Appl. Environ. Microbiol.* **1999**, *65*, 5023–5027.
40. Chung, H.J.; Bang, W.; Drake, M.A. Stress response of Escherichia coli. *Compr. Rev. Food Sci. Food Saf.* **2006**, *5*, 52–64.
41. Foster, J.W.; Richard, H. Escherichia coli glutamate- and arginine-dependent acid resistance systems increase internal pH and reverse transmembrane potential. *J. Bacteriol.* **2004**, *186*, 6032–41.
42. Sharma, M.; Beuchat, L.R. Sensitivity of Escherichia coli O157:H7 to Commercially Available Alkaline Cleaners and Subsequent Resistance to Heat and Sanitizers. *Appl. Environ. Microbiol.* **2004**, *70*, 1795–1803.
43. Kwak, S.J.; Jo, H.J.; Yoon, K.S. Effect of sodium hypochlorite (NaClO) sanitizer-induced stress on growth kinetics and morphological changes in Escherichia coli and Bacillus cereus spores. *Food Sci. Biotechnol.* **2014**, *23*, 815–821.
44. Kim, N.H.; Rhee, M.S. Phytic acid and sodium chloride show marked synergistic bactericidal effects against nonadapted and acid-adapted escherichia coli O157:H7 strains. *Appl. Environ. Microbiol.* **2016**, *82*, 1040–1049.
45. Cao, H.; Wei, D.; Yang, Y.; Shang, Y.; Li, G.; Zhou, Y.; Ma, Q.; Xu, Y. Systems-level understanding of ethanol-induced stresses and adaptation in E. coli. *Sci. Rep.* **2017**, *7*, 1–15.
46. Calado, C.; Monteiro, S.M.S.; Cabral, J.M.S.; Fonseca, L.P. Effect of pre-fermentation on the production of cutinase by a recombinant Saccharomyces cerevisiae. *J. Biosci. Bioeng.* **2002**, *93*, 354–359.
47. Maity, J.P.; Kar, S.; Lin, C.M.; Chen, C.Y.; Chang, Y.F.; Jean, J.S.; Kulp, T.R. Identification and discrimination of bacteria using Fourier transform infrared spectroscopy. *Spectrochim. Acta - Part A Mol. Biomol. Spectrosc.* **2013**, *116*, 478–484.
48. Wang, S.; Deng, K.; Zaremba, S.; Deng, X.; Lin, C.; Wang, Q.; Tortorello, M. Lou; Zhang, W. Transcriptomic response of Escherichia coli O157:H7 to oxidative stress. *Appl. Environ. Microbiol.* **2009**, *75*, 6110–6123.
49. Soufi, B.; Krug, K.; Harst, A.; Macek, B. Characterization of the E. coli proteome and its modifications during growth and ethanol stres. *Front. Microbiol.* **2015**, *6*, 1–11.
50. Hajmeer, M.; Ceylan, E.; Marsden, J.L.; Fung, D.Y.C. Impact of sodium chloride on Escherichia coli O157:H7 and Staphylococcus aureus analysed using transmission electron microscopy. *Food Microbiol.* **2006**, *23*, 446–452.
51. Lee, S.Y.; Kang, D.H. Survival mechanism of Escherichia coli O157: H7 against combined treatment with acetic acid and sodium chloride. *Food Microbiol.* **2016**, *55*, 95–104.
52. Omotoyinbo, O. V; Omotoyinbo, B.I. Effect of Varying NaCl Concentrations on the Growth Curve of Escherichia coli and Staphylococcus aureus. *Cell Biol.* **2016**, *4*, 31–34.
53. Lee, S.Y.; Kang, D.H. Combined effects of heat, acetic acid, and salt for inactivating Escherichia coli O157:H7 in laboratory media. *Food Control* **2009**, *20*, 1006–1012.
54. Richard, H.T.; Foster, J.W. Acid resistance in Escherichia coli. *Adv. Appl. Microbiol.* **2003**, *52*, 167–186.
55. Foster, J.W. Escherichia coli acid resistance: Tales of an amateur acidophile. *Nat. Rev. Microbiol.*

2004, *2*, 898–907.

56. Kanjee, U.; Houry, W.A. Mechanisms of acid resistance in Escherichia coli. *Annu. Rev. Microbiol.* **2013**, *67*, 65–81.
57. Saito, H.; Kobayashi, H. Bacterial responses to alkaline stress. *Sci. Prog.* **2003**, *86*, 271–282.
58. Nyanga-Koumou, A.P.; Ouoba, L.I.I.; Kobawila, S.C.; Louembe, D. Response mechanisms of lactic acid bacteria to alkaline environments: A review. *Crit. Rev. Microbiol.* **2012**, *38*, 185–190.
59. Corte, L.; Antonielli, L.; Roscini, L.; Fatichenti, F.; Cardinali, G. Influence of cell parameters in Fourier Transform InfraRed spectroscopy analysis of whole yeast cells. *Analyst* **2011**, *136*, 2339–2349.
60. Kang, S.-N.; Kim, K.-J.; Park, J.-H.; Kim, K.-T.; Lee, O.-H. Effect of antimicrobial activity of sodium hypochlorite and organic acids on various foodborne pathogens in Korean ginseng root. *African J. Microbiol. Res.* **2013**, *7*, 2724–2729.
61. Aboualizadeh, E.; Bumah, V. V.; Masson-Meyers, D.S.; Eells, J.T.; Hirschmugl, C.J.; Enwemeka, C.S. Understanding the antimicrobial activity of selected disinfectants against methicillin-resistant Staphylococcus aureus (MRSA). *PLoS One* **2017**, *12*, 1–15.
62. Huleihel, M.; Pavlov, V.; Erukhimovitch, V. The use of FTIR microscopy for the evaluation of anti-bacterial agents activity. *J. Photochem. Photobiol. B.* **2009**, *96*, 17–23.
63. Skotti, E.; Kountouri, S.; Bouchagier, P.; Tsitsigiannis, D.I.; Polissiou, M.; Tarantilis, P.A. FTIR spectroscopic evaluation of changes in the cellular biochemical composition of the phytopathogenic fungus Alternaria alternata induced by extracts of some Greek medicinal and aromatic plants. *Spectrochim. Acta - Part A Mol. Biomol. Spectrosc.* **2014**, *127*, 463–472.
64. Sultanbawa, Y.; Cozzolino, D.; Fuller, S.; Cusack, A.; Currie, M.; Smyth, H. Infrared spectroscopy as a rapid tool to detect methylglyoxal and antibacterial activity in Australian honeys. *Food Chem.* **2015**, *172*, 207–212.

Chapter VII

Elucidating the mechanism of action of antibiotics

This chapter is adapted from the manuscript:

Ribeiro Da Cunha, B., Fonseca, L. P. and Calado, C. R. C. (2020) 'Metabolic fingerprinting with Fourier-transform infrared (FTIR) spectroscopy: Towards a high-throughput screening assay for antibiotic discovery and mechanism-of-action elucidation', Metabolites, 10(4). doi: 10.3390/metabo10040145.

Author contribution

Bernardo Ribeiro da Cunha conceptualized the study methodology, conducted the investigation and formal analysis, prepared the original draft, reviewed and edited its final version.

Abstract

The discovery of antibiotics has been slowing to a halt. Phenotypic screening is once again at the forefront of antibiotic discovery, yet mechanism of action (MOA) identification is still a major bottleneck. As such, methods capable of MOA elucidation are required now more than ever, for which Fourier-Transform Infrared (FTIR) spectroscopy (FTIRS) is a promising metabolic fingerprinting technique. A whole-cell FTIRS-based bioassay was developed to reveal the metabolic fingerprint induced by 15 antibiotics on the *Escherichia coli* metabolism. Cells were briefly exposed to four times the minimum inhibitory concentration and spectra were quickly acquired using 96-well plates. After preprocessing optimization, a partial least squares discriminant analysis and principal component analysis were conducted. The metabolic fingerprints obtained with FTIRS were sufficiently specific to allow a clear distinction between different antibiotics, across three independent cultures, with either analysis algorithm. These fingerprints were coherent with the known MOA of all the antibiotics tested, which include examples that target the protein, DNA, RNA, and cell wall biosynthesis. Because FTIRS acquires a holistic fingerprint of the effect of antibiotics on the cellular metabolism, it holds great potential to be used for high-throughput screening in antibiotic discovery and possibly towards a better understanding of the MOA of current antibiotics.

Contents

VII. 1 INTRODUCTION

The discovery of antibiotics has been considered a miracle of modern medicine, but since the golden age of antibiotic discovery, when most classes were introduced, innovation has been slowing to a halt [1]. The genomics era inspired target-based screening but hits generally proved ineffective at reaching their target within the cell and not a single new drug reached the market from target-based screening programs [2]. Nowadays, phenotypic screening is a preferred strategy for antibiotic discovery, mostly because compounds that are effective against whole cells have a higher likelihood of becoming candidate molecules and can target poorly understood metabolic pathways [3].

However, these assays do not reveal the mechanism of action (MOA) of candidate compounds, which requires considerable efforts at a later stage of the discovery process [4]. This results in higher rates of rediscovery, low probabilities of finding compounds with unique biological and/or chemical properties, and limited insight in the pharmacological target. Additionally, phenotypic screening does not explore the chemical grey matter, i.e., compounds capable of inducing some level of phenotypic modulation, but without sufficient potency to induce cell death or growth inhibition, which can be a source of compounds suitable for lead optimization with medicinal chemistry techniques [5,6].

Antibiotic discovery has proven to be a tremendous challenge, but identifying the MOA has proven equally challenging [7]. In fact, determining the MOA of antibiotics is still a bottleneck of the phenotypic screening discovery process, for which metabolomics holds great potential. As such, the ability to rapidly infer MOA and, if possible, the biomolecular target of antibiotics is increasingly important given the pressing need for new antibiotics. An antibiotic discovery program can require screening hundreds of thousands of compounds, in part due to the ease in synthetizing bioactive compounds, and in part given the increasing availability of natural product libraries [8].

Two concepts are relevant when discussing MOA identification. One is determining the molecular pathways affected by a given compound: the drug effects. The second is the specific compound—substrate interactions: the drug target [9]. Although both concepts are very important in antibiotic discovery, given the exploratory purpose of this study, a less stringent definition of MOA identification was used. Herein, we refer to MOA elucidation as an approximation using a mechanism-specific fingerprint, rather than the identification of the

specific targets of a given molecule, and the pathways affected, which formally constitute MOA identification.

Conventional MOA studies are based on macromolecular synthesis assays; however, these assays are typically slow, laborious, low resolution, low accuracy, and low throughput [10]. An equally limited alternative lies in biochemical approaches, for instance the use of affinity chromatography to identify the exact biomolecule to which a candidate molecule binds [11,12]. However, this is somewhat a fishing expedition in the sense that it requires the happy combination of a high-affinity small molecule with a fairly abundant protein receptor [13]. Another important issue of conventional MOA assays is that a large quantity of the test compound is required, which is not always attainable.

Recently, genome-wide transcriptional or translational profiles have been used to reveal the target of candidate molecules, but, more often than not, these profiles overwhelmingly reflect indirect stress responses rather than the specific sequence of events that results in the inhibitory effect [14]. Since the metabolome is at the lower end of the Omics cascade, it reflects the substrates and products of various metabolic enzymes, and thereby can be used to pinpoint drug-induced inhibition. As such, early metabolomics studies into the MOA of antibiotics looked at shifts of metabolite concentrations, induced by a single molecule, to identify its specific target(s) [7,15].

As the understanding of MOA grew more complex, from targets to networks, metabolomics has been increasingly used to build comprehensive multi-parametric profiles of the MOA [16]. These profiles provide a genome-scale characterization of the drug-induced effects, which extends beyond non-metabolic targets [17]. One advantage of metabolic profiling is that it considers both on-target and off-target effects, which together produce the antibiotic effect, via an underlying MOA, of a molecule. Therefore, metabolomics studies on MOA focused on obtaining metabolic profiles. The comparative metabolic profiling of a pair of isogenic methicillin-susceptible and -resistant *Staphylococcus aureus* has emphasized the metabolic alterations that are specific to the MOA of three antibiotics acting on the major biosynthetic pathways, i.e., cell wall, DNA, and protein biosynthesis [18].

Additionally, the metabolic profiling of drug exposure has also been used, together with metabolic and chemogenomic profiles of single-deletion strains, to predict epistatic drug interactions. This enables the rational design of drug combinations by identifying nonantibiotic compounds that, when combined, have antimicrobial activity [19]. Metabolic profiling also plays

a role in the dereplication and guided fractioning of novel natural products with antimicrobial properties [20].

Regarding studies dedicated towards high-throughput MOA elucidation, nuclear magnetic resonance has the advantage of being faster and less expensive, albeit only capable of identifying highly abundant metabolites. One approach to tackle this has been to analyze both the intracellular fingerprints and extracellular footprints, which resulted in more comprehensive and specific metabolic profiles [21]. Another approach relies on Mass Spectrometry (MS), which can be used to identify a wider range of metabolites with higher sensitivity, especially when combined with chromatographic separation techniques. Until recently, MS-based metabolomics could only be applied to study the MOA of individual molecules, but sufficient throughput can be attained with an untargeted approach, where thousands of ion peaks are detected from individual samples, although annotation is still the bottleneck, i.e., identifying metabolites from said peaks [22].

In fact, untargeted metabolomics can achieve the required throughput to systematically ascertain the MOA of moderately large collections of antibiotics [23]. However, to reach the 10–100x increase in throughput required for large-scale studies, there has to be a compromise of either coverage and/or separation, which currently requires dismissing the chromatographic step [24]. As such, by sacrificing resolution, high-throughput untargeted metabolomics using flow injection electrospray has predicted the MOA of uncharacterized antimycobacterial compounds from an industry-scale chemical library [25]. Within said compromise of resolution in favor of throughput, other analytical techniques could be better suited for the rapid handling of a large number of samples, with minimal preparation and manipulation, though they inherently yield less informative data.

Fourier-Transform Infrared (FTIR) spectroscopy (FTIRS) is an established metabolic fingerprinting technique particularly well suited for high throughput, which requires minimal sample handling, and is reagent-free and label-free [26]. Moreover, FTIRS is a metabolic fingerprinting technique that provides relevant chemical information to rapidly and reproducibly discern prominent changes in the metabolome [27], particularly those imposed by stress agents and antibiotics [28]. Additionally, FTIR microscopy has been successfully used to identify antibiotic resistance from clinical isolates [29]. FTIRS is especially promising because its characteristics bridge the gap between the low-throughput/high-information metabolomic assays and the high-throughput/low-information nature of phenotypic screening assays.

While FTIRS does not result in comprehensive data on the metabolite level, the volume of biological information it yields allows for an enhanced assessment of the biomolecular phenomena underlying the antibiotic effect, which has been shown to be suitable in MOA-centric studies. In fact, some studies have dwelled into the ability of FTIRS in detecting antibiotic-specific fingerprints [29–32]. However, these studies either did not explore the full potential of high-throughput or were limited to a small number of antibiotics with considerably distinct MOA. Thus, it remains unclear whether FTIRS can be used to rapidly obtain metabolic profiles with sufficient sensitivity to distinguish very similar MOA, e.g., antibiotics acting on the same major biosynthetic pathway, and within, antibiotics of the same class.

The aim of this study was to explore FTIRS as a metabolic fingerprinting technique for the determination of antibiotics' MOA and pave the way towards a high-throughput protocol. The fingerprint induced by 15 antibiotics on the *Escherichia coli* metabolism was sufficiently specific to allow for a clear distinction between the different antibiotics across three independent cultures, conducted on different days. The range of antibiotics tested includes nine classes acting on key biosynthetic pathways: protein, DNA, RNA, and cell wall synthesis. *E. coli* cells were exposed to the antibiotics at four times the minimum inhibitory concentration for three hours, which ensured an inactivation of at least 90% across all samples. Spectra were preprocessed with the Savitzky–Golay (SG) derivative filter followed by Loopy Multiplicative Scatter Correction (LMSC). After both SG and LMSC parameters were optimized, the dataset was analyzed with Partial Least Squares Discriminant Analysis (PLSDA) and Principal Component Analysis (PCA) so as to consolidate the PLSDA predictions with an unsupervised algorithm. Adequate separation between the MOA at the level of major biosynthetic pathway affected was obtained independently of the analysis algorithm, and this separation extended to the level of antibiotic-specific fingerprints, which is a positive indication that FTIRS is suited to the elucidation of antibiotics MOA. Moreover, all samples exposed to an antibiotic were clearly plotted separately from the controls, which is important to identify novel molecules with an antibiotic effect in the context of a screening assay towards antibiotic discovery.

VII. 2 MATERIALS AND METHODS

VII.2.1 ANTIBIOTIC STOCK SOLUTIONS AND SUSCEPTIBILITY TESTING

Antibiotic stock solutions of 15 antibiotics, belonging to 9 classes acting on 4 key biosynthetic pathways (Table 1), were prepared at a concentration of 4096 μg/mL, adjusted for potency, and kept at –20 °C or 4 °C, per recommendation. For *in vitro* susceptibility testing,

both the CLSI [33] as well as the EUCAST [34] guidelines were considered . In detail, 100 μL of antibiotic solution was serially diluted in flatbottom 96-well plates, to which 100 μL of fresh cation-adjusted Mueller-Hinton broth (MHB) (VWR, Portugal) was added, along with 5 μL of cell suspension to obtain a concentration of 5×10^5 Colony Forming Units per mL (CFU/mL). The bacteria were incubated at 37 °C for 24 h, after which growth inhibition was observed. Minimum inhibitory concentrations (MICs) were determined as the lowest concentration at which no bacterial growth was observed for three independent cultures, and the inoculum size was confirmed by plating on cation-adjusted Mueller-Hinton agar and determining the CFU/mL.

VII.2.2 Bacterial cultures and antibiotic exposure

The bacterium *E. coli* strain JM 101 (ATCC33876) was chosen as a model organism for its ease of manipulation, non-pathogenic nature (biosafety class 1), as well as lack of resistance mechanisms in its genome: supE thi-1 Δ(lac-proAB) [F´ traD36 proAB lacIqZΔM15]. The bacteria were grown in 1 L erlenmeyers, with 400 mL of MHB, in an orbital incubator (TH30 and SM30, Edmund Buhler GmbH) at 37 °C and 250 rpm. The cells were incubated until OD_{590} reached 0.270 ± 0.03, thereby ensuring cells were in the exponential growth phase. For exposure to each antibiotic, 18 mL of culture broth was transferred to a conical centrifugal tube with 7 mL of antibiotic stock solution to obtain a final concentration of four times the MIC. Cells were perturbed for 3-h in an orbital incubator at 250 rpm, 37 °C. The complete procedure was repeated for three independent cultures over different days. After antibiotic exposure, bacterial inactivation was confirmed by counting CFU/mL, as described elsewhere [35].

VII.2.3 Spectral data acquisition, preprocessing, and multivariate analysis

After incubation with the antibiotics, samples were quickly centrifuged at 3000 relative centrifugal force for 10 min at 4 °C (Rotanta 460R, Hettich Zentrifugen, Germany), the supernatant was discarded, and the cell pellet was resuspended in 25 mL of cold 0.9% NaCl (Merck, Germany) to quench the metabolism. Subsequently, the OD_{590} was taken and bacterial inactivation was determined. The cells were, again, pelleted and resuspended in cold 0.9% NaCl to obtain an OD_{590} of 1, from which 2 mL aliquots were further centrifuged for three minutes at 13,000 rpm (13,793 g) (Z160M, Hermle Labortechnik, Germany) and resuspended in 100 μL of cold 0.9% NaCl to obtain an OD_{590} of 20 for FTIR readings. The samples were then plated on an infrared-transparent ZnSe 96 well plate (Bruker, Germany) in quintuplicates.

ZnSe plates were dehydrated for 3 h in a vacuum desiccator with silica, and inserted in a HTS-XT module coupled to a Vertex-70 spectrometer (Bruker Optics, Germany).

Spectra were acquired in transmission mode and consisted of 40 coadded scans at a 4-cm^{-1} resolution. These were then exported as data point table files, which were imported into MATLAB (MathWorks, USA) for subsequent analysis. To reduce spectra heterogeneity originating from operator handling and other undesirable sources of variability, the spectra of mechanical replicates (quintuplicates) were averaged. Subsequently, averaged spectra were preprocessed with the SG filter and then with LMSC. A range of parameters were used for either algorithm, and those that produced the highest successful classification of an LOO-CV PLSDA model were used to build the final PLSDA predictive model and the PCA model.

VII. 3 RESULTS AND DISCUSSION

VII.3.1 MINIMUM INHIBITORY CONCENTRATIONS AND BACTERIAL INACTIVATION FOR FTIR READINGS

To evaluate FTIRS as a metabolic fingerprinting technique suitable for distinguishing the MOA of different antibiotics, *E. coli* was exposed to 15 compounds belonging to different classes and acting on different key biosynthetic pathways. The MIC of each compound was determined using standard methods, and these have been reported (Table 1). Additionally, the average inactivation of independent cultures exposed to antibiotics prior to FTIR readings was calculated (Table 1).

This was done to verify that four times the MIC was sufficient to obtain a cellular inactivation of over 90% for FTIR measurements. This was particularly important because antibiotic exposure for FTIR measurements was conducted at a cell density three orders of magnitude higher than MIC testing, which was required to obtain sufficient biomass for spectra acquisition. Although the relationship between inoculum size and antimicrobial activity is not linear [36,37], previous studies have shown that four times the MIC is a suitable concentration for metabolomics analysis of antibiotics MOA [23].

Moreover, determining the average inactivation also attested that using a relative antibiotic concentration, e.g., four times the MIC, over an absolute concentration, e.g., 500 μg/mL, resulted in an equivalent antibiotic effect (i.e., equivalent inactivation) between different metabolic fingerprints. For instance, in the case of antibiotics that have both a bacteriostatic and bactericidal action, a rule of thumb is that bacteriostatic activity is determined by a ratio of

Table 1. Classification of the antibiotics tested, the biosynthetic pathway targeted and their Minimum Inhibitory Concentration (MIC). The average bacterial inactivation after exposure to 4 X MIC for 3 h was determined, by plate counting colony forming units, and shown as the percentage variation in regard to the control.

Antibiotic	MIC (μg/mL)	Average Inactivation (%)	Class	Biosynthetic Pathway Targeted
Amoxicillin	8	99.8	Beta-lactam	Cell Wall
Ampicillin	8	100	Beta-lactam	Cell Wall
Cephradine	8	99.7	Beta-lactam	Cell Wall
Chloramphenicol	4	94.3	Amphenicol	Protein
Ciprofloxacin	0.5	100	Fluoroquinolone	DNA
Erythromycin	32	93.2	Macrolide	Protein
Isoniazid	256	93	Other	Other
Kanamycin	8	100	Aminoglycoside	Protein
Levofloxacin	0.125	100	Fluoroquinolone	DNA
Metronidazole	128	96.3	Nitroimidazole	DNA
Neomycin	2	100	Aminoglycoside	Protein
Rifampicin	32	100	Rifamycin	RNA
Sulfamethazine	8	99.8	Sulfonamide	DNA
Sulfamethoxazole	32	98.9	Sulfonamide	DNA
Tobramycin	2	100	Aminoglycoside	Protein

the minimum bactericidal concentration to MIC above four [38]. Exposing *E. coli* to four times the MIC should therefore favor bactericidal activity, which is confirmed by the large inactivation reported. Importantly, this consolidates the notion that any spectral differences observed are most likely due to the specificity of the metabolic adaptations induced by each antibiotic, rather than dose-dependent variations.

Furthermore, a 3-h time point was chosen to avoid unspecific stress responses generically triggered by antibiotic exposure and to ensure the absence of growth recovery after antibiotic exposure, as suggested [23]. Unspecific responses are elicited almost immediately, and antibiotic-specific fingerprints typically become more evident as the exposure duration

increases. In fact, a 30-min exposure at a concentration that minimizes cell death and lysis revealed a common metabolic response among bactericidal antibiotics, which evolved to antibiotic-specific metabolic responses at 60 min and more so at 90 min [39]. Increasing the antibiotic concentration well above the MIC seemingly accelerates antibiotic-specific metabolic adaptations, and fingerprints obtained after 30 min of exposure more accurately reflect the MOA [21]. Other approaches combine the data from multiple timepoints [25], but this reduces the overall throughput, hence it is preferably avoided.

Typically, dimethyl sulfoxide (DMSO) is used as a universal solvent for antibiotic screening, given its ability to dissolve both organic and inorganic compounds. However, DMSO inhibits the rapid killing of diverse classes of antibiotics, even at concentrations as low as 1%. In fact, DMSO has been suggested to interfere with antibiotic lethality that is mediated by Reactive Oxygen Species (ROS), in a concentration and exposure duration dependent manner. As such, this protective effect is not constant across antibiotic classes, which complicates MOA elucidation.

Furthermore, this protective effect is not reflected on MIC values, as there are mechanistic differences between transient ROS-mediated killing and MICs. On the other hand, DMSO can alter cell membrane permeability, which is speculated to explain its inhibitory effect, and can result in an apparent increase in potency, as the entry into the cell by certain antibiotics is facilitated. Ultimately, these findings discourage the use of DMSO as a solvent for antimicrobials, especially in rapid-killing assays [40]. Therefore, in this study we only considered antibiotics that are soluble in aqueous solvents, as conducted by others [21].

VII.3.2 FTIR PREPROCESSING OPTIMIZATION

FTIR spectra are composed of a sample-specific component, which ideally is closely related to the biological information of interest, and an unspecific and undesirable component, which is due to variability induced by environmental, experimental, and technical conditions. The objective of spectral preprocessing is to reduce the latter and highlight the biological relevance of the data [41].

Two commonly used preprocessing strategies are derivative filtering, typically with the SG filter, followed by scattering correction, for instance the LMSC algorithm [42]. Importantly, the performance of these preprocessing strategies depends on their parametrization, and this

in turn differs with the system being studied [43]. To identify parameters that yield optimal predictive performance of the PLSDA model, the successful classification after Leave-One-Out Cross-Validation (LOO-CV) was used (Figure 1).

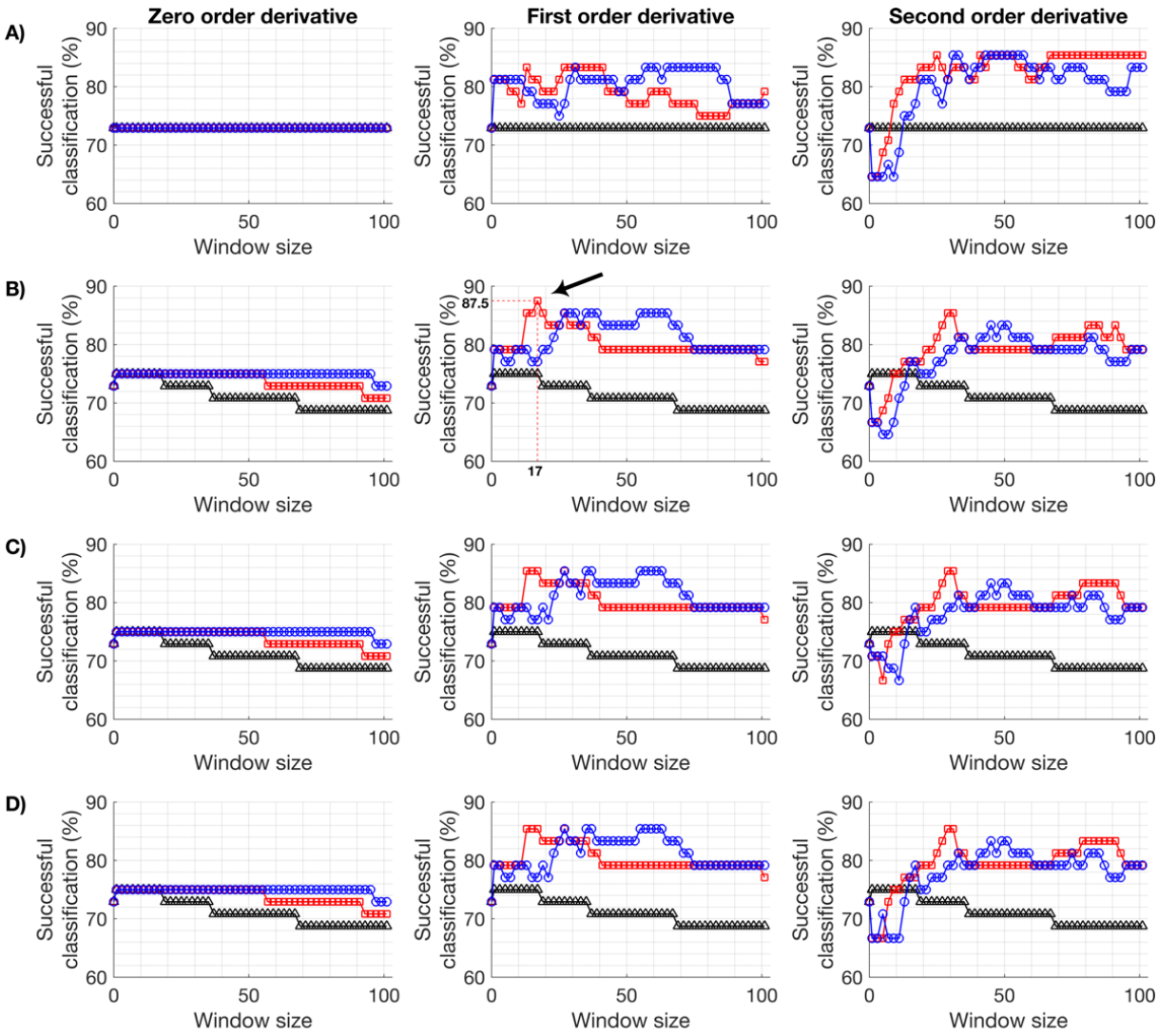


Figure 1. Effect of Savitzky–Golay (SG) and Loopy Multiplicative Scatter Correction (LMSC) parameters on Partial Least Squares Discriminant Analysis (PLSDA) predictive models. For SG filtering, each derivative order (columns) was queried across a constant (black triangle), quadratic (red square), and quartic (blue circle) order polynomial, followed by zero, one, two, and three iterations of LMSC (rows A–D, respectively). The highest successful classification (%) was highlighted (arrow).

A single iteration of LMSC preceded by SG filtering with a window size of 17 datapoints, to which a quadratic polynomial was fitted to determine the first derivative, resulted in the PLSDA model with optimal performance that was discussed over the following sections. The

effect of the optimal preprocessing strategy on the raw spectra has been shown for three antibiotics acting on the major biosynthetic pathways (Figure 2).

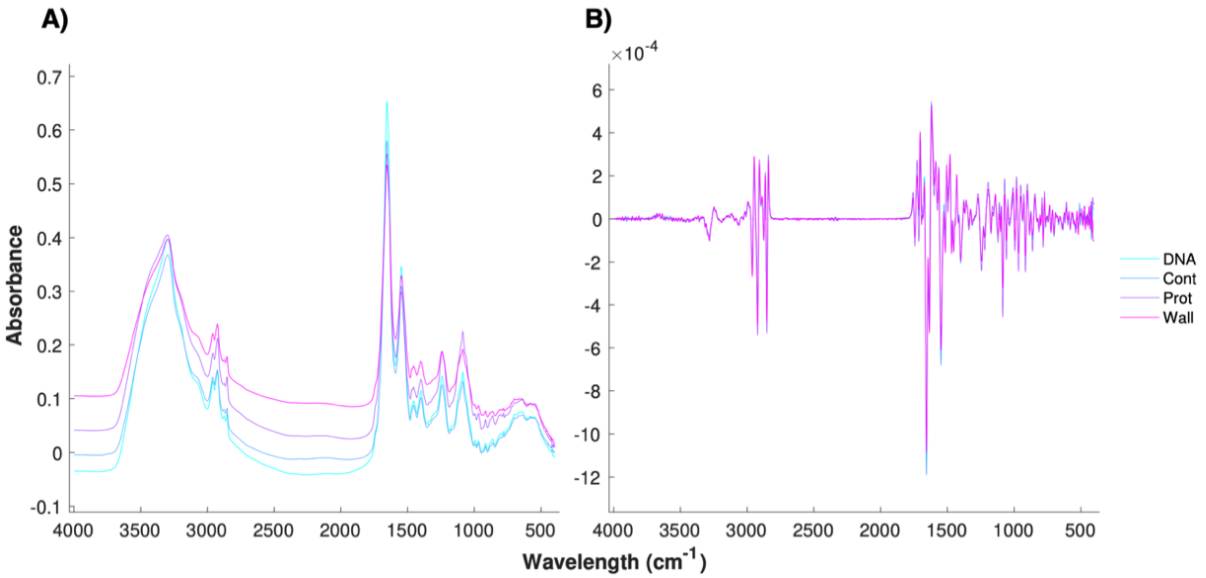


Figure 2. Average spectra of independent cultures exposed to Ciprofloxacin (DNA), Control (Cont), Kanamycin (Prot), and Amoxicillin (Wall) before any manipulation (A) and after the application of the optimal preprocessing strategy (B). These antibiotics were chosen as representatives of the major affected biosynthetic pathways.

VII.3.3 PREDICTING THE MAJOR BIOSYNTHETIC PATHWAY TARGETED

Pinpointing the MOA of a candidate molecule requires identifying its molecular target; however, this is a dauting task that must often be decomposed into smaller elements, the first of which is predicting the major biosynthetic pathway targeted. For that, a PLSDA model was built with the optimized preprocessing parameters (Figure 3). Note that, with these parameters, a successful classification of 87.5% was obtained after LOO-CV (Figure 1).

