Development of new computational methods for fragment-based drug discovery by NMR

Thesis submitted for the degree of
Doctor of Philosophy
at the University of Leicester

by

Luca Gavino Mureddu

Department of Molecular and Cell Biology
University of Leicester

2020

Supervisors:

Prof G. W. Vuister

Prof M. D. Carr

Development of new computational methods for fragment-based drug discovery by NMR

Luca Gavino Mureddu

My project covers the three fundamental steps of fragment-based drug discovery by NMR, (NMR-FBDD): hit identification, binding site identification and hit optimisation. This division enabled me to cover the full NMR-FBDD approach while simultaneously to experience a broad spectrum of techniques as well as to foster collaborations with different labs and levels of expertise.

I first focused on the development of a new software package, called CcpNmr AnalysisScreen, designed for the automated analysis of early stage 1D NMR screening data. AnalysisScreen integrates all the necessary algorithms and routines for supporting the analysis of these data. Furthermore, it provides a novel modular platform for combining tasks in bespoke workflows, and it includes a straightforward mechanism for adding new custom algorithms to the main program. Using a series of simulated spectral datasets with known answers, the performance of the software was assessed. Following this proof of principle, routines were tested using actual experimental data recorded by several collaborators. These analyses prompted the development of novel routines for scoring the outcomes to classify results accurately.

Successively, I focused on the integration of tools for analysing and identifying ligand binding sites. These tools were tested using experimental data recorded by collaborators and successfully used in several postgraduate teaching classes, including national and international workshops.

Lastly, I focused on the optimisation of ligand-binding properties of hits obtained from the initial virtual screening steps. During my internship at the Dana Farber Cancer Institute (Harvard Medical School, Boston, USA), I was involved in a study of the protein-protein complex eIF4G-Mnk. Starting from previous works, which identified transient binding pockets, I conducted a series of studies using a combination of computational tools to design a novel virtual FBDD workflow. This procedure prompted to the generation of potential drug candidates for the inhibition of the eIF4G-Mnk complex.

Aims

1. Assessment and improvement of computational strategies for FBDD by NMR:
    1.1. Critical review of clinical drug candidates and FDA-approved drugs; evaluation of the impact of NMR throughout the discovery and chemical design of case studies, (chapter 1).
    1.2. Development of a new integrative software, CcpNmr AnalysisScreen, for aiding the data analysis through the process of drug discovery by NMR, (chapter 2).
    1.3. Development of a novel customisable tool designed to create bespoke workflows needed in the process of hit identification. Design of novel scoring functions for classifying hits, (chapter 3).
    1.4. Assessment of current algorithms for baseline correction. Development of a new versatile algorithm for NMR screening datasets, (chapter 4).
    1.5. Applications of AnalysisScreen. Comparison of automated and quantitative versus manual and qualitative hit identification results on multiple datasets, (chapter 5).
    1.6. Integration of tools for semi-automatic target-ligands binding site identification in CcpNmr, (chapter 6).
2. Application of virtual FBDD strategies for generating new potential therapeutic molecules for a biological target associated with diseases;
    2.1. Design and application of a novel workflow aimed to disrupt the protein-protein interaction complex eIF4G-Mnk, (chapter 7).

Acknowledgements

Thanks to everyone who has supported me along the way...

List of publications

1. **Mureddu, L.**, Ragan, T.J., Brooksbank, E.J. and Vuister, G.W. CcpNmr AnalysisScreen, a new software programme with dedicated automated analysis tools for fragment-based drug discovery. Journal of Biomolecular NMR, (2020). Reproduced in:
    a. **chapter 1**:

        Figures: 1.1

    b. **chapter 2**:

        Sections: 2.1, 2.3.8, 2.4.1, 2.4.2, 2.4.3

        Figures: 2.2-2.4, 2.6-2.7, 2.11A-B, 2.12B

    c. **chapter 3**:

        Sections: 3.3, 3.4.1, 3.5.1, 3.5.3

        Figures: 3.1-3.7, 3.9-3.10, 3.16-3.18

    d. **chapter 5**:

        Sections: 5.4.3

        Figures: 4.9A-B, 5.10-5.11

2. **Mureddu, L.** & Vuister, G. W. Simple high-resolution NMR spectroscopy as a tool in molecular biology. *FEBS J.* **286**, 2035–2042 (2019). Reproduced in:
    a. **chapter 6**:

        Sections: 6.1-6.4

        Figures: 6.1-6.2

Table of Contents

Chapter 1 Accounts of fragments to drugs by NMR. Where are the successes and where can it be improved? 18

1.1. Abstract 19

1.2. Introduction 20

1.3. Materials and Methods 24

1.3.1. Materials 24

1.3.2. Methods 24

1.4. Case study 1: AZD-3839 and BACE-1 26

1.5. Case study 2: ABT-199 (Venetoclax) and BCL-2/xL 29

1.6. Case study 3: S64315/MIK665 and MCL-1 33

1.7. Discussion and Conclusions 37

1.8. Figures 42

Chapter 2 Design of CcpNmr AnalysisScreen 52

2.1. Abstract 53

2.2. Introduction 54

2.3. Materials and Methods 58

2.3.1. Materials 58

2.3.2. Methods 59

2.4. Results and Discussion 64

2.4.1. Parsing and importing NMR data and metadata 64

2.4.2. Assessment of spectral quality by PCA decomposition 65

2.4.3. Generation of spectral mixtures with minimal overlaps 67

2.5. Conclusions 71

2.6. Figures 74

2.7. Acknowledgments 89

2.8. Addresses 89

Chapter 3 Pipelines on simulated datasets 90

3.1. Abstract 91

3.2. Introduction 92

3.2.1. Ligand-observed techniques 93

3.3. Materials and Methods 97

3.3.1. Materials 97

3.3.2. Methods 98

3.4. Results and Discussion 103

3.4.1. Pipeline design 103

3.4.2. Testing pipelines 106

3.4.3. Determining S/N ratio for STD hit identification 107

3.5. Conclusions 109

3.6. Figures 111

3.7. Acknowledgments 131

3.8. Addresses 131

Chapter 4 A versatile baseline correction for 1D NMR screening 132

4.1. Abstract 133

4.2. Introduction 134

4.3. Materials and Methods 136

4.3.1. Materials 136

4.3.2. Methods 136

4.3.3. New baseline correction algorithm steps 139

4.4. Results and Discussion 141

4.4.1. Qualitative analysis of the various methods 141

4.4.2. Quantitative analysis of the various methods 144

4.5. Conclusions 149

4.6. Figures 151

4.7. Acknowledgments 167

4.8. Addresses 167

Chapter 5 Practical application of AnalysisScreen ________________ 168

5.1. Abstract ________ 169

5.2. Introduction ________ 170

5.3. Materials and Methods ________ 172

5.3.1. Materials ________ 172

5.3.2. Methods ________ 173

5.4. Results and Discussion ________ 177

5.4.1. Dataset-1 ________ 177

5.4.2. Dataset-2 ________ 179

5.4.3. Dataset-3 ________ 181

5.4.4. Dataset-4 ________ 182

5.5. Conclusions ________ 185

5.6. Figures ________ 188

5.7. Acknowledgments ________ 204

5.8. Addresses ________ 204

Chapter 6 Binding Sites Identification Tools in AnalysisScreen: The Chemical Shift Mapping Module ________ 205

6.1. Abstract ________ 206

6.2. Introduction ________ 207

6.3. Chemical shift and exchange ________ 208

6.3.1. Chemical shift perturbation in practice ________ 211

6.4. Conclusions ________ 217

6.5. Figures ________ 218

6.6. Appendix ________ 221

6.6.1. Case study: Minimal shift analysis ________ 221

6.7. Appendix Figures ________ 222

6.8. Acknowledgements ________ 224

6.9. Addresses ________ 224

Chapter 7 Generating high-affinity leads from fragments using NMR-guided dockings and virtual click-chemistry. Application to the eukaryotic Initiation Translation Factor eIF4G 225

7.1. Abstract 226

7.2. Introduction 227

7.3. Materials and Methods 230

7.3.1. Materials 230

7.3.2. Methods 231

7.4. Results and Discussion 239

7.4.1. 4G-Mnk interactions 239

7.4.2. Dockings 239

7.4.3. Molecular Dynamic simulations 244

7.5. Conclusions 246

7.6. Figures 249

7.7. Appendix 264

7.8. Acknowledgments 272

7.9. Addresses 272

Chapter 8 Final Remarks 273

8.1. What have I learnt about FBDD by NMR and how CcpNmr AnalysisScreen can aid future discoveries? 274

8.2. How can the development of new drugs from fragments can be shortened? 276

Chapter 9 Bibliography 278

List of Tables

Table 3.1 Currently available pipes grouped by type and functionalities. ___ 105

Table 4.1 Baseline correction algorithms parameters. ___ 137

Table 6.1 NMR software packages. ___ 216

Table 7.1 Relaxation steps summary.___ 264

Table 7.2 Relaxation steps configuration parameters. ___ 264

Table 7.3 MD simulation configuration parameters. ___ 265

List of Figures

Figure 1.1 NMR-aided fragment-based drug discovery (FBDD). ___________ 42

Figure 1.2 Usage of FBDD methods in the development of new molecules. _ 44

Figure 1.3 Role of NMR-FBDD methods in drug discovery. _______________ 45

Figure 1.4 Optimisation pathway: from NMR hits to AZD-3839. ____________ 46

Figure 1.5 Molecular similarities and interactions for AZD-3839 precursors. _ 47

Figure 1.6 Optimisation pathway: from NMR hits to ABT-199. _____________ 48

Figure 1.7 Molecular similarities and interactions for AZD-3839 precursors. _ 49

Figure 1.8 Optimisation pathway: from NMR hits to S64315. ______________ 50

Figure 1.9 Molecular similarities and interactions for S64315 precursors. ___ 51

Figure 1.10 Molecular similarities comparison for three case-studies. _______ 51

Figure 2.1 CcpNmr Analysis Version 3.0. ______________________________ 74

Figure 2.2 Excel file loading examples. ________________________________ 75

Figure 2.3 CcpNmr AnalysisScreen sidebar and various pop-ups. _________ 76

Figure 2.4 CcpNmr core objects used in AnalysisScreen. ________________ 77

Figure 2.5 PCA on a testing BMRB dataset. ____________________________ 78

Figure 2.6 Principal component analysis (PCA) of 1760 reference spectra. _ 79

Figure 2.7 PCA Q-Scores and T2-Scores for experimental dataset. ________ 80

Figure 2.8 Decomposition module. ____________________________________ 81

Figure 2.9 Mixtures generation on simulated datasets. ___________________ 82

Figure 2.10 Mixtures analysis on simulated datasets. ____________________ 83

Figure 2.11 Mixtures analysis on experimental datasets. _________________ 84

Figure 2.12 Mixtures performance summary on experimental datasets. ____ 85

Figure 2.13 Experimental mixtures examples. __________________________ 86

Figure 2.14 Mixtures generation setup and graphical tools. _______________ 87

Figure 2.15 Mixtures analysis graphical tools. __________________________ 88

Figure 3.1 Common NMR methods for detecting ligand binding. __________ 111

Figure 3.2 CcpNmr AnalysisScreen tools for manual visual inspection. ____ 113

Figure 3.3 CcpNmr AnalysisScreen Pipeline and Hit Analysis module. 114

Figure 3.4 Pipeline architecture and code examples. 115

Figure 3.5 ^{1}H line-broadening detection pipeline. 117

Figure 3.6 Hit analysis results for component 1. 118

Figure 3.7 Hit analysis results for component-3. 119

Figure 3.8 WaterLOGSY hit detection pipeline. 120

Figure 3.9 Hit analysis results for component-1 by WaterLOGSY. 121

Figure 3.10 Hit analysis results for component-3 by WaterLOGSY. 122

Figure 3.11 STD creation pipeline. 123

Figure 3.12 STD hit detection pipeline. 124

Figure 3.13 Hit analysis results for component-1 by STD. 125

Figure 3.14 Hit analysis results for component-3 by STD. 126

Figure 3.15 Hit analysis module on large dataset. 127

Figure 3.16 Peak and hit detection assessment using simulated spectra. 128

Figure 3.17 Peak detection statistics. 129

Figure 4.1 Experiment types spectra for the component-314. 151

Figure 4.2 CWBC algorithm workflow overview. 151

Figure 4.3 CWBC algorithm: step 1. 152

Figure 4.4 CWBC algorithm: steps 2-4. 153

Figure 4.5 CWBC algorithm: steps 5-6. 154

Figure 4.6 CWBC algorithm: steps 7-8. 155

Figure 4.7 Polynomial orders corrections for Pr-314. 156

Figure 4.8 Whittaker smoother corrections for Pr-314. 157

Figure 4.9 ALS corrections for Pr-314. 158

Figure 4.10 airPLS corrections for Pr-314. 159

Figure 4.11 arPLS corrections for Pr-314. 160

Figure 4.12 BDC corrections for Pr-314. 161

Figure 4.13 New baseline corrections (CWBC) for Pr-314. 162

Figure 4.14 Qualitative analysis for Pr-314 $T_{1\rho}$ after applying the CWBC. 163

Figure 4.15 Analytic comparison of methods applied to Pr-314 spectra. 164

Figure 4.16 Analytic comparison of methods applied to the whole dataset. 165
Figure 4.17 Qualitative analysis for 467-Ref ^{1}H. 166
Figure 5.1 Dataset-1, multiple experiments for component-314. 188
Figure 5.2 Dataset-1, multiple experiments for component-467. 189
Figure 5.3 Dataset-1, multiple experiments for component-7373. 190
Figure 5.4 Dataset-1, scorings summary. 191
Figure 5.5 Dataset-2. Singleton scorings summary. 192
Figure 5.6 Dataset-2. Mixtures scorings summary. 192
Figure 5.7 Dataset-2. SpectrumHits statistics. 193
Figure 5.8 Dataset-2. Peak detection statistic scores. 194
Figure 5.9 Dataset-3. Automated *versus* manual SpectrumHit detection. 195
Figure 5.10 Dataset-3. Re-referencing of spectra datasets. 196
Figure 5.11 Dataset-3. Examples of automatically detected SpectrumHits. 197
Figure 5.12 Dataset-4. STD scores. 198
Figure 5.13 Dataset-4. STD spectra examples. 199
Figure 5.14 Dataset-4. CPMG scores. 200
Figure 5.15 Dataset-4. CPMG results. 201
Figure 5.16 Dataset-4. Example of CPMG SpectrumHits. 202
Figure 5.17 Dataset-4. ^{1}H scores. 203
Figure 6.1 Chemical shift and exchange. 218
Figure 6.2 Chemical shift perturbation analysis (CSP) of the binding of Clip2 RNA (AAAUAA) to the Tstar-KH domain. 219
Figure 6.3 Mutation analysis using the Chemical Shift Mapping module. 222
Figure 6.4 Minimal shift analysis using CcpNmr version-3. 223
Figure 6.5 3D Ribbon representation of collaborator's macromolecule. 223
Figure 7.1 eIF4G-Mnk interaction background. 249
Figure 7.2 Project development workflow. 250
Figure 7.3 HADDOCK scorings. 251
Figure 7.4 eIF4G docking sites representations and docking libraries. 252
Figure 7.5 Single docking results for Site-1 and Site-2. 253

Figure 7.6 Site-1 and Site-2 most buried residues. 254
Figure 7.7 Atom interactions and fragments linkage preparation. 255
Figure 7.8 Fragment linkers. 256
Figure 7.9 Multiple-site docking results. 257
Figure 7.10 Ligand 423437 and development history. 258
Figure 7.11 Variants similarities for the ligand 423437. 259
Figure 7.12 Surface representations for optimised ligands. 259
Figure 7.13 MD simulations for 423437. 260
Figure 7.14 423437 interactions during simulation. 261
Figure 7.15 423437 angle torsions analysis. 262
Figure 7.16 423437 surface area analysis. 263

Abbreviations

1D	one-dimensional
2D	two-dimensional
3D	three-dimensional
airPLS	adaptive iteratively reweighted Penalized Least Squares
arPLS	asymmetrically reweighted Penalized Least Squares
ALS	Asymmetric Least Squares
BACE	Beta-Site APP Cleaving Enzyme
BCL	B-Cell Leukaemia/Lymphoma
BDC	Distribution-Based Classification
BINANA	BINding ANAlyzer
BMRB	Biological Magnetic Resonance Data Bank
CcpNmr	Collaborative Computing Project for NMR (software)
CPMG	Carr Purcell Meiboom Gill
Cryo-EM	Cryogenic Electron Microscopy
CSP	Chemical Shift Perturbation
CSV	Comma Separated Values
CWBC	Correlated Weighted Baseline Correction
eIF	eukaryotic Initiation Translation Factor
eq	equivalent
FBDD	Fragment-Based Drug Discovery
FDA	Food and Drug Administration
GPCR	G-Protein-Coupled Receptors
GUI	Graphical User Interface
HADDOCK	High Ambiguity Driven biomolecular DOCKing
HSQC	Heteronuclear Single Quantum Coherence Spectroscopy
IRES	Internal Ribosome Entry Site
JSON	JavaScript Object Notation

KDE	Kernel Density Estimation
L-RMSF	Ligand Root Mean Square Fluctuation
LogP	Logarithm of the Partition Coefficient
MBP	Maltose-Binding Protein
MCL	Myeloid Cell Leukaemia
MD	Molecular Dynamic
Mnk	Mitogen-activated Protein Kinase(MAPK) interacting Kinase
ms	millisecond
MW	Molecular Weight
NEF	NMR-Exchange Format
NMR	Nuclear Magnetic Resonance
NN2	Neural Network-2
NME	New Molecular Entities
NOE	Nuclear Overhauser Effect
NOESY	Nuclear Overhauser Effect Spectroscopy
PAINS	Pan-Assay Interference Compounds
PCA	Principal Component Analysis
PDB	Protein Data Bank
Ppm	parts per million
PQN	Probabilistic Quotient Normalisation
Pr-*	Protein-*
PSA	Polar Surface Area
RF	Radio Frequency
RMSD	Root Mean Square Deviation
ROC	Receiver Operating Characteristic
ROI	Region of Interest
S/N	Signal-to-Noise
SAR	Structure Activity Relationship
SF	Sample Reference
SMILES	Simplified Molecular-Input Line-Entry System

SP	Sample Protein
SPR	Surface Plasmon Resonance
STD	Saturation-Transfer Difference
TROSY	Transverse Relaxation-Optimized Spectroscopy
Tstar	Testis-Signal Transduction and Activation of RNA
Water-LOGSY	Water-Ligand Observed via Gradient SpectroscopY
WS	Whittaker Smoother

Amino Acid Abbreviations

Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic Acid	Asp	D
Cysteine	Cys	C
Glutamine	Gln	Q
Glutamic Acid	Glu	E
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

Chapter 1

Accounts of fragments to drugs by NMR. Where are the successes and where can it be improved?

Keywords: Fragments Based Drug Discovery, FBDD NMR, Venetoclax, BCL-2, BACE-1, MCL-1.

1.1. Abstract

Over the last century, the definitions of pharmaceutical drug and drug discovery have changed considerably. Nowadays, drug discovery involves several distinct, yet sometimes interconnected, stages aimed to obtain suitable molecules able to interact with a biomolecular target triggering a biological response. For each stage, typically a large range of different techniques are used to drive the project forward into the next phase in the fastest possible way.

High throughput screening (HTS) and fragment-based drug design (FBDD) are the two main approaches for the identification of drug-like candidates in the early stages of drug discovery. Nuclear magnetic resonance (NMR) spectroscopy has many applications in FBDD and is used extensively in industry as well as in academia.

In this chapter I discuss the development of molecules in which NMR had a crucial role; their efficacy proved either to be successful or unsuccessful in a clinical setting. I specifically focus on the techniques used, describing strengths and weaknesses for each stage by examining several case studies. More precisely, I examine the development from the primary screening to the final lead optimisation of AZD-3839 and interactions to its target BACE-1, ABT-199 and interactions to BCL-2/xL, and lastly, S64315 and its interactions to MCL-1. From the analysis, I derive conclusions regarding their chemical development patterns and express personal suggestions on improvements that can be made to the FBDD by NMR.

1.2. Introduction

Fragment Based Drug Discovery, FBDD, is nowadays a solid and common approach vastly adopted by a multitude of pharmaceutical companies and academic groups[1]. The rationale behind FBDD has been extensively reviewed and entire books have been written including some references to clinical candidates[2,3]. The general concept of FBDD is straightforward. It starts by the generation of libraries of small molecules called “fragments”. The size of these libraries vary from few hundreds to thousands of small molecules (for industrial cases)[4], that usually follow the so-called “rule of three”, i.e. a molecular weight < 300Da, LogP < 3, maximum hydrogen bond donors and acceptors < 3[5].
Binding of the molecules comprising the libraries are evaluated against a target of interest, usually in so-called “mixtures” of 5 to 10 compounds[6]. The strategy of using small molecules instead of large entities allows for a more efficient exploration of chemical space, defined as the ensemble of all possible molecular conformations which present drug-like properties (~10^{60} molecules)[7]. The approach provides a greater chemical variety that also brings other benefits, such as cost and time reduction in the data analysis.

The use of fragments at starting points in the early stages of drug discovery has been demonstrated to be a viable approach for producing compounds that are highly tailored to their targets[8]. This method also increases the novelty of standard drugs and provides for the possibility of monitoring the chemical path of optimisation, for example, restricting the lipophilicity issue observed in HTS-derived drugs[1,9]. However, since the fragments are much smaller compared to traditional lead-like molecules, their binding affinity to a target of interest is nearly always low (µM to mM). Therefore, only a highly sensitive technique, such as NMR, can detect these weak interactions[10,11].

A variety of NMR methods have been developed over the years to cover each step of drug development, which can be broadly divided in the primary hit detection, the binding site identification, the binding mode elucidation and fragment optimisation (Fig. 1.1).

In this chapter I focus exclusively on the relevance and impact of NMR spectroscopy on the generation of new clinical drugs. Through the analysis of three case studies, I discuss the various techniques used and discuss at which stages of drug development these techniques were crucial to the various projects.

To appreciate the influence on NMR in the current developments, a search on the PubMed database using the keywords “FBDD NMR” was performed. It revealed over 600 journal publications for the last five years (Fig. 1.2A). It is to be expected that not every breakthrough for the NMR technique will have been shared in the public domain; often the origin of new molecular entities remains obscure or difficult to trace. For example, there have been instances of journal articles that described new molecular discoveries without explicitly indicating the relevant molecular names in their titles. A further search through the FDA database revealed, over the same five-year time span, that only two approved drugs were unmistakably obtained from fragments (Fig. 1.2B)[12,13]. Whilst this may seem to be a very low number relative to the total number of FDA approvals, this is partially explained by the fact that more than 40% of the total number of FDA approvals pertains to antibodies (Fig. 1.2C). Not surprisingly, cancer-related diseases register a much larger number of small molecules when considered as a single area of treatment, followed by neural- and cardiac diseases (Fig. 1.2D). In addition, the FBDD pathway is more complex than the high-throughput screening[14], as it requires several steps before achieving a final drug-like compound, as a result, so far it was necessary an average of 13 years of development after the first low affinity hits were detected (Fig. 1.3A).

I performed an in-depth analysis of all FBDD-derived molecules that are or have been clinical candidates at any time in the past up to December 2019 and for which relevant information was publicly accessible through journal articles tables or the Erlanson blog[15]. The analysis showed that NMR is predominantly used in the primary screening for the initial hit identification. For the subsequent FBDD stages of binding site identification and hit optimisation, NMR is increasingly less often used (Fig. 1.3B). This shift indubitably parallels the increased usage of alternative techniques, such as X-ray crystallography, that proves to be the most preferred method for the hit optimisation stage[4] (Fig. 1.3C).

NMR-derived compounds were identified mostly by ligand-detected 1D NMR techniques, such as Water-LOGSY, saturation transfer difference, STD, or $T_{1\rho}$, whereas target-based 2D NMR techniques, such as the chemical shift perturbation (CSP) experiment, were used for the hit and/or binding site validation. Lastly, the so-called SAR by NMR method, which employs mostly NOE-related techniques and multi-dimensional NMR experiments, was mainly used for hit-growing and linking guidance during the optimisation stage of the FBDD process (Fig. 1.3D).

The 1D ligand-detected techniques, such as STD and Water-LOGSY, are used as a gold standard in NMR screening as these do not require expensive protein labelling and therefore can be used for a broad range of molecular targets[16]. Furthermore, the various expression systems of the target, e.g. bacteria, insects or human-derived cells, and other common limitations, such as molecular weights, are not of critical importance for these 1D techniques[17]. In addition, they can also be used in difficult cases where expression and/or purification of the target macromolecule is a limiting factor and only nM concentrations can be attained. Most importantly, the richness of information acquired in a small amount of time (i.e. minutes per sample), allows to perform the analysis in a high-throughput fashion[18]. However, 1D ligand-detect experiments are not suitable for detecting binding sites of interactions, and higher dimensionality NMR techniques, including CSP, are often required. The latter enables the monitoring of target residues that are most likely to be interacting to the fragments, providing precious information for binding validation as well as guidance on the next stage of development[19].

Fragment optimisation is best achieved where a high-resolution 3D molecular structure of the target is available. There are several techniques capable of resolving molecular structures; however, the simplicity and the generally rapid throughput associated with X-ray crystallography, renders this method as the most preferred whenever possible[20] (Fig. 1.3C). In reality, however, very often targets of interest cannot be assessed by X-ray crystallography. For example, complexes displaying a highly flexible mode of interaction can be truly inspected only by NMR[21,22], as crystal packing forces preclude the required molecular adaptation for complex formation. Moreover, the crystal lattice also might not allow the ligand to permeate through to the binding pockets[23]. In contrast, the NMR technique can provide unambiguous information on the various orientations of the ligand with respect to the target, referred to as poses; these poses can be combined with computational methods for designing drug-like compounds with improved binding and pharmacological properties.

In this chapter I present three case studies that employed a variety of NMR techniques and therefore can be considered models of FBDD by NMR.

The first case revises the compound AZD-3839 development path. It was originated by fragments identified by a ligand-detected primary screening using the Water-LOGSY technique. The second case examines the history of an FDA-approved drug which was derived by FBDD, ABT-199, commercially called Venotoclax. The development of this compound is linked to a variety of target-detected NMR methods. Finally, the analyses of S64315 highlights a combination of ligand-detected and target-based NMR techniques in one of the latest clinical drugs derived by FBDD.

1.3. Materials and Methods

1.3.1. Materials

FDA information for New Molecular Entities (NMEs) and original biologics were extracted from https://www.fda.gov/drugs/drug-approvals-and-databases/drug-trials-snapshots. For each entry, a literature search was conducted to determine whether small molecules (only) were derived from an FBDD-approach.

The list of molecules in various stages of clinical trials was reproduced from a number of web sources and published reviews[1,4,24,25]. Some web-based materials were extracted from a detailed analysis of the website Practical Fragments[26], where blog articles dating up to 31st December 2019 and a table as listed in 2015 and 2018 posts[15,27] provided the starting point for this study.

A literature review was conducted for each compound, filtering out only molecules in which NMR had been involved at some stage of the drug discovery process. Subsequently, the exact NMR technique used was noted wherever possible.

1.3.2. Methods

Fragments molecular structures were reproduced using our in-house software ChemBuild[28]. Molecules were then exported in PDB or MOL2 formats, re-aligned and exported in PNG formats from PyMOL[29].

OpenBabel or iBabel 3.6 was used to convert PDB and MOL2 to SMILES format[30],[31]. Smiles were created in the canonical (xc) format; hydrogens and pH were excluded from the calculations of the various molecular properties. Molecular weights, polar surface areas and other properties were calculated using the online tools available at http://www.cheminfo.org.

A collection of scripts for analysing smiles and plotting molecular similarities were written in Python using the Pandas[32], Numpy[33], SciPy[34], Matplotlib[28] libraries. The

Pybel[30] package was used for calculating the molecular fingerprints from SMILES and the Tanimoto coefficient[35].

PDB codes. AZD-3839 and BACE-1: 4B05; BCL and ligands: 6O0L, 4LVT, 2YXJ; MCL1 and ligands: 6QXJ, 6QYK, 6QYL, 6QZ5, 6QZ7, 6QZ6, 6QZB, 6QYN, 6QZ8, 6QYP, 6QYO.

1.4. Case study 1: AZD-3839 and BACE-1

β-Site Amyloid precursor protein Cleaving Enzyme-1 (BACE-1) was identified over twenty years ago as a key component in Alzheimer Disease (AD) pathogenesis[36,37]. BACE-1 is responsible for the initial cleavage of the amyloid precursor protein to its smaller amyloid β-peptides (Aβ), in which accumulation in the brain cells is believed to be one of the underlying causes of AD progression[38]. Not surprisingly, BACE-1 is a therapeutic target and a number of academic groups and pharmaceutical companies have placed considerable efforts into the research and development of new inhibitors in the hope of limiting or blocking the formation of Aβ[39–42].

BACE-1 is characterised by an internal groove created by two lobes (S1 and S2), modulated by a loop ("flap") which reveals the aspartyl catalytic site. The flap is highly dynamic and upon the presence of inhibitor can determine the state of "open" or "close" of the macromolecule giving access to the catalytic pocket. The identification of two crucial aspartic acid residues, i.e. Asp32 and Asp228, has for many years driven the drug development process and optimisation of fragments[39]. An exhaustive list of early fragments and their respective primary screening technique is given by Erlanson and Jahnke[39].

A great example of the history of a complete development is provided by the compound AZD-3839, where the initial fragment was identified from 1D NMR studies. According to Geschwindner *et al.* the choice of NMR for this case provided a compromise between scalability of large fragments libraries and sufficient data-output, while simultaneously assuring a robust method for detecting very weak bindings at low ligand concentration. By eliminating non-specific binders, this process would eventually have the advantage of reducing false positives from the analysis. The original screen using the Water-LOGSY 1D NMR technique was conducted on a 2000-compound library with four fragments per mixture, yielding a relative low hit rate of 0.5%. Compound-1 (Fig. 1.4A, 1) was identified as a binding hit. An intensity "sign-flip" of signals relative to this compound was clearly distinguishable from the NMR spectra, suggesting a binding event to the macromolecular target[43].

Crucially, as a control the authors performed a competition experiment in presence of a stronger known binder, showing a noticeable intensity reduction (more negative) only for the singlet peak from the isocytosine aromatic H^5 proton at around 5.65 ppm (Fig. 1.4A, 1, blue circle). This validation assay reduced potential false positives by identifying fragments that displayed weak binding Water-LOGSY responses but did not show any changes upon the addition of the competitor. Compound-1 was eventually selected for further optimisation steps. Meanwhile, through parallel crystallographic studies performed by Astex Therapeutics[44] an optimised compound that preserved the original amidine motif was developed. The amidine motif was confirmed to be responsible for the strong interaction to the catalytic aspartates (Fig. 1.5A). Consequentially, through a series of scaffold hopping substitutions, the molecule was morphed into the isoindole present in the final compound. Furthermore, the introduction of fluoro atoms improved the permeability of the molecule and the brain exposure by "shielding" the reactive amidine. Lastly, an additional molecular growing, using aromatic cores, gained the extra surface needed for interacting to the adjacent S3 and the flap[45] (Fig. 1.5A), concluding the hit optimisation for this compound.

The various steps of the hit optimisation clearly show how the initial NMR-derived fragment hit has undergone a series of dramatic changes. The magnitude of these changes can be assessed from the similarities of the molecular fingerprints for each component as calculated by the Tanimoto coefficient scoring[35]. A prominent drop of the Tanimoto coefficient is observed from the first molecule to compound-2 and further for compound-3 (Fig. 1.5B). However, from compound-4 to the final AZD-3839 molecule, a much less variable score is observed. A different trend was observed for the molecular weight (MW) of the successive compounds which showed a constant increment up to compound-4, followed by only minor changes towards the final form AZD-3839. Interestingly, the final compound was characterised by a smaller molecular weight (MW) than its previous version, yet presented an increased polar surface area (PSA; Fig. 1.5B). AZD-3839 appeared to be a very promising drug candidate and underwent clinical phase-1 trial. Unfortunately, it was stopped from patient administration, probably due to its high affinity to the hERG ion channel and potentially related side-effects[46]. This case, nevertheless, demonstrated that NMR was crucial for determining the first fragment hit from the primary screening. NMR unambiguously established the identification of the initial amidine-fragment. This motif was revealed of a critical importance in forming interactions to BACE-1, and as a result it was preserved through the long path of chemical optimisations that resulted in the final AZD-3839 compound.

1.5. Case study 2: ABT-199 (Venetoclax) and BCL-2/xL

The second case study is an analysis of the molecular background for the compound ABT-199, commercially referred as Venetoclax. The history of this drug was selected for two reasons: the large impact of NMR throughout its development pathway and the fact that the Abbott NMR group, where earliest studies began, has been pioneering the "SAR by NMR" method, which culminated in the FDA-approved Venetoclax drug in 2016[47,48].

ABT-199 is an inhibitor of the anti-apoptotic proteins BCL-2, BCL-xL and BCL-w[49]. These proteins play a pivotal role in the cell survival; not surprisingly, they are over-expressed in many cancers and they are directly linked to initiation, progression and therapy resistance occurrences[50]. The various BCL members are α-helical proteins. Two protein, i.e. BCL-2 and BCL-xL, share four domains, BH4 (α1), BH3 (α2), BH1 (partially α4) and BH2 (partially α6 and α7) plus the transmembrane, TM, motif. The disposition of two-central hydrophobic helices (α5 and α6) together with the amphipathic α1-4 and α7 form an elongated hydrophobic groove in the BH1, BH2, BH3 region[51], (Fig. 1.7A). The BH3 region, in particular, is responsible for the interaction with other proapoptotic proteins such as BAK and BAX, rendering it a druggable site of interest[52].

The early inhibitor-discovery process was started by screening a large library of small molecules using 2D target-detected approaches which led to the identification of several candidate molecules or "hits", (Fig. 1.6, compounds 1 and 2). The hypothetical binding mechanism was elucidated through ^{15}N-HSQC chemical shift perturbations (CSP) experiments[53]. From the CSP results it was possible to derive that the fluoro-biaryl acid region of the selected compound-1 interacted with the BCL-xL hydrophobic groove. In fact, a series of large shifts were observed for the peaks assigned to residues G94, G138 and G196, located in this groove[53]. However, the study of the complex of BCL-xL with its binding partner BAK suggested the existence of an additional binding interface.

Therefore, a second screening was carried-out in the presence of a large excess of the compound-1, with the aim of saturating the first site of interaction and screening for potential hits to the second interface[53]. Compound-2 (Fig. 1.6) was identified and eventually chosen to be used in a chemical linkage to the compound-1[53]. To do so, multiple linkers and various compound-1 poses were explored in order to improve the overall potency of the resulting molecule. Finally, a ~200-fold improvement in binding affinity was established for compound-3 when compared to the original biaryl-acid (Fig. 1.6).

The first model of the complex of BCL-xL with compound-3 was then developed on the basis of nine intermolecular NOEs[53]. Although these NOEs were indicative of an interaction with both binding interfaces, it was concluded that compound-3 did not adopt optimal or ideal conformations. Consequently, new linkers and a new set of chemical reactions were explored.

Compound-4 was eventually synthesised and evaluated structurally by combining multiple protein-ligand NOEs extracted from 3D ^{13}C-edited and ^{12}C-filtered NOESY spectra to dock the molecule in the BH3 mediated groove[53]. Compound-4 was further optimised in parts that were solvent-exposed. These parts of the molecule were replaced with polar substituents, including a 2-dimethylaminoethyl group in the linker. In addition, the insertion of a new piperazine ring led to the compound ABT-737 (Fig. 1.6).

ABT-737 displayed increased potency and exhibited activity in the presence of human serum. An *in-vivo* analysis suggested that a synergetic therapy was required for inhibiting the anti-apoptotic activity of the BCL family while promoting the pro-apoptotic proteins (BAX and BAK)[54]. Studies by the Abbot group proceeded towards the development of the ABT-263 molecule. After an initial positive assessment on multiple cellular lines, where ABT-263 reported stronger inhibitory actions, presumably by targeting both BCL-xL and BCL-2, advanced clinical studies unfortunately revealed major pitfalls such as thrombocytopenia[55]. Eventually, a crystal structure of BCL in complex with ABT-737, resolved at 2.2 Å, validated the original NMR-determined inhibitor binding pose[56]. The crystallographic model showed that ABT-737 interacted with the two binding interfaces formed by the hydrophobic pockets, P4 and P2, of BCL-2 and BCL-XL; including two hydrogen bonds from the thiophenyl and the 1-chloro-4-(4,4-dimethylcyclohex-1-enyl)benzene moieties to residues G138 and E96, respectively.

Meanwhile, the project continued at the AbbVie group together with a number of collaborators who designed the final compound ABT-199 (Venetoclax) from ABT-263 through a series of substitutions[55]. In addition, new 3D molecular structures of BCL-2 in complex with various ligands were made publicly available. The ABT-199 molecule incorporated several crucial modifications compared to its predecessor. Through an intermediate compound in the series, a pivotal H-bond was identified to D103 (corresponding to G96 in BCL-xL), thus providing an increased affinity to both the BCL-2 and BCL-xL P4 hydrophobic pockets[49] (Fig. 1.7A).

A molecular fingerprint analysis was performed for all available molecules in the development process. However, as the initial NMR-detected fragments underwent a linkage step, the molecular similarities were assessed with respect to compound-3. Similar to the AZD-3839 case study, a drastic drop in the Tanimoto coefficient was observed from compound-3 to the successive optimised forms (Fig. 1.7B). Interestingly, ABT-199 showed a reduced molecular weight and increased polar surface area compared with its predecessor compounds, yet keeping an overall structural similarity starting from the compound 4, although differed notably to the first NMR hits. (Figs. 1.6 and 1.7B).

The ~20 years history of development for the ABT-199 compound revealed a multitude of challenges, including the impossibility of obtaining crystals of complexes with the first leads and some other *in-vivo* difficulties which were not predictable from a structural point of view. However, I am confident to state that the success of Venetoclax could not have been achieved without the crucial data resulting from the usage of multiple NMR techniques during the early stages of the drug discovery process.

1.6. Case study 3: S64315/MIK665 and MCL-1

The final case study presents an overview of the most crucial optimisation steps in the development of the molecule S64315, also known as MIK665[57] (Fig. 1.8). S64315 is one of the latest inhibitors currently being tested in clinical trials for targeting the BCL anti-apoptotic family, MCL-1[58]. A series of studies indicated MCL-1 is over-expressed in many cancer types (multiple myeloma, lymphomas, leukaemia); therefore, it is widely recognised as a druggable target[59]. MCL-1 shares the highly conserved BH3 binding groove with other members of the family such as BCL-xL and BCL-2. This groove is essential for interacting and sequestering the pro-apoptotic proteins resulting in an increased cell survival.

The development of specific inhibitors for MCL-1 targeting the BH3 groove has proven to be challenging[60]. The Vernalis group, together with collaborators, engaged in large efforts in their studies of this complex. The results of the studies allowed for maximising the potency of an initial hit obtained from a ligand-detected NMR screening resulting in the most sophisticated MCL-1 inhibitor to date, S64315/MIK665[57]. However, several difficulties had to be overcome in the expression and purification of the macromolecule in human-derived cell. Consequently, this resulted in a lack of 3D atomic structures which hampered the FBDD studies[57]. Despite this, the protein availability was adequate for the initial NMR-based screening.

The assessment of over 1000 fragments, pooled in groups of eight, using 1D STD, Water-LOGSY and CPMG NMR experiments revealed several potential binding hits. Due to low signal-to-noise ratio, hits were further validated using 2D NMR ^{15}N-HSQC titrations. In addition, to overcome the absence of a detailed 3D molecular structure a new approach for determining ligand poses and guiding the optimisation process, referred to as NMR-guided model (NGM), was developed. The NGM approach employs 3D NMR methodology, i.e. X-filtered NOESY (13C-edited, 13C, 15N-filtered NOESY), to identify crucial NOEs between ligands and the target. The resulting information was combined with high-throughput computational docking studies, allowing for a more accurate classification of binding poses. From the NMR results, multiple compounds with various chemical functionalities were explored, of which a class of compounds comprising a thienopyrimidine group was believed to be the most promising. Particularly, compound-1a was used as the initial fragment towards the development of the S64315 drug (Fig. 1.8). Following a series of substitutions for the compound's ethyl group, multiple variants were tested on BCL-2, BCL-xL and MCL-1. Some of the newly synthesised molecules showed comparable affinity toward all three targets[57]. Using the ^{15}N-HSQC technique it was possible to estimate K_d values for most of these, which ultimately allowed to select compound-5d as the highest affinity binder towards MCL-1.

The NOEs derived from the analysis of the compound-5d/MCL-1 complex indicated several potential contacts. In particular, contacts between the naphthyl ring (Fig. 1.8, 5d, green circle) to the side chains of A227, M231, V249, V253, T266 were observed as well as between a methyl group (Fig. 1.8, 5d, blue circles) to the side chains of M231, V249, V253, L267. To further investigate the BH3 binding region's molecular flexibility, the authors assessed various possible docking poses using multiple structural ensembles. This approach allowed a better estimation of possible allowed geometries, yet consistent with the crucial experimental NOEs information. Ultimately, the preferred molecular orientation consisted of the carboxylic acid pointing toward the solvent region, and the naphthyl group toward the S2 pocket. Different conformational changes for the hydrophobic groove were also assessed. This was achieved by inserting various substituents to the original small molecule core and testing the different rotational property of the aryl ethers and anilines (Fig. 1.8, 8d-15, black circles).

Lately, crystallographic structures of the MCL-1 complex and some variants became available (PDB codes listed in material and methods), allowing for alternative studies for several fragments and their binding modes.

Multiple optimisation steps were carried out, eventually leading to the final state-of-the-art S64315. This final compound presented new crucial ortho-substituents, such as the fluorobenzene and methoxyphenyl-pyridine group, which were responsible for the increased selectivity for MCL-1 compared to its precursor (Fig. 1.8, orange circles). A model of MCL-1 in complex with compound-18a was generated using the MCL-1 crystal structure (PDB code 6QYO) to manually dock the S64315 compound (Fig. 1.9A). The model shows the thienopyrimidine motif, already observed in the original compound-1a, deeply buried in the hydrophobic groove of the BH3 binding domain.

As for the previous case-studies, the molecular fingerprint patterns using the Tanimoto score were inspected for all available compounds (Fig. 1.9B). In line with the observations for ABT-199 and AZD-3839, albeit somewhat less prominent, the results again show the characteristic initial decrease in the Tanimoto score from the first fragment to the following variants, indicating the significant changes during the initial steps of development. Interestingly, several compounds mid-way through the development (i.e. compounds 5d, 8d) showed a higher Tanimoto coefficient compared to the initial optimised fragments, suggesting a more careful optimisation process rather than a revolutionary approach to the first fragment hit. Starting from compound-10 only smaller changes occur together with increased PSA scores. Surprisingly, the final compound appeared to differ the most from its direct precursors. This compound also showed an increased MW and a reduced PSA score compared to its previous three variants.

The search for an MCL-1 inhibitor started several years ago from the identification of a first hit obtained through primary screening by NMR. The process illustrates the huge amount of work required to bring an initial hit to a final lead drug candidate, which included the efforts of several academic and industrial laboratories. The failure of crystallisation trials during the early stages of the project, plus the inherent flexibility of the MCL-1 BH3 binding groove, made NMR spectroscopy uniquely capable of driving the project forward. The S64315 compound is currently under evaluation in the clinical phase-1 trials, which provide hope for patients affected by a variety of cancer-related diseases.

1.7. Discussion and Conclusions

In this chapter, I explored the development histories of the AZD-3839, ABT-199 and S64315 compounds, from the primary screening to the final lead optimisation, focusing on their target interactions and the rationale behind their optimisations. These cases highlighted the underlying role of NMR techniques during the drug discovery phases and their impact throughout each stage.

With multiple compounds in clinical phases, NMR has demonstrated a key role in the process of fragment-based drug discovery[4]. In 2016, ABT-199, commonly known as Venetoclax, was the first confirmed FDA-approved drug derived near exclusively by NMR-FBDD[26,49,61,62]. The development of other fragment-derived drugs were driven by both X-ray crystallography, NMR spectroscopy and optional other techniques, e.g. Vemurafenib (approved in 2011)[63] and Erdafitinib (released in 2019)[13,64].

Throughout the years, different methodologies have been explicitly developed for enhancing the success rate of drug discovery[65]. The great flexibility and adaptability of NMR provides for qualitative and quantitative insights at each point of the drug progression[17,65]. However, NMR also has a number of drawbacks[66]. Starting from the primary screening, the usage of only a single ligand-detected 1D technique for identifying binding fragments may prove to give erroneous results. Hence, it is recommended the use of at least two parallel methods such as STD and Water-LOGSY[66].

In the case of STD experiments, false positives or false negatives can originate from wrongly irradiating directly the ligand, or by not completely saturating the target, respectively. Interpretation errors or ambiguities can also result from contributions arisen by bound and unbound states in the case of Water-LOGSY or from abnormal relaxation series for CPMG experiments[66]. Furthermore, an inevitable consequence of using fragments in primary screening is the so-called “non-specific” binding event[67]. A simple strategy for alleviating this issue was employed in the development of AZD-3839. By recording competition experiments using a potent known ligand for the same target, non-specific binding molecules were identified and excluded from further development. It was not clear whether a similar approach was addressed for assessment of the initial hits in the S64315 discovery.

An additional key step before expensive and laborious optimisation processes of candidate compounds begins is a minute hit chemical-assessment; for example, by employing Pan-Assay Interference Compounds (PAINS) protocols[68]. Applying these filters to hits or family of hits can help in identifying fallacious binders. PAINS-flagged molecules can exert photo-reactivity, redox-activity and other undesirable chemical phenomena which can lead to non-specific biological activities. Unfortunately, it is also wise not to rely solely on PAINS filters. A recent analysis showed that many PAINS-flagged molecules had been wrongly evaluated by the applied filters; the study included, in fact, numerous examples of false negatives and false positive[69].

Ayotte *et al.* proposed the use of CPMG series to detect potential aggregation of compounds in mixtures[70], in order to remove the offending ones from mixtures. In addition, they also pointed out that aggregation can be solvent-dependent, and thus a minor adjustment of the sample composition might improve the outcomes of screening experiments[70]. However, this optimisation approach can be very laborious and time consuming, either in the practical preparation and data analysis stages.

From the analysed case studies, it appeared that analytical power of 1D NMR spectroscopy often was not fully exploited. Instead, the NMR data appear to be used solely as a binary response, probably also due to a lack of proper computational and data analysis tools. The data obtained from Water-LOGSY and STD experiments have shown to offer further quantitative information[71,72]. The SAR by Water-LOGSY, for example, suggests a scoring factor as a mean to identify the most exposed portion of the molecule. Assessing all data that can be derived from 1D NMR experiments can eventually provide insights of the ligand binding pose[73].

Upon validation of fragment hits, the next stage is usually the exploration of potential binding sites on the target. Chemical shift mapping (CSM) is so far the most popular technique used for this task. CSM has been widely used as a standard for molecules that progressed into clinical phases. Nevertheless, a CSM analysis might drive researchers in wrong directions, and final conclusions should not be based on this approach only. Common errors observed in practice include wrongly determining what is relevant based on subjective judgments or misinterpretation of crowded regions of the spectra. Furthermore, in some instances compounds have been shown to change the pH of the solution, resulting in false positive CSPs[66]. By performing appropriate control experiments errors such as these might hopefully be avoided.

Despite clear benefits shown by NMR, it has often been associated with a requirement for daunting and time-consuming data analysis. There may be a multitude of other undescribed factors, but NMR's lack of modern, more practical, quicker, and unbiased methods, alongside with automated data analysis routines has comparatively slowed the entire NMR-FBDD process, prompting a need for improvements in all these aspects. These improvements are fundamental so that molecules are designed appropriately from the early stages onward. Ultimately, only *in-vivo* and clinical trials can determine if a drug candidate has to start an optimisation cycle again, as exemplified by ABT-263 which induced high thrombocytopenia or by the unexpected side-effects of AZD-3839 that potentially led to the premature end of its clinical trials[74].

From the analysis of the three cases discussed in this chapter, it appeared that molecular optimisations are still guided by multiple manual moiety substitutions following by their chemical synthesis and re-evaluation. Although this might generate potential leads, it can be argued that the usage of computational approaches could have accelerated the process further. Molecular docking studies aided the final lead generation of the ABT-199 compound; however, combining NMR and molecular docking can still introduce many mistakes and docking alone cannot be trusted (see Yu-Chian Chen for a detailed discussion[75]). In addition, scoring functions in docking protocols present a highly variable issue and even the newest scoring functions implemented using artificial intelligence (AI) could present limitations[76], which may arise from incomplete or erroneous classification of existing experimental data; information that will be necessary for providing accurate scoring of succeeding experiments[77].

I firmly believe that newer chemo-informatics and AI algorithms, together with improved high-performance computing facilities, will replace in large scale the optimisation stages for FBDD and SBDD. Several algorithms can already complement the experimental data validation and enhance initial binding hits[78]. Whereas most algorithms are included in commercial or proprietary packages, such as Schrodinger[79] and MOE[80], others are of difficult installation or require user-unfriendly command-line programs, for example GANDI[81], and AutoGrow[82]. Furthermore, they clearly do not replace the role of a chemist and can only assist in the initial design phases.

Ideally, computational approaches should also take the biological aspects into consideration simultaneously. For example, the robustness of ligand for potential drug resistance could be taken into consideration. By doing so, multiple fields of pharmacology can be integrated in a way that scientists alone could not have attained.

One question has arisen from the analysis of these study cases: when optimisations should stop and what can define a viable drug? Perhaps, until clinical trials only computational intelligence can truly help in giving an answer to this query.

By performing a simple molecular fingerprint analysis, I have identified some correlations in the development patterns among the three models studied. Comparing the Tanimoto coefficients for the three final compounds (Figs 1.5B, 7B, 9B) I could speculate that by expanding the molecule (via growing/linking methods) did not improve the binding affinity (not shown). This could suggest that just covering the conformational molecular interaction space does not necessarily lead to higher affinity drugs. Fig 1.10 displays the normalised Tanimoto scores for the three different fragment-to-drug evolutions with interpolated optimisation steps. Clearly, the development of AZD-3839 and ABT-199 display a highly similar pattern, which is more dissimilar from S64315. An obvious conclusion could be that they simply differ in their optimisation protocols, but equally it could suggest that S64315 is still in a middle of an active evolution and multiple changes will occur before “converging” to the final drug.

In conclusion, although only a few drugs approved by FDA have a fragment-based trackable history that is easily accessible from the public domain data, a publication count over the last 5 years shows an impressive number of journal articles reporting on new discoveries in the pre-clinical stages in which NMR had a crucial role. In all, this provides for hope that in the near future new potent and selective drugs will become available that will be developed on a much shorter time scale when compared to time taken for the currently FDA approved FB-derived drugs.

1.8. Figures

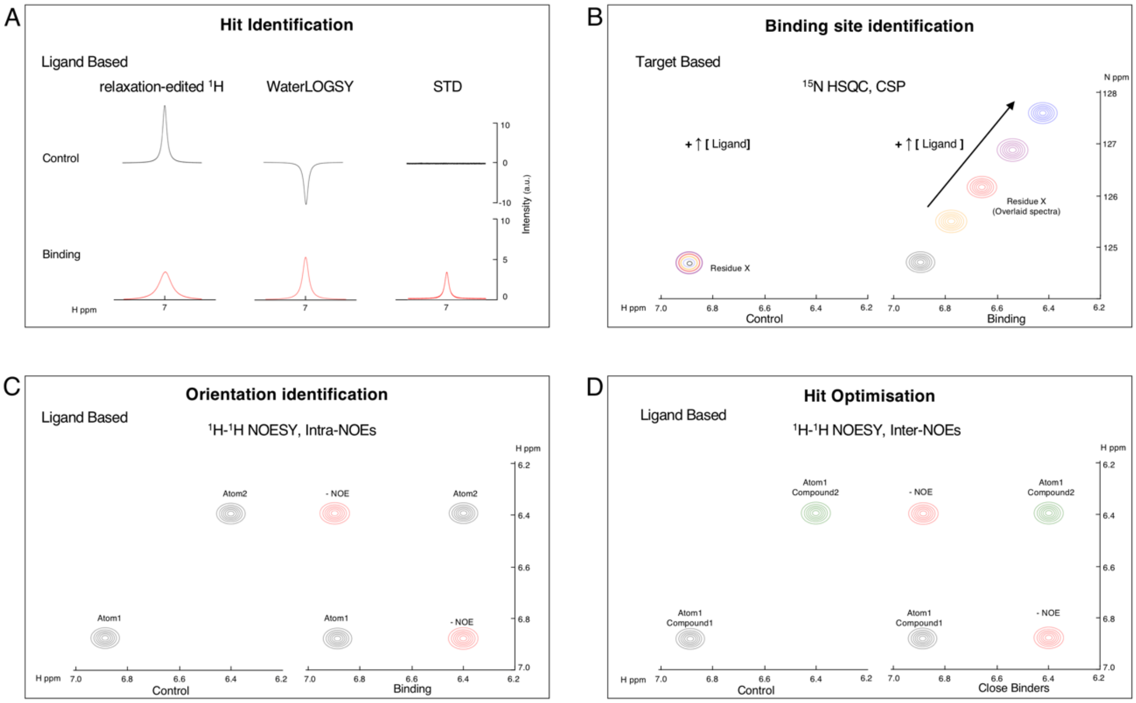


Figure 1.1 NMR-aided fragment-based drug discovery (FBDD).

Ligand-detected hit identification (**A**), chemical-shift mapping (CSM) binding site identification (**B**), fragment orientation identification and optimisation (**C-D**) aided by several ligand-detected NMR techniques1, mainly Nuclear Overhauser Effects (NOE) experiments, Inter-ligand NOEs for Pharmacophore Mapping (INPHARMA), and Interlined Overhauser effect NMR (ILOE-NMR), but also using specific labelling schemes as in the Selective Labelling STD experiment (SOS-STD). Note that all chemical shift scales are arbitrary.

A Common ligand-detected hit identification NMR methods and their respective simplified spectral appearances as manifested by recording ligand spectra in absence (control) and presence of the macromolecular target. (Left) The ^{1}H-relaxation-edited control spectrum ligand signals characterised by narrow lines, which broaden and consequently reduce in intensity as a result of the increased relaxation rate due to binding the much larger target. (Middle) In the WaterLOGSY experiment saturation of the bulk water is transferred via chemical and dipolar exchange to the ligand. The binding event is identified by an inversion of the ligand signals compared to the control. (Right) In the saturation-transfer difference (STD) experiment, saturation of target resonances is transferred via dipolar exchange to the ligand. In the STD spectrum, obtain after subtraction of the “off-resonance” from the “on-resonance” spectrum, only the signals of the bound ligand are observed. **B** Chemical shift mapping (CSM) for detection of the ligand binding pockets on the target. In the 15N-HSQC spectrum of a target protein, peaks are commonly used as a distinctive identifying characteristic of individual residues. Upon titration of a small molecule at increasing concentrations, a ligand binding event will result in a modification of the spectral peak patterns, which can be easily tracked in case of so-called fast exchange. By assignment of the affected peaks to their corresponding residues, and optionally mapping these changes to the protein’s three-dimensional structure, it is possible to perform an evaluation of most affected residues in the potential ligand binding pocket (right panel, Binding). **C** Schematic overview of a transferred-NOE ^{1}H-^{1}H NOESY spectrum for detection of a ligand pose. In case of binding, the ligand takes on the NOE properties of the target (rotational correlation time, τc), showing strong negative NOE, i.e. positive peaks (red) revealing its ligand-bound conformation. In principle, protein-ligand inter-molecular NOEs could be observed as well; however, as the protein concentration is usually much lower compared to the ligand, these NOE will be (much) weaker and difficult to interpret (not shown). **D** Theoretical ^{1}H-^{1}H NOESY spectrum for a sample in the presence of two ligands. Inter-molecular NOEs peaks between molecules are enhanced if they bind in close proximity on the target (red peaks).

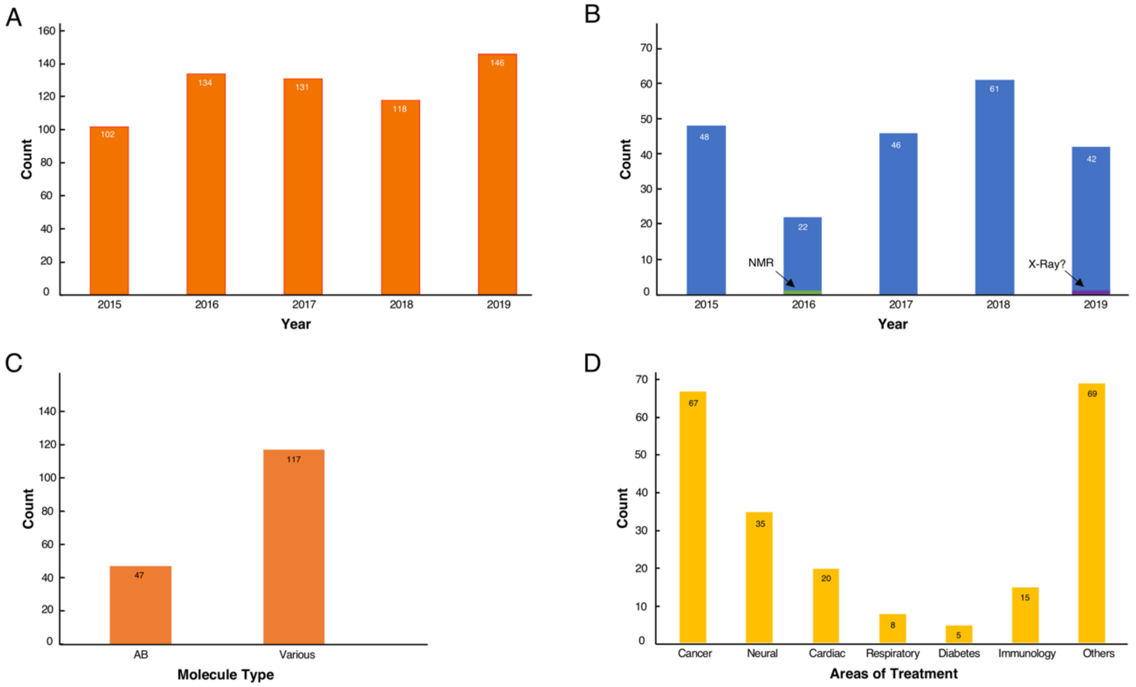


Figure 1.2 Usage of FBDD methods in the development of new molecules. **A** total count of journal papers from Jan-2015 to Dec-2019 retrieved by querying "NMR" and applying a custom filter: "Fragment based" in the database PubMed. N.B. Not all articles have been fully accessed to verify the pertinence to the subject. **B** Total count of New Molecular Entities (NMEs), and original biologics for the same range of time approved by the FDA. In green, the NME of which Fragments origins were derived by NMR studies. In purple, the latest NME derived by FBDD but of undisclosed or not sufficient public data available regarding the methodologies used in early stages of development. **C** Schematic division of FDA approvals, for "various" are intended small molecules and everything else is not an anti-body (AB). **D** Areas of treatment for the various FDA approvals NMEs and original biologics.

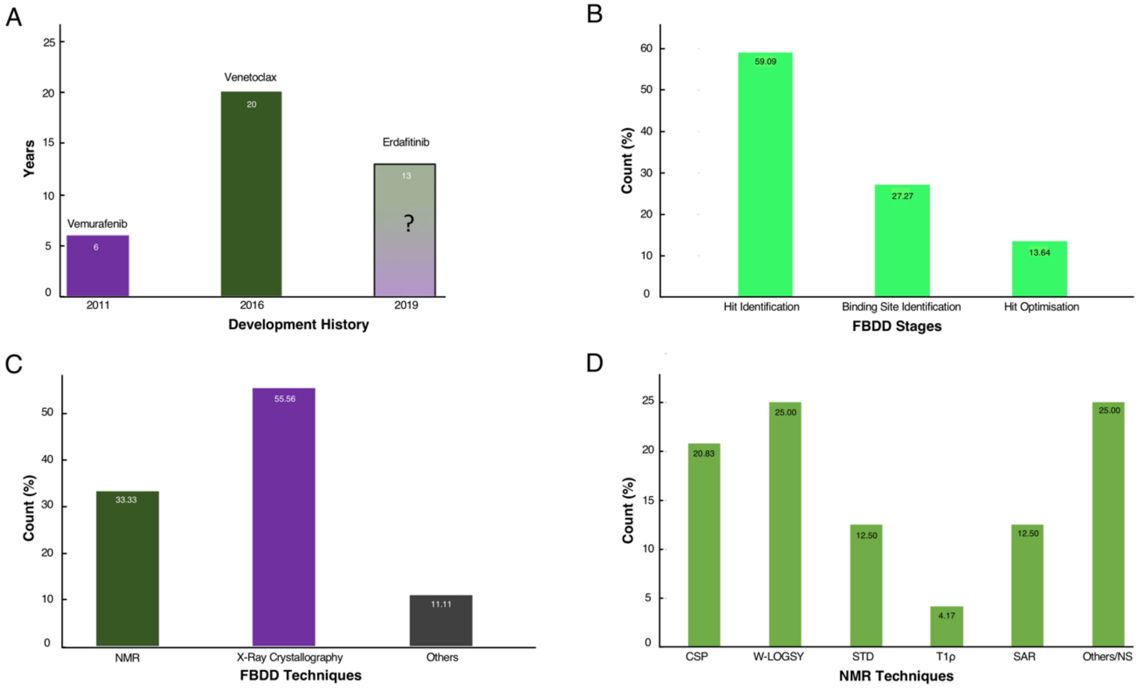


Figure 1.3 Role of NMR-FBDD methods in drug discovery.

A Years of development for three different drugs where FBDD techniques were reported throughout the development. For Venetoclax development were adopted mostly NMR studies[12], while for Erdafitinib the development history and adopted techniques remain unclear or undisclosed[64,83].

B Normalised score (%) of the relative usage of NMR spectroscopy as a technique in the discovery and development of molecules which are or have been under clinical studies. **C** Normalised score (%) of the predominant methodologies used in FBDD. **D** Normalised score (%) for the total count of the various NMR techniques used throughout the drug discovery process.

Statistics were derived from publicly available resources, including databases and web blogs, therefore, they could include errors, inaccuracies or be incomplete.

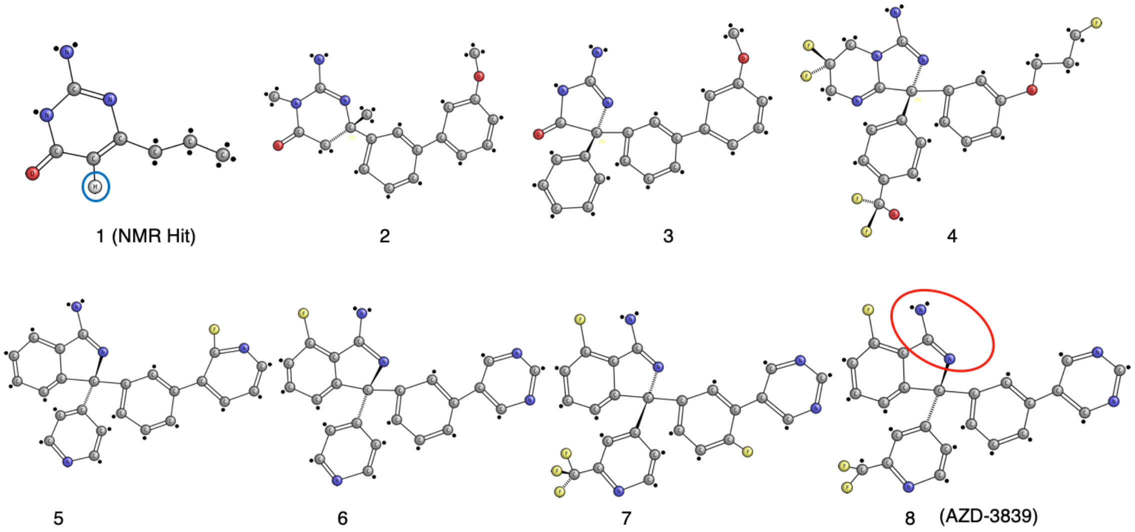


Figure 1.4 Optimisation pathway: from NMR hits to AZD-3839.

Compound **1** represents the initially identified hit from the Water-LOGSY NMR study. The blue circle highlights the isocytosine aromatic proton that was crucial in identifying the hit from the NMR spectrum. Compounds were optimised through a series of crystallographic-based methods to yield the final compound **8** (AZD-3839)[43,84], yet preserving the original amidine motif (red circle) already present in compound **1**.

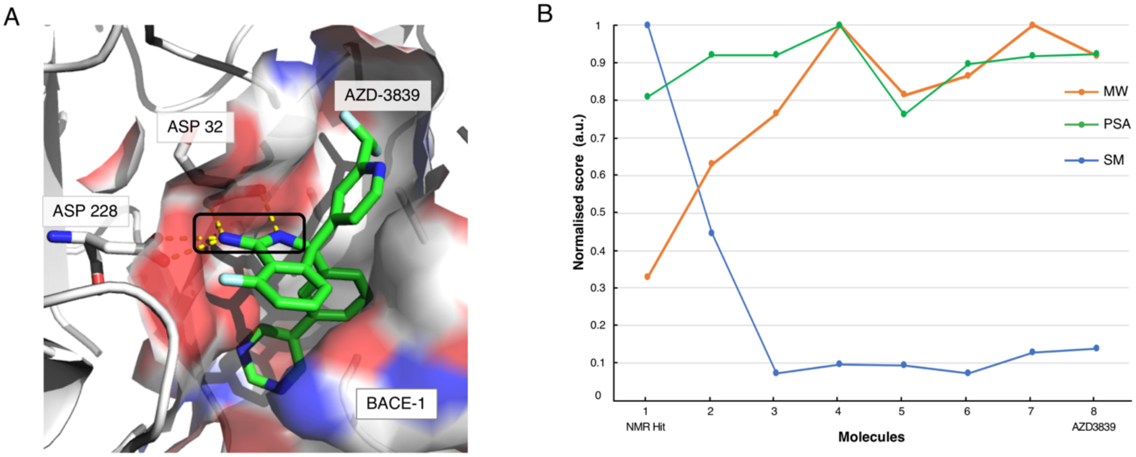


Figure 1.5 Molecular similarities and interactions for AZD-3839 precursors.

A Structure representation of BACE-1 (PDB code: 4B05) and the main interaction between the catalytic groove (Asp32 and Asp228) and the amidine group of AZD-3839, firstly observed in the NMR-discovered hit (black rectangle). **B** Molecular similarity (SM, Blue) "Tanimoto", scaled MW (orange) and PSA (green) scores for the eight compounds on the development path of AZD-3839 (cf. Fig 1.4).

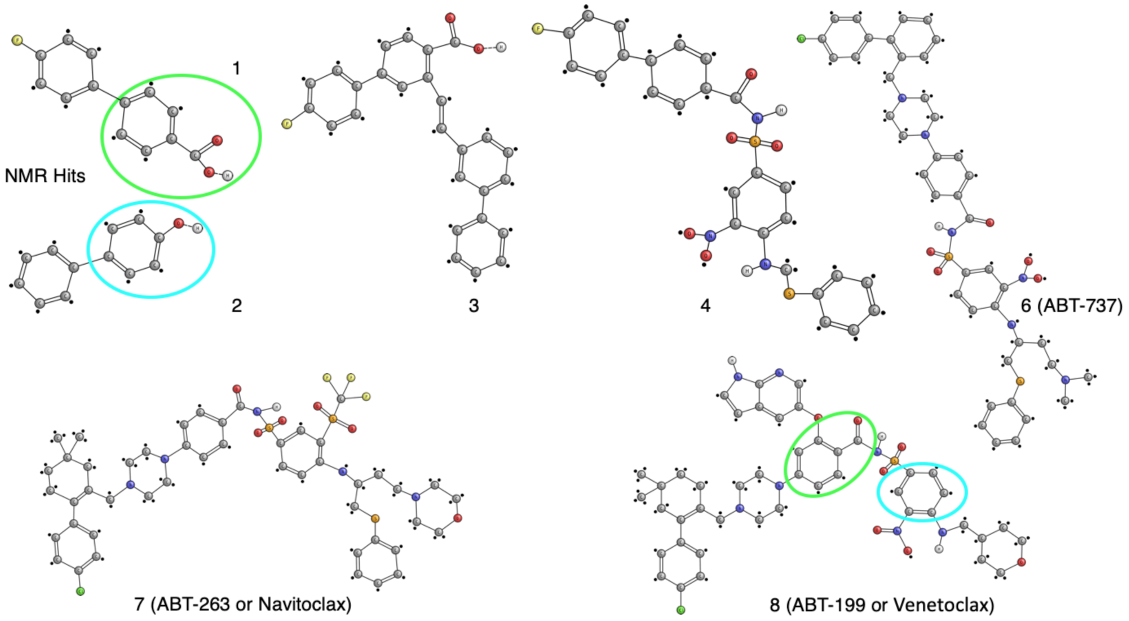


Figure 1.6 Optimisation pathway: from NMR hits to ABT-199.

Hits with aromatic cores (green and cyan circles) were originally identified as interacting with the S1 and S2 of BCL-XL. Compounds 1, 2, 3, 4 were identified and optimised through NMR methodologies, whereas the latest ABT compounds optimisations benefited by X-ray crystallography techniques[49,53,85].

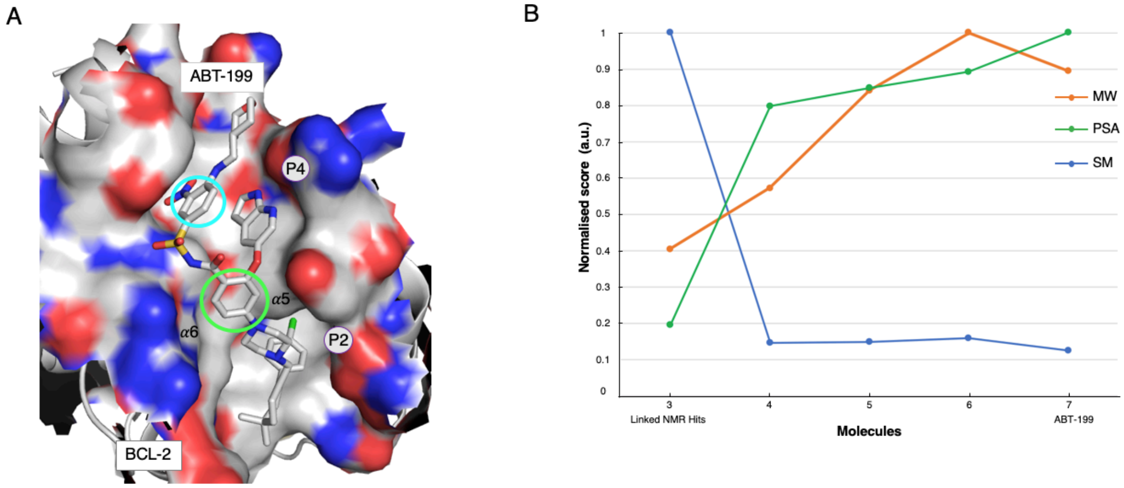


Figure 1.7 Molecular similarities and interactions for AZD-3839 precursors.

A Structure representation of BCL-2 in complex with Venetoclax (PDB code: 6O0L). Green and cyan circles indicated the aromatic motifs originally identified through the NMR primary screening. **B** Molecular similarity (SM, blue) “Tanimoto”, scaled MW (orange) and PSA (green) scores for compounds 3-7 on the development path of ABT-199 (cf. Fig. 1.6).

