

Andreia Espírito Santo Cunha

Towards the discovery of novel anti-Influenza agents using a reverse virtual fragment screening approach

Mestrado em Química Medicinal

Departamento de Química

FCTUC

Setembro 2017



UNIVERSIDADE DE COIMBRA

Andreia Espírito Santo Cunha

**Towards the discovery of novel anti-Influenza
agents using a reverse virtual fragment
screening approach**

**Dissertação apresentada para provas de Mestrado em Química
Medicinal**

Orientador: Prof. Doutor Rui M. M. Brito

Co-orientador: Doutor Carlos J. V. Simões

Setembro 2017

Universidade de Coimbra

“I'd take the awe of understanding over the awe of ignorance any day.”

Douglas Adams

Acknowledgments

I would like to thank all the people who contributed in some way to the work described in this thesis. First, I thank my academic advisor, Prof. Dr. Rui M. M. Brito, for the intellectual freedom, scientific guidance and the possibility to perform the present work. Additionally, I thank Dr. Carlos Simões for all the guidance and knowledge transmitted during this last year.

The experimental results described in this thesis were accomplished with the help and support of fellow lab-mates. Zaida Almeida, I thank you for your scientific insight and the help in the different stages of the project. Pedro Cruz, thank you for sharing your knowledge and always being available to help me during this project, especially regarding NMR.

To Cristiana Pires, my classmate, thank you for your support, humour, company and exchange of ideas, because two heads are better than one!

To my friends, thank you for your support, for never leave even when I asked to, for all the laughs that scares away all the frustrations and tears.

Last, but not least, a special thanks to my family, because without them I wouldn't get this far. To my big brother, thank you for your support and jokes, that always lighten up every weekend, thank you for believing in me every step of my way that lead here. To my parents, thank you for the investment, trust and support and for the values that you teach me, making the person I am today.

Table of contents

Acknowledgments	5
Table of figures.....	9
Index of tables	11
Abbreviations.....	13
Resumo.....	15
Abstract.....	17
1. General Introduction.....	19
1.1. Influenza A Virus	20
1.1.1.Classification	20
1.1.2.Genome and Structure.....	20
1.1.3.Replication Cycle	21
1.1.3.1.Virus entry into the host cell.....	22
1.1.3.2.Entry of vRNPs into the nucleus.....	23
1.1.3.3.Transcription and replication of the viral genome	23
1.1.3.4.Assembly and budding of the viral proteins	24
1.2. Antiviral Drugs	24
1.2.1. Inhibitors of haemagglutinin.....	24
1.2.2.M2 ion channel inhibitors.....	25
1.2.3. Inhibitors of viral RNA polymerase.....	25
1.2.4.Inhibitors of neuraminidase.....	26
1.3. NS1 as a therapeutic target.....	28
2. Aims and Project Workflow.....	28
3. Target Identification and Characterization.....	32
3.1. Introduction.....	32
3.1.1. NS1 structural description	32
3.1.2. Structure-based drug design.....	34
3.1.3. Hot Spot Analysis	35
3.1.4. Experimental techniques for protein characterization.....	37
3.1.4.1.DSC.....	37
3.1.4.2.CD	37
3.1.4.3.SEC-MALS	39
3.1.4.4.NMR.....	40
3.1.4.5.Fluorescence spectroscopy.....	43
3.2. Methods	44
3.2.1.Structural information retrieval and quality assessment	44
3.2.2. Binding site identification and druggability assessment.....	44
3.2.3.Experimental characterization of NS1 cloned domains (ED and RBD) ..	45

3.2.3.1. DSC.....	45
3.2.3.2. CD	45
3.2.3.3. SEC-MALS.....	45
3.2.3.4. NMR	46
3.2.3.5. Fluorescence	46
3.3. Results and Discussion.....	46
3.3.1. Structural information retrieval and quality assessment	46
3.3.2. Pocket and subpocket identification and <i>druggability</i> assessment.....	47
3.3.3. Experimental characterization of both domains of NS (ED and RBD) ...	49
3.3.3.1. DSC.....	49
3.3.3.2. CD	49
3.3.3.3. SEC-MALS.....	51
3.3.3.4. NMR	52
3.3.3.5. Fluorescence	55
4. Virtual Fragment Screening.....	57
4.1. Introduction.....	57
4.1.1. Fragment-based lead discovery	57
4.1.2. Computational methods for FBLD	59
4.1.3. An innovative approach to fragment screening based on IsoMIF	61
4.1.4. NMR for FBLD	62
4.2. Methods	64
4.2.1. Assembly of fragment-bound protein screening library.....	64
4.2.2. Fragment screening via binding-site similarity.....	65
4.2.3. Fragment hit confirmation via docking	67
4.2.4. Fragment hit validation via NMR.....	68
4.2.4.1. Experimental procedure for hit validation via NMR	68
4.3. Results and discussion	68
4.3.1. Assembly of fragment-bound protein screening library.....	68
4.3.2. Fragment Screening via binding-site similarity	73
4.3.3. Fragment hit confirmation via docking	75
4.3.3.1. Fragment hit validation via NMR.....	77
5. Conclusions	83
6. Future perspectives.....	85
7. Bibliography	87
8. Annexes.....	91