Importantly, the control samples, i.e., those exposed to the solvent but not the antibiotic, were predicted as different from all the other samples. This is particularly important as it allows one to differentiate the cases where no biosynthetic pathway was affected; therefore, the lack of drug effect can be predicted for candidate molecules that have no metabolic effect. Additionally, the metabolic fingerprints induced by antibiotics targeting each of the major biosynthetic pathways were separated with as little as two latent variables, as a very simple model was sufficient to explain over 99% of the spectral variability. To ensure that the observed

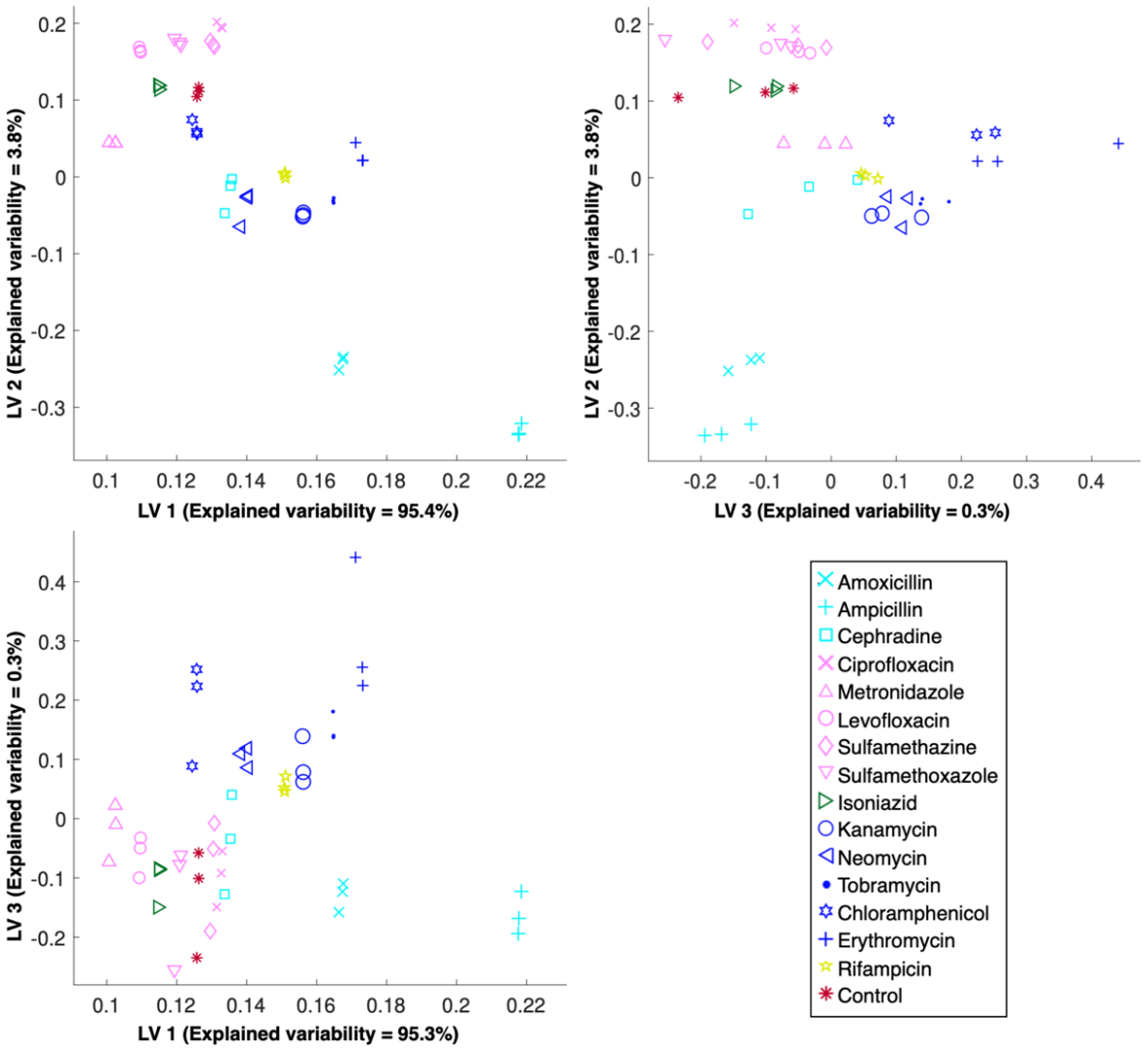


Figure 3. Representation of the metabolic fingerprints induced by antibiotics acting on the major biosynthetic pathways after Partial Least Squares Discriminant Analysis (PLSDA), preceded by an optimized combination of preprocessing algorithms. The variability explained by each Latent Variable (LV) is reported on each corresponding axis. Antibiotics acting on the same major biosynthesis pathway are represented using the same color, and individual antibiotics are represented with a unique symbol.

clusters are indeed intrinsic to the spectra, since PLSDA score plots can often be misleading and misinterpreted [44], a PCA was conducted on the same dataset, preceded by the exact same optimal preprocessing (Figure 4).

The objective of this analysis was to reinforce the PLSDA conclusions, rather than derive new ones. Unlike the predictive model built with PLSDA, PCA is an unsupervised technique that directly reflects the inherent structure of the data. As such, a slightly higher

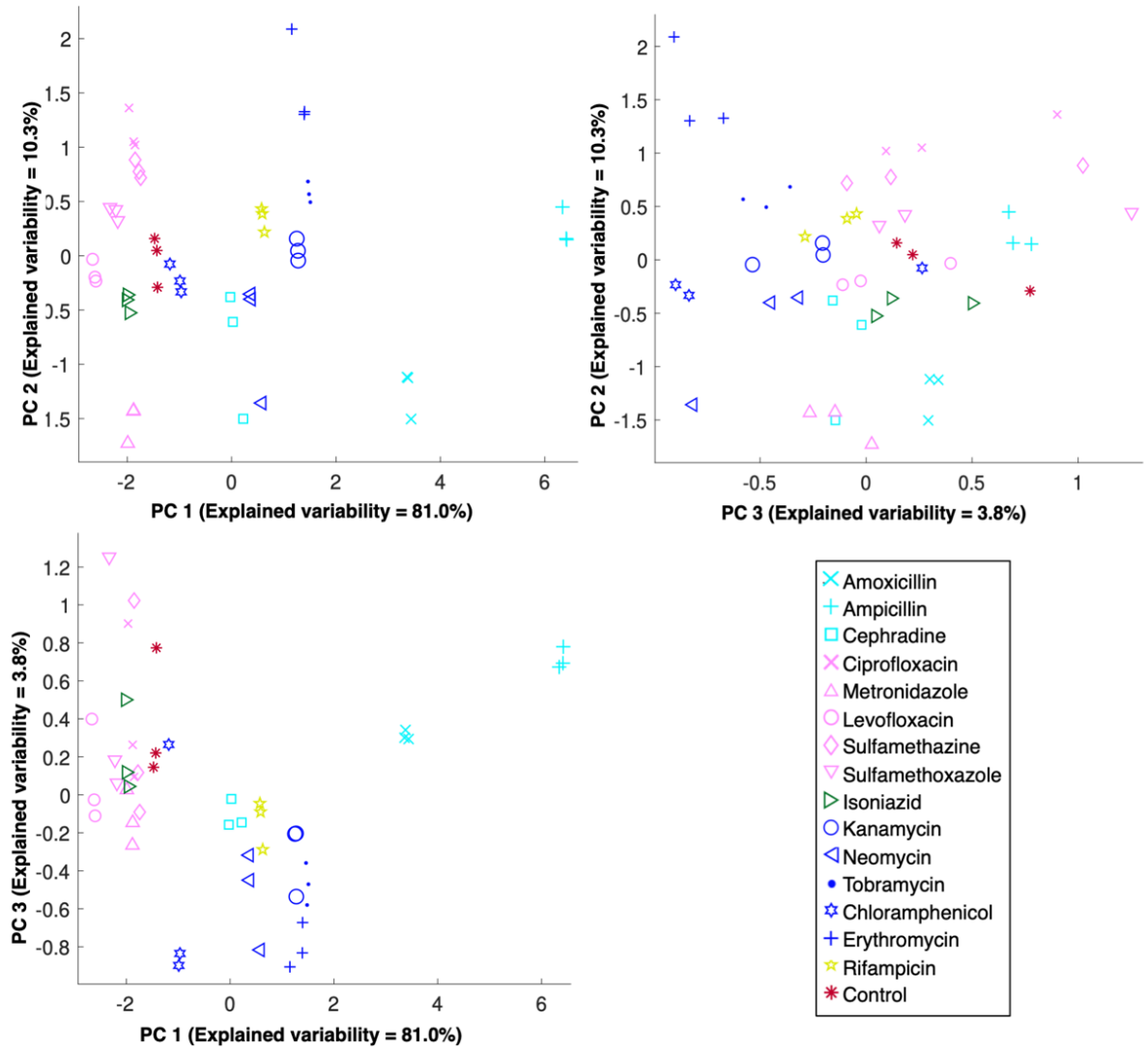


Figure 4. Representation of the metabolic fingerprints induced by 15 antibiotics acting on the major biosynthetic pathways after Principal Component Analysis (PCA), preceded by an optimized combination of preprocessing algorithms. The variability explained by each Principal Component (PC) is reported on each corresponding axis. Antibiotics acting on the same major biosynthesis pathway are represented using the same color, and individual antibiotics are represented with a unique symbol.

intra-replica variability is observed, i.e., biological replicas are slightly more disperse. Ultimately, the similarity between the PLSDA and PCA is a good indicator that the observed results are not the results of a fortunate combination of preprocessing and PLSDA or an artifact of using the PLS algorithm for classification instead of calibration, for which it was originally implemented. As such, these results suggest that the observed results are a direct consequence of the capability of the proposed screening bioassay coupled with high-

throughput FTIR spectra acquisition in detecting metabolic fingerprints, particularly those induced by the exposure to different antibiotics acting on the major biosynthetic pathways.

VII.3.4 DISCRIMINATING THE METABOLIC FINGERPRINT OF PROTEIN SYNTHESIS INHIBITORS

The next logical step in pinpointing the MOA of a candidate molecule is to discriminate between molecules that act with a similar MOA, e.g., on the same biosynthetic pathway but on a different point of the pathway. Within the clusters of metabolic fingerprints representing the major biosynthetic pathways targeted, there are sub-clusters coherent with the antibiotic classes tested (Figure 3 and Figure 4).

For instance, for antibiotics that act on protein biosynthesis, those belonging to the aminoglycoside class (kanamycin, neomycin and tobramycin) have a metabolic fingerprint that is more similar among them than those belonging to the amphenicol (chloramphenicol) and the macrolide (erythromycin) classes. Interestingly, antibiotics of the aminoglycoside class bind to the 30S ribosomal unit, more specifically at the A-site, where they mimic the stabilization induced by cognate tRNA, thereby allowing noncognate tRNA to bind to the A-site, resulting in mRNA misreading and faulty protein synthesis. Additionally, allosteric binding sites affect ribosomal subunit mobility, which reduces translational activity and impairs ribosomal recycling. However, the specific relationship of these effects and cell death are not fully understood [45].

On the other hand, chloramphenicol, which belongs to the amphenicol class of antibiotics, has been considered a ‘general’ translation elongation inhibitor. Chloramphenicol was assumed to be a competitive inhibitor of aminoacyl-tRNA binding in the peptidyl transferase center of the 50S subunit A site, but recent studies suggest an MOA closer to that of macrolides, namely a sequence-specific inhibition of translation elongation [46].

Similarly, macrolides where thought to indiscriminately block protein elongation via a ‘plug-in-the-bottle’ mechanism, where binding to the tunnel close to the peptidyl transferase center physically obstructs nascent chain progression, but recent studies indicate that several proteins can bypass this blockage, thereby suggesting a sequence-specific mechanism [47].

Regardless of the specific MOA of each class, the antibiotics tested that target protein biosynthesis act at the elongation step; therefore, the fact that these were reproducibly plotted separately for three independent cultures suggests FTIRS is not only capable of detecting metabolic fingerprints with sufficient sensitivity to elucidate MOA beyond targeting protein

biosynthesis, but can conceivably be used to elucidate different mechanisms that disrupt protein elongation.

VII.3.5 Discerning the metabolic fingerprints of DNA synthesis inhibitors

Regarding the antibiotics that inhibit DNA biosynthesis, the fluoroquinolones (levofloxacin and ciprofloxacin) were clustered together with the sulfonamides (sulfamethazine and sulfamethoxazole), but these were clearly distinct from metronidazole (Figure 3 and Figure 4). Fluoroquinolones block the progression of the enzyme–DNA complex formed during replication, which ultimately impairs DNA synthesis and induces rapid bacterial death. Specifically, fluoroquinolones MOA is based on the disruption of two enzymes: DNA gyrase, which introduces negative superhelical twists that facilitate the separation of daughter chromosomes and allows for the binding of initiation proteins; and topoisomerase IV, which is responsible for removing the interlinking of daughter chromosomes, therefore allowing their segregation into the daughter cells at the end of a replication round [48].

On the other hand, sulfonamides (sulfamethoxazole and sulfamethazine) are known as non-classical antifolates. This class of molecules are competitive inhibitors with p-aminobenzoic acid, preventing its entrance to the reaction site of dihydropteroate synthase and forming an analogue that cannot be used in the subsequent reactions, thereby greatly reducing folate levels. Because bacteria cannot absorb exogenous folate, thymine depletion occurs, and ultimately DNA biosynthesis errors, which result in the observed antibiotic effect [49]. Thymine depletion induces thymineless death, for which a consensual mechanism has not been proposed. One possible mechanism revolves around stalled replication forks [50], which, if confirmed, implies that the MOA of fluoroquinolones and sulfonamides could be more similar than traditionally acknowledged, which is in accordance with the obtained results. Alternatively, it could be that, despite having dissimilar MOAs, the metabolic fingerprint captured by FTIRS is not sufficiently specific to distinguish between said MOAs.

Lastly, although the MOA of metronidazole is still unclear, it is believed that metronidazole is intracellularly reduced to a short-lived nitroso free radical, which is not only cytotoxic, but also inhibits DNA synthesis and causes DNA damage by oxidation, which results in DNA degradation and eventually cell death [51]. This, in turn, is a considerably different MOA from both fluoroquinolones and sulfonamides, which is coherent with the results obtained. As a note, metronidazole is only intracellularly reduced in the presence of a sufficiently

negative redox potential, and it could therefore be that the experimental setup utilized induced sufficient anaerobic conditions to obtain an antibiotic effect reflected on the metabolic fingerprint, since the facultative anaerobe model organism utilized, i.e., *E. coli*, can be susceptible to metronidazole [52] and apparently was (Table 1), but could also justify the proximity between samples exposed to metronidazole and the control samples.

In sum, while the proposed FTIRS bioassay is apparently not the most adequate tool to reach conclusions regarding the MOA of sulfonamides, it is possible that FTIRS captures the metabolic fingerprint induced by antibiotics with sufficient sensitivity to distinguish those targeting DNA biosynthesis via different mechanisms.

VII.3.6 Differentiating the metabolic fingerprints of cell wall biosynthesis inhibitors

Unlike the antibiotics described so far, those targeting cell wall biosynthesis had to be analyzed differently. Specifically, the standard concentration of four times the MIC resulted in extensive cell lysis, which in turn implied a considerable loss of intensity of the FTIR spectra. As such, cells were exposed to ampicillin at the MIC and to amoxicillin at 25% of the MIC. Cells were exposed to cephradine at the standard concentration, i.e., four times the MIC. Although this might explain the higher dispersion obtained for these antibiotics, in comparison with those targeting either DNA or protein biosynthesis, it was necessary to obtain spectra with a sufficient signal-to-noise ratio for analysis.

Regardless of the distance between clusters of independent cultures exposed to the same antibiotic, there was coherence in the within-cluster distance, i.e., independent cultures were grouped together (Figure 3 and Figure 4), which is a positive indication. Briefly, amoxicillin, ampicillin, and cephradine are beta-lactam antibiotics and only differ in their affinities and/or molecular target. Beta-lactams inhibit transpeptidases and prevent cross-linking, thereby inducing structural deficiencies in the cell wall that results in cell lysis [53]. However, the mechanism of cell death induced by beta-lactams has been shown to extend beyond cell lysis. In fact, it seems that the cell wall synthesis machinery is recruited to a futile cycle of synthesis/degradation that depletes cellular resources and bolsters the bactericidal activity of beta-lactams [54].

Although the issue of the antibiotic concentration is a question that still lingers, and must be attended for the industrial application of FTIRS as a viable screening technology for

antibiotic discovery, it is interesting to note that cephradine, which belongs to the cephalosporins sub-class of antibiotics, is clustered further away from the aminopenicillins (amoxicillin and ampicillin). If further validated, this could be another positive indication of the metabolic sensitivity of fingerprints obtained with FTIRS.

VII.3.7 DIFFERENTIATING OTHER METABOLIC FINGERPRINTS

Interestingly, the samples exposed to isoniazid were clustered closely to those exposed to DNA synthesis inhibitors (Figure 3 and Figure 4). Isoniazid enters the cell as a pro-drug and exerts its antibiotic effect by disturbing various macromolecular syntheses, of which the most frequently discussed is mycolic acid synthesis; therefore, isoniazid is the preferred therapeutic for tuberculosis [55]. Although the MOA is still unclear, the peroxidative activation of isoniazid by the mycobacterial enzyme KatG forms potent inhibitors of lipid and nucleic acid biosynthesis, as well as inducing oxidative stress [56].

Regardless, its inhibitory effect on *E. coli* has long been reported as being dependent on the initial cell concentration, antibiotic concentration, and medium composition [57], and experimentally confirmed (Table 1). Apparently, the metabolic fingerprint detected with FTIRS more closely reflects the inhibition of DNA biosynthesis; however, further validation of this observation is required. On the other hand, rifampicin samples were clustered together with protein synthesis inhibitors. Since rifampicin binds with high affinity to the bacterial DNA-dependent RNA polymerase, this results in its inhibition, ultimately causing a lethal disruption of RNA biosynthesis at the elongation step [58]. This suggests that the clustering observed is coherent given this type of inhibition. In other words, inhibition at the level of transcription should provide a metabolic effect that is closer to protein synthesis inhibition, which is at the level of translation, in comparison with DNA biosynthesis inhibition that occurs at the level of replication. Moreover, this is particularly distinct from cell wall biosynthesis inhibition, which is at a distant end of the spectrum of cellular metabolic responsibilities.

VII. 4 CONCLUSIONS

Given the importance of MOA identification in phenotypic screening, and the role of the latter for the success of antibiotic discovery, methods capable of combining MOA elucidation with high-throughput screening of whole cells are required now more than ever. Here, we explored FTIRS as a metabolic fingerprinting technique regarding its sensitivity towards

elucidating MOA, ranging from its looser definition of drug effects, e.g., the major biosynthetic pathway affected, through to the stricter drug target that individual antibiotics inhibit. Because FTIRS requires extensive data analysis, a combination of parameters of commonly applied preprocessing algorithms were optimized. This ensured that the performance of predictive models was maximized.

In general, the metabolic fingerprints obtained with FTIRS were closely related to the MOA of all the antibiotics tested, which include examples that target protein, DNA, and cell wall biosynthesis. Additionally, the metabolic fingerprints induced by exposure to an RNA biosynthesis inhibitor was similar with those of protein synthesis inhibitors, which is coherent with the metabolic effect expected. Pending further validation, these fingerprints could help to elucidate the MOA of known drugs, for instance isoniazid, which was clustered close to DNA synthesis inhibitors, and also the thymineless death induced by sulfonamides.

Ultimately, these results demonstrate that there is great potential in using FTIRS as a tool to acquire a holistic picture of the effect of different antibiotics on the cellular metabolism, which can be used not only for antibiotic discovery but also towards a better understanding of the MOA of current antibiotics.

VII. 5 Acknowledgments

The present work was partially conducted in the Engineering & Health Laboratory, which results of a collaboration protocol established between Universidade Católica Portuguesa and Instituto Politécnico de Lisboa.

This research was funded by Instituto Politécnico de Lisboa, grant number IDI&CA/IPL/2017/DrugsPlatf/ISEL and IDI&CA/IPL/2018/RenalProg/ISEL.

VII. 6 References

1. Ribeiro da Cunha; Fonseca; Calado Antibiotic Discovery: Where Have We Come from, Where Do We Go? *Antibiotics* **2019**, *8*, 45.
2. Fields, F.R.; Lee, S.W.; McConnell, M.J. Using bacterial genomes and essential genes for the development of new antibiotics. *Biochem. Pharmacol.* **2017**, *134*, 74–86.
3. Lewis, K. Platforms for antibiotic discovery. *Nat. Rev. Drug Discov.* **2013**, *12*, 371–387.
4. Ohki, Y.; Sakurai, H.; Hoshino, M.; Terashima, H.; Shimizu, H.; Ishikawa, T.; Ogiyama, T.; Muramatsu, Y.; Nakanishi, T.; Miyazaki, S.; et al. Perturbation-Based Proteomic Correlation Profiling as a Target Deconvolution Methodology. *Cell Chem. Biol.* **2019**, *26*, 137–143.

5. Bantscheff, M.; Drewes, G. Chemoproteomic approaches to drug target identification and drug profiling. *Bioorganic Med. Chem.* **2012**, *20*, 1973–1978.

6. Kurita, K.L.; Glassey, E.; Linington, R.G. Integration of high-content screening and untargeted metabolomics for comprehensive functional annotation of natural product libraries. *Proc. Natl. Acad. Sci.* **2015**, *112*, 11999–12004.

7. Birkenstock, T.; Liebeke, M.; Winstel, V.; Krismer, B.; Gekeler, C.; Niemiec, M.J.; Bisswanger, H.; Lalk, M.; Peschel, A. Exometabolome analysis identifies pyruvate dehydrogenase as a target for the antibiotic triphenylbismuthdichloride in multiresistant bacterial pathogens. *J. Biol. Chem.* **2012**, *287*, 2887–2895.

8. French, S.; Ellis, M.J.; Coutts, B.E.; Brown, E.D. Chemical genomics reveals mechanistic hypotheses for uncharacterized bioactive molecules in bacteria. *Curr. Opin. Microbiol.* **2017**, *39*, 42–47.

9. Iorio, F.; Bosotti, R.; Scacheri, E.; Belcastro, V.; Mithbaokar, P.; Ferriero, R.; Murino, L.; Tagliaferri, R.; Brunetti-Pierri, N.; Isacchi, A.; et al. Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proc. Natl. Acad. Sci.* **2010**, *107*, 14621–14626.

10. Nonejuie, P.; Burkart, M.; Pogliano, K.; Pogliano, J. Bacterial cytological profiling rapidly identifies the cellular pathways targeted by antibacterial molecules. *Proc. Natl. Acad. Sci.* **2013**, *110*, 16169–16174.

11. Sato, S. ichi; Murata, A.; Shirakawa, T.; Uesugi, M. Biochemical Target Isolation for Novices: Affinity-Based Strategies. *Chem. Biol.* **2010**, *17*, 616–623.

12. Nishiya, Y.; Hamada, T.; Abe, M.; Takashima, M.; Tsutsumi, K.; Okawa, K. A new efficient method of generating photoaffinity beads for drug target identification. *Bioorganic Med. Chem. Lett.* **2017**, *27*, 834–840.

13. Burdine, L.; Thomas, K. Target Identification in Chemical Genetics: The (Often) Missing Link. *Chem. Biol.* **2004**, *11*, 593–597.

14. Hutter, B.; Schaab, C.; Albrecht, S.; Borgmann, M.; Brunner, N.A.; Freiberg, C.; Ziegelbauer, K.; Rock, C.O.; Ivanov, I.; Loferer, H. Prediction of mechanisms of action of antibacterial compounds by gene expression profiling. *Antimicrob. Agents Chemother.* **2004**, *48*, 2838–2844.

15. Halouska, S.; Chacon, O.; Fenton, R.J.; Zinniel, D.K.; Raul, G.; Powers, R. Use of NMR Metabolomics to Analyze the Targets of D-cycloserine in Mycobacteria : Role of D-Alanine Racemase. *J Proteome Res* **2007**, *6*, 4608–4614.

16. Kohanski, M.A.; Dwyer, D.J.; Collins, J.J. How antibiotics kill bacteria: From targets to networks. *Nat. Rev. Microbiol.* **2010**, *8*, 423–435.

17. Zampieri, M. From the metabolic profiling of drug response to drug mode of action. *Curr. Opin. Syst. Biol.* **2018**, *10*, 26–33.

18. Schelli, K.; Zhong, F.; Zhu, J. Comparative metabolomics revealing Staphylococcus aureus metabolic response to different antibiotics. *Microb. Biotechnol.* **2017**, *10*, 1764–1774.

19. Campos, A.I.; Zampieri, M. Metabolomics-Driven Exploration of the Chemical Drug Space to Predict Combination Antimicrobial Therapies. *Mol. Cell* **2019**, *74*, 1291-1303.e6.

20. Wu, C.; Choi, Y.H.; van Wezel, G.P. Metabolic profiling as a tool for prioritizing antimicrobial compounds. *J. Ind. Microbiol. Biotechnol.* **2016**, *43*, 299–312.

21. Hoerr, V.; Duggan, G.E.; Zbytnuik, L.; Poon, K.K.H.; Große, C.; Neugebauer, U.; Methling, K.; Löffler, B.; Vogel, H.J. Characterization and prediction of the mechanism of action of antibiotics through NMR metabolomics. *BMC Microbiol.* **2016**, *16*, 1–14.

22. Chaleckis, R.; Meister, I.; Zhang, P.; Wheelock, C.E. Challenges, progress and promises of metabolite annotation for LC–MS-based metabolomics. *Curr. Opin. Biotechnol.* **2019**, *55*, 44–50.

23. Vincent, I.M.; Ehmann, D.E.; Mills, S.D.; Perros, M.; Barrett, M.P. Untargeted Metabolomics To Ascertain Antibiotic Modes of Action. *Antimicrob. Agents Chemother.* **2016**, *60*, 2281–2291.

24. Zampieri, M.; Sekar, K.; Zamboni, N.; Sauer, U. Frontiers of high-throughput metabolomics. *Curr. Opin. Chem. Biol.* **2017**, *36*, 15–23.

25. Zampieri, M.; Szappanos, B.; Buchieri, M.V.; Trauner, A.; Piazza, I.; Picotti, P.; Gagneux, S.; Borrell, S.; Gicquel, B.; Lelievre, J.; et al. High-throughput metabolomic analysis predicts mode of action of uncharacterized antimicrobial compounds. *Sci. Transl. Med.* **2018**, *10*, 1–12.

26. Marques, V.; Cunha, B.; Couto, A.; Sampaio, P.; Fonseca, L.P.; Aleixo, S.; Calado, C.R.C. Characterization of gastric cells infection by diverse Helicobacter pylori strains through Fourier-transform infrared spectroscopy. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2019**, *210*, 193–202.

27. Goodacre, R.; Vaidyanathan, S.; Dunn, W.B.; Harrigan, G.G.; Kell, D.B. Metabolomics by numbers: Acquiring and understanding global metabolite data. *Trends Biotechnol.* **2004**, *22*, 245–252.

28. Ribeiro da Cunha, B.; Fonseca, L.P.; Calado, C.R.C. A phenotypic screening bioassay for Escherichia coli stress and antibiotic responses based on Fourier-transform infrared (FTIR) spectroscopy and multivariate analysis. *J. Appl. Microbiol.* **2019**, *127*, 1776–1789.

29. Sharaha, U.; Rodriguez-Diaz, E.; Riesenberg, K.; Bigio, I.J.; Huleihel, M.; Salman, A. Using Infrared Spectroscopy and Multivariate Analysis to Detect Antibiotics' Resistant Escherichia coli Bacteria. *Anal. Chem.* **2017**, *89*, 8782–8790.

30. Huleihel, M.; Pavlov, V.; Erukhimovitch, V. The use of FTIR microscopy for the evaluation of anti-bacterial agents activity. *J. Photochem. Photobiol. B Biol.* **2009**, *96*, 17–23.

31. Xuan Nguyen, N.T.; Sarter, S.; Hai Nguyen, N.; Daniel, P. Detection of molecular changes induced by antibiotics in Escherichia coli using vibrational spectroscopy. *Spectrochim. Acta - Part A Mol. Biomol. Spectrosc.* **2017**, *183*, 395–401.

32. Wharfe, E.S.; Winder, C.L.; Jarvis, R.M.; Goodacre, R. Monitoring the effects of chiral pharmaceuticals on aquatic microorganisms by metabolic fingerprinting. *Appl. Environ. Microbiol.* **2010**, *76*, 2075–2085.

33. Institute, C. and L.S. *Methods for Dilution Antimicrobial Susceptibility Tests for Bacteria That Grow Aerobically; Approved Standard — Ninth Edition*; Pennsylvania, USA, 2012; Vol. 32; ISBN 1-56238-784-7.

34. EUCAST The European Committee on Antimicrobial Susceptibility Testing. Routine and extended internal quality control for MIC determination and disk diffusion as recommended by EUCAST. *Version 10.0* **2020**.

35. Tian, X.; Yu, Q.; Wu, W.; Li, X.; Dai, R. Comparative proteomic analysis of Escherichia coli O157:H7 following ohmic and water bath heating by capillary-HPLC-MS/MS. *Int. J. Food Microbiol.* **2018**, *285*, 42–49.

36. Li, J.; Xie, S.; Ahmed, S.; Wang, F.; Gu, Y.; Zhang, C.; Chai, X.; Wu, Y.; Cai, J.; Cheng, G. Antimicrobial activity and resistance: Influencing factors. *Front. Pharmacol.* **2017**, *8*, 1–11.

37. Bidlas, E.; Du, T.; Lambert, R.J.W. An explanation for the effect of inoculum size on MIC and the growth/no growth interface. *Int. J. Food Microbiol.* **2008**, *126*, 140–152.

38. Pankey, G.A.; Sabath, L.D. Clinical Relevance of Bacteriostatic versus Bactericidal Activity in the Treatment of Gram-Positive Bacterial Infections. *Clin. Infect. Dis.* **2004**, *38*, 864–70.

39. Belenky, P.; Ye, J.D.; Porter, C.B.M.; Cohen, N.R.; Lobritz, M.A.; Ferrante, T.; Jain, S.; Korry, B.J.; Schwarz, E.G.; Walker, G.C.; et al. Bactericidal Antibiotics Induce Toxic Metabolic Perturbations that Lead to Cellular Damage. *Cell Rep.* **2015**, *13*, 968–980.

40. Mi, H.; Wang, D.; Xue, Y.; Zhang, Z.; Niu, J.; Hong, Y.; Drlica, K.; Zhao, X. Dimethyl Sulfoxide Protects Escherichia coli from Rapid Antimicrobial-Mediated Killing. *Antimicrob. Agents Chemother.* **2016**, *60*, 5054–5058.

41. Lasch, P. Spectral pre-processing for biomedical vibrational spectroscopy and microspectroscopic imaging. *Chemom. Intell. Lab. Syst.* **2012**, *117*, 100–114.

42. Windig, W.; Shaver, J.; Bro, R. Loopy MSC: A simple way to improve multiplicative scatter correction. *Appl. Spectrosc.* **2008**, *62*, 1153–1159.

43. Zimmermann, B.; Kohler, A. Optimizing savitzky-golay parameters for improving spectral resolution and quantification in infrared spectroscopy. *Appl. Spectrosc.* **2013**, *67*, 892–902.

44. Brereton, R.G.; Lloyd, G.R. Partial least squares discriminant analysis: Taking the magic away. *J. Chemom.* **2014**, *28*, 213–225.

45. Becker, B.; Cooper, M.A. Aminoglycoside antibiotics in the 21st century. *ACS Chem. Biol.* **2013**, *8*, 105–115.

46. Volkov, I.L.; Seefeldt, A.C.; Johansson, M. Tracking of single tRNAs for translation kinetics measurements in chloramphenicol treated bacteria. *Methods* **2019**, *162–163*, 23–30.

47. Davis, A.R.; Gohara, D.W.; Yap, M.N.F. Sequence selectivity of macrolide-induced translational attenuation. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111*, 15379–15384.

48. Blondeau, J.M. Fluoroquinolones: Mechanism of action, classification, and development of resistance. *Surv. Ophthalmol.* **2004**, *49*, 1–6.

49. Fernández-Villa, D.; Aguilar, M.R.; Rojo, L. Folic Acid Antagonists: Antimicrobial and Immunomodulating Mechanisms and Applications. *Int. J. Mol. Sci.* **2019**, *20*, 1–30.

50. Kuong, K.J.; Kuzminov, A. Stalled replication fork repair and misrepair during thymineless death in Escherichia coli. *Genes Cells* **2010**, *15*, 619–634.

51. Löfmark, S.; Edlund, C.; Nord, C.E. Metronidazole Is Still the Drug of Choice for Treatment of Anaerobic Infections. *Clin. Infect. Dis.* **2010**, *50*, S16–S23.

52. Jackson, D.; Salem, A.; Coombs, G.H. The in-vitro activity of metronidazole against strains of Escherichia coli with impaired DNA repair systems. *J. Antimicrob. Chemother.* **1984**, *13*, 227–236.

53. Bardal, S.K.; Waechter, J.E.; Martin, D.S. Chapter 18 - Infectious Diseases. In *Applied Pharmacology*; Saunders: Philadelphia, PA, USA, 2011; pp. 233–291 ISBN 978-1-4377-0310-8.

54. Cho, H.; Uehara, T.; Bernhardt, T.G. Beta-lactam antibiotics induce a lethal malfunctioning of the bacterial cell wall synthesis machinery. *Cell* **2014**, *159*, 1300–1311.

55. Unissa, A.N.; Subbian, S.; Hanna, L.E.; Selvakumar, N. Overview on mechanisms of isoniazid action and resistance in Mycobacterium tuberculosis. *Infect. Genet. Evol.* **2016**, *45*, 474–492.

56. Timmins, G.S.; Deretic, V. Mechanisms of action of isoniazid. *Mol. Microbiol.* **2006**, *62*, 1220–1227.

57. Tritz, G.J. Protection of Escherichia coli from isoniazid inhibition. *Antimicrob. Agents Chemother.* **1974**, *5*, 217–222.

58. Campbell, E.A.; Korzheva, N.; Mustaev, A.; Murakami, K.; Nair, S.; Goldfarb, A.; Darst, S.A. Structural mechanism for rifampicin inhibition of bacterial RNA polymerase. *Cell* **2001**, *104*, 901–912.

Chapter VIII

Simultaneous prediction of antibiotic mechanism of action and potency

This chapter is adapted from the manuscript:

Ribeiro da Cunha, B., Fonseca, L.P. & Calado, C.R.C. Simultaneous elucidation of antibiotic mechanism of action and potency with high-throughput Fourier-transform infrared (FTIR) spectroscopy and machine learning. Appl Microbiol Biotechnol (2021). https://doi.org/10.1007/s00253-021-11102-7.

Author contribution

Bernardo Ribeiro da Cunha conceptualized the study methodology, conducted the investigation and formal analysis, prepared the original draft, reviewed and edited its final version.

Abstract

The low rate of discovery and rapid spread of resistant pathogens have made antibiotic discovery a worldwide priority. In cell-based screening, the mechanism of action (MOA) is identified after antimicrobial activity. This increases rediscovery, impairs low potency candidate detection, and does not guide lead optimization. In this study, high-throughput Fourier-transform infrared (FTIR) spectroscopy (FTIRS) was used to discriminate the MOA of 14 antibiotics at pathway, class, and individual antibiotic level. For that, the optimal combinations and parametrizations of spectral preprocessing were selected with cross-validated partial least squares discriminant analysis, to which various machine learning algorithms were applied. This coherently resulted in very good accuracies, independently of the algorithms, and at all levels of MOA. Particularly, an ensemble of subspace discriminants predicted the known pathway (98.6%), antibiotic classes (100%), and individual antibiotics (97.8%) with exceptional accuracy, and similar results were obtained for simulated novel MOA. Even at very low concentrations (1 μg/mL) and growth inhibition (15%), over 70% pathway and class accuracy were achieved, suggesting FTIRS can probe the grey chemical matter. Prediction of inhibitory effect was also examined, for which a squared exponential Gaussian process regression yielded a root mean square error of 0.33 and a R^2 of 0.92, indicating that metabolic alterations leading to growth inhibition are intrinsically reflected on FTIR spectra beyond cell density.