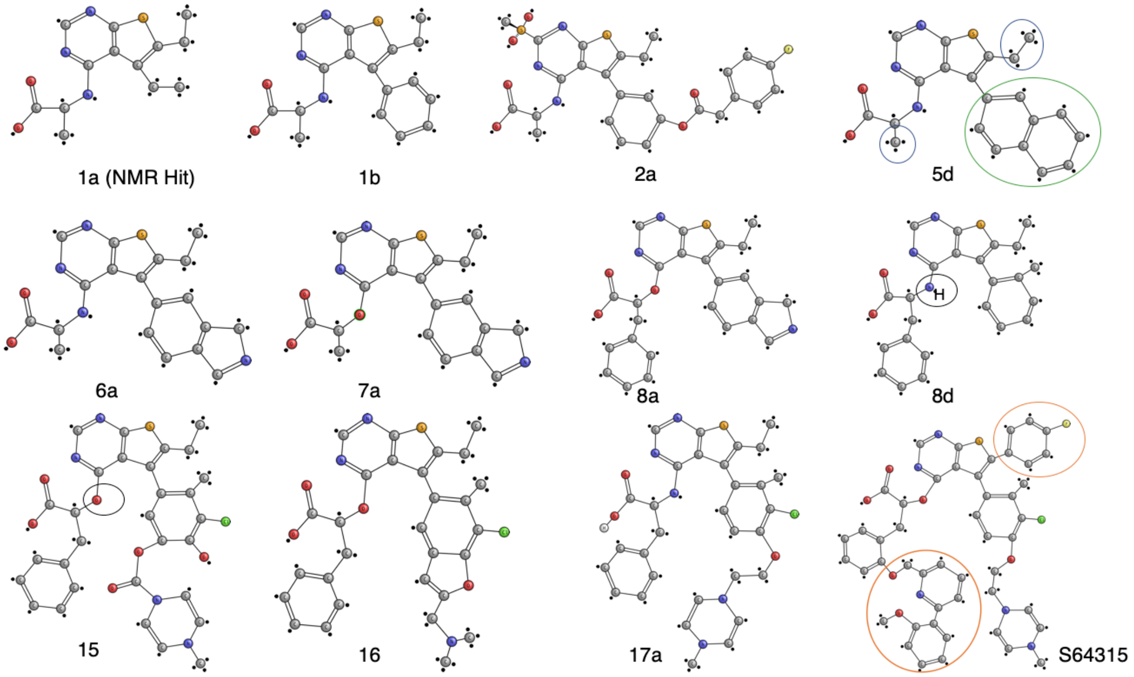


Figure 1.8 Optimisation pathway: from NMR hits to S64315.

All compound nomenclatures are identical to those used in the original manuscript[57].

Compound **1a** represents the initially identified thienopyrimidine core by ligand-detected 1D NMR techniques. The green and blue circles for compound **5d** highlight the chemical groups that gave rise to crucial NOEs for suggesting initial molecule binding poses[57].

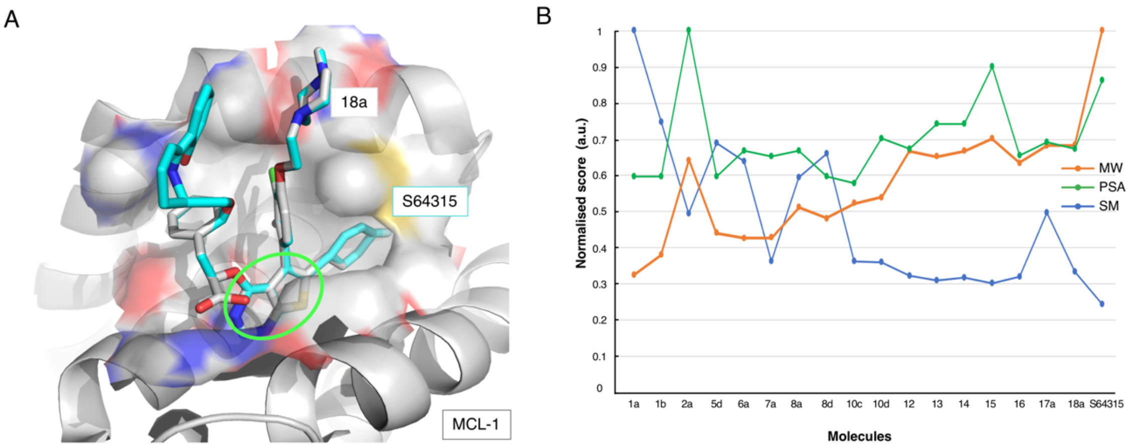


Figure 1.9 Molecular similarities and interactions for S64315 precursors.

A 3D representation of a model of MCL-1 in complex with compound **18a** (PDB code: 6qyo). The green ellipse highlights the original thienopyrimidine motif first identified by an 1D-NMR screening experiment[57]. **B** Molecular similarity (SM, Blue) “Tanimoto”, scaled MW (orange) and PSA (green) scores for the twelve compounds on the development path of S64315 AZD-3839 (cf. Fig 1.8).

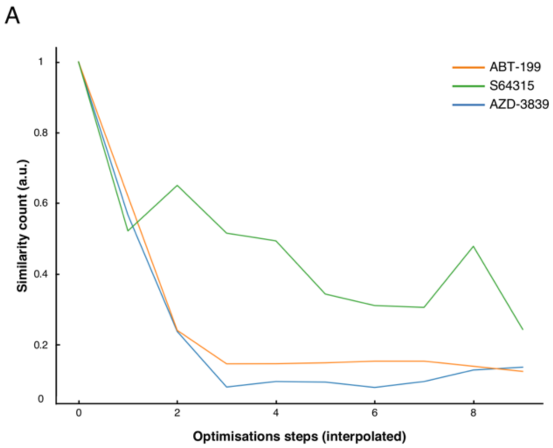


Figure 1.10 Molecular similarities comparison for three case-studies.

A Normalised similarity scores for three different fragment-to-drug developments with interpolated optimisations steps. The orange curve represents the ABT-199 pathway; the green curve represents the S64315 pathway; the blue curve represents the AZD-3839 pathway.

Chapter 2

Design of CcpNmr AnalysisScreen

See List of Publications: 1.

Keywords: Fragments Based Drug Discovery, FBDD NMR, CcpNmr Software, PCA, Mixtures.

2.1. Abstract

Fragment-based drug discovery or FBDD is one of the main methods used by industry and academia for identifying drug-like candidates in early stages of drug discovery. NMR has a significant impact at any stage of the drug discovery process, from primary identification of small molecules to the elucidation of binding modes for guiding optimisations. The essence of NMR as an analytical tool, however, requires the processing and analysis of relatively large amounts of single data items, e.g. spectra, which can be daunting when managed manually. One bottleneck in FBDD by NMR is a lack of adequate and well-integrated resources for NMR data analysis that are freely available to the community. Thus, scientists typically resort to manually inspecting large datasets and relying predominantly on subjective interpretations. In this chapter, I introduce CcpNmr AnalysisScreen, a software package that provides computational tools for automated analysis of FBDD data by NMR. I describe the initial steps and tools required for the hit identification analysis, starting from how data and metadata can be imported and loaded into the main program for the analysis. I then outline how the quality of collected reference spectra can be quickly evaluated using a new dedicated software module. Finally, I illustrate tools I implemented for reducing the amount of time required for the optimal design of NMR samples for screening purposes, which will facilitate the successive data analysis.

2.2. Introduction

Over the years, the versatility of NMR as a non-destructive and adaptable analytical tool has encouraged the development of multiple fragment-based drug discovery (FBDD) approaches by NMR[86]. Nowadays, it is possible, albeit not frequently done, to conduct the entire drug discovery process by NMR (Fig. 1.1). However, as I have shown in chapter 1, the review of the latest FDA approvals and currently tested drugs in clinical studies indicates a substantial contribution of the various NMR-based techniques to the entire drug discovery process. Assuming the target of interest has been already identified, hit identification is the first step in the drug discovery process. This can be achieved by NMR using a number of common ligand-detected NMR methods[86], namely ^{1}H-relaxation-edited (commonly called ^{1}H), Saturation Transfer Difference (STD)[87], Water-LOGSY[88] (Fig. 2.1A), and relaxation experiments ($T_{1\rho}$, T_2). In addition, a number of complementary techniques, i.e. Target Immobilised NMR Screening (TINS)[89], Spin Labels analysis[90], Paramagnetic Relaxation Enhancement (PRE)[91] and ^{19}F experiments[92] have been successfully used in the primary hit identification process. In the next chapter, some of these experiments will be discussed in more detail.

In spite of all the powerful NMR experiments used for NMR-based FBDD[93], inefficient evaluation of the primary hit screening data can disrupt or postpone any of the later phases, such as binding site identification and hit optimisation. Primary screening is routinely performed manually by comparing spectral information derived from thousands of STD, WaterLOGSY and relaxation-edited experiments. Manual analysis of these data inevitable results in human errors or subjective inconsistencies, in addition to problems arising from commonly occurring experimental errors, such as improper alignment and scaling of spectra.

In this chapter I introduce the CcpNmr AnalysisScreen software programme, or AnalysisScreen for short, which is part of the Analysis version-3 software suite[28] (Fig. 2.1).

Analysis version-3 is written in Python 3, a powerful object-oriented programming language. Among the major benefits of Analysis version-3 package is the ability to use numerous third-party software modules and scientific libraries that allow users to exploit complex data manipulation in addition to the already diverse capabilities of Analysis version-3, thus making it an excellent software and methodology development platform[28].

AnalysisScreen maintains the same organisational framework and working areas of CcpNmr AnalysisAssign, called modules. Modules are containers designed to visualise, inspect and perform actions on all varieties of data the project might have; they are discussed throughout the chapters. Furthermore, AnalysisScreen presents a dedicated version of the program suite, which presents an additional section with functionalities in the main menu bar. These functionalities allow for direct access to main screening tasks implemented in the software yet maintain the strategic functionalities of AnalysisAssign (Fig. 2.1B-D).

In this section, I describe the first three fundamental steps and relative computational tools I designed and implemented for performing the ligand-detected screening analysis. Firstly, I introduce a new data-loader mechanism, which allows to parse and a load large quantity of data and metadata that are normally required for screening commercial libraries. The creation of a generic data-loader was essential as each laboratory employs different data structures for storing or representing their spectral information and associated metadata. I then discuss the benefits of using the properties of the principal component analysis (PCA) for assessing spectral quality before carrying out more advanced data manipulation steps. The software also includes dedicated graphical tools for easy inspection, analysis and selection. Finally, I illustrate a strategy for reducing the expensive NMR acquisition time and further spectral deconvolution requirements, by the implementation of a mixture generation tool. This tool aids the design of samples composed of multiple molecules, in a way that NMR minimises spectral peak overlap in order to avoid the appearance of false positives and false negatives in later stages of analyses. The core algorithm was originally described by the NmrMix authors[94] and remained unaltered, but several limiting factors and crucial enhancements were included, such as alleviating the cumbersome reliance on external software packages for the creation of input data as well as a substantial optimisation of speed and extendibility.

CcpNmr AnalysisScreen aims to be the ultimate NMR platform capable of handling multiple tasks and all routines currently available and be flexible enough to easily include any new and emerging methodologies needed for performing the fragment-based drug discovery analysis by NMR in a way that can tackle a variety of user's workflows and requirements.

Code Contributions

The development of all modules which I designed, implemented and included in AnalysisScreen was made possible thanks to the previous work by the CCPN team, such as Dr Fogh, Dr Boucher, Dr Ragan and subsequentially by Dr Brooksbank. They have built the underpinning top core objects and frameworks (Fogh, Ragan, Brooksbank) of CcpNmr Analysis (Fig 2.4), including the generic modules such as the NMR spectra parsers (Boucher), the OpenGL module needed for the visualisation of spectra (Brooksbank) and the core algorithms for the PCA module (Ragan).

From the 31st September 2016 until 31st May 2020, there have been more the 5000 code commits on the CcpNmr Version Control repositories (BitBucket), where the top two committers were Edward Brooksbank and myself with 3120 and 1424 commits, followed by G. Vuister (362), W. Boucher (300), R. Fogh (220) and T.J. Ragan (65).

My commits included all newly developed algorithms and graphical user interface contributions, either for AnalysisScreen and general development, for a total of 67800 new Python lines, including at least 115 new Python files.

2.3. Materials and Methods

2.3.1. Materials

Datasets

Two kinds of datasets were used for testing the importer, decomposition and mixture module. The first type was composed by synthetically generated spectra and the second was kindly provided by an industrial collaborator, referred within the text as "experimental".
The simulated datasets were generated using in-house written scripts (macros) in Python, employing the AnalysisScreen Python environment. Using these macros, I was able to create an arbitrary number of 1D spectral peaks at random positions, heights, and linewidths.
Spectra contained random peaks in the aromatic regions (6-8 ppm), water signal (4-5 ppm) and solvent/ aliphatic signals (4-0 ppm). Each spectrum was linked to a virtual substance or sample, containing random set properties; including SMILES, concentrations, pH, chemical names, etc. All simulated datasets and metadata generated for this work were used only for testing purposes and did not contain any biological significance.
Other spectra, used for testing the decomposition module, were retrieved from the BMRB database[95].

The experimental dataset consisted of a library of 1760 small-molecule compounds, for which a processed one-dimensional reference spectrum was provided in Bruker format. From this library, 1548 fragments had been used to create 310 samples containing four to five, randomly selected small ligands at ~200 µM each and an unnamed target at ~4 µM. A processed STD spectrum for each sample was provided. Although all the crucial data needed for the assessment of the AnalysisScreen routines was available, the biological information and detailed experimental conditions were confidential and not shared with us.

2.3.2. **Methods**

Computational Libraries

AnalysisScreen is written in the Python 3 programming language. The large number of external packages and scientific libraries used within the software are described below by functionalities and usage.

Importers and exporters

Pandas has been used extensively for importing and exporting metadata[32]. Pandas is a Python package providing fast, flexible, and expressive data structures for tabular data. The library provides integrated, intuitive routines for performing common data manipulations and analysis on such datasets[32]. Files can be imported in AnalysisScreen in the format *xls*, whereas, they can be exported in several formats, including *xlsx, xls, csv*, *tsv and json*.

Algorithms and data analysis

Synthetic datasets, implemented algorithms, routines and macros, were written using open-source scientific libraries such as Numpy[33], SciPy[34] and Sci-kit Learn[96] which are included or have been added in the main CcpNmr environment for the specific AnalysisScreen development. NumPy is the fundamental package for scientific computing with Python. It adds support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays. Several critical routines have been speed-optimised using the Numba functionalities[97]. Numba is a powerful tool that translates Python code at runtime using the LLVM compiler[98], improving the execution time by several orders of magnitude.

Graphical user Interface

PyQt5[99], PyQtGraph[100], Matplotlib[101] and Seaborn[102], have been employed for plotting and analysing results as well as for building custom widgets into the main programme.

Widgets represent the building blocks for the final creation of complex modules and pop-ups that are described in this work. They were created mainly using PyQt5, which is a Python binding of the cross-platform GUI toolkit Qt.

Matplotlib and Seaborn have been largely used for plotting one-dimensional spectra during the testing process. PyQtGraph[100] has been used to create custom plots for the decomposition module.

PCA

Principal component analysis, PCA, is a linear conversion of multiple dimensionalities parameters into a lower dimensional space yet maintaining the maximum amount of information about the original input[103].

PCA tools implemented in AnalysisScreen are built using core routines previously developed in our laboratory as part of the development of CcpNmr Analysis Metabolomics. However, I have designed and developed a graphical user interface for facilitating the analysis of large datasets typical for screening.

For analysing the experimental dataset, the entire spectral width was included (-14 to +14 ppm) and the following parameters were used: “Descaling” was enabled; the normalisation method was selected to “PQN”; the centring mode was set to “mean”; finally, the scaling method was set to “Pareto”.

IDE

The CcpNmr program is developed using the integrated development environment (IDE) called PyCharm[104]. The professional edition provided me several precious features and tools needed for the “every-day” code writing, code analysis, profiling, debugging and unit testing. Furthermore, it provides a seamless integration with the major version controls, such as GitHub, which allowed the group to share the code base efficiently and, to smoothly resolve coding conflicts when occasionally we worked simultaneously on common parts of the project (such as GUI widgets).

One of the most useful and used tools of PyCharm in the development of Screen was the CProfile Snapshot. As initially the CcpNmr programme was built to work with a limited amount metadata at the time, e.g. less than 10-20 spectra, it failed to load and/or handle thousands of spectra and metadata associated with industrial screening trials, or to perform algorithms on those spectra in a reasonable amount of time. The statistics and Call Graphs within PyCharm allowed me to inspect and improve/optimise the major bottlenecks present in the code, for example by replacing/reducing nested “for/while loops” with vectorised operations.

Spectral mixtures

The generation of spectral mixtures was implemented using the simulated annealing (SA) algorithm described by Stark *et al.* in the NmrMix software[94]. Although the crucial simulated-annealing algorithm steps were unaltered as in the original package, it has been speed-optimised. It also includes the ability to preserve the best-scored mixtures and includes an option for their use as input for subsequent generations, while retrieving them if ameliorated solutions could not be achieved.

The simulated annealing is one of the most common heuristic methods for optimising model parameters. In general, the method is characterised by searching optimal solutions through several iterations, starting from an “Initial Temperature” to a “Final Temperature”. For each iteration, a solution is randomly generated, and in the case of the mixture generation implementation, spectra are randomly mixed in different pools and total scores are then calculated using the original NmrMix protocol[94]. Broadly speaking, new mixtures will be accepted if a better solution is obtained (a lower score is found) and rejected if a worse solution is encountered (a higher score is found). However, not favourable mixtures are still maintained in the initial phases, but stricter discarded while the process continues or in the algorithm terms: “while the Temperature cools down”. This mechanism allows to reduce the chances of being trapped in a local minimum. In AnalysisScreen, the probability of rejecting or accepting is fine-tuned by the “Temp Constant” parameter.

The total score is given by the summation of single scores for each component[94]. The single component score is based on the proportion of peaks in the spectrum, so that overlaps in spectra containing large quantity of peaks are weighted more compared to spectra with smaller number of peaks. It is defined as:

$$S_c = k\frac{O_c}{P_c}$$

Eq. 2-1

Where Sc is the single score for each component, k is a scaling factor set to 1.0 as default; O_c is the overlap count, two peaks are considered as overlapping if the difference in their chemical shifts are below a threshold value; P_c is the total peak count for the spectrum.

Whereas, the total overlap count and total score is simply given by the sum of the single scores.

Mixtures calculation parameters in the experimental dataset

The total number of mixtures created was 309. The threshold value for considering two peaks overlapped was set to 0.01 ppm. The region of interest was set between 10.0 to 5.0 ppm. Other SA parameters were: Initial Temperature=1e6; Final Temperature=0.01; Max Steps=1e6; Temp Constant=50.0; Cooling Method=Linear; number of iterations=1e6.

Ultimately, spectra were divided in 1, 4 and 10 groups and mixtures were re-generated using the same settings as described above.

2.4. Results and Discussion

2.4.1. **Parsing and importing NMR data and metadata**

Typically, an NMR based FBDD screening experiment requires the handling of a large volume of spectral data and metadata. To address this problem, I included in AnalysisScreen the option to use spreadsheets in an Excel format as a data-loader mechanism, alongside an improved and faster manual loading of multiple spectra, compared to earlier testing versions. The programme can natively read, parse and load files with multiple sheets (Figs 2.1A-C), where column-based keywords define the relevant pieces of information.

Upon parsing and importing into AnalysisScreen, commonly used parameters and information associated within a sample, e.g. different experimental conditions, are immediately available within the sidebar of the AnalysisScreen programme (Fig. 2.3A). All metadata is retained with the relevant CcpNmr object, such as experiment types of spectra or SMILES and other chemical properties of molecules, named Substances in the programme nomenclature. All objects used for screening analysis can also be graphically inspected, edited or deleted using dedicated pop-ups (Figs 2.3B-D).

To further simplify the data analysis preparation, the data loader also includes an automatic path recognition ability so that specifying the absolute spectral data locations is no longer required (Fig. 2.1A). In addition, spectra can be automatically grouped into so-called SpectrumGroups; these are user-defined collections of spectra, designed in such a way that multiple routines can be applied uniformly to all their items (Fig. 2.3C). SpectrumGroups follow the same philosophy of single spectra when it comes to visualisation, and can, therefore, be displayed and manipulated as single entities. Samples, SampleComponents, Substances, SpectrumGroups and SpectrumHits objects are internally connected, forming the underpinning core objects of the AnalysisScreen programme (Fig. 2.4).

2.4.2. **Assessment of spectral quality by PCA decomposition**

Commonly, NMR primary screening studies rely on a collection of one-dimensional spectra acquired for each compound in the screening library, called the reference spectra or reference library. The reference library is typically recorded in an automated fashion and its data are used throughout the analysis. Therefore, ensuring its suitability by filtering out any potentially compromised spectra is essential. Nonetheless, inspecting spectra individually for large libraries can be a time-consuming task.

Principal Component Analysis, PCA[105], can be used for the assessment of spectra, without pre-knowledge of spectral line shapes or other peculiarities. AnalysisScreen offers an integrated PCA decomposition module, capable of effortlessly performing a PCA on large libraries.

I have initially tested the PCA routines using four regular spectra along with the same set but presenting post-processing errors, such as baseline distortions or offsets (Fig. 2.5A). As expected, the resulting first two principal components showed a large variance between the original unmodified spectra clustered in a region around -2 PC1 and -2 PC2, and their modified versions, which appeared as severe outliers with a PC1 distance up to 200 units (Fig. 2.5B). After applying a newly designed baseline correction algorithm, (discussed in chapter 4), spectra were re-assessed by a new PCA analysis (Figs. 2.5C-D). From the new plot, it was clearly visible a great improvement on the overall scoring as outliers vanished and components were clustered in a close proximity to each other (Fig. 2.5D). The promising outputs suggested that the PCA analysis could be applied to experimental datasets as a quick evaluation of spectral artefacts and other post-processing distortions.

The decomposition module was then employed for inspecting the large collaborator dataset, prior to the screening analysis which will be discussed in the chapter 5.

Figure 5 displays the result of a PCA analysis performed on a SpectrumGroup consisting of 1760 experimental reference spectra. The result of this study showed a high variance dispersion among the first two PCA components, enabling quick identification of any outliers. Intriguingly, I could identify several groups of spectra that displayed similar processing defects or other spectral imperfections (Fig. 2.5, sections b, c and d), such as phasing artefacts, inadequate solvent suppression or even the absence of signal data all together. Also, very high values of the Q-Score, a metric commonly used for evaluating variations outside of the PCA model[103], easily identified most of the irregular spectra (Fig. 2.7A). The module also provide the T^2-Scores outputs, a metric used for evaluating variations inside of the PCA model[103] . However, this information was not of particular relevance for this analysis. By plotting the T^2-Scores, outliers were not clearly located in a particular region, but scattered across the sigmoidal curve therefore, I couldn't express a clear conclusion (Fig. 2.7B).

I carefully designed the decomposition module graphical user interface in such a way that it includes multiple features, enabling users to perform a PCA analysis easily and in the shortest time possible (Fig. 2.8A). Either spectra or SpectrumGroups can be loaded via drag and drop from the sidebar; by doing so, results become immediately and automatically available in the tabbed area below. In the first tab is shown as a scatter plot the PCA output for the loaded SpectrumGroup.

Each data-point in the PCA space is linked to its corresponding spectrum, so it can be easily accessed, inspected, removed from the project, or corrected using other tools such as pipes (discussed in the next chapter) present in the package. Successive tabs display principal component vectors, as a 1D plot (Fig. 2.8B) and the global variance, as a scatter plot (Fig. 2.8C). Furthermore, the module allows the possibility of creating new simulated spectra from PC vectors or export the various scores in multiple formats for extended analysis (Fig. 2.8D).

2.4.3. Generation of spectral mixtures with minimal overlaps

Following the quality assessment of the reference library, its reference spectra form the basis for generating mixtures on the basis of their peaks. In fact, to reducing the experimental resources required for NMR-based screening, i.e. samples, NMR time, etc., a common approach is to analyse several compounds simultaneously against a target in a so-called mixture, which should be carefully designed to minimise spectral overlap. Manually generating random mixtures can result in overcrowded spectra, which are difficult to interpret, error prone and time-consuming when it comes to deconvoluting single signal entities to identify possible binders. AnalysisScreen includes optimisation tools that allow the user to create and edit mixtures, thus minimising spectral overlaps. The core engine of the AnalysisScreen mixtures module uses the powerful NmrMix simulated annealing algorithm[94]. However, I significantly boosted the execution speed of the procedure by defining more "isolated" Python functions in simpler numeric terms and decoupling them from the generic module; this allowed me to employ the powerful toolkit provided by the Numba package. Numba, therefore, could convert "on-the-fly" the Python routines to an optimised machine code using the LLVM compiler library[98,106].

The mixture generation tool also guarantees that mixtures and scores are internally preserved during all iterations and eventually the best-scoring solutions are presented to the users. AnalysisScreen can create mixtures *de-novo* starting from reference spectra, but it can also be used to score existing mixtures, such as the one provided by our collaborators. The latter was generated randomly without any further optimisation.

I have initially tested the algorithm using a simulated dataset consisting of 1000 spectra, with randomly generated peaks. Spectra were grouped in 100 random mixtures of 10 components each. Although the original mixtures were created randomly, some mixtures displayed very overcrowded regions (Fig. 2.9A). In contrast, using the mixture generation tool resulted in mixtures that displayed a substantial reduction of peak overlaps (Fig. 2.9B). A further analysis of the initial results showed that the first randomly generated mixtures presented scores in the range of 4 to 6, and outliers over 10 (Fig. 2.10A). Scores, defined as the summation of the total mixtures scores (see Methods and Eq. 2.1), were substantially improved from a total of 515.46 to 468.21 by using a setup that involved ten successive iterations (total score graph not shown). Fig. 2.10 illustrates the optimisation progress, where both the median ("-" symbol) and mean ("x" symbol) were below the control experiment, 5.0 to 4.7 for median and 5.2 to 4.7 for the mean. A further optimisation was also performed starting from the finalised mixtures. The total overlap score was improved to 451 (total overlap graph not shown), and more minimal overlapping mixtures were found, however, the scores distribution was slightly wider than the previous run. This might be caused by running the optimisation using different algorithm parameters, such as the probability of acceptance higher intermediate scores.

As well as the score, AnalysisScreen also reports the total overlap score per mixture and component. The analysis of this dataset showed a reduction in the total overlap score from 899 computed for the control, to 846 for the first run and 834 to the final optimised set (Fig. 2.10B).

Following the validation of the algorithm using simulated spectra, the experimental library of compounds and the spectral reference data were used to generate new experimental mixtures, and these were compared to the previously manually composed ones. This library consisted mainly of aromatic compounds; therefore, the resulting NMR spectra were characterised by crowded downfield regions around 7 ppm. Furthermore, the PCA analysis of the reference spectra of this library revealed a non-optimal solvent suppression in the upfield region (Fig. 2.6). Chemical shifts positions were therefore considered only in the aromatic regions, in line with the collaborator preferences.

I assessed the mixture generation tool with an initial 1000-iterations calculation and calculated the total overlap score for each iteration (Fig. 2.11A). The evolution of the simulation shows the pattern of this stochastic algorithm, with the overlap score reaching several minima just above a value of 1250, which is notably better than value of 1381 obtained for the original randomly created mixtures. However, some iterations displayed considerably inferior values; those solutions were obviously discarded. To assess the influence of the size and the nature of the dataset, I divided our original input into either four or ten random SpectrumGroups and performed the calculations followed by joining the results in a single clustered output. This simple strategy showed a further progressive reduction in total overlaps and scores (Figs 2.11B-D, 2.12A-B). Although this result is somewhat counterintuitive, I speculate that by introducing four or ten random groups, I have increased the overall randomness of the sampling algorithm with respect to relevant spectral regions of interest. Nonetheless, the findings demonstrated the importance of running a large number of iterations to establish an optimal mixture, rather than relying on a few single individual optimisations. Using the automated approach, significantly optimised mixtures were generated when compared to the original randomly generated one. Importantly, I find both a shift to lower values in the distribution of the scores of each mixture as well as a reduction in the number and lowering of the most poorly scoring mixtures, i.e. those with the most problematic overlap. It is to be expected that the latter represent the most challenging mixtures in the analysis of the data (as discussed in chapter 5).

From a visual inspection of different mixtures outputs was possible to notice a large degree of overlap for the manually (randomly) created mixtures (an example of such mixture is shown in Fig. 2.13A), whereas only medium (~2.5 score) or no overlap was observed for other mixtures that were automatically generated and optimised (Figs 2.13B-C).

In addition to the improvements to the core implementation of the algorithm, it is the newly designed graphical interface that affords for major benefits and usability thus truly enabling the power of the algorithm for the creation of quick results that can be immediately transferred to the wet lab.

Starting with the setup, a full set of user-adjustable parameters is available in the Mixture Generation pop-up; this also includes direct access to a built-in peak picker algorithm. The dedicated section provides a list of pre-set solvent regions (in the ^{1}H ppm scale) that can be used as a reference for defining custom regions of exclusions from the calculations (Figs 2.14A-B). After the process is completed, results can be easily inspected and edited within an interactive module, called Mixture Analysis (Fig. 2.14D). The driving element displayed in the module is the main scoring table. Each selection controls the right part of the window where a series of tabs display relative information regarding the mixtures, and which allow for a dynamic inspection of spectra within a mixture.

A summary of the 2D molecular structures and other properties is also available as dedicated tabs (Figs 2.15A-B), which allows for a manual assessment of mixtures from the chemical point of view. Most importantly, the module allows the user to manually edit the final mixtures by simple drag-and-drop operations. This tab also indicates a single score for each component thus allowing the most overlapping item to be identified and the mixture to be modified accordingly. Obviously, this could potentially be counterproductive for a large dataset. For this reason, an extra optimisation module is present (Fig. 2.15D). The module allows to perform a new optimisation run starting from the current mixtures; the user can subsequently accept or reject the newly created mixtures.

2.5. Conclusions

With numerous techniques developed over the years, NMR has been invaluable in all stages of FBDD leading to promising drug-like molecules[107].

The versatility of NMR spectroscopy has made it possible to tackle all aspects of drug discovery, with the 1D primary screening the most employed technique of all available possibilities (cf. chapter 1)[86]. However, NMR data analysis, especially when containing a large number of experimental spectra that need to be examined, can be daunting and time-consuming. The vast amount of data generated for each screening trial and the lack of freely available software capable of dealing with this data leaves scientists setting up and repeating tiresome operations that could inadvertently lead to human errors. Moreover, users might rely only on qualitative assessments, which can further increase the probability of misinterpreting the data. In this chapter, I introduced CcpNmr AnalysisScreen, a software developed specifically for analysing fragment-based drug discovery data derived by NMR spectroscopy.

AnalysisScreen is able to cope with very large datasets, with a magnitude of tens of thousands of one-dimensional spectral entries and associated metadata, including projects with over 1 million peaks, providing fast and reproducible results. This was made possible by optimising the underpinning core functionalities of the main package CcpNmr Analysis, which was originally designed for dealing only with a limited number of spectra, typically necessary for macromolecules assignments[28]. The design and creation of a MS Excel reader has further enhanced the metadata handling capabilities of AnalysisScreen (Figs 2.2-2.3).

I have shown the utility of performing PCA analysis on large datasets and discussed the graphical tools I have designed and included in the software.

Indeed, by using the decomposition module as a quick quality control method, entire reference spectral libraries can be evaluated before performing the screening analysis, thus alleviating using potentially compromised spectral data (Figs 2.5-2.8).

Following the acquisition of reference spectra for each compound in the library, it is a common practice to prepare samples containing a combination of components, usually in cocktails of 5-10 molecules[6]. Although this procedure might save NMR time as well as reducing the amounts of target protein required by generating and collecting data for fewer samples, often it results in an increased analysis time for data analysis. This increased data analysis time originates from the need of deconvolving single signal components from crowded spectra or worst, the necessity of re-performing the entire screening trial with fewer components per mixture, maybe to the point of a single compound at the time.

In order to overcome this problem, I have included in the package a tool to create automatically spectral mixtures that allows to minimise the total overlap. The tool uses the simulated annealing algorithm as introduced in the NmrMix programme[94]. It enabled me to create mixtures directly from large simulated and experimental datasets, that presented a significant reduction in spectral overlaps compared to randomly manually created pools (Figs 2.9-2.13). Afterwards, I discussed graphical tools I designed and implemented for the generation and optimisation of mixtures within the AnalysisScreen package aided to increase the productivity compare to the original software NmrMix (Figs 2.14-2.15).

From a programmatic point of view, both for the mixture generation as well as the decomposition module, it has been crucial to separate the development of the core algorithms from the GUI routines. Thus, new filters and scoring functions could be easily added to the original algorithms and together this setup ensures the future maintainability of the platform.

An example of a future enhancement for the mixture generation algorithm will be the introduction of scoring functions based on chemical properties of the compounds in the mixture, such as chemical structural diversity or pKa. The latter is essential in determining possible adverse reactions among the compounds which could lead, for instance, to molecular aggregation, or to a series of other issues that ultimately might give the rise to spectral artefacts or fallacious outputs in the subsequent data analyses.

In the next chapter, I discuss the common NMR 1D screening methods and the software architecture which has been developed for assisting in the analyses of spectra derived by these techniques.

2.6. Figures

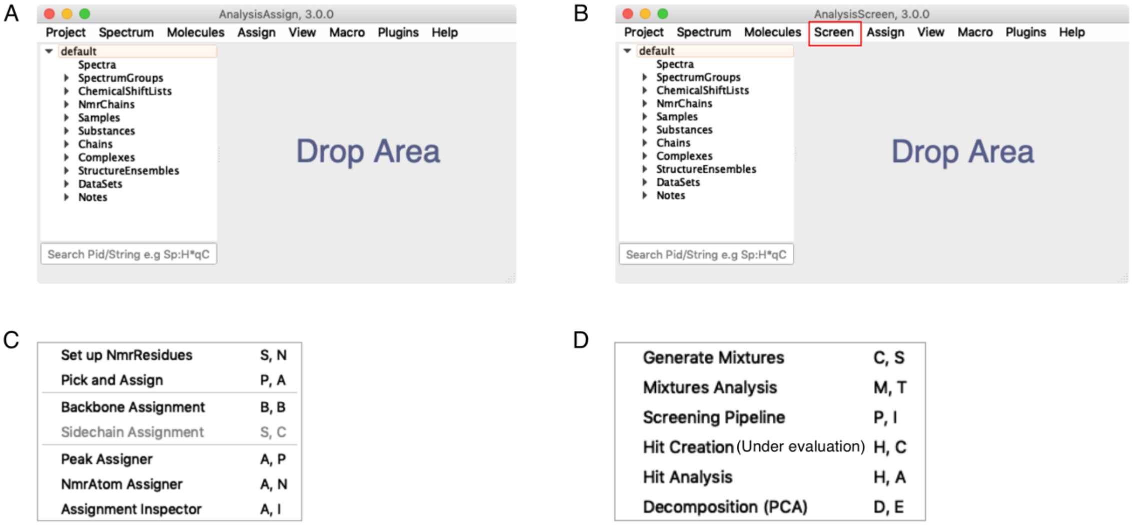
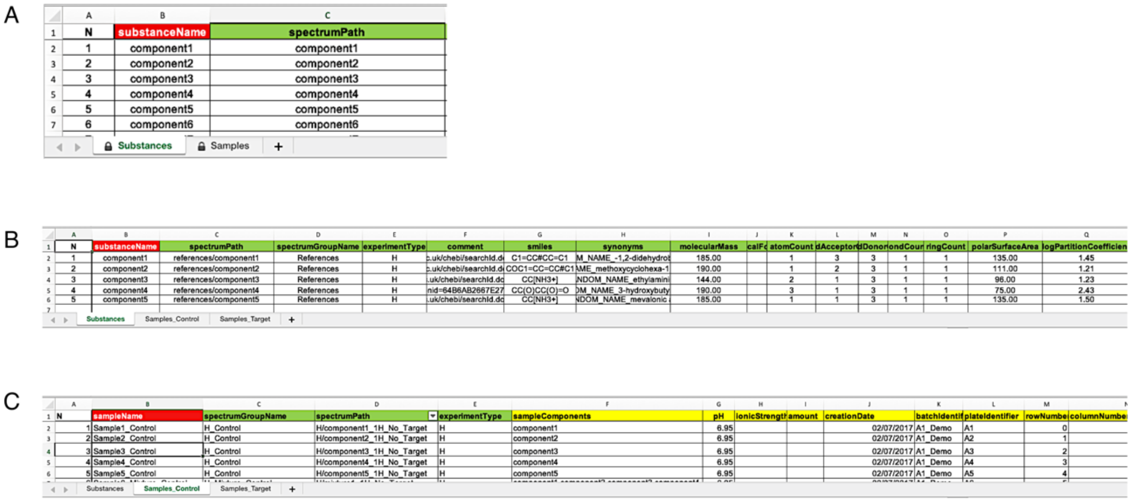


Figure 2.1 CcpNmr Analysis Version 3.0.

A Main window for AnalysisAssign 3.0, standard application for Analysis version-3. **B** Main window for AnalysisScreen 3.0; an additional menu entry present in this application is highlighted by the red rectangle. **C** Main menu listing functionalities for AnalysisAssign. **D** List of extra functionalities available for AnalysisScreen.



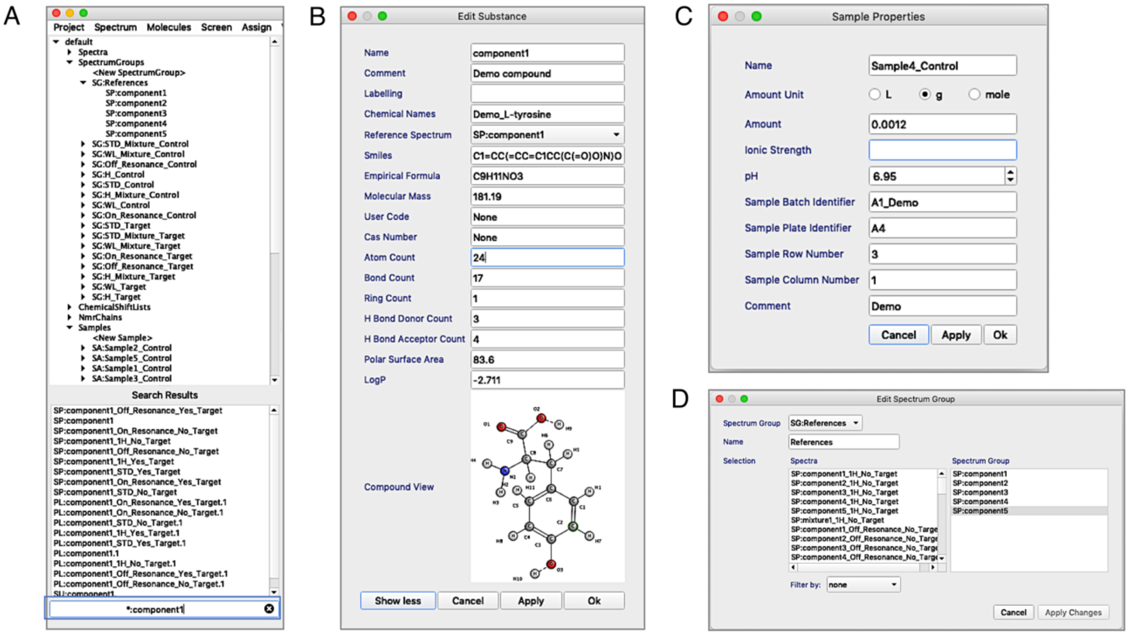
N	substanceName	spectrumPath
1	component1	component1
2	component2	component2
3	component3	component3
4	component4	component4
5	component5	component5
6	component6	component6

N	substanceName	spectrumPath	spectrumGroupName	experimentType	comment	smiles	synonyms	molecularMass	calFc	atomCount	dAcceptor	dDonor	ondCoun	ringCount	polarSurfaceArea	logPartitionCoefficien
1	component1	references/component1	References	H	c.uk/chebi/searchId.d	C1=CC#CC=C1	M_NAME_-1,2-didehydrot	185.00		1	3	3	1	1	135.00	1.45
2	component2	references/component2	References	H	c.uk/chebi/searchId.d	COC1=CC=CC#C1	AME_methoxycyclohexa-1	190.00		1	2	3	1	1	111.00	1.21
3	component3	references/component3	References	H	.uk/chebi/searchId.dc	CC[NH3+]	NDOM_NAME_ethylamini	144.00		2	1	3	1	1	96.00	1.23
4	component4	references/component4	References	H	nid=64B6AB2667E27	CC(O)CC(O)=O	DM_NAME_3-hydroxybuty	190.00		3	1	3	1	1	75.00	2.43
5	component5	references/component5	References	H	.uk/chebi/searchId.dc	CC[NH3+]	NDOM_NAME_mevalonic	185.00		1	1	3	1	1	135.00	1.50

N	sampleName	spectrumGroupName	spectrumPath	experimentType	sampleComponents	pH	ionicStrength	amount	creationDate	batchIdentif	plateIdentifier	rowNumbe	columnNumbe
1	Sample1_Control	H_Control	H/component1_1H_No_Target	H	component1	6.95			02/07/2017	A1_Demo	A1	0	
2	Sample2_Control	H_Control	H/component2_1H_No_Target	H	component2	6.95			02/07/2017	A1_Demo	A2	1	
3	Sample3_Control	H_Control	H/component3_1H_No_Target	H	component3	6.95			02/07/2017	A1_Demo	A3	2	
4	Sample4_Control	H_Control	H/component4_1H_No_Target	H	component4	6.95			02/07/2017	A1_Demo	A4	3	
5	Sample5_Control	H_Control	H/component5_1H_No_Target	H	component5	6.95			02/07/2017	A1_Demo	A5	4	

Figure 2.2 Excel file loading examples.

A Example of an excel file the program can read, parse and load. Files need to include the words "Sample" or "Substance" in the sheet names to activate AnalysisScreen reading capabilities. Spectra path are automatically recognised by only providing the spectrum name in the relative cell. **B** Substance, can contain metadata associated with small molecules, including the relative spectrum path for the processed spectrum used as references for a screening trial. The highlighted red column header indicates the mandatory field. **C** Sample Excel sheet; these sheets contain all metadata associated with particular samples and their components. For example, in a screening study the sample might contain multiple spectra recorded in different experimental conditions or experiment types.



default
Spectra
SpectrumGroups
<New SpectrumGroup>
SG:References
SP:component1
SP:component2
SP:component3
SP:component4
SP:component5
SG:STD_Mixture_Control
SG:WL_Mixture_Control
SG:Off_Resonance_Control
SG:H_Control
SG:STD_Control
SG:H_Mixture_Control
SG:WL_Control
SG:On_Resonance_Control
SG:STD_Target
SG:STD_Mixture_Target
SG:WL_Mixture_Target
SG:On_Resonance_Target
SG:Off_Resonance_Target
SG:H_Mixture_Target
SG:WL_Target
SG:H_Target
ChemicalShiftLists
NmrChains
Samples
<New Sample>
SA:Sample2_Control
SA:Sample5_Control
SA:Sample1_Control
SA:Sample3_Control

Search Results

SP:component1_Off_Resonance_Yes_Target
SP:component1
SP:component1_On_Resonance_No_Target
SP:component1_1H_No_Target
SP:component1_Off_Resonance_No_Target
SP:component1_1H_Yes_Target
SP:component1_STD_Yes_Target
SP:component1_On_Resonance_Yes_Target
SP:component1_STD_No_Target
PL:component1_On_Resonance_Yes_Target.1
PL:component1_On_Resonance_No_Target.1
PL:component1_STD_No_Target.1
PL:component1_1H_Yes_Target.1
PL:component1_STD_Yes_Target.1
PL:component1.1
PL:component1_1H_No_Target.1
PL:component1_Off_Resonance_Yes_Target.1
PL:component1_Off_Resonance_No_Target.1
SU:component1

*:component1

Name: component1
Comment: Demo compound
Labelling:
Chemical Names: Demo_L-tyrosine
Reference Spectrum: SP:component1
Smiles: C1=CC(=CC=C1CC(C(=O)O)N)O
Empirical Formula: C9H11NO3
Molecular Mass: 181.19
User Code: None
Cas Number: None
Atom Count: 24
Bond Count: 17
Ring Count: 1
H Bond Donor Count: 3
H Bond Acceptor Count: 4
Polar Surface Area: 83.6
LogP: -2.711

Name: Sample4_Control
Amount Unit: ○ L ● g ○ mole
Amount: 0.0012
Ionic Strength:
pH: 6.95
Sample Batch Identifier: A1_Demo
Sample Plate Identifier: A4
Sample Row Number: 3
Sample Column Number: 1
Comment: Demo

Spectrum Group: SG:References
Name: References
Selection
Spectra:
SP:component1_1H_No_Target
SP:component2_1H_No_Target
SP:component3_1H_No_Target
SP:component4_1H_No_Target
SP:component5_1H_No_Target
SP:mixture1_1H_No_Target
SP:component1_Off_Resonance_No_Targe
SP:component2_Off_Resonance_No_Targe
SP:component3_Off_Resonance_No_Targe
SP:component4_Off_Resonance_No_Targe
Spectrum Group:
SP:component1
SP:component2
SP:component3
SP:component4
SP:component5
Filter by: none

Figure 2.3 CcpNmr AnalysisScreen sidebar and various pop-ups.

A Screenshot of the sidebar state after parsing and loading an Excel file containing spectral metadata. Objects are automatically created and are listed on various branches. The regex-enabled search widget (blue rectangle) allows for quick scanning of project metadata through the tree, an essential feature when handling several hundred entries of a typical NMR screening dataset. **B** Small molecule metadata are stored into the CcpNmr software as Substances. Substances are a representation of chemical properties of the reference compound. They can be visualised and edited in the Substances pop-up. If SMILES are provided, molecular structures are also shown in this window. **C** The Samples properties pop-up enables users to insert and edit information regarding particular experimental conditions, such as concentration and pH or other sample identifiers. **D** The SpectrumGroup editor pop-up allows users to quickly and easily group spectra using drag-and-drop features. SpectrumGroups can be displayed as single entities in displays or be used as input data for several tools throughout the programme.

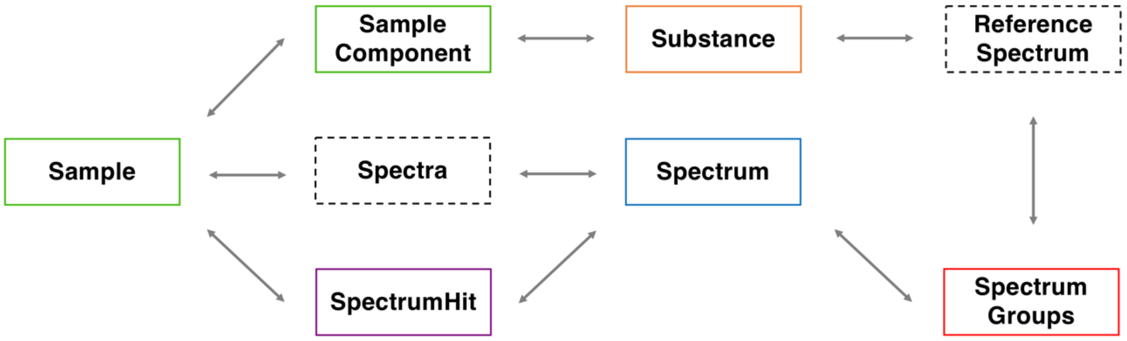


Figure 2.4 CcpNmr core objects used in AnalysisScreen.

Core objects ensure data is accurately maintained across multiple invocations of the programme; furthermore, the linkage among various items creates a robust data-access strategy for implementation of the data analysis routines available in AnalysisScreen.

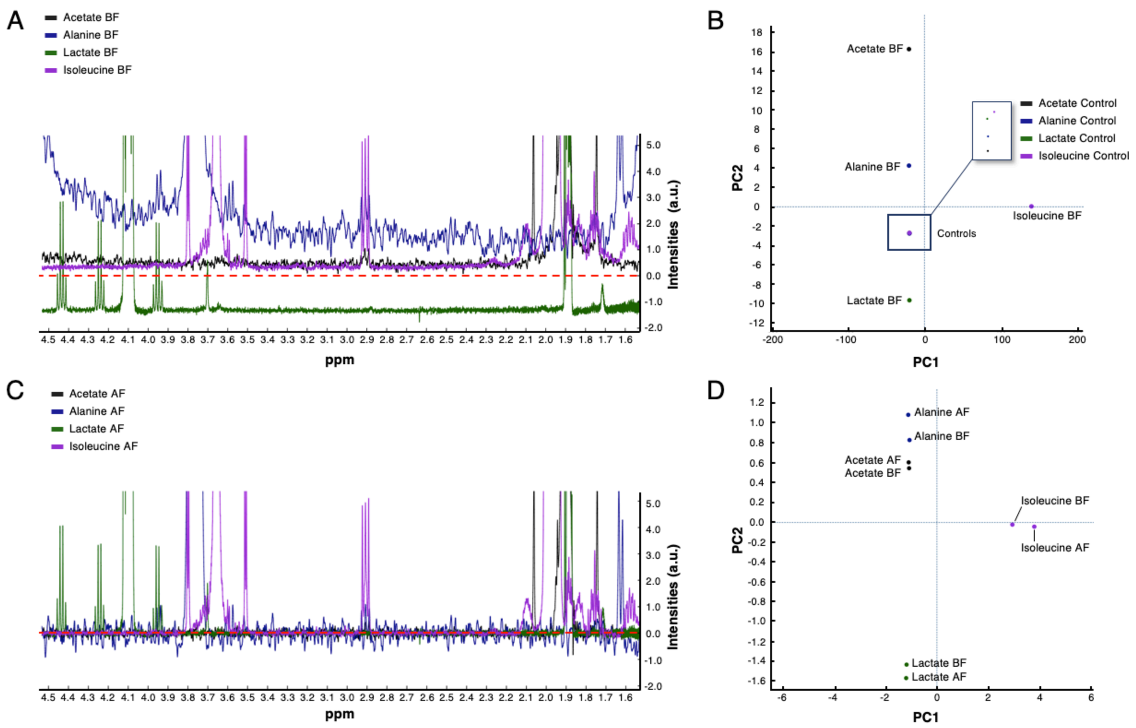


Figure 2.5 PCA on a testing BMRB dataset.

A 1D spectra overlays for acetate (dark blue), alanine (blue), lactate (green), isoleucine (purple), downloaded from the BMRB database[95] and showing a baseline offset. The standard and expected baseline is represented by the red dotted line. **B** First two principal components (PC1, PC2) of the PCA for the same spectra (shown in **A**) together with the original unmodified references used as controls (without any baseline distortion, spectra not shown in **A**). The blue square indicates the control spectra cluster, composed by the original spectra and presenting the same colours as the modified spectra (insertion box); whereas, the remaining components represent the modified spectra which appear in the PCA plot as outliers. **C** Spectra after applying a baseline correction algorithm. **D** First two principal components (PC1, PC2) of the PCA recalculated after applying the spectral correction together with controls. Values for PC1 and PC2 for all spectra are now clustered in close proximity, showing minor variance between before and after the correction.

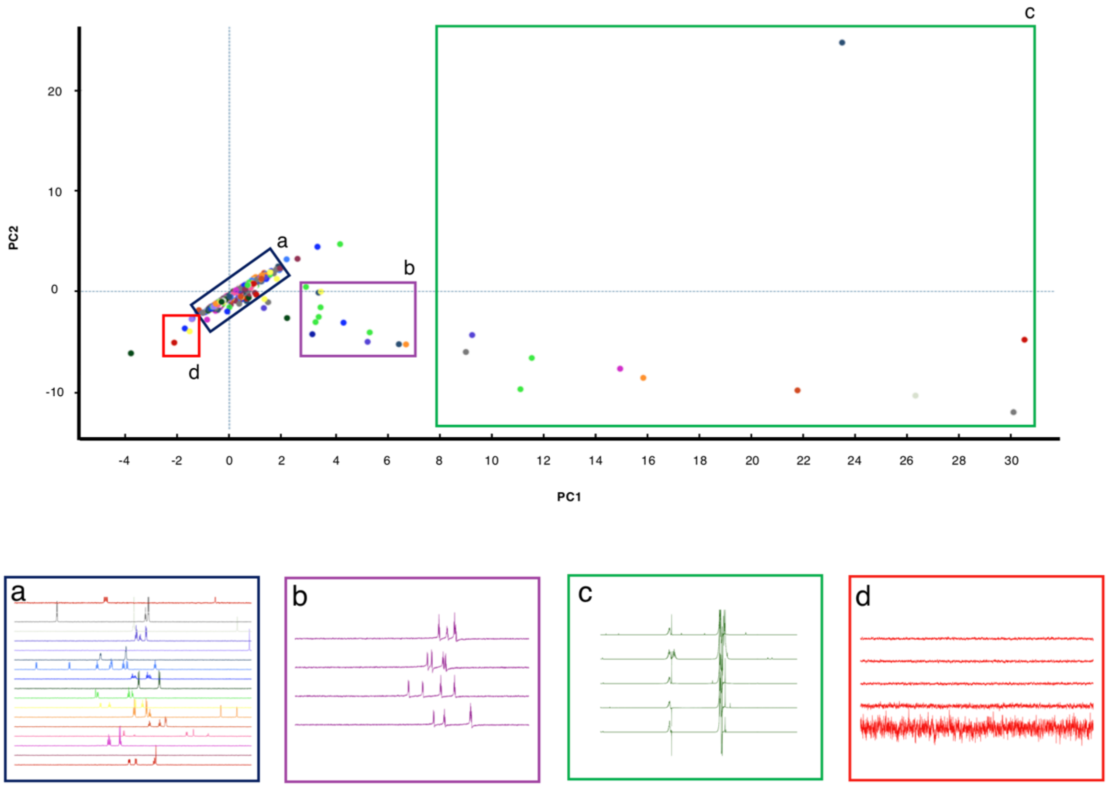


Figure 2.6 Principal component analysis (PCA) of 1760 reference spectra. Most of spectra were uniformly grouped around PCA origins, (blue rectangle, panel **a**); for spectra in the region 3 < PC1 < 7 (purple rectangle, panel **b**) large phasing errors were observed; the spectra in the region PC1 > 8 (green rectangle, panel **c**) appeared highly distorted, probably due to inadequate solvent suppression. Finally, spectra presenting only noise were discovered in the region indicated by the red square (panel **d**).

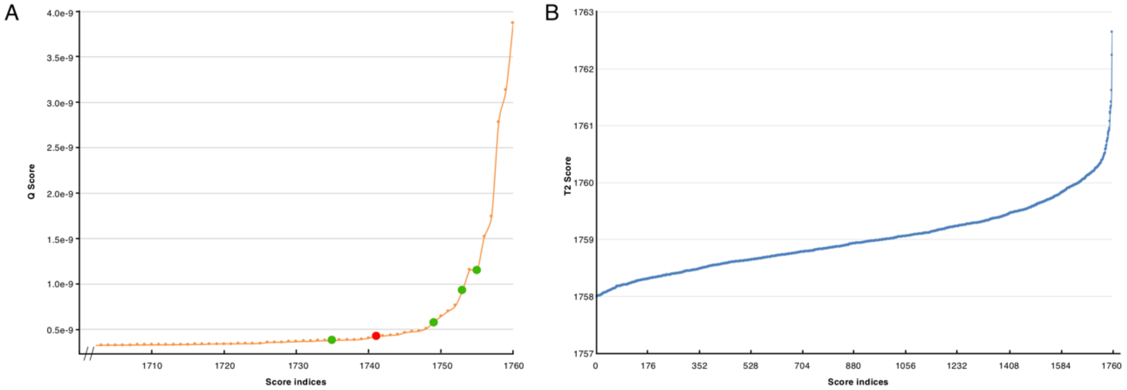
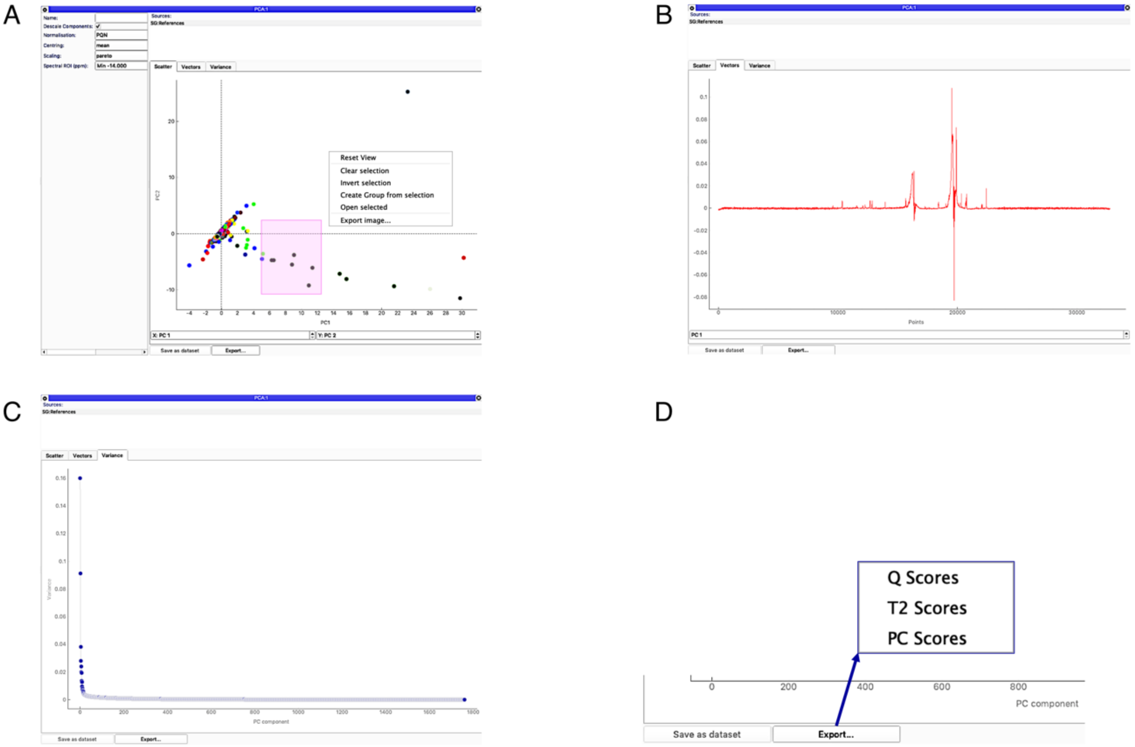


Figure 2.7 PCA Q-Scores and T2-Scores for experimental dataset.

A PCA Q-Scores. Q-scores sorted in descending mode, showing the last 60 values for the experimental dataset. The highest scores were associated with spectral issues, including solvent suppression distortions; highlighted in green and red, some examples showed in Fig. 2.6 (panel c and d). **A** T^2-Scores for the dataset. Previously identified outliers were found across the curve and not in a specific range as in the case of the Q-Scores curve.



Name:
Descale Components: ☑
Normalisation: PQN
Centring: mean
Scaling: pareto
Spectral ROI (ppm): Min -14.000

Sources:
SG:References

Figure 2.8 Decomposition module.

A Main window for the PCA module, the left panel displays the available settings for performing the calculations; top area is the input data section where spectra or a SpectrumGroup can be dropped to start the machinery; the central area displays PCs scores. The pink box shows a selection command and, the adjacent context menu shows possible actions that can be performed on selected items. **B** Screenshot of the PC vector tab display; this section allows to navigate through all resulting data generated by the PCA analysis. The first principal component for the experimental dataset is displayed; the poor solvent suppression is clearly noticeable. **C** The PCA variance tab; first few components show the maximum variance for the dataset. **D** Extra functionalities available in the module for an in-depth analysis of the datasets.

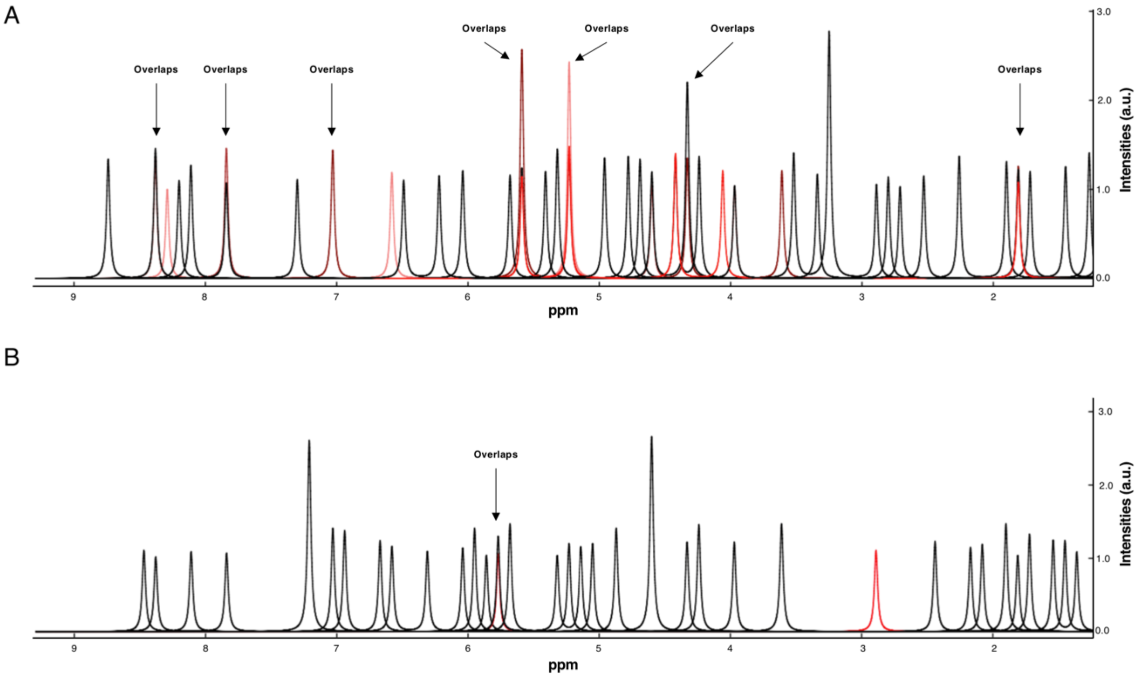


Figure 2.9 Mixtures generation on simulated datasets.

A Randomly created mixture of 10 simulated spectra presenting several regions of overlapped peaks (red spectra and arrow-pointed). **B** An optimised mixture recreated after 10 run of simulated annealing calculation. Only one overlapped region was recorded for the ten spectral pool.

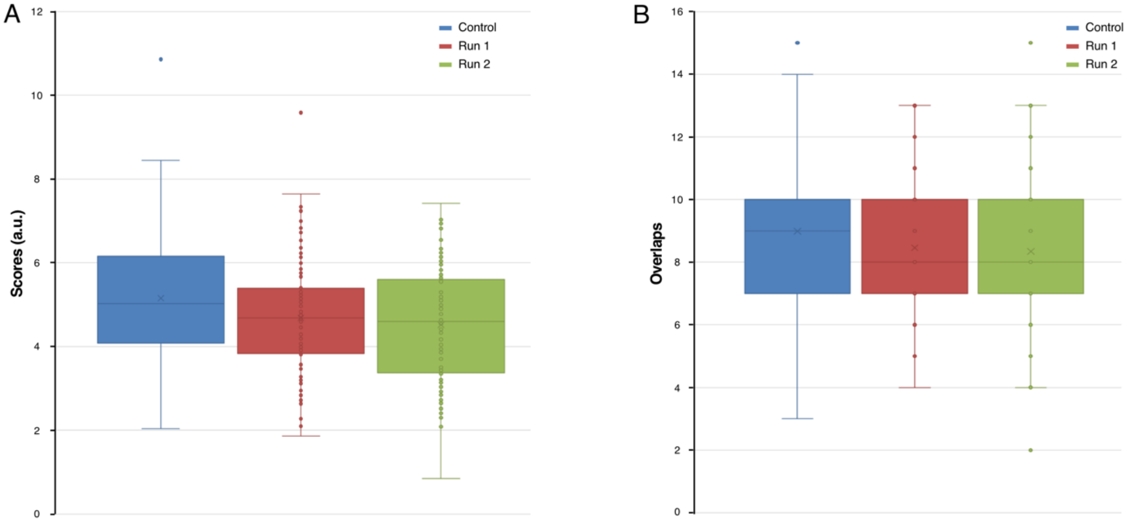


Figure 2.10 Mixtures analysis on simulated datasets.

A Box plots of the score distributions for 100 mixtures, for the control (blue), after 10 iterations for run 1 (red), and after an optimisation on the run 2 (green). The rectangular boxes represent the interquartile range (IQR); the “X” symbol inside the IQR represents the mean; long horizontal bar in the middle of the dataset represents the median (second quartile, Q2), the area below and above indicates the first (Q1) and the third quartile (Q3). Q1, Q2 and Q3 are also referred as 25th, 50th, 75th percentile. Finally, circles indicate outliers in the distributions. The maximum is calculated as Q3 +1.5*IQR and minimum as Q1-1.5*IQR[108]. **B** Overlap count represented as a standardised distribution described as previously.

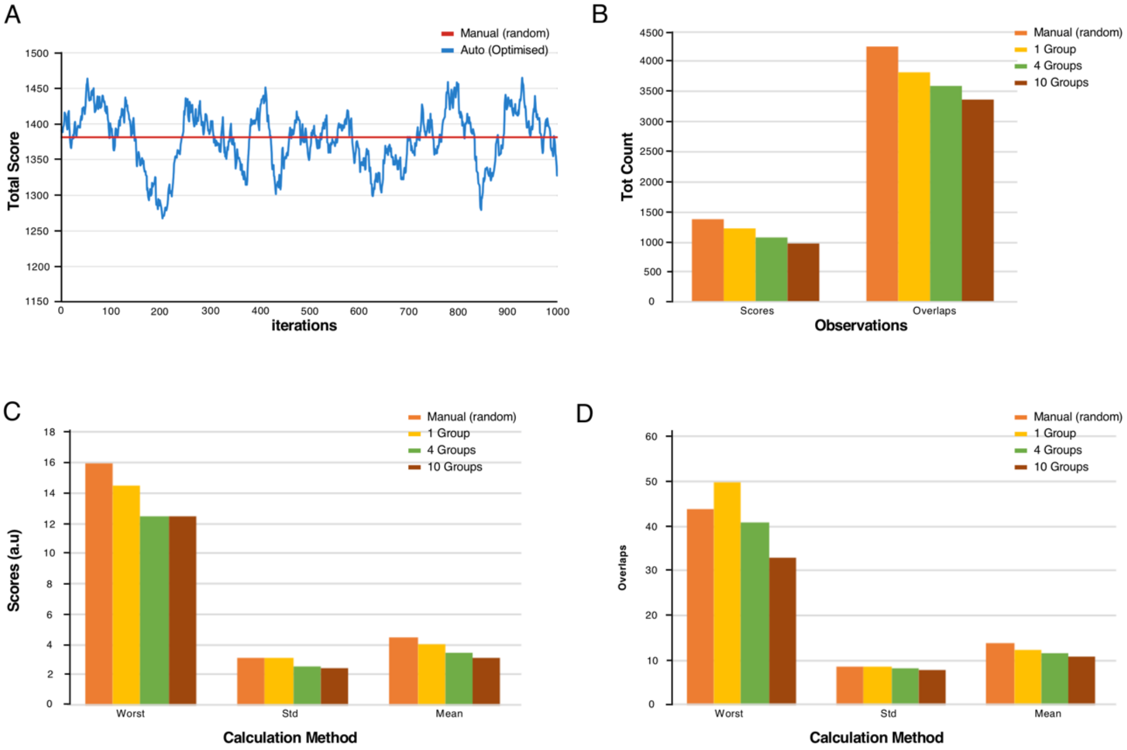


Figure 2.11 Mixtures analysis on experimental datasets.

A Evolution of the total overlap score over 1000 simulated-annealing iterations of the mixtures generation algorithm using 1548 library reference spectra as input data. The red line represents the total overlap score derived from manual randomly created mixtures. **B** Total overlap scores and overlap counts for manual randomly created mixtures (orange), automatically generated and optimised using the whole dataset as input (yellow) or automatically generated on the basis of 4 (green) or 10 groups (brown). **C** Statistical scores for the various groups, indicating the “worst” as the mixture presenting the highest score; the standard deviation (std) and average (mean). **D** Overlaps statistical analysis described in the same terms as **B.**

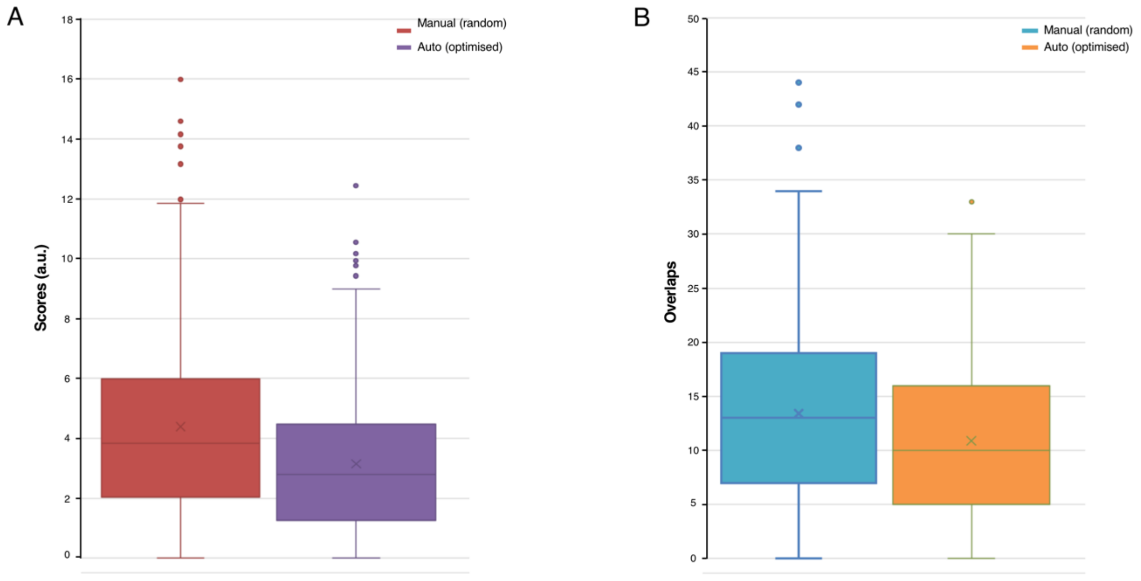


Figure 2.12 Mixtures performance summary on experimental datasets.

A The box plot summaries the score distributions for 310 mixtures manually created (dark red) and for 309 (+1 empty) mixtures after subdividing the dataset in 10 groups and performing a 1000 iterations (purple). **B** Total number of overlaps for the original randomly created mixtures and for the new optimised mixtures generated by the mixture generation module. Overlaps and other mixture scores were calculated as in NmrMix[94].