Table of figures

Figure 1. Schematic structure of influenza A virus ⁸	21
Figure 2. Schematic diagram of the influenza viral life cycle ¹⁰	22
Figure 3. Influenza Ribonucleoprotein Particle (RNP) ¹³	23
Figure 4. Chemical structures of derivatives of adamantane, known as inhibitors of M2 ion channels. A. Amantadine. B. Rimantadine.	25
Figure 5. Commercially available inhibitors of viral RNA polymerase. A. Favipiravir. B. Flutamide.	26
Figure 6. Commercially available inhibitors of Neuraminidase. A. Zanamivir. B. Oseltamivir. C. Peramivir. D. Laninamivir.	27
Figure 7. Workflow of the project presented in this dissertation, with every main step highlighted in blue. The preliminary step of target selection is shown in green.....	31
Figure 8. A. NMR structure of the RNA-binding domain with arginine 38 highlighted (PDB code 2N74). B. Crystallographic structure of the RNA-binding domain bound with dsRNA (PDB code 2ZKO).	32
Figure 9. Crystallography structure of NS1's ED domain. (PDB code 2GX9).	33
Figure 10. Schematic representation of the monomeric form of the structure of NS1 highlighting the RNA-binding domain (RBD, in blue), the Effector domain (ED, in cyan) and the linker region (LR, in yellow). The position of residue 71 is represented by de red bead.	33
Figure 11. Circular Dichroism spectra of proteins with different representative secondary structures.	38
Figure 12. The effect of an external magnetic field B_0 on the orientation of the magnetic moment for a nucleus with $I=1/2$. The nuclei with orientation along the external field have lower energy. The difference in energy between the two energy fields is given by ΔE	41
Figure 13. Jablonski diagram and simplified representation of the fluorescence process. S_0 represents the ground singlet electronic state; S_1 and S_2 are the successively higher energy excited singlet electronic states. T_1 is the lowest energy triplet state.	43
Figure 14. Ramachandran plot for the NMR structure of NS1-RBD (PDB code 2N74; chain A) validated by the WHAT_CHECK/WHAT_IF program.....	47
Figure 15. Druggable pockets of NS1-RBD(PDB code 2N74) identified and analysed by DoGSiteScorer.	49
Figure 16. Differential Scanning Calorimetry of both structural domains of NS1. A - DSC heating curve of the NS1 Effector Domain. The T_m of unfolding of the ED is 49.6°C. B - DSC heating curve of NS1 RNA-binding domain. The T_m of unfolding of RBD is 63.8°C.	49
Figure 17. A. Cartoon representation of the 3D structure of NS1-ED (PDB ID 3RVC). B. Far-UV CD spectra of NS1-ED in PBS buffer, at pH 7.4.....	50
Figure 18. A. Cartoon representation of the 3D structure of NS1-RBD (PDB ID 2N74). B. Far-UV CD spectra of NS1-RBD, in PBS buffer, at pH 7.4.	51
Figure 19. Size-exclusion chromatography coupled to a multi-angle laser light scattering (SEC-MALLS) instrument of the RNA-binding domain of NS1. The chromatogram was run at a flow rate of 0.5 mL/min, in 20 mM sodium phosphate buffer, 100 mM sodium chloride, pH 6.9. Peak D, with an elution of approximately 21 minutes, corresponds to a molecular species a molecular weight of 22.2 kDa, coinciding with RBD in dimeric state and peak M corresponds to a molecular weight of 10.6 kDa, coinciding with RBD in a monomeric form.....	52
Figure 20. ^1H - ^{15}N HSQC spectrum of ^{15}N -labelled NS1-ED. The spectrum was collected at a ^1H frequency of 500 MHz with a 0.26 mM protein sample in PBS, pH 6.9.	53
Figure 21. ^1H - ^{15}N HSQC spectrum of the RBD of NS1 with residue-specific backbone assignments indicated and detailed in Annex A. The spectrum was collected at a ^1H frequency of 600 MHz with a 0.40 mM protein sample in PBS, pH 6.9.	54