Contents

VIII. 1 INTRODUCTION

The rise in deaths and disabilities caused by antimicrobial-resistant pathogens has made infectious diseases in general, and multidrug resistant pathogens in particular, a worldwide priority [1–3]. Many factors contributed to the upsurge and proliferation of antimicrobial resistance [4,5]. Importantly, the rate at which novel antimicrobials enter the market has been stagnated at alarmingly low rates since the antibiotic golden age [6], when most classes in use today were discovered [7]. The success of the antibiotic golden age was fueled by the Waksman platform, but it eventually yielded redundant discoveries [8]. Semi-synthesis filled the pipeline with iterative generations of antibiotics [9], but the number of iterations is limited [10]. While fully-synthetic routes have extended the modifiable chemical space, very few novel structures have been chemically synthetized [11], and mechanism of action (MOA) diversity is essential to outpace the rise of resistance mechanisms and to avoid cross-resistance [12].

During the genomics era, the industry turned to a target-based screening approach in search of novel MOA. However, fewer-than-expected molecules were identified, and those were incapable of entering the bacterial cell and reaching their target [13,14]. Additionally, resistance mechanisms have a higher probability of arising in single-protein targets [15]. Therefore, target-based screening was abandoned in favor of cell-based screening. In the latter, a characteristic is monitored to identify compounds that alleviate the disease phenotype, e.g., killing bacteria or halting their proliferation [16], thus the term phenotypic screening. Because compounds are screened against whole cells, there's a higher likelihood of identifying suitable candidates, as well as those targeting poorly understood pathways and multiple targets. In the cell-based approach, positive hits from phenotypic screening assays advance to MOA identification, after which known or similar compounds are excluded, thus increasing rediscovery rates [17]. Above all, late MOA identification postpones the exclusion of undesirable toxic compounds with unspecific MOA [18], cannot guide lead optimization where medicinal chemistry programs are paramount to improve the chances of commercialization, and limits the detection of low potency candidates that are underexplored but suitable for chemical optimization [19]. Finding new antibiotics is very challenging and identifying MOA has proven equally defiant.

Macromolecular synthesis assays have been used for MOA identification for nearly half a century [20]. Higher throughput formats have been proposed [18], but these are still laborious,

have low resolution and low throughput. Biochemical approaches increase resolution but require a high-affinity small molecule and an abundant protein receptor, and are also laborious, slow and low-throughput [21,22]. Alternatives are bacterial accumulation assays [23], visualizing morphological alterations and leakage of intracellular material [24], or kinetic dose response [25]. Interestingly, many microbial processes were understood with chemical genetics methods applied to MOA identification [26], but these are limited by large mutant libraries [27,28] and few experimentally characterized genes [29]. Promoter-reporter assays are better suited to probe cellular pathways and off-target effects [30] and offer better temporal resolution [26], but are laborious to setup, have low-throughput and have limited MOA resolution [31]. More recently, transcriptomics [32], proteomics [33] and metabolomics [34] have been explored. However, these are still costly, laborious, generally low-throughput, and methodically complex. This impairs screening hundreds of thousands of compounds, which is a reasonable throughput of a drug discovery program [31]. An alternative is high-content screening, an image-based approach where morphological changes reveal the MOA. However, not all MOA are reflected on cell morphology and not all morphological changes trace back to a molecular phenotype [35], so another approach to MOA identification is vital.

Fourier-transform infrared (FTIR) spectroscopy (FTIRS) is an established high-throughput technique that requires minimal sample preparation and is reagent-less, easy to operate and label free [36]. In addition to these operational advantages, at its core, FTIRS measures the vibrational energy of molecular bonds after their interaction with infrared radiation at specific wavelengths. Thus, FTIRS directly reveals the molecular composition of a given sample. More importantly, a single FTIR spectrum can provide molecular insight into most biologically relevant molecules, e.g., nucleic acids, proteins, carbohydrates, lipids, phospholipids, etc. [37]; therefore, most if not all molecular phenotypes are detected with this technique. Although other techniques provide more detailed data at a given step of the “omics” cascade, FTIRS provides comprehensive chemical information to rapidly and reproducibly classify samples based on their biological origin, i.e., metabolic fingerprinting [38]. Therefore, FTIRS can fill the void between high-throughput/low-information phenotypic screening assays (e.g., cell viability) and low-throughput/high-information MOA characterization assays. It has been shown that FTIRS detects antibiotic-specific fingerprints [39–42], but an extensive assessment of FTIRS as a high-throughput tool to accurately predict MOA is lacking. More importantly, an in-depth evaluation of accuracy, sensitivity, specificity, and precision is required, as well as an evaluation of resolution towards known and novel MOA. Also, the effect of antibiotic concentration on MOA resolution has not been addressed and is particularly relevant

to explore the grey chemical matter, i.e., compounds that modulate metabolism without clear growth inhibition and are promising candidates for medicinal chemistry programs. The possibility of estimating growth inhibition from FTIR spectra has also not been explored and could allow for the prediction of antimicrobial activity and MOA in a single assay.

As such, *Escherichia coli* was exposed to twofold dilutions of 14 antibiotics, belonging to 7 classes and acting on 4 major biosynthetic pathways, using a high-throughput microplate-based assay suitable for automation. Firstly, two spectral preprocessing techniques were evaluated, scatter correction and derivative filtering. In particular, multiplicative scatter correction (MSC), its improved version called extended-MSC (EMSC), or EMSC with replicate correction (EMSCrep), were applied alone or in combination with the Savitzky-Golay (SG) derivative filtering. Although both SG and either type of MSC aim to reduce the non-discriminatory signal component and highlight the sample-specific spectral component, because they do so differently, they are typically applied in combination, which results in some level of synergetic improvements [43]. Moreover, the parametrization of these algorithms was optimized using leave-one-out cross-validation (LOO-CV) of partial least squares discriminant analysis (PLSDA). This ensured an unbiased and quantitative evaluation of the effect of preprocessing and revealed “optimal” preprocessing conditions. PLSDA models created with optimally preprocessed spectra were validated, e.g., with principal component analysis (PCA), to ensure spectra contained intrinsic patterns that reflected the biological phenomena being studied. After removal of uninformative spectral regions, various machine learning algorithms were applied to the “optimally” preprocessed and “cleaned” spectra to predict either the major biosynthetic pathway, antibiotic class, or individual antibiotic each sample was exposed to. Considering different algorithms not only enabled further improvements to MOA prediction, but also validated FTIRS for MOA identification beyond the limitations of any given algorithm. The top-performing machine learning algorithm was used to evaluate MOA resolution across antibiotic concentrations, especially at sub-inhibitory concentrations. Lastly, growth inhibition was regressed from FTIR spectra in order to evaluate the simultaneous determination of MOA and potency.

VIII. 2 MATERIALS AND METHODS

VIII.2.1 CHEMICALS AND MATERIAL

Amoxicillin, ampicillin, cephradine, chloramphenicol, clarithromycin, erythromycin, gentamicin, levofloxacin, metronidazole, neomycin, rifampicin, kanamycin and tobramycin

were purchased from Sigma-Aldrich, and ciprofloxacin from Bayer. The major biosynthetic pathway targeted, class, MOA, the minimum inhibitory concentration that inhibited at least 50% of bacterial growth (MIC50), maximum growth inhibition, and recommended solvent of each antibiotic were presented (Table 1). Stock solutions were prepared at 2000 μg/mL, filtered with a 0.2-μm cellulose filter, and stored at -20ºC, unless specified otherwise by the supplier.

VIII.2.2 BACTERIAL GROWTH CONDITIONS AND ANTIBIOTIC EXPOSURE

Escherichia coli (ATCC 33876) cultures were grown in Bacto Tryptic Soy Broth (BD) at 37ºC, 200rpm, until mid-exponential phase. Then, 500 μL was transferred to each well of a 2-mL 96-deep-well microtiter plate (Nunc), to which 800 μL of fresh growth media were previously added, along with 500 μL of antibiotic solution to obtain twofold dilutions ranging from 512 μg/mL through to 0.03 μg/mL. A growth control exposed only to the solvent, and a sterility control without any bacteria, were also analyzed. Antibiotic exposure lasted 16h, at 37ºC, 200 rpm, on at least three independent cultures on different days. Immediately after exposure, 100 μL of the contents of each well was transferred to a flat-bottomed 96-well microtiter plate (Nunc) to read the absorbance at 600nm (ABS_{600}) (SPECTRAmax, Molecular Devices). The percentage of growth inhibition was determined as the difference in ABS_{600} between control and exposed samples [44] and were used to estimate the MIC50.

VIII.2.3 SPECTRAL DATA ACQUISITION

The contents of each well were centrifuged for 3 minutes at 13,000 rpm (13,793g) (Z160M, Hermle Labortechnik, Germany). The supernatant was discarded, and the pellet was resuspended in sterile water to obtain an ABS_{600} of 36 and plated on a ZnSe 96-well plate (Bruker, Germany) in at least duplicates, which were averaged to reduce operator-induced variability. Samples were dehydrated for 1h in a vacuum desiccator with silica, and spectra were acquired in a HTS-XT module coupled to a Vertex-70 spectrometer (Bruker Optics, Germany) in transmission mode, and consisted of 40 coadded scans at a 4 cm^{-1} resolution. Raw spectra were exported with the OPUS software (Bruker, Germany), as data point table files and imported into MATLAB (MathWorks, USA).

Table 1. Mechanism-of-action of 14 antibiotics belonging to 7 classes and acting on 4 major biosynthetic pathways. The minimum inhibitory concentration (MIC50), maximum growth inhibition (MGI), and recommended solvent are also reported.

Biosynthetic pathway	Antibiotic class	Mechanism of action	Antibiotic	MIC50 (μg/mL)	MGI (%)	Solvent
Cell Wall	Beta-lactam	Bind to penicillin binding proteins and prevent cross-linking during peptidoglycan biogenesis. This induces structural deficiencies in the cell wall, crippling its integrity and leads to lysis. However, it has been suggested that the cell wall synthesis machinery is recruited to a futile cycle of synthesis and degradation, which depletes cellular resources, and explains or bolsters bactericidal activity [45].	Amoxicillin	8	82,8	H_20
			Ampicillin	16	77,3	H_20
			Cephradine	> 512	48,7	H_20
DNA	Fluoroquinolone	Block replication fork progression by binding to either the enzyme-DNA complex formed by DNA gyrase or topoisomerase IV, inducing rapid bacterial death [46].	Ciprofloxacin	> 512	42,6	Hydrochloric acid (0.01M)
			Levofloxacin	16	48,6	H_20
	Nitroimidazole	Metronidazole is intracellularly reduced to a short-lived nitroso free radical that is cytotoxic and causes oxidative damage to DNA, resulting in breaks that lead to DNA degradation and eventually cell death [47].	Metronidazole	> 512	66,9	Acetic acid (0.01M)
RNA	Rifamycin	Binds to bacterial DNA-dependent RNA polymerase with high affinity, deep within the DNA/RNA channel but away from the active site, sterically blocking nascent transcripts at the 5' end, leading to a lethal disruption of RNA biosynthesis [48].	Rifampicin	> 512	48,6	H_20

Table 1 (Continued). Mechanism-of-action of 14 antibiotics belonging to 7 classes and acting on 4 major biosynthetic pathways. The minimum inhibitory concentration (MIC50), maximum growth inhibition (MGI), and recommended solvent are also reported.

Biosynthetic pathway	Antibiotic Class	Mechanism-Of-Action	Antibiotic	MIC50 (μg/mL)	MGI (%)	Solvent
Protein	Aminoglycoside	Aminoglycosides bind to the A-site of the 30S ribosomal subunit and stabilize the internal loop, as would a cognate tRNA, which allows the binding of noncognate tRNA and therefore results in mRNA misreading and faulty protein synthesis. Allosteric binding sites reduce ribosomal subunits' mobility, impairing translation factor binding and activity, and limiting ribosome recycling. Different mechanisms have been proposed for reduced translation fidelity and cell death [49].	Gentamycin	128	55,6	H_20
			Kanamycin	16	55,0	H_20
			Neomycin	128	68,0	H_20
			Tobramycin	8	39,7	H_20
	Amphenicol	Competitive inhibitor of aminoacyl-tRNA in the peptidyl transferase center of the 50S ribosomal subunit A site. Recent studies suggest the mechanism is dependent on nascent peptide sequence and the A-site acceptor amino acid [50].	Chloramphenicol	> 512	50,6	Ethanol (7% v/v)
	Macrolide	Physically obstruct the ribosome tunnel close to the 50S subunit peptidyl transferase center, blocking nascent chain progression. Some proteins bypass this blockage, thus a sequence-specific mechanism is more likely: either some proteins can thread through the tunnel, despite macrolide binding; or some proteins cause the dissociation of the antibiotic [51].	Clarithromycin	256	45,2	Acetone (7% v/v)
			Erythromycin	> 512	58,3	Ethanol (7% v/v)

VIII.2.4 DATA ANALYSIS

Firstly, spectra were submitted to quality control routines [52] and a few spectra were discarded accordingly. On the remaining samples, MSC, EMSC and EMSCrep were applied using the Afseth and Kohler toolbox [53]. Machine learning was employed with MATLAB's Classification Learner App. Briefly for prediction, one-vs-all Support Vector Machines (SVM) was applied on standardized data. For fine, medium and coarse Gaussian, the kernel scale was 11, 43, and 170 respectively, and set to automatic for other kernel functions. All SVM models considered a box constrain of 1. K-Nearest Neighbor (KNN) considered standardized data using Euclidean distance with equal weight for 1, 10 and 100 neighbors for the fine, medium and KNN respectively. Both Cosine and Cubic KNN considered 10 neighbors, with Cosine and Minkowski distances, respectively. Weighted KNN applied squared inverse weights to a Euclidean distance considering 10 neighbors. Ensemble of decision trees used either Bag or RSUBoost with 30 learners at 0.1 rate and 20 splits. Subspace ensembles were determined for discriminant analysis and KNN used 934 subspaces and 30 learners. Accuracy was determined as the ratio between the predicted true positives plus the true negatives over the total number of observations. Precision was calculated as the true positive rate over the true positive rate plus the false positive rate. Recall, or sensitivity, was determined as the true positive rate over the sum of the true positive rate plus the false negative rate. Specificity was calculated as the true negative rate over the sum of true negative plus false positives. Accuracy, precision, recall and specificity were determined for one-vs-all prediction.

For regression, MATLAB's Regression Learner App was used. Fine, medium and coarse trees were applied without surrogate decision splits and a minimum leaf size of 4, 12 and 36, respectively. Linear, quadratic, and cubic kernel SVM were applied on standardized data, considering an automatic box constrain, automatic kernel scale and automatic Epsilon. Fine, medium or coarse Gaussian used a kernel scale of 11, 43 and 170, respectively. An ensemble of boosted and bagged trees used a minimum leaf size of 8 and 30 learners, respectively, with a learning rate of 0.1. Gaussian Process Regression (GPR) applied exponential, squared exponential, matern 5/2 or rational quadratic isotropic kernel on standardized data, with a constant basis function, and automatic kernel scale.

VIII. 3 RESULTS

VIII.3.1 ANTIBIOTIC EXPOSURE AND FTIR SPECTRA ACQUISITION

Antibiotic exposure was designed to allow FTIR spectra acquisition from a microtiter susceptibility methodology. Thus, a high-throughput assay based on microtiter plates, along with a serial dilutions format, was conducted with a considerably higher cell density to ensure an adequate FTIR signal-to-noise ratio. Given the inhibitory effect of antibiotics is dependent on the inoculum size, the simultaneous acquisition of FTIR spectra and growth inhibition measurement was preferred so the metabolic effects that resulted in antibiotic-induced growth inhibition, across the range of concentrations tested, were directly reflected on the spectra. Also, this compromise allowed probing the dose-response of MOA identification, and from that to estimate the potency. The maximum observed growth inhibition, and the minimum concentrations to inhibit at least 50% of bacterial growth of unexposed samples (MIC50), were determined (Table 1). Although these are not comparable with the literature, given the mentioned compromise, it is important to note that the maximum concentration of some antibiotics inhibited just over 50% of bacterial growth. As such, using spectra from these samples avoided discrepancies between the expected growth inhibition, based on MIC values obtained with industry-accepted protocols, and the observed inhibitory effect of the tested antibiotics at the reported cell density. Moreover, this average inhibition was deemed adequate for the purpose of MOA identification, as drastic growth inhibition could have resulted in catastrophic cellular responses unspecific to the individual antibiotic-induced stress.

VIII.3.2 SPECTRAL PREPROCESSING OPTIMIZATION

Spectral preprocessing is the first step of the data analysis workflow. It has been shown that preprocessed spectra outperform raw spectra in both classification and quantitative analysis [54]. Typically, the first preprocessing method to be applied is spectral differentiation. Here, the SG filter is most often used given its band-pass properties, which reduces both high-frequency signal components, e.g., homoscedastic noise, and broader signal components, e.g., curving baselines. Another commonly used spectral preprocessing routine is MSC, which separates physical and chemical information by modeling the physical “interferants”, thereby counteracting additive effects resulting from signal acquisition and unfavorable sample properties. Its improved version, EMSC, also accounts for multiplicative effects. Further improvements to EMSC include replicate correction, EMSCrep, which minimizes between-

replicate variation. An example of the effect of preprocessing is presented on Figure 1. However, the performance of these preprocessing methods is affected by their parametrization, which varies for the system under investigation, and requires objective optimization [43].

To that end, the successful classification with PLSDA after LOO-CV was used to evaluate the effect of various combinations of SG parameters alone and coupled with MSC, EMSC, or EMSCrep, on MOA identification. For each combination, a model was constructed from the data subset containing samples exposed to the maximum antibiotic concentration, i.e., 512 μg/mL, and all control samples, i.e., 0 μg/mL. This minimized the number of models required for optimization and reduced computation time. Additionally, a higher concentration ensured antibiotic-induced metabolic fingerprints were imprinted on the spectra, given this corresponded to an average growth inhibition around 50%. This data subset was then matched with an external variable categorizing samples according to either the major biosynthetic pathway (Supplementary Figure 1), the antibiotic class (Supplementary Figure 2), or each antibiotic (Supplementary Figure 3) that *E. coli* cells were exposed to, along with a control group. Independently of the external variable used to build the PLSDA models, i.e., pathway, class or antibiotic, a second derivative SG coupled with EMSCrep always outperformed the other preprocessing algorithms. So, these models were closely examined across a range of SG parameters for the three external variables (Figure 2). Note that window size 0 implies that the SG filter was not used. In short, preprocessing optimization resulted in a 10% improvement at the pathway level, 14% improvement of class prediction, and 21% better prediction of the specific metabolic fingerprint of antibiotics.

It's noteworthy that EMSC alone actually yielded a PLSDA model with poorer performance than MSC and raw spectra, albeit it being more typically used. In regard to the SG parameter optimization, there was considerable performance variability among polynomial orders, and also window size, even within those typically employed (9-25). These results highlight the importance of preprocessing optimization, which varies among different datasets, but more importantly varies according to the variable being predicted from the same dataset, as was the case. Also, preprocessing had a more vigorous effect when the sample-specific component of the spectra was likely more nuanced. In other words, when less prominent spectral differences are expected, e.g., between individual antibiotics, preprocessing had a greater effect than when more prominent differences among spectra exist, e.g., between groups of antibiotics that affect a biosynthetic pathway. Since the latter were more prominent

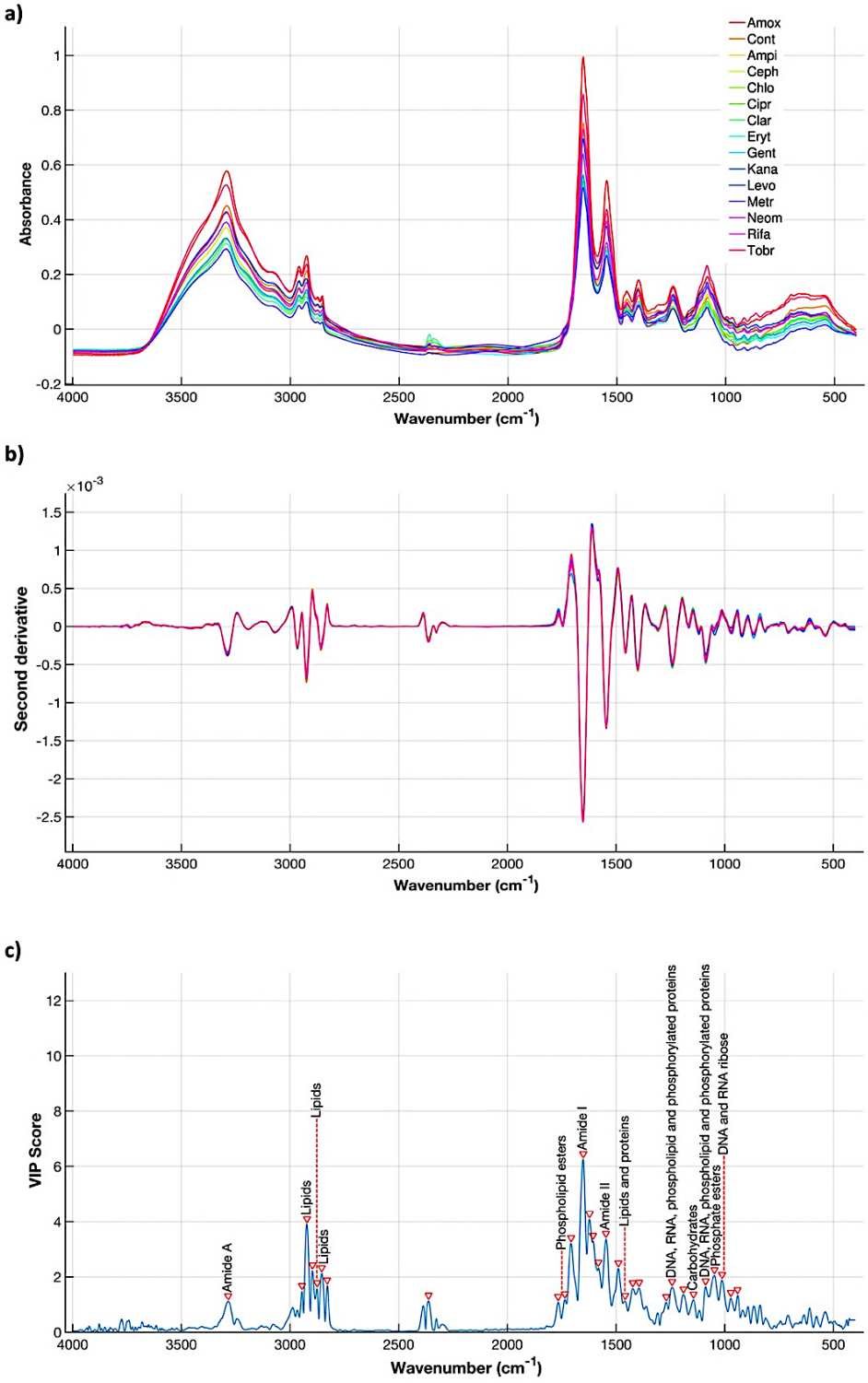


Figure 1. Average of raw spectra of independent *E. coli* cultures exposed to 14 antibiotics (a), and the effect of Savitzky-Golay filtering followed by extended multiplicative scatter correction with replicate correction (b). The variable importance in projection of partial least squares discriminant analysis (c) revealed the spectral regions with higher discrimination potential towards the fingerprints induced by specific antibiotics, which were identified (inverted triangles) and, where available, annotated with their biomolecular origin [55–58].

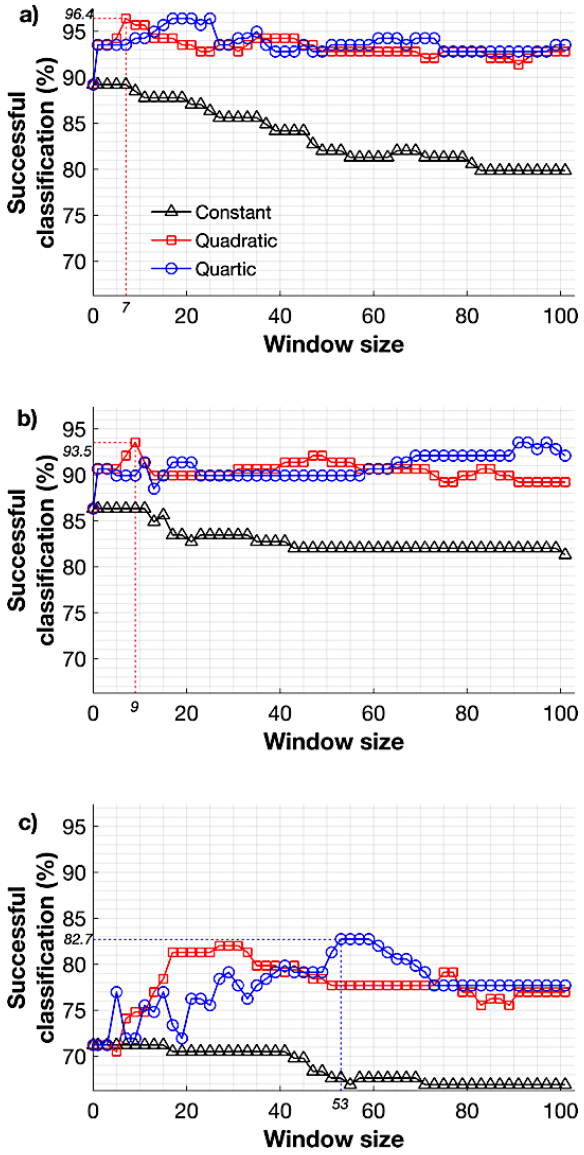


Figure 2. Successful classification of the major biosynthetic pathway (a), antibiotic class (b), and antibiotic-specific metabolic fingerprints (c) induced on *E. coli* cells after exposure to 14 antibiotics. Successful classification was evaluated with leave-one-out cross-validation of partial least squares discriminant analysis of FTIR spectra preprocessed with Savitzky-Golay second derivative filtering followed by extended multiplicative scatter correction with replicate correction.

even in raw spectra, preprocessing optimization had less of an effect. In all cases, the performance gains justify the extent of optimization put in place, as optimal performance wouldn't be fortuitously achieved with "standard" parameters. Even within "standard" parameters there was substantial performance variability, especially for prediction of individual antibiotics.

VIII.3.3 MODEL VALIDATION

Validation is essential to determine the merit of predictive models. For that, a LOO-CV was used since it has the smallest asymptotic bias and variance of the *k*-fold CV strategies [59], and is therefore suitable for an unbiased comparison of the performance of similar models. Another important issue in model validation is overfitting, especially when a large number of Latent Variables (LVs) are considered. To avoid this phenomenon, the previously described models used a generally accepted 10 LVs. Although the models were not used directly for prediction, but rather for preprocessing optimization, it was important to ensure that the data was not being grossly overfitted. To confirm this, the successful classification of either the biosynthetic pathway, antibiotic class or individual antibiotics over a larger range of LVs was determined for the optimal preprocessing combinations (Supplementary Figure 4).

It was also important to validate MOA identification beyond PLSDA by using a different algorithm prior to the application of machine learning strategies. So, PCA was performed on the spectra after preprocessing with the optimal parameters for the prediction of the effect of individual antibiotics (Figure 3). Unlike PLSDA, the objective of an unsupervised PCA is to reveal intrinsic data patterns, therefore it avoids the common pitfalls of PLSDA [60]. However, PCA results are not quantitative, e.g., the successful classification after LOO-CV, but rather qualitative, e.g., visually interpreted from score plots. PCA revealed an acceptable grouping of samples exposed to antibiotics, and these were mostly consistent with the major biosynthetic pathway targeted, which is coherent with PLSDA. Additionally, complementary PCA were conducted considering the preprocessing combinations that yielded optimal classification of antibiotic classes and the major biosynthetic pathway targeted by the antibiotics (Supplementary Figure 5 and 6, respectively). Intriguingly, these were very similar, indicating that PCA was not as sensitive to preprocessing as PLSDA. In general, these results suggest that PLSDA was suitable for preprocessing optimization, and more importantly FTIRS detected specific metabolic alterations after exposure to different antibiotics, independently of the algorithm used.

PCA was also used to address the premise of grouping all controls. Although most antibiotics were dissolved in water, a few required a different solvent (Table 1). To ensure that PCA score plots are not reflecting the metabolic effect of the different antibiotic solvents used, particular attention was given to erythromycin and chloramphenicol, dissolved in ethanol, as well as to ciprofloxacin and metronidazole, both dissolved in an acidic solvent. There is no

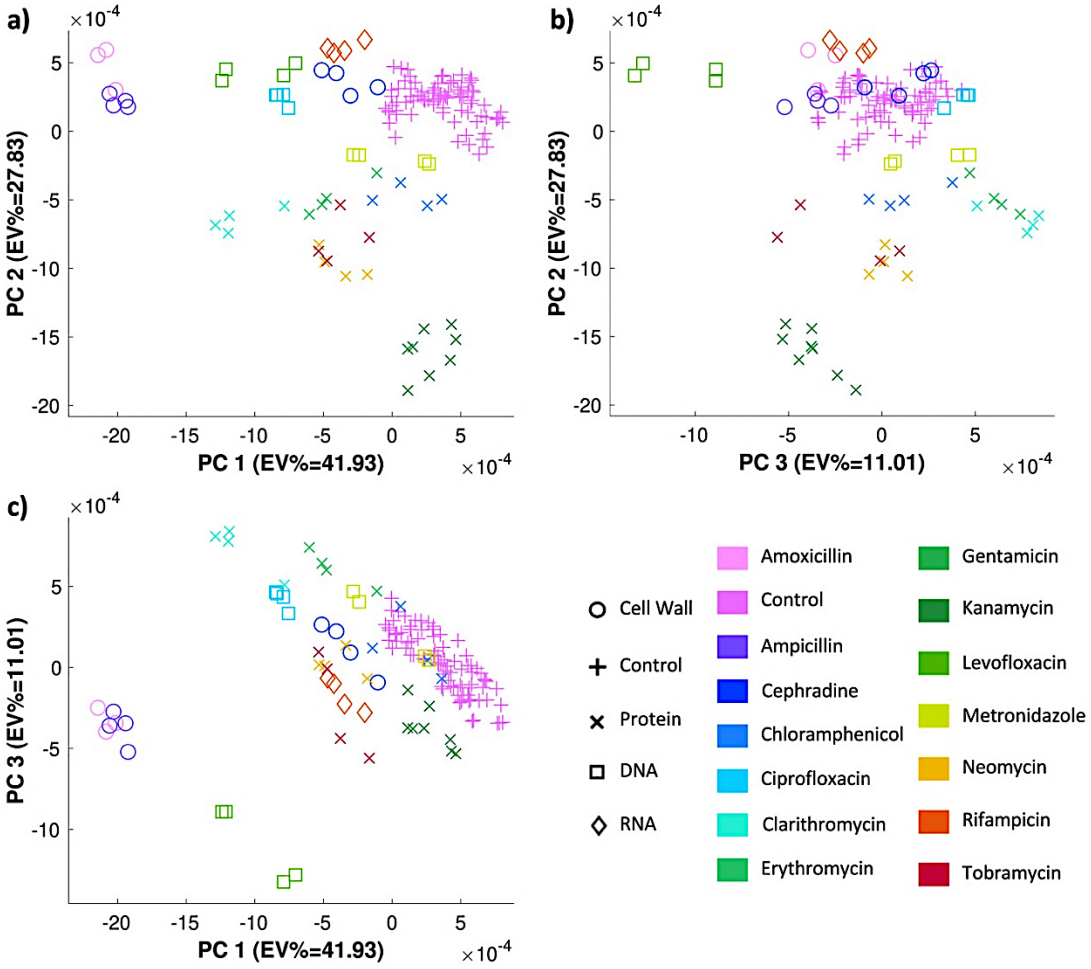


Figure 3. Score plots and explained variability of FTIR spectra of *E. coli* cells exposed to 14 antibiotics, highlighted in different colors, acting on 4 biosynthetic pathways, highlighted with different symbols. Spectra were evaluated with Principal Component Analysis after preprocessing with optimal parameters for antibiotic prediction using a Savitzky-Golay second derivative filter, followed by extended multiplicative scatter correction with replicate correction.

evident grouping according to the antibiotic solvent, which indicates that FTIR spectra of bacteria exposed to different antibiotics more closely reflects their MOA than the effect of their solvent. Moreover, another PCA was done exclusively on the control samples (Supplementary Figure 7), and again a clear pattern of solvent-induced sample separation in the absence of antibiotics was not observed, i.e., the control samples were not grouped together according to the antibiotic solvent.

VIII.3.4 PREDICTION OF KNOWN MECHANISMS OF ACTION

Prior to prediction of known MOA, a few issues were addressed. Firstly, one advantage of using PLS models is the possibility of calculating the variable importance in projection (VIP) scores. These reflect the importance of each variable, i.e., wavenumber, for the outcome of a given PLS model. As such, the VIP scores were overlapped with the spectra to highlight the spectral regions that most contribute to MOA prediction, as exemplified on Figure 1. Here, it became clear that the spectral region associated with atmospheric carbon dioxide (e.g. 2362cm^{-1}) was contributing to MOA classification, so this region was removed from the subsequent analysis, as conducted by others [61]. Secondly, it is reassuring to see that of the spectral bands that most contributed to PLSDA classification, many are typically associated with known biomolecules, e.g., lipids, proteins, carbohydrates, DNA, RNA, phospholipids and phosphorylated proteins (Figure 1 – C). Even so, when using machine learning algorithms, better predictions of individual antibiotics were obtained by removing the least informative spectral regions, i.e., keeping the regions that contributed to the top 2/3 of the VIP scores. The effect of these variable reduction methods is discussed regarding the dose-response of MOA identification, but it is important to note that machine learning after removal of uninformative spectral regions further improved MOA prediction at the antibiotic level by 17%, which coupled to a 21% improvement with preprocessing optimization, amount to an overall improvement of 38% in regard to raw spectra.

Regarding known MOA identification, even though PLSDA proved to be a robust algorithm for the elucidation of antibiotics MOA, better performance was achieved using alternative, and more advanced, predictive algorithms. Additionally, using different predictive algorithms for the same biological query provides higher confidence in the fact that FTIR spectra strongly reflect the antibiotics MOA. As such, optimally preprocessed spectra of *E. coli* cells exposed to the highest concentration of antibiotics were analyzed with a multitude of algorithms (Table 2). In general, MOA resolution was higher at the pathway and class level than at the antibiotic level, independently of the predictive algorithms. Interestingly, the difference in accuracy between pathway and class was mostly smaller than that of class and specific antibiotic effect. The observed coherence among different algorithms, validated with a stringent CV strategy, and substantially distinctive underlying mathematical methods, suggests that there are intrinsic data patterns in the FTIR spectra, and that these have been properly highlighted with an optimized parametrization of preprocessing.

Table 2. Prediction accuracy for the major biosynthetic pathway, antibiotic class, and individual effect of each antibiotic on spectra of *E. coli* cells preprocessed with the optimal parameters of second derivative Savitzky-Golay filtering followed by extended multiplicative scatter correction with replicate correction. Accuracy was evaluated with leave-one-out cross-validation.