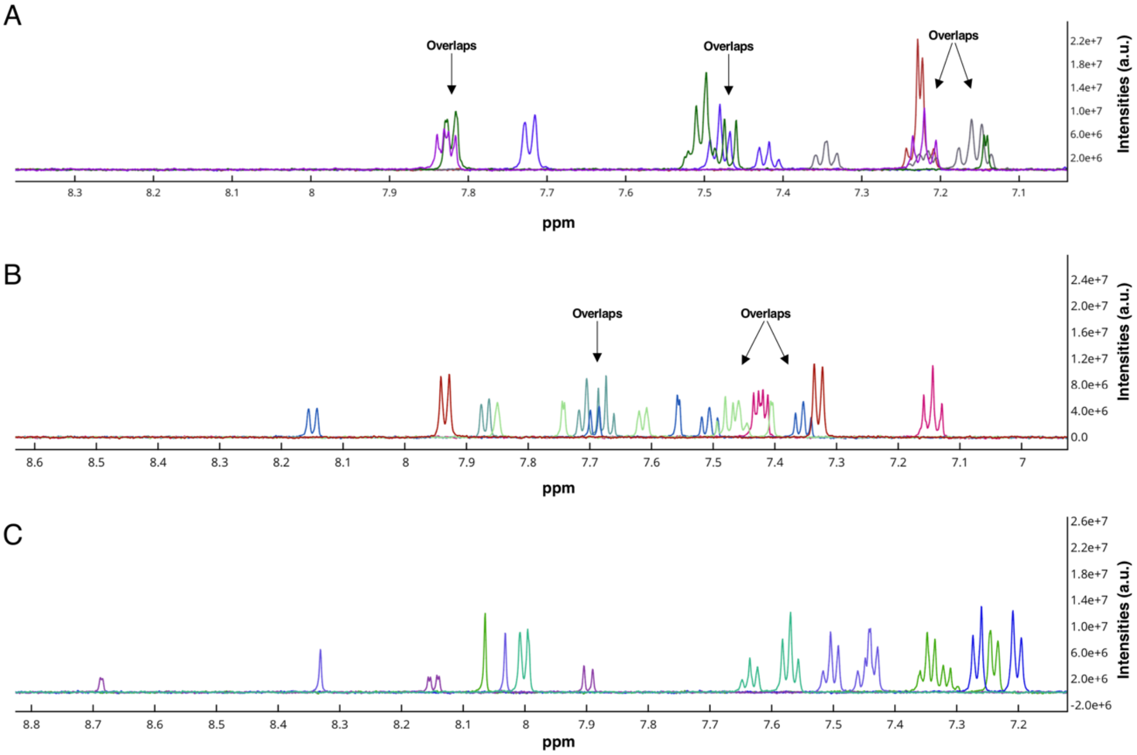
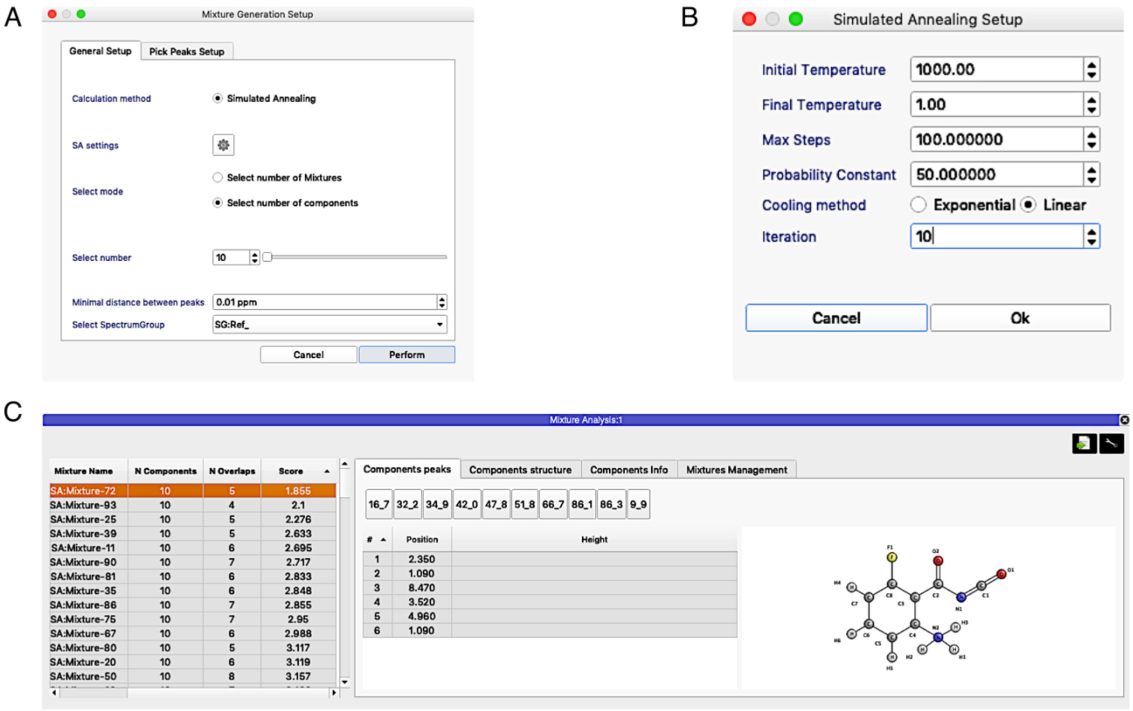


Figure 2.13 Experimental mixtures examples.

A Example of a highly overlapped mixture, manually randomly created, with a calculated overlap score of ~16. **B** Example of a medium overlapped mixture, automatically generated and optimised, with a calculated overlap score of ~2.5; **C** Example of an optimal mixture, automatically generated and optimised, with a calculated overlap score of 0. Arrows indicates regions of maximum overlaps.



Calculation method: ● Simulated Annealing

SA settings

Select mode: ○ Select number of Mixtures ● Select number of components

Select number: 10

Minimal distance between peaks: 0.01 ppm

Select SpectrumGroup: SG:Ref_

Initial Temperature: 1000.00

Final Temperature: 1.00

Max Steps: 100.000000

Probability Constant: 50.000000

Cooling method: ○ Exponential ● Linear

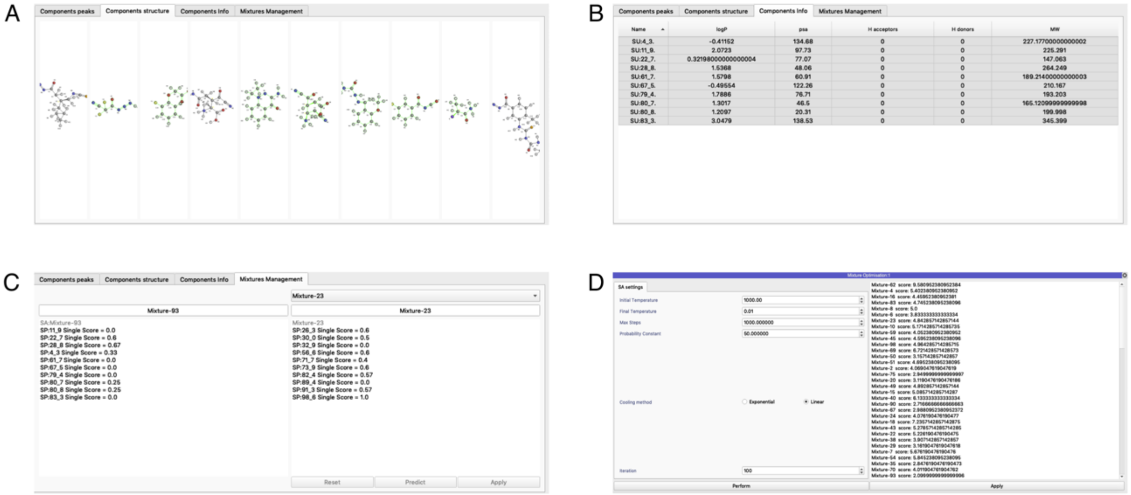
Iteration: 10

Mixture Name	N Components	N Overlaps	Score
SA:Mixture-72	10	5	1.855
SA:Mixture-93	10	4	2.1
SA:Mixture-25	10	5	2.276
SA:Mixture-39	10	5	2.633
SA:Mixture-11	10	6	2.695
SA:Mixture-90	10	7	2.717
SA:Mixture-81	10	6	2.833
SA:Mixture-35	10	6	2.848
SA:Mixture-86	10	7	2.855
SA:Mixture-75	10	7	2.95
SA:Mixture-67	10	6	2.988
SA:Mixture-80	10	5	3.117
SA:Mixture-20	10	6	3.119
SA:Mixture-50	10	8	3.157

#	Position	Height
1	2.350	
2	1.090	
3	8.470	
4	3.520	
5	4.960	
6	1.090	

Figure 2.14 Mixtures generation setup and graphical tools.

A Graphical user interface for the main calculation pop-up window. **B** The current pop-up allows users for a fine adjustment of simulating-annealing algorithm parameters. The figure shows the parameters used for running the first calculation on the simulated dataset, 1000 steps of randomly created mixtures times 10 total iterations. **C** Mixture analysis module showing the newly generated mixtures. The selected tab, *Components peaks* shows a custom peak table, and the 2D structure for the molecule under examination.



Name	logP	psa	H acceptors	H donors	MW
SU:4_3.	-0.41152	134.68	0	0	227.17700000000002
SU:11_9.	2.0723	97.73	0	0	225.291
SU:22_7.	0.32198000000000004	77.07	0	0	147.063
SU:28_8.	1.5368	48.06	0	0	264.249
SU:61_7.	1.5798	60.91	0	0	189.21400000000003
SU:67_5.	-0.49554	122.26	0	0	210.167
SU:79_4.	1.7886	76.71	0	0	193.203
SU:80_7.	1.3017	46.5	0	0	165.12099999999998
SU:80_8.	1.2097	20.31	0	0	199.998
SU:83_3.	3.0479	138.53	0	0	345.399

Mixture-23

Mixture-93

SA:Mixture-93
SP:11_9 Single Score = 0.0
SP:22_7 Single Score = 0.6
SP:28_8 Single Score = 0.67
SP:4_3 Single Score = 0.33
SP:61_7 Single Score = 0.0
SP:67_5 Single Score = 0.0
SP:79_4 Single Score = 0.0
SP:80_7 Single Score = 0.25
SP:80_8 Single Score = 0.25
SP:83_3 Single Score = 0.0

Mixture-23

Mixture-23
SP:26_3 Single Score = 0.6
SP:30_0 Single Score = 0.5
SP:32_9 Single Score = 0.0
SP:56_6 Single Score = 0.6
SP:71_7 Single Score = 0.4
SP:73_9 Single Score = 0.6
SP:82_4 Single Score = 0.57
SP:89_4 Single Score = 0.0
SP:91_3 Single Score = 0.57
SP:98_6 Single Score = 1.0

Initial Temperature 1000.00
Final Temperature 0.01
Max Steps 1000.000000
Probability Constant 50.000000
Cooling method ○ Exponential ● Linear
Iteration 100

```
Mixture-62  score: 9.580952380952384
Mixture-4  score: 5.402380952380952
Mixture-16  score: 4.459523809523811
Mixture-83  score: 4.745238095238096
Mixture-8  score: 5.0
Mixture-6  score: 3.8333333333333334
Mixture-23  score: 4.842857142857144
Mixture-10  score: 5.171428571428573
Mixture-59  score: 4.052380952380952
Mixture-45  score: 4.595238095238096
Mixture-98  score: 4.964285714285715
Mixture-69  score: 6.721428571428573
Mixture-50  score: 3.157142857142857
Mixture-51  score: 4.695238095238095
Mixture-2  score: 4.069047619047619
Mixture-75  score: 2.9499999999999997
Mixture-20  score: 3.119047619047618
Mixture-49  score: 4.892857142857144
Mixture-15  score: 5.085714285714287
Mixture-40  score: 6.133333333333334
Mixture-90  score: 2.7166666666666663
Mixture-67  score: 2.988095238095237
Mixture-24  score: 4.076190476190477
Mixture-18  score: 7.235714285714287
Mixture-43  score: 5.278571428571428
Mixture-22  score: 5.226190476190475
Mixture-38  score: 3.907142857142857
Mixture-29  score: 3.161904761904761
Mixture-7  score: 5.676190476190476
Mixture-54  score: 5.845238095238095
Mixture-35  score: 2.847619047619047
Mixture-70  score: 4.011904761904762
Mixture-93  score: 2.0999999999999996
```

Figure 2.15 Mixtures analysis graphical tools.

A Analysis module showing the 2D structural representation of molecules in the currently analysed mixture. **B** Chemical property summary tab. **C** Mixture editor tab. Single component name and relative score facilitates the identification of crucial unfavourable elements in the mixtures; library manipulation is possible via drag-and-drop features. **D** Mixture optimisation module; the left-side panel shows various settings and the right-side panel shows the predicted new scores.

2.7. Acknowledgments

I thank Dr Christine Prosser[a,b] for providing the experimental dataset needed for validating the routines presented in this chapter. Special thanks go out to Dr. TJ Ragan[a] for comments and suggestions on implementing the simulating annealing algorithm from the literature and for his previous work on the PCA core architecture.

2.8. Addresses

[a] Leicester Institute of Structural and Chemical Biology, Department of Molecular and Cell Biology, University of Leicester, Henry Wellcome Building, Lancaster Road, Leicester LE1 7HN, United Kingdom.
[b] UCB Celltech, Bath Road, Slough, Berkshire SL1 3WE, United Kingdom.

Chapter 3

Pipelines on simulated datasets

See List of Publications: 1.

Keywords: FBDD NMR, CcpNmr Software, STD, WaterLOGSY, CPMG

3.1. Abstract

Fragment-based drug discovery, FBDD, is one of the key methods in the early phases of drug development for identifying drug-like candidates. NMR techniques, in particular 1-dimensional (1D) ligand-detected methods, have mostly been employed in the early stages of drug discovery. However, the data analysis can be often daunting, and limited freely available software packages for qualitative and quantitative spectral assessment is one of the current bottlenecks in FBDD. In this chapter I present the development of screening tools that were included in the CcpNmr AnalysisScreen. Here, I discuss the pipeline architecture, and validation of core algorithms, called pipes, aimed to create guided workflows for the hit identification in the early process of drug discovery by NMR.

3.2. Introduction

The great advances in technology, including sample automation on the NMR spectrometers, have made it possible to probe the protein-ligand complex using multiple experimental NMR methodologies using the same sample in a single screening trial. Although not routinely done in every project, this strategy is believed to minimise the potential downfall of each technique depending on the system under examination and provide for a quantitative scoring. Based on the count of positive observations for each experiment, a molecule is defined as a level-1 hit if it appears as a binder in one experiment, a level-2 hit when two experiments confirm its interaction and so on. In common practice some NMR techniques, i.e. relaxation-edited , STD and WaterLOGSY, are more widely used than others[109,110]; for this reason in this chapter more attention has been dedicated to these techniques and on the development of the corresponding automated analysis tools present in CcpNmr AnalysisScreen.

3.2.1. **Ligand-observed techniques**

Binding hit detection is facilitated by common methods such as relaxation-edited ^{1}H, Saturation Transfer Difference (STD), WaterLOGSY, and in some circumstances by Target Immobilized NMR Screening (TINS) and Spin labels[88,90,111,112].

A small-molecule ligand engaged in a fast-exchange complex with a macromolecule partially acquires the spectroscopic NMR properties, e.g. T_1/T_2 relaxation and ^{1}H-^{1}H cross-relaxation rates, of the macromolecule. When there is a sufficiently large molar excess of the small molecule ligand, this typically results in the detection of chemical shifts of the ligand free-state, but with modified relaxation properties more reminiscent of the bound state[17] (Fig. 3.1A left). For example, small molecules tumble fast in solution and hence their NMR resonance lines are characterised by long transversal relaxation times (T_2) that result in narrow lines. In contrast, when bound to a slowly tumbling macromolecules the NMR lines of the small molecule are significantly broader. Therefore, in the case of fast exchange of the small molecule between the free and bound states, its NMR signals will become broadened (Fig. 3.1A, right).

The saturation transfer difference (STD) experiment relies on the efficient spin-diffusion of saturated proton magnetisation in the macromolecule through measurement of the so-called “on-resonance” and “off-resonance” experiments. In the “on-resonance” experiment, selected ^{1}H resonances of the macromolecule that are non-overlapping with those of the ligand are saturated using a train of RF pulses. The saturation propagates rapidly through the macromolecule and to the bound ligand as a result of efficient intramolecular and intermolecular ^{1}H-^{1}H cross-relaxation, respectively[113] (Fig. 3.1B, left). As the ligands are in rapid exchange between their bound and free states, they maintain their saturated state resulting in attenuated or even absent signals in the resulting “on-resonance” spectra. In the “off-resonance” control experiment, the macromolecular resonances are not saturated resulting in signals with original intensities. Subtraction of the “off-resonance” spectrum from the “on-resonance” spectrum yields the STD spectrum, in which only saturated ligand resonances will be observable (Fig. 3.1B, right). The signals of the macromolecule will be minimal or absent, as a result of the much smaller concentration of the latter in comparison to the ligand, thus greatly simplifying spectral analysis.

In an alternative approach, the so-called WaterLOGSY experiments[88,114] (Fig. 3.1C, left), the ligand and macromolecular target are saturated indirectly through the bulk water magnetisation. The saturation is transferred from the bulk water to the ligand through several mechanisms, in particular by direct ^1H-^1H intermolecular cross-relaxation between water molecules in close proximity to the binding pocket and the bound ligand. Alternative mechanisms include the direct exchange with macromolecular NH and OH protons within the binding site and the ligand, or indirectly, through a spin-diffusion mechanism. In both cases, NMR properties of the bulk water are transferred to the bound ligand, and the resulting spectrum displays inverted signals for bound ligands compared to the unbound ligands (Fig. 3.1C, right). The detection of ligands that bind to macro-molecules with a relatively low density of protons might benefit from the WaterLOGSY technique[9]. Furthermore, WaterLOGSY experiments have displayed higher sensitivity for detecting binding molecules compared to STD experiments when used to screen very large biomolecules at low concentrations[115]. Antanasijevic *et al.* believed that this is caused by the higher concurrent (direct and indirect) saturation of various sites in the binding complex[115].

A third approach exploits the altered T_1/T_2 relaxation properties of ligands that bind to a macromolecular target (*vide supra*). In the so-called ^1H-relaxation-edited experiment, also referred to as the $T_{1\rho}$ experiment, a series of spectra are recorded in which the ligand signals are subjected to varying durations (typically in a range of 0 to 200 ms) of transverse relaxation, i.e. either as R_2 or $R_{1\rho}$. Bound ligands will exhibit faster R_2 or $R_{1\rho}$ rates, i.e. shorter T_2 or $T_{1\rho}$ relaxation times, and their signals will be significantly attenuated in the spectra compared to ligands that do not bind to the macromolecular target (Fig. 3.1D).

In spite of all the powerful NMR experiments used for NMR-based FBDD[93], inefficient evaluation of the primary hit screening data can disrupt or postpone any of the later phases, such as binding site identification and hit optimisation.

Primary screening is routinely performed manually by comparing spectral information derived from thousands of STD, WaterLOGSY and relaxation-edited experiments. Manual analysis of these data inevitable results in human errors or subjective inconsistencies, in addition to problems arising from commonly occurring experimental errors, such as improper alignment and scaling of spectra. The latter are detrimental to the accurate assessment of any datasets, whether manual or automated. Even when using computational routines, several inherent difficulties to the data analysis process still remain. The different nature of each NMR screening experiment translates into fundamentally different spectral patterns. Consequently, it requires robust algorithms, such as those employed for peak detection or peak matching, that ideally require no fine tuning of algorithms via adjustable parameters as this would slowdown, complicate and reduce the reproducibility of whole data analysis. Accurate peak detection is also fundamental for the generation of the most optimal mixtures on the basis of the library of spectra of the compounds, as subsequent deconvolution of their spectra is a key step in the identification of potentially binding compounds.

Currently, only a limited number of tools that provide support for NMR screening exist, such as Bruker TopSpin[116] or MestreLab MNova Screen[117], both of which are often not affordable for occasional or academic users. Alternatively, NmrGlue[118], a freely available collection of NMR library functions, could serve as the building blocks for creating stand-alone custom scrips for expert users, but to the best of my researches no such efforts have been documented. In this chapter I describe the tools I introduced in AnalysisScreen aimed to facilitate the hit identification process, automation of common processing and analysis workflows. As a result, AnalysisScreen assists in both qualitative and quantitative inspection of NMR data, reducing false negatives (wrongly missed or rejected hits) and false positives (wrongly accepted hits). The AnalysisScreen core is implemented with the requirements of speed and customisation in mind, thus offering users a platform capable of easy adaptations, following any future NMR methods that might emerge.

3.3. Materials and Methods

This chapter contains only simulated datasets and a technical description of computational routines. Algorithms, synthetic datasets and macros, were written using the CcpNmr Python environment, as described in chapter 2. In addition, the ABC Python library[119] has been used for the creation of the underling architecture of the pipeline and pipes. Lastly, the Nmrglue[118] package has been employed for the implementation of generic algorithms, such as the phase correction method.

3.3.1. Materials

The ^{1}H, STD, WaterLOGSY pipelines (described above in section 3.4.2) were tested using three types of simulated datasets.

The first dataset was used for testing the screening routines in the pipelines. I created semi-automatically a dataset that included: 5 reference components, 30 control samples and 30 samples miming the presence of an interacting target as singletons or as mixtures of 5 components. The adjustable parameters included: the selection of various chemical shifts, the degree of broadening and intensity changes for each signal in the relative spectrum, and the creation of non-overlapping mixtures. This strategy ensured me a full control over the spectral quality and the most robust way of validating the results. Sixty-five different spectra and their metadata were simulated for each of the three common NMR screening methods using an internal Python script and the spectral data stored as CcpNmr HDF5 formatted files.

Finally, the collection of data was loaded into the main programme using excel files. These files contained several sheets including substances and samples, as described in chapter 2, and they are now part of the tutorial present in the main software distribution.

The second dataset was essential for the testing of the hit analysis module and evaluate its speed handling capacity of the software. Thus, I generated a dataset of 8000 1D spectra, containing 34606 peaks at random positions, with random heights and linewidths. Out of the 8000 spectra, 3000 randomly selected spectra were flagged as SpectrumHits.

For both datasets, each simulated spectrum was linked to a virtual substance or sample, containing randomly set properties; including SMILES, concentrations, pH, chemical names, etc. All simulated datasets and metadata generated for this work were used only for testing and validation, therefore were not associated with real biological experiments.

Lastly, the third dataset was necessary for testing the STD hit detection routines on a dataset that presented a more realistic experimental pattern. To this purpose, I simulated a typical STD spectrum for 100 compounds and created it in 300 different randomly generated variants at various Signal-to-Noise (S/N) ratios. The peak picker routine was expected to find a total of 100 known true positive peaks and 100 true negative. Total true negatives were set arbitrarily to 100 to avoid an unbalanced dataset.

3.3.2. **Methods**

The initial implementation of the AnalysisScreen peak picker was based on the algorithm described by Boucher and Stevens[120]. The core routine uses the SciPy's multi-dimensional image-processing function *maximum_filter* (MF) to detect local maxima points in a 1D data array. This function, and therefore the picking peak routine, requires several parameters for a correct behaviour, namely the size of the search box and "mode" of dealing with boundaries of the input array. The size determines the region to include when searching for maxima, which ultimately will include the number of nearest neighbours (either side of the maximum) considered as a peak. A smaller size value will increase the total number of peaks found, whereas a larger value will exclude real signal peaks. A size of 10 was considered appropriate for most of the simulated dataset; however, for noisy simulated and experimental recorded datasets this value was impossible to determine correctly *a priori*.

I have further extended the routine so that also regions with negative signals could be detected, e.g. as needed for analysing WaterLogsy spectra. I further added the ability to mask regions, (e.g. solvent signals), and to auto-detect positive and negative noise threshold values (see below) so that potential false positive could be limited when detecting real signals.

However, after failing several attempts in automatically estimating all required parameters when dealing with large datasets, I decided to implement and explore a different algorithm based on the method described by Billauer[121].The Billauer algorithm is based on the detection of two local minima to establish the maximum between them. My implementation consisted of the removal of the detection of local minima as the valley points between maxima, and by enabling the detection of true NMR negative signals. Furthermore, I optimised it for handling larger NMR dataset using Numba's properties, reducing the processing speed from seconds to milliseconds per spectrum. Lastly, I inserted extra filters, such as masked regions (to be ignored from the analysis).

Positive and negative noise thresholds are estimated automatically as following:

$$N_{Th} = \alpha\sigma N * N_{Max}$$

Eq. 3-1

Where N is a defined downfield region of the spectrum, default 10% of the total datapoint count; σ is its standard-deviation and α is the adjustment factor. N_{Min}, instead, is used for calculating the negative threshold.

Negative and positive noise threshold values were also used for calculating the Signal-to-Noise ratio as

$$\mathrm{SN_{Ratio}} = \alpha * \frac{\mathrm{S}}{\mathrm{N_{Max}} - \mathrm{N_{Min}}}$$

Eq. 3-2

Where S is the peak height; α is the adjustment factor. N_{Max} and N_{Min} are the positive and negative noise threshold values.

Matching and relative scores for the hit identification were calculated as

$$\mathrm{S_{Rel}} = |\mathrm{A_{Med}}| * \mathrm{A_{Tot}}$$

Eq. 3-3

Where A_{Med} represents the median for the absolute observations (peak heights or Δppm positions for matching scores) and A_T the total count. If only two values A are present in the array, then only the min value is taken:

$$\mathrm{S_{Rel*}} = |\mathrm{A_{Min}}| * \mathrm{A_{Tot}}$$

Eq. 3-4

Hit Scores were normalised to values in a range 0-100 by:

$$\mathrm{S_{Tot}} = 100 * \frac{\mathrm{S} - \mathrm{S_{Min}}}{\mathrm{S_{Max}} - \mathrm{S_{Min}}}$$

Eq. 3-5

Where S are the relative scores calculated using equations 3.3 and/or 3.4.

STD efficiencies were calculated as described in the literature[122]:

$$\mathrm{E_{STD}} = \frac{\mathrm{I_O} - \mathrm{I_{Sat}}}{\mathrm{I_O}} = \frac{\mathrm{I_{STD}}}{\mathrm{I_O}}$$

Eq. 3-6

Where I, is the intensity for a give H signal in the Off-resonance (I_O) and On-resonance (I_{Sat}) spectra.

Pipelines and running parameters

In this work the ^{1}H, STD, WaterLOGSY data analyses were simulated and implemented in their respective pipelines.

The ^{1}H line-broadening detection pipeline was composed of the following pipes:

- *Noise Threshold Pipe*; values were set manually using the built-in GUI widgets to 0-57190.06 (a.u. for intensity).
- *Calculate Integrals*; in which the minimal linewidth to yield a valid integral was set to a value of 0.01 ppm. Peaks were automatically detected. The ignored regions option for this pipe was at the time of testing still under implementation; hence, peaks below 6 ppm were manually deleted.
- *Peak Broadening Hit Finder*; control and target spectrum group were selected according to the dataset; references spectra were calculated automatically from the SampleComponents linked to the samples; control and target peaks were matched to each other and references if their chemical shifts were within a range of 0.010 ppm. Finally, the minimal volume variation ratio was set to 0.20 (20%).

The STD data analysis comprised two pipelines: one to calculate the STD efficiency and one to identify the STD hits.

The first pipeline was composed of:

- *STD Creator*; which created the STD spectra by subtracting the On-resonance to the Off-resonance spectrum, creating a new SpectrumGroup containing the newly available STD spectra.
- *Exclude Regions*; one large region was selected from 1.435 to 6.259 ppm.
- *Noise Threshold*; values were graphically selected and included values from -19946.09 to 9973.05.

- *Peak Picker* pipe; it included negative peaks (although not necessary in this case), noise level factor, and filter size of 10. Filter mode was set to “wrap”.

The subsequent STD hit analysis pipeline consisted of two pipes:

- *STD Efficiency*; SpectrumGroup entries were selected as the dataset; peaks tolerance was set as 0.03 ppm;
- *STD hits*; only the target was used for this experiment and references were retrieved from the SG:References; peaks were matched using a tolerance of 0.03 ppm, finally the minimal efficiency was set to 1%.

The WaterLOGSY data analysis consisted of the following pipes:

- *Exclude Regions*; one graphically selected region from 0 to 6.460 ppm
- *Noise Threshold*; values were graphically selected and included values from –82071.56 to 38390.76.
- *Peak Picker* pipe; it included negative peaks; noise level factor was set to 9.70; filter size was set to 10. Filter mode was set to “wrap”.
- *WaterLOGSY hits*; the mode was set to “intensity changed”; SpectrumGroups were selected according to the dataset; matching tolerances were set to 0.10, finally, the minimal intensity change was set to 10.

3.4. Results and Discussion

In order to facilitate the manual analysis of screening dataset, I firstly implemented a graphical interface that allowed users to compare the spectra derived from the different NMR screening experiments. The interface used the built-in spectrum displays and sidebar capabilities as underpinning elements.
Spectra associated to each sample (previously loaded from excel files) recorded according to a particular experiment type, can be automatically displayed in a so-called stacked mode together with their spectral references (Fig. 3.2A). This approach has the main advantage of comparing multiple experimental data at once and provides for easy access to all items by simply using the directional keys. This method is particularly beneficial from a quick and qualitative assessment of ^{1}H, STD and WaterLogsy experiments; however, for time-series experiments, such as the one shown in Fig. 3.2B, this approach is less suitable. The nature of the time-series data required the development of a dedicated analysis window (Fig. 3.2C); spectral crowdedness can make the manual analysis nearly impossible. 1D time-series can also already be analysed manually using the Chemical-Shift-Mapping module tool I specifically developed for the target-based drug discovery methodologies (described in chapter 6).

3.4.1. Pipeline design

In addition to the above simplified manual inspection of spectra, the true strength of AnalysisScreen, comes with quantitative and automated analysis routines for common 1D experiments. The heterogeneity of NMR techniques for 1D screening translates in the need of specific analysis workflows for each method. I addressed this by designing and implementing the AnalysisScreen pipeline module (Figs 3.3A-B). It permits users to apply multiple tasks or algorithms, called pipes, to single spectra or all spectra contained in a SpectrumGroup.

The pipeline architecture has been implemented with great consideration following the "Clean-Code" design principles, and rigidly following the "clean architecture" design[123] for its framework (Fig. 3.4A). The clean architecture, originally introduced by R. Martin, focuses on the structure and relation across code components which should be independent to each other[124]. The general structure is best represented by concentric circles, where each circle is a different layer or code component. These are organised in a way the outer layers are lower code levels, e.g. user settings, general GUIs, which define mechanism; and the inner layers are higher code levels, e.g. abstract classes, which define policies[125] (Fig 3.4A).

The clear separation between each layer of the underline machinery, allowed me to design the pipes as generic items, independent from the metadata and analysis required. Consequently, a collection of pipes creates a pipeline queue that effectively implements a user-defined workflow. Furthermore, new pipes and new algorithms can be added without altering the functionality of the software. This feature is called in object-orientated programming, the "open-closed" principle, which is another instance of the clean-code concept.

An example of the implementation of a pipe and its associate GUI is shown in Figs 3.4B-C.

The pipeline module itself is built in a such way that it can be run from the AnalysisScreen command-line interface, if so desired by an (expert) user; however, I also designed and developed a full graphical interface (Fig. 3.3B). The pipeline GUI module has three main sections: settings; pipe selection, and the main working area, that contains the selected pipes. AnalysisScreen features application-specific pipes, such as those for line broadening analysis, WaterLOGSY and STD hit detection, as well as a set of other more generic data manipulation pipes that are now shared across all version-3 programmes[28] (Table 3.1).

Table 3.1 Currently available pipes grouped by type and functionalities.

Type	Functionality	Available	Under Development
Ccpn Spectra	Performs actions on Ccpn spectra. **In:** Ccpn Spectra **Out:** Ccpn Spectra	• Global Alignment 1D • Peak Picker: 1D, nD • Integration 1D • Referencing 1D • Auto Scaling 1D • Copy peaks: 1D, nD • Refit peaks: 1D, nD • Auto phasing 1D • Baseline Correction 1D • Duplicate spectra **AnalysisScreen 1D** • Hit detection by line-broadening • STD spectrum creator • STD Efficiency • Hit detection by STD • Hit detection by WaterLogsy • Peak Height, Linewidth, Volume % change • General peak matching	• Global Alignment nD • Peak Filter nD
Pandas Dataset	Represents project or general data in tables like format **In:** Ccpn Spectra, Pandas Dataset **Out:** Pandas Dataset	• Output results	
Generic	Performs generic actions on project **In:** None, Same as Input **Out:** None, Same as Input	• Exclude Solvent Regions 1D • Noise threshold 1D	

Lastly, any pipeline can be saved as a JSON file for re-usage or exchange with other users of the CcpNmr Analysis suite. The architecture of a such file reflects the general pipeline GUI structure and includes all parameters which were last used. The loading and restoring mechanisms for pipes and widgets, was subsequently also adopted for saving and restoring the layouts of main CcpNmr version-3 programme. This feature is believed to save precious time when executing the same routines on different datasets and increases the automation capabilities of the software.

3.4.2. **Testing pipelines**

Pipelines were initially tested on a series of small datasets, where for each sample, multiple spectral patterns were simulated according to the ^{1}H, WaterLOGSY and STD typical lineshapes. The synthetic dataset did not necessitate the application of the common post-processing pipes, such as those for baseline correction, phasing or spectral alignment.

The first pipeline was applied to a ^{1}H dataset with the aim of detecting line-broadening. The pipeline was built from the following pipes: Noise Threshold, Calculate integrals and Peak Broadening Hit Finder pipes (Fig. 3.5A). These pipes automatically detected and integrated signal regions from the control and target simulated spectra, finally only peaks that resulted in a line broadening were matched to their spectral references.
After the pipeline was run, two out of two known SpectrumHits were correctly found: Component-1 at 7.74 ppm (Fig. 3.6), Component-3 at 8.198 (Fig. 3.7).

The second pipeline was built to test the detection of intensity changes typical of WaterLOGSY experiments. The pipeline comprised the following pipes: Exclude Regions, Noise Threshold, Peak Picker and finally the WaterLOGSY specific pipe (Fig. 3.8). After the initial signal detection, the WaterLOGSY pipe evaluates signals from the control and target spectra, and filters only peaks that undergo a significant change, followed by the matching of those peaks to their reference chemical shifts. The pipe successfully detected multiplets at ~7.74 and ~8.20 ppm relative to component-1 and component-3 (Figs 3.9-3.10).

The STD analysis consisted of two separated pipelines. The first was needed to create the STD spectrum from the On- and Off-resonance spectra, followed by the peak detection for all input spectrum groups (Fig. 3.11). The successive pipe calculated the STD efficiency as defined by eq. 3.6 and finally detected the STD positive signal and matched this to the reference spectra. For this experiment as well, peaks from the mixture were matched correctly to the Component-1 and Component-3 at the expected positions ~7.74 and ~8.20 ppm respectively (Figs 3.14-3.15).

SpectrumHits results can be accessed and inspected graphically by the Hit Analysis module. This module allows for dynamic navigation through the spectra and peaks using the best-matched references and SpectrumHits. Furthermore, the main table of this module allows for a quick filtering of the best results by several scores and the module displays all associated hit metadata. For testing and validating the performance of the module, I simulated a larger dataset, containing 3000 SpectrumHits at various hit levels, e.g. level 3 if in a given component was marked as a SpectrumHit in three simultaneous experiments (Fig. 3.15).

3.4.3. **Determining S/N ratio for STD hit identification**

After pipelines were tested on small datasets simulating ideal spectral patterns for STD, WaterLOGSY, and ^{1}H experiments, I then created a larger dataset of simulated spectra, at various signal-to-noise ratios (S/N) to determine the sensitivity of the procedure for the S/N of the spectral data. I aimed to determine the S/N regime for which observations could be accepted reliably as True Positive (TP) SpectrumHits (Fig. 3.16A). Using these simulated spectra, I also evaluated the peak picker algorithm for its accuracy and sensitivity for correctly distinguishing and locating the spectral signal from the noisy part of the spectrum. The results of the original peak picker algorithm (MF) (as described in section 3.4.1) were not encouraging due to the need for manual optimization of its multiple adjustable parameters, required for to control the occurrence of false positive and false negative outcomes. This finding prevented the full automation of the data analysis pipeline.

Consequently, I developed and used a new peak detector (PD) algorithm dependent on only one adjustable threshold parameter. Using the noise level threshold routine detection, the simplified algorithm was able to detect over 90% of true positive observations down to an estimated S/N of ~1.5 (Fig. 3.16B and Fig. 3.17A). Decreasing threshold parameter in an attempt of including more True Positive observations at lower S/N resulted in a decreased general accuracy and precision, which is, obviously, not favourable (Figs 3.16C-D and Figs 3.17A-D).

The threshold value was also inspected using the receiver operating characteristic curve (ROC) curve which allows the capacity of a binary classifier followed to be determined by an adjustment in the threshold value[126]. The ROC curve (Fig. 3.17D) shows the default threshold value to be located in the most favourable region of the ROC curve, suggesting it can be used as a reliable threshold for the automatic peak picking routine.

3.5. Conclusions

The nature of NMR high-throughput screening involves the analysis of a large quantity of data, which can be both intimidating and time-consuming. The lack of free and user-friendly software packages, capable of handling such a large NMR dataset, forces researchers to set up and repeat tiresome procedures manually, a process that might unintentionally lead to mistakes. Furthermore, the manual process also depends on qualitative judgments by the users, which increases the likelihood of misinterpretation of results. In this chapter, I described solutions which are now included in the CcpNmr AnalysisScreen programme. AnalysisScreen is able to cope efficiently with very large datasets; I successfully used tens of thousands one-dimensional spectral entries and their related metadata, including projects with over 1 million peaks yet providing fast and reproducible results. To achieve this, I designed and developed a pipeline module. The pipeline is rigorously designed in a modular way, such as that multiple tasks, called pipes, can be arranged depending on the user requirements. This results in a very flexible platform for custom implementations and bespoke workflows, such as the one needed for analysing different experiments in the ligand-detected NMR screening trials. I created several general post-processing pipes, such as spectral alignment, spectrum referencing, phase and baseline correction, and signal quantification pipes, such as two peak detection and integration. Finally, application specific pipes, such as line broadening, WaterLOGSY and STD hit detection that are needed to analyse spectral mixtures or single spectra in the presence and the absence of a target. I have then tested these workflows using simulated datasets for each experiment type. The STD, WaterLogsy and relaxation-edited ^{1}H experiments retrieved the two expected SpectrumHits. The relative peaks that classified the spectrum as a SpectrumHit were correctly matched to their known reference spectra. Results were easily accessible using the graphical tools that I meticulously designed and developed. Lastly, I have designed a model based on the spectral signal-to-noise ratio capable of estimating a threshold needed for filtering quantitatively SpectrumHits observed in STD spectra.

AnalysisScreen is actively developed and publicly available within the Analysis Version 3.0 release. Future releases of AnalysisScreen will include pipes for automating the analysis of 1D relaxation experiments (with T1/T2 CPMG routines currently under testing as macros) and automatic analysis of STD amplification factor and K_d's. In addition, a more exhaustive Hit Analysis module is envisioned, that integrates cheminformatic tools for classifying hits by functional groups and supports the Pan-Assay Interference Compounds (PAINS) filters[68].

3.6. Figures

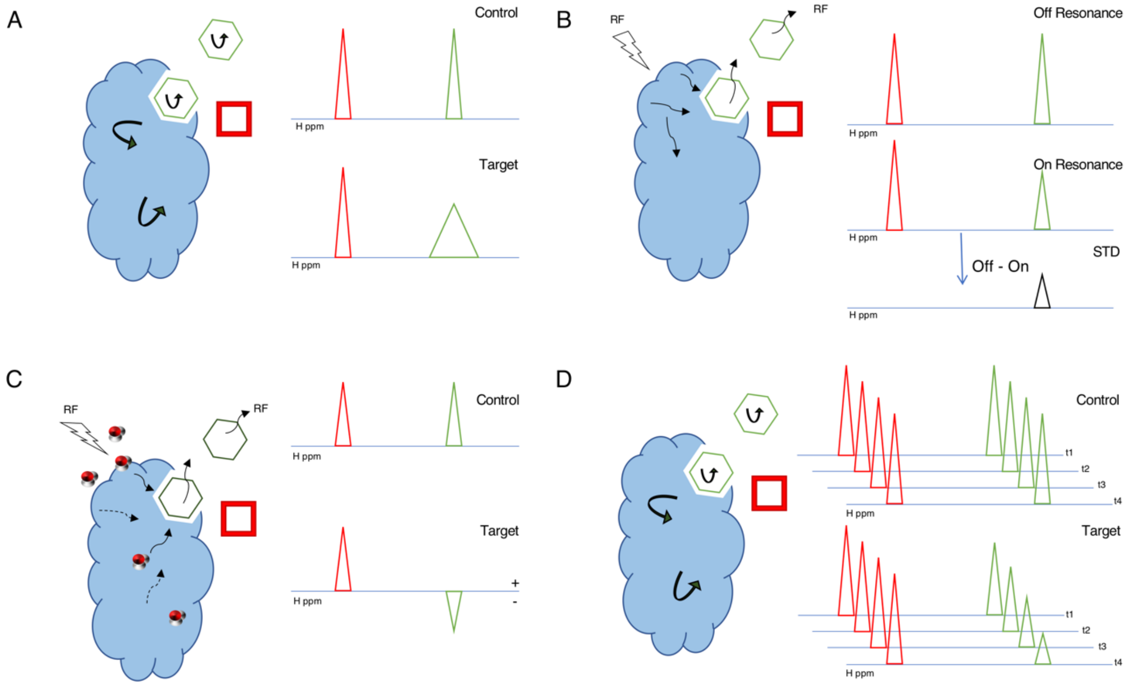


Figure 3.1 Common NMR methods for detecting ligand binding.

Common NMR methods for detecting ligand binding[93] to a large macromolecular target (blue motif). The binding and non-binding compounds (small molecules) are displayed as a green hexagons and red squares, respectively **A** ^{1}H Relaxation-edited experiment. The peaks of both compounds in the control spectrum are characterised by narrow resonance lines. In the presence of a target, a binding compound partially acquires the NMR properties of the macromolecule, resulting in a broadening of its resonance line (green peak). The effect does affect a non-binding compound. **B** In the on-resonance experiment of a saturated transfer difference (STD) experiment, a saturating RF field is applied to the target and saturation is transferred to the binding compound, resulting in a slightly lower intensity of its resonance line. In the off-resonance control experiment no such effect occurs; consequently, only the resonance of the binding compound will be visible in the STD spectrum. **C** In the WaterLOGSY experiment saturation is transferred to the target through saturation of the bulk water molecules and passed on to the binding compound. Its resonance line in the spectrum in the presence of the target will have the opposite sign compared to the control spectrum. **D** In the $T_1\rho$ experiments a series of spectra are recorded with different relaxation durations. For the binding compound, spectral intensities will attenuate at a faster rate compared to the non-binding compound.

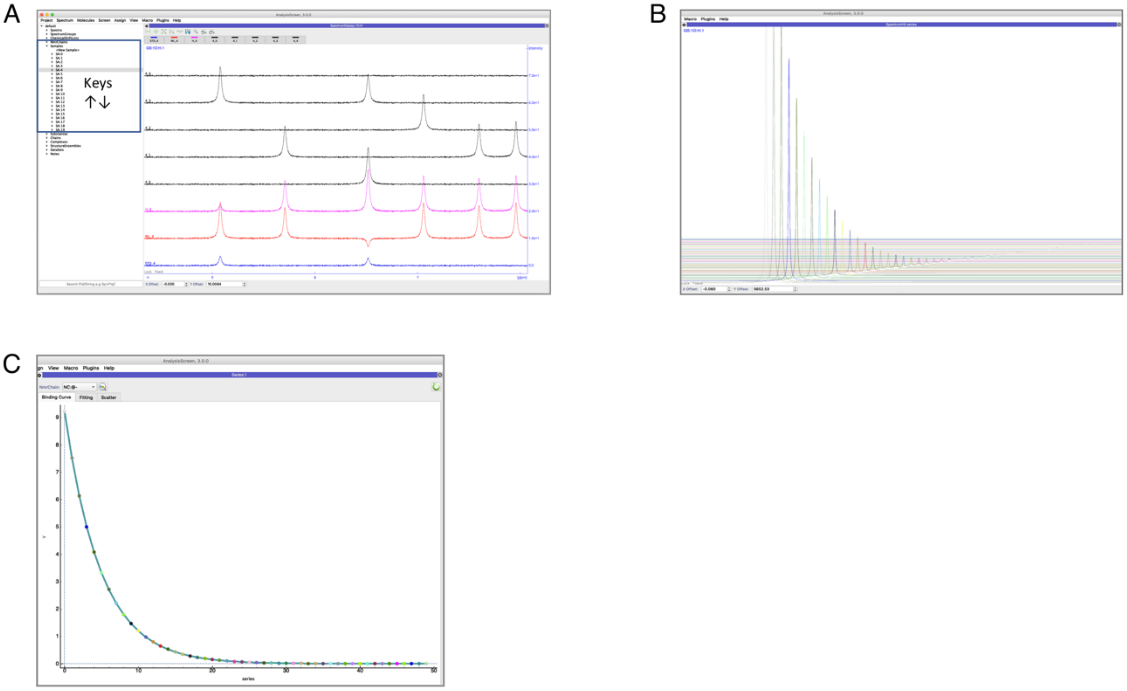
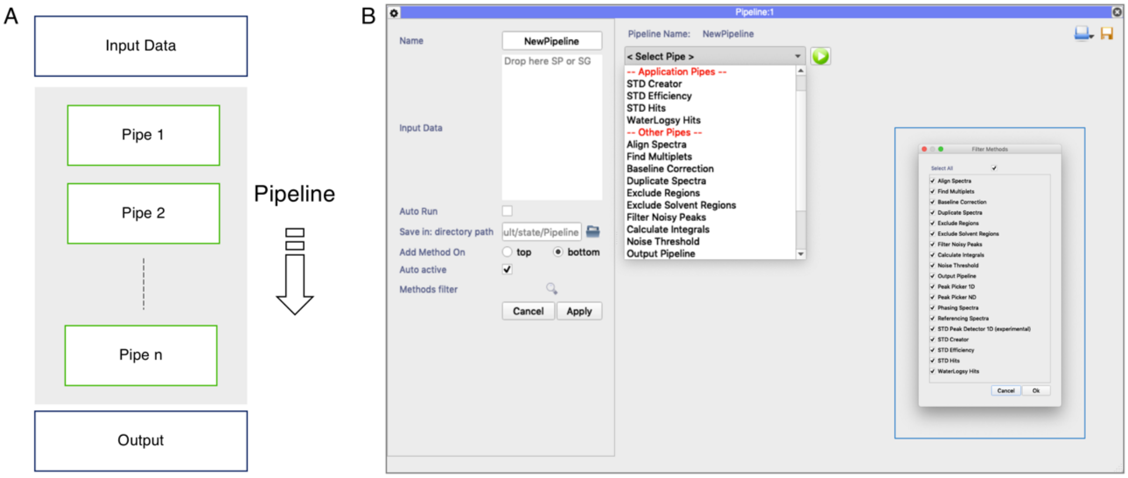


Figure 3.2 CcpNmr AnalysisScreen tools for manual visual inspection.

A Spectrum display module showing a simulated SpectrumGroup comprised of a series of 1D ^{1}H spectra with vertical and horizontal offsets to facilitate a quick inspection. **B** Example of usage of the Chemical Shift Mapping for assessing 1D spectra series. The figure shows a binding curve for the spectra displayed in **A** using peak heights as the input for calculations. **C** The sidebar of AnalysisScreen can be used for easy navigation through data items using up/down keys (indicated by the rectangular box). This option allows to display automatically stacked spectra recorded for the selected sample and their references for a quick and manual visual inspection. NB. this feature is currently enabled using a command line macro.



Pipeline Name: NewPipeline

Name: NewPipeline

Input Data: Drop here SP or SG

Auto Run: ☐

Save in: directory path: ult/state/Pipeline

Add Method On: ○ top ● bottom

Auto active: ☑

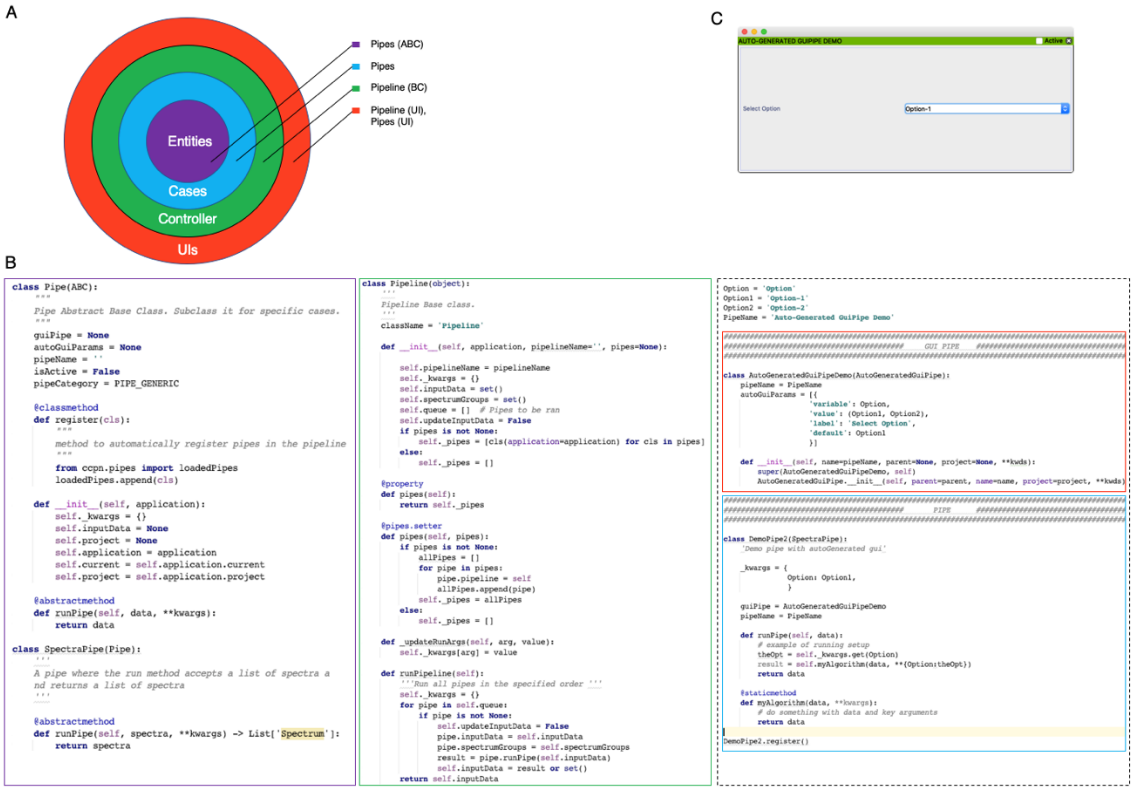
Methods filter

Select All ☑

- ✓ Align Spectra
- ✓ Find Multiplets
- ✓ Baseline Correction
- ✓ Duplicate Spectra
- ✓ Exclude Regions
- ✓ Exclude Solvent Regions
- ✓ Filter Noisy Peaks
- ✓ Calculate Integrals
- ✓ Noise Threshold
- ✓ Output Pipeline
- ✓ Peak Picker 1D
- ✓ Peak Picker ND
- ✓ Phasing Spectra
- ✓ Referencing Spectra
- ✓ STD Peak Detector 1D (experimental)
- ✓ STD Creator
- ✓ STD Efficiency
- ✓ STD Hits
- ✓ WaterLogsy Hits

Figure 3.3 CcpNmr AnalysisScreen Pipeline and Hit Analysis module.

A Schematic representation of a pipeline. The pipeline is able to handle SpectrumGroups as well as single spectra as the input data. Each pipe performs a dedicated action on the spectra and returns a new set of spectra which are used as input for each successive pipe. Finally, a result or report pipe provides information on performed actions. **B** Current graphical user interface for assembling and executing a Pipeline. The left side shows the available settings affecting the execution of the pipeline. Pipelines are constructed by simply selecting pipes from the main pull-down; the grey area underneath displays the selected pipes. On the right side, a pop-up is shown highlighted in blue, which can be used to customise the main selection pull-down. Pipelines can also be saved and restored, including last used parameters, as a JSON file that can be shared with other AnalysisScreen users.



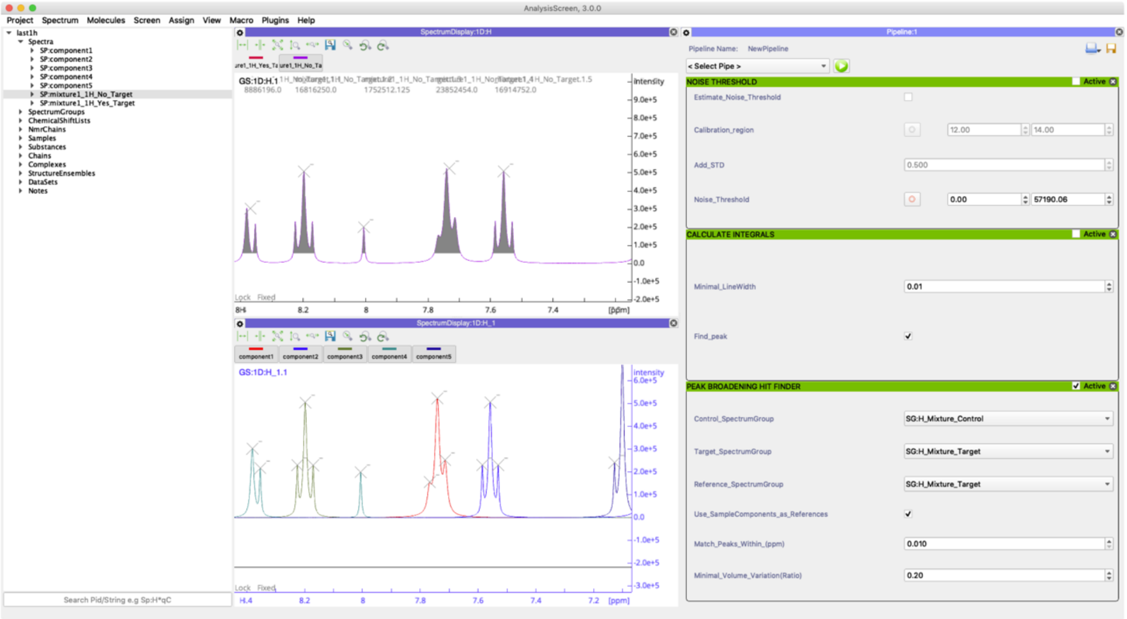
```python
class Pipe(ABC):
    """
    Pipe Abstract Base Class. Subclass it for specific cases.
    """
    guiPipe = None
    autoGuiParams = None
    pipeName = ''
    isActive = False
    pipeCategory = PIPE_GENERIC

    @classmethod
    def register(cls):
        """
        method to automatically register pipes in the pipeline
        """
        from ccpn.pipes import loadedPipes
        loadedPipes.append(cls)

    def __init__(self, application):
        self._kwargs = {}
        self.inputData = None
        self.project = None
        self.application = application
        self.current = self.application.current
        self.project = self.application.project

    @abstractmethod
    def runPipe(self, data, **kwargs):
        return data

class SpectraPipe(Pipe):
    '''
    A pipe where the run method accepts a list of spectra a
    nd returns a list of spectra
    '''

    @abstractmethod
    def runPipe(self, spectra, **kwargs) -> List['Spectrum']:
        return spectra
```

```python
class Pipeline(object):
    '''
    Pipeline Base class.
    '''
    className = 'Pipeline'

    def __init__(self, application, pipelineName='', pipes=None):

        self.pipelineName = pipelineName
        self._kwargs = {}
        self.inputData = set()
        self.spectrumGroups = set()
        self.queue = []  # Pipes to be ran
        self.updateInputData = False
        if pipes is not None:
            self._pipes = [cls(application=application) for cls in pipes]
        else:
            self._pipes = []

    @property
    def pipes(self):
        return self._pipes

    @pipes.setter
    def pipes(self, pipes):
        if pipes is not None:
            allPipes = []
            for pipe in pipes:
                pipe.pipeline = self
                allPipes.append(pipe)
            self._pipes = allPipes
        else:
            self._pipes = []

    def _updateRunArgs(self, arg, value):
        self._kwargs[arg] = value

    def runPipeline(self):
        '''Run all pipes in the specified order '''
        self._kwargs = {}
        for pipe in self.queue:
            if pipe is not None:
                self.updateInputData = False
                pipe.inputData = self.inputData
                pipe.spectrumGroups = self.spectrumGroups
                result = pipe.runPipe(self.inputData)
                self.inputData = result or set()
        return self.inputData
```

```python
Option = 'Option'
Option1 = 'Option-1'
Option2 = 'Option-2'
PipeName = 'Auto-Generated GuiPipe Demo'

##########################################################################################
########################################      GUI PIPE     ###############################
##########################################################################################

class AutoGeneratedGuiPipeDemo(AutoGeneratedGuiPipe):
    pipeName = PipeName
    autoGuiParams = [{
                    'variable': Option,
                    'value': (Option1, Option2),
                    'label': 'Select Option',
                    'default': Option1
                    }]

    def __init__(self, name=pipeName, parent=None, project=None, **kwds):
        super(AutoGeneratedGuiPipeDemo, self)
        AutoGeneratedGuiPipe.__init__(self, parent=parent, name=name, project=project, **kwds)

##########################################################################################
########################################        PIPE       ###############################
##########################################################################################

class DemoPipe2(SpectraPipe):
    'Demo pipe with autoGenerated gui'

    _kwargs = {
              Option: Option1,
              }

    guiPipe = AutoGeneratedGuiPipeDemo
    pipeName = PipeName

    def runPipe(self, data):
        # example of running setup
        theOpt = self._kwargs.get(Option)
        result = self.myAlgorithm(data, **{Option:theOpt})
        return data

    @staticmethod
    def myAlgorithm(data, **kwargs):
        # do something with data and key arguments
        return data

DemoPipe2.register()
```

Figure 3.4 Pipeline architecture and code examples.

A Representation of the pipeline "onion" architecture. The central circle (purple) corresponds to highest code levels, called "entities"; here are present classes which define generic rules and behaviours. Entity are for example the pipe abstract base-classes (ABC). The second layer (light blue), called "cases", are the specific pipe implementations, these are the various screening pipe discussed in the chapter. These can be added or removed freely without altering the architecture. The following layer is the so called "controller". The pipeline base-class (BC) acts as a controller, it is responsible of queuing pipes and running the machinery. The last layer, red circle, corresponds to the user interface code, both for pipeline and pipes. It is a lower level, in the sense that it is the most likely to change during the software development. Changes in this layer will not affect any of the inner layers. **B** On the left box, (purple), a screenshot of the Python source code for the pipe abstract classes; on the inner box, (green), a screenshot of the pipeline base class; on the right box, (dotted black), a file containing the Python source code template for an auto-generated GUI pipe. The file contains the GUI class (red box), and the specific pipe class (blue box). Pipes files containing the appropriate code elements become automatically usable in the pipeline module **C** A simple demo pipe with an auto-generated GUI outputted from the code showed in figure **B** (dotted box).



Pipeline Name: NewPipeline

NOISE THRESHOLD ☐ Active

Estimate_Noise_Threshold ☐

Calibration_region 12.00 14.00

Add_STD 0.500

Noise_Threshold 0.00 57190.06

CALCULATE INTEGRALS ☐ Active

Minimal_LineWidth 0.01

Find_peak ✓

PEAK BROADENING HIT FINDER ✓ Active

Control_SpectrumGroup SG:H_Mixture_Control

Target_SpectrumGroup SG:H_Mixture_Target

Reference_SpectrumGroup SG:H_Mixture_Target

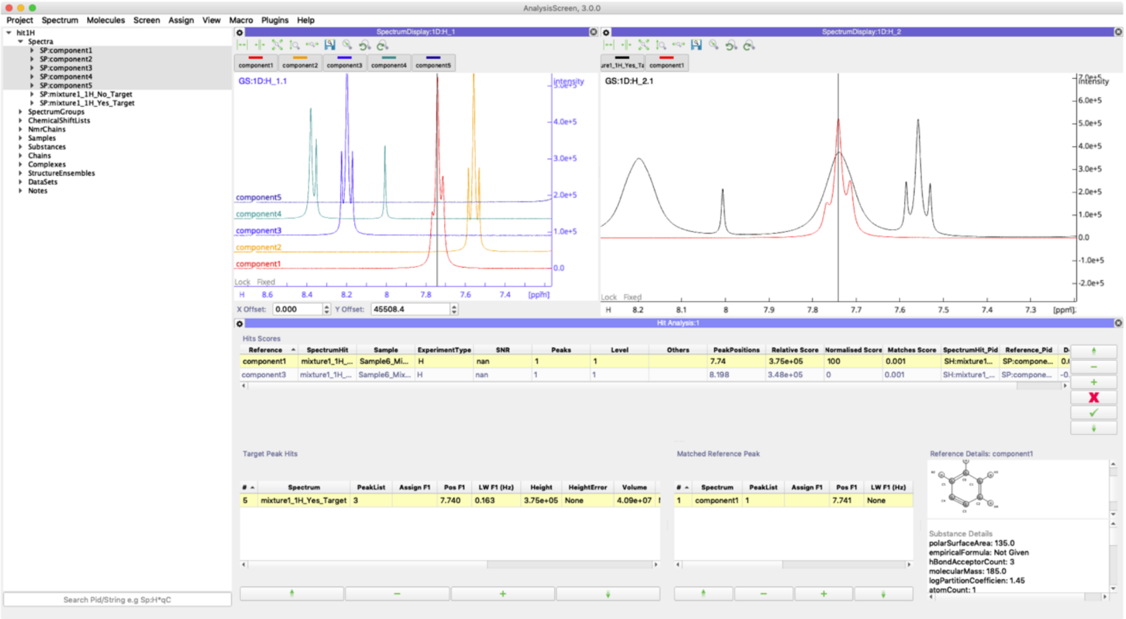
Use_SampleComponents_as_References ✓

Match_Peaks_Within_(ppm) 0.010

Minimal_Volume_Variation(Ratio) 0.20

Figure 3.5 ^{1}H line-broadening detection pipeline.

Simulated ^{1}H relaxation-edited spectra in presence and absence of a target and display of automatically detected integrals (top spectrum display module), library reference spectra (lower spectrum display module), and peak-broadening detection pipeline used in the data analysis (right-hand side).



Hits Scores

Reference	SpectrumHit	Sample	ExperimentType	SNR	Peaks	Level	Others	PeakPositions	Relative Score	Normalised Score	Matches Score	SpectrumHit_Pid	Reference_Pid	D
component1	mixture1_1H_...	Sample6_Mi...	H	nan	1	1		7.74	3.75e+05	100	0.001	SH:mixture1...	SP:compone...	0.0
component3	mixture1_1H_...	Sample6_Mix...	H	nan	1	1		8.198	3.48e+05	0	0.001	SH:mixture1_...	SP:compone...	-0.

Target Peak Hits

#	Spectrum	PeakList	Assign F1	Pos F1	LW F1 (Hz)	Height	HeightError	Volume
5	mixture1_1H_Yes_Target	3		7.740	0.163	3.75e+05	None	4.09e+07

Matched Reference Peak

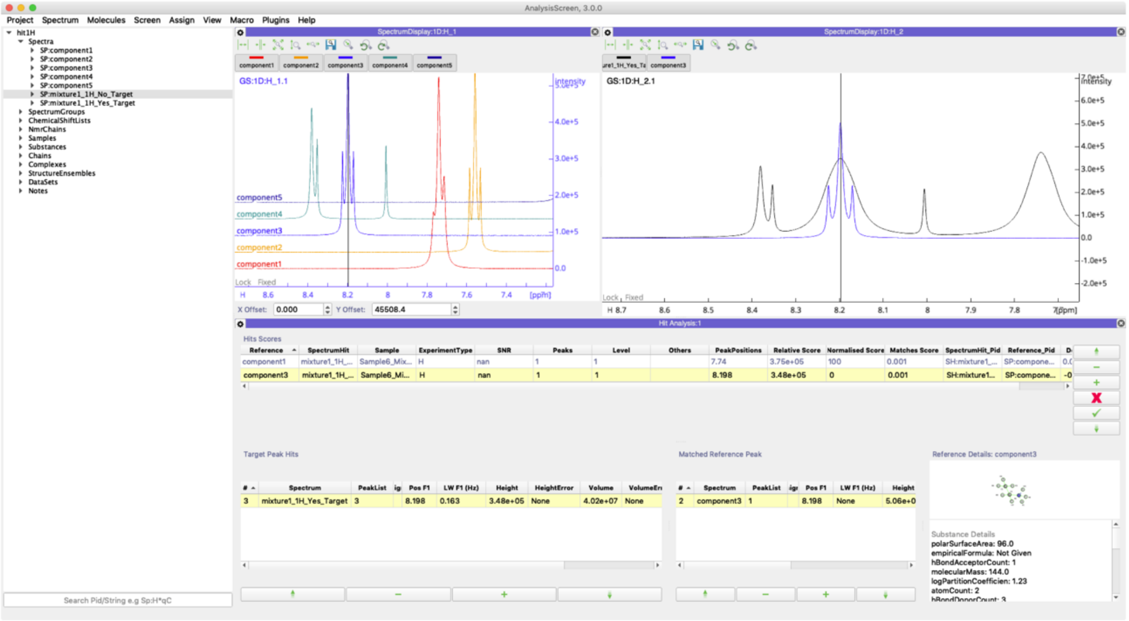
#	Spectrum	PeakList	Assign F1	Pos F1	LW F1 (Hz)
1	component1	1		7.741	None

Reference Details: component1

Substance Details
polarSurfaceArea: 135.0
empiricalFormula: Not Given
hBondAcceptorCount: 3
molecularMass: 185.0
logPartitionCoefficien: 1.45
atomCount: 1

Figure 3.6 Hit analysis results for component 1.

Hit analysis after running a ^{1}H broadening detection pipeline. Top left spectrum display module shows a stack of reference spectra for the sample under examination. Top right spectrum display module shows the ^{1}H spectra for the target experiment and its matched library reference, as calculated by integral value change. The hit analysis module (bottom) displays a summary of various scores for SpectrumHits and matched references properties in the analysis module.



Hits Scores

Reference	SpectrumHit	Sample	ExperimentType	SNR	Peaks	Level	Others	PeakPositions	Relative Score	Normalised Score	Matches Score	SpectrumHit_Pid	Reference_Pid	D-
component1	mixture1_1H_...	Sample6_Mix...	H	nan	1	1		7.74	3.75e+05	100	0.001	SH:mixture1_...	SP:compone...	0.0
component3	mixture1_1H_...	Sample6_Mi...	H	nan	1	1		8.198	3.48e+05	0	0.001	SH:mixture1...	SP:compone...	-0

Target Peak Hits

#	Spectrum	PeakList	ig	Pos F1	LW F1 (Hz)	Height	HeightError	Volume	VolumeEr
3	mixture1_1H_Yes_Target	3		8.198	0.163	3.48e+05	None	4.02e+07	None

Matched Reference Peak

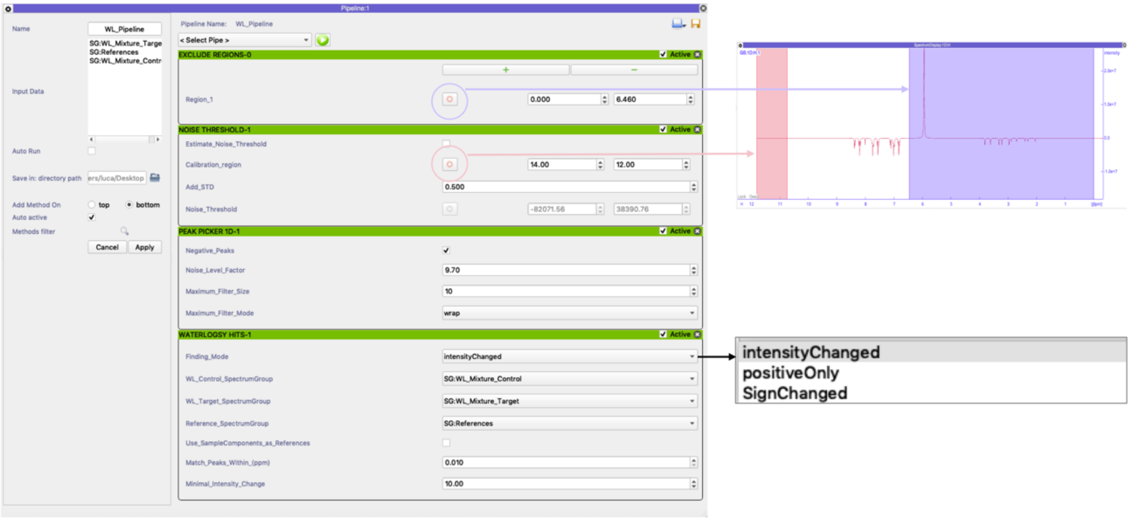
#	Spectrum	PeakList	ig	Pos F1	LW F1 (Hz)	Height
2	component3	1		8.198	None	5.06e+0

Reference Details: component3

Substance Details
polarSurfaceArea: 96.0
empiricalFormula: Not Given
hBondAcceptorCount: 1
molecularMass: 144.0
logPartitionCoefficien: 1.23
atomCount: 2
hBondDonorCount: 3

Figure 3.7 Hit analysis results for component-3.

Results after running a ^{1}H line-broadening detection pipeline displaying component-3 in the Hit Analysis module.



Name: WL_Pipeline

Input Data: SG:WL_Mixture_Targe
SG:References
SG:WL_Mixture_Contr

Auto Run: □

Save in: directory path: ers/luca/Desktop

Add Method On: ○ top ● bottom

Auto active: ☑

Methods filter

Pipeline Name: WL_Pipeline

EXCLUDE REGIONS-0 ☑ Active

Region_1: 0.000 6.460

NOISE THRESHOLD-1 ☑ Active

Estimate_Noise_Threshold: □

Calibration_region: 14.00 12.00

Add_STD: 0.500

Noise_Threshold: -82071.56 38390.76

PEAK PICKER 1D-1 ☑ Active

Negative_Peaks: ☑

Noise_Level_Factor: 9.70

Maximum_Filter_Size: 10

Maximum_Filter_Mode: wrap

WATERLOGSY HITS-1 ☑ Active

Finding_Mode: intensityChanged

WL_Control_SpectrumGroup: SG:WL_Mixture_Control

WL_Target_SpectrumGroup: SG:WL_Mixture_Target

Reference_SpectrumGroup: SG:References

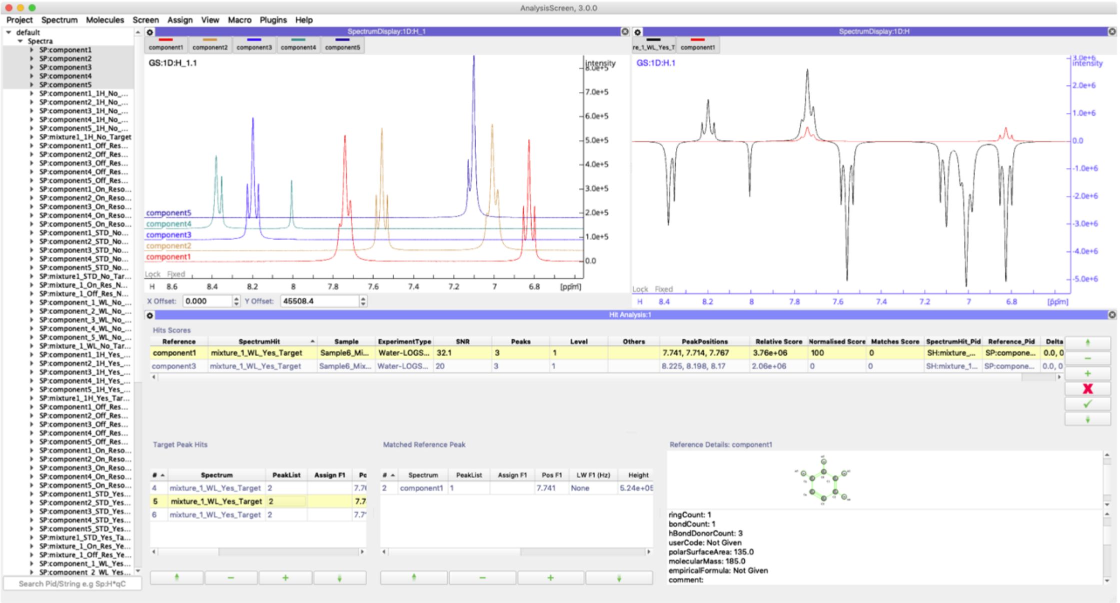
Use_SampleComponents_as_References: □

Match_Peaks_Within_(ppm): 0.010

Minimal_Intensity_Change: 10.00

Figure 3.8 WaterLOGSY hit detection pipeline.