Figure 22. ¹ H DOSY spectrum of the RBD-NS1 at 25°C. Spectrum was collected at a ¹ H frequency of 400 MHz with a protein sample at a 0.12 mM concentration.	55
Figure 23. Fluorescence spectra of NS1-RBD in the presence of different percentages of DMSO at pH 6.9. Spectra were collected at 25°C. The RBD concentration was 15 μM, the excitation wavelength was 280 nm and cuvette pathlength was 5x5 mm. Baseline spectra with buffer and various percentages of DMSO were also acquired and subtracted to the raw data.	56
Figure 24. Schematic representation of SAR by NMR ⁷³	64
Figure 25. Scheme of the filtration step in the assembly of the fragment-bound library.....	69
Figure 26. Physicochemical property distribution of the fragments contained in the assembled fragment-bound library.	70
Figure 27. Plot of BEI vs LEI for the 509 fragments in the assembled virtual library.....	71
Figure 28. Distribution of Surface Efficiency Index across the fragments retained in our screening library.	72
Figure 29. Distribution of Ligand Lipophilicity Efficiency across the fragments retained in our screening library.....	73
Figure 30. Surface representation of RBD and molecular interaction fields calculated using MIF software. The colour cyan corresponds to the hydrophobic probe; orange corresponds to the aromatic probe; blue corresponds to the hydrogen bond donor probe; red corresponds to the hydrogen bond acceptor probe; the positive charge probe is represented by the colour green and the negative charge probe is represented by the colour magenta.	73
Figure 31. ¹ H- ¹⁵ N HSQC of free RBD and first (1:1) and second additions (1:2) of 4-hydroxyphenylacetate. Free RBD is represented in red, first addition in green and second addition in grey.	78
Figure 32. Representation of the RBD surface with the subpocket composed by the amino acids Ile 54 and Glu55 represented in dark-blue.	78
Figure 33. Representation of the RBD surface with the subpocket composed by amino acids Ala57 and Arg59 represented in orange.	79
Figure 34. ¹ H- ¹⁵ N HSQC of free RBD and first (1:1) and second additions (1:2) of 7-chloro-3,4-dihydroisoquinolin-1(2H)-one. Free RBD is represented in red, first addition in green and second addition in grey.	80
Figure 35. Representation of the RBD surface with the subpockets in the proximity of amino acids Trp16, Arg35, Leu36, Asp39, Gln40 and Leu43 represented in red and blue....	80

Index of tables

Table 1. Influenza A virus genomic segments and major proteins.....	20
Table 2 . The five major pockets of the RBD (PDB ID 2N74) predicted by DoGSiteScorer.	48
Table 3. Parameters derived from “Rule of three” and used to construct the fragment library.....	65
Table 4. Summary of calculated ligand efficiency indexes, LE and BEI, for the top-5 fragments extracted from protein binding sites holding the highest similarity (Tanimoto score) with the main cavity in NS1-RBD.	74
Table 5. Summary of the calculated ligand efficiency indexes, LE and BEI, for the 5 fragments extracted from protein binding sites holding the lowest similarity (Tanimoto score) with the main cavity in NS1-RBD.	75
Table 6. Docking scores for the top-5 fragments extracted from protein binding sites holding the highest Tanimoto Scores.	76
Table 7. Docking scores for fragments extracted from protein binding sites holding the lowest Tanimoto Scores.	77

Abbreviations

CD – Circular Dichroism

CSP – Chemical Shift Perturbation

DSC – Differential Scanning Calorimetry

dsRNA – double stranded ribonucleic acid

EF – Effector domain

FBLD – Fragment-based lead discovery

HA – Haemagglutinin

HSQC – Heteronuclear Single Quantum Coherence

IFN – Interferon

ITC – Isothermal Titration Calorimetry

K – Lysine

LR – Linker region

M1 – Matrix protein 1

M2 – Matrix protein 2

MIF – Molecular Interaction Fields

mRNA – messenger Ribonucleic Acid

NEP – Nuclear Export Protein

NLS – Nuclear Localization Signals

NMR – Nuclear Magnetic Resonance

NP – Nucleoprotein

NS1 – Non-Structural protein 1

PA – Polymerase acid protein

PB1 – Polymerase basic protein 1

PB1-F2 – Polymerase basic protein 1 F2

PB2 – Polymerase basic protein 2

PBD – Protein Data Bank

PBS – Phosphate Buffered saline

PPR – Pattern Recognition Receptor

PSA – Polar Surface Area

R – Arginine

RBD – Ribonucleic acid binding domain

RNA – Ribonucleic acid

SAR – Structure-Activity Relationship

SEC-MALS – Size Exclusion Chromatography coupled to a Multi-Angle Light Scattering detector

SPR – Surface Plasmon Resonance

ssRNA – single stranded ribonucleic acid

SVM – Support Vector Machines

UV – Ultraviolet

vRNA – viral ribonucleic acid

vRNP – viral ribonucleoprotein particle

Resumo

Os vírus influenza são agentes patogênicos responsáveis por doenças respiratórias – gripe – que afetam a população mundial e caracterizadas por elevada morbidade e significativa mortalidade. As infecções provocadas pelo vírus da gripe podem ser controladas através de vacinação e medicamentos antivirais. No entanto, as vacinas precisam de reformulação e administração anuais e oferecem proteção limitada. O rápido aparecimento de estirpes de influenza resistentes aos fármacos atualmente comercializados contra o vírus influenza A realça a necessidade de desenvolver novas classes de antivirais.