Model	Type	Pathway accuracy (%)	Class accuracy (%)	Antibiotic accuracy (%)
SVM	Linear	91.4	96.4	74.1
	Quadratic	97.1	98.6	79.9
	Cubic	99.3	98.6	83.5
	Fine Gaussian	61.9	61.9	63.3
	Medium Gaussian	97.1	98.6	72.7
	Coarse Gaussian	81.3	84.2	61.9
KNN	Fine	99.3	97.8	91.4
	Medium	86.3	87.8	67.6
	Coarse	60.4	60.4	60.4
	Cosine	92.1	87.8	68.3
	Cubic	85.6	87.1	67.6
	Weighted	95.7	97.8	88.5
Ensemble	Bagged trees	97.8	97.1	89.9
	RUSBoosted trees	87.1	81.3	67.6
	Subspace discriminant	98.6	100	97.8
	Subspace KNN	100.0	99.3	88.5

In the case where a known compound is being tested, the models would have already seen the metabolic fingerprint induced by said compound. This scenario was simulated perfectly with LOO-CV since it maximized the number of training samples for all models and allowed an unbiased comparison of different models. In this scenario, prediction of the pathway affected by antibiotics was perfect with an ensemble of a subspace KNN classifiers. At the class level, an ensemble of subspace discriminant classified MOA perfectly. It is interesting to note that the same classifier at the pathway level only failed for a sample

exposed to a protein synthesis inhibitor, which was predicted as a DNA synthesis inhibitor, and an RNA synthesis inhibitor that was predicted as a cell wall synthesis inhibitor. Regarding MOA prediction at the antibiotic level, an ensemble of subspace discriminant achieved excellent accuracy, precision, recall and specificity for each antibiotic (Figure 4). Interestingly, the misclassified RNA inhibitor at the pathway level with an ensemble of subspace discriminant was no longer misclassified at the class or antibiotic level, although the protein synthesis inhibitor chloramphenicol was still classified as a DNA inhibitor, which reinforces the effect of preprocessing on the outcome of predictive algorithms, and therefore the importance of its optimization. Additionally, a sample exposed to tobramycin was predicted as exposed to neomycin and another sample in reverse. Since both tobramycin and neomycin are aminoglycosides, thus target the 30S ribosomal subunit in a similar fashion, it could be the case that MOA resolution becomes limited at the level of individual antibiotics that target the same biomolecule. Ultimately, the classification of known MOA at the level of individual antibiotics was generally very good regarding accuracy, precision, sensitivity and specificity, and despite the few misclassifications discussed, these results clearly encourage the application of FTIR-based MOA identification in antibiotic discovery.

VIII.3.5 Prediction of novel mechanisms of action

Equally important is the scenario where a compound with a novel MOA, either similar or considerably different from known MOA, would be tested. To simulate this, all samples exposed to a given antibiotic were withheld from model training and subsequently MOA was predicted at the antibiotic level with the top performing model for known MOA identification (Figure 5). Notice that four random control samples were selected to simulate a compound with no discernable metabolic effect. Because samples exposed to any given antibiotic were not used to train the model, the objective was for the antibiotic with closest MOA to be identified by the model. For instance, when predicting the MOA of ampicillin, the model should return that of amoxicillin and vice-versa, and likewise for e.g., erythromycin and clarithromycin, etc. In that regard, the most common independent predictions were considered as a consensus prediction, from which the final model prediction was determined. Considering that some MOAs had few fingerprints, this scenario was expected to be more challenging. For instance, samples exposed to rifampicin, an RNA synthesis inhibitor, would always be misclassified as the model would have never seen the metabolic fingerprint of RNA synthesis inhibition. To some extent, this simulates the event of having a compound that disrupts a metabolic pathway that no other compound targets. On the other hand, the model always returned a prediction,

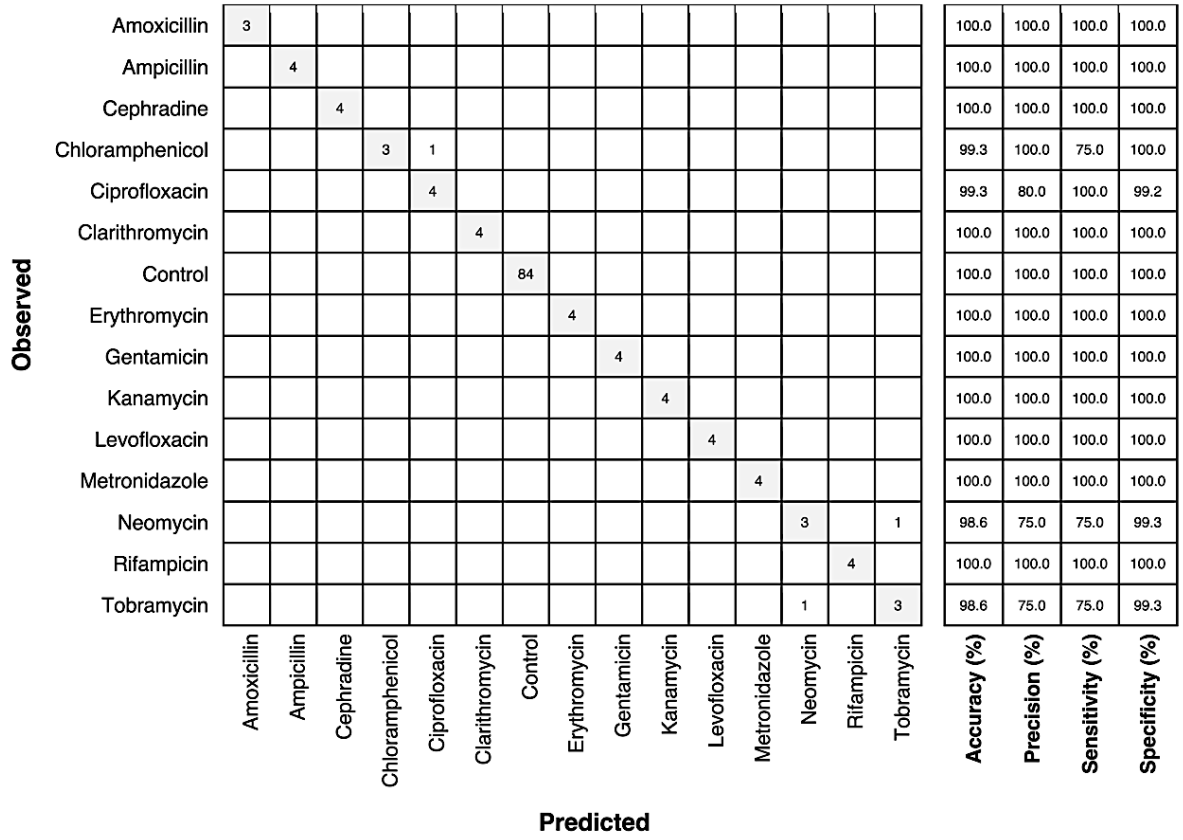


Figure 4. Confusion matrix after leave-one-out cross-validation of the one-versus-all classifications of an Ensemble of discriminant analysis used to predict the effect of individual antibiotics on *E. coli*. Accuracy, precision, sensitivity and specificity are highlighted for each antibiotic. FTIR spectra were preprocessed with optimal parameters of a Savitzky-Golay second derivative filter, followed by Extended Multiplicative Scatter Correction with replicate correction.

e.g., rifampicin-exposed samples were predicted as being exposed to cephradine, which impaired an evaluation of MOA identification with the previously discussed performance indicators. Moreover, even in cases that MOA cannot be comparatively identified, if a metabolic effect or growth inhibition is detected, then follow-up methodologies with higher resolution and lower throughput can be used to pinpoint the MOA.

Identification of novel MOA was surprisingly positive as all consensus predictions obtained matched those expected, excluding samples exposed to cephradine, metronidazole and rifampicin. While the case of rifampicin has been discussed, metronidazole is interesting because it was consistently predicted as a protein synthesis inhibitor, even with PCA (Figure 3). Indeed, the MOA of metronidazole is still not fully understood, and some suggestions have been made that its MOA might be more complex than simply DNA synthesis inhibition [62],

Observed	Predicted				Consensus
Amoxicillin	Ampicillin	Ampicillin	Ampicillin		*Ampicillin*
Ampicillin	Amoxicillin	Amoxicillin	Amoxicillin	Amoxicillin	*Amoxicillin*
Cephradine	Control	Control	Rifampicin	Control	*Control*
Chloramphenicol	Erythromycin	Metronidazole	Erythromycin	Tobramycin	*Erythromycin*
Ciprofloxacin	Metronidazole	Metronidazole	Levofloxacin	Levofloxacin	*Metronidazole/ Levofloxacin*
Clarithromycin	Erythromycin	Erythromycin	Erythromycin	Erythromycin	*Erythromycin*
Control	Control	Control	Control	Control	*Control*
Erythromycin	Clarithromycin	Clarithromycin	Chloramphenicol	Clarithromycin	*Clarithromycin*
Gentamicin	Kanamycin	Kanamycin	Kanamycin	Kanamycin	*Kanamycin*
Kanamycin	Gentamicin	Gentamicin	Gentamicin	Gentamicin	*Gentamicin*
Levofloxacin	Ciprofloxacin	Ciprofloxacin	Rifampicin	Ciprofloxacin	*Ciprofloxacin*
Metronidazole	Chloramphenicol	Chloramphenicol	Clarithromycin	Erythromycin	*Chloramphenicol*
Neomycin	Tobramycin	Tobramycin	Gentamicin	Tobramycin	*Tobramycin*
Rifampicin	Cephradine	Cephradine	Cephradine	Cephradine	*Cephradine*
Tobramycin	Neomycin	Neomycin	Neomycin	Chloramphenicol	*Neomycin*

Figure 5. Predicted MOA of independent cultures withheld from training of a one-versus-all ensemble of discriminant analysis. The consensus prediction, i.e., the most frequently predicted MOA, was used as the overall prediction. FTIR spectra were preprocessed with optimal parameters of a Savitzky-Golay second derivative filter, followed by extended multiplicative scatter correction with replicate correction.

which could justify its poor predictions in this study. As such, it is reasonable to assume that the poorer prediction obtained for either rifampicin or metronidazole could be unrelated to the suitability of FTIRS for MOA identification. For instance, two ciprofloxacin-exposed samples were predicted as being exposed to metronidazole, which suggests that the metabolic fingerprint induced by metronidazole has large resemblances with that of other DNA synthesis inhibitors, which is coherent with the expected MOA. On the other hand, this could be due to both ciprofloxacin and metronidazole being dissolved in an acidic solvent, which is more unlikely given that the solvent did not seem to be reflected on the metabolic fingerprints. Either way, a larger set of DNA synthesis inhibitors, other than fluoroquinolones, could shed some light on this issue, as well as the use of a universal solvent. The case of cephradine is more challenging to interpret. Because of the PCA results, along with the LOO-CV PLSDA and machine learning models for similar MOA identification, it seems that cephradine induced a

consistent metabolic fingerprint across independent cultures. Moreover, said fingerprint resulted in coherent predictions across different algorithms when the cephradine-exposed samples were included in the training data set, which suggests that this fingerprint was reproducibly captured with FTIRS, but if excluded from training, this fingerprint was not predicted as similar to the other beta-lactams. Although unlikely, it could be the case that the MOA of cephradine is sufficiently distinct from that of amoxicillin and ampicillin to induce considerably different metabolic fingerprints. Training the models with more fingerprints of cephalosporins and other cell wall inhibitors could shed some light on the nature of this misclassification. On a different note, the fact that chloramphenicol, the single representative of the amphenicol class of antibiotics, was predicted as erythromycin is a positive indication, given both target the 50S ribosomal subunit. In the end, the classification of novel MOA at the level of individual antibiotics was largely very good, particularly when the model was presented with similar metabolic fingerprints. To tackle the cases with poorer performances, holistic and extensive databases of metabolic profiles for FTIR-based MOA identification will be required.

VIII.3.6 Dose-response of MOA identification

So far, the analysis has considered spectra obtained from bacteria exposed to the highest concentration of antibiotics tested. This was done to reduce the prohibitively large computation time of testing the extensive preprocessing parameter combinations for the wide range of antibiotic concentrations to which bacteria were exposed to, as well as to maximize the antibiotic-induced metabolic fingerprint component of the spectra. After determining the optimal preprocessing parameters, addressing the possibility of overfitting, validating the supervised classification obtained with unsupervised PCA, removal of uninformative spectral regions, and the application of several machine learning strategies, the top-performing algorithm was used to query if the highest concentration was the most adequate for elucidating the MOA of antibiotics. As such, the effect of antibiotic concentration on MOA identification was evaluated for each twofold dilution of antibiotic concentration, which ranged from 512 μg/mL through to 0.03 μg/mL, in addition to the grouped control samples (Figure 6). Clearly, higher antibiotic concentrations lead to improved MOA identification at either the pathway, class or antibiotic level. Therefore, the premise of considering spectra obtained from bacteria exposed to the highest concentration of antibiotics tested seems to hold true and supports the fact that these spectra were used for preprocessing optimization and to build more advanced predictive models.

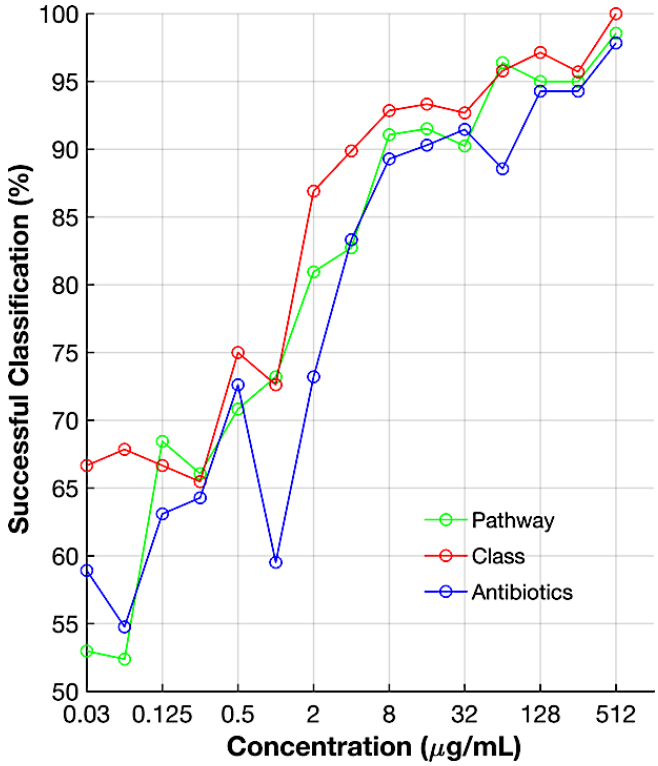


Figure 6. Effect of the antibiotic concentration on the successful classification of the major biosynthetic pathway, antibiotic class and antibiotic-specific metabolic fingerprints induced on *E. coli* cells after exposure to 14 antibiotics. Successful classification was evaluated with leave-one-out cross-validation of an ensemble of discriminant analysis of FTIR spectra preprocessed with the respective optimal parameters of Savitzky-Golay second derivative filtering followed by extended multiplicative scatter correction with replicate correction.

It is interesting to note that the successful classifications of pathway, class and individual antibiotics differ very slightly. Prior to the removal of uninformative spectral regions, and using PLSDA, MOA identification at the antibiotic level was clearly poorer than at either the class or pathway level (Supplementary Figure 8). This suggests that the workflow employed was particularly important to improve MOA resolution at the antibiotic level, irrespectively of the antibiotic concentration. Additionally, this dose-response consolidates the notion that FTIR spectra have intrinsic data patterns that are relevant for elucidating antibiotics MOA, and these patterns appear even at very low antibiotic concentrations that only induce a slight growth inhibition. For instance, over 70% accurate MOA prediction at the pathway and class level was achieved with an antibiotic concentration of just 1 μg/mL, which only induced slightly over 15% growth inhibition. This is particularly relevant if the high cell density is considered, which would correspond to an even lower antibiotic concentration in a typical

susceptibility assay. In the end, these results are a positive indicator that FTIRS can be used to explore the grey chemical matter, although it is important to stress that lower concentrations of antibiotics might not be an accurate interpretation of compounds from the grey chemical matter. Remembering that these compounds induce some phenotypic modulation but only present low activity in typical screening conditions, their low activity is most likely due to other factors beyond concentration, e.g., binding affinity, which can be improved with medicinal chemistry.

VIII.3.7 Estimation of Antimicrobial Potency

After dwelling into MOA prediction, the ability of the proposed method in determining if a candidate molecule has an antibiotic effect was examined. In the simplest terms, a molecule with an antibiotic effect has antagonistic properties on the growth of microbes [63]; therefore, a good estimator of the antibiotic effect should probe microbial growth, which was achieved by measuring the ABS_{600} simultaneously with FTIR spectra (Supplementary Table 1). Note that after measuring ABS_{600}, but prior to FTIR spectra readings, the biomass of each sample was normalized, therefore all spectra had approximately the same cellular density. The optimized preprocessing parameters discussed thus far were used for regression, but instead of EMSCrep, EMSC was used given that EMSCrep suppressed some of the spectral variability related to growth inhibition, since replicates more closely reflect a similar MOA, than growth inhibition, given the latter is subject to higher variability. As such, a second-degree polynomial, fitted over a window size of 9 points, was used to determine the second derivative spectra with the SG filter, followed by EMSC.

In regard to model validation, a PLS regression (PLSR) model was used to estimate the optical density of each sample, which decreased after exposure to any of the drugs tested (Table 1), from which overfitting was ruled out using the coefficient of determination (R^2) and the Root-Mean-Square Error (RMSE) (Supplementary Figure 9). The RMSE can be thought of as a normalized distance between predicted and observed values; thus, it is a good estimator of the standard deviation of the prediction error. Since a PLSR with 10 LVs produced an R^2 of 0.84 and a RMSE of 0.46, and that control samples had an average ABS_{600} of 4.00, while those exposed to the maximum tested concentration had an average ABS_{600} of 1.86, even at two standard deviations away from said means there was virtually no overlap between these two sample groups, assuming a normal distribution. This roughly means a correct

prediction of growth inhibition above 95% for samples exposed to the highest tested antibiotic concentration.

The optimized and validated dataset was explored with more advanced algorithms (Table 3) towards better performance, and to consolidate that FTIR spectra strongly reflected the antagonistic effect of antibiotics on microbial growth. This was especially important to validate because the biomass of each sample was set to a predefined value prior to spectral acquisition, thus all spectra had approximately the same cellular density. So, if growth inhibition was intrinsically reflected on the FTIR metabolic fingerprints, it was not just a function of cellular concentration. A squared exponential Gaussian process regression yielded a RMSE of 0.33 and a R^2 of 0.92, which was the best performance of all algorithms considered and whose predicted values variability were visualized as a function of the observed data (Figure 7). Here, the variability of the predicted data was not clearly larger for a particular range of ABS_{600}; in other words, it seems randomly distributed. This suggests there was little or no bias of the model for over- or under-estimation. This does not mean that the model did not over- or under-estimate the actual value; rather, this occurred in a random fashion. In fact, it is possible that some of this random noise was introduced with the biomass normalization step of the bioassay and could therefore be reduced. Ultimately, these results suggest FTIR spectra reflected the ABS_{600}, and so these could be used to identify bacterial growth inhibition, which in turn can be used to estimate the potency of antibiotics similarly to a spectrophotometric susceptibility testing method. Importantly, potency could be estimated despite the normalization of biomass prior to spectral acquisition, which is most likely due to the metabolic fingerprints detected with FTIRS reflecting the alterations induced by antibiotics, which lead to the growth inhibition detected when measuring the ABS_{600}. As growth inhibition can be detected, promising compounds can be revealed even in cases where MOA identification was not as successful, for which follow-up methodologies can be used to pinpoint the MOA. As such, this opens the door on a single step assay capable of simultaneously determining the MOA of antibiotic candidates and their potency.

VIII. 4 Discussion

Antibiotic discovery has plummeted to frightening rates. High-throughput methods capable of probing the complex MOA of antibiotics are urgently needed [64]. Metabolomics is an especially promising technique for MOA identification because it reflects the downstream products of biomolecular processes; hence, it is a closer reflection of phenotype than, for

Table 3. Performance of regression algorithms regarding the coefficient of determination (R^2), the Root-Mean-Square Error (RMSE), the Mean Square Error (MSE) and the Mean Absolute Error (MAE). Regression was performed with spectra of *E. coli* cells preprocessed with the optimal parameters of second derivative Savitzky-Golay filtering followed by extended multiplicative scatter correction against the absorbance at 600nm after a 16h exposure to twofold dilutions of antibiotics. All models were evaluated with 50-fold cross-validation.

Model	Type	RMSE	R^2	MSE	MAE
Tree	Fine	0.54	0.78	0.29	0.36
	Medium	0.55	0.77	0.30	0.37
	Coarse	0.60	0.73	0.36	0.43
SVM	Linear	0.85	0.46	0.72	0.46
	Quadratic	0.40	0.88	0.16	0.31
	Cubic	0.40	0.88	0.16	0.29
	Fine Gaussian	1.14	0.02	1.29	0.94
	Medium Gaussian	0.39	0.89	0.15	0.29
	Coarse Gaussian	0.80	0.52	0.64	0.52
Ensemble	Boosted trees	0.42	0.87	0.18	0.31
	Bagged trees	0.37	0.89	0.14	0.27
Gaussian Process Regression	Squared exponential	1.15	-0.00	1.32	0.95
	Matern 5/2	1.15	0.00	1.32	0.95
	Exponential	0.78	0.54	0.60	0.55
	Rational quadratic	0.33	0.92	0.11	0.25

instance, transcriptomics. Often, transcriptomic profiles reflect indirect responses overlaid with the specific antibiotic responses, thus are noisier and more challenging to use when pinpointing the MOA [65]. As such, Halouska et al. used Nuclear Magnetic Resonance (NMR) metabolomics to obtain biochemical signatures that reflect antibiotic efficacy, selectivity and toxicity and used PLSDA to predict the MOA of 3 known drugs, with unknown mechanism, by comparing said signatures to those obtained from 12 known drugs with known mechanisms [66]. Similarly, Vincent et al. suggested the use of Liquid Chromatography Mass Spectrometry

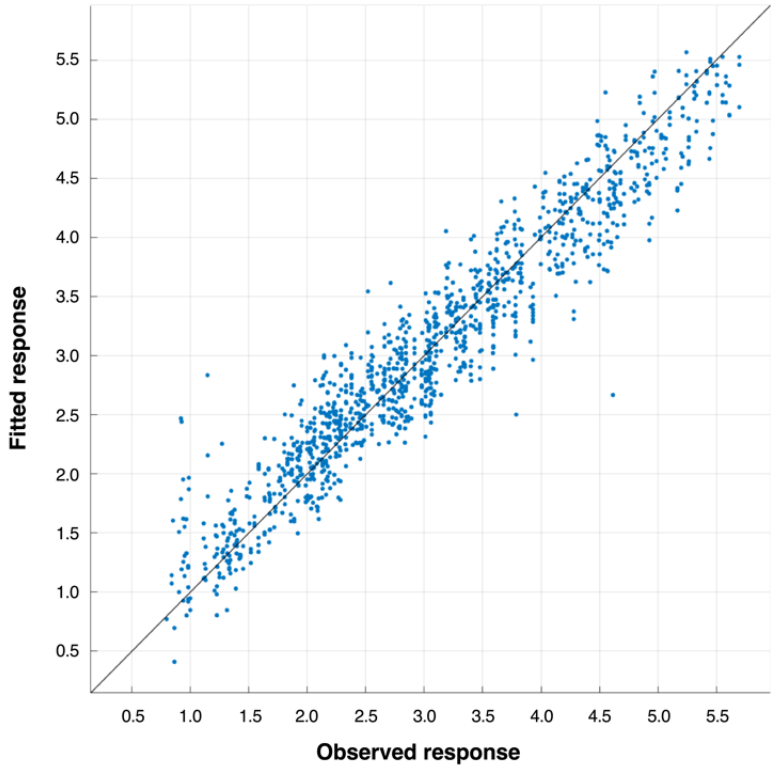


Figure 7. Observed *versus* fitted response of a rational quadratic Gaussian process regression performed over spectra of *E. coli* cells preprocessed with the optimal parameters of second derivative Savitzky-Golay filtering followed by extended multiplicative scatter correction against the absorbance at 600nm after a 16h exposure to twofold dilutions of 14 antibiotics.

(LC-MS) to overcome the constraint of NMR to highly abundant metabolites [67]. Although the exact target of 4 compounds were identified, 2 other drugs presented metabolic profiles with predominant off-target effects that impaired MOA identification, and a compound that did not affect any metabolic pathway did not present a distinguishable profile, despite inhibiting bacterial growth. Moreover, considerable variation of metabolic profiles was observed in the control samples, which reinforces the need for stringent controls and prudent analysis with MS, and also exposes a limitation of metabolomics-based MOA identification. Often, MS-based metabolomics with extremely high sensitivity renders it difficult to operate and challenging to interpret.

Recently, there has been an effort towards high-throughput metabolomics. One such method has been suggested by Zampieri et al., who applied MS to predict the MOA of uncharacterized antimycobacterial compounds [68]. In said study, a reference of metabolic responses was built and then used to query the MOA of over 200 compounds of a

GlaxoSmithKline library. To achieve high throughput, metabolites were extracted directly from the whole culture broth, without cell separation, and then were directly injected into a time-of-flight mass spectrometer. Omission of chromatography and direct sample injection into the electrospray interface had been shown to reduce cycle time to 1 minute per sample [69]. In the study by Zampieri et al., the large drug-metabolome collection of profiles enabled the authors to identify the MOA of compounds that did not target metabolic pathways. However, given the sensitivity of MS, divergent metabolic profiles after drug exposure were detected. As such, decoupling metabolic changes that are drug-specific from those that are indirect adaptations remained challenging, and required time-resolved metabolic profiles that imply lower-throughput. In addition, the effect of compounds that interact with low expression or inactive targets might be below the limit of detection, but this is a transversal issue to the field of antibiotic discovery.

An alternative to metabolomics is Bacterial Cytological Profiling (BCP), a method based on microscopic observations of drug-induced effects. For instance, O'Driscoll et al. visualized the effect of colistin with scanning electron microscopy, namely cellular aggregation, from which DNA synthesis was excluded as the likely target, since DNA damage induces filamentation, which was not observed [70]. However, in the context of high throughput, other microscopy techniques are more relevant. Nonejuie et al. used fluorescence microscopy to capture the cytological profile of 41 antibiotics [71]. From these, inhibitors of different pathways, as well as those acting on the same pathway, were distinguished visually on PCA score plots and statistically with Spearman's rank correlation. Yet, the authors reinforced that their methodology did not identify the precise molecular target and therefore would require a follow-up study to pinpoint the MOA, for instance with a panel of resistant mutants or of sensitized strains.

Similarly, Zoffman et al. also used fluorescence microscopy to screen 1.5 million compounds from the Roche library, after it had been reduced down to 750 compounds with dose-response screens [72]. The authors captured over 100 imaging features per sample, which were fed to machine learning algorithms that yielded a similarity score in regard to a set of MOA of reference compounds. Even with careful calibration, and an extremely large set of reference features, the authors reported a large intrinsic variability of the microscopic parameters. Additionally, because phenotypic differences become evident below the MIC, the authors suggest that their method can be used to identify low potency hits. However, the authors also recognized that morphology is not affected by all antibacterial compounds, thus

complemented their methodology with sequencing-derived molecular information. This proved to be useful as some compounds only led to molecular or morphological differences but highlights a key limitation of BCP.

FTIRS has been suggested as an innovative technique suitable for the detection of metabolic fingerprints induced by antibiotics. In comparison with BCP, FTIRS presents the advantage of not requiring reagents beyond growth media and the compounds, e.g., no need for fluorescent dyes. Also, FTIRS is less laborious, especially regarding the hands-on time. For instance, the proposed method only requires one washing step (e.g., centrifugation and resuspension) after incubation with the antibiotics and prior to FTIR spectra acquisition. On the other hand, the method suggested by Zoffman et al. requires addition of fluorescent dyes before the end of incubation with the compounds. After incubation, PBS had to be added to each sample and samples needed to be washed. Followed another incubation step, another washing step, sample dilution and transfer of a small volume to a multi-well plate for imaging.

Moreover, FTIRS provides a better reflection of the biomolecular phenomena underlying the antibiotic effect, since spectra derive from vibrations of different biomolecular bonds present in the sample, rather than the higher-level morphological alterations that BCP detects. As seen, not all antibiotics had a clear effect on morphology, so BCP will be inherently biased towards those that do. Metabolomics-based MOA identification is another key technique. Although MS is generally accepted as a more sensitive technique than FTIRS, this does not make it necessarily more suitable for any given purpose. In the case of bacterial typing at the species level, MS has been shown to be a superior technique, however for subspecies discrimination FTIRS is preferable. This is because it includes information regarding several types of biomolecules, such as lipids, nucleic acids, carbohydrates and proteins [73]. In other words, although FTIRS's sensitivity is not comparable to that of MS, its ability to detect a wider range of molecules deems FTIRS more suitable for certain applications, as could be the case for MOA identification. Also, as previously discussed, the high sensitivity of MS-based techniques often results in data with high variability that is challenging to interpret.

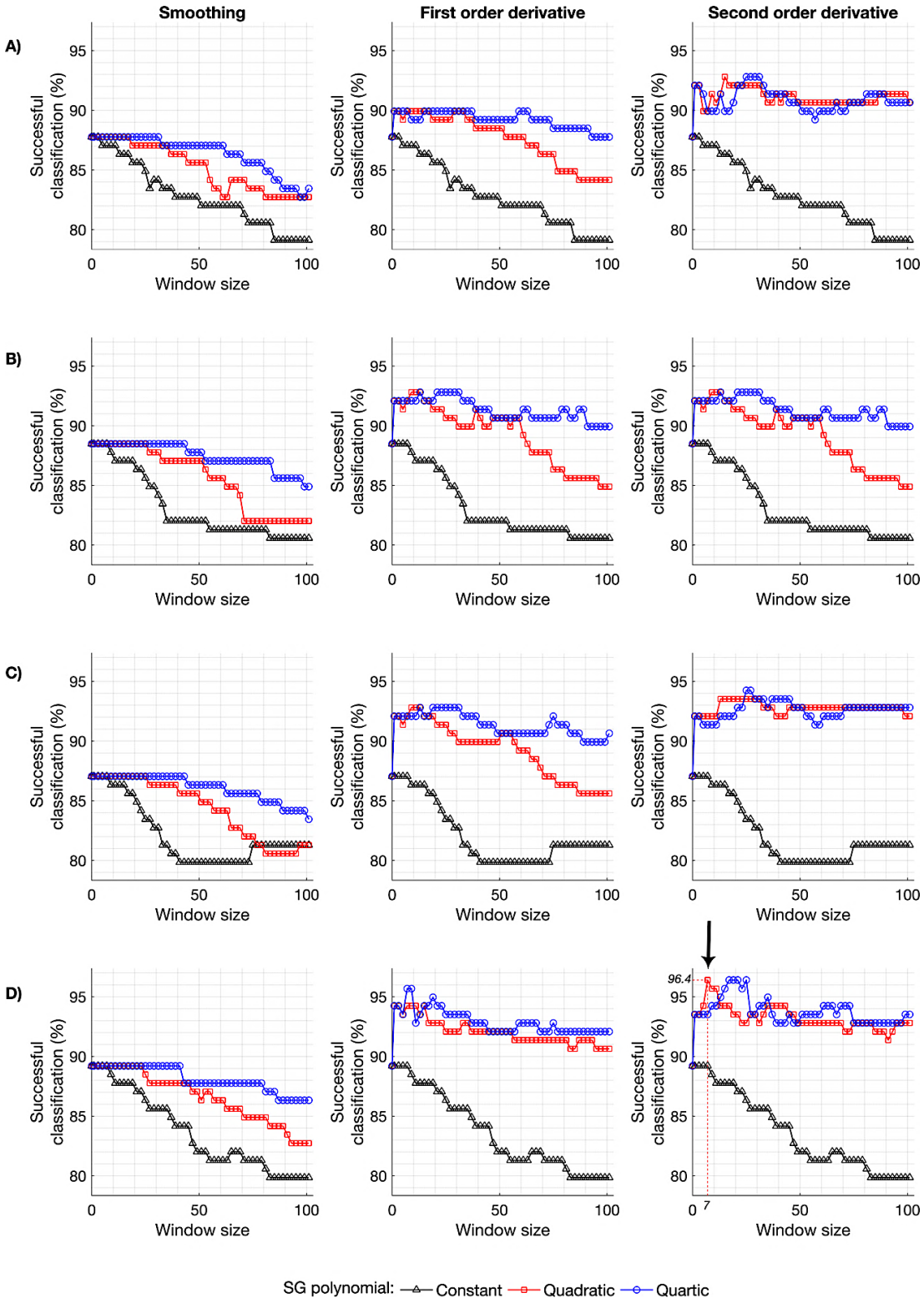
In this study we aimed to consolidate FTIRS as a metabolic fingerprinting technique that is highly suitable for MOA identification. In particular, MOA resolution was evaluated at the level of individual antibiotics, different antibiotic classes, as well as antibiotics acting on different biosynthetic pathways. Also, preprocessing optimization was shown to be a center stone of FTIRS studies in general, and those that focus on MOA in particular, given up to 20%

improvements were obtained in PLSDA accuracy. Moreover, a plethora of predictive algorithms were employed in an attempt to improve PLSDA performance, and also to ascertain that the results were not dependent on the mathematical formulations of the different algorithms. In that regard, various algorithms consistently matched or exceed the predictive performance obtained with PLSDA for the three levels of metabolic sensitivity queried.