WaterLOGSY data analysis pipeline for detecting hits by comparing signal intensity changes and peak matching by chemical shifts. Interactive manipulation of the exclude regions or noise threshold is facilitated by dedicated buttons highlighted in purple and red in the corresponding pipes, respectively. Different calculation modes for the WaterLOGSY hit detection pipe are displayed in the inset box.


AnalysisScreen, 3.0.0
Project Spectrum Molecules Screen Assign View Macro Plugins Help
SpectrumDisplay:1D:H_1
component1 component2 component3 component4 component5
GS:1D:H_1.1
intensity
component5
component4
component3
component2
component1
H 8.6 8.4 8.2 8 7.8 7.6 7.4 7.2 7 6.8 [ppm]
SpectrumDisplay:1D:H
GS:1D:H.1
3.0e+6 2.0e+6 1.0e+6 0.0 -1.0e+6 -2.0e+6 -3.0e+6 -4.0e+6 -5.0e+6
Hit Analysis:1

X Offset: 0.000 Y Offset: 45508.4

Hits Scores

Reference	SpectrumHit	Sample	ExperimentType	SNR	Peaks	Level	Others	PeakPositions	Relative Score	Normalised Score	Matches Score	SpectrumHit_Pid	Reference_Pid	Delta
component1	mixture_1_WL_Yes_Target	Sample6_Mi...	Water-LOGS...	32.1	3	1		7.741, 7.714, 7.767	3.76e+06	100	0	SH:mixture_...	SP:compone...	0.0, 0
component3	mixture_1_WL_Yes_Target	Sample6_Mix...	Water-LOGS...	20	3	1		8.225, 8.198, 8.17	2.06e+06	0	0	SH:mixture_1...	SP:compone...	0.0, 0

Target Peak Hits

#	Spectrum	PeakList	Assign F1	Pc
4	mixture_1_WL_Yes_Target	2		7.7(
5	mixture_1_WL_Yes_Target	2		7.7
6	mixture_1_WL_Yes_Target	2		7.7'

Matched Reference Peak

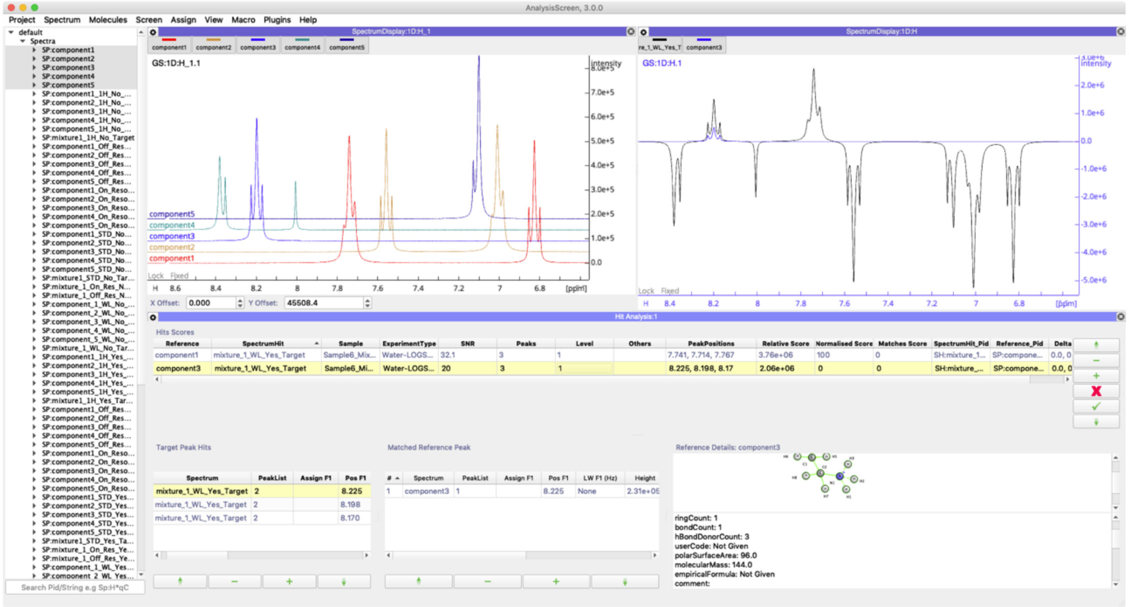
#	Spectrum	PeakList	Assign F1	Pos F1	LW F1 (Hz)	Height
2	component1	1		7.741	None	5.24e+05

Reference Details: component1

ringCount: 1
bondCount: 1
hBondDonorCount: 3
userCode: Not Given
polarSurfaceArea: 135.0
molecularMass: 185.0
empiricalFormula: Not Given
comment:

Figure 3.9 Hit analysis results for component-1 by WaterLOGSY.

Hit analysis after running a WaterLOGSY hit detection pipeline. Top left spectrum display module shows the reference spectra for the sample under examination. Top right spectrum display shows the WaterLOGSY spectrum for the target experiment and its matched reference hit. The hit analysis module (bottom) displays a summary of various scores for SpectrumHits and matched references properties in the analysis module.



X Offset: 0.000 Y Offset: 45508.4

Hits Scores

Reference	SpectrumHit	Sample	ExperimentType	SNR	Peaks	Level	Others	PeakPositions	Relative Score	Normalised Score	Matches Score	SpectrumHit_Pid	Reference_Pid	Delta
component1	mixture_1_WL_Yes_Target	Sample6_Mix...	Water-LOGS...	32.1	3	1		7.741, 7.714, 7.767	3.76e+06	100	0	SH:mixture_1...	SP:compone...	0.0, 0
component3	mixture_1_WL_Yes_Target	Sample6_Mi...	Water-LOGS...	20	3	1		8.225, 8.198, 8.17	2.06e+06	0	0	SH:mixture_...	SP:compone...	0.0, 0

Target Peak Hits

Spectrum	PeakList	Assign F1	Pos F1
mixture_1_WL_Yes_Target	2		8.225
mixture_1_WL_Yes_Target	2		8.198
mixture_1_WL_Yes_Target	2		8.170

Matched Reference Peak

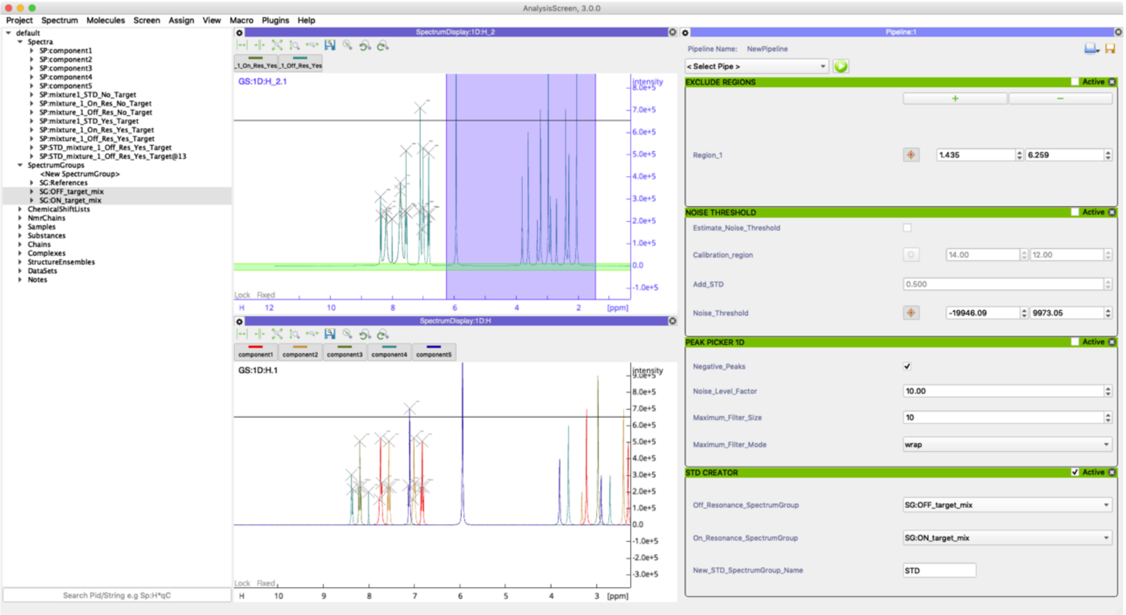
#	Spectrum	PeakList	Assign F1	Pos F1	LW F1 (Hz)	Height
1	component3	1		8.225	None	2.31e+05

Reference Details: component3

ringCount: 1
bondCount: 1
hBondDonorCount: 3
userCode: Not Given
polarSurfaceArea: 96.0
molecularMass: 144.0
empiricalFormula: Not Given
comment:

Figure 3.10 Hit analysis results for component-3 by WaterLOGSY.

Results after running a WaterLogsy hit detection pipeline displaying component-3 in the Hit Analysis module.



Pipeline Name: NewPipeline

EXCLUDE REGIONS □ Active

Region_1 1.435 6.259

NOISE THRESHOLD □ Active

Estimate_Noise_Threshold □

Calibration_region 14.00 12.00

Add_STD 0.500

Noise_Threshold -19946.09 9973.05

PEAK PICKER 1D □ Active

Negative_Peaks ☑

Noise_Level_Factor 10.00

Maximum_Filter_Size 10

Maximum_Filter_Mode wrap

STD CREATOR ☑ Active

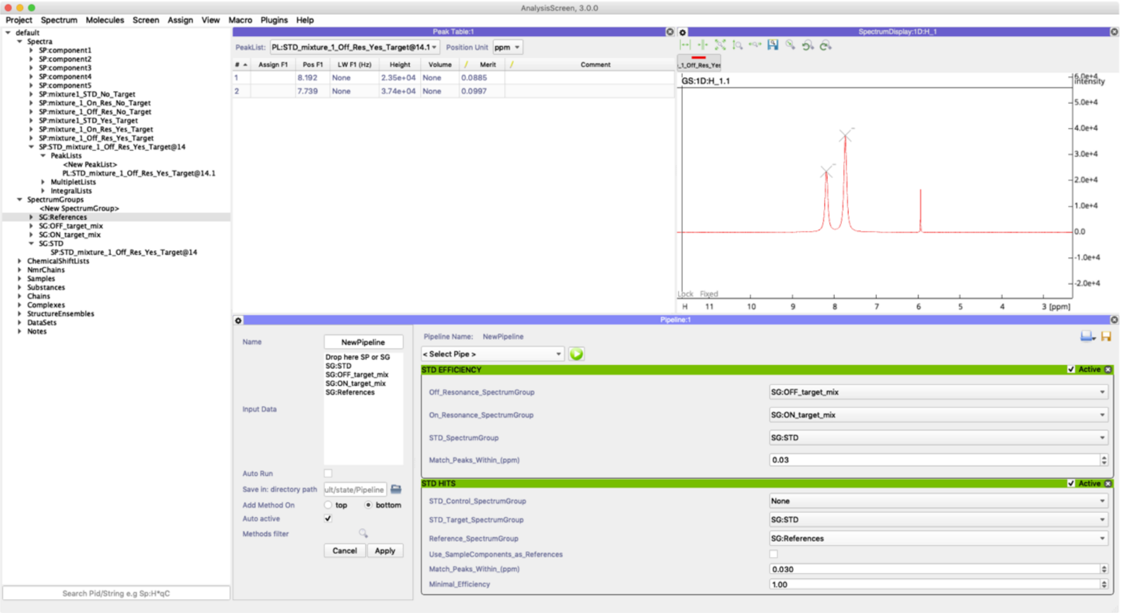
Off_Resonance_SpectrumGroup SG:OFF_target_mix

On_Resonance_SpectrumGroup SG:ON_target_mix

New_STD_SpectrumGroup_Name STD

Figure 3.11 STD creation pipeline.

Simulated On/Off resonance spectra and interactive items needed to drive the peak detection routine (top spectrum display module), library reference spectra (lower spectrum display module), and an STD creator pipeline used in the data analysis (right-hand side).



PeakList: PL:STD_mixture_1_Off_Res_Yes_Target@14.1 Position Unit ppm

#	Assign F1	Pos F1	LW F1 (Hz)	Height	Volume	Merit	Comment
1		8.192	None	2.35e+04	None	0.0885	
2		7.739	None	3.74e+04	None	0.0997	

Name: NewPipeline

Input Data: Drop here SP or SG
SG:STD
SG:OFF_target_mix
SG:ON_target_mix
SG:References

Auto Run □

Save in: directory path ult/state/Pipeline

Add Method On ○ top ● bottom

Auto active ☑

Methods filter

Cancel Apply

Pipeline Name: NewPipeline

< Select Pipe >

STD EFFICIENCY ☑ Active

Off_Resonance_SpectrumGroup: SG:OFF_target_mix

On_Resonance_SpectrumGroup: SG:ON_target_mix

STD_SpectrumGroup: SG:STD

Match_Peaks_Within_(ppm): 0.03

STD HITS ☑ Active

STD_Control_SpectrumGroup: None

STD_Target_SpectrumGroup: SG:STD

Reference_SpectrumGroup: SG:References

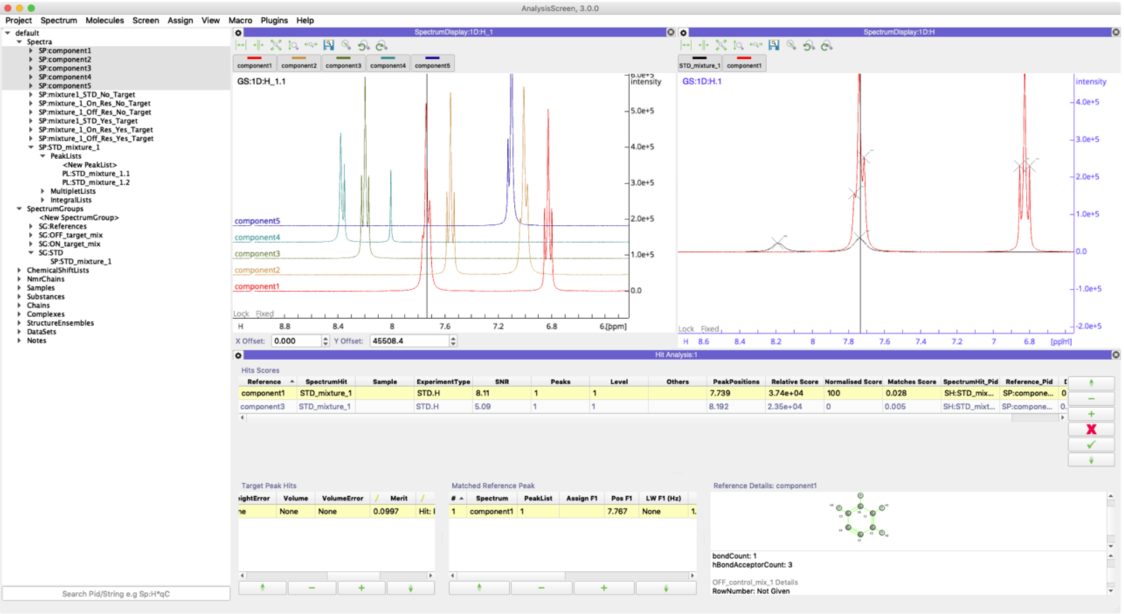
Use_SampleComponents_as_References □

Match_Peaks_Within_(ppm): 0.030

Minimal_Efficiency: 1.00

Figure 3.12 STD hit detection pipeline.

Top left: a peak table displaying calculated STD efficiencies as peak “figure of merit”. Top right: a newly created STD spectrum. Bottom: an example of pipeline built for calculating the STD efficiency and hit detection.



X Offset: 0.000 Y Offset: 45508.4

Reference	SpectrumHit	Sample	ExperimentType	SNR	Peaks	Level	Others	PeakPositions	Relative Score	Normalised Score	Matches Score	SpectrumHit_Pid	Reference_Pid	D
component1	STD_mixture_1		STD.H	8.11	1	1		7.739	3.74e+04	100	0.028	SH:STD_mix...	SP:compone...	0
component3	STD_mixture_1		STD.H	5.09	1	1		8.192	2.35e+04	0	0.005	SH:STD_mixt...	SP:compone...	0.

ightError	Volume	VolumeError	/	Merit	/
ne	None	None		0.0997	Hit: I

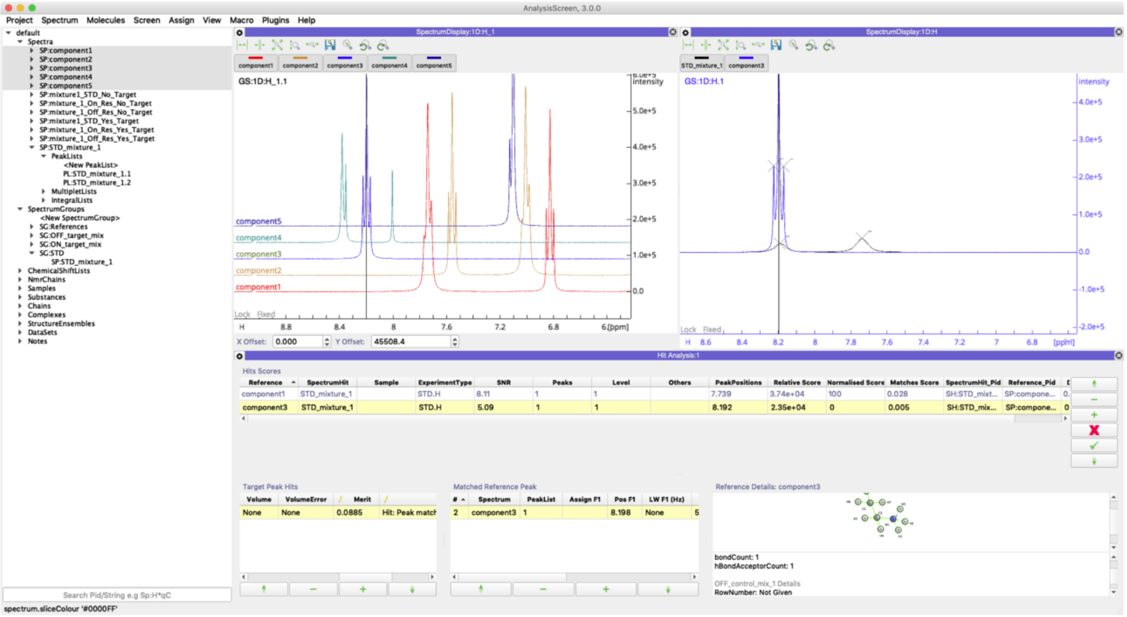
#	Spectrum	PeakList	Assign F1	Pos F1	LW F1 (Hz)	
1	component1	1		7.767	None	1.

bondCount: 1
hBondAcceptorCount: 3

OFF_control_mix_1 Details
RowNumber: Not Given

Figure 3.13 Hit analysis results for component-1 by STD.

Hit analysis after running an STD hit detection pipeline. Top left spectrum display module shows the reference spectra for the sample under examination. Top right spectrum display shows the STD spectrum for the target experiment and its matched reference hit. The hit analysis module (bottom) displays a summary of various scores for SpectrumHits and matched references properties in the analysis module.



AnalysisScreen, 3.0.0
Project Spectrum Molecules Screen Assign View Macro Plugins Help
default
Spectra
SP:component1
SP:component2
SP:component3
SP:component4
SP:component5
SP:mixture1_STD_No_Target
SP:mixture_1_On_Res_No_Target
SP:mixture_1_Off_Res_No_Target
SP:mixture1_STD_Yes_Target
SP:mixture_1_On_Res_Yes_Target
SP:mixture_1_Off_Res_Yes_Target
SP:STD_mixture_1
PeakLists
<New PeakList>
PL:STD_mixture_1.1
PL:STD_mixture_1.2
MultipletLists
IntegralLists
SpectrumGroups
<New SpectrumGroup>
SG:References
SG:OFF_target_mix
SG:ON_target_mix
SG:STD
SP:STD_mixture_1
ChemicalShiftLists
NmrChains
Samples
Substances
Chains
Complexes
StructureEnsembles
DataSets
Notes
Search Pid/String e.g Sp:H*qC
spectrum.sliceColour '#0000FF'
SpectrumDisplay:1D:H_1
component1 component2 component3 component4 component5
GS:1D:H_1.1
component5
component4
component3
component2
component1
Lock Fixed
H 8.8 8.4 8 7.6 7.2 6.8 6.[ppm]
intensity 5.0e+5 4.0e+5 3.0e+5 2.0e+5 1.0e+5 0.0
X Offset: 0.000 Y Offset: 45508.4
SpectrumDisplay:1D:H
STD_mixture_1 component3
GS:1D:H.1
intensity 4.0e+5 3.0e+5 2.0e+5 1.0e+5 0.0 -1.0e+5 -2.0e+5
Lock Fixed
H 8.6 8.4 8.2 8 7.8 7.6 7.4 7.2 7 6.8 [ppm]
Hit Analysis:1
Reference Details: component3
+ − ✗ ✓

Hits Scores

Reference	SpectrumHit	Sample	ExperimentType	SNR	Peaks	Level	Others	PeakPositions	Relative Score	Normalised Score	Matches Score	SpectrumHit_Pid	Reference_Pid	D
component1	STD_mixture_1		STD.H	8.11	1	1		7.739	3.74e+04	100	0.028	SH:STD_mixt...	SP:compone...	0.
component3	STD_mixture_1		STD.H	5.09	1	1		8.192	2.35e+04	0	0.005	SH:STD_mix...	SP:compone...	0

Target Peak Hits

Volume	VolumeError	/	Merit	/
None	None		0.0885	Hit: Peak match

Matched Reference Peak

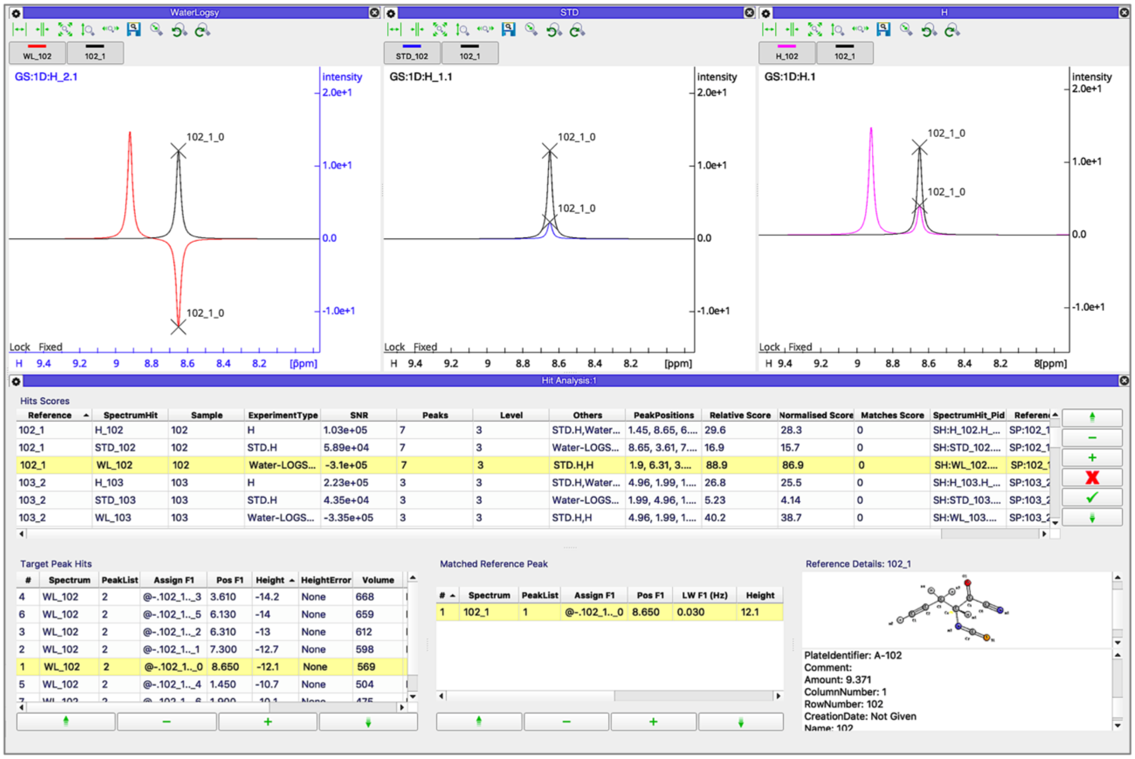
#	Spectrum	PeakList	Assign F1	Pos F1	LW F1 (Hz)	
2	component3	1		8.198	None	5

bondCount: 1
hBondAcceptorCount: 1

OFF_control_mix_1 Details
RowNumber: Not Given

Figure 3.14 Hit analysis results for component-3 by STD.

Results displayed in the hit analysis module for the component-3 after running an STD hit detection pipeline.



Hits Scores

Reference	SpectrumHit	Sample	ExperimentType	SNR	Peaks	Level	Others	PeakPositions	Relative Score	Normalised Score	Matches Score	SpectrumHit_Pid	Referen
102_1	H_102	102	H	1.03e+05	7	3	STD.H,Water...	1.45, 8.65, 6....	29.6	28.3	0	SH:H_102.H_...	SP:102_1
102_1	STD_102	102	STD.H	5.89e+04	7	3	Water-LOGS...	8.65, 3.61, 7....	16.9	15.7	0	SH:STD_102....	SP:102_1
102_1	WL_102	102	Water-LOGS...	-3.1e+05	7	3	STD.H,H	1.9, 6.31, 3....	88.9	86.9	0	SH:WL_102....	SP:102_1
103_2	H_103	103	H	2.23e+05	3	3	STD.H,Water...	4.96, 1.99, 1....	26.8	25.5	0	SH:H_103.H_...	SP:103_2
103_2	STD_103	103	STD.H	4.35e+04	3	3	Water-LOGS...	1.99, 4.96, 1....	5.23	4.14	0	SH:STD_103....	SP:103_2
103_2	WL_103	103	Water-LOGS...	-3.35e+05	3	3	STD.H,H	4.96, 1.99, 1....	40.2	38.7	0	SH:WL_103....	SP:103_2

Target Peak Hits

#	Spectrum	PeakList	Assign F1	Pos F1	Height	HeightError	Volume
4	WL_102	2	@-.102_1..._3	3.610	-14.2	None	668
6	WL_102	2	@-.102_1..._5	6.130	-14	None	659
3	WL_102	2	@-.102_1..._2	6.310	-13	None	612
2	WL_102	2	@-.102_1..._1	7.300	-12.7	None	598
1	WL_102	2	@-.102_1..._0	8.650	-12.1	None	569
5	WL_102	2	@-.102_1..._4	1.450	-10.7	None	504
7	WL_102	2	@-.102_1..._6	1.900	-10.1	None	475

Matched Reference Peak

#	Spectrum	PeakList	Assign F1	Pos F1	LW F1 (Hz)	Height
1	102_1	1	@-.102_1..._0	8.650	0.030	12.1

Reference Details: 102_1

PlateIdentifier: A-102
Comment:
Amount: 9.371
ColumnNumber: 1
RowNumber: 102
CreationDate: Not Given
Name: 102

Figure 3.15 Hit analysis module on large dataset.

Current Hit Analysis module graphical user interface containing a report of 1000 simulated samples for three different experiment types. The Hit Analysis module allows interactive inspection and assessment of SpectrumHits showing spectra, scores and associated metadata. Furthermore, custom peak tables (bottom) allow quick navigation through the peak hits in the selected spectrum display. A summary for the sample and SpectrumHit properties is shown in the bottom right corner.

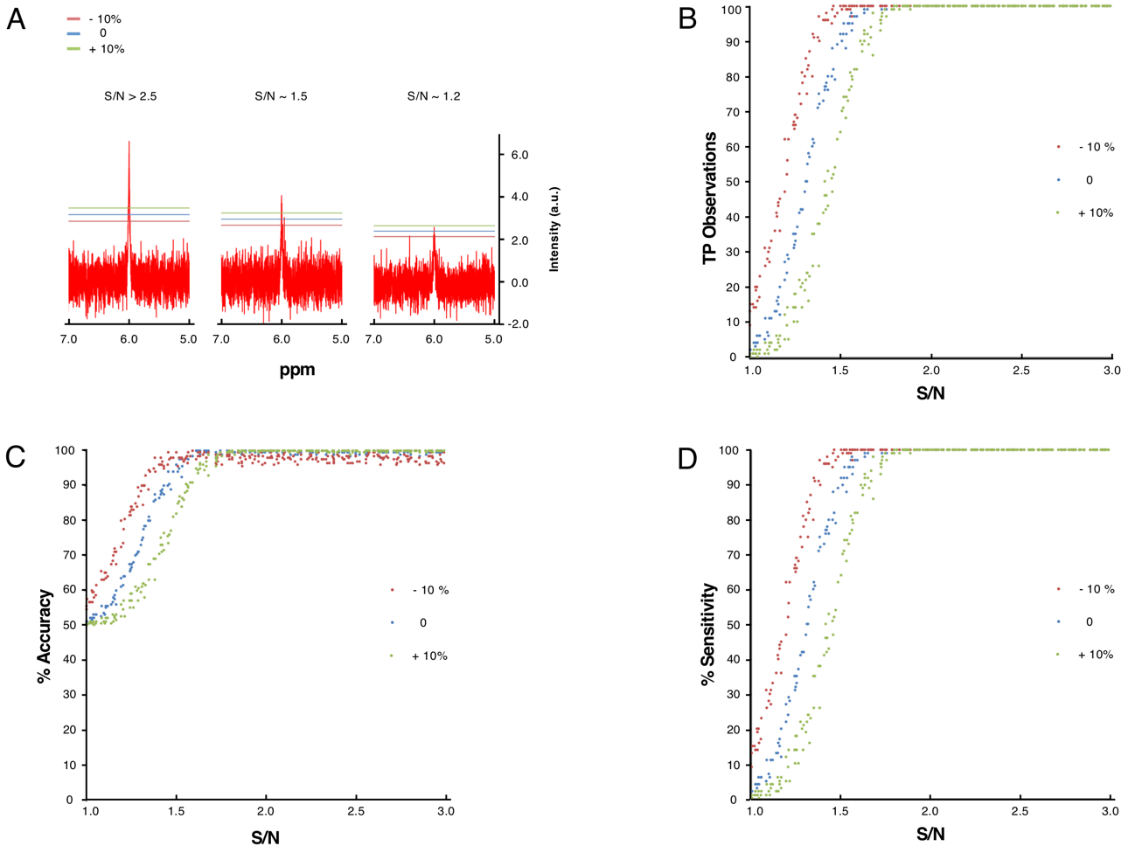


Figure 3.16 Peak and hit detection assessment using simulated spectra.

A Simulated ^{1}H spectra at different signal-to-noise ratios and estimated positive noise thresholds calculated using Eq. 3.1, with α set to 1.5 (blue), relative adjustment N_{Th+10} = +10% N_{Th} (green) and N_{Th-10} = -10% (red). The left panel shows typical spectral peaks with an S/N greater than 2.5. Peak intensities are well above threshold values and peaks are correctly identified. At around a S/N of 1.5, most of the peaks are still identified, although a larger number of artefacts can be mistakenly included as real peaks. At very low S/N it is generally difficult to distinguish genuine peak-shapes from the spectral noisy distortions. **B** Total count of correctly identified observations for 100 simple spectra simulated at over 20000 different S/N variations. **C** Total accuracy for the peak picker on simulated spectra at different delta values. Accuracy (A) was defined as A = (TP+TN) / (TP+FN+FP+TN). **D** Total sensitivity for the peak picker on simulated spectra. Sensitivity (S) was calculated as S = TP / (TP+FN), with TP, TN, FP and FN denoting true positive, true negative, false positive, and false negative values, respectively.

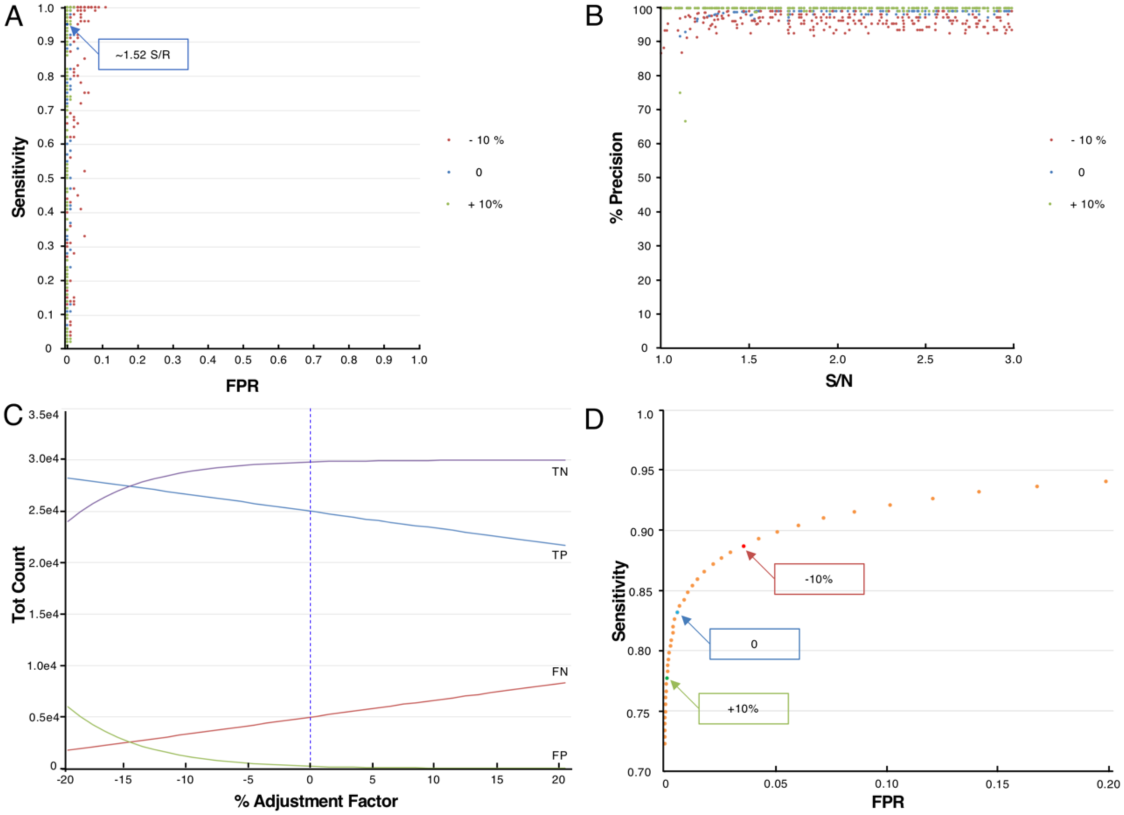


Figure 3.17 Peak detection statistics.

The correctness of the automatically determined peak detection noise threshold value was inspected using receiver operating characteristic (ROC) scoring. It allows to determine the capacity of a binary classifier followed by an adjustment in the threshold value[127]. In the ROC curve the Sensitivity is plotted against the False Positive Rate (FPR). Sensitivity, False Positive Rate, Specificity and Precision, are calculated from the true positive (TP), true negative (TN), false positive (FP), false negative (FN) values. **A** ROC plot for performance of the algorithm at different S/N ratios. Blue arrow indicates the score for a spectrum at ~1.5 S/N. The Sensitivity is calculated as TP/(TP+FN); the False Positive Rate, FPR is calculated as 1- Specificity, where the Specificity is calculated as TN/(TN+FP); **B** Precision of the peak picker for spectra at different S/N ratios.

Precision was calculated as TP/(TP+FP). Notably, the precision is badly compromised when lowering the threshold value by 10% (blue data-points; -10%) at all S/N ratios. **C** Total count for true positives, true negatives, false positives and false negatives for each run of peak picking using different adjustments of the automatically determined noise threshold value. **D** ROC curve for the total sensitivity and false positive rate, as expected the automatically determined threshold value is located in the most optimal location of the ROC curve, suggesting it can be used reliably in the automatic peak picking routine without need for adjustment.

3.7. Acknowledgments

I thank Dr TJ Ragan[a] for comments and suggestions on implementing the initial pipeline architecture and its GUI representation.

3.8. Addresses

[a] Leicester Institute of Structural and Chemical Biology, Department of Molecular and Cell Biology, University of Leicester, Henry Wellcome Building, Lancaster Road, Leicester LE1 7HN, United Kingdom.

Chapter 4

A versatile baseline correction for 1D NMR screening

Keywords: FBDD NMR, Baseline Correction, CcpNmr.

4.1. Abstract

Flat baselines are an absolute requirement for any manual or automatic FBDD screening protocol. Several previously published algorithms failed to properly correct our collaborator's experimental NMR screening datasets. In this chapter I present a new baseline correction algorithm which is capable of correcting NMR spectra that are highly distorted without the need of any user parameter input or adjustment, making it an ideal choice in any automated screening analysis workflow. The algorithm was tested on 36 experimental NMR spectra and corrections were compared qualitatively and quantitatively to results outputted by other several published algorithms.

4.2. Introduction

One-dimensional (1D) fragment-based NMR screening studies are characterised by manual or automated comparison of reference or control spectra to those spectra that form the experimental observations to detect a potential ligand binding to a target. Prior to any analysis, it is fundamental to inspect and correct any experimental artefacts in the datasets in order to avoid false positives and negatives. Using the decomposition module (cf. chapter 2) it is possible to detect various spectral artefacts (Figs 2.5-2.6). These errors can slow down the whole process of drug discovery, or in severe cases they can void the entire trial. One of the most common spectral artefacts is an irregular baseline pattern.

Theoretically, the experimental noise should follow a random Gaussian distribution with a zero mean, resulting in spectra where regions devoid of signal display randomly scattered points as a function of frequency centred around zero. This is commonly referred to as a straight or flat baseline. However, a series of phenomena, such as a hardware instability and the corruption of the initial data-points in the free induction decay (FID) can cause underpinning baseline errors[128]. For the early time-domain points of signal acquisition, in fact, the electronics can still be recovering from the application of the RF pulse, and improper optimisation of the instrument parameters can result in errors in the Fourier-transformed spectrum.

Automatic algorithms for signal correction can be divided therefore in two groups: time-domain correction and frequency domain correction[128–130]. Time-domain methods aim to correct the corrupted data-points of the FID, whereas frequency domain methods are based on the reconstruction of the baseline only and its subsequent use in correction of the original spectrum.

Over the years, several algorithms and manual baseline correction tools have been described[128–130] and included in commercial software packages such as MestreNova[131].

In this chapter, I discuss the testing of several previously published automatic frequency-domain algorithms, including the Whittaker Smooth algorithm, the Asymmetric Least Squares Smoothing[132] ALS, the adaptive iteratively reweighted Penalized Least Squares airPLS[133], the asymmetrically reweighted Penalized Least Squares Smoothing arPLS[134] and finally the Distribution-Based Classification DBC[135].

I applied these methods to our datasets and compared their outputs to the results obtained from a newly proposed approach referred to as Correlated Weighted baseline correction (CWBC; *vide infra*). Importantly, in contrast to the published methods the CWBC method does not require pre-knowledge of signal location and most importantly, does not require any user-adjustable parameters. From a qualitative and quantitative analysis, the CWBC method yielded superior results when tested on an experimental screening dataset.

4.3. Materials and Methods

4.3.1. **Materials**

The experimental dataset used for testing the set of algorithms was composed of 13 ^{1}H, 11 $T_{1\rho}$ and 12 WaterLOGSY one-dimensional spectra. In total, 5 substances were reported (named: 314, 441, 467, 7372, 7373); for each substance a reference spectrum, plus spectra acquired in presence and absence of a protein target using three different experiment types was reported. Target and substances molecular structures were undisclosed and not required for any of validations performed in this work. Spectra files were provided as CSV files and they have been processed by the collaborators.

For each algorithm tested, a new corrected spectrum was created using the CcpNmr Analysis core capabilities.

4.3.2. **Methods**

Spectra and algorithms were analysed using CcpNmr Analysis Screen 3.0. Additional scripts encoding the various algorithms and for quantitative analysis of the outputs were written using the *Signal*, *Stat* and *Spatial* functions present in the SciPy included in the libraries of the CcpNmr Python environment.

The tested algorithms, such as WhittakerSmooth, airPLS, arPLS and ALS were recreated from various sources found in the public domain and inserted in the main CcpNmr library. The core algorithm DBC was natively distributed in the package NMRGlue[118]. All algorithms tested are now available for users from the CcpNmr AnalysisScreen distribution.

The parameters used for the comparison, were selected as suggested by the original authors, or optimised to achieve the best level of qualitative correction:

Table 4.1 Baseline correction algorithms parameters.

Method	Parameter 1	Parameter 2	Parameter 3
als	lambda =10 ** 2	p=0.001	nIter=1
WhittakerSmooth	lambda_=100	differences=1	
arPLS	lambda=5.e5	ratio=1.e-6	nIter =50
airPLS	lambda=100	porder=1	nIter =15
dbc	wd=20		

Correlated Weighted baseline correction (CWBC) method

The CWBC method avails mainly of the linear cross-correlation function[136] to fit the original spectral data points, $\bar{O}$, and an array of "ones", $\bar{1}$, to estimate a baseline vector as an array $\bar{E}$:

$$\bar{E} = \frac{\bar{O} \otimes \bar{1}}{c}$$

Eq. 4-1

Where $\otimes$ denotes the cross-correlation between two discrete arrays $\bar{O}$ and $\bar{1}$, c is a constant minimised to 0.0039.

Scorings

To quantify the differences from the corrected spectrum to its original, I used the Spearman rank-order correlation coefficient, r_s. The Spearman correlation measures the monocity of the relationship between two curves[137]. This type of correlation differs from others, like the Pearson, as the two comparing curves do not need to be normally distributed, and the correlation score is given by the measurement of correspondence between ranking. It is calculated as following:

$$r_s = \frac{cov(rg_x, rg_y)}{\delta rg_x \, \delta rg_y}$$

Eq. 4-2

Where $rg_{x/y}$ denote the rank values of the data points of curves X and Y, respectively, *cov* is the covariance and δ is the standard deviation[138,139].

To quantify the level of baseline correction in a region where signal is not expected, e.g. the first 1000 data points of the spectrum, I estimated a noisy region by creating a random normal Gaussian distribution centred around zero Y intensity, with same shape and length as the input experimental observations. I, then, computed the Euclidean distance between the two vectors. The distance was calculated as

$$E = \sqrt{(Y_1 - Y_2)^2} = |Y_1 - Y_2|$$

Eq. 4-3

Where Y_1 and Y_2 are the two curves.

The final score was normalised as following:

$$Es = \left|\frac{E_{max}}{E}\right|$$

Eq. 4-4

Where E_{max} is the max distance recorded for the various experiments.

Two Spearman ranking scores were calculated to quantify the correlation between lineshapes, one for the whole Spectrum S_w, and one for a portion of the aromatic field, S_a, and the baseline score E_s. The total score was calculated as:

$$S = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Sw + Sa + Es}{3}\right)$$

Eq. 4-5

where n denotes the total number of available experiments.

4.3.3. **New baseline correction algorithm steps**

The new method consists of several estimations of the baseline pattern while excluding the signal from the spectrum. The first baseline approximation is calculated through a cross correlation method. The subtraction of the estimated baseline curve from the original spectrum allows to distribute uniformly the noise and signal across the zero-y-axis coordinate. This permits to estimate the positive and negative noise threshold values, which are then used to establish where the signal might reside in the spectrum. Detected signal regions are weighted to a zero value. This information is then transferred to the original spectrum, where estimated signal regions are removed from the spectrum. The spectrum now consists only of noise and "gaps" which are filled with interpolated values; the interpolation is necessary to preserve the original signal intensity pattern. A further cross correlation of this spectrum ensures a smoother estimation of the baseline, removing other potential signal artefacts which might compromise the final result. Lastly, the subtraction of the latter estimation from the original data results in a baseline-corrected spectrum.

The eight steps, describing the CWBC method, are summarised in Fig. 4.2:

1. Initial baseline estimation: create an array containing a subset of the discrete linear cross-correlation, E_1, between the Original data-points $\bar{O}$, and a simulated array of the same shape $\bar{1}$, Fig. 4.3A.
    1.1. Correct the left and right edges on the E_1 curve. $E_{1L/R}$ 50 pts.

1.2. Create a subset of arrays $E_1s_{L/R}$, by linear correlation for both $E_{1L/R}$, Fig. 4.3A-insert.

1.3. Reshape $E_1s_{L/R}$ by linear interpolation (20 pts), to the original input length (50 pts), Fig. 4.3B.

2. Subtract Estimation from Original data, $\bar{R}_1 = \bar{O} - \bar{E}_1$, Fig. 4.4A.
3. Estimate the baseline noise thresholds (minimum and maximum) from R_1, Fig. 4.4A-insert.

3.1. Thresholds are obtained from subdividing R_1 and calculating the mean of minimum and maximum values of bins where no signal is found.

4. Assign (weight) a zero value for $\bar{R}_1$ signal, $\bar{S}$, above the threshold values and adjacent points. The new vector $\bar{E}_2$, is an estimation of the noise without the signal: $\bar{E}_2 = \bar{R}_1 \not\subseteq \bar{S}$, Fig. 4.4B.

4.1. Adjust outliers among groups of zeros. Find if small groups of non-zeros are between two large groups of zeros, then set them to zero, Fig. 4.4B-insert.

5. Find the Zero values (corresponding to the signal) in E_2 and transfer this information to E_1. The new array E_3 will present masked "gaps". Fig. 4.5A.
6. Correct gaps using a linear interpolation, E_{3i}, of adjacent points so that the intensities are preserved in the following step, Fig. 4.5B.
7. Final baseline Estimation E_4. E_4 created by cross correlating the latest E_{3i} and a simulated array as in step 1, Fig. 4.6.A
8. Final correction, the corrected $\bar{R}_2$ is given by subtracting the latest cross correlated output E_4 from the original $\bar{O}$, $\bar{R}_2 = \bar{O} - \bar{E}_4$, Fig. 4.6B.

4.4. Results and Discussion

4.4.1. Qualitative analysis of the various methods

The qualitative analysis of the performance of the various baseline correction methods were carried out on the ^{1}H, $T_{1\rho}$ and WaterLOGSY spectra of component-314 in the absence (control) and in the presence (target) of a macromolecule.

Figure 4.7 shows the results of the application of a set of 1st, 2nd and 3rd order polynomial fitting functions. These functions were used as a control before applying more complex algorithms and were not expected to apply any level of valid correction. In fact, the polynomial fitting resulted in heavily distorted baselines for all three experiment types, irrespective of the polynomial order of the fit. Where the 1st order polynomial fit resulted in the baseline below the zero intensity, albeit keeping the original spectral appearance intact, the higher polynomial orders introduced a “rolling” distortion across the whole spectrum. This phenomenon was even more pronounced in the ^{1}H experiment, which already suffered from phasing errors.

I next tested the Whittaker smooth algorithm (Fig. 4.8). This produced an almost perfect flat baseline across the whole spectrum for each of the three experiments. However, this method appeared to be a destructive approach with respect to data content as it resulted in a highly compromised data-points in the signal-containing regions of the spectra and the appearance of artefacts of opposite sign to the original signal. Therefore, the resulting spectra could not be used for any reliable screening interpretations.

The following tested algorithm was the Asymmetric Least Squares Smoothing, ALS[132]. This algorithm is an implementation of the Whittaker smoother, which is used to estimate the baseline, followed by a numerical weighting of signals where positive observations are weighted less than negative signals. It requires two parameters, one for the flexibility of the baseline and one for the position. These parameters cannot be automatically selected and authors claimed that human judgment is always required[132].

The ALS method (Fig. 4.9) appeared to be a more suitable algorithm as it was able to rectify the baseline for both experimental spectra. It correctly preserved the shapes for positive peaks; however, it did not appear to be optimised for dealing with negative peaks simultaneously. The algorithm produced a significant distortion on the negative peaks, making the algorithm unusable for experiments like WaterLOGSY, for which the distinction between intensity signs is fundamental for the hit analysis classification. Therefore, this type of error will result in the introduction of false positives upon the hit identification process. Furthermore, rolling artefacts appeared in all signal-free regions (cf. Fig. 4.9A). Lastly, the baseline noise was not correctly distributed around zero but instead appeared erroneously to be increased to positive values.

In 2009 Zhang *et al.* proposed the adaptive iteratively reweighted Penalized Least Squares airPLS[133]. The algorithm uses a weighting system for the signal regions, and it works by iteratively adjusting weights of sum squares errors between the experimental curve and its fitted baseline. These weights are then derived by using the difference between the earlier fittings and the original regions of interests. The performance of this method is dependent on a crucial parameter for the smoothness, usually referred as *lambda* (λ), and the iteration count[133].

All the previous methods present a major issue in portions of spectra when no peaks are detected which can result in an underestimated correction[134]. Baek, S.J. *et al* (2014) proposed the partially balanced but asymmetric weights, arPLS algorithm[134] algorithm as a solution. The algorithm is based on the assumption that in spectral regions without peaks the noise is distributed uniformly above and below a baseline and is given a weight, whereas is not given if signals are greater than the baseline. As the previous methods, arPLS requires a proper value of a critical parameter *lambda*, which is used to tune the balance between fitness and smoothness of the baseline. In addition, the user-optimised *ratio* parameter determines the end to the iterations when the weights have reached a minimal value[134].

However, similar problems to the ASL were observed using both airPLS and its alternative implementation arPLS (Figs 4.10-4.11). Both these algorithms did not perform as expected in presence of positive and negative peaks in the same spectrum. The inadequate spectral solvent suppression further enhanced the weakness of these methods, resulting in severe spectral distortions across all experiments. Spectra with a confined positive signal, such as the ^{1}H and $T_{1\rho}$ aromatic fields, were corrected adequately and peak-shapes were preserved. However, since the variegate nature of NMR datasets, were spectral patterns can differ hugely among the trial, for example due to differences in signal to noise ratio, phasing or solvent suppression, these two methods cannot used on automated screening routines as they will require a continuous optimisation of parameters to obtain acceptable results; as a consequence they should be used by expert users only for the manual analysis of limited cases.

A dedicated NMR approach aimed to correct the baselines of 1D ^{1}H metabolomics data was described by Wang and *et al.*[135]. Their method estimates the baseline pattern by calculating a distribution of standard deviations describing the spectrum's noise. The method, called Distribution-Based Classification (DBC), calculates the standard deviations of the spectral intensities across a sliding-window of the spectrum. Only a single user-defined window-size parameter is required for this algorithm, making it more promising for automation in screening by NMR.

The method was the last tested algorithm. It performed adequately on the $T_{1\rho}$ experiment (Fig. 4.12A) but showed various degrees of baseline rolling when applied to the ^{1}H and WaterLOGSY spectra (Figs 4.14B-C). Furthermore, large artefacts appeared at the edge of the $T_{1\rho}$ and ^{1}H spectra, similar to the so-called spectral “smile” typically observed in the spectra originating from the NMR manufacturer Bruker. This artefact could be considered only to be an aesthetic aspect that does not compromise a manual inspection of spectra. However, it could introduce false positive signals in automated signal recognition routines.

The CWBC method did not result in any of the previously described issues (Fig. 4.13). For all three experiment types baselines were restored around zero intensity, smiles were not present, and spectral signals were correctly preserved. Only the large phase error in the ^{1}H experiment prevented a full baseline correction around the water signal; however, the line-shapes of actual signal containing peaks were not distorted.

Unarguably, phasing errors derive from fundamentally different spectroscopic phenomena, and therefore different algorithms should be used to correct this error, preferably prior to the application of any baseline correction methods.

Using the $T_{1\rho}$ spectrum of component-314 the line-shapes of the original spectrum and the CWBC spectrum were compared (Fig. 4.14). The aromatic region between 8.8 and 6 ppm was perfectly devoid of any offset and the rolling artefact was completely eliminated (Fig. 4.14A). The negative signal in the region at around 5 ppm was maintained and adjacent positive peaks were correctly present without any distortions, proving the ability of the algorithm to correct accurately both positive and negative signals simultaneously (Fig. 4.14B). Finally, complex spectral patterns, both in the aliphatic region around 3.5 ppm and the aromatic region at 7.2 ppm (Figs 4.14C-D), were properly maintained in the corrected spectrum, thus confirming that the algorithm was able to reliably correct the baseline errors and preserve accurately the crucial signals needed for a qualitative and quantitative analysis as required in screening by NMR.

4.4.2. **Quantitative analysis of the various methods**

The level of correction for each algorithm has been assessed through three different approaches: similarities among the whole spectral intensities, referred as “whole value”, similarities for a confined downfield signal region, referred as “aromatic value” and, baseline distances in known noise regions between original and corrected spectra, referred as “baseline value”. The higher baseline distance value suggests a higher degree of transformation of the spectrum; the value is inversely proportional to the distance measured to the simulated Gaussian curve at the zero intensities axis. High values relative to the whole and aromatic indicate a closer pattern between the two observations. These scorings have been designed so that in a scale 0 to 1, the perfect correction should show all three values as high as possible.

The first test has been carried out on spectra relative to the component-314 for the three experiment types in the presence of the target. The choice of using spectra derived by samples in presence of the target assured to have a higher amount of signal and therefore assessing more potential critical areas. Although, the receptor signal should have properly suppressed using appropriate filters for this type of ligand-based experiments, and control and target spectra should have shown (about) the same number of spectral peaks.

The quantitative analysis confirmed the visual examination for the $T_{1\rho}$ experiment, Fig. 4.15A. The CWBC algorithm appeared to be consistently superior in all three aspects tested for this experiment, with over 70% of the total spectral and over 90% of aromatic line shapes preserved, plus an excellent correction of the baseline. As expected, the three polynomials did not apply any valid correction. The airPLS was the second best in the terms of consistency of the three parameters but lacked in the line-shape accuracy and the general baseline re-referencing. Similarly, for the arPLS where the baseline recorded a high value, but the spectral patterns were compromised by their inefficacy to handle positive and negative signals in the same spectrum. The DBC and ALS methods preserved in a similar manner the crucial aromatic regions, but the lower baseline values suggested artefact or under-correcting the baseline. Lastly, the WS registered the highest value in the noisy region but lower values in the signal regions did not make this algorithm excel in the complex.

A different scenario was shown by analysing outputs from the ^{1}H spectrum. This spectrum was characterised by a very large degree of phasing error, especially in the solvent regions. The new baseline correction method didn't preserve fully the whole spectral patterns compared to 1st order polynomial and DBC, but the almost null baseline correction for the two methods could not make them a superior method compared to the CWBC; this was especially true for the polynomial fit which, as previously seen, only shifted the whole spectrum along the intensity axis. However, for the CWBC algorithm the critical aromatic field was at ~63%, similar to the DBC and for the airPLS, but the significantly higher baseline value, suggested an overall favourable level of correction, Fig. 4.15B.

A similar case to the $T_{1\rho}$ was observed for the analysis of WaterLogsy spectrum. Fig. 4.15C. This spectrum did not have noticeable post-processing errors and minimal baseline offsets with slight rolling distortions. Once more, the CWBC baseline method recorded the overall highest values for whole and aromatic patterns similarities, at 0.89 and 0.97, following by a perfectly flat baseline in the noise region as suggested by the 0.98 score in the baseline reports. Whereas, other methods such as the airPLS, arPLS and ALS, recorded low values in aromatic and whole regions, indicating a general alteration of the original signals. Notably, DBC performed relatively well on aromatic and whole spectrum, but under-performed on the noisy baseline level, Fig. 4.15C.

Lastly, the average of all three observations was analysed for the three spectra, Fig. 4.15D. Not surprisingly the CWBC algorithm achieved the highest level of correction, at 0.81 units, followed by the arPLS, WS DBC with values of 0.67, 0.63, 0.62 respectively.

Subsequently, the full dataset was analysed as described above and the normalised values for whole, aromatic and baseline distances were reported for each spectrum, Fig. 4.16A. As expected, the 1st order polynomial scored the highest coefficient of correlations when comparing the spectral patterns as a whole region, followed by the CWBC and the BDC, at ~0.92 and ~0.89 respectively. The CWBC scored the highest value on the aromatic region, followed by the BDC method and polynomial, whereas, in the baseline distance values, the WS registered the highest score, followed by the New and airPLS, that showed fractionally lower values.

The average of these three scoring methods were also inspected, Fig. 4.16B. In line with previous observations, the New baseline correction algorithm scored the top merit at 97%, followed by the arPLS and BDC at 74% and 68% reciprocally.

Scorings were finally grouped by experiment types in a heat map, Fig. 4.16C. From the plot it is possible to appreciate how the CWBC has evenly performed well in all spectra relative to the ^{1}H, $T_{1\rho}$ and WLOGSY, with scores constantly above 0.75 and only occasionally lower in the ^{1}H type. In contrast, other methods tested showed intermittent high scores among the dataset. For example, the arPLS performed better on $T_{1\rho}$ compared to the BDC, but the latter was superior on WaterLOGSY spectra. The lowest score for the new algorithm was recorded for the spectrum ^{1}H 467 at 0.45. However, a qualitative inspection, Fig. 4.17A, showed that the crucial aromatic and aliphatic signals were perfectly corrected, Figs 4.17B-C, and only fewer peaks around the solvent region were lost due to the high phasing errors, Fig. 4.17D. This spectrum in fact was the most distorted across the dataset.

Lastly, the computing time of the tested algorithms were evaluated, Fig. 4.16D. Excluding the polynomials, which obviously were computationally less demanding; the CWBC algorithm was significantly faster than the top two methods, BDC and arPLS, by over 5 and 20 orders of magnitude and only a fraction of a second slower than the WS, with a total executing time of 0.60s for correcting 36 spectra. This value might sound insignificant for this dataset, but it only consisted of a portion of the original dataset recorded by the collaborators, whereas, typical screening dataset can be up to 10 thousand spectral entities.

4.5. Conclusions

NMR screening data analysis consists in the detection and interpretation of signals recorded for samples composed by small molecules and macromolecular target(s)[86]. Low affinity binders can result in difficult to assess spectral differences which can lead to false positive and negative observations. It is, therefore, necessary that control, and experimental spectral settings are cautiously optimised for screening trials before proceeding to automated acquisitions. This is in practice not always possible, and datasets might present a series of spectral errors that can make the entire dataset completely unusable and unreliable for extracting crucial screening information. One of the most common processing artefacts that prevents a trustworthy data analysis is an improper baseline distribution and the presence of any degree of offsets among spectra. Over the year several baseline correction strategies were described both for NMR and other spectroscopic techniques[128–130]. In this chapter I tested the most recent and promising algorithms found in the public domain.

I have assessed outputs from a variegate collection of experimental spectra after applying a series of algorithms such as the Whittaker Smooth, ALS[132], ArPLS[134] and AirPLS[133] and BDC[135]. Furthermore, simple polynomial fittings were used as a control test. Each algorithm showed advantages and disadvantages, some of which performed particularly well in a determined experiment type but not in another for the same dataset, but most importantly, all relied on a fine and bespoke adjustment of internal parameters depending on the spectral input. These adjustments and selection of a correct algorithm for large screening dataset are simply prohibitive and might create unnecessary delays and reproducibly issues.

Most of the algorithms tested produced serious artefacts that compromised the integrity of the dataset depending on the nature of the experiment type.
To solve the various issues encountered during the correction of the dataset, I implemented a new method capable of correcting complex baseline artefacts. The approach uses a double cross-correlation approach to estimate the baseline and a weighted model for estimating the signal regions. I therefore propose the term correlated weighted baseline correction (CWBC).

The newly developed method was capable of correcting highly distorted spectra and preserving signal patterns without the need of any user-adjustable parameter. Furthermore, it required less computational timing compared to others, making it an ideal solution for large and multivariate dataset such as the ones required in screening by NMR. Previous published algorithms included mostly qualitative assessment of any previously developed method, whereas, for this new approach I quantitatively classified the performance, including a comparison among different experiment types. However, extra testing will be necessary for determining threshold limits up to the algorithm can successfully correct compromised spectra. Lastly, future plan will include extending its employability to higher dimensionalities experiment types and larger datasets, so that can be used as a robust and versatile tool for NMR post-processing routines.

4.6. Figures

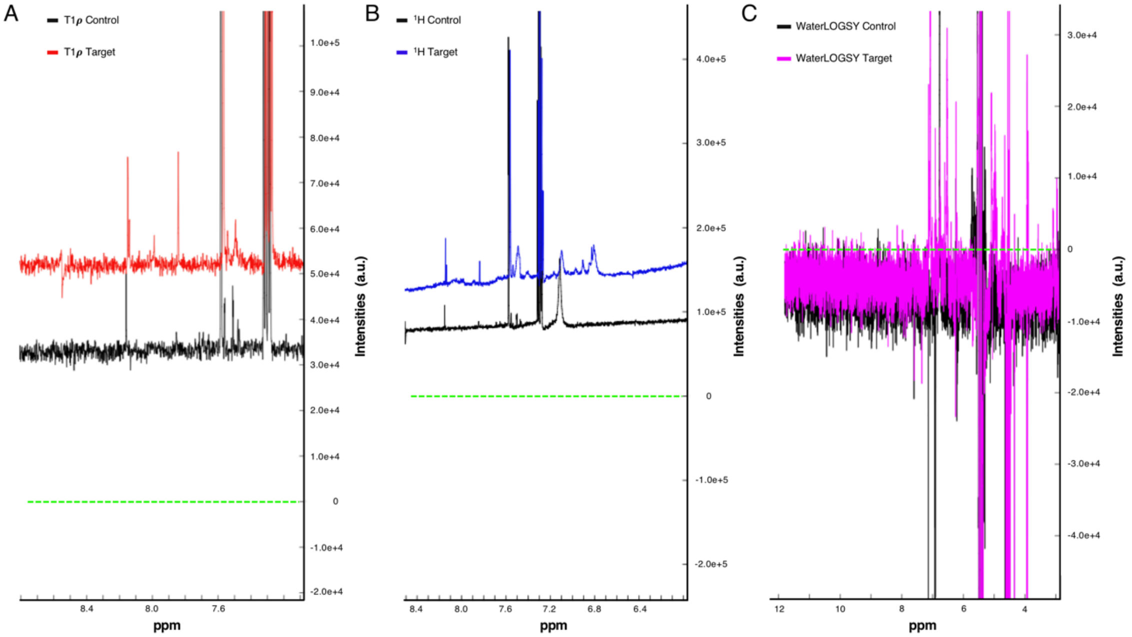


Figure 4.1 Experiment types spectra for the component-314.

A A portion of the aromatic field of spectra used for testing the various baseline correction algorithms. In black the control $T_{1\rho}$ spectrum and in red the $T_{1\rho}$ spectrum recorded in presence of the macromolecular target. **B** ^{1}H spectra recorded as the control in black and in presence of the target in blue. **C** WaterLOGSY spectra with the control displayed in black and target in magenta. For each experiment type the intensity comparison was impossible to assess due to the large offset between the control and experimental observations. The dotted green line represents the expected 0 baseline line.

Figure 4.2 CWBC algorithm workflow overview.

Schematic representation of the new proposed algorithm, CWBC; each rectangle consists of a crucial operation required to correct the spectral baseline.

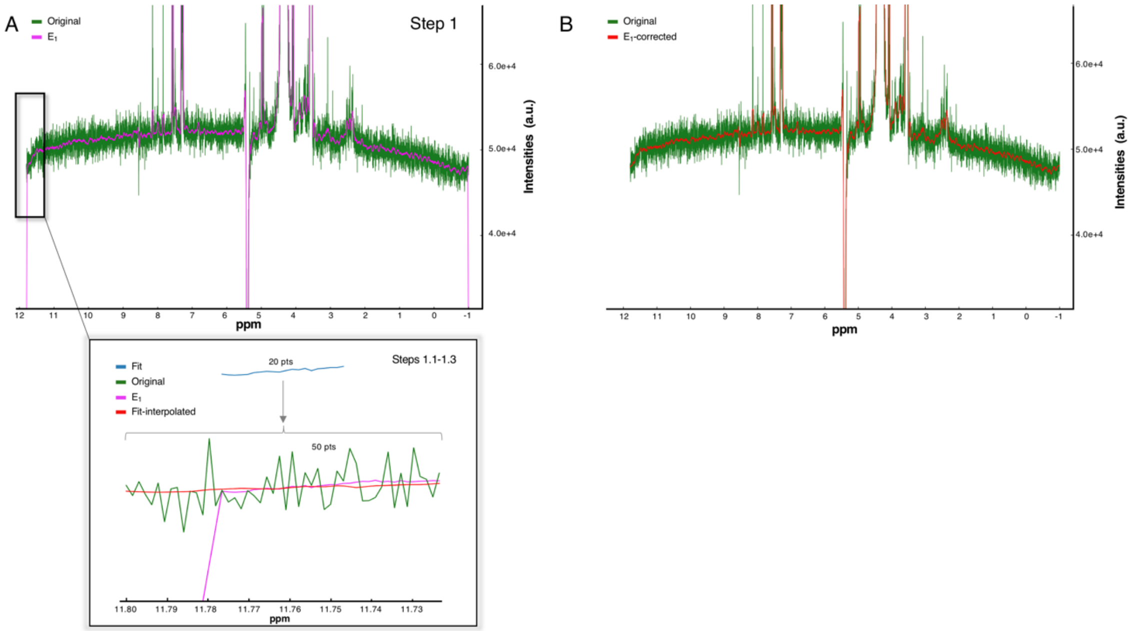


Figure 4.3 CWBC algorithm: step 1.

A Step 1. Overlap of the original spectrum, in green, with the cross-correlated curve fit E_1, in magenta. **A-Insert** Steps 1.1-1.3. "Smile" correction for the downfield region of the spectrum. In magenta, the artefact created in the E_1 cross-correlation using the engine mode "same". In blue the new cross-correlation for only a small portion of the spectrum using engine mode "valid", which produced a smaller curve of 20 points. In red, the interpolated curve for the fit consisting of the same length of the original region of the spectrum in examination, which is shown in green. **B** Overlap of the original spectrum, in green, with the corrected cross-correlated curve fit E_1 in red.

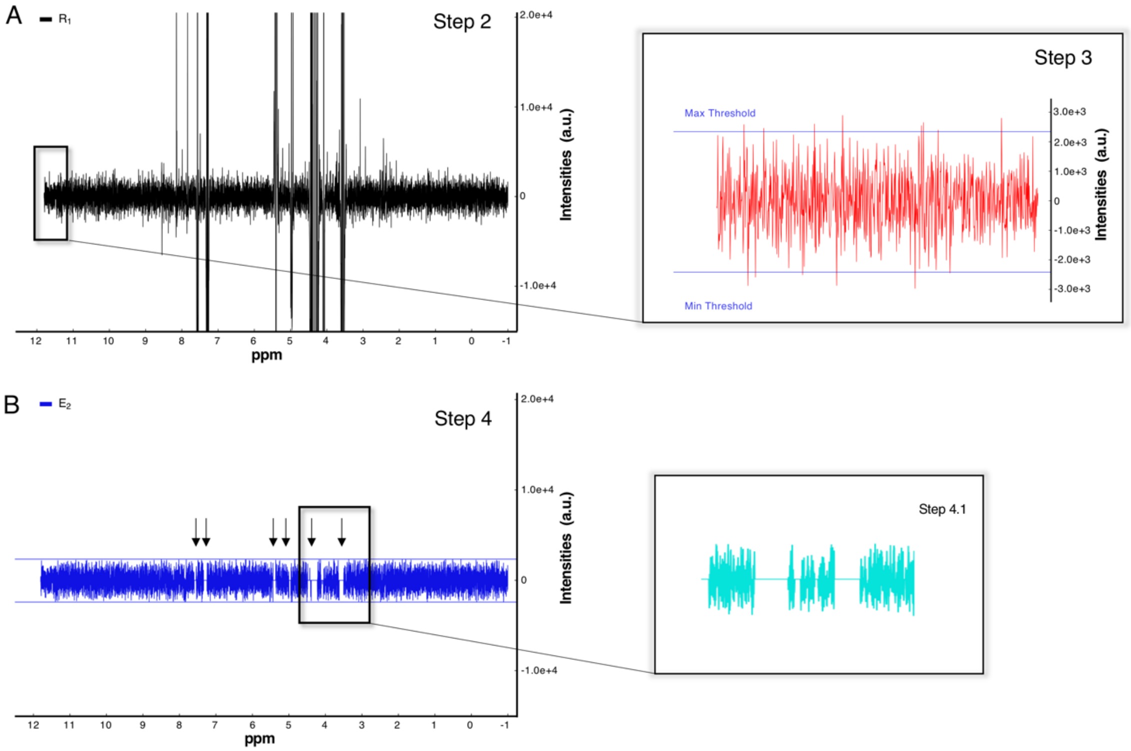


Figure 4.4 CWBC algorithm: steps 2-4.

A Step 2. R_1 spectrum created by the subtraction of the corrected E_1 from the original datapoints. **B** Step 3. Identification of noise threshold values on R_1. **C** Step 4. Removal of signals above and below thresholds on R_1; intensities in these regions are set to a value of 0, black arrows. **C-insert** Step 4.1. Amendment of "zero groups", removal inliers and increment of outliers.

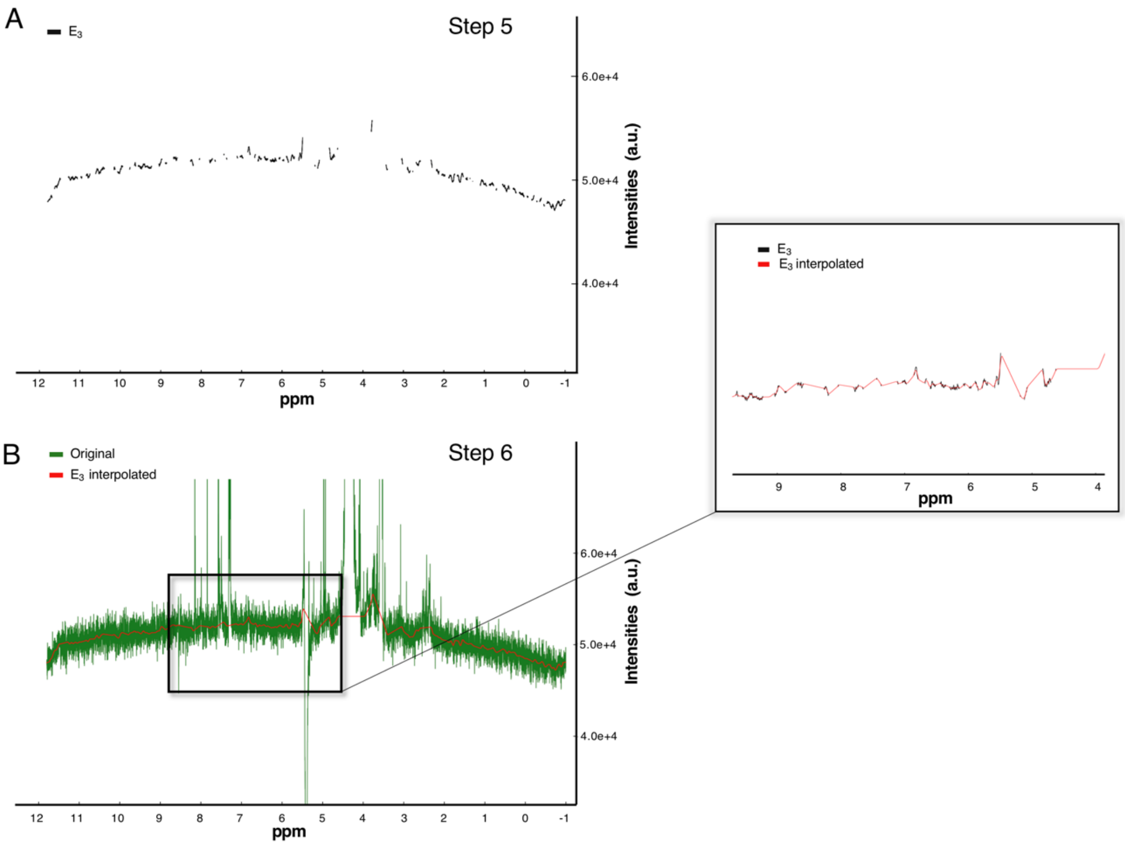


Figure 4.5 CWBC algorithm: steps 5-6.

A Step 5, Transferral of zero points from R_1 to E1. The resultant spectrum, E_3, displays a fragmented pattern due to the masked signal. **B** Step6, interpolation of masked signal. Overlay of the original spectrum, green, and $E_{3interpolated}$ spectrum, red. **B-insert**, overlay of E_3 (black), and the resulting spectrum after the linear interpolation of masked signals (red).

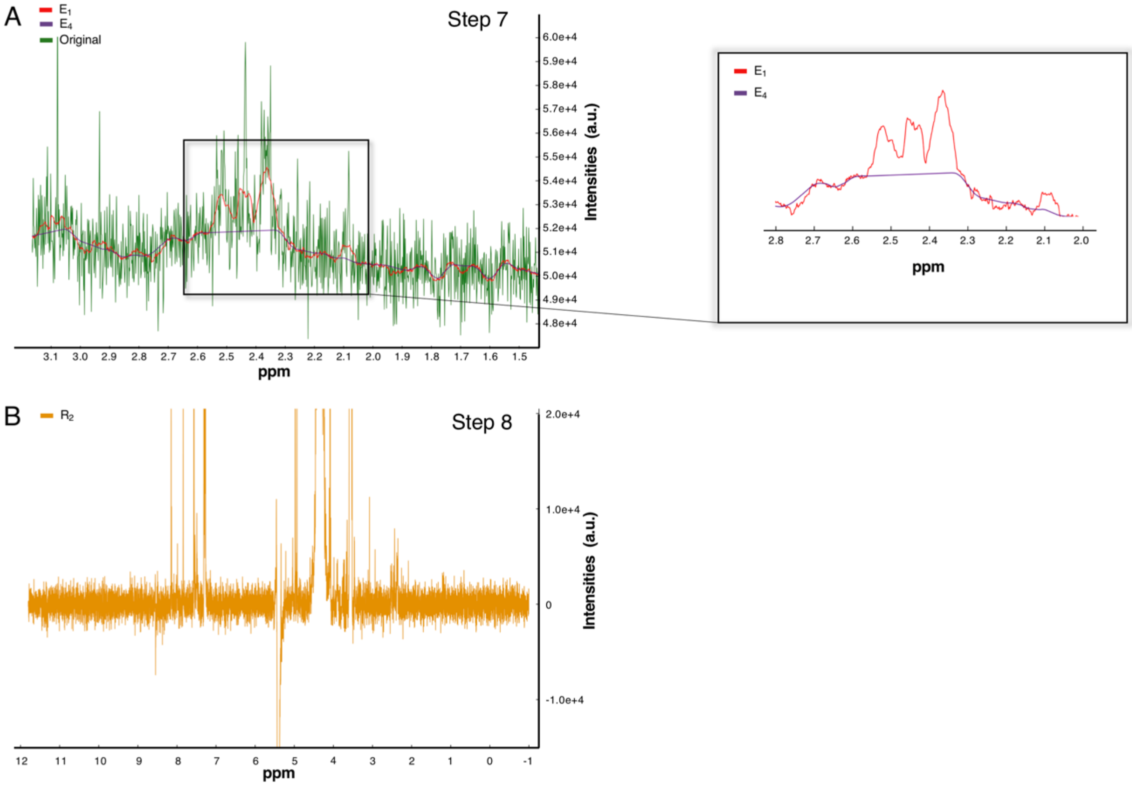


Figure 4.6 CWBC algorithm: steps 7-8.

A Overlay of the first, E_1 and final cross-correlated curve E_4 (red and dark purple) together with the original spectrum (green). **A-insert**, a zoomed region of the spectrum denoting the fitting differences between the first, E_1 and final cross-correlated curve E_4 (red and dark purple). **B** Overview of the final R_2 spectrum, created by the subtraction of the original and the final baseline estimation.

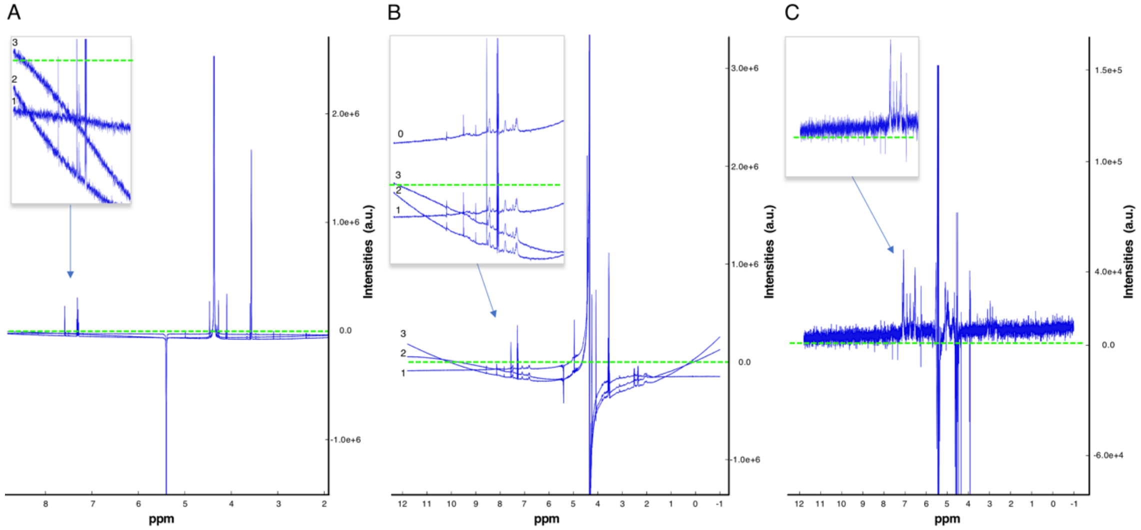


Figure 4.7 Polynomial orders corrections for Pr-314.

A The resulting $T_{1\rho}$ spectra after applying the polynomial routines. Every polynomial order over-corrected the baseline by offsetting it below the 0-threshold line (dotted green line). **B** The resulting ^{1}H spectra after applying the first three polynomial orders to the ^{1}H spectra. As previously observed, the polynomial over-corrected or introduced large baseline rolling artefacts. In the top left box, 0 corresponds to the original spectrum, 1, 2, 3 refer to the 1st, 2nd and 3rd polynomial order outcomes. **C** WaterLOGSY spectrum after the application of the 1st order polynomial fit. The spectrum was slightly corrected in the first data-points in the downfield region, but under corrected in the upper region, resulting in a compromised spectrum. 2nd and 3rd orders not displayed for clarity, results were in line with A and B, showing large rolling errors across the spectrum.

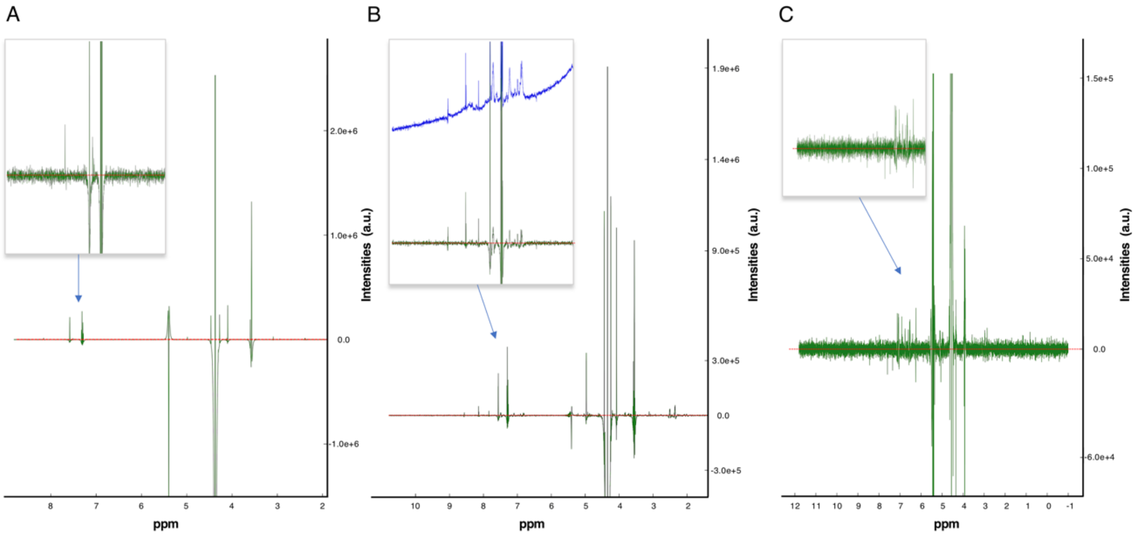


Figure 4.8 Whittaker smoother corrections for Pr-314.

A The $T_{1\rho}$ spectrum after applying the Whittaker Smoother (WS) filter to the original data-points. Although a perfect baseline was achieved at the zero intensities axis (red dotted line), all spectral peak-shapes were compromised. **B** The resulting ^{1}H spectrum after the application of the WS method. Top left panel shows the aromatic region for the original spectrum, blue, and the WS corrected, (green). The over-correction produced negative intensity peak which were artefacts. **C** WaterLOGSY spectrum after the application of the WS method. It showed negative artefacts and decreased intensity values as in the previous experiment types.

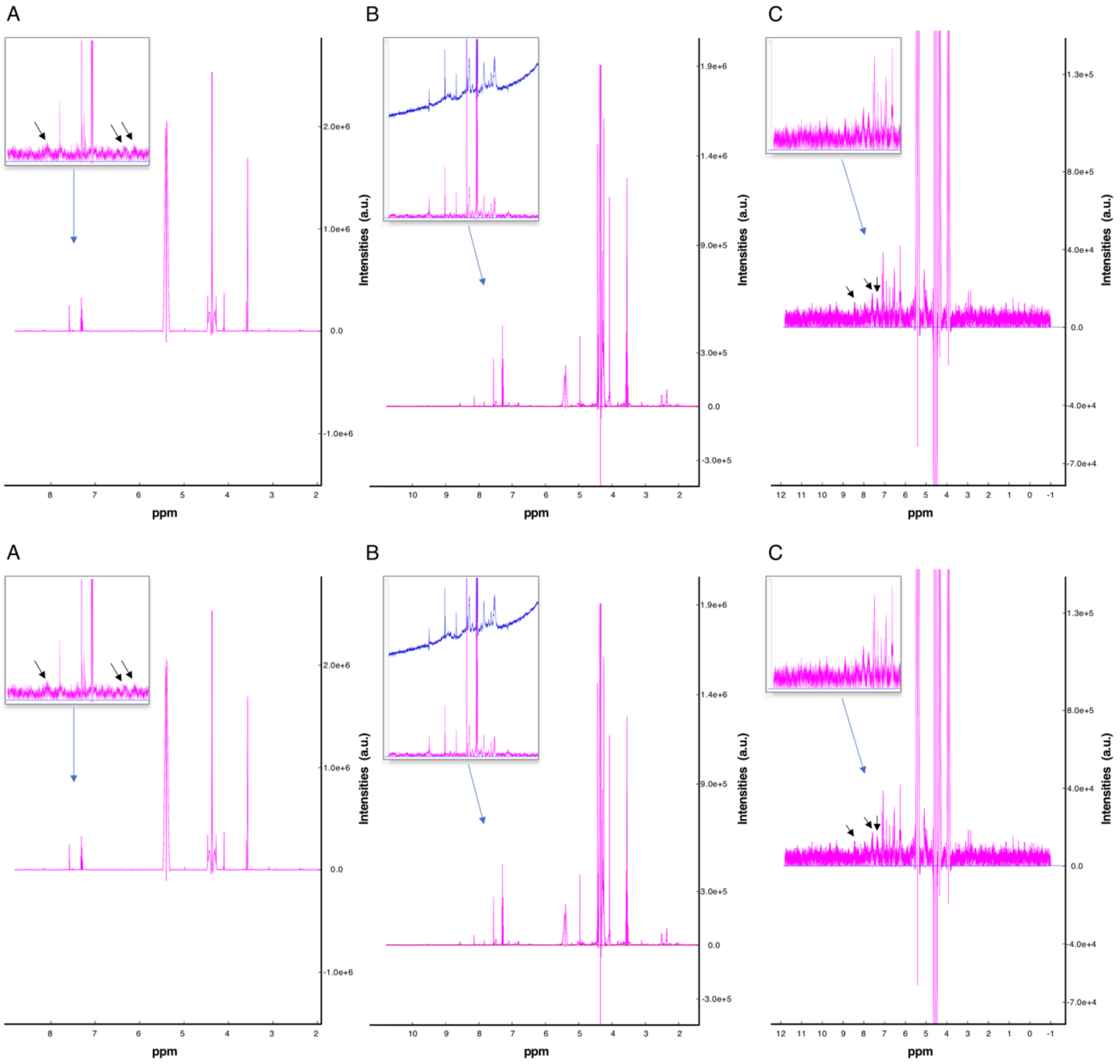


Figure 4.9 ALS corrections for Pr-314.

A $T_{1\rho}$ spectrum after applying the Asymmetric Least Squares Smoothing, ALS, algorithm. The spectrum showed low and broad artefact signals across the baseline, black arrow; furthermore, negative peaks indicated an overcorrection resulting in distortions. **B** ^1H spectrum outcome showed similar patterns to the $T_{1\rho}$ experiment. **C** The WaterLOGSY spectrum after the applied algorithm. It showed severe positive artefacts and negative overcorrection typical of the algorithm.

Baselines appeared under-corrected and not normally distributed along the zero intensities axis (blue dotted line) for all three experiment types.

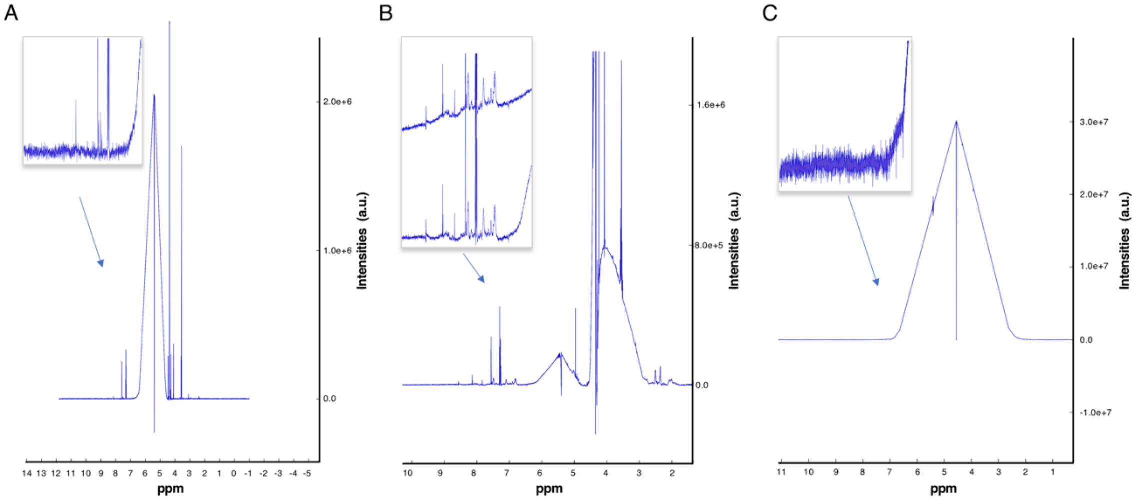


Figure 4.10 airPLS corrections for Pr-314.

A $T_{1\rho}$ spectrum after applying the adaptive iteratively reweighted Penalized Least Squares, airPLS. The algorithm corrected relative well the positive signals but badly compromised the rest of the spectrum by over-fitting the negative solvent region. **B** ^{1}H spectrum after the application of the airPLS method. It showed global spectral distortions due to solvent (negative) regions. **C** The WaterLOGSY spectrum after the applied algorithm. This spectrum too, showed a pattern warp around the water region at 5 ppm.

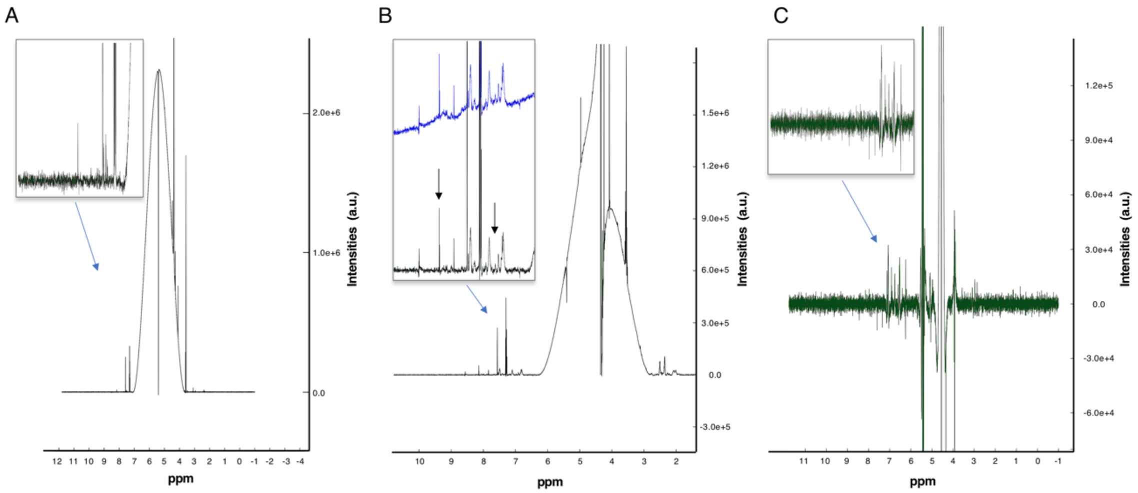


Figure 4.11 arPLS corrections for Pr-314.

A $T_{1\rho}$ spectrum after applying the asymmetrically reweighted Penalized Least Squares Smoothing, arPLS[134]. The algorithm eliminated the baseline error in the aromatic region, but performed poorly in the water region, compromising a large portion of the spectrum between 6 to 4 ppm. **B** The ^{1}H spectrum manifested severe issues in the solvent fields. On the top left square, it is possible to notice the result of an over-correction for some positive aromatic peaks (black arrows); intensities appeared wrongly decreased compared to the original blue spectrum, arising potential false positive screening hits. **C** The WaterLOGSY spectrum after the applied algorithm. The spectrum showed compromised negative peaks both in the aromatic and in the solvent regions.

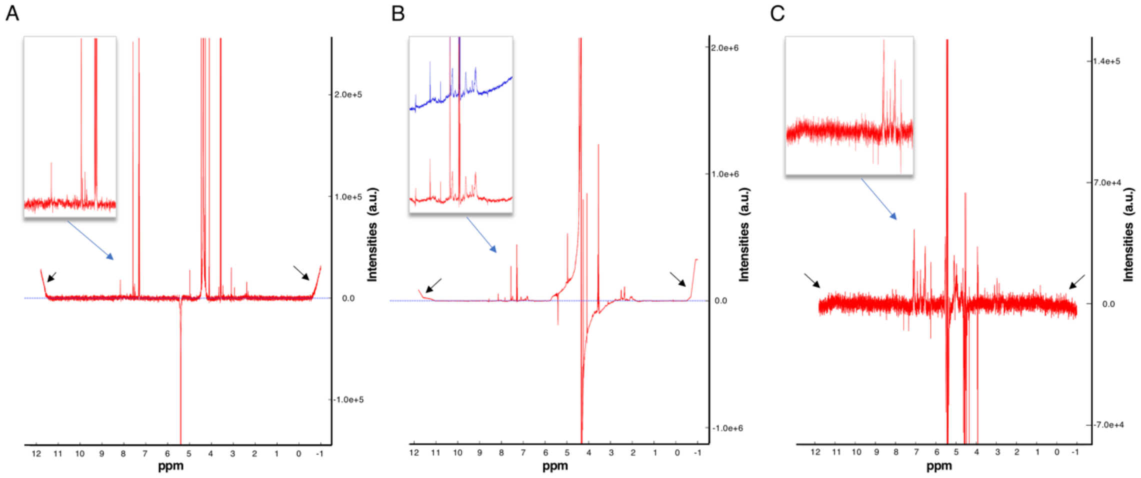


Figure 4.12 BDC corrections for Pr-314.

A $T_{1\rho}$ spectrum after applying the Distribution-Based Classification, DBS, baseline correction algorithm. The spectrum appeared correctly re-referenced to the zero intensities threshold line in most of the data-points. A large "smile" artefact was visible at both edges of the spectrum (black arrows) **B** The ^{1}H spectrum showed the same "smile" errors (black arrows) and mild baseline rolling in the signal regions. **C** The WaterLOGSY spectrum after the applied algorithm. It was corrected relatively well except for the constant artefacts at both ends.

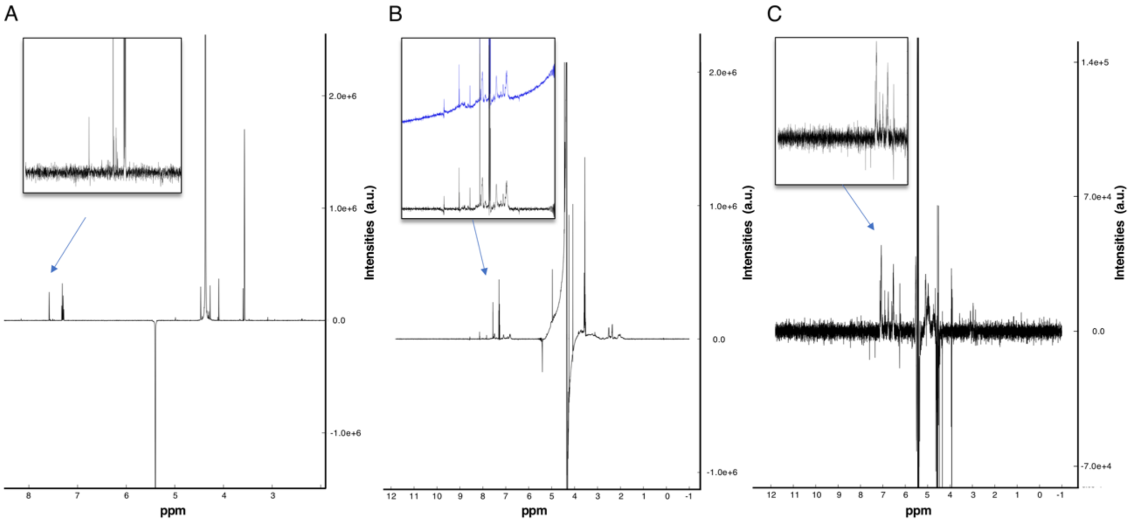


Figure 4.13 New baseline corrections (CWBC) for Pr-314.

A $T_{1\rho}$ spectrum after applying the CWBC algorithm. Spectral patterns were correctly maintained across data-points. **B** The resulting ^{1}H spectrum after running the new baseline correction method. The spectrum showed an accurate correction in aromatic and aliphatic regions; some difficulties were observed in the unphased water region at ~5 ppm. **C** The WaterLOGSY spectrum after applying the newly developed algorithm. It was meticulously corrected and did not show any induced artefacts.

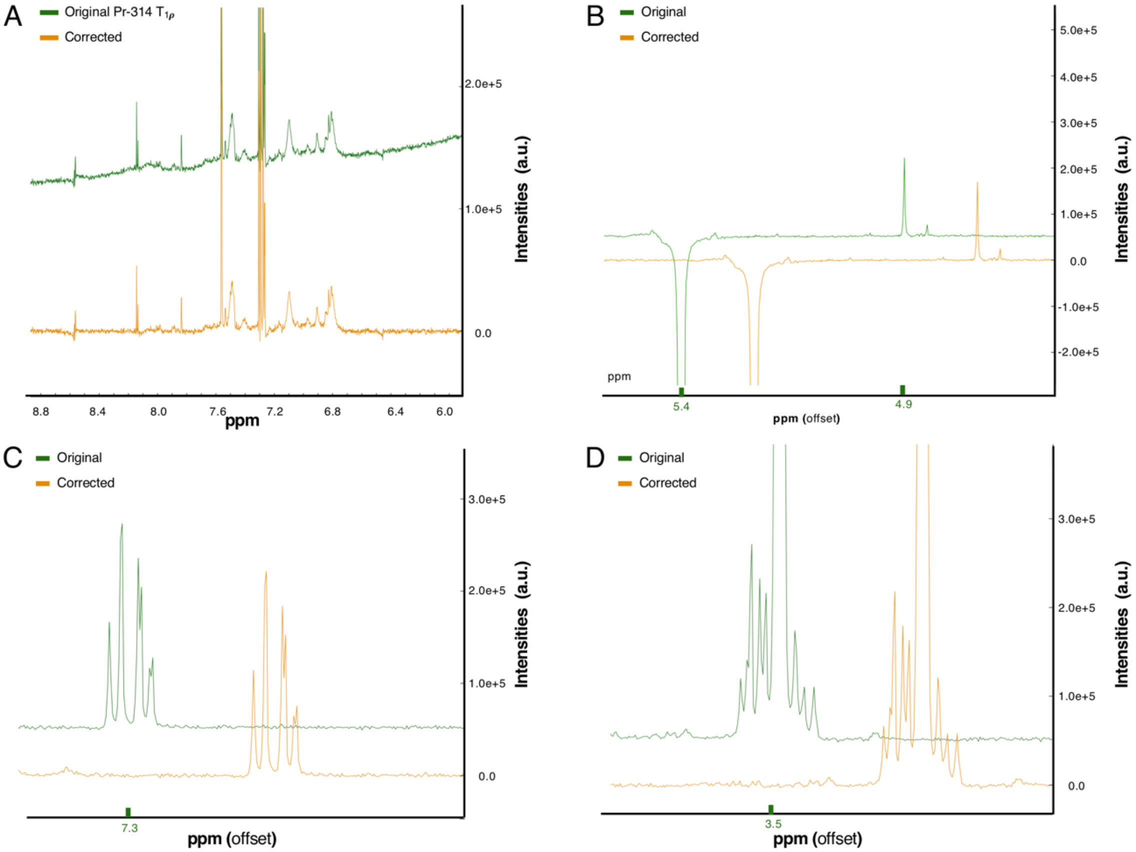


Figure 4.14 Qualitative analysis for Pr-314 T$_{1\rho}$ after applying the CWBC.

A Spectral overlay for the original spectrum, in green, and the corrected spectrum, in orange. The figure shows a high accuracy in preserving peak-shapes in the aromatic region of the spectrum. **B** A zoomed portion of the spectra where negative and positive peaks are in close proximity. The corrected spectrum is shifted by ~0.15 ppm in the x-axis for clarity. **C** An example of a multiplet in the aromatic field. The corrected spectrum is offset by ~0.15 ppm in the x-axis. **D** An example of a complex multiplet pattern in the aliphatic region. The corrected spectrum is offset by ~0.15 ppm in the x-axis.

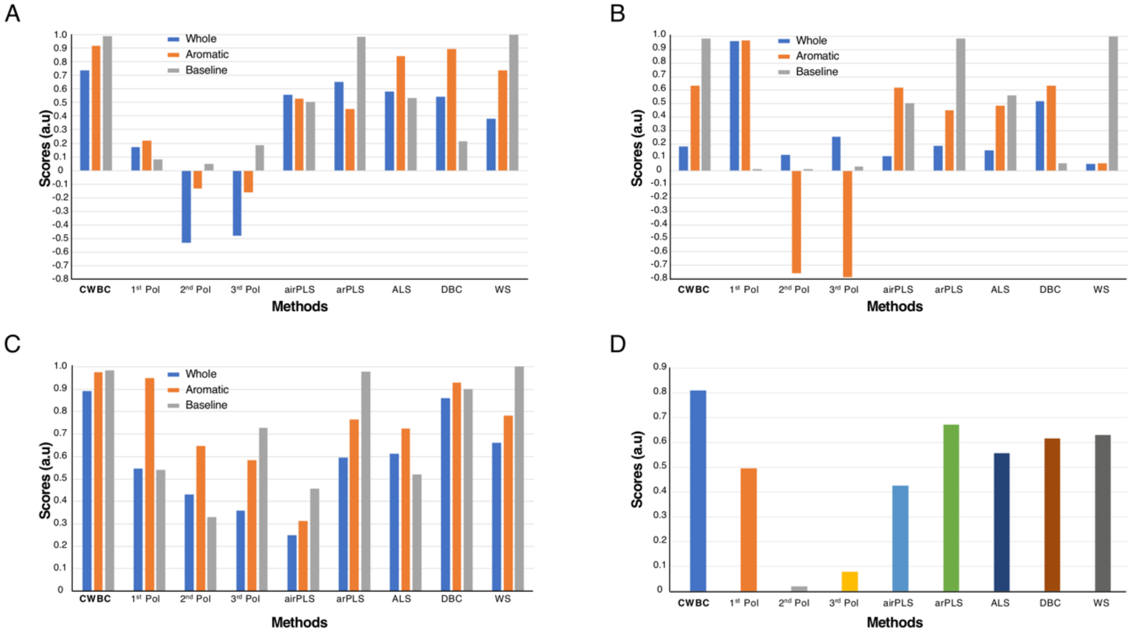


Figure 4.15 Analytic comparison of methods applied to Pr-314 spectra. Scoring in arbitrary units for whole (blue) and aromatic only lineshapes (orange), and baseline distance (grey), calculated as in Eq. 4.2. Values close to 1 for the whole and aromatic signify high similarities between original and corrected. Values close to 1 for the baseline bar represent closer distance between the corrected baseline and simulated spectrum, calculated as in Eq. 4.4. **A** Results for the $T_{1\rho}$ experiment type spectrum. **B** Results for the ^{1}H spectrum. **C** Results for the WaterLogsy spectrum. **D** Average scoring calculated as in Eq. 4.5 for each method tested.

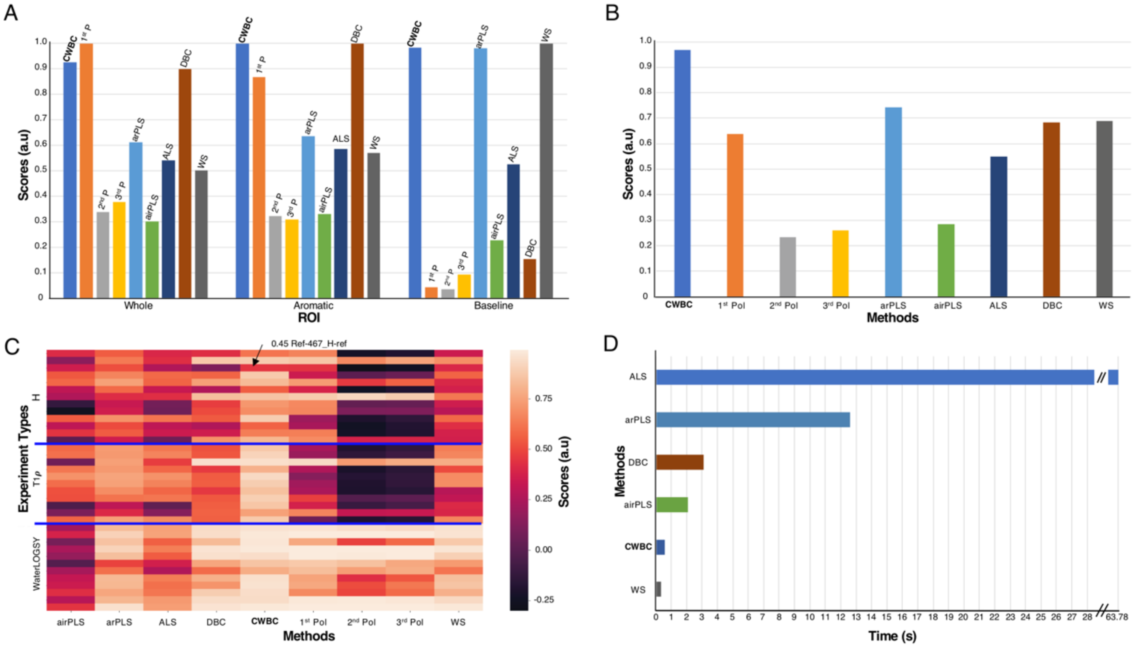


Figure 4.16 Analytic comparison of methods applied to the whole dataset. **A** Average scoring for the whole, aromatic and baseline applied to all 36 spectra in the dataset. Each colour represents the different method tested. **B** Average scoring calculated as in Eq. 4.5 for the total spectra. **C** Heat map plot for shows single average scores (whole, aromatic, baseline) for each spectrum. Spectra are sorted by experiment types; starting from the top: ^{1}H, followed by $T_{1\rho}$ and finally WaterLOGSY. Higher scores are represented by lighter colours. **D** Executing timing for each algorithm tested for this dataset.

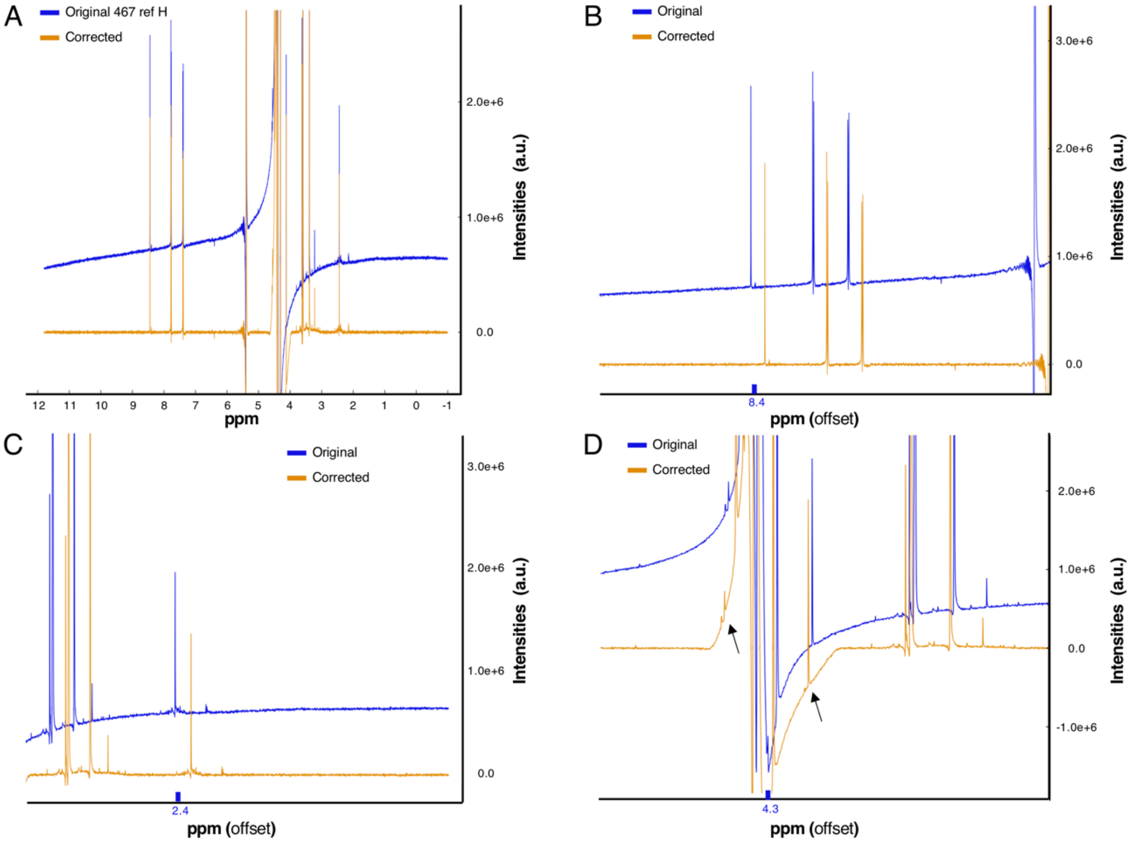


Figure 4.17 Qualitative analysis for 467-Ref ^{1}H.

A Spectral overlay for the original spectrum, in blue, and the corrected spectrum, in orange. The original spectrum was characterised by a large intensity offset and a severe phasing issue. **B** The figure shows a high accuracy in preserving the peak-shapes in the aromatic region of the spectrum. The corrected spectrum is offset by ~0.15 ppm in the x-axis. **C** Zoom-in in the aliphatic region of 467; the image suggests a high level of accuracy in preserving the peak-shapes. The corrected spectrum is offset by ~0.15 ppm in the x-axis. **D** Zoom-in in the solvent region around 4.3 ppm. The corrected spectrum is offset by ~0.02 ppm in the x-axis.

4.7. Acknowledgments

I thank Drs Ken Kameron[a] and Andrea Gohlke[a] for providing the dataset required for these analyses.

4.8. Addresses

[a] Beatson Institute for Cancer Research, Garscube Estate, Switchback Road, Bearsden, Scotland G61 1BD, United Kingdom.

Chapter 5

Practical application of AnalysisScreen

See List of Publications: 1.

Keywords: WaterLOGSY, STD, CPMG, CcpNmr AnalysisScreen.

5.1. Abstract

In this chapter I discuss results obtained by analysing four experimental datasets using pipelines and tools available in CcpNmr AnalysisScreen (cf. chapters 2, 3). Our collaborators, both industrial and academic, who provided us the essential datasets for validating AnalysisScreen, have previously relied largely on a visual inspection of 1-dimensional spectra for assessing binding events of fragments to their targets. This strategy did highlight the major spectral differences between a control and a spectrum acquired in presence of a target and therefore classified molecules as binding hits; however, it lacked any quantitative evaluation.

More generally, the visual inspection of spectra has several drawbacks, such as a difficulty in reproducing experimental results and inaccuracies in matching complex spectral signals across different datasets. In this chapter, I discuss pipelines and scorings obtained by performing automated and semi-automated analysis routines on a series of different datasets, including multiple experiment types, such as ^{1}H Relaxation-edited, WaterLOGSY, STD, CPMG. I examine results from the automated routines by comparing these to the results obtained through visual inspection.

5.2. Introduction

Apart from AnalysisScreen (cf. chapters 2, 3), currently only a few commercial software packages provide dedicated support for NMR screening[116,140]. In spite of the large premium for software like Bruker TopSpin[116] and MestreLab MNova Screen[140,141], which often cannot be justified for occasional users and students, these packages provide little or difficult customisation of workflows. For instance, TopSpin tools only provide for a qualitative analysis of hits using binary scores, such as ”binding” or “not binding” hits, whereas MestreLab reports an overall intensity percentage change above a certain user threshold. The aim on CcpNmr AnalysisScreen is to provide a full support, not only for an easy and quick qualitative inspection of NMR data but also to provide a reproducible and detailed scoring of each dataset. In line with computational docking, where several scoring functions and algorithms developed during the years allowed to score and filter results[142–144], screening by NMR should be assessed quantitatively in the same manner. In fact, the various compounds to be tested can be expected to bind the target with different orders of magnitudes and it is essential to differentiate a weaker binder from a stronger binder using the primary screening[145,146]. This approach will guarantee a reduction in human mis-interpretation of data, facilitate the deconvolution of a large quantity of information and most importantly it will aid in the reproducibility of results.

CcpNmr AnalysisScreen architecture and features have been described in chapters 2 and 3. In this chapter I report on four case studies which were analysed using the new package. Each case study contained experimental datasets that were recorded by industrial and academic collaborators as part of a genuine FBDD effort. The biological information of these studies was not revealed, as it was not required in the development and validation of the software at this stage. Furthermore, some of the spectra and associated metadata was proprietary and not shared with us. The datasets were mostly comprised of STD experiments, but also ^{1}H-Relaxation-edited, WaterLOGSY, $T_{1\rho}$ and CPMG at multiple relaxation times. Several thousand spectra were analysed, with the two largest libraries amounting to 1548 and 1632 reference components each. For each dataset, scores were given according to the experiment type under examination, e.g. intensity changes, shifts distribution, chemical shifts matching, signal-to-noise ratios versus efficiencies and in the case of CPMG analysis, detailed PDF files with extended analysis properties.

Here, I will describe tools I used to tackle the peculiarities of each dataset, i.e. moderate-to-severe post-processing issues, phasing, scaling and referencing issues. The latter is a common problem present in NMR due to variations in experimental conditions when acquiring in multiple stage screening samples and their reference compound spectra, e.g. at different spectrometers, temperature, solvent compositions, etc. The pipeline implemented in CcpNmr AnalysisScreen therefore included re-referencing and global alignment pipes, that are capable of automatically detecting and applying shifts to each individual spectrum or, alternatively, setting a specific parameter simultaneously for all spectra.

Although these projects were only used for a software validation purpose, the diversity of each case was fundamental for gaining invaluable information for developing tools needed for assessing multiple screening research approaches.

All datasets were initially assessed only visually and qualitatively. Subsequently, by comparing these results with those obtained from automated or semi-automated procedures, several discrepancies were observed, including the appearance of potential hits that were completely misjudged during the visual, non-automatic analysis.

5.3. Materials and Methods

5.3.1. Materials

AnalysisScreen capabilities were assessed using four experimental datasets that were kindly supplied by four different laboratories. Three datasets consisted of one-dimensional spectra recorded with Bruker spectrometers, whereas one dataset was provided in a CSV file format.

Dataset-1. The first dataset was composed of 27 spectra relative to three compounds, identified as 314, 467, 7373. The spectra recorded for these compounds displayed (according to the data-owner) the strongest binding properties, therefore, they should function as a good validation model.

For each substance there was a reference spectrum plus three ^{1}H, six $T_{1\rho}$, six WaterLOGSY and three STD spectra, acquired in the presence and the absence of a protein target. Target and substance structures were undisclosed. Spectra files were provided as CSV files and they had been processed by the collaborators themselves.

Each component was inspected independently (singleton), and only the aromatic region of the spectrum was considered for detecting hits.

Dataset-2. The second dataset was composed of two subsets of spectra. The first subset included spectra relative to 35 undisclosed small compounds. For each compound a reference ^{1}H spectrum was provided alongside with the control STD and the relative STD acquired in presence of the target. The second subset included eleven samples as mixtures of four compounds each. It was not clear if the mixtures included the same compounds previously recorded as singletons or a different library was used. All spectra had already been processed and biological properties were unknown.

Dataset-3. Part of this dataset was already introduced in chapter 2; it consisted of 1548 fragments used for creating 310 mixtures of approximately five components each. For each fragment a processed ^{1}H reference spectrum was provided in addition to a total of 309 processed STDs recorded for each mixture in the presence of the target. Ligands were prepared at a concentration of ~200 µM and the target at ~4 µM.

Dataset-4. The dataset consisted of 1632 reference compounds divided in 168 mixtures, with each mixture containing approximately 10 compounds. Data from three different NMR screening experiments were available: ^{1}H, On and Off resonance, and a series of seven CPMGs at 0, 45, 50, 100, 300, 500, 800 ms decay time.

Ligands were used at a concentration of approximately 300 µM, and the target at 5-10 µM.

Each sample was recorded as a control (compounds only), referred as SF, and for the ^{1}H and CPMG experiments, recorded at 0, 45, 50, 100, 300, 500 and 800 ms relaxation time, in the presence of the target, referred to as SP. The total number of spectra for this dataset was 4692, plus 168 STD spectra which were generated automatically as part of the analysis.

5.3.2. **Methods**

The analyses have been carried out using the built-in pipelines and bespoke macros employing packages included in the AnalysisScreen Python environment.

For detecting spectral signals, I have used the two peak picking algorithms described in chapter 3; the first implementation was based on the *maximum_filter* algorithm[120]*,* referred as MF, and the second method was implemented from the *Peak Detector algorithm*[121], referred as PD throughout this chapter.

Dataset-1 was analysed semi-automatically; peaks were picked manually in all spectra and intensities were compared between control and spectra recorded in the presence of a target using a macro.

Percentage changes in peak intensity heights were calculated as:

$$Hc = \frac{|H_1 - H_2|}{(H_1 + H_2)/2}$$

Eq. 5-1

Where H_1 and H_2 are the two peak heights being evaluated. Changes in peak heights were used to score the signals in the $T_{1\rho}$ and WaterLOGSY spectra.

Dataset-2 was analysed using a combination of manual and automated pipes.
The parameters used for the MF in the first subgroup (singletons) were: *size* = 10; *mode* = wrap; *noise level factor* = 2.
The PD had one adjustable parameter only, delta, which was set to: *delta* = 1.6; whereas, the ignored regions were: 4.8 to 4.6 ppm and 2.65 to 2.55 ppm. For both methods, negative peaks were excluded.
Peaks were identified using the MF using the following parameters:
Reference Spectra: *noise level* = 500.000, *filter size* = 10
The ignored regions were: 4.9 to 4.65 ppm, 3.49 to 3.45 ppm and 2.7 to 2.650 ppm. For the STD Spectra: *noise level factor* = 2, *filter size* = 10.
To the same spectra was applied the PD with a *delta* value of 1.6. Whereas, the ignored regions were: 4.8 to 4.6 ppm, 2.78 to 2.50 ppm, 1.20 to 1.00 ppm. The STD hit detection pipe, used to match references to the STD spectra, had a tolerance of 0.03 ppm.

Dataset-3 was analysed using a pipeline comprised of: exclude regions, noise threshold, peak picker, and SpectrumHit detection with matching references.

Automated peak detection on all spectra was achieved using the PD peak picker with a delta value of 1.5. For this dataset the aliphatic region of the spectrum, i.e. 5 to 0 ppm, was excluded. Reference peaks were matched to the STD spectral peaks using a tolerance of 0.03 ppm for the first run and 0.01 for the final run.

Dataset-4 was analysed using a combination of macros and GUI pipes. The ^{1}H analysis macro consisted of the following steps:

1. Peak detection on reference spectra, (*delta*: 1.5; *excluded regions*: 5.1 to -4 ppm with automatic noise threshold calculation);
2. Re-reference references to experimental. The manually calculated shift was 0.079 ppm;
3. Peak detection on control spectra (parameters as 1);
4. Copy control peaks to spectra with the target, “refit” maxima within a tolerance of 0.0035 ppm on both peak edges;
5. Calculate percentage changes as in Eq. 5.1;
6. Match references to target spectra within a tolerance of 0.05 ppm;

The STD macro was similar to that used for the analysis of the ^{1}H data, except that the control and target spectra were replaced by the On-Off resonance spectra and efficiencies were calculated as in Eq. 3.6. Furthermore, STD spectra were created from the On-Off resonances and the signal-to-noise ratio was calculated accordingly.

The CPMG macro consisted of the following steps:

1. Peak detection on reference spectra, (*delta*: 1.5; *excluded regions:* 5.1 to -4 ppm with automatic noise threshold calculation);
2. Re-reference references to experimental. The applied shift was 0.079 ppm;
3. Peak detection on control spectra at 0 ms (parameters as 1);

4. Propagate peaks to all spectra in the control and target series, “refit”; maxima within a tolerance of 0.0035 ppm;
5. Calculate exponential fit for each peak in the series;
6. Calculate changes in merit for the control and target fitted slopes by:

$$S = 1 - \frac{|T|}{|C|}$$

Eq. 5-2

Where T is the target and C the control parameter obtained from the fitting routine;

7. Match references to target spectra within a tolerance of 0.05 ppm.

Excel sheets were used to load and parse the metadata for all datasets (cf. chapter 2). For the last dataset, however, Excel files with all relevant data were automatically generated from the experimental data structure using in-house written python scripts. These allowed to divide and organise the large dataset in three different files for an easier handling and include extra needed metadata.

5.4. Results and Discussion

5.4.1. **Dataset-1**

The first dataset tested was relatively small as it contained only spectra corresponding to compounds that the collaborators considered to be binding to the target, subsequently referred to as SpectrumHits in the AnalysisScreen terminology. The availability of data for three different NMR screening experiments, i.e. $T_{1\rho}$, WaterLogsy and STD, renders this set-in principle suitable to validate the corresponding automated detection pipelines. Moreover, the usage of multiple NMR screening experiments is deemed to improve on the reliability of the screening procedure and the concept of level-1, level-2 and level-3 in the hit analysis procedure has previously been proposed[147]. As explained in chapter 3, based on the count of positive observations for each experiment, a molecule is defined as a level-1 hit if it appears as a binder in one experiment, a level-2 hit when two experiments confirm its interaction and so on.

Upon a first visual inspection, I noticed that all spectra showed severe phasing issues, referencing mismatches, baseline errors, scaling and large peak shifts, likely due to referencing errors (data not shown). These problems made the use of a fully automated pipeline for classifying the SpectrumHits on the basis of the original data impossible. Consequently, using the relevant pipes in AnalysisScreen the spectra were phased, baseline corrected, and subsequently peaks were picked and carefully inspected for each experiment. The resulting spectra for three compounds, i.e. 314, 467 and 7373 are displayed in Figs 5.1-3. The first component, 314, showed a notable intensity reduction in the $T_{1\rho}$ spectrum recorded in the presence of the target compared to the control (Fig. 5.1B). In particular, peaks at 7.58 ppm and 7.32 (control spectrum) showed the largest differences with 36% and 39% intensity changes, respectively. Changes for the same peaks were also observed in the WaterLOGSY spectra (Fig. 5.1C).

Peaks changed intensity sign upon binding to the target, and effects are clearly visible at 7.14 and 6.92, with a percentage change of 33% and 42%. Lastly, the STD spectrum (Fig. 5.1D,) showed easily distinguishable signals at 7.13, 6.89, 6.93, 6.92, 6.90 which had an average signal-to-noise ratio of 2.97. This suggested a significant STD response. The presence of significant effects across all three experiment types indicate that this compound could be classified as a level-3 binding SpectrumHit.

The second component, 467, showed a significant intensity reduction in the $T_{1\rho}$ Fig. 5.2B, with the largest percentage of 40% for the peak at 8.4 ppm, followed by 36% at 7.73 ppm. The WaterLOGSY spectra showed intensity changes for the same peaks, Fig. 5.2C. However, due to a large offset, ~0.87 ppm between the control spectrum and the target spectrum, and poor target signal suppression, results should be examined cautiously. Similarly, the STD spectrum Fig. 5.2D, showed a low average signal-to-noise ratio of ~1.41 and peaks were barely recognisable at 7.79, 7.25 and 6.96 ppm, suggesting a very weak STD response. Last group of spectra available for this dataset was relative to the component 7373 Fig. 5.3A. The $T_{1\rho}$ spectra showed very large percentage of changes for the peak at 8.01 and 6.35 ppm but most importantly, a significant shift among all peaks.

The next experiment type, WaterLOGSY Fig. 5.3C, showed sharp positive peaks in the spectrum for the sample recorded with the target, but unfortunately the control spectrum did not show any signal, apart for solvent, so a full comparison was not possible. Lastly, the STD spectrum, Fig. 5.3D, showed the highest average signal-to-noise ratio of 4.34. The clear singlet at 7.87 and two doublets at 7.5 and 7.2 ppm suggested this component to be an STD SpectrumHit.

The spectra for compounds 467 and 7373 were analysed in a similar fashion. A summary of the observed effects for the three components and each experiment type is shown in Fig. 5.4A. In the $T_{1\rho}$ experiment, all three components showed a high average intensity change for peaks in the aromatic region, with compound 467 recording the highest score of 0.30. This component, however, revealed the lowest scores for the WaterLOGSY and STD experiments with values of 0.14 and 0.03, respectively. The previous manual assessment classified this compound as a level-3 SpectrumHit; however, the very weak STD score, with a S/N < 1.5, should be carefully included as a binding spectrumHit.

Compound 7373, had its WaterLOGSY score arbitrarily set to 0.35 as it showed the sharpest positive peaks in the target spectrum; together with the highest STD score of 0.17 this compound could be considered the most prominent SpectrumHit of the dataset. Surprisingly, and in contrast to the automated result, the previous manual assessment classified this compound as a level-2 SpectrumHit.

The spectra of all compounds also displayed large shifts of the peaks (cf. Fig. 5.1). It is unclear if these shifts were due to different experimental conditions among the samples, e.g. control and in presence of the target, or that these shifts were genuine and the result of the interactions of the compound with the target.

In conclusion, the quality of the dataset was such that additional experiments would be necessary to confirm any of three components as binding to the target.

5.4.2. **Dataset-2**

The second dataset used to test AnalysisScreen comprised a group of 35 singleton STD spectra and a small group of eleven mixtures of 4 compounds each. Dataset-2 did not present any of the issues as previously described for dataset-1, and automatic routines could be tested using built-in pipelines. I specifically used this dataset to examine the effect of two automated peak picking algorithms, referred to as the *maximum_filter* algorithm[120] (MF) the *Peak Detector algorithm*[121] (PD).

Peaks were firstly picked in all reference spectra using the MF peak picker and inspected manually to ensure that peaks originating from the noise were not included. Secondly, peaks were identified manually, as well as using the MF and PD algorithms, in the STD control spectrum and STD spectrum in the presence of the target (Fig. 5.5A). There were no major differences between the two algorithms, although the MF algorithm required several trials before finding the optimal parameters. In fact, non-optimal parameters resulted in a large number of false positive or false negative peaks (not shown). A total of 29 SpectrumHits were found after running the newer PD algorithm, and only 27 when using the MF algorithm. Using manually picked peaks 28 SpectrumHits were observed (Fig. 5.5B).

This dataset also included eleven mixtures which were analysed using the same workflow as discussed previously. The total number of SpectrumHits resulting from this analysis is shown in Fig. 5.6. In particular, four spectra were flagged as SpectrumHit by the collaborator upon a visual inspection (VI col), whereas my personal visual inspection classified 14 SpectrumHits (VI). The latter result was taken as a reference point for comparing the various algorithms. The MF algorithm for the peak picking resulted in only four SpectrumHits to be identified, whereas the PD peak picker identified fourteen SpectrumHits. Peaks were also determined manually and using this information fifteen SpectrumHits were counted. I speculate that the collaborator flagged fewer SpectrumHits as a result of cross validating the data with other techniques. A successive analysis compared similarities between SpectrumHits found automatically or semi-automatically and those derived from my visual inspection (Fig. 5.7). Notably, all four SpectrumHits identified by the MF algorithm can be considered true positives; the PD algorithm performed significantly better, with a total of 13 correctly matched SpectrumHits. However, peaks identified manually, and those identified with the PD peak picker showed an additional SpectrumHit not identified through the visual inspection, which was considered being a false positive (Fig. 5.7B). Importantly, by missing ten of the known SpectrumHits the MF algorithm displayed a very high false negative rate when compared with the PD peak picker (Fig. 5.7A). A further statistical analysis between these two methods showed a slightly higher precision count for the MF peak picker, but a significant reduction of sensitivity, accuracy and specificity compared to the PD peak picker (Fig. 5.8).

These results confirmed that the PD peak picker was a better alternative to the originally developed MF peak picker and therefore used as preferred automated algorithm for the analysis of other available datasets.

5.4.3. **Dataset-3**

I next tested the performance of the automated STD analysis implemented in AnalysisScreen using dataset-3 containing 310 experimental STD spectra, acquired for samples in the presence of a biological target and mixture compositions of up to five components. Firstly, spectral peaks were manually picked for all available spectra. Using simplified tools also available in AnalysisScreen (as described in chapter 3), each STD spectrum was visually inspected by comparing it to its single spectral reference. A total of 18 compounds displaying STD effects were considered being true positive SpectrumHits (Fig. 5.9A). The same number of SpectrumHits were found by using an automated matching routine. However, from the analysis reported on by the AnalysisScreen Hit Analysis module (cf. chapter 3), I noticed that most of STD spectra were uniformly misaligned to their corresponding reference spectra (Figs 5.10A-B), suggesting a referencing issue. For the dataset under examination, a total shift of 0.0075 ppm was determined (Fig. 5.10B) and applied to the STD spectra. Finally, they were re-matched to the reference data and re-evaluated.

Ultimately, a complete pipeline, consisting of automatic peak picking, re-referencing, and hit detection pipes was applied to the dataset. A total of 29 SpectrumHits were found (Fig. 5.9A). Using the Hit Analysis module, the previously not identified SpectrumHits were inspected and confirmed as true positive SpectrumHits, albeit some had very low scores (Fig. 5.11). However, four compounds previously flagged as SpectrumHits were now no-longer founds (Figs 5.9B, 5.11), usually as a result of being below some pre-set threshold values, e.g. the spectral Signal-to-Noise Ratio, S/N. Some spectra, in fact, appeared to be very noisy and difficult to interpret. In line with the simulated datasets (cf. chapter

3), for the experimental STD data SpectrumHits for peaks with a S/N lower than 1.5 were barely recognisable from the overall noise, and therefore were excluded as true positives. Lastly, a graphical summary of the scores versus S/N is shown in Fig. 5.9C.

5.4.4. **Dataset-4**

Dataset-4 differed considerably compared to datasets 1-3 described above. Dataset-4 contained multiple processed spectra for each compound, in addition to spectra recorded for three different experiment types, i.e. STD, ^{1}H and a CPMG relations series. The latter prompted the inclusion of its relevant parameters in the excel reader (discussed in chapter 2), as well as additional changes in the AnalysisScreen core data structure in order to accommodate the additional information. Furthermore, since the workflow was novel, some steps necessary for the analysis were not (yet) available in the GUI pipeline and were therefore performed using the built-in command-line interface of AnalysisScreen.

Starting from the STD analysis, peaks were initially identified in the Off-resonance spectra and transferred to the On-resonance spectra followed by a peak refitting to establish the new local maximum. STD spectra were obtained by the common subtraction of On- and Off-resonance spectra and peak efficiencies were calculated as in Eq. 3.6. To identify any potential SpectrumHits, it would be required that the signal-to-noise would be sufficient, while simultaneously the STD effect should be significant. Hence, for each STD spectrum the S/N was calculated and plotted against the total STD efficiency (defined as in eq. 3.6), Fig. 5.12A. A small cluster of reference spectra that display a S/N value >1.5, in accordance with the simulated data of chapter 3, and > 0.5 total efficiency score. This region is expected to contain the potential SpectrumHits, which can be further analysed in the subsequent drug discovery steps, for example using parallel NMR experiments, such as the Chemical Shift Mapping analysis discussed in the following chapter.

In order to identify any potential similarities between the SpectrumHits, a Kernel Density Estimation (KDE) between the single peak efficiency and the peak chemical shift was calculated (Fig. 5.12B). The KDE plot shows that the highest scoring region is around 7.25 to 7.5 ppm. It shows a central efficiency score of ~10% and outliers up to 30%. Example of STD spectra are shown in Fig. 5.13; a visual inspection of the spectral appearances suggests that the proposed categorisation (cf. Fig. 5.12A) adequately classifies the STD spectra.

Next, the spectra acquired using the CPMG technique were inspected. For each mixture, peaks were initially identified in the control spectra recorded at 0 ms relaxation time and the peaks were then transferred to all other six spectra in the series, for both control and target spectra. From assessing the decay of the peak heights as a function of relaxation time, $T_{1\rho}$ spectrumHits were determined from Eq. 5.3 (Fig. 5.14A). Only the top 50 scoring spectra were used for further investigation, which corresponded to compounds with scores above 3 units. Again, a KDE plot correlating the chemical shift position and the score was also calculated (Fig. 5.14B). As also observed for the STD result (cf. Fig. 5.12B), the highest probability of identifying SpectrumHits is centred between ~7.3 and 7.5 ppm. To establish the optimal sensitivity of the CPMG method, a KDE was also calculated for the decay time and the ppm positions (data not shown). This analysis showed the highest density in a time region between 400 and 200 ms with a normally distributed ppm positions and a maximum at around 7.4 ppm.

From the results it was possible to differentiate between non-hits and potential SpectrumHits (Fig. 5.15). Putative SpectrumHits were characterised by an estimated $T_{1\rho}$ time between 250 and 400 ms, with a median of ~330 and an average of ~376 ms, whereas, non-binding fragments showed a $T_{1\rho}$ time between 400 and 880 ms, with an estimated median of ~510 and an average of ~1819 ms. Fig. 5.16 displays typical examples of both the spectra and the resulting decays curves.

Lastly, the plain ^{1}H SpectrumGroup was assessed by comparing changes in intensities between control and target spectral signals. Scorings were calculated as previously discussed from Eq. 5.1. The scores distribution was similar to the CPMG, with a maximum peak at 1.5 total score, also in this case, only the top 50 compounds were arbitrarily advanced for further inspection, Fig. 5.17A. Finally, the single score and ppm position for each peak corresponding to the top SpectrumHits were calculated in a KDE.

For this experiment type, scores were localised at a value of ~0.7 for peak SpectrumHits recorded in a region between 7.4-7.5, which was consistent with previously reported CPMG and STD SpectrumHits-range.

I believe that the KDE report, when performed with the spectral peak assignments, might help in elucidating the binding orientations of a specific cluster of compounds, therefore providing an initial structure activity relationship (SAR) from the primary hit identification.

Finally, for the top 50 SpectrumHits the confidence level was determined using all experiment types. Only one compound was classified as level-3 SpectrumHit, common to all experiments. Seven compounds showed a level-2, common in at least two experiments and 133 compounds were classified as level-1.

5.5. Conclusions

The variety of datasets, and experiment types, personal working preferences, and other user customisation of datasets as demonstrated by the four examples in this chapter, renders a one-off solution for data analysis nearly impossible. Instead, specific workflows need to be created for each case. The flexibility of CcpNmr AnalysisScreen allows for a straightforward adaptation to specific user's needs with bespoke pipelines or simple macros, that can be written and run within the program.

In this chapter I described the analysis of four different datasets, which were provided by different collaborators. Each dataset presented a completely different data-structure, including the mapping of nomenclature and associated metadata. However, the presence of an Excel file-parser made it possible to parse and load within the programme all datasets with ease.

Every dataset presented some peculiarity, that made the initial analysis more challenging than expected, especially if it was compared to the analysis of a simulated dataset as discussed in chapter 3.

The dataset-1 was characterised by severe phasing issues, severe referencing misalignment across the different spectral types. Furthermore, weak ligand signals, improper protein and solvent signal suppression, together with a portion of the reference dataset missing, made the design and execution of a fully automate analysis very challenging. Instead, spectra were re-phased using a single pipeline and peaks were picked manually with the built-in manual selection of AnalysisScreen. Finally, intensities signals were compared using a custom macro and scores were reported (Fig. 5.4). In contrast to the collaborator's qualitative definition of SpectrumHits, AnalysisScreen yielded a much more exhaustive scoring classification.

The dataset-2 was composed of singleton STDs, mixtures and their spectral references. The collaborator shared its own qualitative judgment on the singleton analysis but could not fully identify SpectrumHits from the mixtures due to overcrowded regions and difficulties in spectral deconvolution.

This dataset was primarily used for testing the performance of two peak picker algorithms, denoted as MF and PD. The PD peak picker algorithm proved to be more reliable with higher sensitivity and accuracy scores (Figs 5.8 B-C). Compared to the MF algorithm, the PD algorithm is also easier to optimise for performing an automated screening analysis as it only requires one user adjustable parameter. By optimising the speed of the algorithm from seconds to milliseconds in execution time per spectrum, it now presents a superior practical choice.

Dataset-3 was used for testing the performance of the automated STD analysis. SpectrumHits were grouped by a scoring function based on signal intensity and estimated signal-to-noise ratio, validating the model built using the simulated dataset as previously discussed in chapter 3. Notably, the hit analysis indicated referencing issues between the spectral references and experimental STD (Fig. 5.10B), resulting in a global re-alignment. The fully automated hit-detection pipeline identified more genuine SpectrumHits (Fig. 5.11) compared to previous manual visual inspections by both our collaborator and myself.

Dataset-4 was used to evaluate both ^{1}H, STD and CPMG series. This dataset also presented some troubling features, including very high STD efficiencies in solvent regions, ^{1}H target spectral intensities that were higher than their control, and CPMG series with scattered intensity outliers that prevented an optimal performance of fitting routines. However, this dataset was fundamental for developing the AnalysisScreen routines and assisted in further validating the various scoring functions, identifying potential SpectrumHits (Fig. 5.13).

In this chapter, I have shown how automated computational tools can drastically reduce both the time and bias to determine the output of screening by NMR compared with manual analysis, including reducing false positive and false negative observations. The inherent flexibility of AnalysisScreen allows for its continuous development. The collaborations with both academic and industrial partners will ensure the completion of a robust package capable of dealing with a variety of NMR datasets.

5.6. Figures

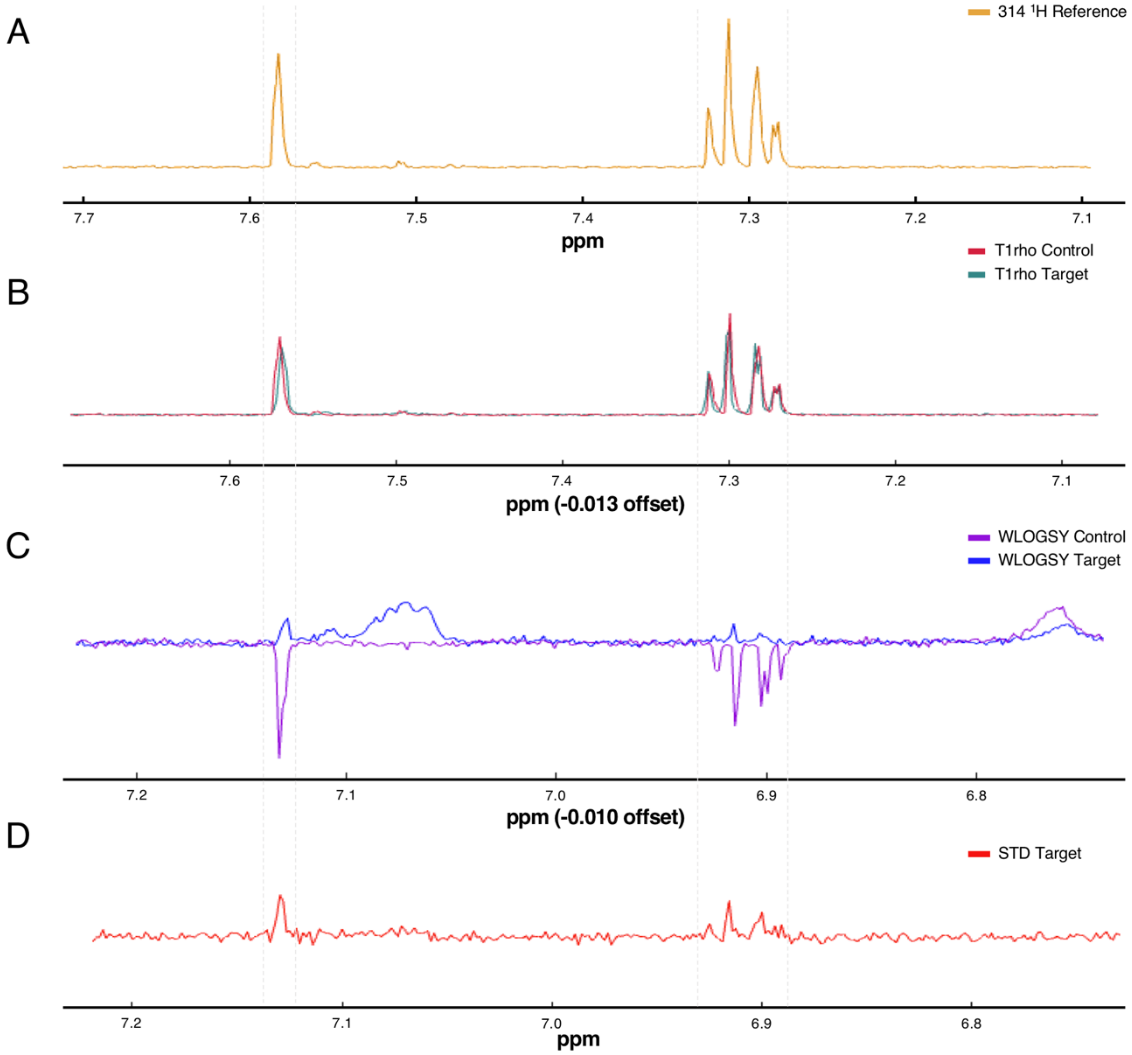


Figure 5.1 Dataset-1, multiple experiments for component-314.

A ^{1}H reference spectrum for the substance 314. **B** $T_{1\rho}$ control spectrum in red, and in presence of a target, light green. Spectra have an offset of ~0.0013 pm to each other. **C** WaterLOGSY control spectrum, purple, and after the addition of the target, blue. Positive peaks relative to the compound-314 appeared in several positions. Spectra are shifted by 0.010 ppm to each other. **D** STD spectrum. The dotted lines highlight the aligned areas where major spectral changes occurred. For all spectra: proper uniform referencing was likely severely compromised (see text for details).

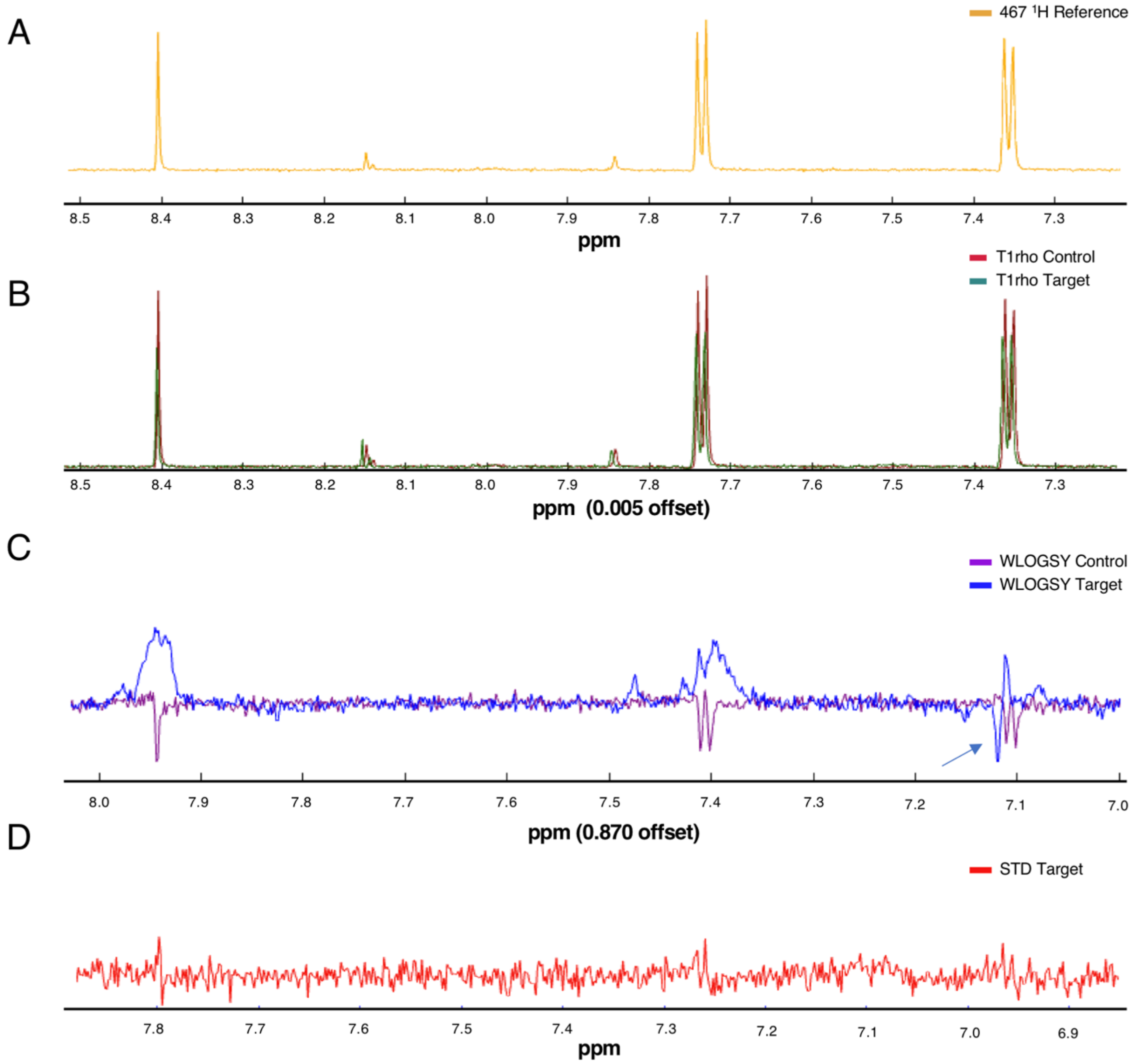


Figure 5.2 Dataset-1, multiple experiments for component-467.