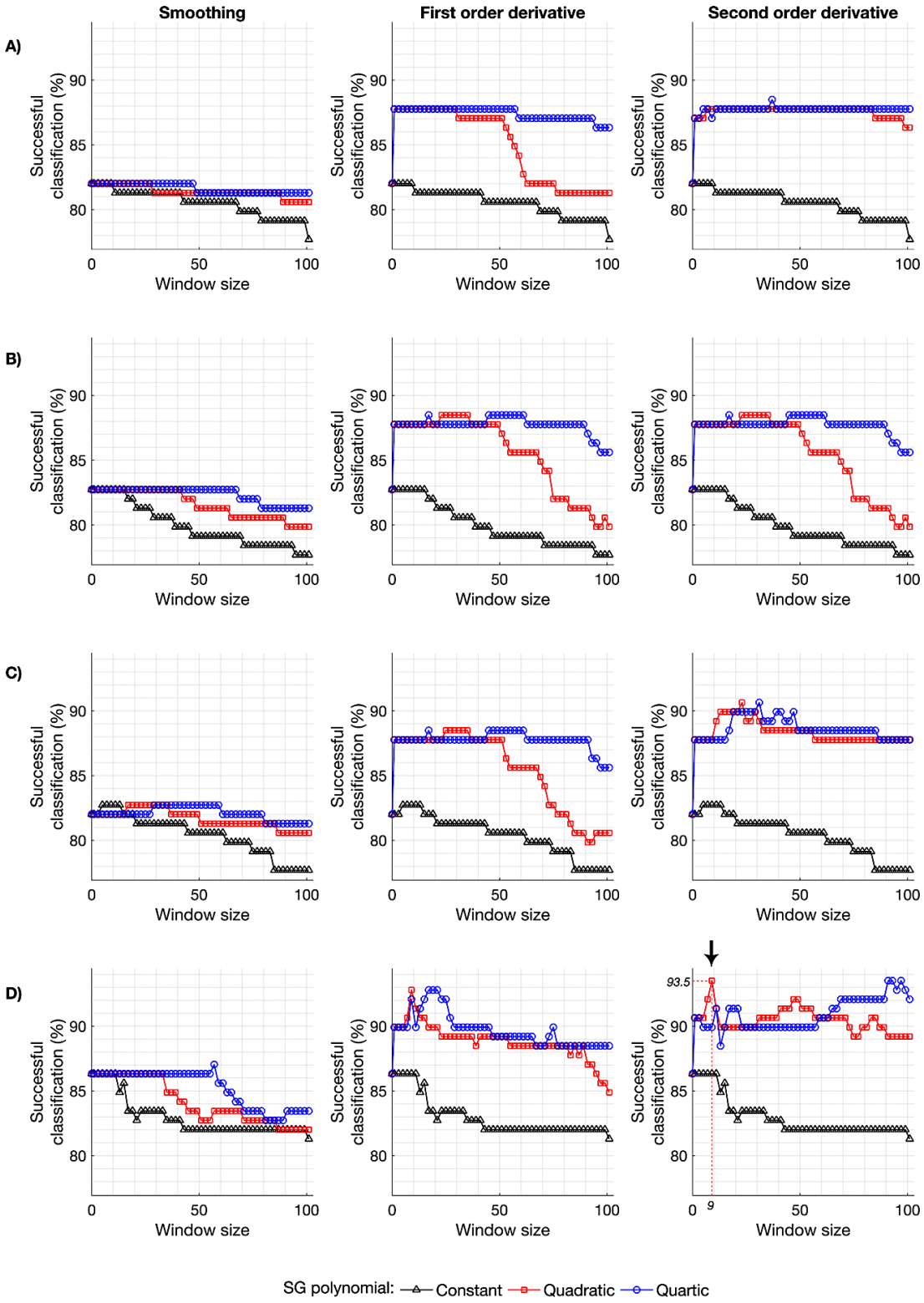
In the end, using the optimal preprocessing combination and an ensemble of discriminant analysis yielded a near-perfect MOA prediction at the pathway (98.6%), class (100%), and individual antibiotics (97.8%) level. This corresponds to an additional 17% improvement of antibiotic prediction on top of those obtained with preprocessing optimization. Looking deeper at the very few samples that were misclassified reveals that MOA resolution might only be slightly limited when predicting the effect of compounds of the same class with very similar MOA. Additionally, by looking at the dose-response of MOA prediction, we have shown that for low antibiotic concentrations, MOA can still be accurately detected at all MOA levels, which opens the door for FTIRS-based explorations into the grey chemical matter. For that, it is of paramount importance to validate these results with compounds from the grey chemical matter, as these might induce a very different metabolic response in comparison to low concentrations of antibiotics. Ultimately, FTIRS is a very promising technique to identify MOA and has possible applications to probe the grey chemical matter, which warrants further efforts.

In addition to MOA identification, we have shown that FTIRS can quickly estimate antibiotic potency. Besides resulting in a substantial improvement of PLSR performance, using various machine learning algorithms, with different mathematical formulations, was again crucial to consolidate the fact that the metabolic fingerprints detected with FTIRS reflect the alterations induced by antibiotics, which eventually lead to the growth inhibition that was detected when measuring the ABS_{600}. This is particularly important because it allows the combination of two steps of the phenotypic screening workflow in a single assay, namely the identification of compounds that inhibit cell growth along with the underlying mechanism of said inhibition. In combination with the reported ability to identify MOA at sub-inhibitory concentration, our results indicate that FTIRS is highly accurate towards the MOA of antibiotics, as well as to the degree that said MOA is affecting the cellular metabolism and inhibiting bacterial growth, beyond cell density.

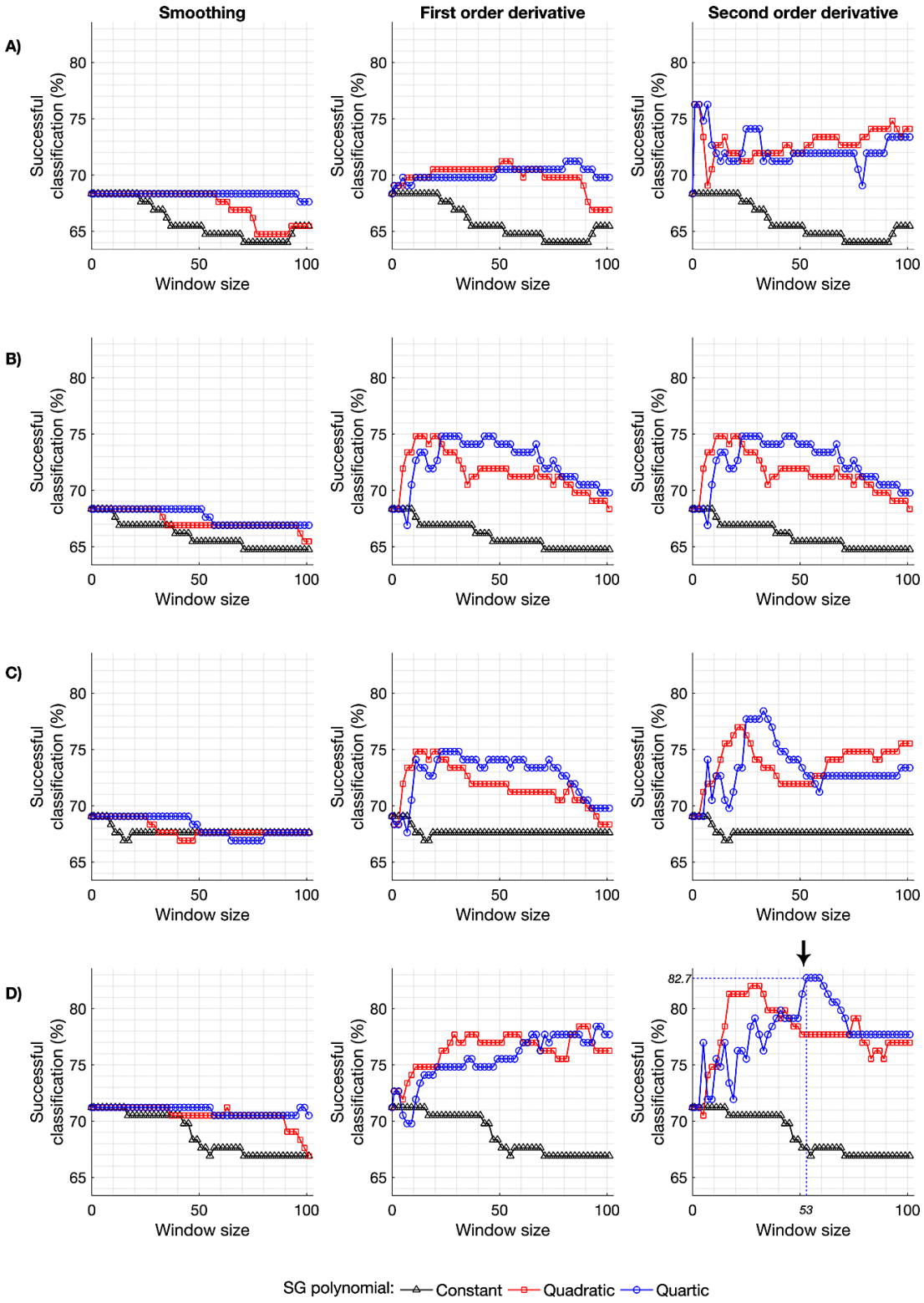
VIII. 5 SUPPLEMENTARY DATA



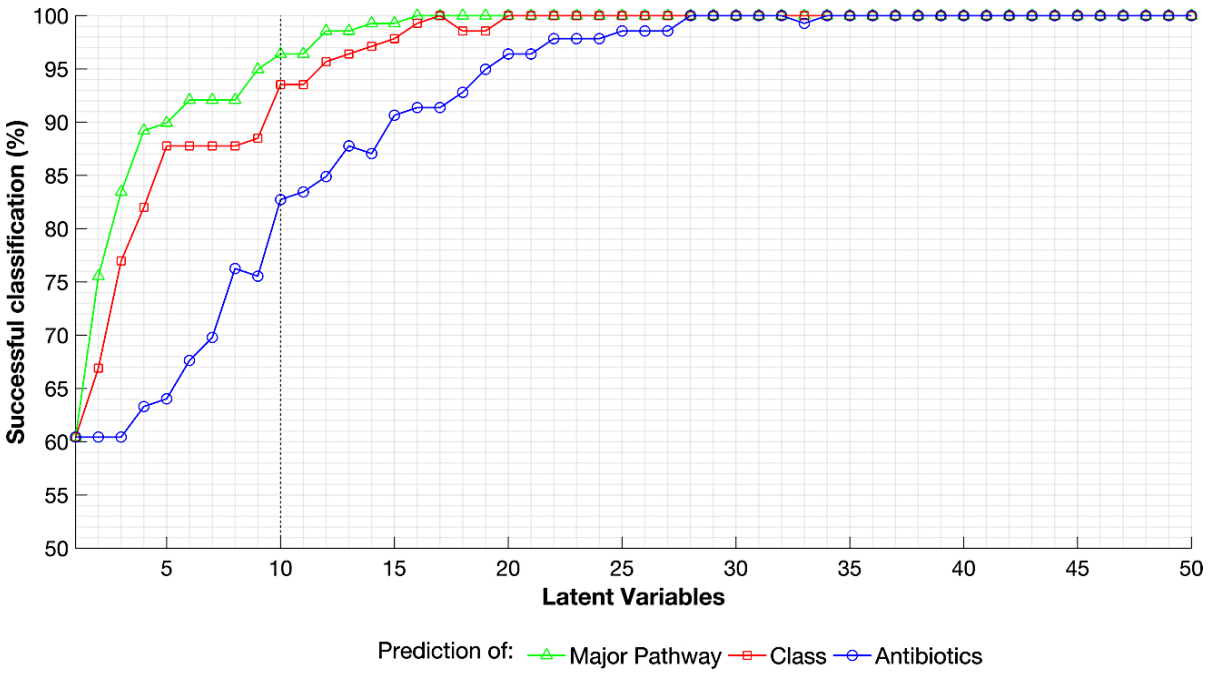
Supplementary Figure 1. Successful classification of the major biosynthetic pathway disrupted on *E. coli* cells after exposure to 14 antibiotics. Successful classification was evaluated with leave-one-out cross-validation of partial least squares discriminant analysis of FTIR spectra preprocessed with Savitzky-Golay filtering (A), followed by Multiplicative Scatter Correction (B), or Extended Multiplicative Scatter Correction (C), or Extended Multiplicative Scatter Correction with replicate correction (D).



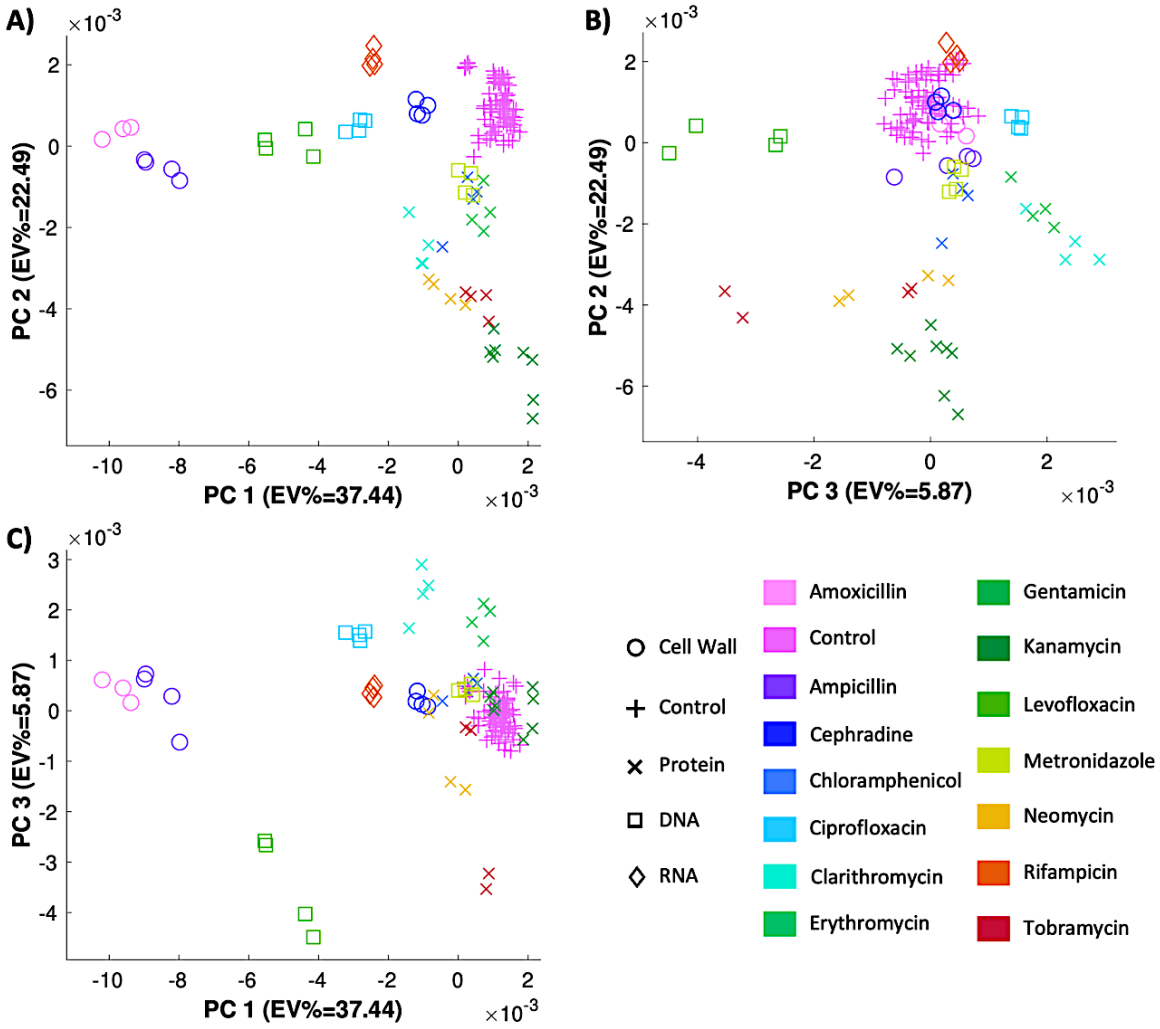
Supplementary Figure 2. Successful classification of the antibiotic class from spectra of *E. coli* cells exposed to 14 antibiotics. Successful classification was evaluated with leave-one-out cross-validation of partial least squares discriminant analysis of FTIR spectra preprocessed with Savitzky-Golay filtering (A), followed by Multiplicative Scatter Correction (B), or Extended Multiplicative Scatter Correction (C), or Extended Multiplicative Scatter Correction with replicate correction (D).



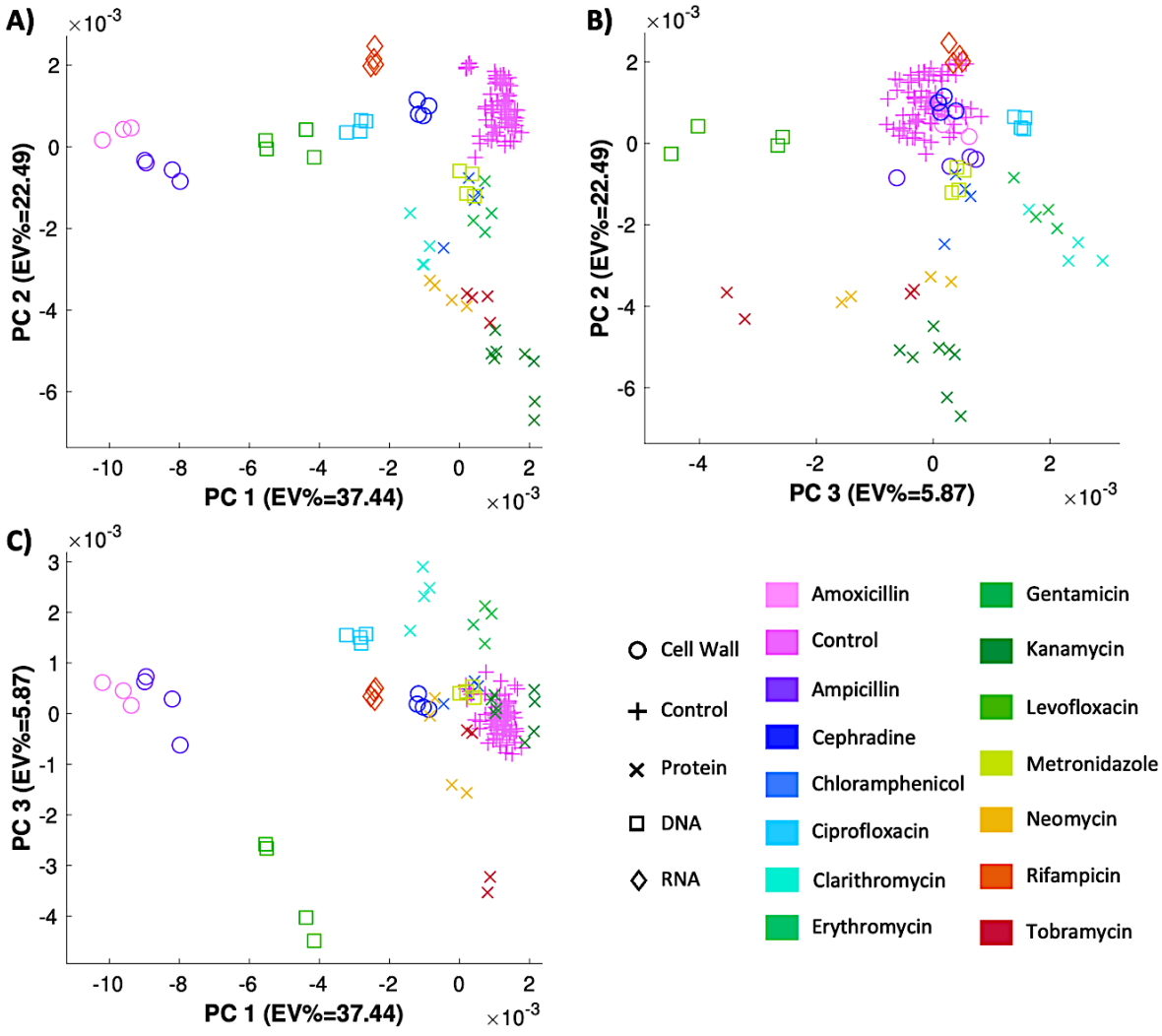
Supplementary Figure 3. Successful classification of the antibiotic-specific metabolic fingerprint imposed by 14 antibiotics on *E. coli* cells. Successful classification was evaluated with leave-one-out cross-validation of partial least squares discriminant analysis of FTIR spectra preprocessed with Savitzky-Golay filtering (A), followed by Multiplicative Scatter Correction (B), or Extended Multiplicative Scatter Correction (C), or Extended Multiplicative Scatter Correction with replicate correction (D).



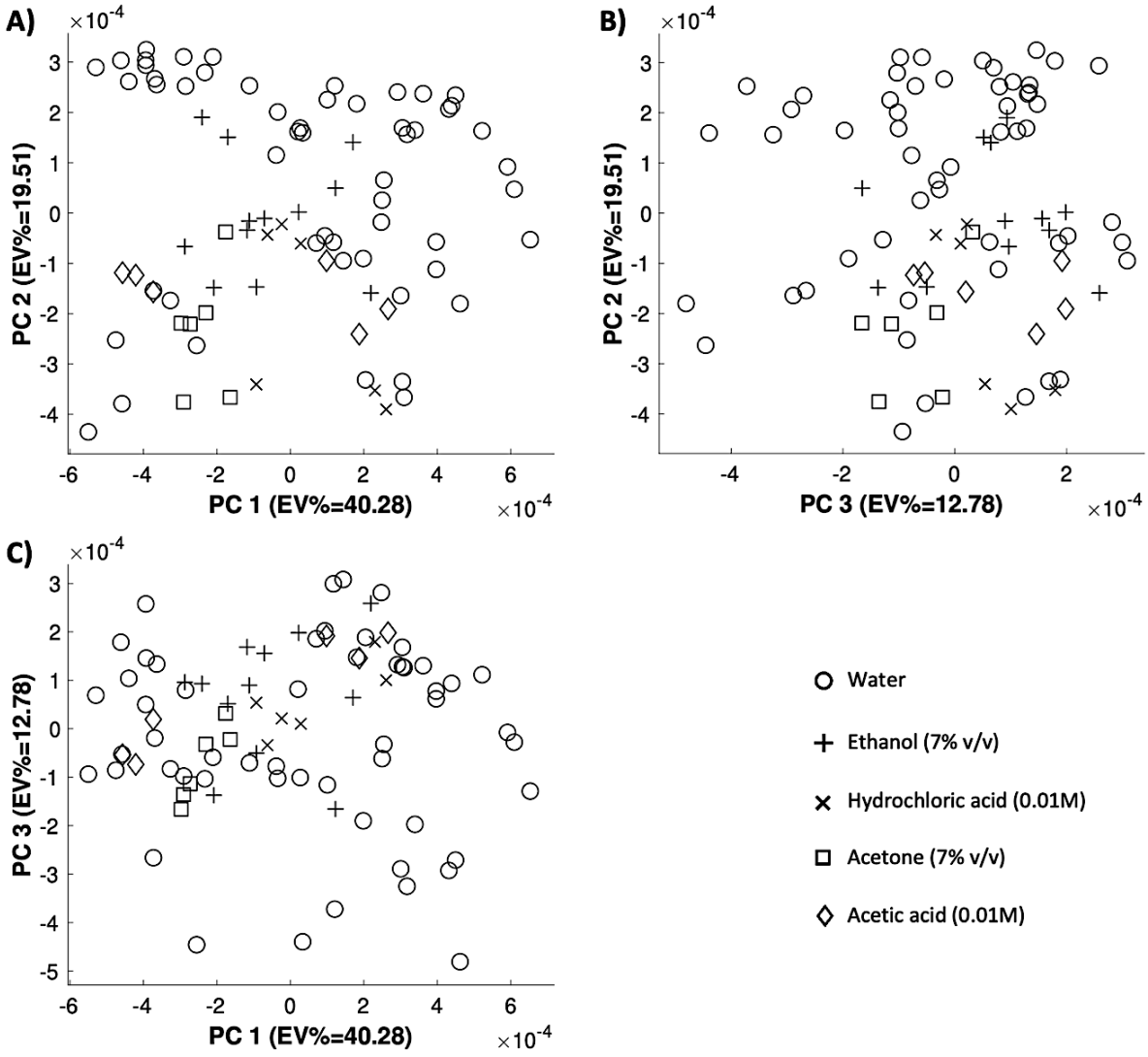
Supplementary Figure 4. Effect of the number of latent variables on the successful classification of the pathway affected, antibiotic class and antibiotic-specific metabolic fingerprints induced on *E. coli* cells. Successful classification was evaluated with leave-one-out cross-validation of partial least squares discriminant analysis of Fourier-transform infrared spectra preprocessed with optimized Savitzky-Golay second derivative filtering followed by extended multiplicative scatter correction with replicate correction. Vertical dashed line indicates the number of latent variables used for preprocessing optimization.



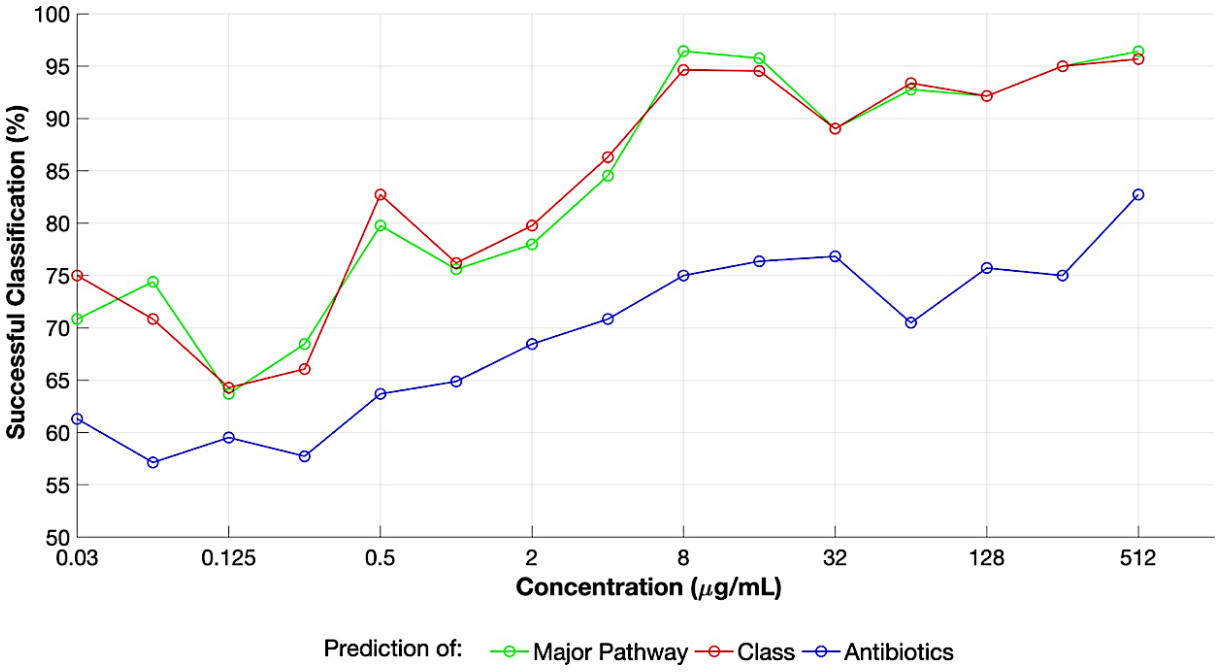
Supplementary Figure 5. Score plots and explained variability of Fourier-transform infrared spectra of *E. coli* cells exposed to 14 antibiotics, highlighted in different colors, acting on 4 biosynthetic pathways, highlighted with different symbols. Spectra were evaluated with principal component analysis after preprocessing with optimal parameters for antibiotic class prediction for Savitzky-Golay second derivative filtering, followed by extended multiplicative scatter correction with replicate correction.



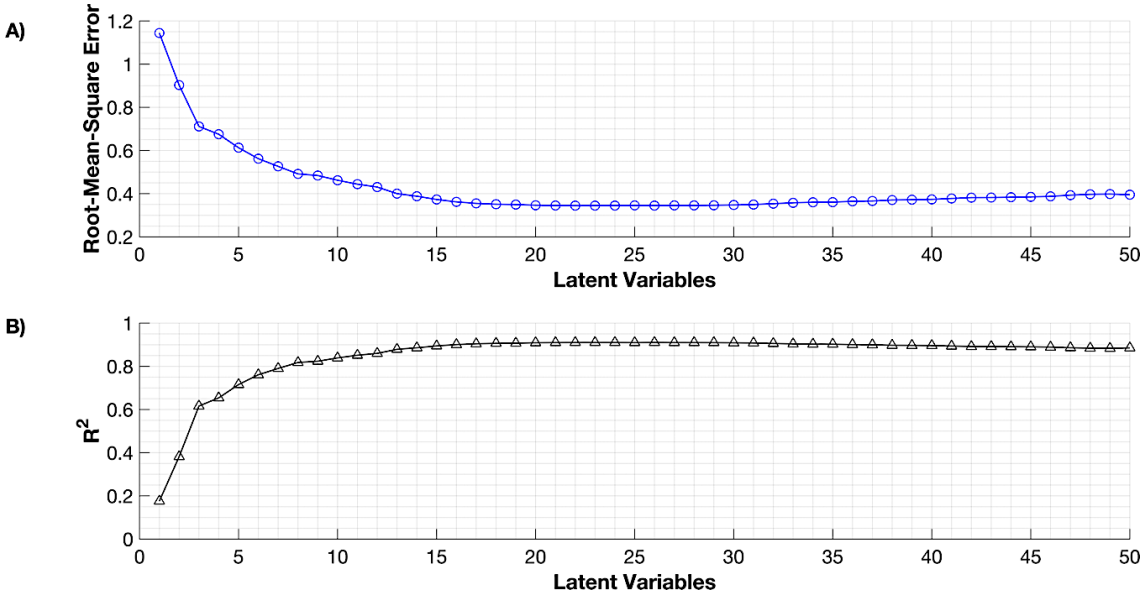
Supplementary Figure 6. Score plots and explained variability of Fourier-transform infrared spectra of *E. coli* cells exposed to 14 antibiotics, highlighted in different colors, acting on 4 biosynthetic pathways, highlighted with different symbols. Spectra were evaluated with principal component analysis after preprocessing with optimal parameters for the prediction of the major pathway affected using a Savitzky-Golay second derivative filtering, followed by extended multiplicative scatter correction with replicate correction.



Supplementary Figure 7. Score plots and explained variability of Fourier-transform infrared spectra of *E. coli* cells not exposed to antibiotics, i.e., the control samples, dissolved in different solvents, highlighted with different symbols. Spectra were evaluated with principal component analysis after preprocessing with optimal parameters for antibiotic prediction for Savitzky-Golay second derivative filtering, followed by extended multiplicative scatter correction with replicate correction.



Supplementary Figure 8. Effect of the antibiotic concentration on the successful classification of the major biosynthetic pathway, antibiotic class and antibiotic-specific metabolic fingerprints induced on *E. coli* cells after exposure to 14 antibiotics. Successful classification was evaluated with leave-one-out cross-validation of partial least squares discriminant analysis of Fourier-transform infrared spectra preprocessed with the respective optimal parameters of Savitzky-Golay second derivative filtering followed by extended multiplicative scatter correction with replicate correction.



Supplementary Figure 9. Effect of the number of latent variables on the coefficient of determination (R^2) and the root-mean-square error of a partial-least-squares regression evaluated with leave-one-out cross-validation. FTIR spectra of *E. coli* cells were preprocessed with optimized Savitzky-Golay second derivative filtering followed by extended multiplicative scatter correction.

Supplementary Table 1. Adjusted absorbance at 600nm of *E.coli* after 16h of exposure to serial dilutions of 14 antibiotics. Results were averaged over biological replicates.

Antibiotic	**Concentration (μg/mL)**															
	0	**0,03**	**0,06**	**0,125**	**0,25**	**0,5**	**1**	**2**	**4**	**8**	**16**	**32**	**64**	**128**	**256**	**512**
Amoxicillin	4,69	4,76	4,71	4,57	4,66	4,28	4,34	3,59	2,56	2,01	1,52	1,21	0,93	0,91	0,83	0,84
Ampicillin	4,97	4,92	4,64	4,71	5,10	4,60	4,88	4,20	3,40	2,77	2,10	1,75	1,52	1,35	1,24	1,19
Cephradine	5,39	5,49	5,32	5,58	5,11	5,56	5,51	5,06	5,19	4,97	4,47	4,68	4,33	4,17	3,55	2,92
Chloramphenicol	4,37	3,81	3,52	3,45	3,19	3,40	3,47	2,85	2,62	2,36	2,33	2,39	2,38	2,33	2,68	2,81
Ciprofloxacin	3,86	3,39	3,43	3,22	2,96	3,03	2,99	2,93	2,68	2,46	2,31	2,19	2,21	2,22	2,03	2,08
Clarithromycin	2,80	2,37	2,23	2,32	2,22	2,28	2,44	2,30	2,42	2,19	1,92	1,86	1,75	1,48	1,24	1,14
Erythromycin	3,86	3,39	3,43	3,22	2,96	3,03	2,99	2,93	2,68	2,46	2,31	2,19	2,21	2,22	2,03	2,08
Gentamicin	2,80	2,63	2,91	3,03	2,79	2,77	2,45	2,20	2,07	1,84	1,84	1,70	1,48	1,40	1,31	1,36
Kanamycin	2,89	2,79	2,93	2,69	2,32	1,98	2,00	1,94	1,73	1,51	1,17	1,33	1,28	1,28	1,26	1,36
Levofloxacin	2,92	2,71	2,46	2,23	2,18	1,93	2,04	1,98	1,94	1,57	1,39	1,12	0,97	0,99	0,95	0,98
Metronidazole	4,26	4,12	4,38	4,51	4,24	4,39	3,72	3,48	3,57	3,21	3,34	3,17	3,19	2,82	2,62	2,73
Neomycin	3,81	3,73	3,98	3,62	3,48	3,66	3,46	3,50	3,16	2,46	2,31	2,16	1,97	1,86	2,00	1,97
Rifampicin	4,81	4,41	4,45	4,36	3,97	4,21	4,68	4,26	4,11	4,04	3,77	3,15	3,53	2,82	2,63	2,65
Tobramycin	4,58	5,06	4,88	3,97	3,83	3,70	3,11	2,61	2,46	2,19	2,06	1,95	1,63	1,71	1,94	2,00
Average	4,00	3,83	3,80	3,68	3,50	3,49	3,43	3,13	2,90	2,57	2,34	2,20	2,10	1,97	1,88	1,86

VIII. 6 ACKNOWLEDGMENTS

The present work was partially conducted in the Engineering & Health Laboratory, which results of a collaboration protocol established between Universidade Católica Portuguesa and Instituto Politécnico de Lisboa. This study was funded by Fundação para a Ciência e Tecnologia (grant number PTDC/BIO/69242/2006), Agência de Inovação (grant CLARO) and Instituto Politécnico de Lisboa (grants IDI&CA/IPL/2017/DrugsPlatf/ISEL and IDI&CA/IPL/2018/RenalProg/ISEL).