A ^{1}H reference spectrum for the substance 467. **B** $T_{1\rho}$ control spectrum, red, and in presence of a target, light green. Spectra have an offset of ~0.005 pm to each other. **C** WaterLOGSY control spectrum, purple, and after the addition of the target, blue. Positive peaks relative to the compound-467 appeared in several positions, although an unexpected negative peak of the doublet at 7.1 ppm occurred (blue arrow). Spectra are shifted by 0.870 ppm. **D** The weak STD spectrum relative for this component.

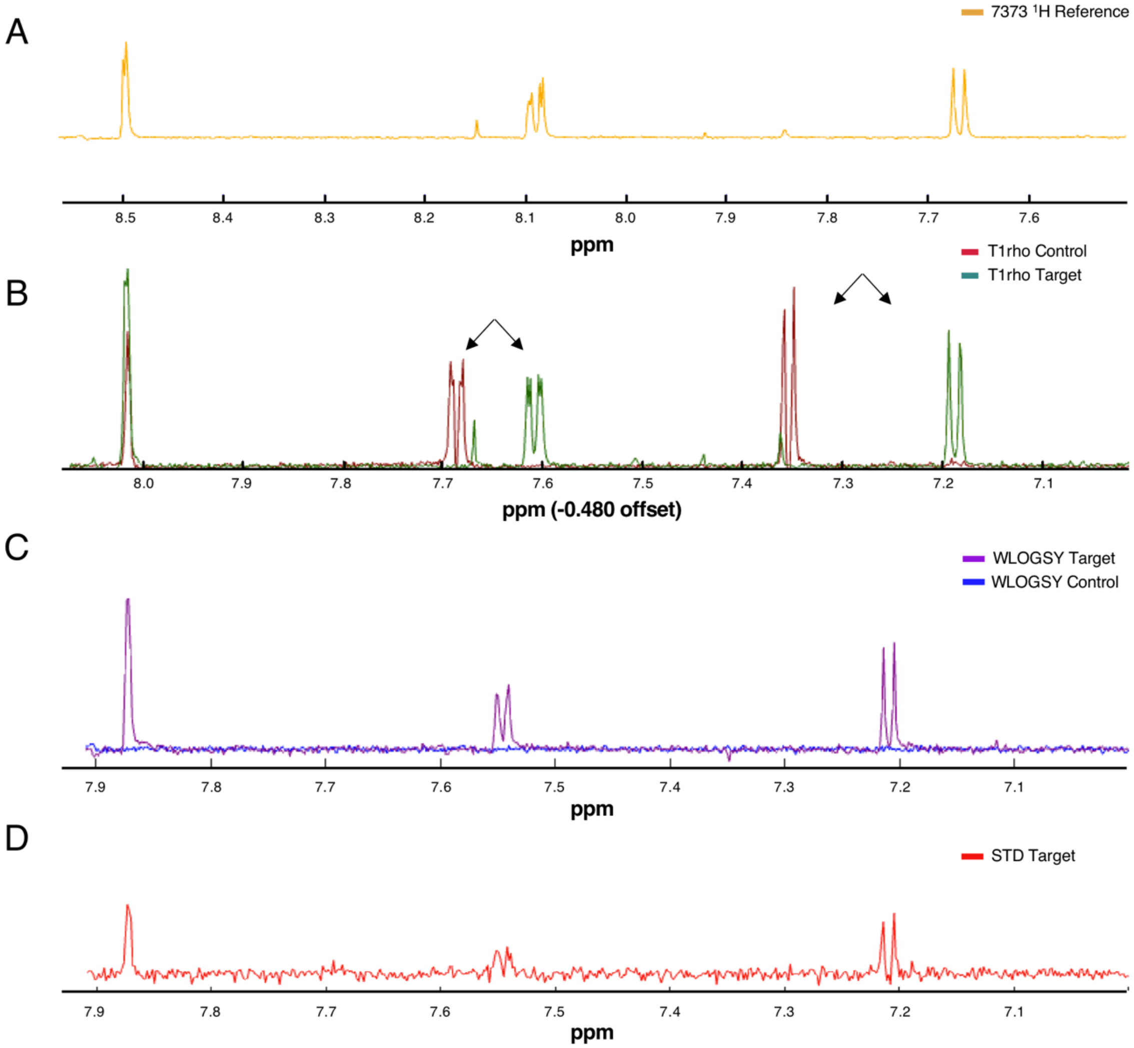


Figure 5.3 Dataset-1, multiple experiments for component-7373.

A ^{1}H reference spectrum for the substance 7373. **B** $T_{1\rho}$ control spectrum, red, and in presence of a target, light green. Spectra have an offset of ~0.480 pm to each other and aligned to the singlet at 8.01 ppm, large peak shifts were observed for these spectra (black arrows). **C** WaterLOGSY control spectrum, blue, and after the addition of the target, purple. Note the control spectrum was characterised by the total absence of the ligand signals due to improper solvent suppression. **D** STD spectrum relative for this component.

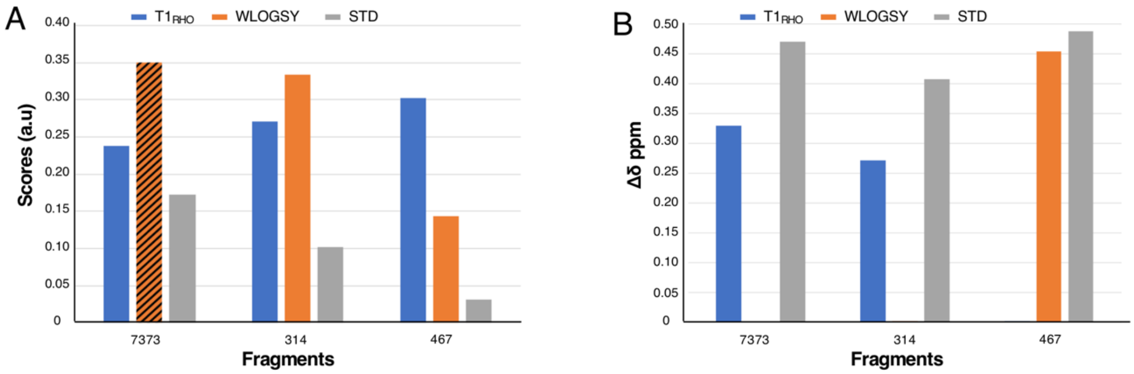


Figure 5.4 Dataset-1, scorings summary.

A Scores for $T_{1\rho}$ (blue), WaterLOGSY (orange), STD (grey) for the three analysed components. **B** Recorded peak shifts for $T_{1\rho}$ (blue), WaterLOGSY (orange), STD (grey) for the three analysed fragments.

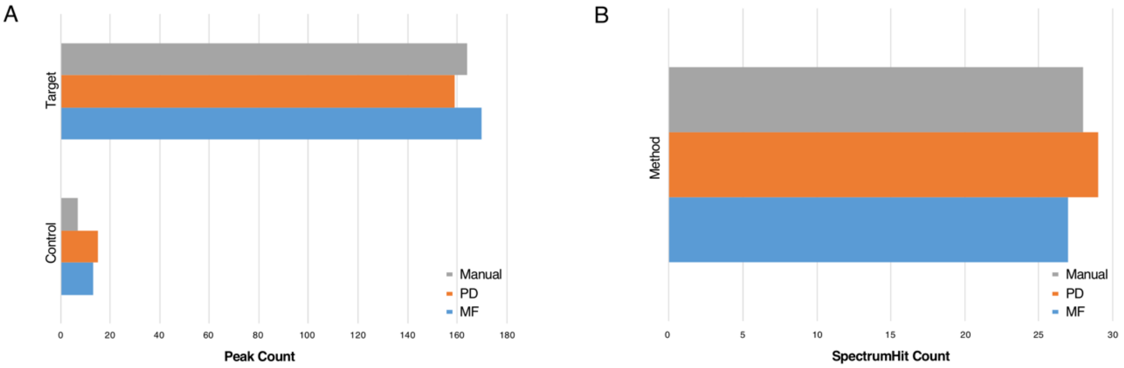


Figure 5.5 Dataset-2. Singleton scorings summary.

A Total peak count for the target and control spectra using three different methods: Manual (grey), MF peak picker (blue) and PD peak picker (orange). After several trials in optimising running parameters for the MF algorithm for this dataset, it found most of the spectral signals, including lower S/N ratio peaks (blue). On the contrary, the PD algorithm found a slighter higher number of signals without manual adjustment of parameters (orange). **B** Total SpectrumHits count using peaks determined by the manual picking, (grey), the MF peak picker (blue) and the PD peak picker (orange). The latter identified all the manually detected SpectrumHits plus a lower S/N observation, which was missed by the manual inspection (grey) and the MF algorithm (blue).

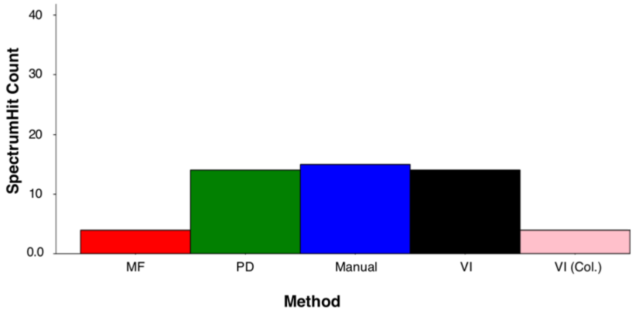


Figure 5.6 Dataset-2. Mixtures scorings summary.

Total SpectrumHit counts using peaks determined by the MF peak picker (red), the PD peak picker (green), a manual picking, (blue), my personal visual inspection (VI), (black) and the collaborator visual inspection, VI Col. (pink).

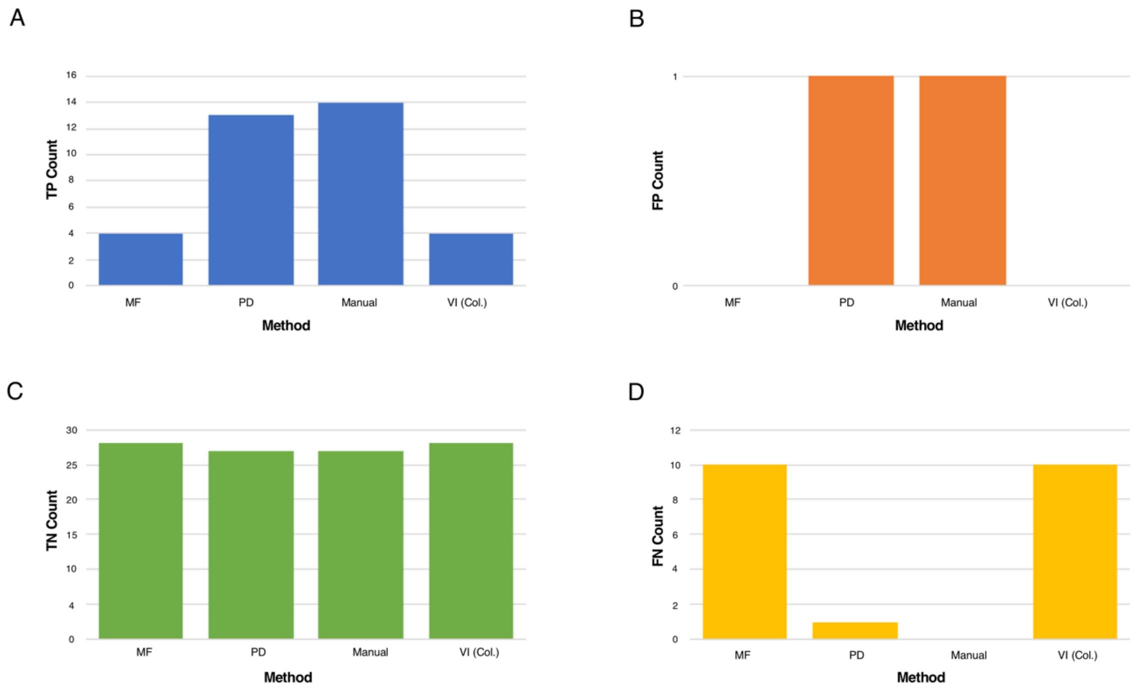


Figure 5.7 Dataset-2. SpectrumHits statistics.

A True positive SpectrumHit count compared to the personal visual inspection. **B** False positive SpectrumHit count. **C** True negative SpectrumHits count. **D** False negative SpectrumHit count. For this experiment, my manual inspection was used as true reference count, as I speculated the collaborators were biased by additional results observed using parallel hit identification techniques, such as X-ray crystallography.

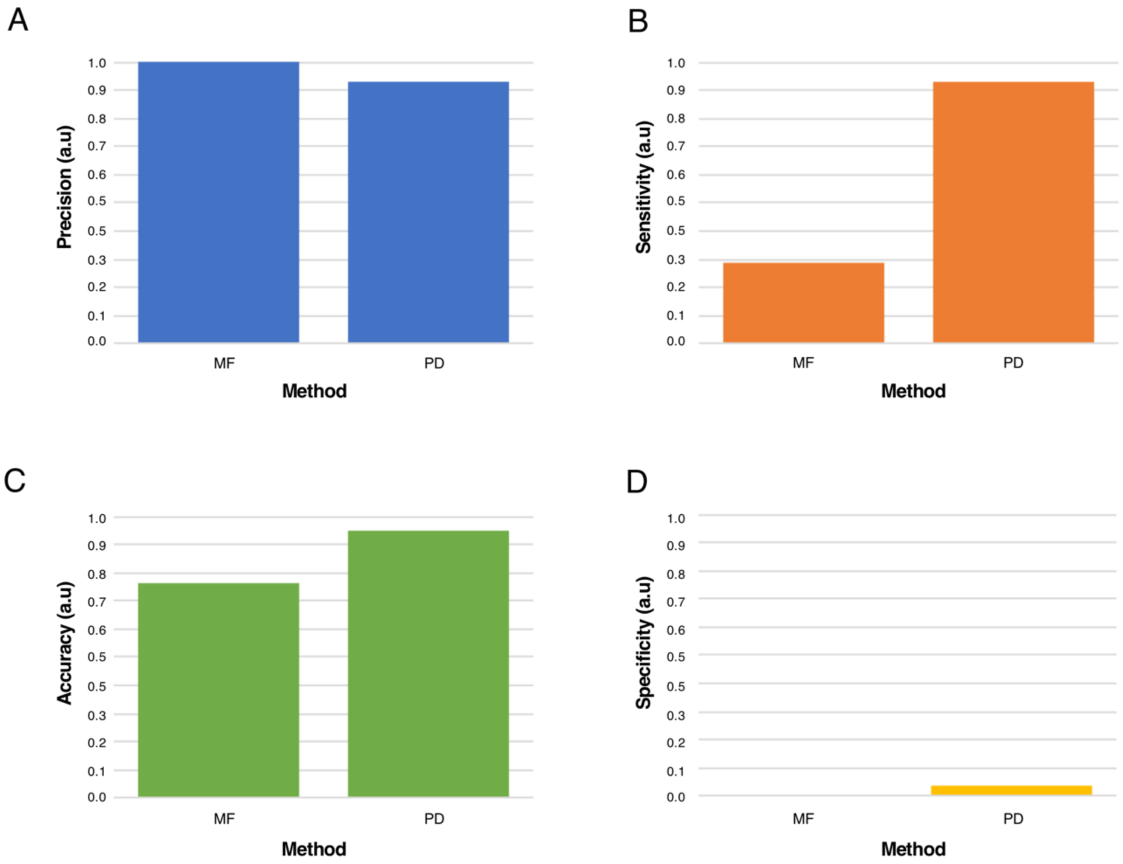


Figure 5.8 Dataset-2. Peak detection statistic scores.

A Precision, **B** Sensitivity, **C** Accuracy **D** Specificity in scale 0 to 1 units for the two automatic peak picker algorithms.

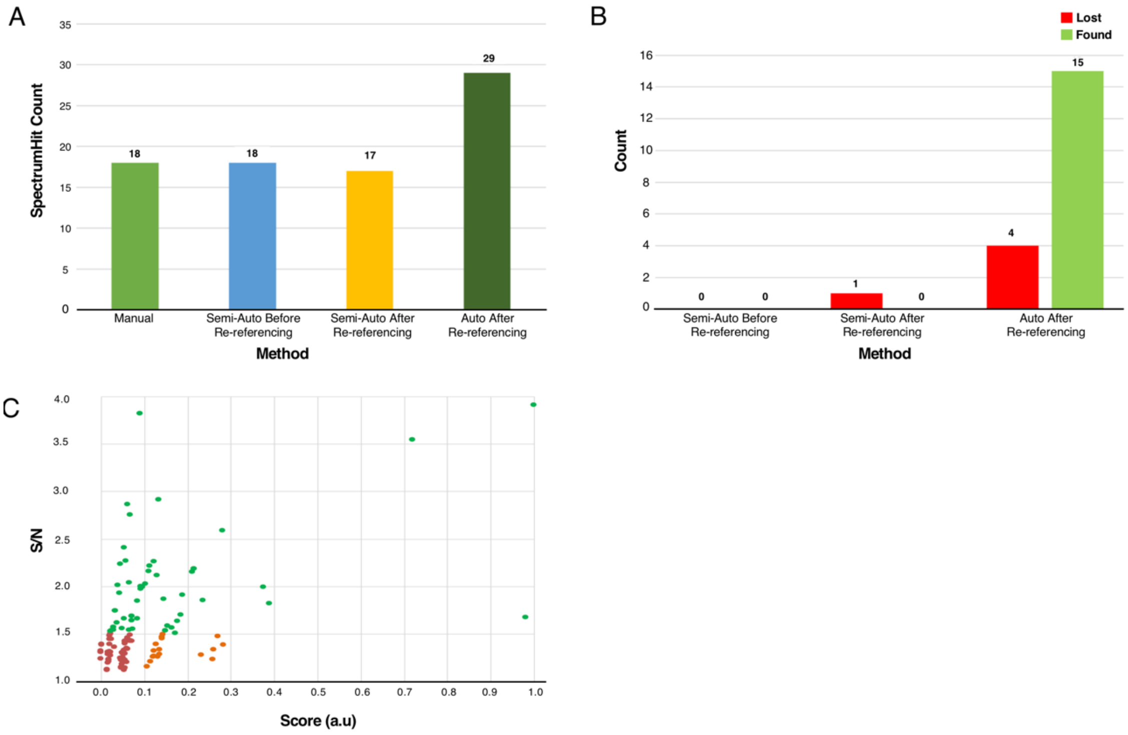


Figure 5.9 Dataset-3. Automated *versus* manual SpectrumHit detection.

A Total number of SpectrumHits obtained by a visual inspection using manually picked peaks (light green bar); SpectrumHits obtained by the SpectrumHit detection pipeline before and after re-referencing, using the same previously manually picked peaks (blue and yellow bars) and SpectrumHits obtained after re-referencing and automatic peak detection using default parameters (dark green). **B** Newly detected and lost SpectrumHits counts between the four methods. Notably, the automatic approach showed 15 new potential SpectrumHits which were missed upon manual analysis. **C** Global STD scores versus S/N. Red dots represent compounds that are unlikely to be SpectrumHits, orange dots, dubious SpectrumHits, whereas green dots denoted confirmed SpectrumHits.

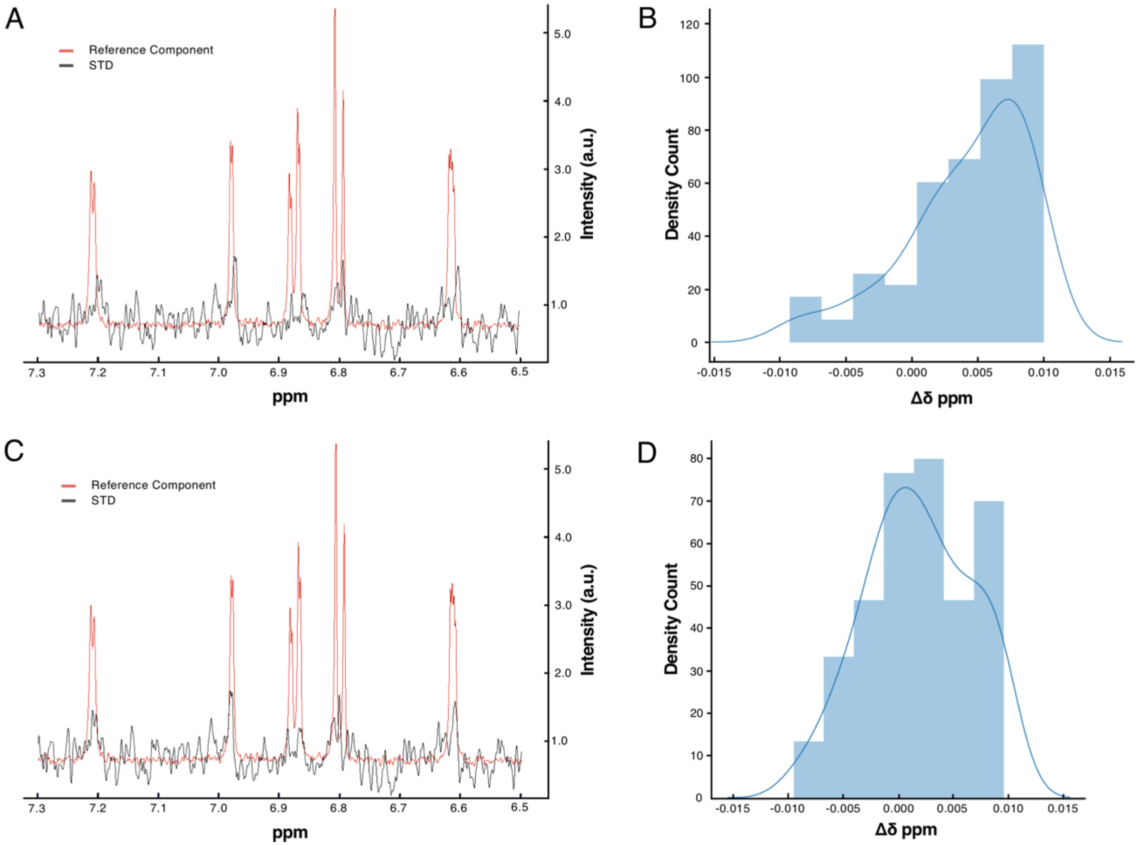


Figure 5.10 Dataset-3. Re-referencing of spectra datasets.

A and **C** show an example of an STD SpectrumHit and its best-matched reference before and after applying a re-referencing pipe. **B** and **D** illustrate peak shift distributions of experimental STD spectra to their reference spectra before and after a re-referencing pipe was applied. The maximum of the distribution, ~0.0075 ppm (extracted from **B**), was used to calculate the total adjustment needed to re-reference the STD spectra to their references. **D** New distribution after the adjustment was applied, with a maximum centred around ~0.000 ppm

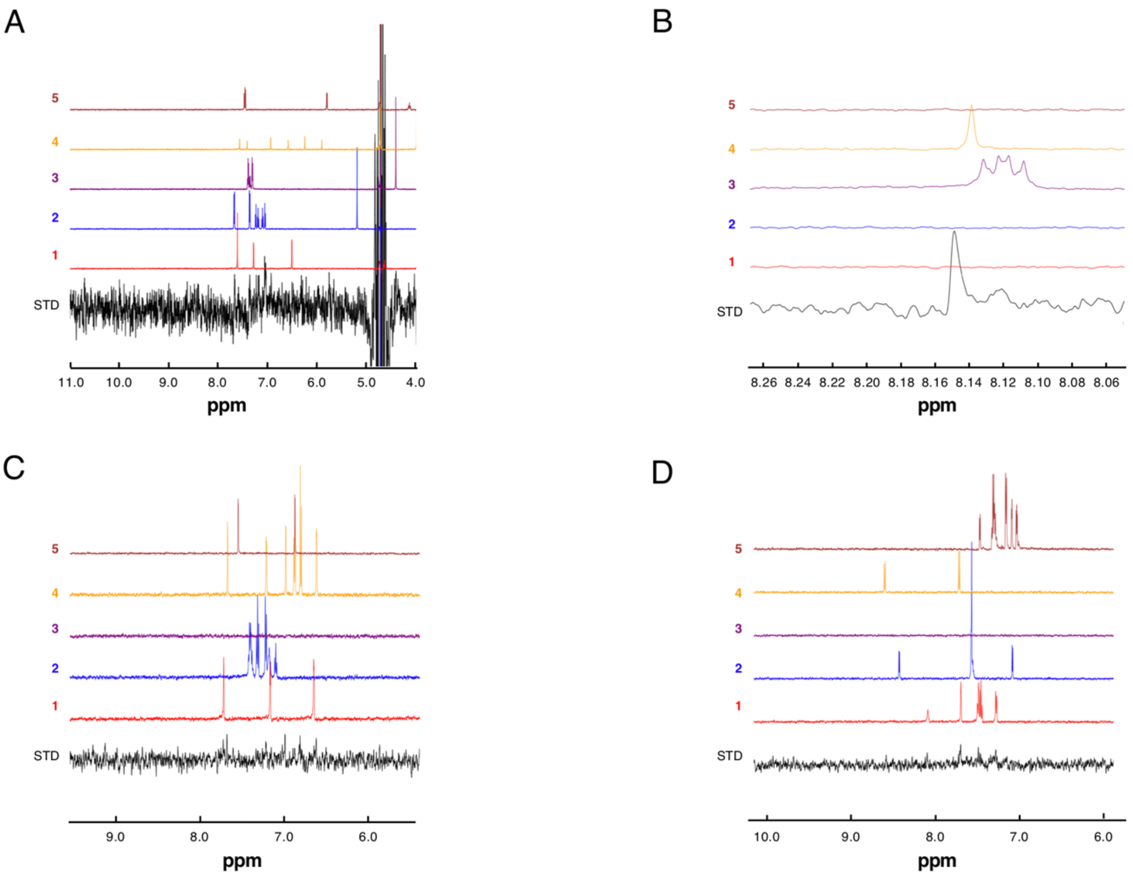


Figure 5.11 Dataset-3. Examples of automatically detected SpectrumHits. Example experimental STD spectra (black) and five relevant reference spectra are shown. **A** Example of an STD spectrum which was discarded as a true positive on the basis of a low S/N. **B** Example of an STD spectrum with a peak at 8.149 ppm which did not match any of the reference spectra and therefore was excluded as a true positive SpectrumHit. Examples of STD spectra (**C** and **D**) with very weak matching signals previously not identified during manual inspection.

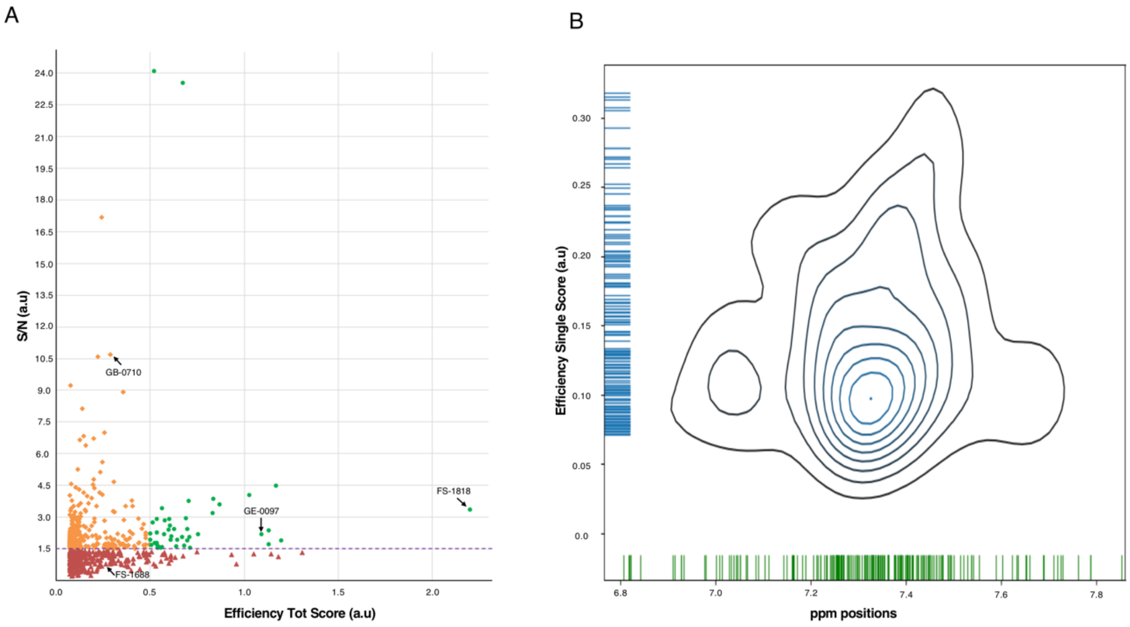


Figure 5.12 Dataset-4. STD scores.

A Total STD efficiency score versus Signal-to-noise (S/N) ratios per reference spectrum. Green dots denote expected SpectrumHits, above a threshold S/N of 1.5 and total efficiency of 0.5; orange dots denote compounds which showed STD spectra showed a S/N over 1.5 but lower than 0.5 total efficiency; in red, compounds below the threshold limits. The arrows indicate an example of spectra for each group and are displayed in the next figure (Fig. 5.13). **B** Kernel Density Estimate plot of STD single efficiencies and ppm positions for each peak determined in the top SpectrumHits.

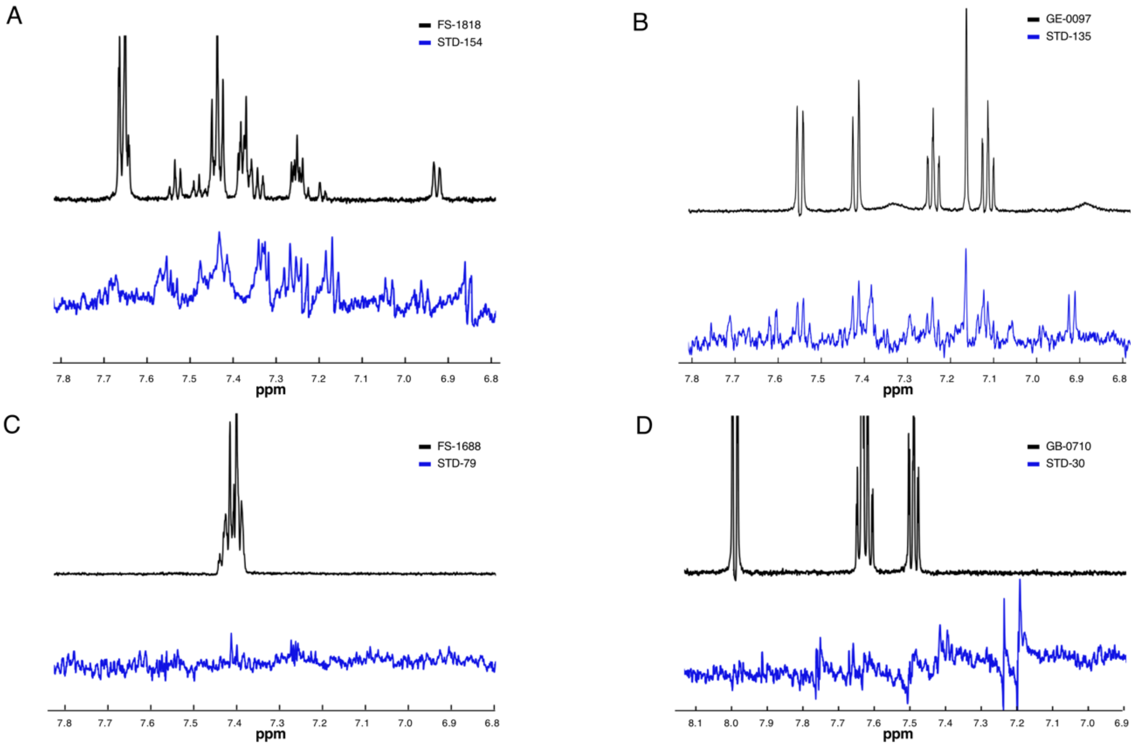


Figure 5.13 Dataset-4. STD spectra examples.

A STD and matched reference for the top-scoring SpectrumHit, (high S/N and high efficiency score). **B** STD and matched reference spectrum for a typical SpectrumHit, (medium S/N and medium efficiency score). **c** STD and matched reference for a random compound in the red cluster of Fig. 5.12A, (low S/N and low efficiency score). **D** STD and matched reference for a random compound in the orange cluster of Fig. 5.12A, (high S/N and low efficiency score).

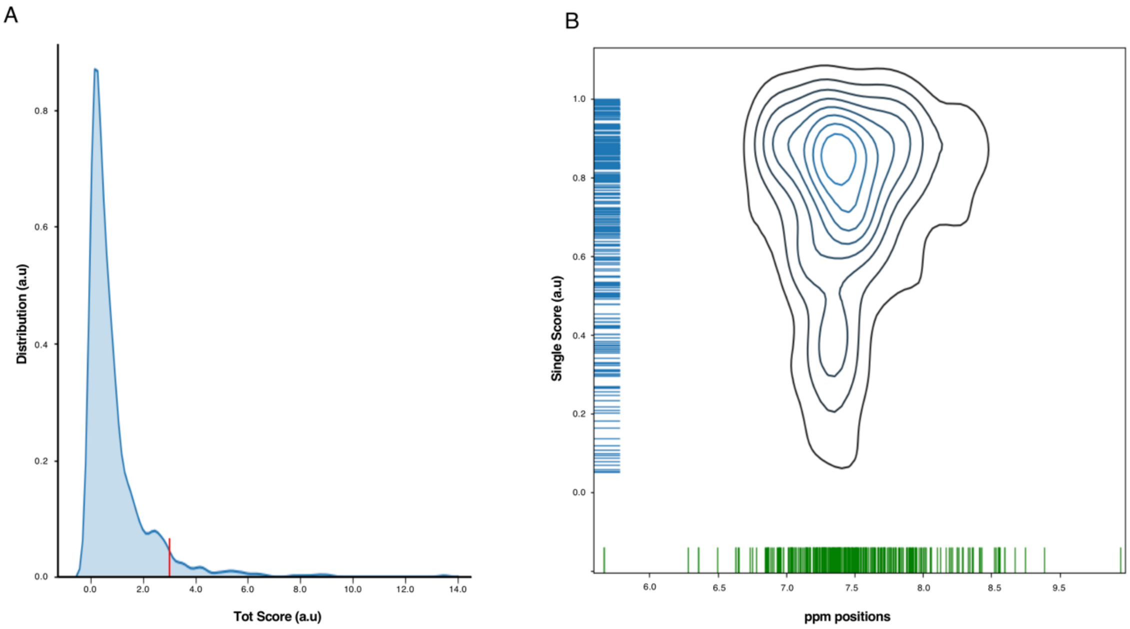


Figure 5.14 Dataset-4. CPMG scores.

A Total CPMG score distribution, values above the red vertical line corresponds to the threshold value up to where consider spectra as SpectrumHits. **B** KDE plot of CPMG single efficiencies and ppm positions for each peak determined in SpectrumHits above the threshold value.

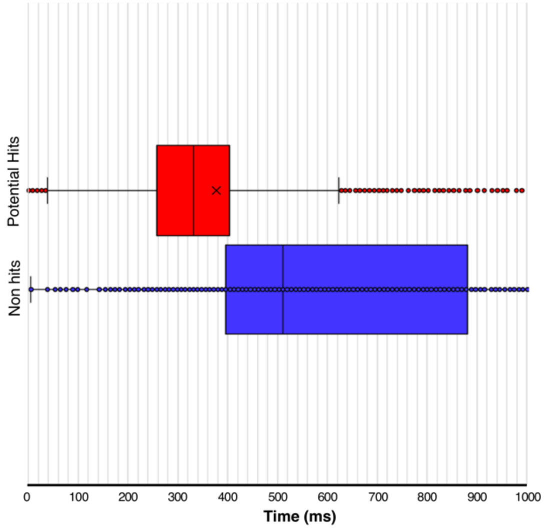


Figure 5.15 Dataset-4. CPMG results.

Representation of the CPMG results as a box plot. Red box the potential SpectrumHits, blue, spectra unlikely to be SpectrumHits. Vertical line within the box indicates the median, whereas the “X” indicates the average value.

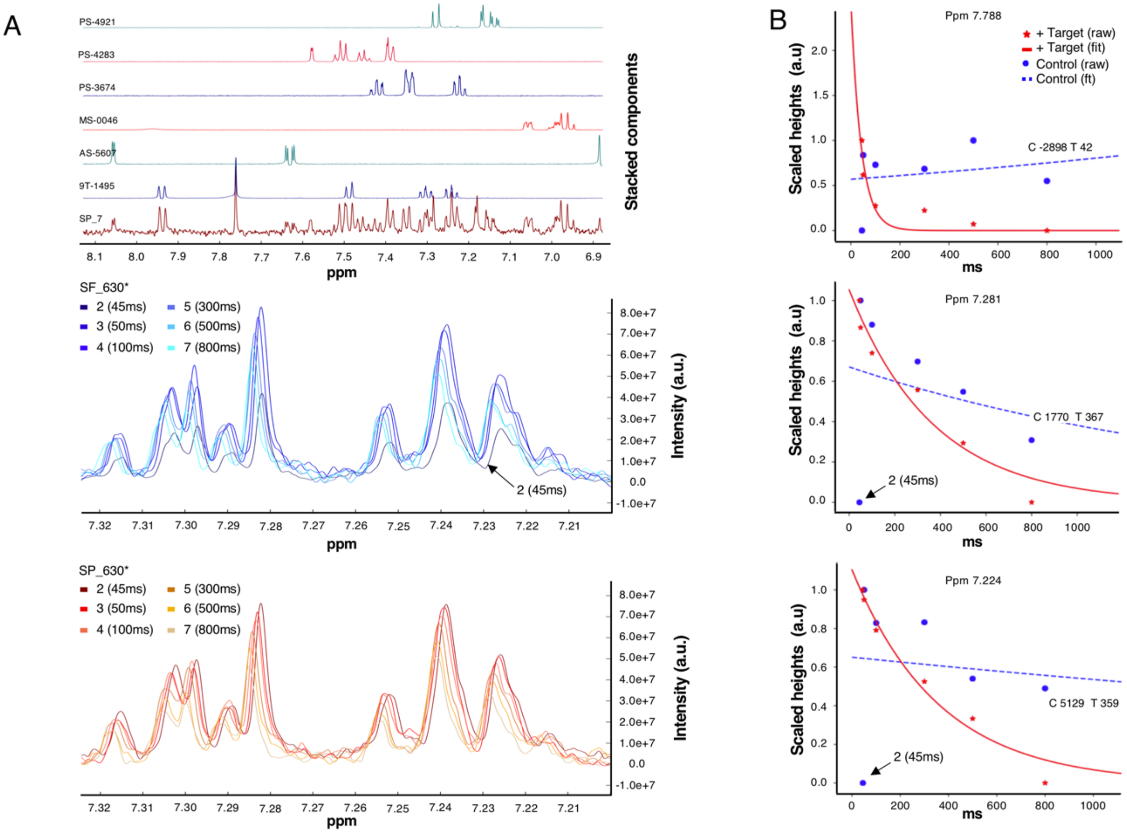


Figure 5.16 Dataset-4. Example of CPMG SpectrumHits.

A top panel, references for the mixture-63 re-referenced to the last spectrum of CPMG series. Middle panel, control spectra, prefix "SF". Spectra are coloured in a blue gradient; darker colours correspond to shorter times. Lower panel, spectra recorded in presence of the target, prefix "SP". Spectra are coloured in a red/yellow gradient, dark red colours correspond to shorter times. **B** Example of fitting plots were the higher pattern changes were indicated. Relaxation plots are automatically generated for each peak in the dataset and exported as a pdf file from further inspections. A systematically significantly lower intensity was observed in spectra recorded at 45 ms relaxation time, especially for the control series.

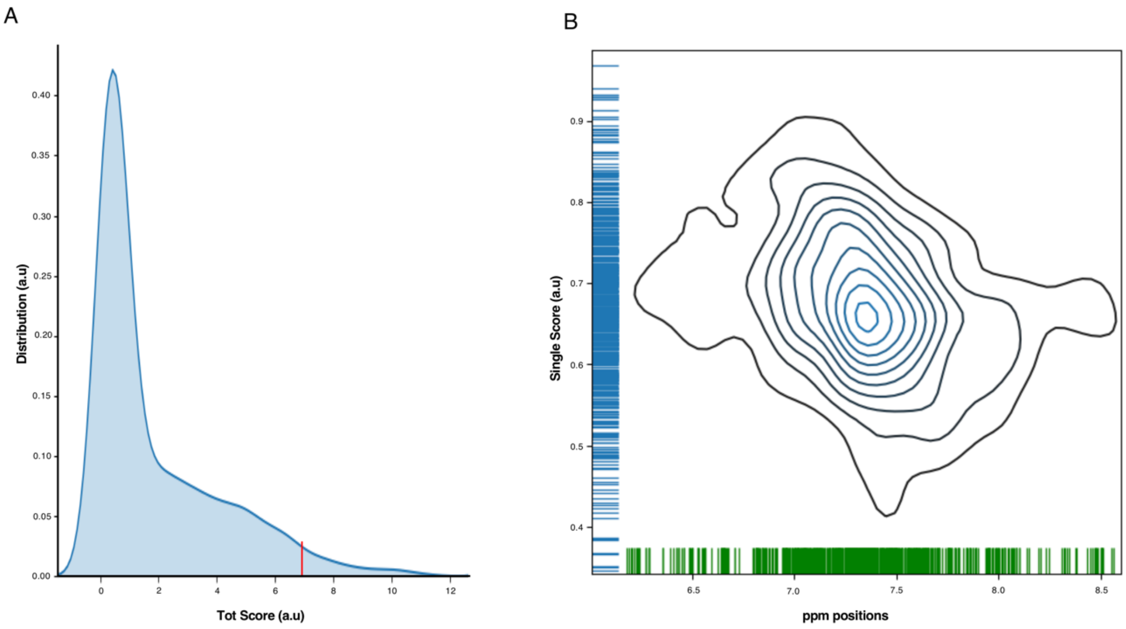


Figure 5.17 Dataset-4. ^{1}H scores.

A Total ^{1}H score distribution, red vertical line corresponds to the threshold value up to where consider spectra as SpectrumHits. **B** KDE plot of ^{1}H single efficiencies and ppm positions for peak hits.

5.7. Acknowledgments

Thanks to Drs Ken Kameron[a], Andrea Gohlke[a], Patrick McIntyre[b], Christine Prosser[b,c] for providing the dataset required for these analyses. Special thanks to Dr Hari Arthanari[d] for providing the largest and more complete dataset from the HMS NMR-core facilities.

5.8. Addresses

[a] Beatson Institute for Cancer Research, Garscube Estate, Switchback Road, Bearsden, Scotland G61 1BD, United Kingdom.

[b] Leicester Institute of Structural and Chemical Biology, Department of Molecular and Cell Biology, University of Leicester, Henry Wellcome Building, Lancaster Road, Leicester LE1 7HN, United Kingdom.

[c] UCB Celltech, Bath Road, Slough, Berkshire SL1 3WE, United Kingdom.

[d] Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, USA.

Chapter 6

Binding Sites Identification Tools in AnalysisScreen: The Chemical Shift Mapping Module

See List of Publications: 2.

Keywords: NMR, Chemical Shift Mapping, Tstar, CcpNmr AnalysisAssign, AnalysisScreen, 15N-HSQC.

6.1. Abstract

Nuclear Magnetic Resonance (NMR) is one of the major techniques for investigating the structure, dynamics and interactions between biomolecules. However, non-experts often experience NMR experimentation and data analysis as intimidating. I discuss a simple yet powerful NMR technique, the so-called chemical shift perturbation (CSP) analysis, as a tool to elucidate macromolecular interactions in small- and medium-sized complexes, including protein-protein, protein-drug, and protein-DNA/RNA interactions. I discuss current software packages for NMR data analysis and present a new interactive graphical tool implemented in CcpNmr Analysis version-3, which can drastically reduce the time required for the CSP analysis. Lastly, I illustrate the usefulness of a protein 3D structure for interpretation of the CSP data.

6.2. Introduction

It is the ultimate aim of the molecular biologist to understand cellular functioning in its molecular context. As such, it is imperative to know at which time and place specific biomolecules are active to exert their function. At the root of our understanding, however, is the realisation that the interactions between individual molecules, that together form active complexes of sometimes intricate complexity, constitute the underpinning basis of all the biological processes. Structural biology is the field of science which aims to describe such interactions between biologically relevant molecules at an atomic level. It is based on the notion that the interactions are facilitated by the specific molecular shapes and, as has nowadays become evident, also their dynamical changes. Together these are crucial in determining the affinities that drive the assembly of the macromolecular complexes[148].

Nuclear Magnetic Resonance (NMR) is one of the three major techniques that provides structural, dynamical and also interaction data[149]. In this work I will illustrate how a simple yet powerful experimental NMR technique, the so-called chemical shift perturbation (CSP) analysis, can be used to investigate interactions between biomolecules or biomolecules and small drug-like compounds. Since it was first proposed, the CSP analysis has become well-established, as illustrated by the increasing number of papers referring to the technique (Fig. 6.1A), with currently ~80 references annually. In this chapter, I also discuss how current NMR software packages can facilitate the CSP data analysis. Particular focus will be given to the CcpNmr AnalysisAssign version-3, which provides several user-friendly tools for retrieving the relevant data thus, providing for invaluable biological information.

6.3. Chemical shift and exchange

NMR is a spectroscopic technique that employs an inherent property of many nuclei called “spin” to yield spectra of various nuclei of biological interest, e.g. ^{1}H, ^{13}C, ^{15}N and ^{31}P. NMR’s exquisite (bio-)chemical usefulness originates from its ability to discriminate the different nuclei in a biomolecule. The electronic environment of each nucleus slightly modifies its exact resonance frequency through a process called chemical shielding and consequently the positions of various peaks in the NMR spectrum are specific for each nucleus in the molecule. The position of a peak in an NMR spectrum is commonly referred to as the “chemical shift” and denoted by the symbol δ. The chemical shift constitutes one of the most important parameters and in NMR provides a powerful tool for a biochemist; as it not only allows us to discriminate one nucleus from another, but also provides information about their conformation and nearby chemical environment. For example, an analysis using recorded chemical shifts from the Biological Magnetic Resonance Data Bank (BMRB) reveals distributions with median values of ~8.24, ~4.32 ppm and ~1.39 ppm for the Alanine H^{N}, H^{α} and H^{β}, respectively (Fig. 6.1B). The spread of each of the three distributions reflects the different conformations, i.e. the chemical environments, and dynamics, i.e. the change in these environments, of each of the nuclei in the various Alanine residues in their respective proteins.

In practice, the analysis of one-dimensional biomolecular spectra is prohibitive because of spectral overlap. To overcome this problem, it is possible to correlate one nucleus with another, generating two-dimensional (2D) or even higher-dimensional (3D, 4D, nD) NMR spectra. For proteins in particular, a simple and very informative example is the 2D heteronuclear ^{15}N-HSQC[150] (or alternatively for larger proteins the ^{15}N-TROSY-HSQC[151]) experiment. The resulting 2D ^{15}N-HSQC spectrum affords greater resolution and valuable information as peaks can be used as a “fingerprint” of a protein. In practice, nearly each peak represents a backbone amide group of an individual residue, with the exception of peaks originating from the HN containing side chains of amino acids Asn, Gln, His, Trp, Lys and Arg. However, except for His, the signals from the sidechain moieties are easily recognisable. Furthermore, proline residues are absent in 2D heteronuclear ^{15}N-HSQC spectra due to the lack of an amide group. The chemical shift values of the various nuclei can be extremely useful when used as a proxy to monitor protein-ligand interactions, a process called chemical shift mapping or chemical shift perturbation (CSP).

When a protein is titrated with a ligand, e.g. with a drug or another biomolecule, the chemical shift of the nitrogen and proton nuclei of the residues that are in close proximity to the binding site will be most affected. Thus, the binding of the ligand results in changes in the chemical shifts of these nuclei, causing the resulting peaks to alter their position in the NMR spectrum. By recording a series of NMR experiments at varying stoichiometries of protein and ligand, the resulting series of spectra conveys information regarding the affinity of the ligand, as well as identifying the important residues involved in the interaction.

To briefly explain the theoretical aspect of this phenomenon, consider for simplicity a peak for a nucleus *i* in a protein *P* at 7.0 δ ppm in a one-dimensional NMR spectrum (Fig. 6.1C). For a simple two-state binding process with ligand *L*:

$$P + L \underset{k_{off}}{\overset{k_{on}}{\rightleftharpoons}} PL$$

Eq. 6-1

the [P]:[L] = 1:0 equivalent condition (Fig. 6.1C, 1:0 eq) represents the left side of equation 1, where the protein is in its unbound state. Upon addition of a high-affinity ligand, under the condition that the k_{off} is small, the equilibrium lies fully towards the bound state. Consequently, addition of the ligand causes the *i* peak in the unbound state to decrease in intensity, whereas a new peak at a different position appears; this new peak represents the bound state of the same protein nucleus *i*. At a [P]:[L] = 1:0.5 stochiometry (Fig. 6.1C, 1:0.5 eq) peaks for both the unbound and bound states will be present with equal intensities (neglecting dynamic effects). The difference between the bound- and unbound peak position is called $\Delta\delta_i$. At a [P]:[L] = 1:1 stochiometry, only the peak of the bound state will be present, as the peak for the unbound state will have disappeared (Fig. 6.1C, 1:1 eq).

In NMR, the above situation when the k_{off} is much smaller than $\Delta\delta_i$ is called the "slow exchange" regime. In contrast, in a "fast exchange" regime, when k_{off} is much greater than $\Delta\delta_i$, in each of the spectra recorded at different [P]:[L] stochiometries the peak position for nucleus *i* represents the population weighted average of its free and bound positions. Consequently, the peak appears to be "moving" from its original position (Fig. 6.1D 1:0 eq) towards its bound position as the ligand concentration increases (Fig. 6.1D, 1:0.5 eq to 1:2 eq). In cases where $k_{off} \approx \Delta\delta_i$, the peak typically disappears due to line broadening effects and this situation is called "intermediate-exchange" (not shown).

The real power of the CSP method lies in the identification of protein residues most affected by the ligand, i.e. those residues with nuclei that display large $\Delta\delta_i$ values. In case of multi-dimensional spectra, such as the ^{15}N-HSQC spectrum (Fig. 6.2A), the total chemical shift change involving all dimensions is usually taken (*vide infra*, equation 2) and used as a proxy for the importance of the specific residue in the interaction. Furthermore, for the fast exchange regime K_d can be determined from the observed $\Delta\delta_i$ values as a function of ligand concentration[19].

6.3.1. Chemical shift perturbation in practice

The first step in a CSP procedure is the assignment of individual peaks to a specific residue in the protein. Different software packages and algorithms have been developed to establish the assignments of the backbone nuclei, i.e. H^N, N, C^α, C', either by manual or automated approaches[152]. According to Lee and Markley, based on BMRB statistics in 2014 Sparky was still the most widely used NMR data analysis tool for backbone assignment[153]. A much more recent programme, not included in their list, is CcpNmr AnalysisAssign version-3[28], which provides tools for simple and semi-automated backbone assignments and a dedicated interactive CSP analysis module, from here on referred to as the CSP module (Fig. 6.2C). An *ab-initio* protein backbone assignment can be a time-limiting factor, as it requires some effort in terms of sample preparation, NMR spectra recording and data analysis. However, assignments can potentially be retrieved from databases, such as the BMRB, and serve as the starting point, adjusting them as required by the user's experimental conditions.

Table 1 displays an overview of functionalities from the various NMR data analysis programmes that are relevant for a CSP analysis. After the initial setup, either Sparky[153], CcpNmr Analysis Version 2[154], NmrView[155] and CARA[156] can automatically calculate the $\Delta\delta$ from chemical shifts. However, only AnalysisAssign version-3 has capabilities for interactive inspection of the $\Delta\delta$ data in relation to the underpinning spectra, interactive thresholding and easy adaptation of various relevant parameters. Moreover, apart from AnalysisAssign version-3 users will have to manually plot and/or export to other software packages for further analysis. This typically will involve multiple manual steps before retrieving the biological relevance from their initial NMR dataset. In addition, parameter adjustments will necessitate repeating these various steps, thus increasing the time required for the whole analysis and inevitably increasing the risk of introducing human errors. The AnalysisAssign version-3 CSP module, contains all required functionalities for a full CSP analysis, including automatically generated bar charts of CSP values as a function of residue number which are linked to the underpinning spectra, peak tables and a binding plot. Additionally, if the protein molecular structure information is available, the annotations and selections can be mirrored to a graphical molecular visualisation, such as PyMOL[29].

I tested the AnalysisAssign CSP module by exploring the binding of Clip2 RNA, AAAUAA, to the Tstar-KH domain[157,158]. I imported both the assignments of the free Tstar-KH domain in addition to a series of five ^{15}N-HSQC spectra of uniformly ^{15}N-labelled Tstar-KH at 0.0, 0.5, 1.0, 1.5, and 2.0 equivalent of Clip2 ligand, directly from the original Sparky data using the inherent data conversion routines of AnalysisAssign. The assignments of the KH domain were propagated from the spectrum at 0.0 equivalent of ligand to all other spectra using the simple drag-and-drop feature of AnalysisAssign for copying peak lists. As expected, some of the peaks had changed their positions upon titration with the ligand. Using the interactive tables and spectrum displays it was possible to easily identify shifted peaks and correct their position in the target spectra, either individually, or as a group of peaks across multiple spectra, or automatically. Fig. 6.2B shows the overlay of the five spectra, with trajectories of shifted peaks indicated for selected residues.

Fig. 6.2C shows the overlay of the five ^{15}N-HSQC spectra at the position of G78. The gradual change in peak position upon increasing ligand concentration is evident. Importantly, to ensure a valid analysis all spectra should be properly referenced as changes in peak positions could otherwise be misinterpreted. Fortunately, AnalysisAssign has routines to establish, and where needed report, on the spectral alignment that functions even in the case of non-fully identical spectra[28]. Using an automated analysis, in which all the peaks were accurately matched to their extrema, CSP_i values are calculated for each Tstar-KH residue i using Eq. 6.2 and displayed automatically as a bar plot in the CSP module interface.

$$CSP_i = \sqrt{(\Delta\delta_{Hi})^2 + \alpha(\Delta\delta_{Ni})^2}$$

Eq. 6-2

Where α denotes the relative weighting of chemical shift changes of the ^{15}N nuclei relative to the ^{1}H nuclei, by convention set to 0.14[159], and $\Delta\delta_{Hi}$ and $\Delta\delta_{Ni}$ denote the observed changes of the proton and nitrogen chemical shifts for residue *i*, respectively. Crucially, the threshold below which the CSP_i values are deemed not significant needs to be established. A value of 1σ derived from the distribution of all CSP values is set by default as the first estimate without a need of other filters[19]. Rapid manual inspection of the affected residues establishes if this threshold needs upward or downward adjustment. In the case of G78, the peak follows a consistent trajectory upon ligand titration with well-defined changes in peak positions, rendering its CSP an appropriate threshold value. In contrast, another residue with a similar CSP value, A138, is located in a crowded region of the spectrum (cf. Fig. 6.2B) and thus its peak movements in such areas are potentially compromised by mis-assignment of the peaks. Such residues should be flagged, barring further NMR data confirming their proper assignment, they should be excluded from the analysis. It is wise to reduce any possible false positives, until the moment that their inclusion appears warranted, e.g. after careful inspection of the structure (*vide infra*) or on the basis of other data that confirms that they can be considered relevant for the binding event. Depending on k_{off} and the residue-specific $\Delta\delta_i$, peaks can disappear in any of the spectra with ligand concentrations > 0. Such situation still conveys useful information about the involvement of the specific residue, albeit that the exact magnitude of its CSP_i value cannot be established. In general, the disappearance of a peak for a specific residue is assumed to imply a $\Delta\delta_i$ (CSP_i) value resulting in exchange broadening due to an intermediate exchange regime and hence signifies importance. Consequently, these residues are automatically flagged by the CSP analysis module, e.g. residue I97 in Tstar-KH.

The true power of the CSP analysis is revealed when mapping the CSP results onto a molecular structure. The CSP module can automatically map the annotations and residue selections onto a molecular structure of the biomolecule under investigation using an external molecular visualisation programme, such as PyMOL[29]. The CSP analysis of the interaction of Clip2 RNA with the Tstar-KH domain identified residues F72, V73, G74, K75, L77, G78, G81, S83 T93, I97, R104, K106, K108, E110, R113, Y120 and L130 with significant CSP values (Fig. 6.2C). When mapped onto the structure of Tstar-KH, a clustering of several affected residues is observed across an interface formed by one α-helix, two ß-strands and two loops, corresponding to the KH hydrophobic groove (Fig. 6.2D). In fact, according to Feracci *et al.*[157] residues G78 and I97, (the latter flagged as "missing" by the CSP analysis), belong to a set of crucial residues in the groove that stabilise the interaction with the Clip2 RNA.

In cases where no 3D structure is available, often the structures of homologous proteins can be used as a reference. Protein structure is more highly conserved than primary sequence; therefore, small changes in protein composition usually do not significantly alter the structure of the macromolecule. Alternatively, in many cases protein structures can be fairly accurately obtained using homology modelling, where an existing structure of a homologous protein is used as a template to generate the structure of the protein of interest. Such models can be reliably used for mapping of protein-ligand interactions, assuming the binding interface has not been affected by the mutations. A description of contemporary available homology modelling software packages and servers are discussed in the review by Vyas *et al*[160].

Currently freely available NMR software suites, such as Sparky[153], allow retrieving CSP values from their tables, but users are required to manually export to third party software to carry on the analysis or are limited to single and static plotting, like in the case of CcpNmr Analysis Version 2. The possibility to graphically and interactively inspect the NMR data that identify the residues involved in a protein-ligand interaction, makes the new AnalysisAssign CSP module extremely useful for non-experts and drastically reduces the time required for a CSP analysis. The CSP module is also not limited to ^{1}H-^{15}N as it can accommodate any combination of nuclei, e.g. ^{1}H-^{13}C in case of methyl residues. AnalysisAssign is implemented in a flexible fashion that will facilitate easy adaptation to insert specific calculation modes for $\Delta\delta_i$ values, automatic pre- and user-defined K_d fittings as well as direct links to external auto-docking software such as HADDOCK[161]. For more advanced users, it is also possible to use the AnalysisAssign libraries to create specific, but simple macros to extrapolate further information from the dataset (Table 1). For example, a macro can be used to plot the minimal shift changes resulting from the mutation of a specific residue in a protein, including in this calculation only residues in which the CSP is above the defined threshold value (see Appendix 6.6). Several settings and data exporters based on the NMR-exchange format (NEF)[162] or tabular .xls format have been implemented, thus providing tools to easily export information to other programs for further analysis if so required.

A series of other programmes have also been developed to address specific tasks using peaks in NMR spectra and NMR assignments, such as the programme Farseer, which performs analyses on large and multivariable datasets, including CSP[163]. Other programmes include auto-FACE, which facilitates the identification of binding mechanisms from CSP data[164] and TITAN, which uses peak perturbations trajectories to help the identification of interaction mechanisms[165].

Features	CcpNmr V3	CcpNmr V2	Sparky	NmrView/CARA
Peak Automation[1]	Simultaneous peak selections, copying and re-fitting	Simultaneous peak selections, copying and re-fitting	Simultaneous peak selections, copying and re-fitting	Single peak re-fitting
Interactivity[2]	Selectable plot items and tables with live updates	None or static plots	None	None
Settings[3]	Multiple dimensionality Multiple Atoms Multiple Δδ calculation modes Several GUI parameters	Limited dimensionality	Limited dimensionality	Limited dimensionality
Extras[4]	Link to Molecular Visualisation IPython Console Macro editor	None	User's macros	User's macros
Exports[5]	Images: Various formats Texts: Various formats Software readible: Json	Text: Various formats	Text: Various formats	Text: Various formats

Table 6.1 NMR software packages.

Comparison of common freely available NMR software packages with built-in backbone assignment (as shown on the BMRB statistics in 2014[153]) and CSP analysis capabilities. [1]Graphical peak selection, copying assignments between spectra, peak adjustment and refitting, provisions to follow peaks across titration series. [2]Live updates of results, interactive adjustment of parameters. [3]Adaptable to different experiment types, ability to handle different dimensionalities, ability to handle different peak parameters. [4]Interaction graphical visualisation tools, ability to link to other software packages. [5]Exports to external formats

6.4. Conclusions

Chemical shift perturbation is very useful as a simple tool to elucidate macromolecular interactions in small and medium-sized complexes, including protein-protein, protein-drugs, and protein-DNA/RNA interactions[19], either in solution, or in solid state NMR[166]. The CSP method works best when recording heteronuclear NMR spectra on samples in which the biomolecule, e.g. the protein, is isotopically labelled as to allow for selective detection. Fortunately, protein overexpression and isotope labelling, e.g. by ^{15}N or ^{13}C, has now become routine in *E.coli* and *Pichia Pastoris*, with new and promising developments for expression in higher eukaryotic systems, which guarantees the presence of a more complex folding machinery and post-translational modifications[167]. Together with new developments in NMR technology, e.g. direct ^{15}N-observation[168], the CSP method will find even more widespread application.

In comparing the various programmes for CSP data analysis, the CSP module of AnalysisAssign version-3 provides for a simple and interactive graphical tool that allows users to significantly reduce the time required for a CSP analysis.

6.5. Figures

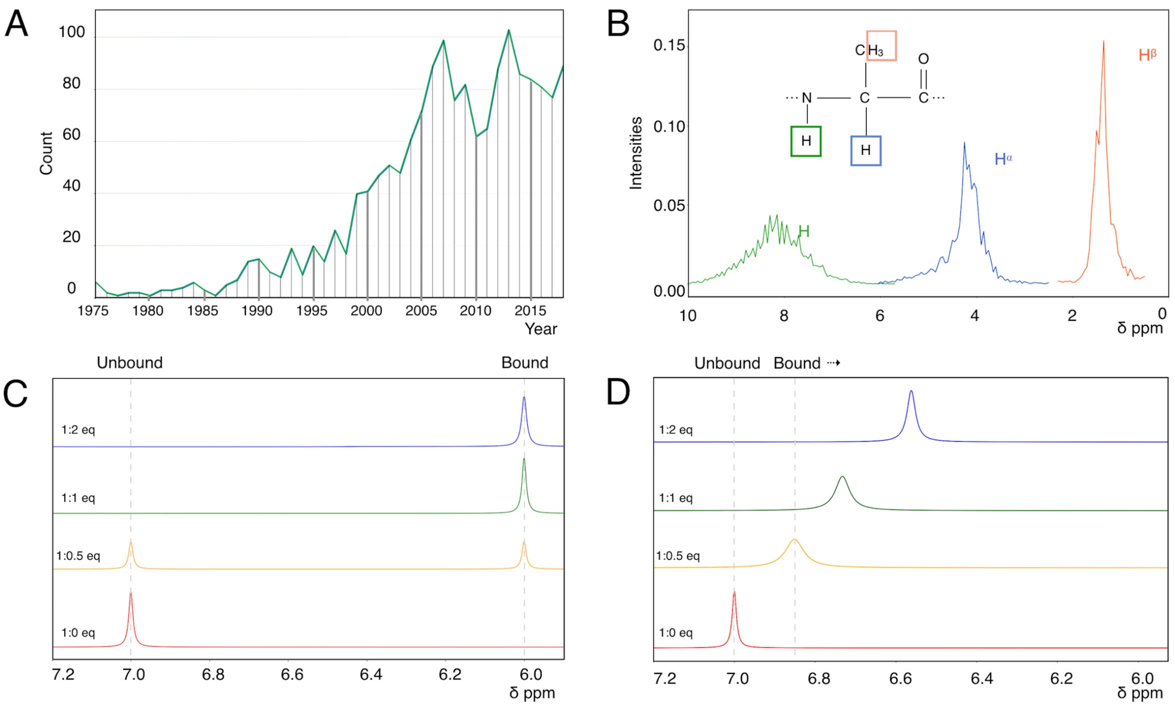
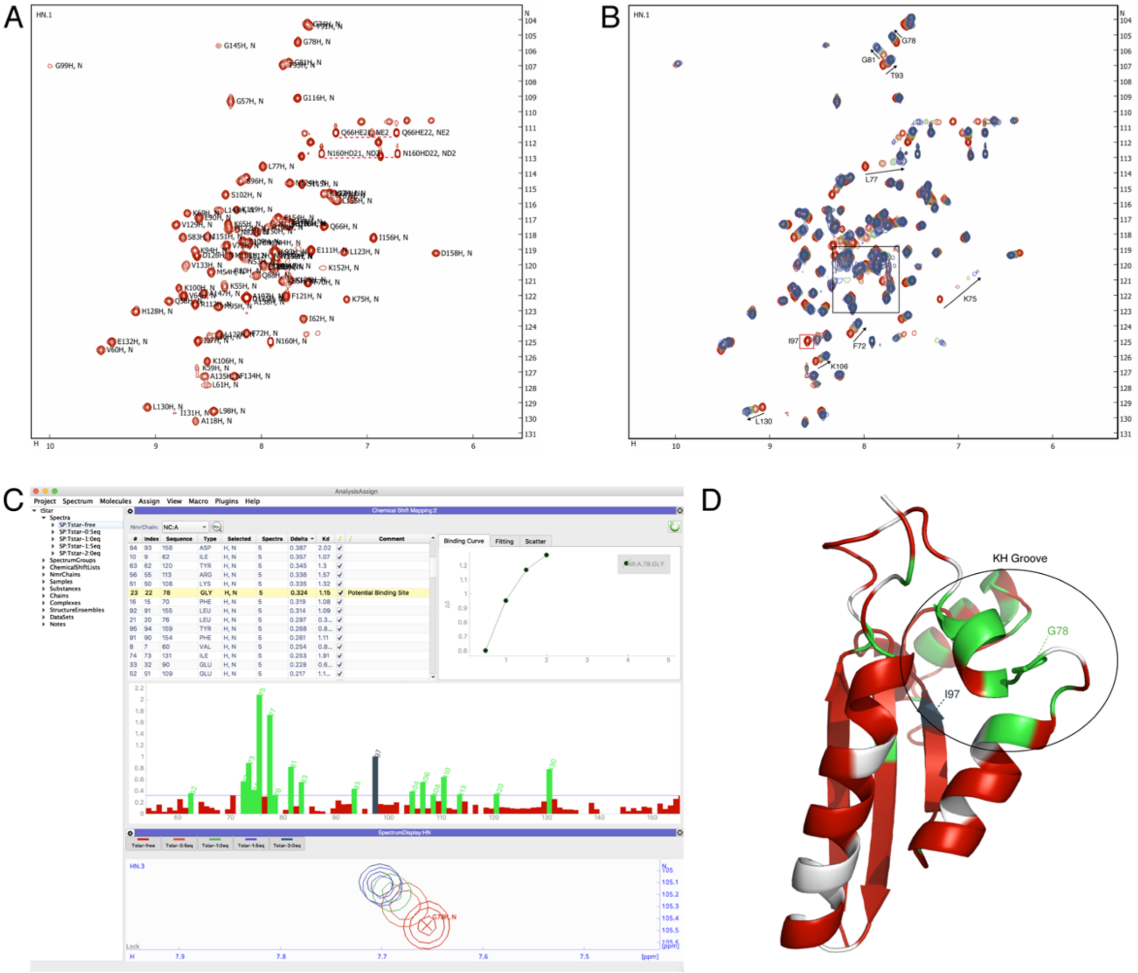


Figure 6.1 Chemical shift and exchange.

A Number of CSP publications as function of year of publication. The plot shows the number of journal articles in the PubMed database by querying for “chemical shift (perturbation or mapping)”. **B** Distribution of deposited chemical shifts for the HN, Hα and Hβ nuclei of Alanine as derived from the Biological Magnetic Resonance Data Bank (BMRB). For historical reasons, the scale in NMR is expressed in relative terms, the so-called ppm scale, which runs from high positive values on the left to low, or negative values, on the right of the scale. **C** Simulated 1D ^{1}H NMR spectra under the slow chemical exchange regime. Spectra are shown at 0.0 (red), 0.5 (orange), 1.0 (green) and 2.0 (blue) equivalent (eq) of (NMR-invisible) ligand. Δδ was assumed to be -1 ppm. **D** Simulated 1D ^{1}H NMR spectra under the fast chemical exchange regime. Spectra are coloured as previously. Peak positions were calculated using equation 6 in reference[19], using Δδ = -1 ppm, [protein] = 100 µM, Kd = 200 µM. Note the gradual shift of the peak as function of ligand concentration. The exchange-induced broadening of peaks at 0.5 equivalent and 1 equivalent are slightly exaggerated for illustrative purposes.



NmrChain: NC:A

#	Index	Sequence	Type	Selected	Spectra	Ddelta	Kd	/	/	Comment
94	93	158	ASP	H, N	5	0.387	2.02	✓		
10	9	62	ILE	H, N	5	0.357	1.07	✓		
63	62	120	TYR	H, N	5	0.345	1.3	✓		
56	55	113	ARG	H, N	5	0.336	1.57	✓		
51	50	108	LYS	H, N	5	0.335	1.32	✓		
23	22	78	GLY	H, N	5	0.324	1.15	✓		Potential Binding Site
16	15	70	PHE	H, N	5	0.319	1.08	✓		
92	91	155	LEU	H, N	5	0.314	1.09	✓		
21	20	76	LEU	H, N	5	0.297	0.3...	✓		
95	94	159	TYR	H, N	5	0.268	0.8...	✓		
91	90	154	PHE	H, N	5	0.261	1.11	✓		
8	7	60	VAL	H, N	5	0.254	0.8...	✓		
74	73	131	ILE	H, N	5	0.253	1.91	✓		
33	32	90	GLU	H, N	5	0.228	0.6...	✓		
52	51	109	GLU	H, N	5	0.217	1.1...	✓		

Figure 6.2 Chemical shift perturbation analysis (CSP) of the binding of Clip2 RNA (AAAUAA) to the Tstar-KH domain.

A 15N-HSQC spectrum of 200 µM uniformly 15N-labelled Tstar KH domain[157]; selected assignments are indicated. **B** Five overlaid spectra of Tstar-KH 15N-HSQC domain. Spectra are shown at 0.0 (red), 0.5 (orange), 1.0 (green), 1.5 (purple) and 2.0 (dark blue) equivalent of Clip2 ligand. The black square box indicates the crowded region. Arrows indicate different peak perturbation trajectories. The I97 peak (red box) is present only at 0.0 equivalent but disappears upon addition of Clip2. **C** The CSP analysis module in AnalysisAssign version-3. Clicking any bar in the bar chart (middle; included residues in green, excluded residues in red) or residue in the residue table (top) will navigate to the peaks of the corresponding residue in the spectra (bottom). The binding curve can be automatically displayed on the right side of the table. Multiple selection overlays related binding curves. All adjustments of parameters and settings of the CSP analysis module, such as setting the threshold line (i.e. horizontal line in middle panel) or excluding a residue from the analysis (check-boxes in the residue table), will result in a real-time update of all plots without the need for any tedious or error-prone manual actions. **D** Ribbon representation of the Tsar-KH domain (PDB code 5EL3) with residues coloured according to their CSP values resulting from interaction of Clip2 RNA. Residues flagged with missing peaks in the spectra containing ligand, e.g. I97, are highlighted in dark blue. The black circle highlights the KH domain clip2 binding groove. Unassigned residues or residues removed from the analysis are indicated in light grey.

6.6. Appendix

6.6.1. Case study: Minimal shift analysis

Binding site identification tools built in Analysis Screen can be applied to several other NMR studies. A practical example is a minimal shift analysis I have performed using the chemical shift mapping module. A collaborator who carried out eight mutations to a wild-type macromolecule (undisclosed), recorded a series of NMR experiments to each individual mutated variant. The collaborator shared with us several ^{15}N-HSQC spectra recorded on a wild type form and after each single residual mutation. However, spectral backbone assignments were not available for this dataset. Luckily a similar complex was deposited in the BMRB database[95]. Therefore, I wrote a custom macro to retrieve the crucial information and transfer backbone atom assignments to the dataset under analysis. After having re-fitted all peaks to their extrema positions using semi-automated routines present in the software, I performed a global alignment of the wild-type spectral signals to all other variants using an in-house alignment algorithm[28]. This allowed for a correction of referencing points among the entire dataset, therefore, reducing the recording of potential false positive shifts. The CSP analysis was performed as a 1:1 between the wild-type and each mutant form, Fig. 6.3. Eventually, delta shifts above one standard deviation threshold were used to calculate the minimal shift analysis summary (Fig. 6.4).

From the analysis reported from the CSM module and the 3D structure representation of the model (Fig. 6.5), our collaborator could confirm previous *in-vitro* assay experiments.

The availability of the chemical shift mapping module allowed for a quick validation of a collaborator biological problem, further demonstrating the power of the CSM as a tool in molecular biology.

6.7. Appendix Figures

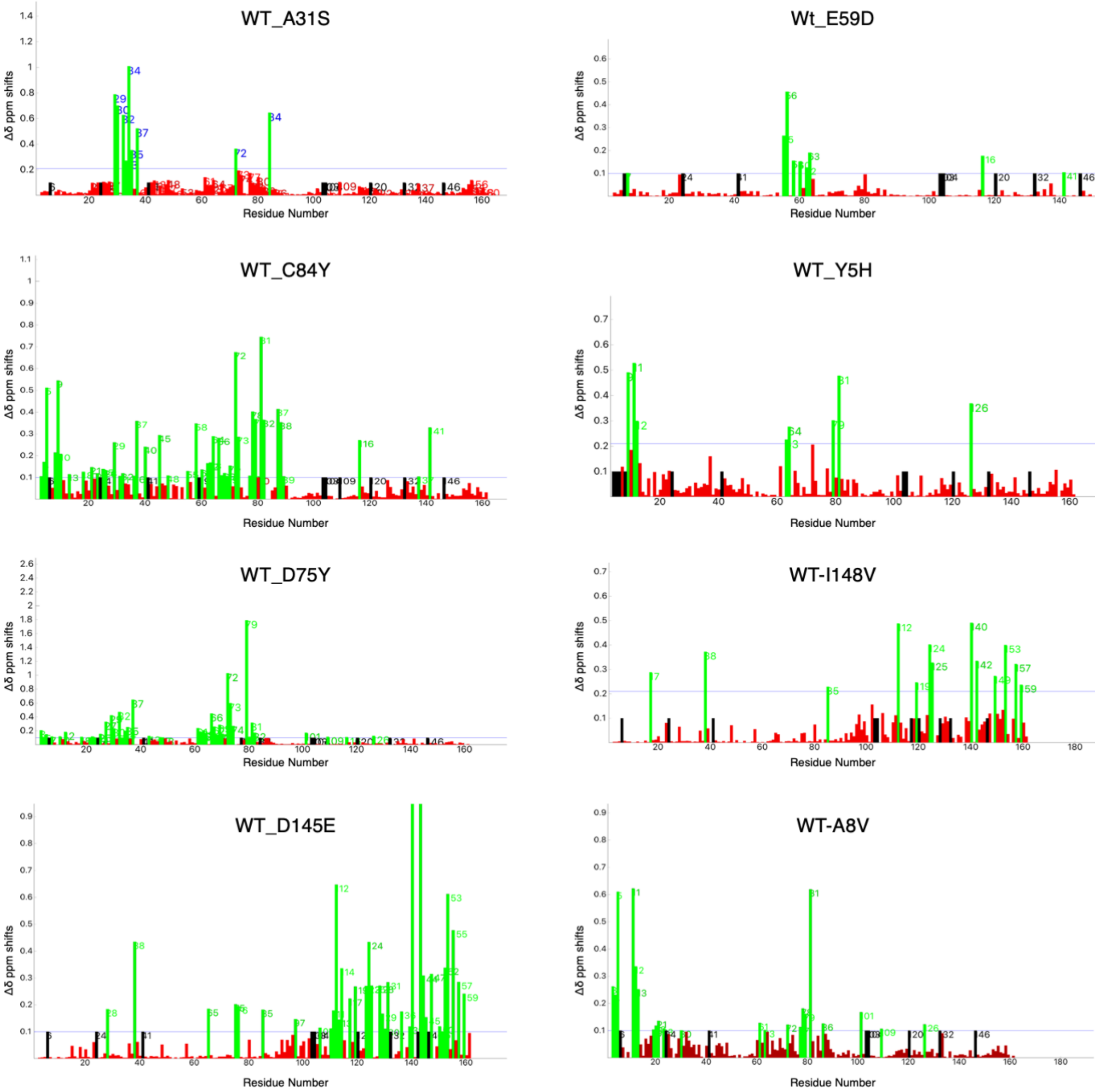


Figure 6.3 Mutation analysis using the Chemical Shift Mapping module.

A series of screenshots of the CSM bar plot for each of the eight single mutations analysis. Green bars represent residues with highest chemical shifts perturbations and above 1σ threshold value (blue horizontal line).

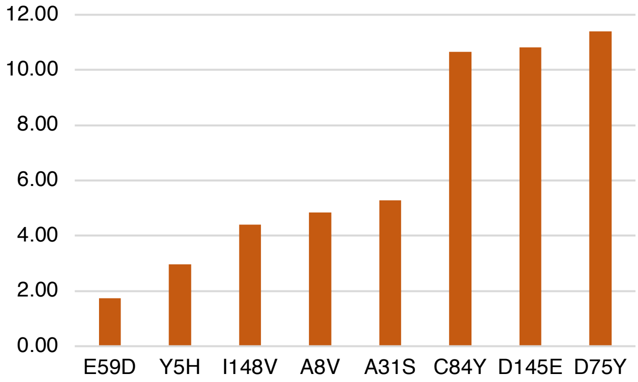


Figure 6.4 Minimal shift analysis using CcpNmr version-3.

The graph summaries the most crucial mutations based on the chemical shift analysis.

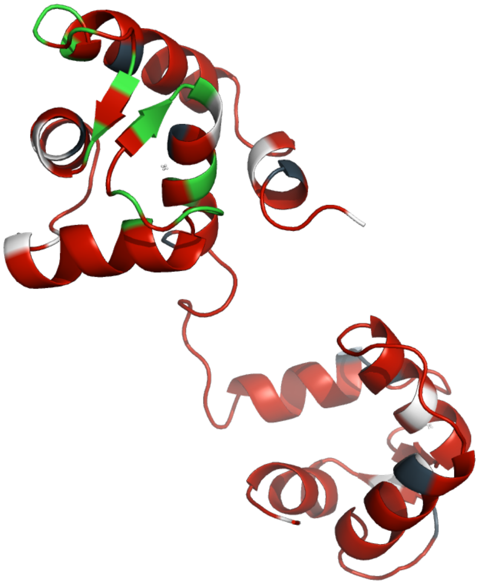


Figure 6.5 3D Ribbon representation of collaborator's macromolecule.

In the figure are represented the regions affected by the mutation D75Y. In green are shown the residues marked by CcpNmr as relevant which correspond to the highest spectral peak perturbations; in dark blue are the residues in which spectral peaks have disappeared after the mutation; in white are excluded residues or unavailable NMR assignments, (PDB code undisclosed).

6.8. Acknowledgements

I thank Drs Cyril Dominguez[a] and Mikael Feracci[a] for the usage of their Tstar-KH/Clip2 data. I also thank Drs Ed Brooksbank[a], Wayne Boucher[a] and Rasmus Fogh[a] for their expert contributions to the CcpNmr Analysis programme. I also thank Manaf Guma[a] for providing the dataset needed for the minimal shift analysis.

6.9. Addresses

[a] Leicester Institute of Structural and Chemical Biology, Department of Molecular and Cell Biology, University of Leicester, Henry Wellcome Building, Lancaster Road, Leicester LE1 7HN, United Kingdom.

Chapter 7

Generating high-affinity leads from fragments using NMR-guided dockings and virtual click-chemistry. Application to the eukaryotic Initiation Translation Factor eIF4G

Keywords: Docking, MD simulations, FBDD, eIF4G, Mnk.

7.1. Abstract

The interaction between eIF4G-Mnk is a key mechanism which leads to elevated eIF4E levels. High concentrations of eIF4E have been observed in many types of cancers, including lymphoma and leukaemia. Previous work at Arthanary's lab identified potential residues that form the protein-protein interaction surface in the eIF4G–MNK complex[169]. Furthermore, molecular dynamics simulations of this complex have shown the presence of binding pockets otherwise not visible in static crystallographic structures, pointing at a possible strategy through a fragment-based drug discovery FBDD approach. FBDD is one of the main methods for the identification of drug-like candidates in the early stages of drug discovery. Once fragment hits have been identified, the next step entails increasing potencies by optimising their binding properties. However, a bottleneck of this technique is a lack of computational automated workflows needed to link, merge and optimise binding fragments in order to generate potential leads. This process is very often performed by manually designing and synthesising new molecular variants and subsequently testing their binding properties, which obviously results in a laborious and time-consuming process.

In this chapter I present a novel *in-silico* workflow which employs a series of computational tools available in the public domain for generating tailored drug-like compounds aimed at disrupting the interaction between eIF4G-Mnk.

7.2. Introduction

Translation regulation controls protein synthesis and translation of specific mRNAs, a process essential to life. Aberrant functioning of these processes can lead directly and indirectly to various forms of tumour activity, such as cell survival, angiogenesis, transformation, invasion and metastases[170,171]. Cap-dependent and IRES-dependent are the two pathways for the initiation of the translation machinery which allows the ribosomal 40S subunit to anchor the mRNA through the eIF1-5 family[172] and code for new protein[173].

The cap-dependent process initiates with the recruitment of eIF proteins 4G, 4E, 4A to the m7GpppN cap at the 3' end of the mRNA. In particular eIF4E has a crucial role starting the machinery by binding directly the RNA[170]. However, eIF4E has to be activated first through the phosphorylation of its Ser209 by the Mnk kinase[174]; furthermore, this kinase can interact with eIF4E only if scaffolded to eIF4G[170].

Not surprisingly, high concentration of eIF4E has been observed in many types of cancers, including lymphoma, leukaemia and many others[175] and has been targeted by several pharmaceutical companies with small molecules[176,177]. It was only in 2014, when Papadopoulos *et al.* at Wagner's lab[a] elucidated the allosteric mechanism of dissociation of the eIF4E/eIF4G complex through crystallographic studies[169]. Nonetheless, an alternative inhibition pathway is under evaluation at the Arthanari lab[a], which involves the inhibition of the protein-protein interaction between eIF4G and Mnk through small molecule inhibitors (Fig. 7.1A). This strategy is believed to minimise side effects and decrease the probability of developing future drug resistance.

A detailed crystal structure of eIF4G that shows the crucial C-terminal region was recently solved (Papadopoulos *et al.*, personal communication). The highly charged C-terminus is believed to play a key role in the interaction with Mnk. An X-ray crystallography structure for a short sequence of Mnk is also available, and this information was used by Dr Gorgulla[a] and co-workers to carry out computational studies on the complex. Long runs of molecular dynamics (MD) simulations (up to 10 μs) were performed with the aim of elucidating any structural changes which could help in understanding potential inhibition mechanisms of the complex and could assist in the design of inhibitors. A qualitative analysis of the MD trajectories indicated the sporadic appearance of transient smaller binding pockets in the proximity of C-terminal region, otherwise not visible in the initial static crystallographic structure (Fig. 7.1B). This region was believed to be an anchor point for the Mnk kinase.

Previous NMR chemical shift perturbation studies (CSP, see also chapter 6) (Arthanari *et al.*, unpublished), identified several crucial residues involved in the eIF4G/Mnk interaction. In particular, the CSP's of Glu1553, Asp1554, Lys1557, Glu1558, Tyr1562, Trp1589, Arg1591, Glu1592, Glu1595, Glu1596 indicated these residues as the most relevant in the binding activity. Conversely, recent CSP studies carried out by Dr. Viennat[a] included the assessments of the Mnk peptide properties upon binding to the eIF4G. The analysis showed relevant effects for residues Lys7, Arg8, Arg9, Lys10, Lys11, Lys12 suggesting the relevance of these residues for complex formation (Figs 7.1C-D).

The absence of a well-established binding pocket for eIF4G resulted in a challenging search for possible inhibitors. Although the MD simulations yielded identification of potential transient binding pockets, a traditional in-silico high-throughput screening cannot guarantee the detection of suitable molecules. Instead, a more beneficial approach would be the usage of a fragment-based drug design protocol that accommodates these transient pockets.

Thus, I designed an in-silico workflow to tackle this issue, outlined in Fig. 7.2A. The protocol starts with the validation of binding sites between Mnk and eIF4G using the HADDOCK[178] and DeepSite[179] algorithms. Subsequently, two parallel sets of docking approaches are performed to principal pockets using small fragment libraries. Top scored compounds by the docking algorithms are then filtered with custom-designed filters based on interaction types. In the following steps, fragments are linked using virtual click-chemistry algorithms. This procedure will create a new library of small molecules that can be re-docked *in-silico*, covering a larger interaction surface of the protein target. Finally, the scored compounds by the latter algorithms and filters, will be further optimised and validated using MD simulations to achieve a higher affinity binder (Figs 7.2A-B).

Compared to the classical *in-silico* approaches, the workflow as outlined above has the advantage of automatically creating highly tailored drug-like compounds starting from readily available fragments, using the advantage of combinatorial multiplication. The workflow will generate and evaluate thousands of compounds and effectively filter out unsuitable molecules. Eventually, only a limited number of molecules will be validated using more advanced and time-consuming computational methodologies, such as MD simulations. Finally, knowledge derived from well-validated candidates, including suggested chemical reactions needed for compound synthesis, can be transferred to the wet-lab for further binding assays, therefore drastically reducing the time and resources associated with a traditional fragment-based drug design pathway.

7.3. Materials and Methods

7.3.1. Materials

Macromolecules

The 3D structure coordinates of eIF4G was obtained by extracting a MD simulation frame from a previous study carried out by Dr. C. Gorgulla[a]. The frame was extracted from the system following an accurate visual inspection of all frames and simulation behaviour. The 3D structures of eIF4G and Mnk were previously solved by Arthanari and co-workers using X-ray crystallography; a PDB code was not available at the time of this work.
FASTA sequence for eIF4G protein is defined as following:

^{1440}*PSEELNRQLEKLLKEGSSNQRVFDWIEANLSEQQIVSNTLV*
RALMTAVCYSAIIFETPLRVDVAVLKARAKLLQKYLCDEQKELQ
ALYALQALVVTLEQPPNLLRMFFDALYDEDVVKEDAFYSWESS
KDPAEQQGKGVALKSVTAFFKWLREAEEESD^{1596}

FASTA sequence for Mnk peptide is defined as following:

^{2}*KRRKKKRKTRAT*^{13}

Small molecules

Several libraries of small molecules were used throughout this work. The first set included a library of approximately 9000 molecules with a molecular weight (MW) up to 250 and was used for the primary screening on site-1 and site-2; an additional library of molecules with an MW up to 150 was used as linkers. Both libraries were available in the release of the AutoGrow package[82]. It was composed by a variety of molecules presenting the following functional groups: acid anhydrides, acyl halides, alcohols, thiols, alkenes, alkynes, amines, azides, carbonochloridates, carboxylates, epoxides, esters, halides, isocyanates, isothiocyanates, sulfonylazides, thio-acids.

A further library was downloaded from the database Zinc15[180] as PDBQT files using a specifically written Python script. This library comprised of approximately 600000 compounds at an MW up to 200, LogP (logarithm of the partition coefficient) in a range of 0-1. The purpose of this library was to cover higher chemical spaces in the process of the fragment optimisations.

7.3.2. **Methods**

Protein-Protein interactions

HADDOCK 2.2 webserver[161,178] "expert mode" was used to predict the potential binding poses of Mnk to the eIF4G structure and validate the residues of interest that were previously established by NMR experiments. In particular, the active residues selected were Glu1553, Asp1554, Lys1557, Glu1558, whereas the passive residues were not indicated. Haddock itself estimated passive residues within a radius of 6.5Å from the active residues. The Mnk active residues were: Lys2, Arg3, Arg4, Lys5, Lys6, Lys7 in the PDB nomenclature (residue Lys7, Arg8, Arg9, Lys10, Lys11, Lys12 in previous NMR CSP analysis enumeration). Furthermore, the C-termini of both Mnk (the remaining of the chain was manually removed from the PDB file) and eIF4G (Tyr1562, Trp1589, Arg1591, Glu1592, Glu1594, Glu1595, Glu1596) were excluded. Ten runs with a cluster size of four were performed using the clustering method *Fraction of Common Contacts (FCC*). The RMSD cut-off for clustering was 0.60 Å. The remaining parameters were retained at the default values.

Haddock scoring

The HADDOCK scoring function[161] is the sum of several energies and buried surface areas. The score is calculated as following:

$$HADDOCK_{score} = E_{vdW} + 0.2E_{Elec} + 0.1E_{AIR} + E_{Desolv}$$

Eq. 7-1

Where E_{vdW} is the van der Waals intermolecular energy; E_{Elec} is electrostatic intermolecular energy; E_{AIR} is the distance restraints energy (only unambiguous and AIR (ambig) restraints); E_{Desolv} is the desolvation energy[161].

DeepSite

DeepSite is a protein-binding site predictor which uses a 3D-convolutional neural networks approach. According to Jiménez *et at.*, it is powered by a training of 7622 protein studies[179] therefore capable of detecting binding pockets with a high accuracy.

Docking preparation

Docking ROI cubic grid coordinates were calculated using the AutoDockTools 1.5.6[142] software. This information was then used to create Json files as input data for running the AutoGrow routines. ROI grids are represented with the labels “size x, y, z and center x, y, z” in the json file attached below. Ligand and receptor were prepared with the available Python scripts included in the package.

Coordinates and other parameters for site-1, site-2 and linked site-1 + site-2 are listed below in the same json formatting style:

Site-1:

```
        {
        "size_x"                        : 12,
        "size_y"                         : 13,
        "size_z"                         : 14,
        "center_x"                        : 21,
        "center_y"                        : -10.639,
        "center_z"                        : -24.639,
        "num_processors"                      : 8,
        "exhaustiveness"                     : 10,
        "directory_of_fragments"                 : "~/fragments/MW_150/",
        "filename_of_receptor"                  : "~/site-1/4G_clean.pdb",
        "use_strict_lipinski_filter"             : true,
        "max_MW"                                                    : 700,
        "maintain_core"                      : true,
        "additional_autoclickchem_parameters"     : "",
        "minimum_core_atoms_required"            : 0,
        "num_generations"                      : 2,
        "scoring_function"                     : "VINA",
        "allow_modification_without_frag_addition": true,
        "use_lipinski_filter"                  : true,
        "score_by_ligand_efficiency"            : true,
        "top_ones_to_advance_to_next_generation"  : 1000,
        "number_of_crossovers_first_generation"   : 0,
        "number_of_crossovers"                  : 0,
        "max_seconds_per_generation"             : 100000000,
        "number_of_mutants_first_generation"      : 0,
        "number_of_mutants"                     : 0,
    "use_ghose_filter"                     : true
         ...}
```

Site-2.

The only difference compared to site-1 parameters was a smaller grid space and a translated centre of the grid.

```
{       "size_x"                         : 12,
        "size_y"                         : 12,
        "size_z"                    : 10,
        "center_x"                   : 24.917,
        "center_y"                   : -15.306,
        "center_z"                   : -23.167,
…}
```

Dockings were carried out using the python core routine present in AutoGrow. The docking engines were VINA[181], and NeuralNetwork[143] (NN2).

Linked fragments site:

```
{       "size_x"                    : 14,
        "size_y"                    : 16,
        "size_z"                    : 16,
        "center_x"                   : 21.583,
        "center_y"                   : -12.611,
        "center_z"                   : -22.667,
        "num_processors"                  : 8,
        "exhaustiveness"                 : 20,
        "directory_of_fragments"               :
        "use_strict_lipinski_filter"           : false,
        "max_MW"                     : 700,
        "maintain_core"                  : true,
        "additional_autoclickchem_parameters"    : "",
        "minimum_core_atoms_required"          : 4,
        "num_generations"                 : 4,
        "scoring_function"                 : "VINA",
        "allow_modification_without_frag_addition": false,
        "use_lipinski_filter"               : true,
        "score_by_ligand_efficiency"            : true,
        "top_ones_to_advance_to_next_generation"  : 100,
        "number_of_crossovers_first_generation"   : 0,
        "number_of_crossovers"               : 0,
        "max_seconds_per_generation"            : 100000000,
        "number_of_mutants_first_generation"     : 4000,
        "number_of_mutants"                 : 4000,
        "use_ghose_filter"                 : false
}
```

File conversions

The program Babel[30,31] was primarily used for file conversions. Numerous conversions were required among PDB, PDBQT, MOL2 and SMILES files and several custom Python scripts were written for each workflow stages as needed.

Interactions

Interactions between eIF4G and ligands were automatically calculated using a command line programme called H-Bind[182] driven by custom Python scripts. H-Bind was compiled on an older OSX platform (10.12) and then copied over to a newer, and more powerful computer running OSX 10.14, as several compiling issues could not be solved for the latter platform.

Custom AutoGrow

AutoGrow is a command line application able to perform click-chemistry reactions starting from two separate libraries of compounds[82,183]. A genetic algorithm enables to perform several generations of so-called mutations and crossovers. A mutation consists of a modification of the core compound by adding or replacing an active group. Whereas, a crossover involves the creation of a new molecule by mixing two initial structures. By selecting the number of generations, the algorithm keeps randomly increasing initial molecular structures and eventually performs virtual dockings; top scored molecules and molecules that passed the druglikeness filter, will then progress to an extra generation[82].

This strategy has two major drawbacks, the first is the impossibility of controlling input fragments during the generations and the second, a very high probability in generating infinite large final compounds, with MW exceeded 1000 Da which would present unrealistic drug candidate molecules.

However, by inspecting the open-source core code, I could apply several crucial modifications that allowed me to customise a specific workflow for my needs.

The modifications included the creation of a Json file reader for inputting the required parameters in a more robust and practical way, such that additional input parameters, such as a maximum MW filter, could be included. Thus, I could direct the program to cap the molecular growth at a determined threshold.

The program was also split into multiple core routines, so that single actions could be performed at a particular time, e.g. only an optimisation step or a re-docking calculation using a different engine. Furthermore, new additional required features were added, for example, limiting the algorithm to create only variants of input molecules without adding additional fragments. Lastly, several limiting programmatic errors were identified, which were promptly resolved.

Other Scripts

Compounds filters and file-parsing routines were written in Python 2.7 using libraries in a custom Anaconda environment, which included Pandas[32], Bio-Pandas[184] and Babel[30].

Docking Scoring

Two scoring methods have been used for docking the ligands to eIF4G. The first scoring method, VINA[181], uses a combination of knowledge-based and empirical measurements. The second scoring method is the so-called NNScore 2.0[143]; it is calculated using a combination of VINA parameters and a novel algorithm from Durrant *et al.,* called BINANA[185]. This algorithm also provides estimated K_d values for each docked compound. However, these were not used to assess results as they varied widely across similar $\Delta G_{binding}$, and therefore they were deemed meaningless for these studies. Both the VINA and NNscore methods were accessible from the original AutoGrow distribution.

Graphical analysis

The PyMOL package[29] was used to visualise and inspect docking results. A custom macro was written to facilitate the inspection of the large quantities of molecules. CcpNmr Chembuild[28] was used to draw small molecules for the final figures. IUPAC names were derived from SMILES using the software package ChemDraw 19.0[186].

MD simulations

A molecular dynamics simulation of the to target eIF4G and ligand 423437 (SMILES: *NC(=O)NC(=O)/C=C/C(=O)OCc1cn(nn1)C(=O)c2c(F)c(F)ccc2F*) complex was performed at the LISCB computational facilities[b], using 1 GPUs (GeForce GTX 1080 Ti) and 1 CPUs. The MD trajectory of 20 ns was calculated using the Desmond Multisim engine available in the Schrodinger Maestro 2019 free package and results were analysed using the GUI available within the same software suite[79].

The ligand and target were contained in a buffer with a counter Ion/Salt of 8 Na^+ with a total charge of +8 and a concentration of 16.23 mM. Water was then used as solvent and included in a cubic-shaped box of size 10 Å (TIP3P) for a total volume of 309354 $Å^3$. The system containing 29491 atoms was then minimised using the default protocol available in the Maestro suite before the simulation. The full configuration is described in tables 7.1 and 7.2, (appendix, section 7.7). Finally the forcefield used was the OPLS_2005[187]. The full log with an exhaustive description of configuration parameters used in the simulation is displayed in the table 7.3, (appendix, section 7.7).

MD Scoring

The Root Mean Square Deviation (RMSD) is commonly referred to establish the average changes in atom coordinates between simulation frames relative to the reference for a given target. RMSDs were calculated as following:

$$\mathrm{RMSD}_x = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(r_i'(t_x) - r_i(t_{ref}))^2}$$

Eq. 7-2

where *N* is the number of atoms; *t* is the time, t_{ref} is the time at the first frame; *r'* correspond to the atom coordinates under evaluation for the frame *x* after superimposing to the reference[79].

The Ligand Root Mean Square Fluctuation (L-RMSF) is used for describing variations in atom coordinates. RMSFs for the ligand were calculated as:

$$\mathrm{RMSF}_i = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(r_i'(t) - r_i(t_{ref}))^2}$$

Eq. 7-3

Where T is the time, T_{ref} is the ref time corresponding to the first frame, r is the position of the ligand atom for a given frame[79].

7.4. Results and Discussion

7.4.1. **4G-Mnk interactions**

In order to validate the potential binding sites of Mnk to eIF4G, a series of HADDOCK calculations were performed using the HADDOCK webserver[178]. Ten clusters were evaluated and scored using the HADDOCK score (Eq. 7.1), the electrostatic energy, Van der Waals interactions and plotted against the *i*-RMSD (Figs 7.3A-C). From the analysis, the clusters 2,3,4,6 appeared to have the best scoring in all three representations. From the calculated poses of Mnk, it was possible to speculate about the relevance of two binding pockets on the eIF4G surface, which are subsequently denoted as site-1 and site-2 (Figs 7.4A-B).

Binding pockets were also established using the DeepSite algorithm[188] (Fig. 7.4C). Obviously, the algorithm suggested more than two pockets, but importantly site-1 and site-2 previously identified on the basis of the NMR studies and MD simulations were also confirmed by DeepSite.

7.4.2. **Dockings**

Site-1 and site-2 were targeted with two identical libraries of 9000 fragments, selected to cover a variety of chemical moieties. Dockings were performed using the VINA[142] and NN2[143] scoring algorithms (Fig. 7.4D). The two methods were performed in parallel for each algorithm to minimise possible false positives and false negatives. In fact, each method's top-ranked 1000 compounds showed only partial similarity, with some of VINA's top-scoring compounds presenting a much lower score for NN2 and *vice versa*. Ultimately, the unique, highest scoring compounds per-site were filtered using a python script for both methods. Interestingly, compounds for site-1 were scored with higher values for both methods compared to site-2, with an average ΔG of -4.5 for the VINA score and

~4.4 for the NN2 score, suggesting a better defined hydrophobic groove for the site-1 compared to site-2 (Figs 7.5A-B).

Hydrogen interactions and salt-bridges between docked compounds and the target were then analysed using the H-Bind method for each pocket. The results for both the VINA and NN2 selected compounds showed that residues Gln1513, Glu1522, Lys1557 were most crucial for site-1 and residues Tyr1551, Asp1554, Glu1558 for site-2. In contrast, residue Asp1559 for site-2 showed varying results for compounds selected by the two algorithms (Figs 7.5C-D).

Following a visual inspection, the two most buried residues for each pocket were selected as a reference point for a further filtering of the compounds, in particular Asp1554, Lys1557 for site-1 and Tyr1551, Asp1554, for site-2 (Fig 7.6). Only ligands that showed at least two interactions to the buried residues were selected and prepared for the linking step (Figs 7.7A-B).

Next, a new series of scripts were used for classifying and tagging fragment atoms involved in contacts to the target (Figs 7 C-D). The tag was crucial for the following linkage of fragments. In practice, the tag instructed AutoGrow in performing any click-chemistry reactions using any atoms as input, except for flagged entities.

This approach had the benefit of maintaining the fragment core structure intact after the reaction was performed. By measuring the distance between some docked compounds, it was obvious that any of the two randomly selected fragments most-often could not be linked directly while preserving crucial target contacts (Fig. 7.8A). Therefore, the usage of linkers was deemed necessary.

Two sets of linker libraries were considered, one consisting of “shards” with MW up to 150 Dalton and one of fragments with MW up to 200 Dalton (Fig. 7.8B). By manually posing representative examples within the gap between site-1 and site-2, I eventually decided to use the library of 200 Dalton, although it might seem slightly bulkier than the available space. Using a larger linker allowed for minimisation of the inevitable atom-loss per fragment, therefore preserving the necessary molecule length needed for reaching the two binding sites. The first reaction joined linkers to site-1 fragments; newly formed molecules were connected to site-2 fragments by a following set of click-chemistry reactions.

The new library was successfully created using a random selection of available reactions and compounds from the site-1 and site-2 core fragments. The 3793 compounds were re-docked as done previously using either the VINA or NN2 scoring functions, but this time, a much larger target area was explored in the docking simulations.

Surprisingly, at this stage the two algorithms diverged significantly in scoring the ligand poses. The VINA-derived ΔG values for the newly linked compound improved considerably compared to the previous, single docked compounds. For the linked fragments, the average and median values were in a range of -5.20 kcal/mol and outliers up to -7 kcal/mol (Fig. 9A). In contrast, scorings reported by the NN2 algorithm were similar to the single docked fragments (Fig. 9B). These NN2 results appeared very dubious; in fact, a visual inspection of the newly generated poses revealed that nearly all top scored compounds presented unrealistic docking poses, almost as if they were simulating a single smaller compound docking pose. I speculate this could be explained by the training set of the Neural Network, which most likely consisted of only small molecules with MW up to 300 Dalton, and thus failed to score correctly larger molecules. Consequently, the NN2 scoring function was not used for following stages. Examples of docked compounds using VINA and NN2 are shown in Figs 7.9C-D.

After a careful manual inspection of the first 1000 VINA top-ranked compounds, 32 were selected based on their general poses. In fact, these compounds presented fragment-derived cores with similar orientations as previously observed when the original compounds were docked to their respective sites, therefore maintaining the initial best promising poses. I started subsequent analyses from compound 423437 (Fig. 7.10A), which was then prepared for further optimisation phases and molecular dynamics simulations.

From the first inspection, this compound passed all the drug-like properties, including the Lipinski rule of five, which were included in the original AutoGrow framework (Fig. 7.10B).

The compound was initially derived by the halide ZINC04290163, initially scored with a ΔG of -4.3 kcal/mol (Fig. 7.10C **1**), followed by the conversion to the azide 415441 (Fig. 7.10C **2**). The new compound was afterwards linked to the site-2 fragment, ZINC01686679, initially scored with a ΔG of -4.3 kcal/mol (Fig. 7.10C **1b**). Following an alkyne to azide reaction, the two fragments generated the final triazole compound 423437, with a final ΔG of -6.3 kcal/mol (Fig. 7.10C **3**).

I used this compound as a starting point for further improving the binding energies. Thus, I modified the AutoGrow algorithms in a way to alter the compound without linking further fragments to the main core. By only consenting click-chemistry reactions to modify existing ligand moieties, it was possible to generate a new library of variants, which were subsequently re-docked using the VINA scorings. An example of improved compound is 3018740 (Fig. 7.10C **4**); the conversion of the amine to an azide allowed to increase the binding affinity to a ΔG of 6.6 kcal/mol.

Unfortunately, the optimization did not generate a large library of variants for further analysis; even using a much larger Zinc15-derived library, the algorithm kept re-creating the same compounds, which were flagged as duplicates and promptly deleted by the algorithm. This could have been caused for several reasons, such as a limited number of atoms available for performing reactions, a limited amount of reactions included in the original algorithm, a programmatic error, for example, in randomly selecting the same reactions and compounds, or most probably, by a combination of both.

To further explore possible chemical conformations for the compound 423437, I performed a run of crossovers among the first run of optimisations. The crossovers enabled to overlay variants previously generated and exchange moieties across the two molecules, maintaining the common core structure. Surprisingly, all the new generated compounds revealed a much lower energy binding than the original non optimised 423437. The compound 34108840 is an example with a ΔG of -5 kcal/mol (Fig. 7.10C **5**).

Eventually, all compounds generated from the two methods, were ranked by fingerprint similarities and compared to the fragment prior optimisation using the Tanimoto coefficient and the binding energy ΔG (Fig. 7.11). The 3018740 was among the highest in terms of similarities and yet presented an improved binding energy compared to 423437. Whereas compounds derived from crossovers were clustered in the lowest region of binding energies and showed the lowest Tanimoto coefficient. Surprisingly, the average ΔG decreased proportionally to the amount of diversities recorded to the original molecule, suggesting that a large degree of modification to the molecule could impact negatively the binding activity to a point in which the use of original fragments was irrelevant for the final lead-like drug, as seen in the three case studies reported in chapter 1.

By inspecting the docking poses for the three compounds 423437, 3018740 and 34108840, they all presented similar spatial conformation, with the trifluoro benzoyl ring nicely interacting with the site-1 groove and the amine/azide group pointing at the site-2. As expected, the reaction preserved the crucial atoms identified as interacting with the target, maintaining intact original cores as needed for hitting the respective sites. Although, the optimised versions 3018740 and 34108840 amide groups presented a slight torsion angle. This conformation adopted by the molecule, opens a new hypothesis on the presence of a potential binding site in the proximity of Glu1558. This site could be called site-3 (Fig. 7.12C-D) and it might be further explored in future studies.

7.4.3. **Molecular Dynamic simulations**

In order to validate the binding energies and establish the interaction mechanisms for the compound 423437, I performed a 20 ns molecular dynamic simulation. Firstly, the RMSD of both for the ligand and the protein was assessed over the trajectory (Fig. 7.13A). Analysing the evolution of the ligand in the complex is fundamental for detecting the stability of the ligand within its binding pockets. The analysis shows that ligand RMSD values were in a range between 0.5 Å and 3 Å, well below the protein RMSD values, suggesting that the ligand was stable with respect to the protein's binding site during at least the first third of the simulation . From the protein-side, a steady increment of RMSD values was observed in the final quarter of the simulation. The protein RMSD values appeared not fully converged in a such a short simulation, I speculate due to potential structural changes that might start to occur only after 15 ns; however, although these were not of interest in the current project, they could have affected negatively the general binding properties to the ligand.

More importantly, an in-dept analysis of the residues involved and the frequency of their contacts showed Tyr1551 and Asp1554 to be most often interacting with the ligand. These residues belong to the site-2 and site-1. Notably, during the simulation, the ligand failed to interact with the Tyr1551, indicating a temporary detachment of the ligand from the binding site. Interesting, the results suggested that compound 423437 potentially could reach a third site, site-3, in proximity of Glu1558. The latter was flagged as the third most interacting residue (Fig. 7.13B). This finding also is consistent with the NMR CSP results and the results from the INDEEP pocket analysis (cf. Fig. 4.C). The analysis revealed that Tyr1551 formed H-bonds, hydrophobic and water bridges with compound 423437, whereas Asp1554 mainly formed H-bonds, with some ionic character. Val1556 in site-2 revealed just H-bonds, whereas Glu1558 and Asp1559 were indirectly interacting with compound 423437 through water bridges (Fig. 7.13C).

The time-dependent analysis of the complex revealed that compound 423437 displayed constant contacts to both the site-1 and site-2 binding pockets, suggesting it was a good candidate for inhibiting the eIF4G-Mnk interaction. However, the inherent flexibility of the protein resulted in only sporadic binding of the trifluoro benzoyl ring to site-1. This could happen due to the nature of the site-1 pocket, being relatively small in size compounded by a conformational change of the protein which could result in its partial closure (Fig. 7.14). Also, the analysis of single atoms through the Ligand Root Mean Square Fluctuation (L-RMSF) confirmed the atoms 20 to 28 (in the Desmond enumeration), corresponding to the benzene core, were the most unstable, followed by 6 and 7 (Fig. 7.13D).

I next explored the conformational freedom of compound 423437 resulting from crucial multiple rotational bonds (Fig 7.15). This analysis suggested only a few highly flexible sections of the ligand. In particular, the bonds between the ester and the triazole ring proved highly flexible during the entire simulation. This information is of a crucial relevance and might give insight for fully automated molecular optimisation phases.

Finally, the Solvent Accessible Surface Area (SASA) was analysed for the full simulation (Fig. 7.16d). The SASA values indicated low accessibility during the first 2.5 ns of the simulation, corresponding to a tight association of the ligand involving all sites. This was followed by a steady increase of the SASA values indicating a progressive detachment of the benzoyl core from the pockets to a steady detached state with only sporadic final contacts, indicating an unbinding event in the process.

7.5. Conclusions

In this work, I explored the generation of inhibitors to disrupt the formation of the scaffolding complex between Mnk and eIF4G, a strategy to target its involvement in oncogenesis. Previous studies at the Arthanari and Wagner labs[a] characterised the eIF4G-Mnk binding mode. Several crucial residues were identified by NMR studies using the ^{15}N-HSQC chemical shift mapping technique (Figs 7.1C-D). Furthermore, a 3D crystallographic structure for eIF4G and a short peptide for Mnk were determined, which allowed for computationally assessing the dynamic of interaction between Mnk and eIF4G (Fig. 7.7.1B). A qualitative analysis of an eIF4G molecular dynamics trajectory revealed the formation of several potential pockets in proximity of the known Mnk binding region. These pockets served as a starting point for designing a virtual FBDD workflow aimed to identify ligands needed for suppressing the formation of the complex (Fig. 7.2).

Simulations using the HADDOCK[178] server and the results of the DeepSite[179] binding site pocket predictor validated two small potential binding pockets, called site-1 and site-2 (Figs 7.4A-C). Using a library of small compounds with MW up to 250 Dalton, comprising of a large diversity of chemical groups, a customised version of the programme AutoGrow[82] was used in two independent virtual docking to these two sites, using two different scoring methods, VINA and NN2. The VINA scoring presents a well validated and robust algorithm, whereas the NN2 method guaranteed faster outputs. Furthermore, I believe this dual strategy could reduce false negatives otherwise resulting from using one single algorithm only. From the initial 9000 compounds, the top 1000 entities obtained from each algorithm were maintained and proceed to the next filter.

An analysis of hydrogen interactions and salt-bridges between the docked compounds and the most buried residues for each pocket, identified Gln1513, Glu1522, Lys1557 for site-1 and residues Tyr1551, Asp1554, Glu1558 for site-2 as crucial interacting residues (Figs 7.5C-D). Using an in-silico linkage molecule library, customised routines present in the AutoGrow package[82] were used to create new molecules by linking fragments from site-1 and site-2. The newly generated linked compounds that passed the druglikeness filters were re-docked to eIF4G. As expected, the binding scores for this new library of compounds were much improved compared to the single fragment energies (Fig. 7.10C) Of this library, compound 423437 was selected as an example for further assessments and optimisations.

Using a larger library of fragments fetched from the Zinc15 database[180], I performed several growing and crossover modifications of compound 423437 using the click-chemistry routines present in the AutoGrow package. However, further modifications on the initial structure did not result in a significant improvement in the binding affinity. Interestingly, more diverse variants resulted in molecules displaying weaker binding (Figs 7.10-7.11).

Finally, to explore the dynamic aspects of the complex between compound 423437 and eIF4G, I performed a 20 ns MD simulation. The MD trajectory showed that the ligand maintained an overall close contact to the macromolecule in the first third of the simulation. However, at least one highly rotatable bond of compound 423437 resulted in a partial binding of the molecule throughout the 20 ns trajectory. In particular, the amide portion of compound 423437 strongly interacted to residues Asp1554 and Tyr1551 of site-1, in line with the previous docking results. However, the aromatic trifluoro benzoyl ring interacted with the site-2 pocket only during the first part of the simulation before floating freely around the solvent space.

In conclusion, albeit compound 423437 was scored by VINA as one of the best interacting to the eIF4G protein, the analysis of its 20 ns MD trajectory indicated that the compound is subject to a large rotational conformational freedom. Consequently, it does not fully bind the protein for the whole simulation and therefore, it cannot be considered being the final compound at this stage but rather a candidate for further optimisation.

Ideally, any future workflows would also be further automated in order to reduce manual intervention, thus decreasing potential human errors and bias, as well as reducing overall analysis and development time.

7.6. Figures

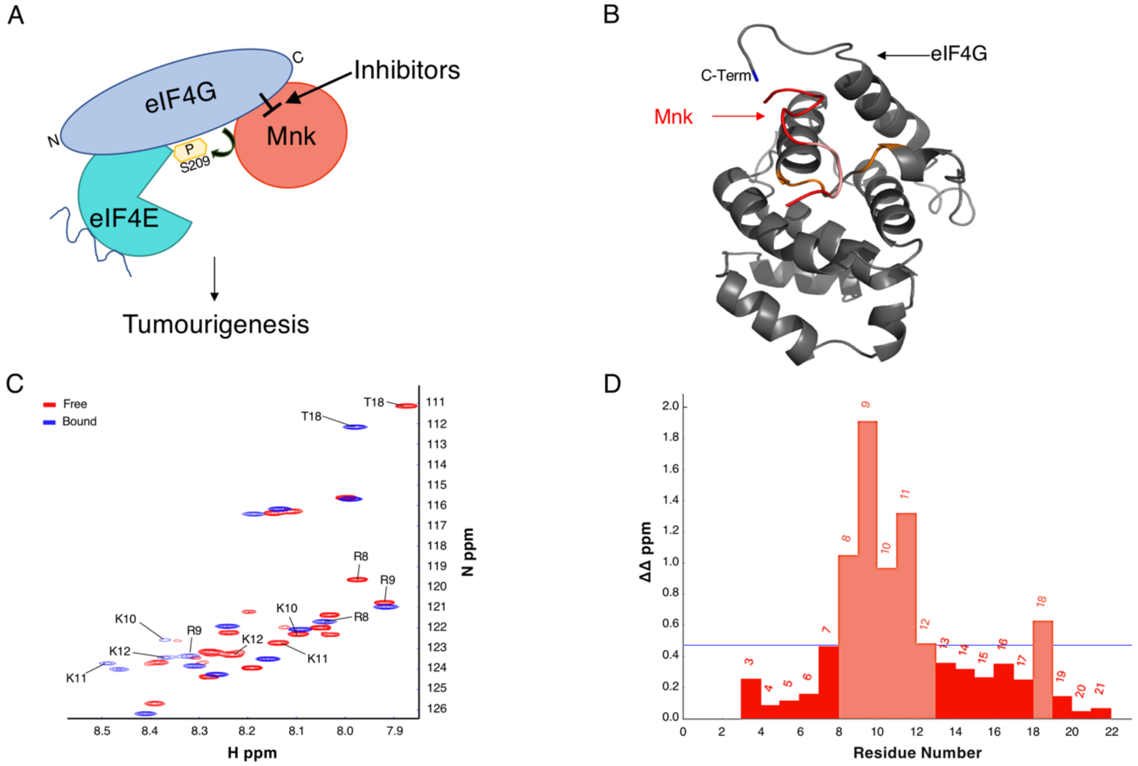


Figure 7.1 eIF4G-Mnk interaction background.

A eIF4e is phosphorylated by Mnk at the S209 activating the translation for proteins. Mnk needs to anchor eIF4G to exert its action, providing for a possible route towards active interference with small molecules of relevance for tumourigenesis. **B** 3D representation of the complex Mnk (red-pink) and eIF4G (dark grey). Both of structures were obtained by X-ray crystallographic studies; colours reflex the NMR chemical shift perturbation analysis. **C** ^{15}N-HSQC spectra of the Mnk peptide in its free form (red) and bound to eIF4G (blue). **D** CSP analysis of the Free-Bound complex using the CcpNmr AnalysisScreen module shows in light orange the residue of Mnk mostly interacting with eIF4G (Arg8, Arg9, Lys10, Lys11, Lys12, Thr18).

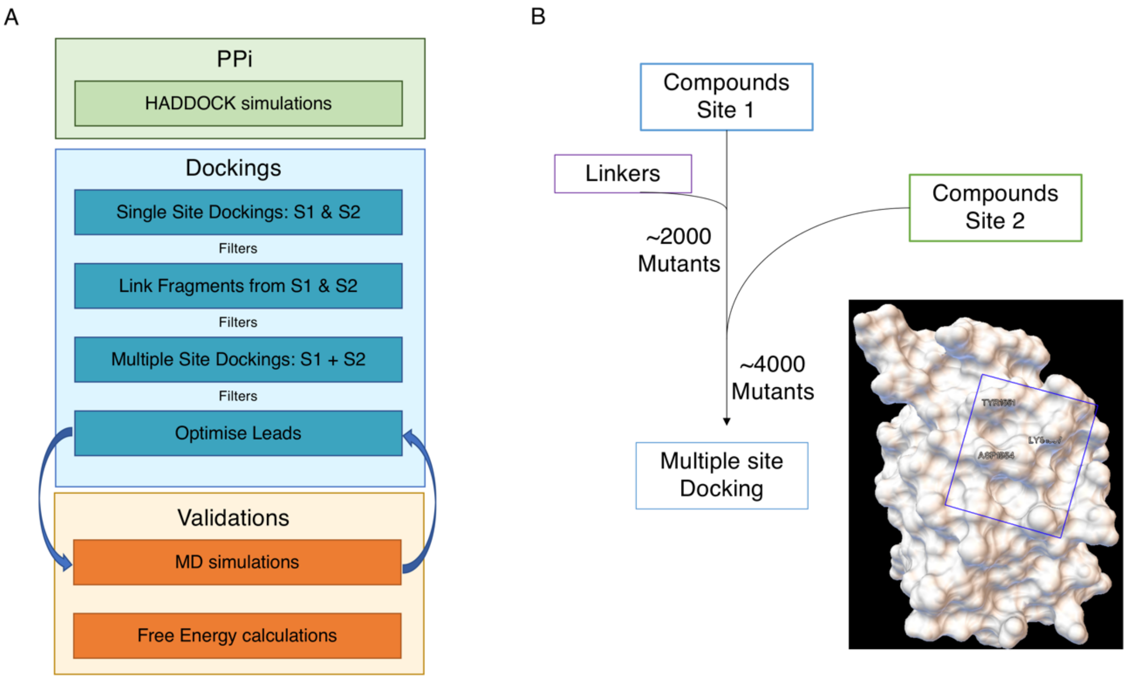


Figure 7.2 Project development workflow.

Proposed stages for the development of inhibitors for the complex eIF4G-Mnk. Each colour represents advancement sections according with different computational methodologies employed. **B** Dockings steps and fragment linkage summary.

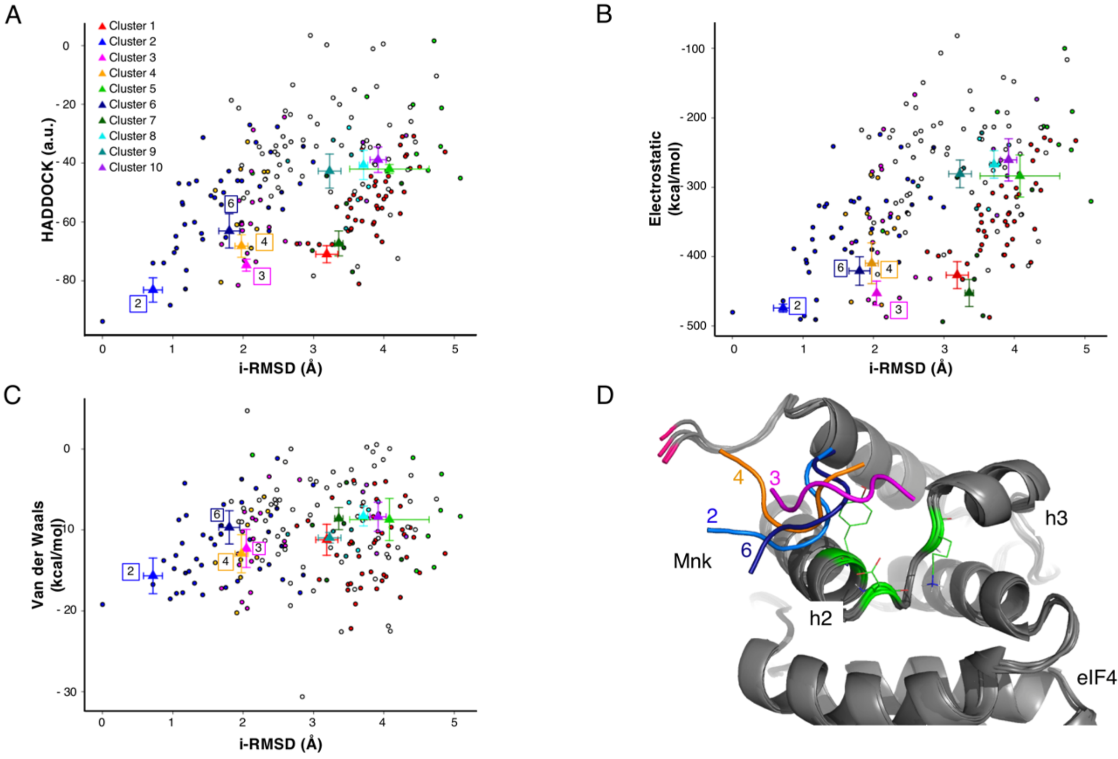


Figure 7.3 HADDOCK scorings.

A interface-RMSD (regarding residues backbone atoms in the interface) vs HADDOCK scores showed the best clusters to be 2, 3, 4, 6 (light blue, pink, yellow, dark blue respectively); most negative HADDOCK scores represent more favourable interaction. **B** Clusters grouped by electrostatic values vs *i*-RMSD confirmed 2, 3, 4, 6 groups being the most preferable. **C** Van der Waals interactions vs *i*-RMSD also suggested cluster 2 as the top model, followed by a closer grouping of clusters 3, 4, 6. Although the Haddock clusters do not define good or bad binding models, the presence of a well-defined cluster, such as the 2, is an indication of a potential binding pose. **D** 3D representations of the aligned four best scored clusters. eIF4G in dark grey, light green correspond to the NMR obtained restraints. It is noticeable how Mnk peptide for cluster 2, 4, 6 orientates toward the C-term of eIF4G (light pink), however a large part of interactions is between the helixes 2 and 3 (site-2).

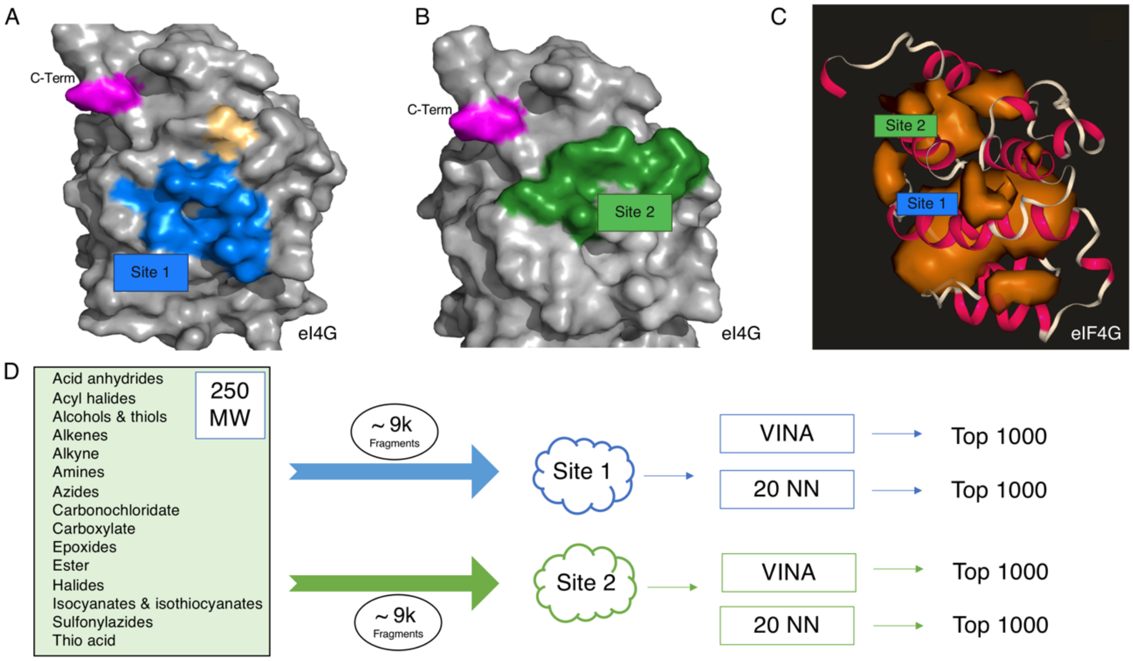


Figure 7.4 eIF4G docking sites representations and docking libraries.

A 3D structure illustration of eIF4G and its site-1 binding pocket (blue). **B** Site-2 binding pocket (green). **C** DeepSite-1 output shows potential binding pockets as an orange surface representation. Site-1 and Site-2 are indicated by the blue and green rectangles. **C** Fragments library description (light green rectangle), and fragment output count per docking algorithms pathway.

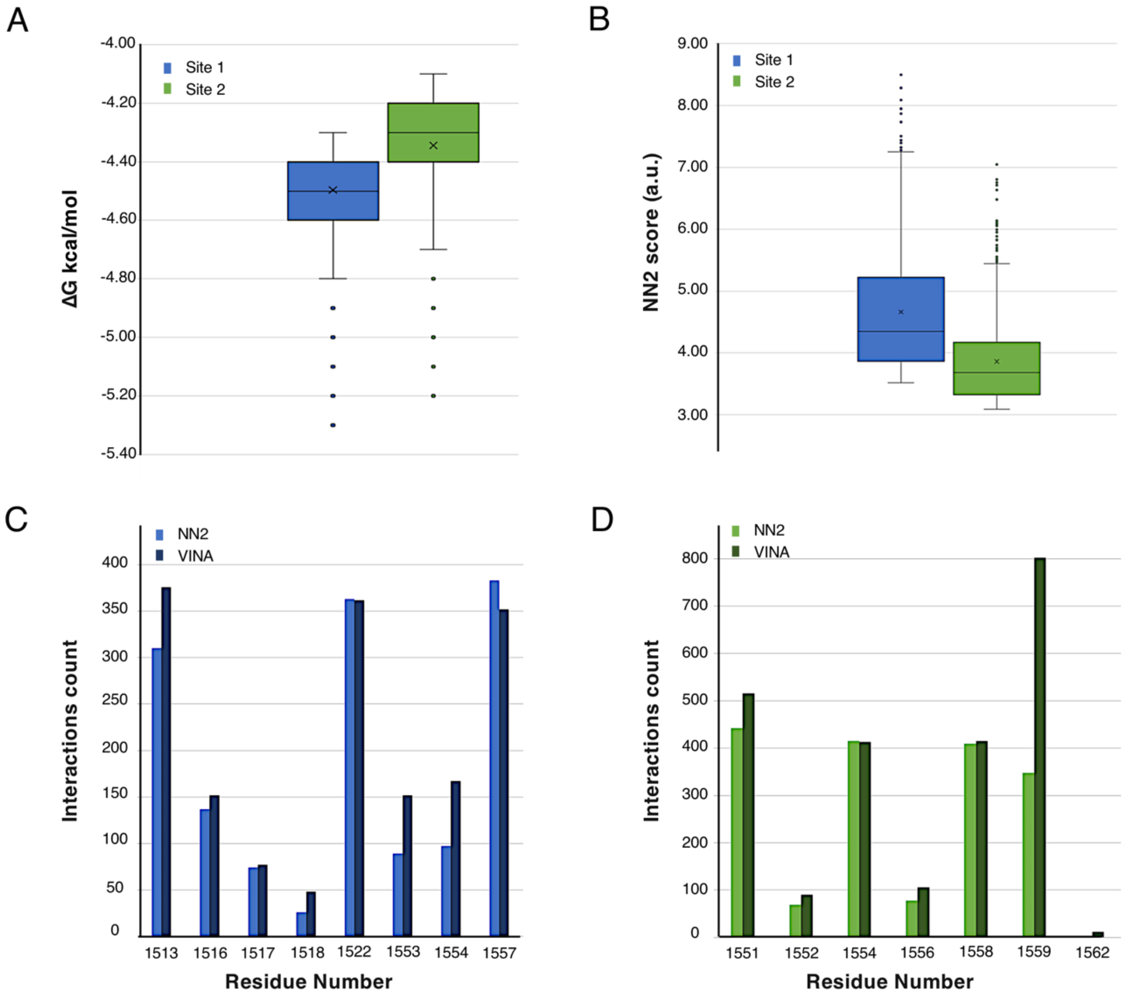


Figure 7.5 Single docking results for Site-1 and Site-2.

A Energy binding ΔG in kcal/mol for site-1 (blue) and for site-2 (green) using AutoDock VINA scoring function. Most negative values represent more favourable interactions. **B** NN2 scores in arbitrary units for site-1 (blue) and for site-2 (green) using a 20-Neural Networks scoring function. Most positive values denote strongest binding. **C-D** Interactions count calculated for each binding pose outputted by the two docking engines in respect to residues of the eIF4G site-1 (**C**) and site-2 (**D)**.

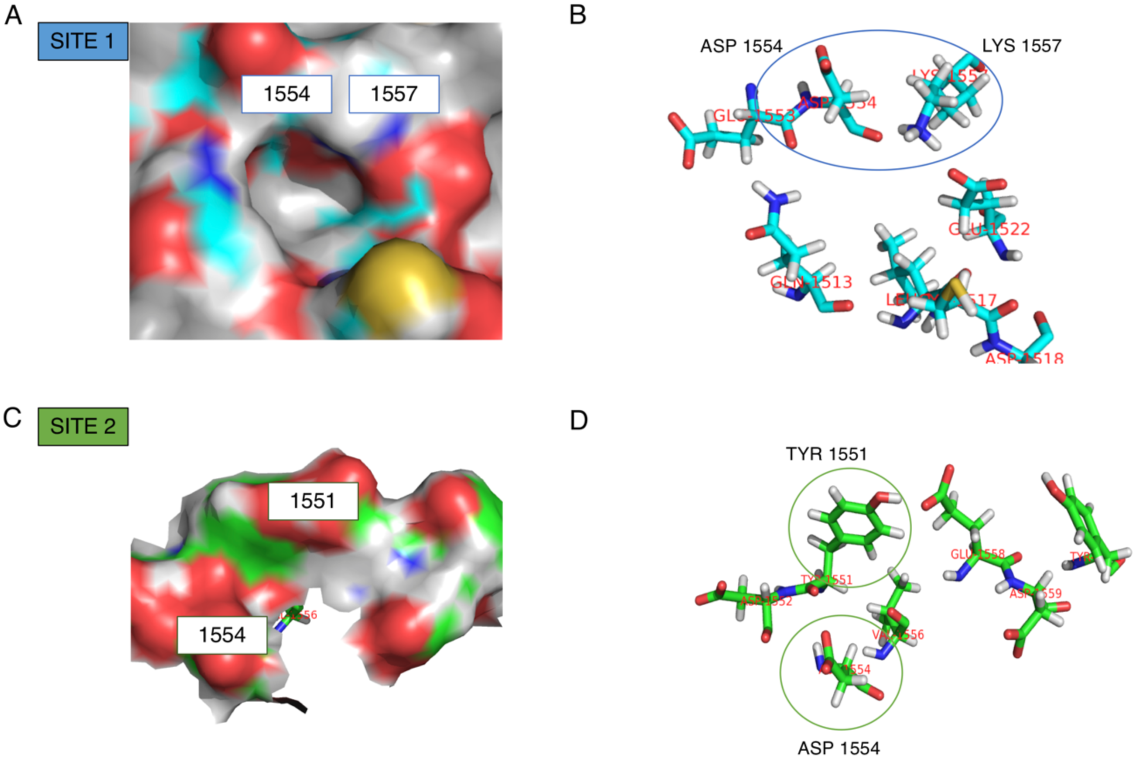
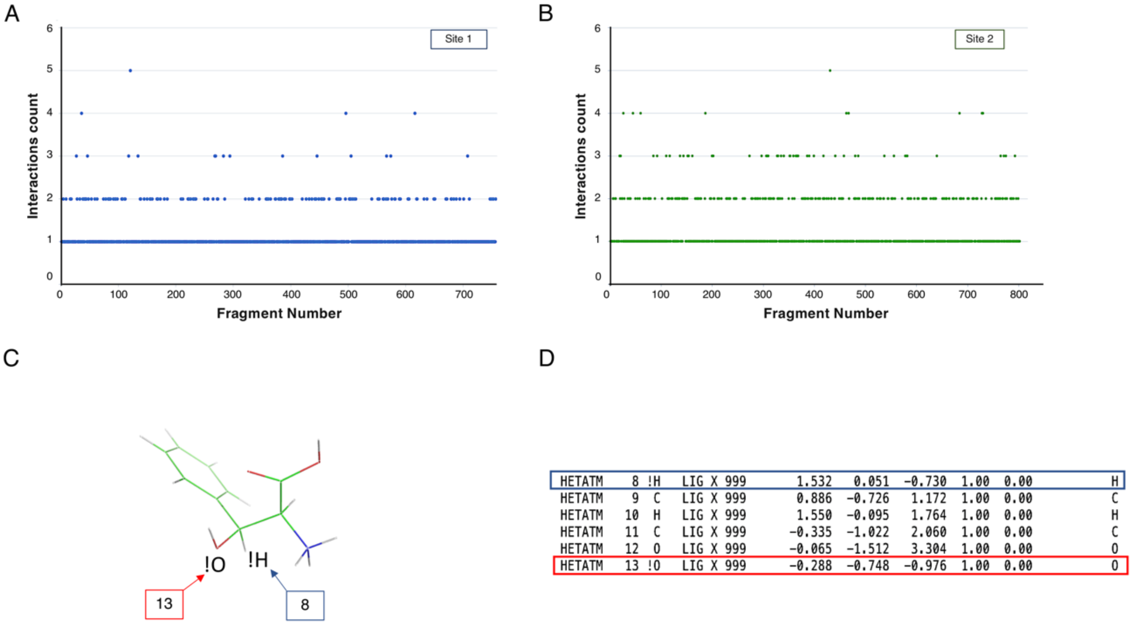


Figure 7.6 Site-1 and Site-2 most buried residues.

A Surface map representation of site-1. **B** Hydrophobic groove of site-1 in PyMOL stick representation, with Asp1554 and Lys1557 being the most buried residue for this pocket (blue circle). **C** Surface map representation of site-2. **B** Site-2 binding pocket in PyMOL stick representation. Asp1554 and Tyr1551 constitute the most buried residues (green circles).



```
HETATM    8 !H   LIG X 999       1.532   0.051  -0.730  1.00  0.00           H
HETATM    9  C   LIG X 999       0.886  -0.726   1.172  1.00  0.00           C
HETATM   10  H   LIG X 999       1.550  -0.095   1.764  1.00  0.00           H
HETATM   11  C   LIG X 999      -0.335  -1.022   2.060  1.00  0.00           C
HETATM   12  O   LIG X 999      -0.065  -1.512   3.304  1.00  0.00           O
HETATM   13 !O   LIG X 999      -0.288  -0.748  -0.976  1.00  0.00           O
```

Figure 7.7 Atom interactions and fragments linkage preparation.

A Count of ligand atoms for the top docked fragments interacting to the most buried residues identified for site-1 and site-2. **C** Example of a 3D lines representation of a fragment with a “!” tag for each crucial atom. **D** New PDB files with amended atoms nomenclatures.

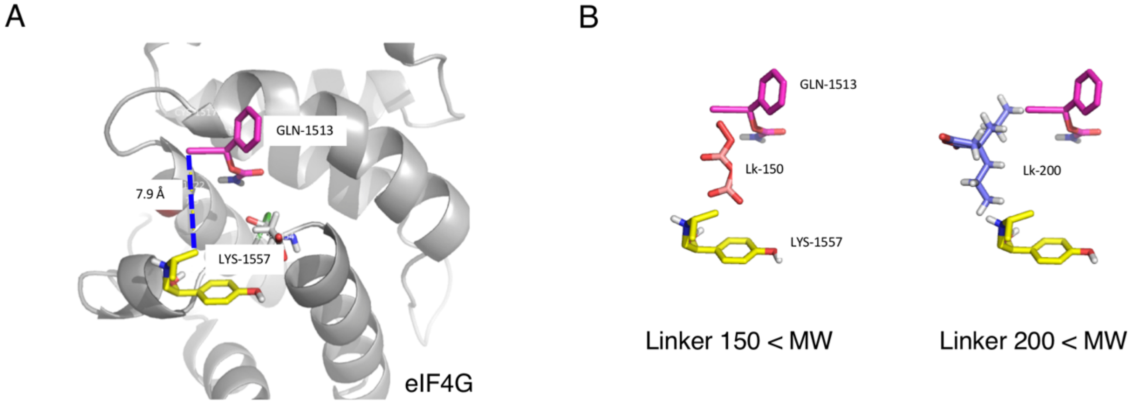


Figure 7.8 Fragment linkers.

A eIF4G and two small fragments docked in site-1 and site-2. Distance measurement has been used to select the appropriate linker. **B** 3D examples of linkers that were randomly selected from libraries of 150 MW and 200 MW fragments.

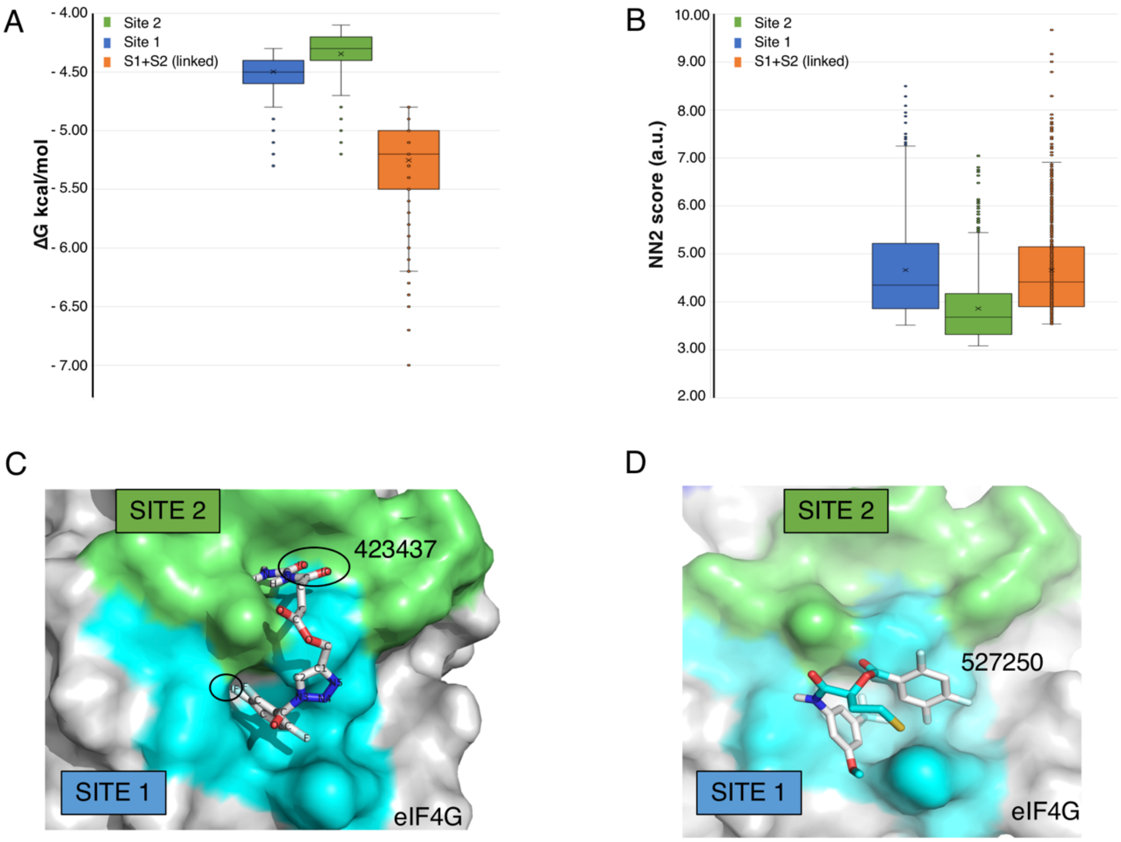
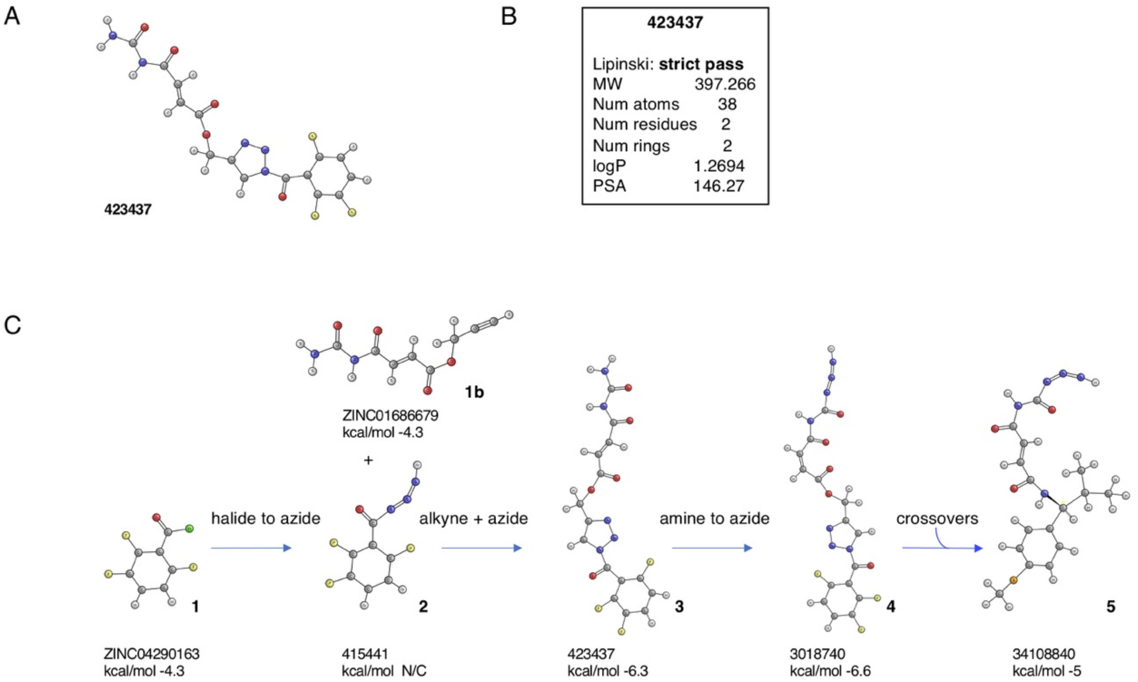


Figure 7.9 Multiple-site docking results.