VIII. 7 REFERENCES

1. Boucher, H.W.; Talbot, G.H.; Bradley, J.S.; Edwards, J.E.; Gilbert, D.; Rice, L.B.; Scheld, M.; Spellberg, B.; Bartlett, J. Bad Bugs, No Drugs: No ESKAPE! An Update from the Infectious Diseases Society of America. *Clin. Infect. Dis.* **2009**, *48*, 1–12.
2. Smith, R.A.; M'ikanatha, N.M.; Read, A.F. Antibiotic Resistance: A Primer and Call to Action Rachel. *Heal. Commun.* **2015**, *30*, 309–312.
3. IACG - The Interagency Coordination Group on Antimicrobial Resistance *No time to wait: Securing the future from drug-resistant infections*; 2019;
4. Laxminarayan, R.; Duse, A.; Wattal, C.; Zaidi, A.K.M.; Wertheim, H.F.L.; Sumpradit, N.; Vlieghe, E.; Hara, G.L.; Gould, I.M.; Goossens, H.; et al. Antibiotic resistance-the need for global solutions. *Lancet Infect. Dis.* **2013**, *13*, 1057–1098.
5. Dodds, D.R. Antibiotic resistance: A current epilogue. *Biochem. Pharmacol.* **2017**, *134*, 139–146.
6. Ribeiro da Cunha; Fonseca; Calado Antibiotic Discovery: Where Have We Come from, Where Do We Go? *Antibiotics* **2019**, *8*, 45.
7. Fernandes, P.; Martens, E. Antibiotics in late clinical development. *Biochem. Pharmacol.* **2017**, *133*, 152–163.
8. Lewis, K. Platforms for antibiotic discovery. *Nat. Rev. Drug Discov.* **2013**, *12*, 371–387.
9. Bush, K.; Pucci, M.J. New antimicrobial agents on the horizon. *Biochem. Pharmacol.* **2011**, *82*, 1528–1539.
10. Fields, F.R.; Lee, S.W.; McConnell, M.J. Using bacterial genomes and essential genes for the development of new antibiotics. *Biochem. Pharmacol.* **2017**, *134*, 74–86.
11. Wright, P.M.; Seiple, I.B.; Myers, A.G. The evolving role of chemical synthesis in antibacterial drug discovery. *Angew. Chemie - Int. Ed.* **2014**, *53*, 8840–8869.
12. Theuretzbacher, U. Global antibacterial resistance: The never-ending story. *J. Glob. Antimicrob. Resist.* **2013**, *1*, 63–69.
13. Payne, D.J.; Gwynn, M.N.; Holmes, D.J.; Pompliano, D.L. Drugs for bad bugs: Confronting the challenges of antibacterial discovery. *Nat. Rev. Drug Discov.* **2007**, *6*, 29–40.
14. Tommasi, R.; Brown, D.G.; Walkup, G.K.; Manchester, J.I.; Miller, A.A. ESKAPEing the labyrinth of antibacterial discovery. *Nat. Rev. Drug Discov.* **2015**, *14*, 529–542.
15. Silver, L.L. Appropriate targets for antibacterial drugs. *Cold Spring Harb. Perspect. Med.* **2016**, *6*, 1–7.
16. Zheng, W.; Thorne, N.; McKew, J.C. Phenotypic screens as a renewed approach for drug

discovery. *Drug Discov. Today* **2013**, *18*, 1067–1073.

17. Ohki, Y.; Sakurai, H.; Hoshino, M.; Terashima, H.; Shimizu, H.; Ishikawa, T.; Ogiyama, T.; Muramatsu, Y.; Nakanishi, T.; Miyazaki, S.; et al. Perturbation-Based Proteomic Correlation Profiling as a Target Deconvolution Methodology. *Cell Chem. Biol.* **2019**, *26*, 137–143.
18. Cunningham, M.L.; Kwan, B.P.; Nelson, K.J.; Bensen, D.C.; Shaw, K.J. Distinguishing on-target versus off-target activity in early antibacterial drug discovery using a macromolecular synthesis assay. *J. Biomol. Screen.* **2013**, *18*, 1018–1026.
19. Bantscheff, M.; Drewes, G. Chemoproteomic approaches to drug target identification and drug profiling. *Bioorganic Med. Chem.* **2012**, *20*, 1973–1978.
20. Crumplin, G.C.; Smith, J.T. Nalidixic acid: an antibacterial paradox. *Antimicrob.Agents Chemother.* **1975**, *8*, 251–261.
21. Sato, S. ichi; Murata, A.; Shirakawa, T.; Uesugi, M. Biochemical Target Isolation for Novices: Affinity-Based Strategies. *Chem. Biol.* **2010**, *17*, 616–623.
22. Nishiya, Y.; Hamada, T.; Abe, M.; Takashima, M.; Tsutsumi, K.; Okawa, K. A new efficient method of generating photoaffinity beads for drug target identification. *Bioorganic Med. Chem. Lett.* **2017**, *27*, 834–840.
23. Six, D.A.; Krucker, T.; Leeds, J.A. Advances and challenges in bacterial compound accumulation assays for drug discovery. *Curr. Opin. Chem. Biol.* **2018**, *44*, 9–15.
24. dos Santos, B.S.; da Silva, L.C.N.; da Silva, T.D.; Rodrigues, J.F.S.; Grisotto, M.A.G.; Correia, M.T. do. S.; Napoleão, T.H.; da Silva, M. V.; Paiva, P.M.G. Application of omics technologies for evaluation of antibacterial mechanisms of action of plant-derived products. *Front. Microbiol.* **2016**, *7*, 1–13.
25. MacNair, C.R.; Stokes, J.M.; French, S.; Myers, C.L.; Iyer, K.R.; Brown, E.D. A cell-based approach to characterize antimicrobial compounds through kinetic dose response. *Bioorganic Med. Chem.* **2016**, *24*, 6315–6319.
26. Barker, C.A.; Farha, M.A.; Brown, E.D. Chemical Genomic Approaches to Study Model Microbes. *Chem. Biol.* **2010**, *17*, 624–632.
27. Kitagawa, M.; Ara, T.; Arifuzzaman, M.; Ioka-Nakamichi, T.; Inamoto, E.; Toyonaga, H.; Mori, H. Complete set of ORF clones of Escherichia coli ASKA library (A complete set of E. coli K-12 ORF archive): unique resources for biological research. *DNA Res.* **2005**, *12*, 291–299.
28. Baba, T.; Ara, T.; Hasegawa, M.; Takai, Y.; Okumura, Y.; Baba, M.; Datsenko, K.A.; Tomita, M.; Wanner, B.L.; Mori, H. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Syst. Biol.* **2006**, *2*, 2006.0008.
29. Chang, Y.C.; Hu, Z.; Rachlin, J.; Anton, B.P.; Kasif, S.; Roberts, R.J.; Steffen, M. COMBREX-DB: An experiment centered database of protein function: Knowledge, predictions and knowledge gaps. *Nucleic Acids Res.* **2016**, *44*, D330–D335.
30. Elad, T.; Seo, H. Bin; Belkin, S.; Gu, M.B. High-throughput prescreening of pharmaceuticals using a genome-wide bacterial bioreporter array. *Biosens. Bioelectron.* **2015**, *68*, 699–704.
31. French, S.; Ellis, M.J.; Coutts, B.E.; Brown, E.D. Chemical genomics reveals mechanistic hypotheses for uncharacterized bioactive molecules in bacteria. *Curr. Opin. Microbiol.* **2017**, *39*, 42–47.
32. Briffotaux, J.; Liu, S.; Gicquel, B. Genome-wide transcriptional responses of Mycobacterium to antibiotics. *Front. Microbiol.* **2019**, *10*, 1–14.
33. Ma, W.; Zhang, D.; Li, G.; Liu, J.; He, G.; Zhang, P.; Yang, L.; Zhu, H.; Xu, N.; Liang, S. Antibacterial mechanism of daptomycin antibiotic against Staphylococcus aureus based on a quantitative bacterial proteome analysis. *J. Proteomics* **2017**, *150*, 242–251.
34. Schelli, K.; Zhong, F.; Zhu, J. Comparative metabolomics revealing Staphylococcus aureus metabolic response to different antibiotics. *Microb. Biotechnol.* **2017**, *10*, 1764–1774.

35. Kurita, K.L.; Glassey, E.; Linington, R.G. Integration of high-content screening and untargeted metabolomics for comprehensive functional annotation of natural product libraries. *Proc. Natl. Acad. Sci.* **2015**, *112*, 11999–12004.

36. Marques, V.; Cunha, B.; Couto, A.; Sampaio, P.; Fonseca, L.P.; Aleixo, S.; Calado, C.R.C. Characterization of gastric cells infection by diverse Helicobacter pylori strains through Fourier-transform infrared spectroscopy. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2019**, *210*, 193–202.

37. Ribeiro da Cunha, B.; Ramalhete, L.; Fonseca, L.P.; Calado, C.R.C. Fourier-Transform Mid-Infrared (FT-MIR) Spectroscopy in Biomedicine. In *Essential Techniques for Medical and Life Scientists: A Guide to Contemporary Methods and Current Applications- Part II*; Tutar, Y., Ed.; Bentham Science Publishers, 2020; pp. 1–39 ISBN 9789811464867.

38. Goodacre, R.; Vaidyanathan, S.; Dunn, W.B.; Harrigan, G.G.; Kell, D.B. Metabolomics by numbers: Acquiring and understanding global metabolite data. *Trends Biotechnol.* **2004**, *22*, 245–252.

39. Huleihel, M.; Pavlov, V.; Erukhimovitch, V. The use of FTIR microscopy for the evaluation of anti-bacterial agents activity. *J. Photochem. Photobiol. B Biol.* **2009**, *96*, 17–23.

40. Ribeiro da Cunha, B.; Fonseca, L.P.; Calado, C.R.C. High-throughput bioassay for mechanism of action determination of antibacterial drugs. In Proceedings of the ENBENG 2017 - 5th Portuguese Meeting on Bioengineering, Proceedings; 2017; pp. 1–4.

41. Ribeiro da Cunha, B.; Fonseca, L.P.; Calado, C.R.C. Metabolic fingerprinting with fourier-transform infrared (FTIR) spectroscopy: Towards a high-throughput screening assay for antibiotic discovery and mechanism-of-action elucidation. *Metabolites* **2020**, *10*.

42. Xuan Nguyen, N.T.; Sarter, S.; Hai Nguyen, N.; Daniel, P. Detection of molecular changes induced by antibiotics in Escherichia coli using vibrational spectroscopy. *Spectrochim. Acta - Part A Mol. Biomol. Spectrosc.* **2017**, *183*, 395–401.

43. Zimmermann, B.; Kohler, A. Optimizing savitzky-golay parameters for improving spectral resolution and quantification in infrared spectroscopy. *Appl. Spectrosc.* **2013**, *67*, 892–902.

44. Wang, J.; Liu, H.; Zhao, J.; Gao, H.; Zhou, L.; Liu, Z.; Chen, Y.; Sui, P. Antimicrobial and antioxidant activities of the root bark essential oil of Periploca sepium and its main component 2-hydroxy-4-methoxybenzaldehyde. *Molecules* **2010**, *15*, 5807–5817.

45. Cho, H.; Uehara, T.; Bernhardt, T.G. Beta-lactam antibiotics induce a lethal malfunctioning of the bacterial cell wall synthesis machinery. *Cell* **2014**, *159*, 1300–1311.

46. Blondeau, J.M. Fluoroquinolones: Mechanism of action, classification, and development of resistance. *Surv. Ophthalmol.* **2004**, *49*, 1–6.

47. Löfmark, S.; Edlund, C.; Nord, C.E. Metronidazole Is Still the Drug of Choice for Treatment of Anaerobic Infections. *Clin. Infect. Dis.* **2010**, *50*, S16–S23.

48. Campbell, E.A.; Korzheva, N.; Mustaev, A.; Murakami, K.; Nair, S.; Goldfarb, A.; Darst, S.A. Structural mechanism for rifampicin inhibition of bacterial RNA polymerase. *Cell* **2001**, *104*, 901–912.

49. Becker, B.; Cooper, M.A. Aminoglycoside antibiotics in the 21st century. *ACS Chem. Biol.* **2013**, *8*, 105–115.

50. Volkov, I.L.; Seefeldt, A.C.; Johansson, M. Tracking of single tRNAs for translation kinetics measurements in chloramphenicol treated bacteria. *Methods* **2019**, *162–163*, 23–30.

51. Davis, A.R.; Gohara, D.W.; Yap, M.N.F. Sequence selectivity of macrolide-induced translational attenuation. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111*, 15379–15384.

52. Lasch, P. Spectral pre-processing for biomedical vibrational spectroscopy and microspectroscopic imaging. *Chemom. Intell. Lab. Syst.* **2012**, *117*, 100–114.

53. Afseth, N.K.; Kohler, A. Extended multiplicative signal correction in vibrational spectroscopy, a tutorial. *Chemom. Intell. Lab. Syst.* **2012**, *117*, 92–99.

54. Siqueira, L.F.S.; Lima, K.M.G. MIR-biospectroscopy coupled with chemometrics in cancer studies. *Analyst* **2016**, *141*, 4833–4847.

55. Bellisola, G.; Sorio, C. Infrared spectroscopy and microscopy in cancer research and diagnosis. *Am. J. Cancer Res.* **2012**, *2*, 1–21.

56. Maity, J.P.; Kar, S.; Lin, C.M.; Chen, C.Y.; Chang, Y.F.; Jean, J.S.; Kulp, T.R. Identification and discrimination of bacteria using Fourier transform infrared spectroscopy. *Spectrochim. Acta - Part A Mol. Biomol. Spectrosc.* **2013**, *116*, 478–484.

57. Moen, B.; Janbu, A.O.; Langsrud, S.; Langsrud, Ø.; Hobman, J.L.; Constantinidou, C.; Kohler, A.; Rudi, K. Global responses of Escherichia coli to adverse conditions determined by microarrays and FT-IR spectroscopy. *Can. J. Microbiol.* **2009**, *55*, 714–728.

58. Sales, K.C.; Rosa, F.; Cunha, B.R.; Sampaio, P.N.; Lopes, M.B.; Calado, C.R.C. Metabolic profiling of recombinant Escherichia coli cultivations based on high-throughput FT-MIR spectroscopic analysis. *Biotechnol. Prog.* **2017**, *33*, 285–298.

59. Zhang, Y.; Yang, Y. Cross-validation for selecting a model selection procedure. *J. Econom.* **2015**, *187*, 95–112.

60. Brereton, R.G.; Lloyd, G.R. Partial least squares discriminant analysis: Taking the magic away. *J. Chemom.* **2014**, *28*, 213–225.

61. Dean, A.P.; Martin, M.C.; Sigee, D.C. Resolution of codominant phytoplankton species in a eutrophic lake using synchrotron-based Fourier transform infrared spectroscopy. *Phycologia* **2007**, *46*, 151–159.

62. Dingsdag, S.A.; Hunter, N. Metronidazole: an update on metabolism, structure-cytotoxicity and resistance mechanisms. *J. Antimicrob. Chemother.* **2018**, *73*, 265–279.

63. Clardy, J.; Fischbach, M.A.; Currie, C.R. The natural history of antibiotics. *Curr. Biol.* **2009**, *19*, 437–441.

64. Zampieri, M.; Sekar, K.; Zamboni, N.; Sauer, U. Frontiers of high-throughput metabolomics. *Curr. Opin. Chem. Biol.* **2017**, *36*, 15–23.

65. Hutter, B.; Schaab, C.; Albrecht, S.; Borgmann, M.; Brunner, N.A.; Freiberg, C.; Ziegelbauer, K.; Rock, C.O.; Ivanov, I.; Loferer, H. Prediction of mechanisms of action of antibacterial compounds by gene expression profiling. *Antimicrob. Agents Chemother.* **2004**, *48*, 2838–2844.

66. Halouska, S.; Fenton, R.J.; Barletta, R.G.; Powers, R. Predicting the in vivo mechanism of action for drug leads using NMR metabolomics. *ACS Chem. Biol.* **2012**, *7*, 166–171.

67. Vincent, I.M.; Ehmann, D.E.; Mills, S.D.; Perros, M.; Barrett, M.P. Untargeted Metabolomics To Ascertain Antibiotic Modes of Action. *Antimicrob. Agents Chemother.* **2016**, *60*, 2281–2291.

68. Zampieri, M.; Szappanos, B.; Buchieri, M.V.; Trauner, A.; Piazza, I.; Picotti, P.; Gagneux, S.; Borrell, S.; Gicquel, B.; Lelievre, J.; et al. High-throughput metabolomic analysis predicts mode of action of uncharacterized antimicrobial compounds. *Sci. Transl. Med.* **2018**, *10*, 1–12.

69. Fuhrer, T.; Heer, D.; Begemann, B.; Zamboni, N. High-throughput, accurate mass metabolome profiling of cellular extracts by flow injection-time-of-flight mass spectrometry. *Anal. Chem.* **2011**, *83*, 7074–7080.

70. O'Driscoll, N.H.; Cushnie, T.P.T.; Matthews, K.H.; Lamb, A.J. Colistin causes profound morphological alteration but minimal cytoplasmic membrane perforation in populations of Escherichia coli and Pseudomonas aeruginosa. *Arch. Microbiol.* **2018**, *200*, 793–802.

71. Nonejuie, P.; Burkart, M.; Pogliano, K.; Pogliano, J. Bacterial cytological profiling rapidly identifies the cellular pathways targeted by antibacterial molecules. *Proc. Natl. Acad. Sci.* **2013**, *110*, 16169–16174.

72. Zoffmann, S.; Vercruysse, M.; Benmansour, F.; Maunz, A.; Wolf, L.; Blum Marti, R.; Heckel, T.; Ding, H.; Truong, H.H.; Prummer, M.; et al. Machine learning-powered antibiotics phenotypic drug discovery. *Sci. Rep.* **2019**, *9*, 1–14.

73. Quintelas, C.; Ferreira, E.C.; Lopes, J.A.; Sousa, C. An Overview of the Evolution of Infrared Spectroscopy Applied to Bacterial Typing. *Biotechnol. J.* **2018**, *13*, 1–10.

Chapter IX

Fast identification of off-target liabilities

This chapter is adapted from the manuscript:

Ribeiro da Cunha, B., Aleixo, S. M., Fonseca, L.P. & Calado, C.R.C. Fast identification of off-target liabilities in early antibiotic discovery with Fourier-transform infrared (FTIR) spectroscopy. (*Submitted*).

Author contribution

Bernardo Ribeiro da Cunha conceptualized the study methodology, conducted the investigation and formal analysis, prepared the original draft, reviewed and edited its final version.

Abstract

Structural modifications of known antibiotic scaffolds have kept the upper hand on resistance, but we're on the verge of not having antibiotics for many common infections. Mechanism-based discovery assays reveal novelty, exclude off-target liabilities, and guide lead optimization. For that, we developed a fast and automatable protocol using high-throughput Fourier-transform infrared (FTIR) spectroscopy (FTIRS). Metabolic fingerprints of *Staphylococcus aureus* and *Escherichia coli* exposed to 35 compounds, dissolved in DMSO or water, were acquired. Our data analysis pipeline identified biomarkers of off-target effects, optimized spectral preprocessing, and identified the top performing machine learning algorithms for off-target liabilities and mechanism of action (MOA) identification. Spectral bands with known biochemical associations more often yielded more significant biomarkers of off-target liabilities when bacteria were exposed to compounds dissolved in water than DMSO. Highly discriminative models distinguished compounds with predominant off-target effects from antibiotics with well-defined MOA (AUROC > 0.87, AUPR > 0.79, F1 > 0.81), and from the latter predicted their MOA (AUROC > 0.88, AUPR > 0.70, F1 > 0.70). The compound solvent did not noticeably affect predictive models. FTIRS is fast, simple, inexpensive, automatable, and highly effective at predicting MOA and off-target liabilities. As such, FTIRS mechanism-based screening assays can be applied for hit discovery and to guide lead optimization during the early stages of antibiotic discovery.

Contents

IX. 1 INTRODUCTION

Thirty years from now, antibiotic resistant infections will cause the death of more people than cancer does now, and carry a cumulative cost of inaction in excess of $100 trillion [1]. Structural modifications of major antibiotic classes, mostly discovered with the Waksman platform, filled the pipeline and temporarily outpaced resistance [2], but a discovery void since the late 1980s [3] has put humanity on the verge of not having therapeutic options for many common infections.

The challenge is finding novelty with the characteristics that make an ideal antibiotic [4], namely high antimicrobial activity and very low toxicity [5]. In fact, nearly one-third of drug discovery attrition is due to toxicity [6]. Antibiotics are toxic due to structural homology between prokaryotes and the host, or other undesirable off-target activity [7]. Even candidates with sufficient antimicrobial activity and acceptable cytotoxic profiles might not be suitable for therapeutic use, as *in vitro* activity often does not correlate with *in vivo* efficacy. The latter is better understood by elucidating a compounds mechanism of action (MOA) [8]. Also, shifting to a mechanism-based discovery approach opens the door to the grey chemical matter [9], which could provide good candidates for medicinal chemistry programs [10]. MOA identification can also guide lead optimization by identifying structural modification that induce predominant off-target activity, i.e., toxic compounds [11]. Although MOA identification is not necessary [12], it increases the probability of success of drug discovery programs [13].

Typically, MOA identification involves techniques that are either slow, low-throughput, difficult to scale, costly, labor-intensive, or a combination thereof [10,11,14–21]. Similarly, off-target screening panels are inefficient, and often unpracticable, to monitor safety profiles across successive generations of structural derivatives. Also, these panels are biased towards known mechanisms of toxicity; only probe one off-target interaction; and evaluate the compound itself, not secondary metabolites [22]. As such, the quest for a fast, cheap, and automatable technique that can be used for the systematic identification of MOA and exclusion of toxic compounds continues.

Fourier-transform infrared (FTIR) spectroscopy (FTIRS) is a technique with such characteristics that is suitable for metabolic fingerprinting in general [23], and identifying antibiotic-induced profiles [24–27]**.** In this study, a FTIRS assay was developed to capture the metabolic fingerprints induced by antibiotics and chemical stressors on *Staphylococcus aureus*

and *Escherichia coli*. The effect of the antibiotic solvent was evaluated by exposing each bacterium to the compounds dissolved in water or dimethyl sulfoxide (DMSO). A data analysis pipeline was implemented to identify biomarkers of off-target effects, to optimize spectral preprocessing, and to select the top performing machine learning algorithms for the identification of off-target liabilities and MOA prediction.

IX. 2 Materials and Methods

IX.2.1 Antibiotics, chemical stressors and materials

Amoxicillin, bacitracin, blasticidin, cephradine, chloramphenicol, clarithromycin, enrofloxacin, erythromycin, gentamicin, levofloxacin, metronidazole, rifampicin, rifapentine, sulfamethazine, sulfanilamide and tetracycline were purchased from Sigma-Aldrich (USA), whereas ciprofloxacin was acquired from Bayer healthcare (Germany) and bleomycin from Cayman Chemical (Germany). Isoniazid, kanamycin, sulfamethoxazole and tobramycin were purchased from Fluka (Belgium), and ampicillin plus neomycin from NZYtech (Portugal). Copper chloride, DMSO, ethanol, sodium chloride, sodium dodecyl sulfate, sodium hydroxide and sodium hypochlorite were acquired from Sigma-Aldrich (USA). Hydrogen peroxide was purchased from Alvita (UK), hydrochloric acid from VWR (Portugal), and ethylenediaminetetraacetic acid from Jose M. Vaz Pereira (Portugal). Stock solutions of all compounds were prepared at 6 mM either in water or DMSO (6% v/v) and vigorously stirred prior to usage.

IX.2.2 High-throughput spectra acquisition

Stock solutions of all compounds (Supplementary Table 1) were prepared at 6 mM either in water or DMSO (6% v/v). Five independent cultures of *E. coli* (ATCC 33876) and *S. aureus* (ATCC 6538P) were grown on 90 mL of cation-adjusted Mueller-Hinton broth (VWR, Portugal) at 37ºC, 250rpm until early-exponential phase (~3h). Bacteria were centrifuged at 3,000 RCF for 10 minutes and resuspended in 6 mL of NaCl 0.9% to obtain an ABS_{600} of 0.75 or 1.5 for *E. coli* and *S. aureus*, respectively. After, cells were dispensed on 96-well microtiter flat-bottomed polystyrene plate (60 μL/well), previously prepared with 60 μL of 3x concentrated growth media and 60 μL of stock solution.

After a 2h incubation at 37ºC, 30 μL were transferred to an infrared transparent ZnSe 96-well plate (Bruker, Germany) in triplicate, which corresponded to mechanical replicates.

ZnSe plates were dehydrated for 1h in a vacuum desiccator with abundant silica, and inserted in a HTS-XT module coupled to a Vertex-70 spectrometer (Bruker Optics, Germany). Spectra were acquired in transmission mode and consisted of 40 coadded scans at a 4 cm^{-1} resolution. Raw spectra were exported from OPUS (Bruker, Germany) as data point tables into MATLAB (MathWorks, USA) for subsequent analysis.

IX.2.3 Quality control routine

Mechanical replicates consistently deemed outliers, across different quality control criteria, were excluded from the dataset. For that, the biofilm thickness, signal-to-noise ratio, and water vapor content were used as quality control criteria [28]. Samples scoring above $q_3 + w \times (q_3 - q_1)$ or below $q_1 + w \times (q_3 - q_1)$, where w is the whisker length (approximately $\pm 2.7\sigma$) and q_1 and q_3 are the 25th and 75th percentiles, for these three quality control criteria, were deemed suspected outliers. If suspected outliers also scored above the 99% confidence limit of a principal component analysis Hoteling's T^2, they were considered outliers and removed from the dataset. No more than two mechanical replicates were considered outliers for any given sample. Mechanical replicates that passed the quality control routines were averaged, so a single spectrum for each of the five biological replicates was used in subsequent analysis.

IX.2.4 Statistical analysis for biomarker identification

An automated script was developed from a previously described workflow [29] to identify biomarkers that distinguish, in mean terms, two populations by applying the most suitable hypothesis tests to the corresponding independent groups of observations. We herein describe the broadest scope of the automated script. Parametric tests were preferred since they're more powerful and are more likely to detect true differences. However, if their assumptions are not met, their results are not statistically valid and may lead to inaccurate conclusions, so non-parametric tests should be used [30–35].

The first step of the automated workflow was the Shapiro-Wilk test, adjusted by Royston, to evaluate normality for samples with at least 20 observations. For smaller samples the goodness-of-fit tests are unlikely to detect non-normality [36–39], in which case nonparametric test were also used. Then, if both samples came from populations with normal distribution, a F test evaluated populational equality of variances. If the population variances

were equal, equality of mean population values was evaluated with a two sample T-test; otherwise, a Welsh T-test was applied. If both samples did not come from populations with normal distribution, two situations may have occurred: either both had size at least 30, in which case the parametric Z test was used to assess the equality of the mean population values, given the central limit theorem; otherwise, both had size less than 30 and the equality of the population medians was evaluated with a non-parametric Mann-Whitney-Wilcoxon test. Lastly, if one of the samples came from a population with normal distribution and the other didn't, then either the sample from the non—normal population had less than thirty observations, in which case the non-parametric Mann-Whitney-Wilcoxon test was used, or if the sample from the non-normal population had more than 30 observations, again using the central limit theorem, the parametric Z test was used to evaluate the equality of the mean population values.

Irrespectively of the test used to evaluate the equality of mean or median population values, if the null hypothesis is rejected, in mean terms, then there are significant differences between the groups of observations, and the ratio is considered a biomarker. All tests were applied for a significance level of 5%. The automated workflow was implemented using Matlab.

IX.2.5 Preprocessing optimization and predictive algorithms

Preprocessing aims to reduce the non-discriminatory sample-specific component of the spectra while highlighting inter-replica variability [40]. The most frequently used preprocessing algorithms are the Savitzky-Golay (SG) filter and MSC, either in its extended version (EMSC) or its extended version with replicate correction (EMSCrep) [41]. EMSCrep was applied using the toolbox provided by Afseth and Kohler [42].The performance of these algorithms depends on their parametrization, which requires extensive optimization [43]. For that, a previously described method was used to optimize preprocessing parameters [44].

Machine learning models were created with MATLAB's Classification Learner App using standard parameters, and then developed with stand-alone scripts. In detail, Support Vector Machines (SVM) was used for one-vs-all multiclass classification with either the linear, quadratic, cubic, fine Gaussian, medium Gaussian or coarse Gaussian kernel function applied on standardized data. For fine, medium and coarse Gaussian, the kernel scale was 11, 43, and 170 respectively, while for the other kernel functions this was set to automatic. All SVM models considered a box constrain of level 1. K-Nearest Neighbor (KNN) was applied on standardized data using the Euclidean distance with equal weight, considering 1, 10 and 100

neighbors for the fine, medium and KNN respectively. Both the Cosine and Cubic KNN considered 10 neighbors, while the distance metrics were the Cosine and Minkowski, respectively. Weighted KNN applied squared inverse weights to a Euclidean distance considering 10 neighbors. Additionally, ensemble models of decision trees were determined using both the AdaBoost and RSUBoost method with 30 learners at a rate of 0.1 and 20 splits, and the Bag method, which uses random predictor selections at each split (i.e., random forests) with 30 learners. Subspace ensemble were determined for discriminant analysis and KNN considering a dimension of 14 subspaces and 30 or 1 learner, respectively. Discriminant analysis considered the diagonal covariance matrix.

Accuracy was determined has the ratio between the predicted true positives plus the true negatives over the total number of observations. Precision was calculated as the true positive rate over the true positive rate plus the false positive rate. Recall, or sensitivity, was determined as the true positive rate over the sum of the true positive rate plus the false negative rate. Specificity was calculated as the true negative rate over the sum of true negative plus false positives. The Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves were calculated using MATLAB's perfcurve function.

IX. 3 Results

IX.3.1 Compound concentration, exposure duration, library solvent and other considerations

In addition to expanding the "screenable" chemical space to compounds with lower antimicrobial activity, the transition to mechanism-based antibiotic discovery has other advantages. Firstly, it dismisses the burden of testing a range of concentrations for each compound, since a large spectrum of MOA can be detected with FTIRS by exposing bacteria to a single concentration [44]. This implies faster and simpler workflows, requiring less compound quantity and other consumables. In addition, because the endpoint is no longer growth inhibition typically observed after 16-24h, shorter time-frames can be considered. Typically, metabolic profiles of antibiotic exposure are mostly generic during the initial 30 minutes, and become more specific 60 minutes after exposure, and even more so 90 minutes after exposure [45]. These profiles become specific faster if the antibiotic concentration is considerably higher than the minimum inhibitory concentration (MIC) [46].

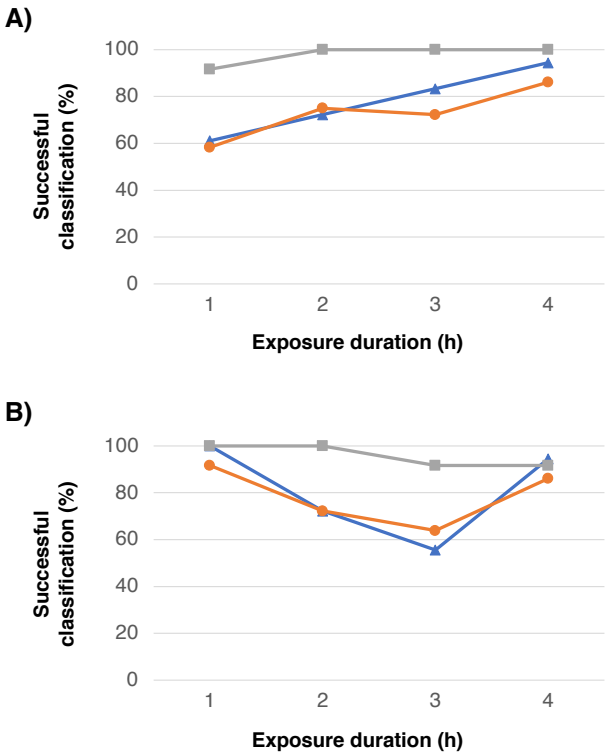


Figure 1. Effect of compound concentration and exposure duration on mechanism of action prediction. Each data point corresponds to the successful classification of a leave-one-out cross-validated partial least squares discriminant analysis of Fourier-transform infrared spectra of *Escherichia coli* (A) and *Staphylococcus aureus* (B), after preprocessing with optimized parameters. Bacteria were exposed to antibiotics at 1,000 µM (grey squares), 100 µM (orange circles) and 10 µM (blue triangles).

Thus, the effect of compound concentration (10 μM, 100 μM and 1 mM) and exposure duration (1, 2, 3, and 4h) on MOA prediction was evaluated (Figure 1). For that, FTIR spectra of *E. coli* and *S. aureus* exposed to amoxicillin and ampicillin, kanamycin and neomycin, as well as sulfamethazine and sulfamethoxazole, were acquired. Spectral preprocessing was optimized for each datapoint (data not shown), and MOA was predicted with a Partial Least Squares discriminant analysis (PLS-DA) after Leave-One-Out Cross-Validation (LOO-CV). To ensure bacteria were exposed to the same number of molecules, molar concentrations were used, which are more pharmacologically representative of their relative potency. In fact, these should arguably become the standard dose system [47].

A 2h exposure at the highest concentration was chosen as this ensured perfect MOA prediction for both strains. FTIR spectra of both bacteria were acquired after exposure to 35 compounds (Supplementary Table 1), 24 of which were antibiotics of 13 classes that act on 6

biosynthetic pathways. Importantly, most major classes used in systemic therapy were included, which are highly desirable by the pharmaceutical industry. Of the remaining compounds, 8 were chemical stressors, as examples of compounds with predominantly off-target effects, and two were controls (the distribution of mechanistic categories of the compounds tested is available on Supplementary Figure 1).

The most commonly used solvent for large compound libraries, DMSO, is known for altering permeability and interfering with reactive oxygen species, which inhibits the rapid killing of some antibiotic classes and increases the potency of others, even at concentrations as low as 1% [48]. Alternatively, we explored water as a solvent. However, many molecules are not water-soluble in their early stages of development, and although medicinal chemistry programs often improve this [49], it could strongly bias the assay. While a single universal solvent would be preferable for assay simplicity and throughput, all experiments were conducted with both solvents. Interestingly, spectral differences were more noticeable between different solvents than between different bacteria (Figure 2).

In a typical FTIRS protocol, samples are washed, to remove the growth media, and resuspended to a predetermined cell concentration, which yields cleaner FTIR spectra with normalized biochemical constituents. To increase throughput, we opted to skip the washing step. For that, bacteria were grown to exponential phase, centrifuged and concentrated to obtain adequate biomass density for spectra acquisition. These were distributed on a 96 well plate previously prepared with the compound library dissolved in both solvents, incubated and transferred to an infrared transparent plate. Because of its simplicity, and given the sample layout was kept the same, our assay is fast and well-suited for automation.