A VINA energy binding ΔG for site-1 (blue), for site-2 (green) previously calculated, and the newly linked compounds library (orange).

B NN2 scores in arbitrary units for site-1 (blue), site-2 (green) and the linked fragments (orange). **C** Surface representation of one of the top VINA scored compounds. In black circles are highlighted the previously flagged atoms interacting with the respectively binding pockets. **D** Surface representation of one of the top NN2 scored compounds.



423437	
Lipinski: **strict pass**	
MW	397.266
Num atoms	38
Num residues	2
Num rings	2
logP	1.2694
PSA	146.27

Figure 7.10 Ligand 423437 and development history.

A 2D illustration of the ligand selected for proceeding with further optimisations. **B** Druglikeness property for the compound 423437. **C** Virtual click-chemistry reactions from original fragments to optimised versions.

Compound-**1**: 2,3,6-trifluorobenzoyl chloride.

Compound-**1b**: prop-2-yn-1-yl(*E*)-4-oxo-4-ureidobut-2-enoate.

Compound-**2**: N-(1λ^4-diazenylidene)-2,3,6-trifluorobenzamide.

Compound-**3**: 1-(2,3,6-trifluorobenzoyl)-1H-1,2,3-triazol-4-yl)methyl (E)-4-oxo-4-ureidobut-2-enoate.

Compound-**4**: 1-(2,3,6-trifluorobenzoyl)-1H-1,2,3-triazol-4-yl)methyl (E)-4-(3-(1λ^4-diazenylidene)ureido)-4-oxobut-2-enoate.

Compound-**5**: (S)-N^1-((1λ^4-diazenylidene)carbamoyl)-N^4-(2-methyl-1-(4-(methylthio)phenyl)propyl)fumaramide.

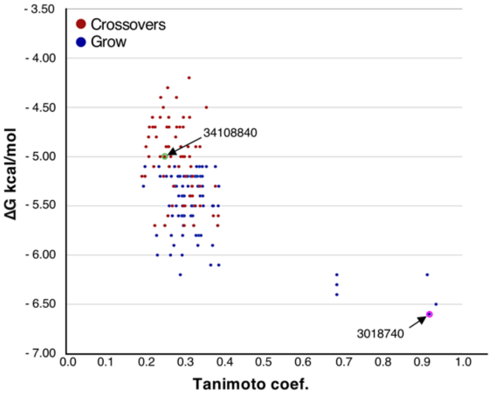


Figure 7.11 Variants similarities for the ligand 423437.

A Tanimoto coefficient versus the binding energies for optimised libraries in respect to the original fragment 423437. Highlighted 3018740 (purple) derived from the fragment grow engine and 34108840 (green), derived from the crossover engine present in AutoGrow[82].

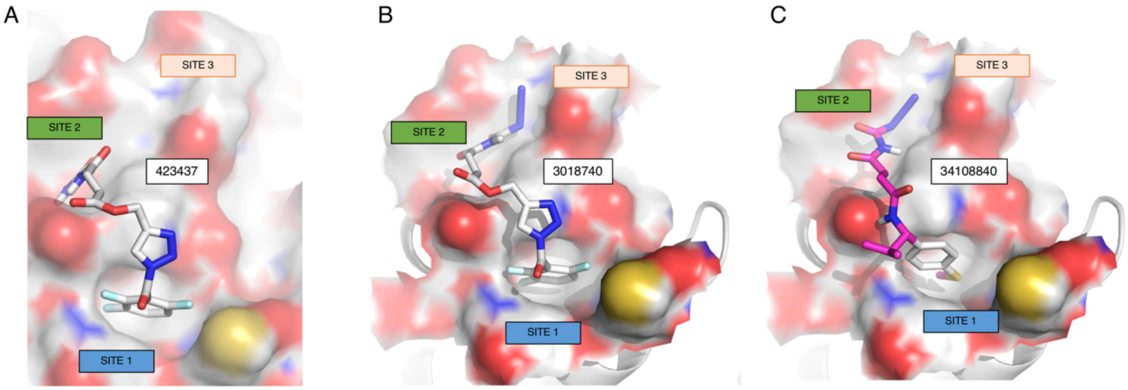


Figure 7.12 Surface representations for optimised ligands.

A 423437 **B** 3018740 **C** 34108840 surface map, ligands interact similarly to eIF4G binding pockets; however, optimised versions 3018740 and 34108840 appear to extend the interaction to a third site (orange rectangle).

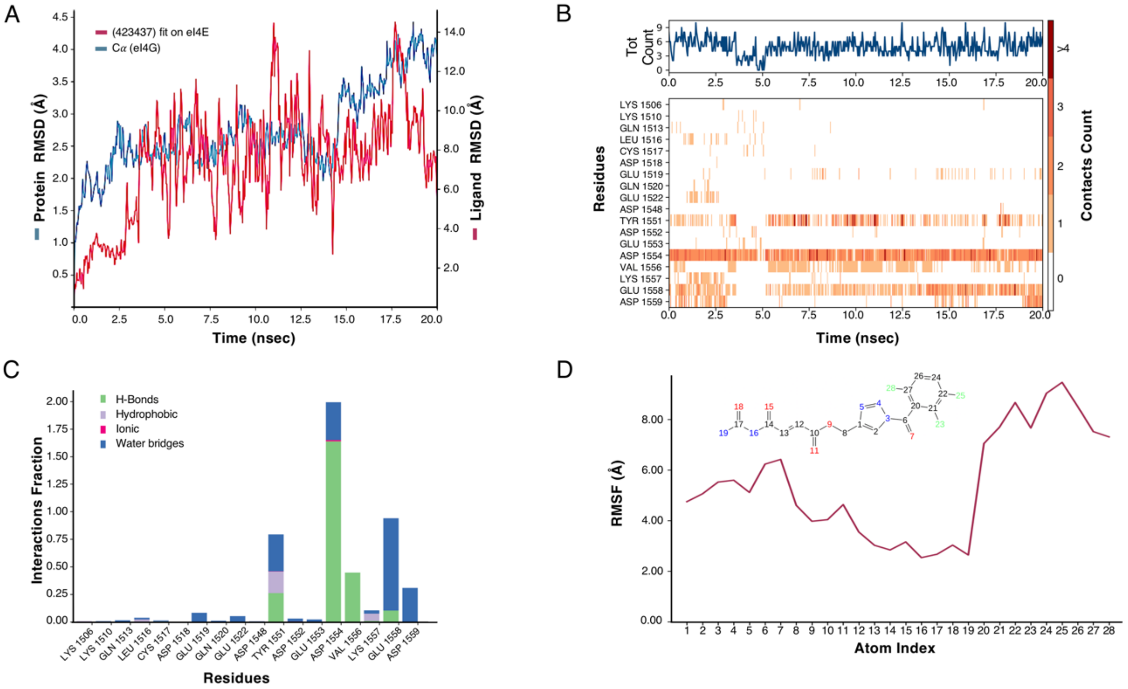


Figure 7.13 MD simulations for 423437.

A RMSD evolution for the protein cα (blue) and ligand (red) calculated independently in a simulation of 20ns. **B** Heat-map with contacts count for each eIF4G residues during the simulation. Dark red correspond to the most occurrences (>4), whereas white gaps signify no interactions between protein and ligand. Top graph, summaries the total count in a 1D curve in respect of time. **C** Protein-ligand interactions divided by groups. Hydrogen bonds, hydrophobic, ionic and water bridges, stacked bar charts represent a normalised value count; a value over 1.0 suggests that the residue might undertake multiple simultaneous contacts with the ligand. **D** Ligand Root Mean Square Fluctuation (L-RMSF) shows changes in the ligand atom positions through the simulation time.

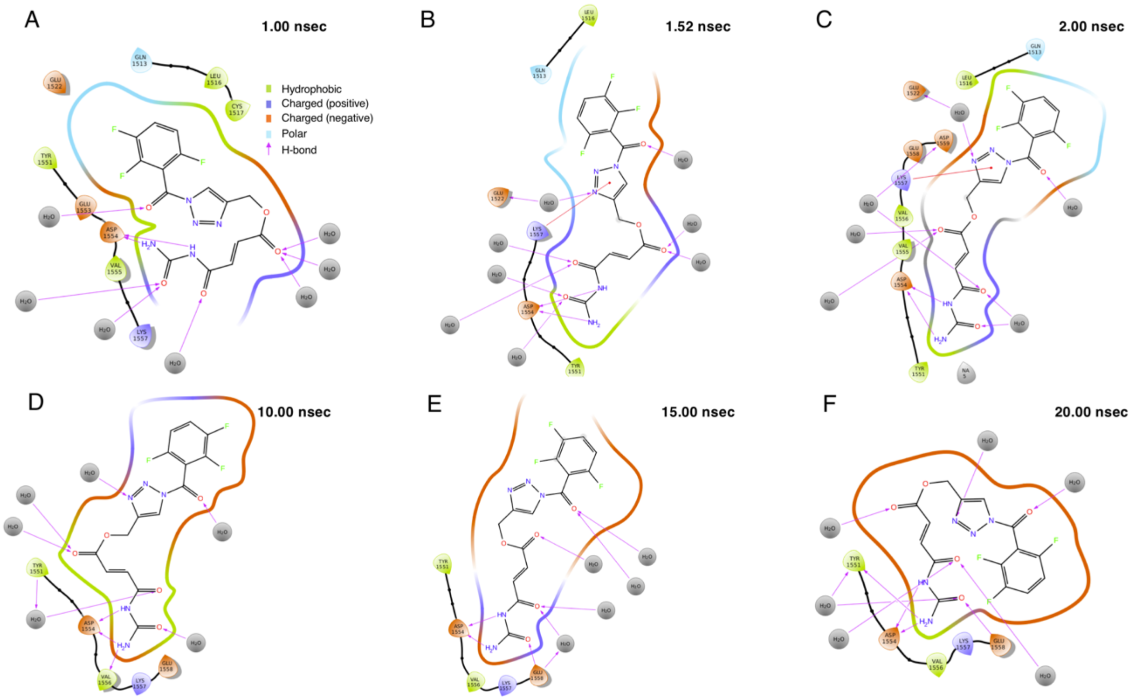


Figure 7.14 423437 interactions during simulation.

Simulation snapshots at 1.00, 1.52, 2.00, 10.00, 15.00, 20.00 nsec (**A-F**) and the multiple ligand-target interaction types for each timeframe.

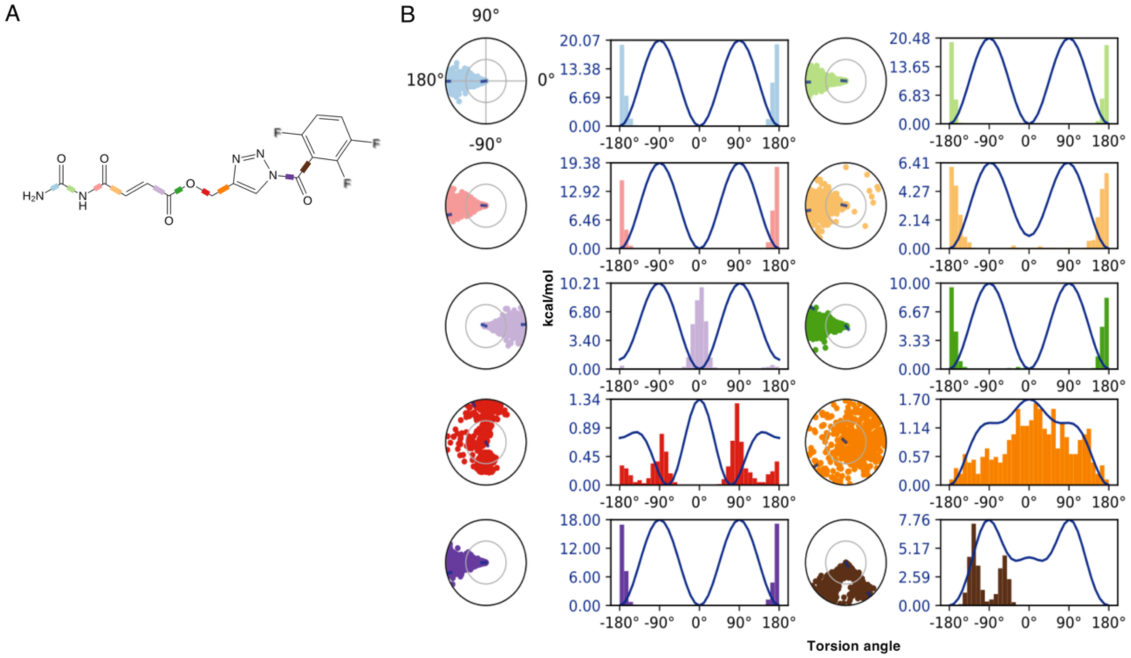


Figure 7.15 423437 angle torsions analysis.

A 2D ligand structure and coloured rotatable bond, each torsion is described in the following B figure by a dial and bar plot of the same colour. **B** Conformational adjustments in degrees for each rotatable bond in the ligand throughout the simulation trajectory. The centre of the dial plot represents the initial frames and expands radially outwards describing the simulation evolution. The adjacent bar plots outline the probability density of the torsion reported in the dials.

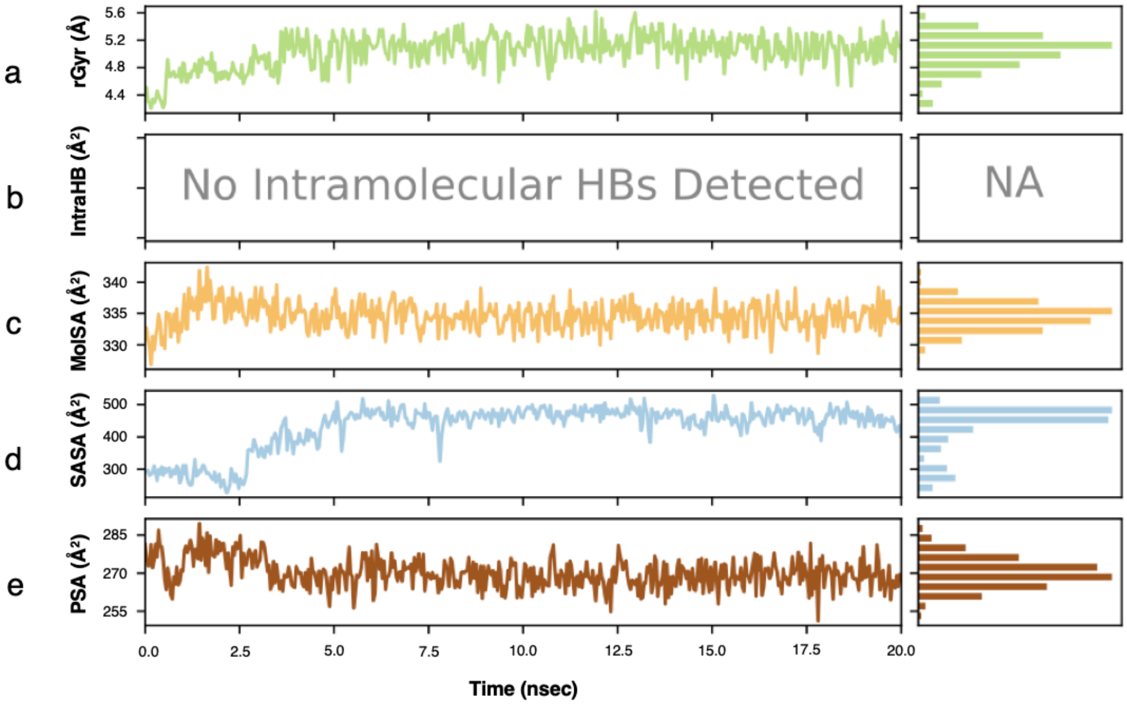


Figure 7.16 423437 surface area analysis.

Summary of the various ligand properties through the 0-20ns simulation. **a** Radius of gyration, rGyr, which is related to its main moment of inertia and compute the “extendedness” of a ligand through the time; **b** indicates the intramolecular Hydrogen Bonds, IntraHB, were not detected for 423437. **c** Molecular Surface Area (MolSA), which corresponds to a van der Waals surface area, **d** Solvent Accessible Surface Area (SASA). **f** Polar Surface Area (PSA), equivalent to the solvent accessible surface area in a molecule for which only oxygen and nitrogen atoms are considered.

7.7. Appendix

Prior to the simulation a series of minimisations were performed using the built-in protocols of Schrodinger Maestro[79]. The stages were run using the default parameters and steps are summarised in table 7.1, whereas the full script parameters are summarised in table 7.2. Lastly the full molecular dynamics simulation parameters are displayed in table 7.3.

Table 7.1 Relaxation steps summary.

Steps	1	2	3	4	5	6
Simulation type ensemble	NVT	NVT	NPT	solvatation	NPT	NPT
thermostat(t) / barostat(b)		Berendsen	Berendsen		Berendsen	Berendsen
Temperature	10K	10K	10K		300K	300K
Pressure		n.d.	1atm		1atm	1atm
Time		12ps	12ps		12ps	24ps
Velocity resampling		1ps	1ps		1ps	
Temperature relaxation		fast	fast		fast	fast
Pressure relaxation const.		slow		slow	normal	
Soluted atoms restrained	non-hydrogen	non-hydrogen	non-hydrogen		non-hydrogen	non-hydrogen

Table 7.2 Relaxation steps configuration parameters.

task	
task	desmond:auto
set_family	
desmond	
checkpt.write_last_step	
simulate1	
title	Brownian Dynamics NVT, T : 10 K, small timesteps, and restraints on solute heavy atoms, 100ps
annealing	off
time	100
timestep	
0.001	
0.001	
0.003	
temperature	10

ensemble	
class	NVT
method	Brownie
brownie	
delta_max	
restrain	
atom	
force_constant	
simulate2	
effect_if	[[==, -gpu, @*.*.jlaunch_opt[-1]], ensemble.method : Langevin]
title	NVT, T : 10 K, small timesteps, and restraints on solute heavy atoms, 12ps
annealing	off
time	12

timestep	
0.001	
0.001	
0.003	
temperature	10
restrain	
atom	
force_constant	
ensemble	
class	
method	
thermostat.tau	
randomize_velocity.interval	1
eneseq.interval	0.3
trajectory.center	
simulate3	
title	NPT, T : 10 K, and restraints on solute heavy atoms, 12ps
effect_if	[[==, -gpu, @*.*.jlaunch_opt[-1]], ensemble.method: Langevin]
annealing	off
time	12
temperature	10
restrain	retain
ensemble	
class	
method	
thermostat.tau	
barostat.tau	
randomize_velocity.interval	1
eneseq.interval	0.3
trajectory.center	
solvate_pocket	
should_skip	
ligand_file	

simulate4	
title	NPT and restraints on solute heavy atoms, 12ps
effect_if	[[@*.*.annealing], annealing = off, temperature = @*.*.temperature[0][0], ensemble.method : Langevin]
time	12
restrain	retain
ensemble	
class	
method	
thermostat.tau	
barostat.tau	
randomize_velocity.interval	1
eneseq.interval	0.3
trajectory.center	
simulate5	
title	NPT and no restraints, 24ps
effect_if	[[@*.*.annealing], annealing =off, temperature : @*.*.temperature[0][0], ensemble.method : Langevin]
time	24
ensemble	
class	
method	
thermostat.tau	
barostat.tau	
eneseq.interval	0.3
trajectory.center	solute
simulate6	
cfg_file	423437_md.cfg
jobname	$MASTERJOBNAME
dir	.
compress	

Table 7.3 MD simulation configuration parameters.

ORIG_CFG	
annealing	false
backend	
app	mdsim
boot	
file	423437_md-in.cms
type	mae
bigger_rclone	false
checkpt	
first	
interval	
name	
write_last_step	
coulomb_method	useries

cpu	1
cutoff_radius	9.0
elapsed_time	0.0
energy_group	false
eneseq	
first	0.0
interval	1.2
name	$JOBNAME$[_replica$REPLICA$].
ensemble	
barostat	
tau	
class	NPT
method	MTK
thermostat	

tau	
glue	solute
maeff_output	
first	0.0
interval	120.0
name	$JOBNAME$[_replica$REPLICA$]-out.cms
periodicfix	true
trjdir	$JOBNAME$[_replica$REPLICA$]_trj
meta	false
meta_file	
model_file	423437_md-in.cms
pressure	
1.01325	
isotropic	
randomize_velocity	
first	0.0
interval	inf
seed	2007
temperature	@*.temperature
restrain	none
simbox	
first	0.0
interval	1.2
name	$JOBNAME$[_replica$REPLICA$]_simbox.dat
surface_tension	0.0
taper	false
temperature	
300.0	
0	
time	20000.0
timestep	
0.002	
0.002	
0.006	
trajectory	
center	
first	0.0
format	dtr
frames_per_file	250
interval	40.0
name	$JOBNAME$[_replica$REPLICA$]_trj
periodicfix	true
write_velocity	false
app	mdsim
argv	
/home/tjr22/luca/schrodinger2019-3/internal/bin/gdesmon	
--include	
423437_md-out.cfg	
boot	
file	423437_md-in.cms
type	mae
config	
ORIG_CFG	
annealing	false
backend	
app	mdsim
boot	

file	423437_md-in.cms
type	mae
bigger_rclone	false
checkpt	
first	
interval	
name	
write_last_step	
coulomb_method	useries
cpu	1
cutoff_radius	9.0
elapsed_time	0.0
energy_group	false
eneseq	
first	0.0
interval	1.2
name	$JOBNAME$[_replica$REPLICA$].
ensemble	
barostat	
tau	
class	NPT
method	MTK
thermostat	
tau	
glue	solute
maeff_output	
first	0.0
interval	120.0
name	$JOBNAME$[_replica$REPLICA$]-out.cr
periodicfix	true
trjdir	$JOBNAME$[_replica$REPLICA$]_trj
meta	false
meta_file	
model_file	423437_md-in.cms
pressure	
1.01325	
isotropic	
randomize_velocity	
first	0.0
interval	inf
seed	2007
temperature	@*.temperature
restrain	none
simbox	
first	0.0
interval	1.2
name	$JOBNAME$[_replica$REPLICA$]_si
surface_tension	0.0
taper	false
temperature	
300.0	
0	
time	20000.0
timestep	
0.002	
0.002	
0.006	
trajectory	

center	
first	0.0
format	dtr
frames_per_file	250
interval	40.0
name	$JOBNAME$[_replica$REPLICA$]_trj
periodicfix	true
write_velocity	false
app	mdsim
boot	
file	423437_md-in.cms
type	mae
config	?
force	
bonded	
exclude	
include	
constraint	
maxit	
tol	
ignore_com_dofs	true
nonbonded	
accuracy_level	0
far	
Nterms	32
kappa	
	0.333333
	0.333333
	0.333333
n_k	
	45
	45
	45
order	
	4
	4
	4
r_spread	4.0
sigma_s	0.85
spreading_style	scatter_gather
type	QuadS
n_zone	1024
near	
correct_average_dispersion	
r_tap	
taper	
type	
r_cut	9.0
r_lazy	10.113508356103871
sigma	2.048076502869348
type	useries
term	
gibbs	
alpha_vdw	0.5
output	
first	
interval	

name	fep.dE
type	none
weights	
bondA	
bondB	
es	
qA	
qB	
qC	
vdw	
vdwA	
vdwB	
window	-1
list	
virtual	
exclude	
include	
global_cell	
clone_policy	rounded
margin	1.1135063561038707
n_replica	1
partition	
	1
	1
	1
r_clone	5.056754178051936
reference_time	0.0
topology	periodic
gui	
ewald_tol	
integrator	
Multigrator	
barostat	
Langevin	
tau	0.020833333
thermostat	
seed	2012
tau	0.016129
type	Langevin
MTK	
tau	0.041666666666666664
thermostat	
NoseHoover	
mts	2
tau	
	0.020833333333333332
	0.020833333333333332
	0.020833333333333332
type	NoseHoover
timesteps	48
type	MTK
nve	
type	Ve
thermostat	
DPD	

seed	2012
Langevin	
seed	2012
tau	0.016129
NoseHoover	
mts	2
tau	
0.08333333333333333	
0.08333333333333333	
0.08333333333333333	
timesteps	12
type	NoseHoover
brownie	
barostat	
delta_max	0.1
tau	1.0
thermostat	
seed	
delta_max	0.1
thermostat	
seed	
brownie_NPT	
barostat	
T_ref	300.0
tau	0.016129
thermostat	
seed	
delta_max	0.1
thermostat	
seed	
brownie_NVT	
delta_max	0.1
thermostat	
seed	
dt	0.002
posre_scaling	1.0
pressure	
P_ref	
isotropy	
max_margin_contraction	
tension_ref	
respa	
far_timesteps	
migrate_timesteps	
near_timesteps	
outer_timesteps	
temperature	
T_ref	
type	Multigrator
mdsim	
checkpt	
first	
interval	
name	
wall_interval	
write_first_step	
write_last_step	
last_time	20000.0
plugin	

anneal	
first	0.0
interval	1.2
schedule	
time	
0.0	
30.0	
60.0	
90.0	
600.0	
value	
0.0	
300.0	
600.0	
900.0	
300.0	
type	anneal
energy_groups	
first	0.0
interval	1.2
name	
options	
corr_energy	
type	energy_groups
write_report	true
eneseq	
first	
flush_interval	
interval	
name	
type	
gcmc	
batch_size	1600
eneseq	
name	system_gcmc.ene
first	0.0
grid	
exclusion_radius	
region_buffer	
spacing	
track_voids	
whole_box_frequency	
interval	3.0
mu_excess	-6.18
name	system.gcmc
nsteps	5000
quiet	true
restore_engrps	false
seed	2007
solvent_density	0.03262
temperature	300.0
type	gcmc
list	
status	
eneseq	
trajectory	

randomize_velocities	
remove_com_motion	
maeff_output	
bootfile	423437_md-in.cms
first	0.0
full_system_only	false
glue	
1872	
2581	
interval	120.0
name	423437_md-out.cms
periodicfix	true
precision	8
trjdir	423437_md_trj
type	maeff_output
write_last_step	true
maeff_snapshot	
first	0.0
interval	1.2
name	
type	maeff_snapshot
randomize_velocities	
first	0.0
interval	inf
seed	2007
temperature	300.0
type	randomize_velocities
remove_com_motion	
first	0.0
interval	inf
type	remove_com_motion
simbox_output	
first	0.0
interval	1.2
name	
type	simbox_output
status	
first	
interval	
type	
trajectory	
center	
first	0.0
format	dtr
frames_per_file	250
glue	
1872	
2581	
interval	40.0
mode	noclobber
name	423437_md_trj
periodicfix	true
type	trajectory
write_first_step	true
write_last_step	true
write_velocity	false
title	Desmond MD simulation
migration	

first	0.0
interval	0.018000000000000002
spatial_order	auto
force	
bonded	
exclude	
include	
constraint	
maxit	
tol	
ignore_com_dofs	true
nonbonded	
accuracy_level	0
far	
Nterms	32
kappa	
0.333333	
0.333333	
0.333333	
n_k	
45	
45	
45	
order	
4	
4	
4	
r_spread	4.0
sigma_s	0.85
spreading_style	scatter_gather
type	QuadS
n_zone	1024
near	
correct_average_dispersion	
r_tap	
taper	
type	
r_cut	9.0
r_lazy	10.113508356103871
sigma	2.048076502869348
type	useries
term	
gibbs	
alpha_vdw	0.5
output	
first	
interval	
name	
type	none
weights	
bondA	
bondB	
es	
qA	
qB	
qC	

vdw	
vdwA	
vdwB	
window	-1
list	
virtual	
exclude	
include	
global_cell	
clone_policy	rounded
margin	1.1135063561038707
n_replica	1
partition	
1	
1	
1	
r_clone	5.056754178051936
reference_time	0.0
topology	periodic
gui	
ewald_tol	
integrator	
Multigrator	
barostat	
Langevin	
tau	0.020833333
thermostat	
seed	2012
tau	0.016129
type	Langevin
MTK	
tau	0.041666666666666664
thermostat	
NoseHoover	
mts	2
tau	
0.020833333333333332	
0.020833333333333332	
0.020833333333333332	
type	NoseHoover
timesteps	48
type	MTK
nve	
type	Ve
thermostat	
DPD	
seed	
Langevin	
seed	2012
tau	0.016129
NoseHoover	
mts	2
tau	
0.08333333333333333	
0.08333333333333333	
0.08333333333333333	

timesteps	12
type	NoseHoover
brownie	
barostat	
delta_max	0.1
tau	1.0
thermostat	
seed	
delta_max	0.1
thermostat	
seed	
brownie_NPT	
barostat	
T_ref	300.0
tau	0.016129
thermostat	
seed	
delta_max	0.1
thermostat	
seed	
brownie_NVT	
delta_max	0.1
thermostat	
seed	
dt	0.002
posre_scaling	1.0
pressure	
P_ref	
isotropy	
max_margin_contraction	
tension_ref	
respa	
far_timesteps	
migrate_timesteps	
near_timesteps	
outer_timesteps	
temperature	
T_ref	
type	Multigrator
mdsim	
checkpt	
first	
interval	
name	
wall_interval	
write_first_step	
write_last_step	
last_time	20000.0
plugin	
anneal	
first	0.0
interval	1.2
schedule	
time	
0.0	
30.0	
60.0	
90.0	
600.0	

value	
0.0	
300.0	
600.0	
900.0	
300.0	
type	anneal
energy_groups	
first	0.0
interval	1.2
name	
options	
corr_energy	
type	energy_groups
write_report	true
eneseq	
first	
flush_interval	
interval	
name	
type	
gcmc	
batch_size	1600
eneseq	
name	system_gcmc.ene
first	0.0
grid	
exclusion_radius	
region_buffer	
spacing	
track_voids	
whole_box_frequency	
interval	3.0
mu_excess	-6.18
name	system.gcmc
nsteps	5000
quiet	true
restore_engrps	false
seed	2007
solvent_density	0.03262
temperature	300.0
type	gcmc
list	
status	
eneseq	
trajectory	
randomize_velocities	
remove_com_motion	
maeff_output	
bootfile	423437_md-in.cms
first	0.0
full_system_only	false
glue	

1872	
2581	
interval	120.0
name	423437_md-out.cms
periodicfix	true
precision	8
trjdir	423437_md_trj
type	maeff_output
write_last_step	true
maeff_snapshot	
first	0.0
interval	1.2
name	
type	maeff_snapshot
randomize_velocities	
first	0.0
interval	inf
seed	2007
temperature	300.0
type	randomize_velocities
remove_com_motion	
first	0.0
interval	inf
type	remove_com_motion
simbox_output	
first	0.0
interval	1.2
name	
type	simbox_output
status	
first	
interval	
type	
trajectory	
center	
first	0.0
format	dtr
frames_per_file	250
glue	
1872	
2581	
interval	40.0
mode	noclobber
name	423437_md_trj
periodicfix	true
type	trajectory
write_first_step	true
write_last_step	true
write_velocity	false
title	Desmond MD simulation
migration	
first	0.0
interval	0.018000000000000002
spatial_order	auto
threader_size	0

7.8. Acknowledgments

I thank Drs Christoph Gorgulla[a], Thibault Viennet[a] for their previous work on eIF4G-Mnk, suggestions and inputs for future directions. Dr TJ Ragan[b] for running the MD simulations at the LISCB[b] facilities. I specially thank Dr Hari Arthanary[a] for hosting me in his lab and providing resources and expertise for this project.

Finally, I thank the MRC-Impact panel for funding the internship with the additional “flexible scheme”, which enabled me to achieve a successful and fruitful experience.

7.9. Addresses

[a] Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, USA.

[b] Leicester Institute of Structural and Chemical Biology, Department of Molecular and Cell Biology, University of Leicester, Henry Wellcome Building, Lancaster Road, Leicester LE1 7HN, United Kingdom.

Chapter 8

Final Remarks

8.1. What have I learnt about FBDD by NMR and how CcpNmr AnalysisScreen can aid future discoveries?

By directly assessing the development history of three commercial drugs from the primary screening to the final lead optimisation, I could understand the rationale behind the selection of each technique needed for tackling each individual progression step and the overall challenges associated with these laborious projects. The study cases discussed in chapter 1 highlighted the underpinning role and impact of NMR techniques throughout the drug discovery phases. More broadly, the inspection of multiple compounds in clinical phases has described how the NMR spectroscopy plays a central role in the fragmental drug discovery process, FBDD[4]. Above all, this spectroscopic technique is the most preferred choice when it comes to the identification of weak binders in the primary screening. In the initial stages of drug discovery, hundreds to thousands of fragments are tested against a macromolecule using a variety of one-dimensional ligand-detected methodologies. As a result, an enormous amount of data is generated in a relatively short time. Although acquisition and processing of NMR spectral information can be performed in an automated fashion, analysis is often attained through rigorous qualitative visual inspections.

Pharmaceutical companies might avail of commercial analysis software, such as ACD-lab[140], MestreNova[141], Top-Spin[116] plugins or build their own custom scripts for assessing their own outputs. However, the high premium requested by these platforms cannot be justified by academic groups or occasional users, thus, scientists typically resort in cumbersome operations and ultimately, they might draw conclusions based upon subjective interpretations.

During these years, I have focused extensively on the design and development of a unique application dedicated for assessing data derived by NMR-FBDD. It is called CcpNmr AnalysisScreen, or AnalysisScreen for short, and it is a part of the main Analysis version-3 software suite[28] (chapter 2).

AnalysisScreen is built using robust software architectures which enabled the inclusion of current NMR analysis routines, new algorithms, and most importantly, it will enable following developers in adding methodologies that might be needed in the future (chapter 3).

The strength and novelty in the software reside in a comprehensive metadata parser, the ability of creating custom workflows, the presence of fast algorithms with automated detection of optimal parameters, multiple scoring functions, and the reproducibility of results (chapters 4-5). Additionally, a modern graphical user interface facilitates the analysis and experience of users, therefore, increasing their productivity.

By working with several academic and industrial groups, I designed AnalysisScreen aiming to be the ultimate NMR software package needed to cover all aspects of fragment-based drug discovery data analysis; as such, it includes an exhaustive tool required for inspecting macromolecular binding sites information, the so-called chemical shift mapping module[189](chapter 6). Chemical shift perturbation analysis is a powerful tool to elucidate macromolecular interactions and has been used extensively in most clinical NMR-derived drugs. It was in my greatest interests in providing the community an intuitive and interactive graphical tool which allows users to drastically reduce the time required for this crucial operation.

From the literature has emerged that NMR played a pivotal role in elucidating small molecules binding poses through the analysis of high dimensionalities spectral observations. Currently AnalysisScreen does not support yet this area. However, future collaborations with industrial partners might originate in concise development plans.

New industrial partnerships are constantly being established within the group and they will be paramount in consolidating the AnalysisScreen capabilities so it is beneficial to the entire NMR community and ultimately it will aid in the development of drugs in a faster timescale.

8.2. How can the development of new drugs from fragments can be shortened?

In 2016, ABT-199, commonly known as Venetoclax, has been the first confirmed FDA-approved drug derived by NMR-FBDD[26,49,61,62]. Its development took nearly 20 years.

It appears that the majority of efforts are put into molecular optimisations, that are mostly guided through multiple molecular adjustments followed by their new chemical synthesis and cyclical experimental re-evaluations. Although this will eventually culminate in efficient drugs, some considerations arise whether the usage of computational approaches could have made the whole process more efficient.

High-performance computers are constantly increasing in computational power and availability to researchers, furthermore, large number of algorithms can aid the drug discovery in unprecedented ways[79,82,183].

Using freely available tools, I have demonstrated how they can be assembled in custom workflows to guide the creation of highly tailored candidates in a limited amount of time (chapter 7).

By experiencing in first person challenges associated with drug-optimisations and the impossibility of using a stand-alone protocol or tool, I firmly believe that fragment-based and structure-based drug design can be aided by developing new platforms and workflows that merge together seamlessly *in-silico* tools such as *de-novo* library creation, dockings, molecular dynamics simulations and experimental knowledge.

Ideally computational methods would take in parallel the biological aspects into consideration. Some of these may be, for example, by enhancing ligand's robustness for possible drug resistance. In doing so, it is possible to combine several pharmacology fields in a way that the scientists alone could not have achieved.

The current pandemic (Covid-19) is sadly showing us that we are not ready in the discovery and development of drugs in a state of global emergency, but I firmly believe that in a near future, potent and selective drugs will be available for patients in a much shorter time scale, especially when compared to the currently FDA approved fragment-derived drugs.

Chapter 9

Bibliography

1. Erlanson, D. A., Fesik, S. W., Hubbard, R. E., Jahnke, W. & Jhoti, H. Twenty years on: The impact of fragments on drug discovery. *Nat. Rev. Drug Discov.* **15**, 605–619 (2016).
2. Rees, D. Fragment-based Drug Discovery: Lessons and Outlook . Edited by Daniel A. Erlanson and Wolfgang Jahnke; Series Editors: Raimund Mannhold, Hugo Kubinyi, and Gerd Folkers. *ChemMedChem* **11**, 1667 (2016).
3. Klon, A. E. *Fragment-Based Methods in Drug Discovery*. (Springer, 2015).
4. Singh, M., Tam, B. & Akabayov, B. NMR-fragment based virtual screening: A brief overview. *Molecules* **23**, 233 (2018).
5. Congreve, M., Carr, R., Murray, C. & Jhoti, H. A 'Rule of Three' for fragment-based lead discovery? *Drug Discov. Today* **8**, 876–877 (2003).
6. Mercier, K. A. & Powers, R. Determining the optimal size of small molecule mixtures for high throughput NMR screening. *J Biomol NMR* **31**, 243–258 (2005).
7. Reymond, J. L., Van Deursen, R., Blum, L. C. & Ruddigkeit, L. Chemical space as a source for new drugs. *MedChemComm* (2010) doi:10.1039/c0md00020e.
8. Hughes, T. V., Baldwin, I. & Churcher, I. Fragment-based Drug Discovery - From Hit Discovery to FDA approval: Lessons Learned and Future Challenges. *Int. Drug Discov.* 34–39 (2011).
9. A decade of drug-likeness. *Nat. Rev. Drug Discov.* **6**, (2007).
10. Vinogradova, O. & Qin, J. NMR as a unique tool in assessment and complex determination of weak protein-protein interactions. *Top. Curr. Chem.* **326**, 35–45 (2012).
11. Ortega-Roldan, J. L. *et al.* Accurate characterization of weak macromolecular interactions by titration of NMR residual dipolar couplings: Application to the CD2AP SH3-C:ubiquitin complex. *Nucleic Acids Res.* **37**, e70 (2009).
12. Deeks, E. D. Venetoclax: First Global Approval. *Drugs* **76**, 979–987 (2016).

13. FDA grants accelerated approval to erdafitinib for metastatic urothelial carcinoma. *Case Med. Res.* (2019) doi:10.31525/fda1-ucm635910.htm.

14. Murray, C. W., Verdonk, M. L. & Rees, D. C. Experiences in fragment-based drug discovery. *Trends Pharmacol Sci* **33**, 224–232 (2012).

15. D. A. Erlanson. Fragments in the clinic: 2018 edition. http://practicalfragments.blogspot.com/2018/10/fragments-in-clinic-2018-edition.html (2018).

16. Ciulli, A. & Abell, C. Fragment-based approaches to enzyme inhibition. *Curr. Opin. Biotechnol.* **18**, 489–496 (2007).

17. Campos-Olivas, R. NMR screening and hit validation in fragment based drug discovery. *Curr Top Med Chem* **11**, 43–67 (2011).

18. Dalvit, C. *et al.* High-throughput NMR-based screening with competition binding experiments. *J. Am. Chem. Soc.* **124**, 7702–7709 (2002).

19. Williamson, M. P. Using chemical shift perturbation to characterise ligand binding. *Prog. Nucl. Magn. Reson. Spectrosc.* **73**, 1–16 (2013).

20. Erlanson, D. A. Personal essay: Fragments in the blogosphere. *RSC Drug Discov. Ser.* (2015) doi:10.1039/9781782620938-fp019.

21. Leone, V., Marinelli, F., Carloni, P. & Parrinello, M. Targeting biomolecular flexibility with metadynamics. *Curr. Opin. Struct. Biol.* **20**, 148–154 (2010).

22. Teague, S. J. Implications of protein flexibility for drug discovery. *Nat. Rev. Drug Discov.* **2**, 527–541 (2003).

23. Zheng, H. *et al.* X-ray crystallography over the past decade for novel drug discovery -where are we heading next? *Expert Opin. Drug Discov.* **10**, 975–989 (2015).

24. Chessari, G. & Woodhead, A. J. From fragment to clinical candidate-a historical perspective. *Drug Discov. Today* **14**, 668–675 (2009).

25. Sheng, C. & Zhang, W. Fragment Informatics and Computational Fragment-Based Drug Design: An Overview and Update. *Med. Res. Rev.* **33**, 554–598 (2013).

26. D. A. Erlanson. Practical Fragments blog.

http://practicalfragments.blogspot.com.

27. D. A. Erlanson. Fragments in the clinic: 2015 edition. http://practicalfragments.blogspot.com/2015/01/fragments-in-clinic-2015-edition.html (2015).

28. Skinner, S. P. *et al.* CcpNmr AnalysisAssign: a flexible platform for integrated NMR analysis. *J. Biomol. NMR* **66**, 111–124 (2016).

29. DeLano, W. L. The PyMOL Molecular Graphics System, Version 1.8. *Schrödinger LLC* http://www.pymol.org (2014) doi:10.1038/hr.2014.17.

30. O'Boyle, N. M. *et al.* Open Babel: An Open chemical toolbox. *J. Cheminform.* **3**, (2011).

31. The Open Babel Package. http://openbabel.org.

32. McKinney, W. pandas: a foundational Python library for data analysis and statistics. *Python High Perform. Sci. Comput.* 1–9 (2011).

33. Van Der Walt, S., Colbert, S. C. & Varoquaux, G. The NumPy array: A structure for efficient numerical computation. *Comput. Sci. Eng.* **13**, 22–30 (2011).

34. Taschini, S. Interval Arithmetic: Python Implementation and Applications. *Proc. 7th Python Sci. Conf. (ScyPy 2008)* (2008).

35. Bajusz, D., Rácz, A. & Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminform.* **7**, (2015).

36. Vassar, R. *et al.* β-Secretase cleavage of Alzheimer's amyloid precursor protein by the transmembrane aspartic protease BACE. *Science (80-.).* **286**, 735–741 (1999).

37. Venugopal, C., Demos, C., Jagannatha Rao, K., Pappolla, M. & Sambamurti, K. Beta-Secretase: Structure, Function, and Evolution. *CNS Neurol. Disord. - Drug Targets* **7**, 278–294 (2008).

38. Vassar, R., Kovacs, D. M., Yan, R. & Wong, P. C. The β-secretase enzyme BACE in health and Alzheimer's disease: Regulation, cell biology, function, and therapeutic potential. *J. Neurosci.* **29**, 12787–12794 (2009).

39. Erlanson, Daniel A.; Janke, W. BACE Inhibitors. in *Fragment-based Drug Discovery, lessons and outlook* 333–338 (2016).

40. Ghosh, A. K., Brindisi, M. & Tang, J. Developing β-secretase inhibitors for treatment of Alzheimer's disease. *J. Neurochem.* **1**, 71–83 (2012).

41. Goyal, M. *et al.* Development of dual inhibitors against Alzheimer's disease using fragment-based QSAR and molecular docking. *Biomed Res. Int.* (2014) doi:10.1155/2014/979606.

42. Hung, S. Y. & Fu, W. M. Drug candidates in clinical trials for Alzheimer's disease. *J. Biomed. Sci.* **24**, 47 (2017).

43. Geschwindner, S. *et al.* Discovery of a novel warhead against β-secretase through fragment-based lead generation. *J. Med. Chem.* **50**, 5903–5911 (2007).

44. Murray, C. W. *et al.* Application of fragment screening by X-ray crystallography to β-secretase. *J. Med. Chem.* **50**, 1116–1123 (2007).

45. Edwards, P. D. *et al.* Application of fragment-based lead generation to the discovery of novel, cyclic amidine β-secretase inhibitors with nanomolar potency, cellular activity, and high ligand efficiency. *J. Med. Chem.* **50**, 5912–5925 (2007).

46. Berg, S. *et al.* Design and synthesis of β-site amyloid precursor protein cleaving enzyme (BACE1) inhibitors with in vivo brain reduction of β-amyloid peptides. *J. Med. Chem.* **55**, 9346–9361 (2012).

47. Fesik, S. W., Shuker, S. B., Hajduk, P. J. & Meadows, R. P. SAR by NMR: an NMR-based approach for drug discovery. *Protein Eng.* **10**, 73 (1997).

48. Shuker, S. B., Hajduk, P. J., Meadows, R. P. & Fesik, S. W. Discovering High-Affinity Ligands for Proteins: SAR by NMR. *Science (80-.).* **274**, 1531–1534 (1996).

49. Souers, A. J. *et al.* ABT-199, a potent and selective BCL-2 inhibitor, achieves antitumor activity while sparing platelets. *Nat. Med.* **19**, 202–208 (2013).

50. Choudhary, G. S. *et al.* MCL-1 and BCL-xL-dependent resistance to the

BCL-2 inhibitor ABT-199 can be overcome by preventing PI3K/AKT/mTOR activation in lymphoid malignancies. *Cell Death Dis.* **6**, e1593 (2015).

51. Kelekar, A. & Thompson, C. B. Bcl-2-family proteins: The role of the BH3 domain in apoptosis. *Trends Cell Biol.* **8**, 324–330 (1998).
52. Letai, A. *et al.* Distinct BH3 domains either sensitize or activate mitochondrial apoptosis, serving as prototype cancer therapeutics. *Cancer Cell* **2**, 183–192 (2002).
53. Petros, A. M. *et al.* Discovery of a Potent Inhibitor of the Antiapoptotic Protein Bcl-xL from NMR and Parallel Synthesis. *J. Med. Chem.* **49**, 656–663 (2006).
54. Oltersdorf, T. *et al.* An inhibitor of Bcl-2 family proteins induces regression of solid tumours. *Nature* **435**, 677–681 (2005).
55. Vandenberg, C. J. & Cory, S. ABT-199, a new Bcl-2–specific BH3 mimetic, has in vivo efficacy against aggressive Myc-driven mouse lymphomas without provoking thrombocytopenia. *Blood* **131**, 2285–2288 (2013).
56. Lee, E. F. *et al.* Crystal structure of ABT-737 complexed with Bcl-xL: Implications for selectivity of antagonists of the Bcl-2 family [1]. *Cell Death Differ.* **14**, 1711–1713 (2007).
57. Szlávik, Z. *et al.* Structure-guided discovery of a selective mcl-1 inhibitor with cellular activity. *J. Med. Chem.* **62**, 6913–6924 (2019).
58. Maragno, A. L. *et al.* Abstract 4482: S64315 (MIK665) is a potent and selective Mcl1 inhibitor with strong antitumor activity across a diverse range of hematologic tumor models. (2019) doi:10.1158/1538-7445.am2019-4482.
59. Albershardt, T. C. *et al.* Multiple BH3 mimetics antagonize antiapoptotic MCL1 protein by inducing the endoplasmic reticulum stress response and up-regulating BH3-only protein NOXA. *J. Biol. Chem.* **286**, 24882–24895 (2011).
60. Soderquist, R. S. & Eastman, A. BCL2 inhibitors as anticancer drugs: A plethora of misleading BH3 mimetics. *Mol. Cancer Ther.* **15**, 2011–2017

(2016).

61. Hubbard, R. . E. *Fragment-based Drug Discovery Lessons and Outlook*. *Fragment-based Drug Discovery: Lessons and Outlook* (2016). doi:10.1002/9783527683604.

62. Erlanson, D. A. Introduction to fragment-based drug discovery. *Top. Curr. Chem.* **317**, 1–32 (2012).

63. Flaherty, K. T., Yasothan, U. & Kirkpatrick, P. Vemurafenib. *Nat. Rev. Drug Discov.* **10**, 811–812 (2011).

64. Perera, T. P. S. *et al.* Discovery & pharmacological characterization of JNJ-42756493 (Erdafitinib), a functionally selective small-molecule FGFR family inhibitor. *Mol. Cancer Ther.* **16**, 1010–1020 (2017).

65. Pellecchia, M. NMR spectroscopy in fragment based drug design. *Top. Med. Chem.* **5**, 125–139 (2010).

66. Davis, B. J. & Erlanson, D. A. Learning from our mistakes: The 'unknown knowns' in fragment screening. *Bioorganic Med. Chem. Lett.* **23**, 2844–2852 (2013).

67. Mercier, K. A., Shortridge, M. D. & Powers, R. A Multi-Step NMR Screen for the Identification and Evaluation of Chemical Leads for Drug Discovery. *Comb. Chem. High Throughput Screen.* **12**, 285–295 (2009).

68. Baell, J. B. & Nissink, J. W. M. Seven Year Itch: Pan-Assay Interference Compounds (PAINS) in 2017 - Utility and Limitations. *ACS Chem. Biol.* **13**, 36–44 (2018).

69. Capuzzi, S. J., Muratov, E. N. & Tropsha, A. Phantom PAINS: Problems with the Utility of Alerts for P an- Assay in terference Compound S. *J. Chem. Inf. Model.* **57**, 417–427 (2017).

70. Ayotte, Y. *et al.* Exposing Small-Molecule Nanoentities by a Nuclear Magnetic Resonance Relaxation Assay. *J. Med. Chem.* **62**, 7885–7896 (2019).

71. Meyer, B. *et al.* Saturation transfer difference NMR spectroscopy for identifying ligand epitopes and binding specificities. *Ernst Schering Res.*

Found. Workshop **44**, 149–167 (2004).

72. Cala, O. & Krimm, I. Ligand-orientation based fragment selection in STD NMR screening. *J. Med. Chem.* **58**, 8739–8742 (2015).

73. Raingeval, C. *et al.* 1D NMR WaterLOGSY as an efficient method for fragment-based lead discovery. *J. Enzyme Inhib. Med. Chem.* **34**, 1218–1225 (2019).

74. Yan, R. Stepping closer to treating Alzheimer's disease patients with BACE1 inhibitor drugs. *Transl. Neurodegener.* **5**, 1–11 (2016).

75. Chen, Y. C. Beware of docking! *Trends Pharmacol. Sci.* **36**, 78–95 (2015).

76. Gabel, J., Desaphy, J. & Rognan, D. Beware of machine learning-based scoring functions-on the danger of developing black boxes. *J. Chem. Inf. Model.* **54**, 2807–2815 (2014).

77. Uçar, M. K., Nour, M., Sindi, H. & Polat, K. The Effect of Training and Testing Process on Machine Learning in Biomedical Datasets. *Math. Probl. Eng.* (2020) doi:10.1155/2020/2836236.

78. Bienstock J., R. Overview: Fragment-Based Drug Design. *Libr. Des. Search Methods, Appl. Fragm. Drug Des.* **1076**, 1–26 (2011).

79. Schrödinger. Maestro l Schrödinger. *Schrödinger Release 2018-1* (2018).

80. Inc., C. C. G. Molecular Operating Environment (MOE), 2015.01. *1010 Sherbooke St.West, Suite #910, Montreal, QC, Canada, H3A 2R7* (2015).

81. Dey, F. & Caflisch, A. Fragment-based de Novo Ligand design by multiobjective evolutionary optimization. *J. Chem. Inf. Model.* **48**, 679–690 (2008).

82. Durrant, J. D., Lindert, S. & McCammon, J. A. AutoGrow 3.0: An improved algorithm for chemically tractable, semi-automated protein inhibitor design. *J. Mol. Graph. Model.* **44**, 104–112 (2013).

83. Markham, A. Erdafitinib: First Global Approval. *Drugs* **79**, 1017–1021 (2019).

84. Jeppsson, F. *et al.* Discovery of AZD3839, a potent and selective BACE1 inhibitor clinical candidate for the treatment of alzheimer disease. *J. Biol.*

Chem. **287**, 41245–41257 (2012).

85. Petros, A. M. *et al.* Discovery of a potent and selective Bcl-2 inhibitor using SAR by NMR. *Bioorganic Med. Chem. Lett.* **20**, 6587–6591 (2010).

86. Dias, D. M. & Ciulli, A. NMR approaches in structure-based lead discovery: Recent developments and new frontiers for targeting multi-protein complexes. *Prog. Biophys. Mol. Biol.* **116**, 101–112 (2014).

87. Mayer, M. & Meyer, B. Characterization of ligand binding by saturation transfer difference NMR spectroscopy. *Angew. Chemie - Int. Ed.* **38**, 1784–1788 (1999).

88. Dalvit, C. *et al.* Identification of compounds with binding affinity to proteins via magnetization transfer from bulk water. *J. Biomol. NMR* **18**, 65–68 (2000).

89. Vanwetswinkel, S. *et al.* TINS, target immobilized NMR screening: An efficient and sensitive method for ligand discovery. *Chem. Biol.* **12**, 207–216 (2005).

90. Jahnke, W. Spin labels as a tool to identify and characterize protein-ligand interactions by NMR spectroscopy. *Chembiochem* **3**, 167–73 (2002).

91. Guan, J. Y. *et al.* Small-molecule binding sites on proteins established by paramagnetic NMR spectroscopy. *J. Am. Chem. Soc.* **135**, 5859–5868 (2013).

92. Dalvit, C. & Vulpetti, A. Technical and practical aspects of 19F NMR-based screening: Toward sensitive high-throughput screening with rapid deconvolution. *Magn. Reson. Chem.* **50**, 592–597 (2012).

93. Sugiki, T., Furuita, K., Fujiwara, T. & Kojima, C. Current NMR techniques for structure-based drug discovery. *Molecules* **23**, 148 (2018).

94. Stark, J. L., Eghbalnia, H. R., Lee, W., Westler, W. M. & Markley, J. L. NMRmix: A Tool for the Optimization of Compound Mixtures in 1D 1H NMR Ligand Affinity Screens. *J. Proteome Res.* **15**, 1360–1368 (2016).

95. Ulrich, E. L. *et al.* BioMagResBank. *Nucleic Acids Res.* **36**, D402-408 (2008).

96. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

97. Lam, S. K., Pitrou, A. & Seibert, S. Numba: A LLVM-based python JIT compiler. *Proc. Second Work. LLVM Compil. Infrastruct. HPC - LLVM '15* **7**, 1–6 (2015).

98. Lattner, C. A. LLVM : An Infrastructure for Multi-Stage Optimization. *University of Illinois at Urbana-Champaign* (2002).

99. Summerfield, M. *Rapid Gui Programming with Python and Qt: The Definite Guid to PyQt Programming*. *Rapid GUI Programming with Python and Qt:* (2008). doi:10.1093/infdis/jit776.

100. Campagnola, L. PyQtGraph. http://www.pyqtgraph.org.

101. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).

102. Waskom, M., Olga, B. & Botvinnik, A. Seaborn. (2017).

103. Mujica, L. E., Rodellar, J., Fernández, A. & Güemes, A. Q-statistic and t2-statistic pca-based measures for damage assessment in structures. *Struct. Heal. Monit.* **10**, 539–553 (2011).

104. JetBrains. PyCharm. (2020).

105. Stoyanova, R. & Brown, T. R. NMR spectral quantitation by principal component analysis. *NMR Biomed.* **154**, 163–175 (2001).

106. Lattner, C. & Adve, V. The LLVM compiler framework and infrastructure tutorial. in *Lecture Notes in Computer Science* (2005). doi:10.1007/11532378_2.

107. Erlanson, D. A. Fragment-based lead discovery: a chemical update. *Curr. Opin. Biotechnol.* **17**, 643–652 (2006).

108. Galarnyk, M. Understanding Boxplots. *towardsdatascience.com* https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51 (2018).

109. Dalvit, C. Theoretical analysis of the competition ligand-based NMR experiments and selected applications to fragment screening and binding

constant measurements. *Concepts Magn. Reson. Part A Bridg. Educ. Res.* **32**, 341–372 (2008).

110. Bhunia, A., Bhattacharjya, S. & Chatterjee, S. Applications of saturation transfer difference NMR in biological systems. *Drug Discov. Today* **17**, 505–513 (2012).

111. Schade, M. Fragment-Based Lead Discovery by NMR. *Front. Drug Des. Discov.* **3**, 105–119 (2012).

112. Begley, D. W., Moen, S. O., Pierce, P. G. & Zartler, E. R. Saturation transfer difference NMR for fragment screening. *Curr. Protoc. Chem. Biol.* **5**, 251–268 (2013).

113. Lepre, C. A., Moore, J. M. & Peng, J. W. Theory and applications of NMR-based screening in pharmaceutical research. *Chem. Rev.* **104**, 3641–3676 (2004).

114. Dalvit, C., Fogliatto, G., Stewart, A., Veronesi, M. & Stockman, B. WaterLOGSY as a method for primary NMR screening: Practical aspects and range of applicability. *J. Biomol. NMR* **21**, 349–359 (2001).

115. Antanasijevic, A., Ramirez, B. & Caffrey, M. Comparison of the sensitivities of WaterLOGSY and saturation transfer difference NMR experiments. *J. Biomol. NMR* **60**, 37–44 (2014).

116. Bruker TopSpin Fragment-based Screening (FBS) tool. https://www.bruker.com/products/mr/nmr/software/fragment-based-screening-with-nmr.html.

117. Perez, M. Using Mnova Screen to Process, Analyze and Report Ligand-Protein Binding Spectra for Fragment-based Lead Design. *Mestrelab Research* (2014).

118. Helmus, J. J. & Jaroniec, C. P. Nmrglue: An open source Python package for the analysis of multidimensional NMR data. *J. Biomol. NMR* **55**, (2013).

119. van Rossum, G. & Drake, F. L. *The Python Language Reference Manual*. *Linux J.* (2009). doi:10.1159/0000113495.

120. Stevens, T. J. & Boucher, W. *Python Programming for Biology*. *Python*

Programming for Biology (2015). doi:10.1017/cbo9780511843556.

121. Billauer, E. Peak Detect. http://billauer.co.il/peakdet.html (2012).

122. Viegas, A., Manso, J., Nobrega, F. L. & Cabrita, E. J. Saturation-transfer difference (STD) NMR: A simple and fast method for ligand screening and characterization of protein binding. *J. Chem. Educ.* **88**, 990–994 (2011).

123. Martin, R. C. *Clean Code - A Handbook of Agile Software Craftmanship*. *Igarss 2014* (2014). doi:10.1007/s13398-014-0173-7.2.

124. Martin, R. C. The Clean Code Blog. (2012).

125. Louie, J. Uncle Bob's Clean Architecture. https://jameslouiecs.blogspot.com/2018/11/uncle-bobs-clean-architecture.html (2018).

126. Chu, S. & Gochin, M. Identification of fragments targeting an alternative pocket on HIV-1 gp41 by NMR screening and similarity searching. *Bioorganic Med. Chem. Lett.* **23**, 5114–5118 (2013).

127. Hajian-Tilaki, K. Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. *Casp. J. Intern. Med.* **4**, 627–635 (2013).

128. Xi, Y. & Rocke, D. M. Baseline correction for NMR spectroscopic metabolomics data analysis. *BMC Bioinformatics* **9**, (2008).

129. Bartels, C., Güntert, P. & Wüthrich, K. IFLAT-A New Automatic Baseline-Correction Method for Multidimensional NMR Spectra with Strong Solvent Signals. *J. Magn. Reson. Ser. A* **117**, 330–333 (1995).

130. Marion, D. & Bax, A. Baseline distortion in real-fourier-transform NMR spectra. *J. Magn. Reson.* **79**, 352–356 (1988).

131. Forezi, L. S. M. & Castelo-Branco, F. S. Editing NMR spectra with MestReNova software: A practical guide. *Rev. Virtual Quim.* **9**, 2650–2672 (2017).

132. Eilers, P. H. C. & Boelens, H. F. M. Baseline Correction with Asymmetric Least Squares Smoothing. *Life Sci.* **75**, 3631–3636 (2005).

133. Zhang, Z.-M., Chen, S. & Liang, Y.-Z. Baseline correction using adaptive

iteratively reweighted penalized least squares. *Analyst* **135**, 1138 (2010).

134. Baek, S.-J., Park, A., Ahn, Y.-J. & Choo, J. Baseline correction using asymmetrically reweighted penalized least squares smoothing. *Analyst* **140**, 250–257 (2015).

135. Wang, K. C., Wang, S. Y., Kuo, C. H. & Tseng, Y. J. Distribution-based classification method for baseline correction of metabolomic 1D proton nuclear magnetic resonance spectra. *Anal. Chem.* **85**, 1231–1239 (2013).

136. Gooch, J. W. Cross Correlation. in *Encyclopedic Dictionary of Polymers* (2011). doi:10.1007/978-1-4419-6247-8_15207.

137. Zar, J. H. Spearman Rank Correlation. in *Encyclopedia of Biostatistics* (2005). doi:10.1002/0470011815.b2a15150.

138. James Lani. Correlation (Pearson, Kendall, Spearman). *Stat. Solut.* (2010).

139. Standard Probability and Statistics Tables and Formulae. *Technometrics* **43**, 249–250 (2001).

140. ACD/Labs. ACD/NMR Workbook, version 2019, Advanced Chemistry Development, Inc., Toronto, ON, Canada, www.acdlabs.com, 2019. (2019).

141. Peng, C. *et al.* Fast and Efficient Fragment-Based Lead Generation by Fully Automated Processing and Analysis of Ligand-Observed NMR Binding Data. *J. Med. Chem.* **59**, 3303–3310 (2016).

142. Morris G.M. & Dallakyan S. AutoDock — AutoDock. *02-27* (2013).

143. Durrant, J. D. & McCammon, J. A. NNScore 2.0: A neural-network receptor-ligand scoring function. *J. Chem. Inf. Model.* **51**, 2897–2903 (2011).

144. Kawatkar, S., Wang, H., Czerminski, R. & Joseph-McCarthy, D. Virtual fragment screening: An exploration of various docking and scoring protocols for fragments using Glide. *J. Comput. Aided. Mol. Des.* **23**, 527–539 (2009).

145. Huang, R., Bonnichon, A., Claridge, T. D. W. & Leung, I. K. H. Protein-ligand binding affinity determination by the waterLOGSY method: An optimised approach considering ligand rebinding. *Sci. Rep.* (2017)

doi:10.1038/srep43727.

146. Maity, S., Gundampati, R. K. & Kumar, T. K. S. NMR methods to characterize protein-ligand interactions. *Natural Product Communications* (2019) doi:10.1177/1934578X19849296.

147. Campos-Olivas, R. NMR Screening and Hit Validation in Fragment Based Drug Discovery. *Curr. Top. Med. Chem.* (2010) doi:10.2174/156802611793611887.

148. Breukels, V. *et al.* Overview on the use of NMR to examine protein structure. *Curr Protoc Protein Sci* **64**, 17.5.2-17.5.44 (2011).

149. Kwan, A. H., Mobli, M., Gooley, P. R., King, G. F. & MacKay, J. P. Macromolecular NMR spectroscopy for the non-spectroscopist. *FEBS J.* **278**, 687–703 (2011).

150. Bodenhausen, G. & Ruben, D. J. Natural abundance nitrogen-15 NMR by enhanced heteronuclear spectroscopy. *Chem. Phys. Lett.* **69**, 185–189 (1980).

151. Pervushin, K., Riek, R., Wider, G. & Wuthrich, K. Attenuated T2 relaxation by mutual cancellation of dipole-dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution. *Proc. Natl. Acad. Sci.* **94**, 12366–12371 (1997).

152. Würz, J. M., Kazemi, S., Schmidt, E., Bagaria, A. & Güntert, P. NMR-based automated protein structure determination. *Arch. Biochem. Biophys.* **628**, 24–32 (2017).

153. Lee, W., Tonelli, M. & Markley, J. L. NMRFAM-SPARKY: Enhanced software for biomolecular NMR spectroscopy. *Bioinformatics* **31**, 1325–7 (2015).

154. Vranken, W. F. *et al.* The CCPN data model for NMR spectroscopy: Development of a software pipeline. *Proteins Struct. Funct. Genet.* **59**, 687–696 (2005).

155. Johnson, B. A. & Blevins, R. A. NMR View: A computer program for the

visualization and analysis of NMR data. *J. Biomol. NMR* **4**, 603–614 (1994).

156. Keller, R. *The computer aided resonance assignment tutorial*. *Goldau, Switzerland* (2004).

157. Feracci, M. *et al.* Structural basis of RNA recognition and dimerization by the STAR proteins T-STAR and Sam68. *Nat. Commun.* **7**, (2016).

158. Foot, J. N., Feracci, M. & Dominguez, C. Screening protein - Single stranded RNA complexes by NMR spectroscopy for structure determination. *Methods* **65**, 288–301 (2014).

159. Ayed, A. *et al.* Latent and active p53 are identical in conformation. *Nat. Struct. Biol.* **8**, 756–60 (2001).

160. Vyas, V. K., Ukawala, R. D., Ghate, M. & Chintha, C. Homology Modeling a Fast Tool for Drug Discovery: Current Perspectives. *Indian J. Pharm. Sci.* **1**, 1–17 (2012).

161. Dominguez, C., Boelens, R. & Bonvin, A. M. J. J. HADDOCK: A protein-protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.* **125**, 1731–7 (2003).

162. Gutmanas, A. *et al.* NMR Exchange Format: a unified and open standard for representation of NMR restraint data. *Nat. Struct. Mol. Biol.* **22**, 433–434 (2015).

163. Teixeira, J. M. C., Skinner, S. P., Arbesú, M., Breeze, A. L. & Pons, M. Farseer-NMR: automatic treatment, analysis and plotting of large, multi-variable NMR data. *J. Biomol. NMR* **71**, 1–9 (2018).

164. Krishnamoorthy, J., Yu, V. C. K. & Mok, Y. K. Auto-FACE: An NMR based binding site mapping program for fast chemical exchange protein-ligand systems. *PLoS One* **5**, e8943 (2010).

165. Waudby, C. A., Ramos, A., Cabrita, L. D. & Christodoulou, J. Two-Dimensional NMR Lineshape Analysis. *Sci. Rep.* **6**, 24826 (2016).

166. Gardiennet, C. *et al.* Solid-state NMR chemical-shift perturbations indicate domain reorientation of the DnaG primase in the primosome of Helicobacter pylori. *J. Biomol. NMR* **64**, 189–195 (2016).

167. Verardi, R., Traaseth, N. J., Masterson, L. R., Vostrikov, V. V. & Veglia, G. Isotope labeling for solution and solid-state NMR spectroscopy of membrane proteins. *Adv. Exp. Med. Biol.* **992**, 35–62 (2012).

168. Takeuchi, K., Arthanari, H., Shimada, I. & Wagner, G. Nitrogen detected TROSY at high field yields high resolution and sensitivity for protein NMR. *J. Biomol. NMR* **63**, 323–331 (2015).

169. Papadopoulos, E. *et al.* Structure of the eukaryotic translation initiation factor eIF4E in complex with 4EGI-1 reveals an allosteric mechanism for dissociating eIF4G. *Proc. Natl. Acad. Sci. U. S. A.* **111**, e3187-3195 (2014).

170. Silvera, D., Formenti, S. C. & Schneider, R. J. Translational control in cancer. *Nat. Rev. Cancer* **10**, 254–266 (2010).

171. Grzmil, M. & Hemmings, B. A. Translation regulation as a therapeutic target in cancer. *Cancer Res.* **72**, 3891–3890 (2012).

172. Obayashi, E. *et al.* Molecular Landscape of the Ribosome Pre-initiation Complex during mRNA Scanning: Structural Role for eIF3c and Its Control by eIF5. *Cell Rep.* **18**, 2651–2663 (2017).

173. Merrick, W. C. Cap-dependent and cap-independent translation in eukaryotic systems. *Gene* **332**, 1–11 (2004).

174. Joshi, S. Mnk kinase pathway: Cellular functions and biological outcomes. *World J. Biol. Chem.* **5**, 321–333 (2014).

175. De Benedetti, A. & Rhoads, R. E. Overexpression of eukaryotic protein synthesis initiation factor 4E in HeLa cells results in aberrant growth and morphology. *Proc. Natl. Acad. Sci. U. S. A.* **87**, 8212–8216 (1990).

176. Lu, C., Makala, L., Wu, D. & Cai, Y. Targeting translation: EIF4E as an emerging anticancer drug target. *Expert Rev. Mol. Med.* **18**, e2 (2016).

177. Thompson, P. A. *et al.* Abstract 2698: eFT226, a potent and selective inhibitor of eIF4A, is efficacious in preclinical models of lymphoma. **2698**, 2698–2698 (2019).

178. Van Zundert, G. C. P. *et al.* The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *J. Mol. Biol.* **428**, 720–

725 (2016).

179. Jiménez, J., Doerr, S., Martínez-Rosell, G., Rose, A. S. & De Fabritiis, G. DeepSite: Protein-binding site predictor using 3D-convolutional neural networks. *Bioinformatics* **33**, 3036–3042 (2017).

180. Sterling, T. & Irwin, J. J. ZINC 15 - Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **55**, 2324–2337 (2015).

181. OLEG TROTT, A. J. O. & Schroer, A. AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J. Comput. Chem.* **31**, 455–461 (2010).

182. Raschka, S., Wolf, A. J., Bemister-Buffington, J. & Kuhn, L. A. Protein–ligand interfaces are polarized: discovery of a strong trend for intermolecular hydrogen bonds to favor donors on the protein side with implications for predicting and designing ligand complexes. *J. Comput. Aided. Mol. Des.* **32**, 511–528 (2018).

183. Durrant, J. D. & McCammon, J. A. Autoclickchem: Click chemistry in silico. *PLoS Comput. Biol.* **8**, e1002397 (2012).

184. Raschka, S. BioPandas: Working with molecular structures in pandas DataFrames. *J. Open Source Softw.* **2**, (2017).

185. Durrant, J. D. & McCammon, J. A. BINANA: A novel algorithm for ligand-binding characterization. *J. Mol. Graph. Model.* (2011) doi:10.1016/j.jmgm.2011.01.004.

186. Brown, T. ChemDraw. *Sci. Teach.* (2014).

187. Harder, E. *et al.* OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins. *J. Chem. Theory Comput.* **12**, 281–296 (2016).

188. Martínez-Rosell, G., Giorgino, T. & De Fabritiis, G. PlayMolecule ProteinPrepare: A Web Application for Protein Preparation for Molecular Dynamics Simulations. *J. Chem. Inf. Model.* **57**, 1511–1516 (2017).

189. Mureddu, L. & Vuister, G. W. Simple high-resolution NMR spectroscopy as a tool in molecular biology. *FEBS J.* **286**, 2035–2042 (2019).