IX.3.2 Biomarkers of off-target effects

Spectral band ratios were built from all combinations of FTIR spectra bands with well-established biochemical assignment [50,51] (Table 1). Prior to calculating spectral ratios, raw spectra were preprocessed with offset correction and EMSCrep. A total of 870 spectral ratios were analyzed. Of these, spectra of *S. aureus* exposed to compounds dissolved in water yielded 639 biomarkers, and 60 biomarkers when compounds were dissolved in DMSO. Similarly, spectra of *E. coli* exposed to compounds dissolved in water yielded 602 biomarkers and 44 when dissolved in DMSO. Despite the drastic effect of the solvent on the number of ratios that are deemed biomarkers, this alone is not very informative. Therefore, for a given

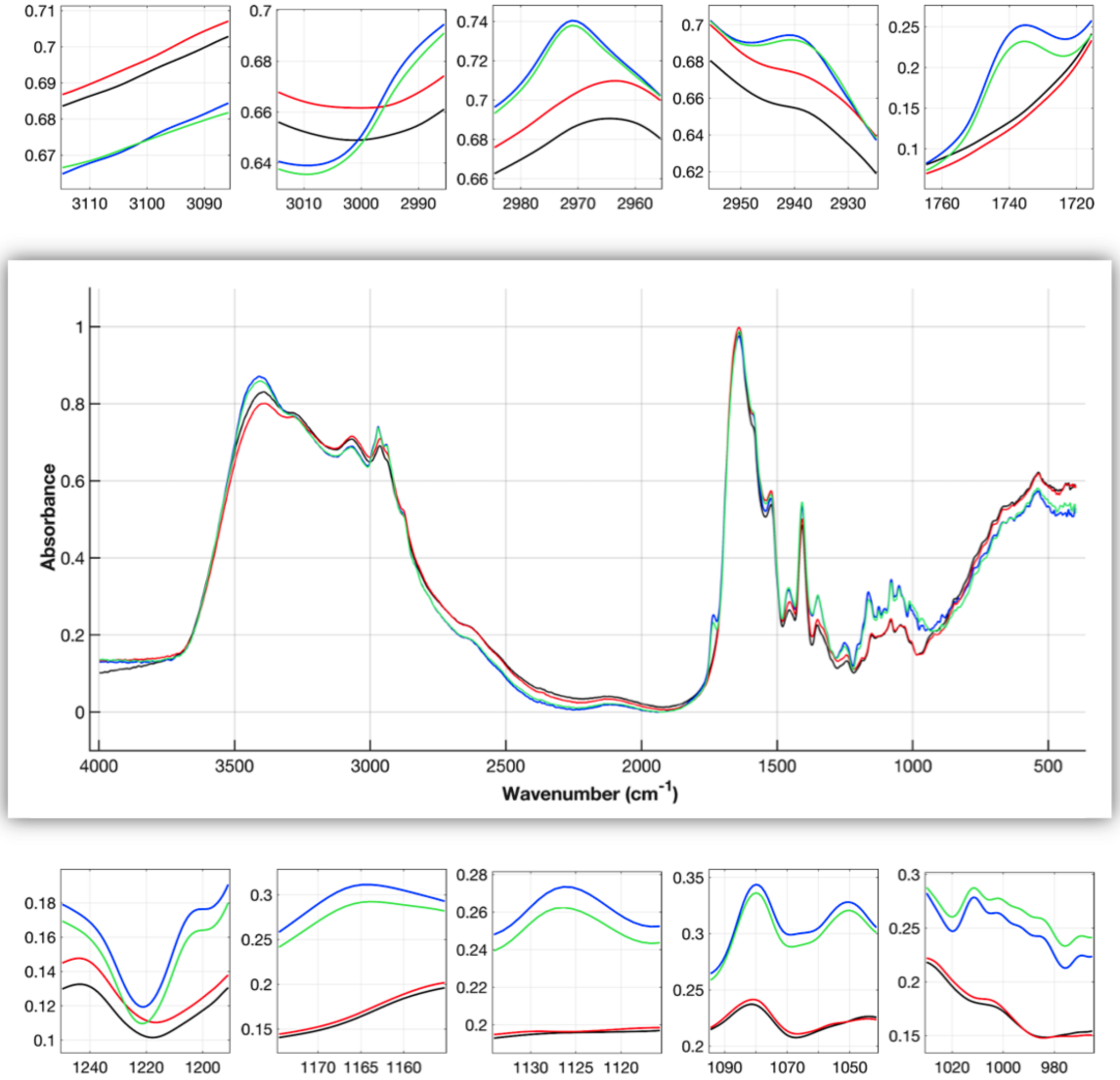


Figure 2. Fourier-transform infrared spectra of *S. aureus* (black and blue) and *E. coli* (red and green) after exposure to enrofloxacin dissolved in water (black and red) or DMSO (blue and green). Spectra are averages of five biological replicates after preprocessing with baseline correction, normalization and extended multiplicative scatter correction with replicate correction.

band present in at least one biomarker, the inverse of the p-value of all biomarkers to which said band contributed was summed. As such, this band importance in biomarker (BIB) score reflects the frequency and weight of a given band across all biomarkers. Bands that more frequently yielded more significant biomarkers presented a higher BIB score. In turn, this allows mapping the spectral origin of the biomarkers for either of the four datasets (Figure 3).

Table 1. Common biochemical assignments of the vibrational modes detected with Fourier-transform infrared spectroscopy.

Wavenumber (cm^{-1})	**Vibrational mode**	**Biochemical assignment**
3500	$\nu(\mathrm{O-H})$	
3300	$\nu(\mathrm{N-H})$	Amide A
3100	$\nu(\mathrm{N-H})$	Amide B
2960	$\nu_{as}(\mathrm{CH_3})$	Lipids
2920	$\nu_{as}(\mathrm{CH_2})$	Lipids
2870	$\nu_{s}(\mathrm{CH_3})$	Lipids
2850	$\nu_{s}(\mathrm{CH_2})$	Lipids
1740	$\nu(\mathrm{C=O})$	Phospholipid esters
1680		Amide I (antiparallel pleated sheets and α-turns)
1655		Amide I (α-helical structures)
1650	$80\%\ \nu(\mathrm{CO});\ 20\%\ \nu(\mathrm{C-N})$	Amide I
1637	-	Amide I (β-pleated -sheet structures)
1550	$60\%\ \gamma(\mathrm{N-H});\ 30\%\ \nu(\mathrm{C-N});$ $10\%\ \nu(\mathrm{C-C})$	Amide II
1515		Tyrosine
1466	$\delta_{as}(\mathrm{CH_3})$	Lipids and proteins
1455	$\delta_{as}(\mathrm{CH_2})$	Lipids and proteins
1380	$\nu(\mathrm{CH_3})$	Phospholipids, fatty acids, triglycerides

Vibrational modes: ν – stretching; γ – wagging, twisting and rocking; δ – bending; and def – deformation. Where as – asymmetrical and s – symmetrical.

Table 1 (Continued). Common biochemical assignments of the vibrational modes detected with Fourier-transform infrared spectroscopy.

Wavenumber (cm^{-1})	Vibrational mode	Biochemical assignment
1240	$\nu_{as}(\mathrm{PO_2^-})$	DNA, RNA, phospholipid, phosphorylated protein
1150	$\nu(\mathrm{CO}), \gamma(\mathrm{COH})$	Carbohydrates
1120	$\nu(\mathrm{C-O})$	RNA ribose
1111	$\nu(\mathrm{C-O})$	RNA ribose
1080	$\nu_s(\mathrm{PO_2^-})$	DNA, RNA, phospholipid, phosphorylated proteins
1050	$\nu(\mathrm{COP})$	Phosphate ester
1030	$def(\mathrm{CHO})$	Carbohydrates
1012	$\nu(\mathrm{CO})$	DNA and RNA ribose
965	$\nu(\mathrm{PO_3^{2-}})$	DNA and RNA
950	$\nu(\mathrm{PO_3^{2-}})$	Phosphorylated proteins
920	$\nu(\mathrm{COP})$	Phosphorylated proteins
841		Fingerprint region
620		Fingerprint region

Vibrational modes: ν – stretching; γ – wagging, twisting and rocking; δ – bending; and def – deformation. Where as – asymmetrical and s – symmetrical.

Greater differences in the BIB score were found depending on the compound library solvent than the bacterial model. When water was the solvent, the three most relevant bands were 1380 cm^{-1}, 1466 cm^{-1}, and 1455 cm^{-1}, which correspond to the stretching of methyl groups habitually found in lipids; to the asymmetric bending of methyl groups commonly found in lipids and proteins; and the asymmetric bending of methylene groups typically found in lipids and proteins, respectively. However, these bands had a neglectable or zero BIB score when DMSO

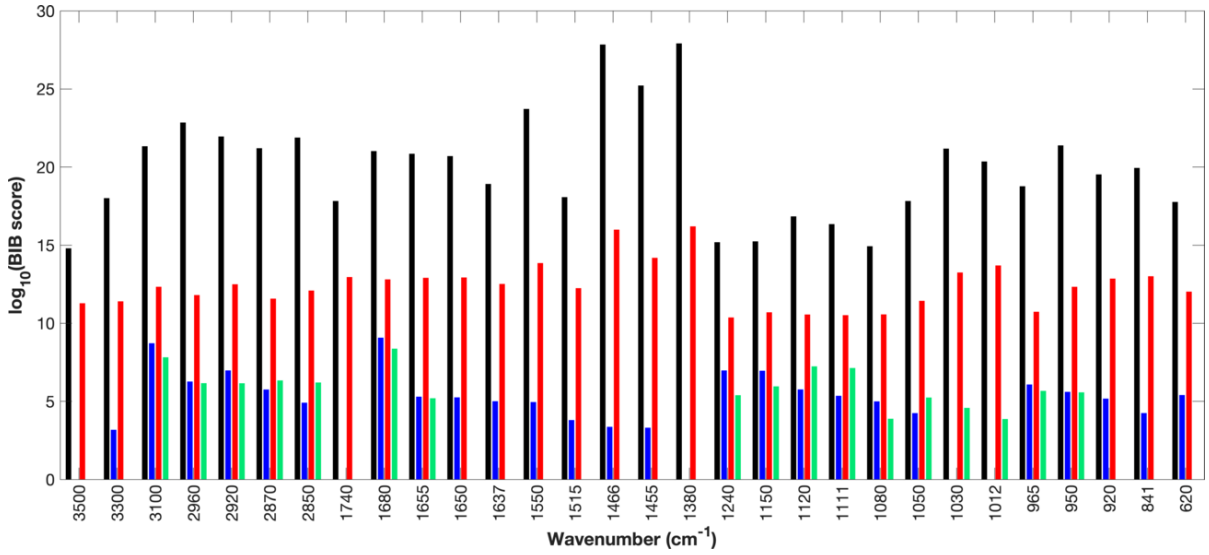


Figure 3. Band importance in biomarker (BIB) scores revealing the contribution of biochemically relevant spectral bands to all identified biomarkers of off-target effects, i.e., spectral ratios with statistically significant differences in mean terms, for Fourier-transform infrared spectra of *S. aureus* (black and blue) and *E. coli* (red and green) exposed to compounds dissolved in water (black and red) or DMSO (blue and green).

was the solvent. In this case, the most significant bands were less coherent across both bacterial models. Here, the three most relevant bands were 1680 cm^{-1}, 3100 cm^{-1}, and 2920 cm^{-1} for *S. aureus* or 1120 cm^{-1} for *E. coli*. These bands are typically associated with antiparallel pleated sheets and α-turns of Amide I; the stretching of the N-H bond of Amide B; the asymmetric stretching of methylene groups associated with lipids or the stretching of the C-O bond typically found in RNA ribose, respectively. Moreover, for the same solvent, *S. aureus* spectral ratios yielded more biomarkers than those of *E. coli*. In theory, an experimental protocol that is sensitive to a wider range of metabolic alterations, which in this case translates to higher BIB scores, should be better suited for the fast exclusion of compounds with off-target effects. However, in a practical sense this proposition is only useful if it can be validated in a predictive scenario.

IX.3.3 Distinguishing on- and off-target effects

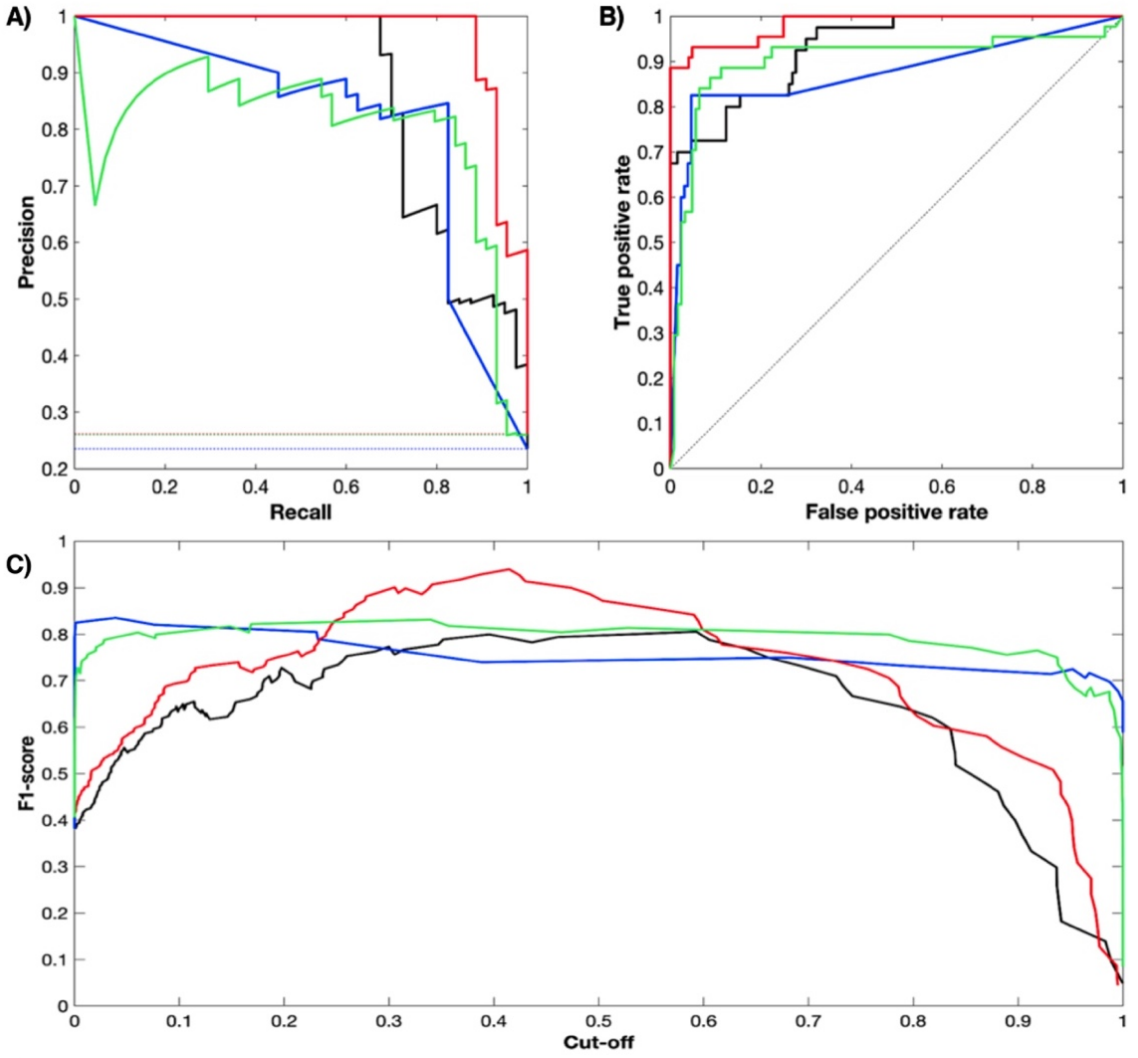
To evaluate the ability of the proposed assay in predicting off-target effects, various machine learning algorithms were applied to spectra preprocessed with optimized parameters. Because EMSCrep combined with second derivative filtering revealed consistently superior to other forms of MSC and derivative orders, other combinations were not considered during

preprocessing optimization. The optimized preprocessing parametrization was identified (Supplementary Figure 2) for each dataset, i.e., spectra of *S. aureus* exposed to compounds dissolved in water, spectra of *S. aureus* exposed to compounds dissolved in DMSO, and likewise for *E. coli*. Moreover, the PLS-DA models built for preprocessing optimization revealed uninformative spectral regions (i.e., VIP score < 1), which were removed to obtain more robust models and further reduce computation time.

Then, various parametrizations of different machine learning algorithms were applied to each dataset, whose accuracy of LOO-CV samples was generally coherent within each dataset (Supplementary Table 2). Minimizing computation time was important as this allowed to increase the number of folds for subsequent cross-validation, which is advantageous since it translates to more training samples that better reflect the underlying sample distribution. While external validation is widely regarded as a more suitable method to estimate the predictive capability of a given model, it is highly dependent on the random data splits, often yielding unstable estimates. In addition, LOO-CV has been shown to perform better than other *k*-fold and external validation algorithms when dealing with small-sample high-dimensional chemometric data [52].

As such, each sample was predicted exactly once using the most accurate algorithm for its dataset, and the posterior probability that it belongs to the positive class, i.e., predominant off-target effects, was used to construct the receiver operator characteristic (ROC) and precision-recall (PR) curves shown on Figure 4 – A and B, from which the F1-score was determined across a range of decision cut-offs (Figure 4 – C). Although ROC curves are typically used to evaluate a classifiers' performance across decision thresholds, PR curves are often more informative when the data is skewed [53] or when there are few positive cases [54]. Similarly, the F1-score was determined since it provides an estimation of accuracy in a way that is less sensitive to class imbalances, and because it penalizes extreme values.

These indicators revealed that, using the same solvent, slightly better predictions were obtained with the *E. coli* datasets than with the *S. aureus* datasets. Moreover, for the *E. coli* dataset, using water as a solvent for the compound library yielded marginally better performance, although the opposite occurred with the *S. aureus* dataset. It is also interesting to note that the F1-scores follow a similar trend as FTIR spectra, i.e., higher similarities were found between different bacteria exposed to the same solvent than between the same bacteria exposed to different solvents. Most importantly, neither the compound library solvent, nor the



D)

Dataset	Window size	Polynomial order	Derivative order	Machine learning algorithm	Area under the ROC curve	Area under the PR curve	Maximum F1-score
S. aureus (H_2O)	137	2	2	Ensemble discriminant	0.9304	0.8672	0.8060
S. aureus (DMSO)	29	2	2	SVM quadratic	0.8737	0.8113	0.8354
E. coli (H_2O)	81	2	2	Ensemble discriminant	0.9822	0.9671	0.9398
E. coli (DMSO)	62	4	2	SVM quadratic	0.9022	0.7948	0.8315

Figure 4. Ability to distinguish antibiotics with well-defined mechanism of action from compounds with predominant off-target effect. The precision-recall curve (A), along the receiver operating characteristic curve (B), and the F1-score across different decision cut-offs (C) were determined after leave-one-out cross-validation. Machine learning was applied to Fourier-transform infrared spectra of *S. aureus* (black and blue) and *E. coli* (red and green) exposed to the compound library dissolved in water (black and red) or DMSO (blue and green) after preprocessing with optimized parameters (D).

bacterial model, had a drastic effect on the ability to distinguish on-and off-target effects, as would be expected considering the observed BIB scores. In the end, the coherent performance indicators suggest that FTIR spectra provide a sufficiently unique phenotypic fingerprint that enables the fast distinction between bacteria exposed to compounds with predominant off-target effects from those exposed to antibiotics with a well-defined MOA.

IX.3.4 GUIDING MEDICINAL CHEMISTRY PROGRAMS: FAST CLASSIFICATION OF ANTIBIOTICS MOA

In addition to distinguishing on- versus off-target activity, the fast identification of off-target liabilities can also be achieved by probing the mechanistic category of antibiotics. For that, the optimized preprocessing parameters, followed by removing uninformative spectral regions, used for distinguishing on- and off-target effects, was used. The most accurate machine learning algorithm, and its parameters, was identified for each dataset after LOO-CV (Supplementary Table 3), which in this case were ensembles of subspace discriminant classifiers across the four datasets.

Because the fast classification of antibiotics MOA is a multiclass prediction problem, a one-versus-all approach was used. As such, ROC and PR curves, as well as F1-scores, were determined for each MOA (Supplementary Figure 3). Figure 5 presents the area under each curve, as well as the maximum F1-score obtained for each MOA category, and for each dataset. Here, it seems that the MOA involving the inhibition of folic acid, mycolic acid and RNA were more challenging to predict. Although this mechanism-specific underperformance was evident in the area under the ROC curve (Figure 4 – A), it was not as notorious in the area under the PR curve and maximum F-score (Figure 5 – B and C). One possible reason for this is that the latter are less sensitive to class imbalances and skewed datasets, and therefore provided a more accurate description of the models' performance, which were overestimated by the ROC curves. Alternatively, since these were the mechanistic categories with fewer representative antibiotics, it could be that this mechanism-specific underperformance was due to class under-representation. However, the fact that there were as many representatives of the control category as those of the RNA category suggests otherwise.

Moreover, it could also be that these MOA are in fact more challenging to predict. For instance, the inhibition of folic acid by sulfonamides induces a thymineless death, whose mechanism could be similar to a stalled replication fork [55], in which case it would resemble the MOA of other DNA synthesis inhibitors such as fluoroquinolones. Similarly, the MOA of

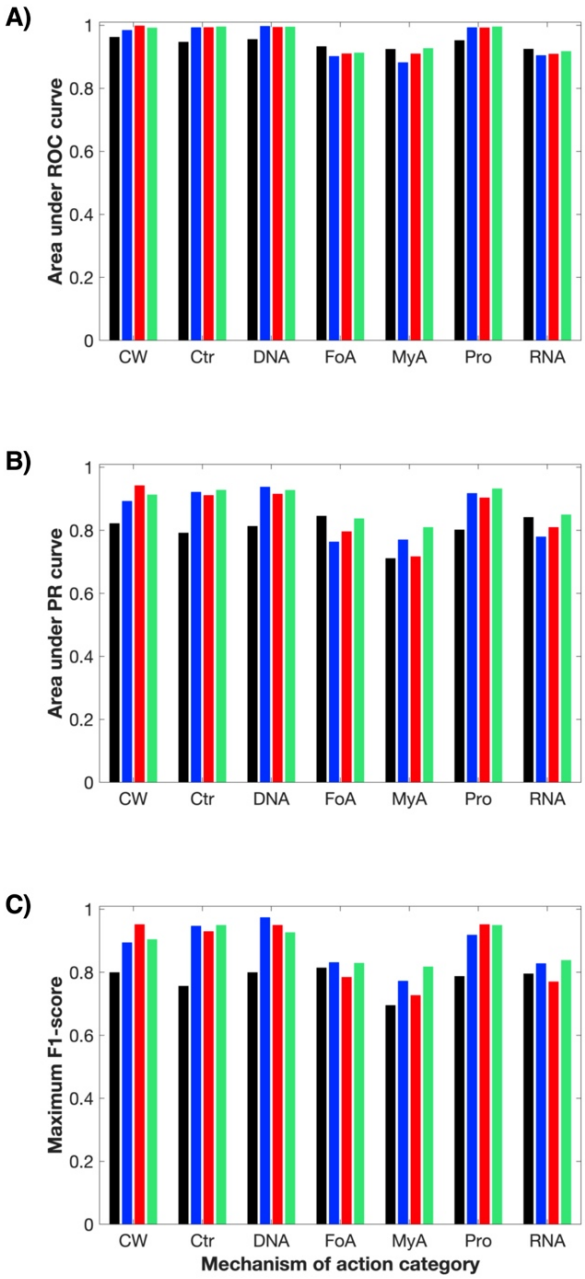


Figure 5. Ability to predict antibiotic mechanistic categories, namely those that target the cell wall (CW), DNA, folic acid (FoA), mycolic acid (MyA), protein (Pro), RNA and control (Ctr). The area under the precision-recall (PR) curve (A), as well as under the receiver operating characteristic (ROC) curve (B), along the maximum F1-score (C) were determined with leave-one-out cross-validation of ensembles of subspace discriminant classifiers applied to Fourier-transform infrared spectra of *S. aureus* (black and blue) and *E. coli* (red and green) exposed to the compound library dissolved in water (black and red) or DMSO (blue and green) after preprocessing with optimized parameters.

isoniazid, which is assumed to be related to the inhibition of mycolic acid synthesis, also impairs the electron transport chain via the accumulation of reactive oxygen species [56]. These have been shown to be correlated with the bactericidal activity of various antibiotics [45], and could therefore yield fingerprints where unspecific response mechanisms overshadow antibiotic-specific events, which could result in poorer MOA prediction.

In general, the models constructed were well suited for the fast classification of antibiotics MOA. Because no single dataset outperformed the remainder, neither the model bacteria nor solvent had a substantial effect on the global predictive performance, as observed for the distinction of on- versus off-target effects. Ultimately, this suggests that phenotypic fingerprints acquired with FTIRS capture the complex biological response when bacteria are exposed to antibiotics of different mechanistic categories and can be used to accurately predict their MOA beyond a given bacterial model or compound library solvent.

IX. 4 Discussion

FTIRS has been shown to detect antibiotic-induced profiles that elucidate their MOA [24–27]. In addition to the prediction of both known and simulated novel MOA, we have shown that FTIRS can be used to estimate potency and to probe the grey chemical matter [44]. In this study we developed and streamlined a protocol for the fast exclusion of off-target liabilities. This involved minimizing the number of steps, and reducing their duration, while ensuring accurate MOA identification. As such, our assay required a short cultivation, exposure, sample dehydration and FTIR spectra acquisition. This means that from frozen bacterial stocks all the way through to exported spectra files took less than 12h, with a hands-on time around 1h. In comparison, a streamlined high-throughput metabolomics protocol took at least 2 full days [57].

Firstly, we developed an automated workflow comprised of a series of hypothesis tests to determine the BIB score of the spectral bands with common biochemical associations. This confirmed that FTIR spectra had intrinsic patterns that reflect the different biological phenomena occurring when bacteria are exposed to antibiotics versus those that occur upon exposure to chemical stressors. Importantly, this was a good starting point prior to machine learning techniques, whose interpretation becomes increasingly complex as algorithms rely on abstract mathematical relationships to extract patterns that are ‘suggested’ by the user, thereby blurring the relationship between data and biological events.

Then, optimally preprocessed FTIR spectra were analyzed with different machine learning classifiers to predict on- and off-target effects. Our data analysis pipeline ensured that the sample-specific spectral component was highlighted for each dataset, and the most accurate machine learning algorithm was applied. As such, robust performance indicators of different nature, i.e., the area under the ROC and PR curves, along the F1-score were achieved for each dataset. The same data analysis pipeline was applied to predict the MOA of the previously identified antibiotics, and notwithstanding a slight performance hit, the resulting models were successful in predicting MOA across the mechanistic categories tested. Although larger compound libraries need to be considered, FTIRS seems to be well suited to rapidly distinguish metabolic profiles induced by chemical stressors from the profiles induced by antibiotics with well-defined MOA, and from the latter predict their mechanistic category.

It is interesting to note that visual inspection of the spectra revealed greater similarities between *E. coli* and *S. aureus* exposed to enrofloxacin dissolved in each of the solvents, than between spectra of either bacteria exposed to the antibiotic dissolved in different solvents. This suggests that the solvent affected the specific antibiotic-induced metabolic alterations. This observation was coherent with the spectral bands that most often, and with greater significance, contributed to biomarkers of off-target effects from which we hypothesized that the experimental setup that reveals a wider range of statistically significant biochemical alterations should be better suited for the fast exclusion of compound with off-target effects. However, neither the compound library solvent, nor the bacterial model, had a drastic effect on the ability to distinguish on-and off-target effects, or to predict the MOA of the antibiotics tested. This is particularly relevant because other techniques suggested for antibiotic discovery are sensitive to the metabolic effect of DMSO, which impairs their application [46]. Moreover, these techniques require additional steps to stabilize the metabolism, which are often are not enough to avoid divergent metabolic profiles across similar samples, which degrades the quality of MOA profiles and thus of subsequent predictive models.

This suggests that FTIRS is a robust technique that could be used for the rapid exclusion of compounds with predominant off-target activity, and simultaneously provides valuable insight into the MOA of compounds, even if only at the level of the major biosynthetic pathway. This level of sensitivity has proven to be sufficient to guide lead optimization [11]; for example, it is equivalent to macromolecular accumulation assays, whose utility is not to pinpoint MOA, but rather as a simple and fast assay for preliminary screening or MOA confirmation. Because our protocol is fast, simple, inexpensive, and highly automatable, it can

be used to monitor MOA profiles and thereby rapidly exclude off-target liabilities, which typically arise during medicinal chemistry efforts, before considerable efforts are put into their development and optimization. As such, FTIRS can fuel a new generation of mechanism-based screening assays that could be an invaluable tool for the early stages of antibiotic discovery.

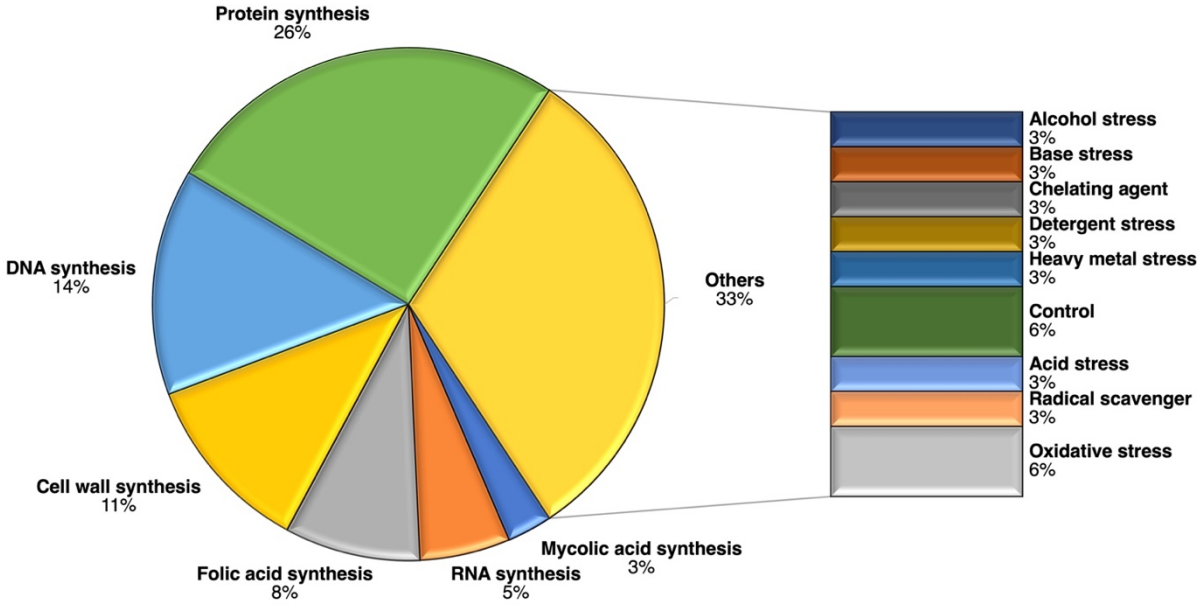
IX. 5 SUPPLEMENTARY DATA

Supplementary Table 1. List of compounds used to stress *Escherichia coli* and *Staphylococcus aureus*. Category reflects the major type of stress or biosynthetic pathway targeted, while class reflects the structural classification of compounds.

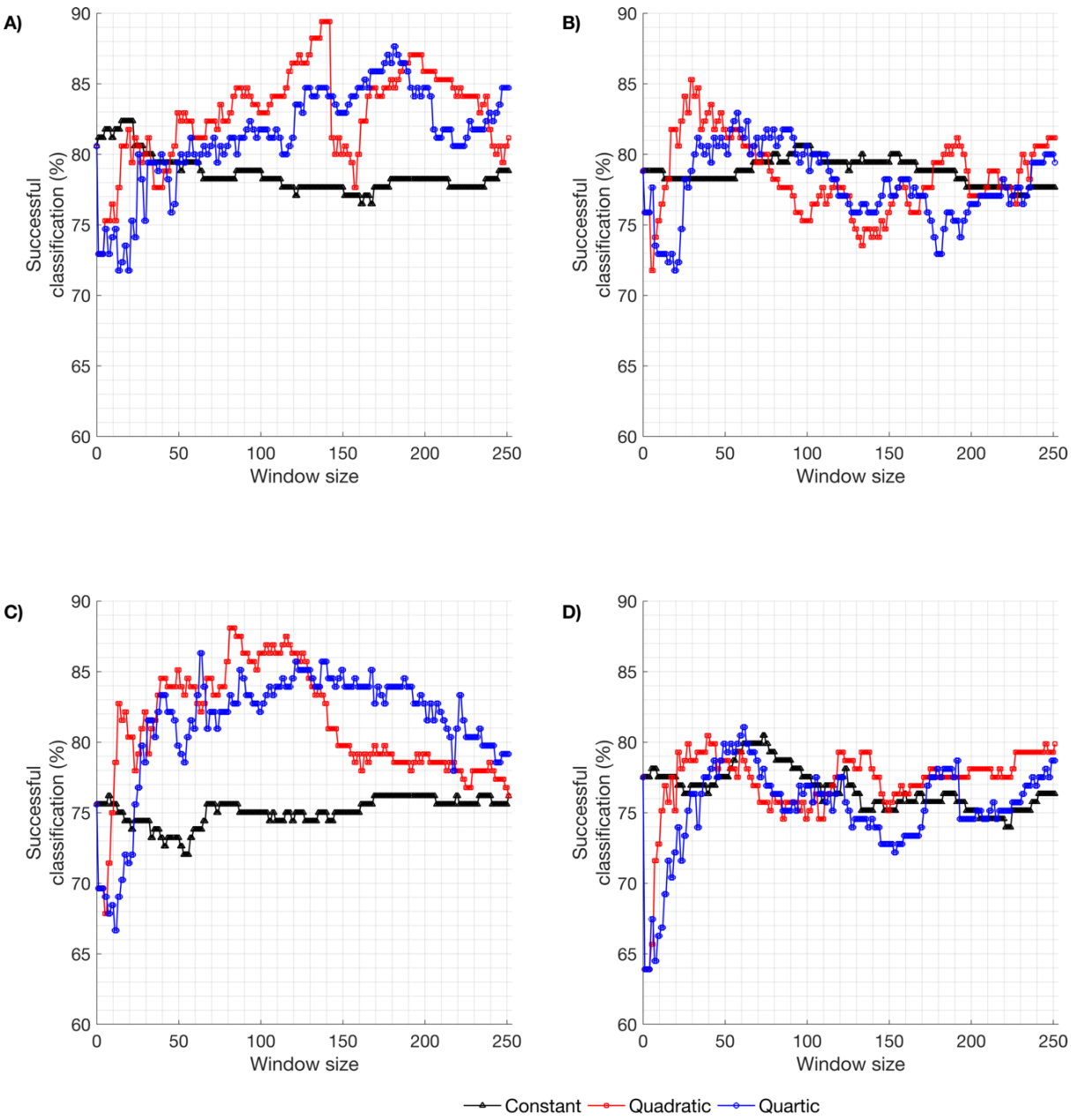
Category	Class	Compound
Acid stress	Other	Hydrochloric acid
Alcohol	Alcohol	Ethanol
Base stress	Other	Sodium hydroxide
Cell wall	AMP	Bacitracin
	Beta-lactams	Amoxicillin
		Ampicillin
	Cephalosporin	Cephradine
Chelating agent	Other	Ethylenediaminetetraacetic acid
Control	-	Water
		Sodium chloride
Detergent stress	Other	Sodium dodecyl sulfate
DNA synthesis	Antineoplastics	Bleomycin
	Fluoroquinolones	Enrofloxacin
		Ciprofloxacin
		Levofloxacin
	Nitroimidazoles	Metronidazole

Supplementary Table 1 (Continued). List of compounds used to stress *Escherichia coli* and *Staphylococcus aureus*. Category reflects the major type of stress or biosynthetic pathway targeted, while class reflects the structural classification of compounds.

Category	Class	Compound
Folic acid biosynthesis	Sulfonamide's	Sulfanilamide
		Sulfamethazine
		Sulfamethoxazole
Heavy metal stress	Other	Copper chloride
Mycolic acid biosynthesis	Pyridine Derivatives	Isoniazid
Oxidative stress	Other	Hydrogen peroxide
		Sodium hypochlorite
Protein synthesis	Aminoacylnucleoside	Blasticidin
	Aminoglycosides	Gentamicin
		Kanamycin
		Neomycin
		Tobramycin
	Amphetamines	Chloramphenicol
	Macrolides	Clarithromycin
		Erythromycin
	Tetracyclines	Tetracycline
Radical scavenger	Other	DMSO
RNA synthesis	Rifamycins	Rifampicin
		Rifapentine



Supplementary Figure 1. Distribution of the mechanistic category of a 35-compound library to which *Escherichia coli* and *Staphylococcus aureus* were exposed to. Categories reflect the major type of stress or biosynthetic pathway targeted by the compounds.



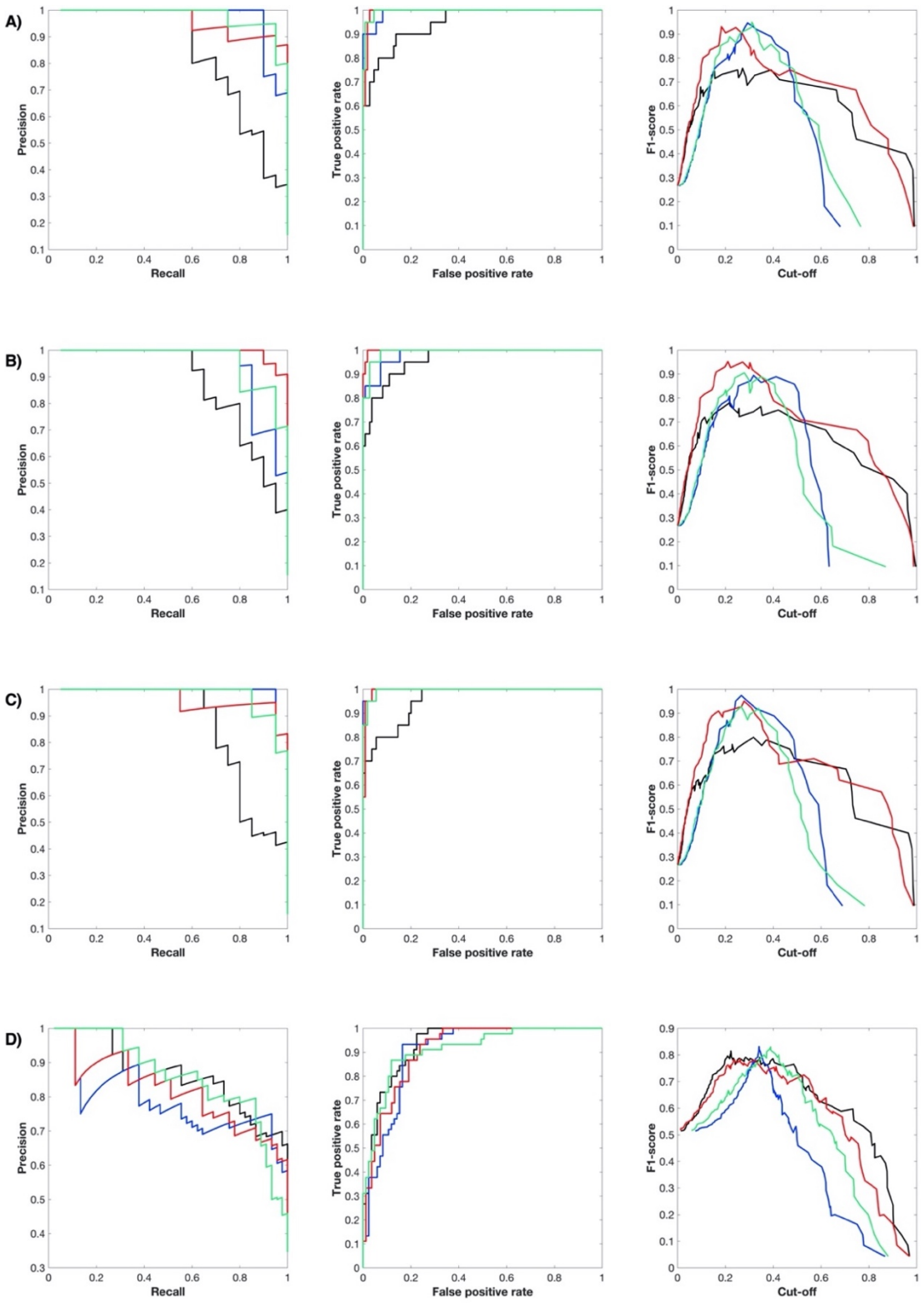
Supplementary Figure 2. Successful classification of on- versus off-target effects with FTIR spectra of *Staphylococcus aureus* (A and B) or *Escherichia coli* (C and D) cells exposed to 35 compounds dissolved in water (A and C) or DMSO (B and D). Successful classification was evaluated with leave-one-out cross-validation of partial least squares discriminant analysis of spectra preprocessed with a second derivative Savitzky-Golay filter followed by extended multiplicative scatter correction with replicate correction.

Supplementary Table 2. Accuracy of machine learning algorithms in distinguishing antibiotics with well-defined MOA from compounds with predominant off-target effects, i.e., toxic compounds. Four datasets were evaluated, which were comprised of two bacterial models and two compound library solvents. Preprocessing was applied with optimized parameters for each dataset. The highest accuracy for each dataset was highlighted in bold to identify the top performing algorithm.

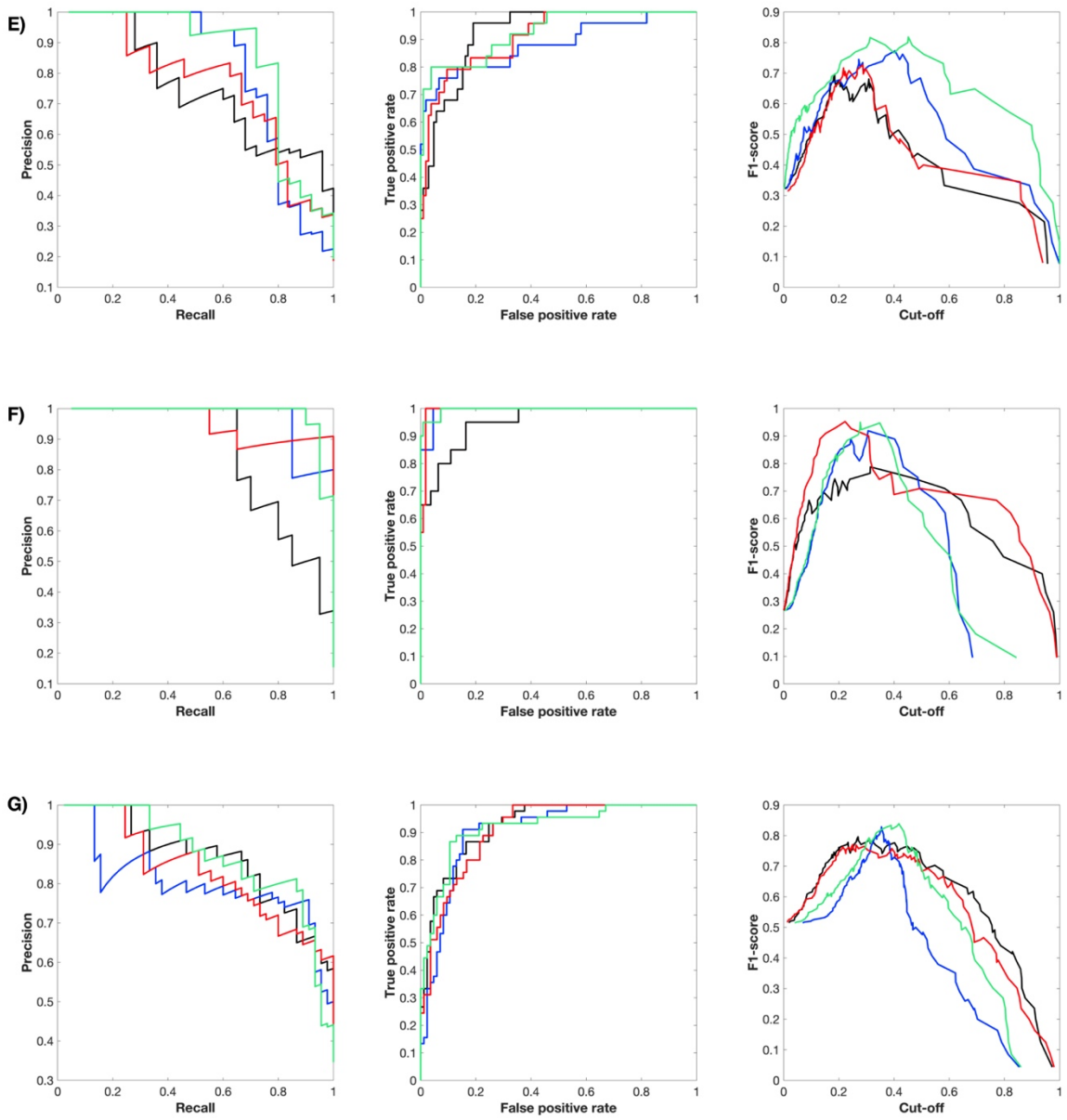
Algorithm	***S. aureus*** **(H_2O)**	***S. aureus*** **(DMSO)**	***E. coli*** **(H_2O)**	***E. coli*** **(DMSO)**
SVM linear	84.71%	84.12%	83.33%	79.88%
SVM quadratic	83.53%	**91.76%**	83.33%	**90.53%**
SVM cubic	82.35%	89.41%	78.57%	89.35%
SVM fine gaussian	76.47%	76.47%	73.81%	73.96%
SVM medium gaussian	78.24%	75.88%	74.40%	78.11%
SVM coarse gaussian	80.59%	81.18%	80.36%	76.33%
KNN fine	85.29%	84.71%	76.19%	79.29%
KNN medium	85.29%	78.82%	82.74%	77.51%
KNN coarse	76.47%	76.47%	73.81%	73.96%
KNN cosine	84.71%	80.59%	86.90%	81.07%
KNN cubic	84.71%	79.41%	83.33%	75.74%
KNN weighted	85.29%	82.35%	80.36%	79.88%
Ensemble boosted trees	75.29%	76.47%	73.21%	73.96%
Ensemble bagged trees	81.76%	85.88%	80.36%	82.84%
Ensemble RSUboosted trees	75.88%	85.88%	74.40%	79.88%
Ensemble discriminant	**85.88%**	83.53%	**87.50%**	84.62%
Ensemble KNN	80.59%	84.71%	75.60%	79.88%
Linear discriminant	69.41%	70.59%	67.26%	63.91%
Quadratic discriminant	72.35%	62.94%	66.67%	53.85%

Supplementary Table 3. Accuracy of machine learning algorithms in classifying the mechanistic category of antibiotics. Four datasets were evaluated, comprised of two bacterial models and two compound library solvents. Preprocessing was applied with optimized parameters for each dataset. The highest accuracy for each dataset was highlighted in bold to identify the top performing algorithm.

Algorithm	***S. aureus*** **(H_2O)**	***S. aureus*** **(DMSO)**	***E. coli*** **(H_2O)**	***E. coli*** **(DMSO)**
SVM linear	64.62%	63.85%	61.24%	63.08%
SVM quadratic	69.23%	75.38%	64.34%	74.62%
SVM cubic	65.38%	63.08%	57.36%	73.85%
SVM fine gaussian	34.62%	34.62%	34.88%	34.62%
SVM medium gaussian	14.62%	20.00%	28.68%	16.92%
SVM coarse gaussian	53.85%	50.00%	53.49%	55.38%
KNN fine	70.77%	73.08%	59.69%	73.85%
KNN medium	74.62%	69.23%	65.12%	73.08%
KNN coarse	34.62%	34.62%	34.88%	34.62%
KNN cosine	70.00%	69.23%	65.89%	66.92%
KNN cubic	70.77%	65.38%	66.67%	73.08%
KNN weighted	73.85%	76.15%	61.24%	74.62%
Ensemble boosted trees	59.23%	21.54%	51.94%	66.15%
Ensemble bagged trees	67.69%	73.85%	63.57%	72.31%
Ensemble RSUboosted trees	52.31%	54.62%	46.51%	53.08%
Ensemble discriminant	**79.23%**	**82.31%**	**75.97%**	**80.00%**
Ensemble KNN	70.00%	73.85%	60.47%	76.15%



Supplementary Figure 3. Detailed performance of machine learning algorithms used to distinguish the effect of a 35-compound library, dissolved in either water (black and red) or DMSO (blue and green), on *Staphylococcus aureus* (black and blue) and *Escherichia coli* (red and green). Performance is presented per compound category, namely controls (A), inhibitors of cell wall synthesis (B), DNA (C), folic acid (D), mycolic acid (E), protein (F), and RNA (G).



Supplementary Figure 3 (Continued). Detailed performance of machine learning algorithms used to distinguish the effect of a 35-compound library, dissolved in either water (black and red) or DMSO (blue and green), on *Staphylococcus aureus* (black and blue) and *Escherichia coli* (red and green). Performance is presented per compound category, namely controls (A), inhibitors of cell wall synthesis (B), DNA (C), folic acid (D), mycolic acid (E), protein (F), and RNA (G).

IX. 6 ACKNOWLEDGMENTS

The authors would like to acknowledge the contribution of Ana Linhares Correia to the data collection process.

This work was supported by Instituto Politécnico de Lisboa (grants IDI&CA/IPL/2017/DrugsPlatf/ISEL, 2018/RenalProg/ISEL and 2020/NephroMD) and was partially conducted in the Engineering & Health Laboratory, established through a collaboration between Universidade Católica Portuguesa and Instituto Politécnico de Lisboa. Sandra M. Aleixo was partly supported by the Fundação para a Ciência e Tecnologia, under Grant UIDB/00006/2020 to CEAUL.

IX. 7 REFERENCES

1. O'Neill, J. Antimicrobial Resistance: Tackling a crisis for the health and wealth of nations The Review on Antimicrobial Resistance Chaired b by J O'Neil, and supported by the Wellcome Trust and the UK Government. **2014**.
2. Lewis, K. Platforms for antibiotic discovery. *Nat. Rev. Drug Discov.* **2013**, *12*, 371–387.
3. Silver, L.L. Challenges of antibacterial discovery. *Clin. Microbiol. Rev.* **2011**, *24*, 71–109.
4. Singh, S.B.; Young, K.; Silver, L.L. What is an "ideal" antibiotic? Discovery challenges and path forward. *Biochem. Pharmacol.* **2017**, *133*, 63–73.
5. Rolain, J.M.; Baquero, F. The refusal of the Society to accept antibiotic toxicity: Missing opportunities for therapy of severe infections. *Clin. Microbiol. Infect.* **2016**, *22*, 423–427.
6. F. Peter Guengerich Mechanisms of Drug Toxicity and Relevance to Pharmaceutical Development. *Drug Metab. Pharmacokinet.* **2011**, *26*, 3–14.
7. Rudmann, D.G. On-target and off-target-based toxicologic effects. *Toxicol. Pathol.* **2013**, *41*, 310–314.
8. Pethe, K.; Sequeira, P.C.; Agarwalla, S.; Rhee, K.; Kuhen, K.; Phong, W.Y.; Patel, V.; Beer, D.; Walker, J.R.; Duraiswamy, J.; et al. A chemical genetic screen in Mycobacterium tuberculosis identifies carbon-source-dependent growth inhibitors devoid of in vivo efficacy. *Nat. Commun.* **2010**, *1*, 1–8.
9. Peach, K.C.; Bray, W.M.; Winslow, D.; Linington, P.F.; Linington, R.G. Mechanism of action-based classification of antibiotics using high-content bacterial image analysis. *Mol. Biosyst.* **2013**, *9*, 1837–1848.
10. Zoffmann, S.; Vercruysse, M.; Benmansour, F.; Maunz, A.; Wolf, L.; Blum Marti, R.; Heckel, T.; Ding, H.; Truong, H.H.; Prummer, M.; et al. Machine learning-powered antibiotics phenotypic drug discovery. *Sci. Rep.* **2019**, *9*, 1–14.
11. Cunningham, M.L.; Kwan, B.P.; Nelson, K.J.; Bensen, D.C.; Shaw, K.J. Distinguishing on-target versus off-target activity in early antibacterial drug discovery using a macromolecular synthesis assay. *J. Biomol. Screen.* **2013**, *18*, 1018–1026.
12. Silver, L.L. Appropriate targets for antibacterial drugs. *Cold Spring Harb. Perspect. Med.* **2016**, *6*, 1–7.
13. Moffat, J.G.; Vincent, F.; Lee, J.A.; Eder, J.; Prunotto, M. Opportunities and challenges in

phenotypic drug discovery: An industry perspective. *Nat. Rev. Drug Discov.* **2017**, *16*, 531–543.

14. Nishiya, Y.; Hamada, T.; Abe, M.; Takashima, M.; Tsutsumi, K.; Okawa, K. A new efficient method of generating photoaffinity beads for drug target identification. *Bioorganic Med. Chem. Lett.* **2017**, *27*, 834–840.
15. Rock, F.L.; Mao, W.; Yaremchuk, A.; Tukalo, M.; Crépin, T.; Zhou, H.; Zhang, Y.-K.; Hernandez, V.; Akama, T.; Baker, S.J.; et al. An antifungal agent inhibits an aminoacyl-tRNA synthetase by trapping tRNA in the editing site. *Science* **2007**, *316*, 1759–1761.
16. Pathania, R.; Zlitni, S.; Barker, C.; Das, R.; Gerritsma, D.A.; Lebert, J.; Awuah, E.; Melacini, G.; Capretta, F.A.; Brown, E.D. Chemical genomics in Escherichia coli identifies an inhibitor of bacterial lipoprotein targeting. *Nat. Chem. Biol.* **2009**, *5*, 849.
17. Phillips, J.W.; Goetz, M.A.; Smith, S.K.; Zink, D.L.; Polishook, J.; Onishi, R.; Salowe, S.; Wiltsie, J.; Allocco, J.; Sigmund, J.; et al. Discovery of kibdelomycin, a potent new class of bacterial type II topoisomerase inhibitor by chemical-genetic profiling in Staphylococcus aureus. *Chem. Biol.* **2011**, *18*, 955–965.
18. Chevereau, G.; Bollenbach, T. Systematic discovery of drug interaction mechanisms. *Mol. Syst. Biol.* **2015**, *11*, 807.
19. Briffotaux, J.; Liu, S.; Gicquel, B. Genome-wide transcriptional responses of Mycobacterium to antibiotics. *Front. Microbiol.* **2019**, *10*, 1–14.
20. Ma, W.; Zhang, D.; Li, G.; Liu, J.; He, G.; Zhang, P.; Yang, L.; Zhu, H.; Xu, N.; Liang, S. Antibacterial mechanism of daptomycin antibiotic against Staphylococcus aureus based on a quantitative bacterial proteome analysis. *J. Proteomics* **2017**, *150*, 242–251.
21. Zampieri, M.; Sekar, K.; Zamboni, N.; Sauer, U. Frontiers of high-throughput metabolomics. *Curr. Opin. Chem. Biol.* **2017**, *36*, 15–23.
22. Brown, D.G.; Smith, G.F.; Wobst, H.J. Promiscuity of in Vitro Secondary Pharmacology Assays and Implications for Lead Optimization Strategies. *J. Med. Chem.* **2020**, *63*, 6251–6275.
23. Goodacre, R.; Vaidyanathan, S.; Dunn, W.B.; Harrigan, G.G.; Kell, D.B. Metabolomics by numbers: Acquiring and understanding global metabolite data. *Trends Biotechnol.* **2004**, *22*, 245–252.
24. Huleihel, M.; Pavlov, V.; Erukhimovitch, V. The use of FTIR microscopy for the evaluation of antibacterial agents activity. *J. Photochem. Photobiol. B Biol.* **2009**, *96*, 17–23.
25. Xuan Nguyen, N.T.; Sarter, S.; Hai Nguyen, N.; Daniel, P. Detection of molecular changes induced by antibiotics in Escherichia coli using vibrational spectroscopy. *Spectrochim. Acta - Part A Mol. Biomol. Spectrosc.* **2017**, *183*, 395–401.
26. Ribeiro da Cunha, B.; Fonseca, L.P.; Calado, C.R.C. A phenotypic screening bioassay for Escherichia coli stress and antibiotic responses based on Fourier-transform infrared (FTIR) spectroscopy and multivariate analysis. *J. Appl. Microbiol.* **2019**, *127*, 1776–1789.
27. Ribeiro da Cunha, B.; Fonseca, L.P.; Calado, C.R.C. Metabolic fingerprinting with fourier-transform infrared (FTIR) spectroscopy: Towards a high-throughput screening assay for antibiotic discovery and mechanism-of-action elucidation. *Metabolites* **2020**, *10*.
28. Lasch, P. Spectral pre-processing for biomedical vibrational spectroscopy and microspectroscopic imaging. *Chemom. Intell. Lab. Syst.* **2012**, *117*, 100–114.
29. Marques, V.; Cunha, B.; Couto, A.; Sampaio, P.; Fonseca, L.P.; Aleixo, S.; Calado, C.R.C. Characterization of gastric cells infection by diverse Helicobacter pylori strains through Fourier-transform infrared spectroscopy. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2019**, *210*, 193–202.
30. Kafadar, K.; Sheskin, D.J. *Handbook of Parametric and Nonparametric Statistical Procedures*; 2006; Vol. 51; ISBN 9781420036268.
31. Chan, Y.H. Biostatistics 102 : Quantitative Data – Parametric. *Significance* **2003**, *44*, 391–396.

32. Lang, T. Twenty statistical errors even you can find in biomedical research articles. *Croat. Med. J.* **2004**, *45*, 361–370.

33. Neideen, T.; Brasel, K. Understanding Statistical Tests. *J. Surg. Educ.* **2007**, *64*, 93–96.

34. Ghasemi, A.; Zahediasl, S. Normality tests for statistical analysis: A guide for non-statisticians. *Int. J. Endocrinol. Metab.* **2012**, *10*, 486–489.

35. Kaur, A.; Kumar, R. Comparative Analysis of Parametric and Non-Parametric Tests. *J. Comput. Math. Sci. | www.compmath-journal.org* **2015**, *6*, 336–342.

36. Royston, P. A Remark on Algorithm as 181: The W-Test for Normality. *J. R. Stat. Soc. Ser. C (Applied Stat.* **1995**, *44*, 547–551.

37. Romão, X.; Delgado, R.; Costa, A. An empirical power comparison of univariate goodness-of-fit tests for normality. *J. Stat. Comput. Simul.* **2010**, *80*, 545–591.

38. Noughabi, H.A.; Arghami, N.R. Monte carlo comparison of seven normality tests. *J. Stat. Comput. Simul.* **2011**, *81*, 965–972.

39. Yap, B.W.; Sim, C.H. Comparisons of various types of normality tests. *J. Stat. Comput. Simul.* **2011**, *81*, 2141–2155.

40. Randolph, T.W. Scale-based normalization of spectral data. *Cancer Biomarkers* **2006**, *2*, 135–144.

41. Martens, H.; Nielsen, J.P.; Engelsen, S.B. Light scattering and light absorbance separated by extended multiplicative signal correction. Application to near-infrared transmission analysis of powder mixtures. *Anal. Chem.* **2003**, *75*, 394–404.

42. Afseth, N.K.; Kohler, A. Extended multiplicative signal correction in vibrational spectroscopy, a tutorial. *Chemom. Intell. Lab. Syst.* **2012**, *117*, 92–99.

43. Zimmermann, B.; Kohler, A. Optimizing savitzky-golay parameters for improving spectral resolution and quantification in infrared spectroscopy. *Appl. Spectrosc.* **2013**, *67*, 892–902.

44. Ribeiro da Cunha, B.; Fonseca, L.P.; Calado, C.R.C. Simultaneous elucidation of antibiotic mechanism of action and potency with high-throughput Fourier-transform infrared (FTIR) spectroscopy and machine learning. *Appl. Microbiol. Biotechnol.* **2021**, 1–18.

45. Belenky, P.; Ye, J.D.; Porter, C.B.M.; Cohen, N.R.; Lobritz, M.A.; Ferrante, T.; Jain, S.; Korry, B.J.; Schwarz, E.G.; Walker, G.C.; et al. Bactericidal Antibiotics Induce Toxic Metabolic Perturbations that Lead to Cellular Damage. *Cell Rep.* **2015**, *13*, 968–980.

46. Hoerr, V.; Duggan, G.E.; Zbytnuik, L.; Poon, K.K.H.; Große, C.; Neugebauer, U.; Methling, K.; Löffler, B.; Vogel, H.J. Characterization and prediction of the mechanism of action of antibiotics through NMR metabolomics. *BMC Microbiol.* **2016**, *16*, 1–14.

47. Chmielewska, A.; Lamparczyk, H. Mass versus molar doses, similarities and differences. *Pharmazie* **2008**, *63*, 843–848.

48. Mi, H.; Wang, D.; Xue, Y.; Zhang, Z.; Niu, J.; Hong, Y.; Drlica, K.; Zhao, X. Dimethyl Sulfoxide Protects Escherichia coli from Rapid Antimicrobial-Mediated Killing. *Antimicrob. Agents Chemother.* **2016**, *60*, 5054–5058.

49. Wu, Y.; Ding, X.; Yang, Y.; Li, Y.; Qi, Y.; Hu, F.; Qin, M.; Liu, Y.; Sun, L.; Zhao, Y. Optimization of biaryloxazolidinone as promising antibacterial agents against antibiotic-susceptible and antibiotic-resistant gram-positive bacteria. *Eur. J. Med. Chem.* **2020**, *185*, 111781.

50. Maquelin, K.; Kirschner, C.; Choo-Smith, L.P.; Van Den Braak, N.; Endtz, H.P.; Naumann, D.; Puppels, G.J. Identification of medically relevant microorganisms by vibrational spectroscopy. *J. Microbiol. Methods* **2002**, *51*, 255–271.

51. Ribeiro da Cunha, B.; Ramalhete, L.; Fonseca, L.P.; Calado, C.R.C. Fourier-Transform Mid-Infrared (FT-MIR) Spectroscopy in Biomedicine. In *Essential Techniques for Medical and Life Scientists: A Guide to Contemporary Methods and Current Applications- Part II*; Tutar, Y., Ed.; Bentham Science Publishers, 2020; pp. 1–39 ISBN 9789811464867.

52. Majumdar, S.; Basak, S.C. Beware of External Validation! - A Comparative Study of Several Validation Techniques used in QSAR Modelling. *Curr. Comput. Aided. Drug Des.* **2018**, *14*, 284–291.

53. Davis, J.; Goadrich, M. The Relationship Between Precision-Recall and ROC Curves. *Proc. 23rd Int. Con- ference Mach. Learn.* **2016**, 233–240.

54. Bleakley, K.; Biau, G.; Vert, J.P. Supervised reconstruction of biological networks with local models. *Bioinformatics* **2007**, *23*, 57–65.

55. Kuong, K.J.; Kuzminov, A. Stalled replication fork repair and misrepair during thymineless death in Escherichia coli. *Genes Cells* **2010**, *15*, 619–634.

56. Zeng, S.; Soetaert, K.; Ravon, F.; Vandeput, M.; Bald, D.; Kauffmann, J.M.; Mathys, V.; Wattiez, R.; Fontaine, V. Isoniazid bactericidal activity involves electron transport chain perturbation. *Antimicrob. Agents Chemother.* **2019**, *63*, 1–17.

57. Zampieri, M.; Szappanos, B.; Buchieri, M.V.; Trauner, A.; Piazza, I.; Picotti, P.; Gagneux, S.; Borrell, S.; Gicquel, B.; Lelievre, J.; et al. High-throughput metabolomic analysis predicts mode of action of uncharacterized antimicrobial compounds. *Sci. Transl. Med.* **2018**, *10*, 1–12.

Chapter X

Concluding remarks & future perspectives

Recent advances in the field of antibiotic discovery, brought by the high-tech approaches introduced during and after the genomics era, have indisputably brought us closer to a new era of antibiotic discovery. New light has been shed on the function of various genes, their products, and their regulatory networks. With that, our understanding of bacterial metabolism and physiology, and the effect that many molecules exert on these, has increased manyfold. Paradoxically, antibiotic discovery has slowed to a halt, and it seems that the more we know the less we can discover. Despite the technical limitations of the high-tech approaches described, the preoccupying state of the antibiotic pipeline cannot be associated with any particular limitation, nor is it due to a lack of effort nor of perseverance.

It seems that a solution for this issue is more likely to stem from a different discovery strategy. In essence, it's not about where to look for antibiotics, but rather how we look for them. Over 90% of clinical antibiotics are the evolutionary outcome of an ongoing tug-of-war between actinomycetes and bacteria. In this struggle for survival, antibiotics have evolved to reach and bind to their target. These, and other natural products have been perfected by nature, and it would be either naïve or condescending to think that we can artificially achieve the same results without tremendous effort. Naturally, designer-drugs, despite all their potential, are extremely laborious and have very low success rates.

Meta-omics studies have shown that we are far from exhausting the complete natural product repository: there are underexplored opportunities, e.g., unexpressed antibiotic gene clusters in known microorganisms, as well as unexplored prospects, e.g., in plants or marine organisms, or even from commensal bacteria. Therefore, it seems logical to continue our pursuit of novel antibiotics from natures' repositories. For instance, a 'smart' synthetic biology approach should allow the usage of heterologous hosts to reach thousands or millions of biosynthetic gene clusters. These could then be screened to identify new compounds that can impair clinically relevant microbes, without having to understand and manipulate the conditions required for the cultivation of 'unculturable' bacteria, or those regarding silenced gene clusters.

Then the question is how to go about a different screening strategy. It is reasonable to pursue a new strategy that capitalizes on our most successful experiences, given our urgent need for new antibiotics. Reviving the cell-based approach was a step forward from the reductionist approach that dominated the genomics-era, where it quickly became evident that considering the biological context was crucial when analyzing single targets. Cell-based screening has historically been the most effective approach for discovering first-in-class

antibiotics, and also allows a ‘brute force’ approach that is more likely to produce results in the short-term. However, the cell-based screening approach typically begins by probing the antimicrobial activity of candidate molecules, so it has been mostly successful at identifying the ‘low-hanging-fruit’.

We argue that a new screening strategy should turn the page on the activity-based paradigm and shift to one based on mechanism. This is not a new idea. In fact, this is a paradigm shift that has been proposed by other researchers. At the core of this paradigm shift lies the concept of grey chemical matter, i.e., compounds capable of inducing some level of phenotypic modulation, but without sufficient potency to induce measurable cell death or inhibit growth. Since it is likely that most compounds with detectable potency have been discovered, or can be identified with the traditional approach, technologies sensitive to phenotypic alterations at sub-inhibitory concentrations are required to probe this ‘not-so-low-hanging fruit’.

In theory, all of the technologies currently in use to probe the mechanism of action (MOA) of antibiotics could be used to probe the grey chemical matter. In practice, the reality is not so linear. Conventional assays are mostly limited by the use of a label, sensitivity, and most importantly, throughput. Novel omics-centered assays hold great potential, but despite offering added-value information on biological events, their reduced throughput and the fact that a single step of the omics cascade is probed with a given technique, limits their application for screening purposes. Additionally, some of these techniques are overly sensitive or noisy, which hinders the interpretation of the data they output. Other techniques, such has bacterial cytological profiling that focus on the end product of the omics cascade, the phenotype, overcome some of these limitations, but because their readouts are not a direct reflection of the biochemical composition of the sample, they’re limited in regard to their biological sensitivity and often require complementary information.

Throughout this Ph.D., the case has been made that Fourier-transform infrared spectroscopy (FTIRS) is an excellent tool that can fill this particular technological gap. FTIRS is a mature technique that can provide highly sensitive biochemical profiles and is amendable to high throughput. Although FTIRS had been explored for antibiotic-related studies, this Ph.D. was the first deep-dive into the key components that are required of a mechanism-based discovery platform. In particular, we have shown that FTIRS can acquire metabolic fingerprints that reflect both stress and antibiotic responses. We then shown that these metabolic fingerprints can be used as profiles of MOA, and that said profiles are sufficiently distinct to

elucidate the MOA of antibiotics at greater detail than the level of the major biosynthetic pathway targeted, given antibiotics of the same class had similar MOA profiles.

We further explored how our FTIRS mechanism-based protocol performed when predicting antibiotics MOA in two scenarios: firstly, when a known compound was being tested, here the predictive models had already 'seen' the metabolic fingerprint induced by said compound; and secondly, when a compound with a simulated novel MOA, either similar or considerably different from known MOA, was blind tested. The exceptional predictive performance of known MOAs, alongside the mechanistically coherent prediction of simulated novel MOA, validated the suitability of FTIRS for MOA identification, which was only slightly limited when distinguishing MOA profiles of compounds that belonged the same class, for which other techniques are better tailored than FTIRS.

Beyond MOA identification, a mechanism-based screening platform should effectively expand the 'screenable' chemical space. Thus, we established the dose-response of our FTIRS bioassay, which served as proxy for the grey chemical matter. While higher antibiotic concentrations improved MOA identification, as expected, it also became apparent that even very low antibiotic concentrations, which only induced a slight growth inhibition, also yielded MOA profiles with sufficient quality for their prediction at the biosynthetic pathway and class level. Moreover, our FTIRS bioassay revealed the degree to which any given MOA affected bacterial metabolism and inhibited bacterial growth, which opened the door to a single step assay capable of simultaneously determining the MOA of antibiotic candidates and their potency. However, there is another critical advantage to transitioning to a mechanism-based screening approach, namely the rapid exclusion of compounds with predominant off-target activity, which tend to be toxic.

The work pursued throughout this ambitious venture had the ultimate goal of finding an innovative solution to a serious world-wide issue affecting society. For that innovation to come to life, it was important to first validate it within the scientific community, and then attempt to bring it to market. As such, participating in the innovation accelerator Lab2Market@Tecnico2020, as team 'ASPIR', was an amazing opportunity to further develop the work not only from a business perspective, but also from an out-of-academia scientific standpoint. In fact, many of the interviews conducted with industry experts throughout this accelerator shaped what this Ph.D. covered, and its future directions. For instance, one key issue that we are yet to tackle is identifying the MOA of different compounds in a mixture, which

would be of tremendous value in the context of natural product discovery. Another impending task is to network our platform into collaborations with research groups, either in the academia or industry, that can supply compound libraries from which we hope to find novel antibiotics. Another area for future action is SpecA. Throughout this scientific endeavor, we went to great efforts to develop a comprehensive and strict data analysis pipeline, which resulted in the creation of SpecA. So far, SpecA has not yet been distributed but it was originally created with that intent, which we still wish to fulfill.

In the end, this Ph.D. thesis established that FTIRS can identify the MOA of compounds; can discover hits from the grey chemical matter, effectively expanding the 'screenable' chemical space; ensures that potent compounds are identified, if not by their mechanism then by their inhibitory effect; and rapidly excludes compounds with predominant off-target effects, either during hit discovery or lead optimization. In its latest iteration, our protocol was made considerably faster and more scalable, which further consolidated its applicability to screening large compound libraries in an industrial setting. As such, we are confident that FTIRS can fuel a new generation of mechanism-based screening assays, which can be both more efficient and efficacious, and may swing the war on infectious diseases back in our favor.