Vários estudos funcionais e estruturais demonstraram que a proteína não-estrutural 1 (NS1) do vírus Influenza é um potencial alvo terapêutico. A NS1 tem um papel principal na replicação do vírus e na supressão do sistema. Esta proteína multifuncional, que participa em interações proteína-RNA e proteína-proteína, e altamente conservada é constituída por dois domínios: o domínio N-terminal constituído por 73 aminoácidos que interage com o ácido ribonucleico (RNA) de dupla cadeia (RBD) e o domínio C-terminal denominado de domínio efetor (ED).

Com o intuito de descobrir compostos orgânicos com capacidade de inibir a função da NS1, no presente projeto foi desenvolvido e implementado um protocolo que combina abordagens computacionais e experimentais. Primeiro, e para permitir a validação experimental de novos inibidores, caracterizou-se os domínios RBD e ED da NS1 através de várias técnicas experimentais, incluindo Calorimetria de Varrimento Diferencial (DSC), Dicroísmo Circular (CD), Cromatografia de Exclusão Molecular acoplada a um detetor multiangular de dispersão de luz (SEC-MALS) e Ressonância Magnética Nuclear (RMN). Os dados experimentais mostram que, em solução, o ED encontra-se em estado monomérico e o RBD em dímero. O espectro de RMN HSQC ^1H - ^{15}N do RBD-NS1 revela uma grande dispersão de desvios químicos, demonstrando o alto potencial desta técnica para o rastreio de compostos ligantes através da análise da perturbação dos desvios químicos.

A componente computacional deste protocolo assenta na identificação de regiões à superfície da NS1 propensos à interação com moléculas com características semelhantes às de fármacos, seguida do rastreio de pequenas moléculas orgânicas (fragmentos) associados a locais de ligação com elevada similaridade química e topológica de uma biblioteca de estruturas proteicas. As potenciais interações dos fragmentos assim identificados com a

NS1 são analisadas mediante acoplagem (*docking*) dos mesmos aos locais de ligação previstos para esta proteína. Num passo final, e após a seleção e obtenção dos fragmentos mais promissores a partir de fornecedores à volta do mundo, realizou-se a validação experimental das interações por RMN. Na fase final deste projeto já foram testados dois fragmentos e é possível concluir a partir destes resultados preliminares o sucesso do protocolo.

Abstract

Influenza viruses are major human pathogens responsible for respiratory diseases affecting millions of people worldwide and characterized by high morbidity and significant mortality. Influenza infections can be controlled by vaccination and antiviral drugs. However, vaccines need annual reformulation and administration, and provide limited protection. The rapid emergence of influenza virus strains resistant to current antiviral drugs directed against influenza A highlights the need for the development of new classes of antivirals. Toward this end, several structural and functional studies of influenza non-structural protein 1 (NS1) have identified this protein as a potential therapeutic target. This highly conserved and multifunctional protein is composed of two distinctive structural domains, a 73-residue N-terminal double stranded RNA-binding domain (RBD) and a C-terminal effector domain (ED).

With the purpose of discover an organic compound with the ability to inhibit NS1, a protocol combining computational and experimental approaches was designed. First, and to allow the experimental validation of new inhibitors, we have characterized the RBD and ED domains of NS1 with several experimental techniques, such as Differential Scanning Calorimetry (DSC), Circular Dichroism (CD), Size-Exclusion Chromatography coupled with Multi-Angle Light Scattering (SEC-MALS) and Nuclear Magnetic Resonance (NMR). The experimental data shows that in solution the ED is in a monomeric state and RBD is a dimer. The ^1H - ^{15}N HSQC spectrum of RBD-NS1 shows large chemical shift dispersion and thus presents high potential to be used in compound screening based on chemical shift perturbation experiments.

The core of the computational protocol is based on the identification of regions in the NS1 surface with molecules with similar features to drugs, followed by screening of small organic molecules (fragments) associated with binding sites with high chemical similarity from a protein database. The possible interactions of the fragments with NS1 are analyzed using docking of the fragments with the binding site of NS1. In the final step, after the selection of the most promising fragments we proceed to the experimental validation using NMR experiments. We tested two fragments and the preliminary results show the success of the protocol.

1. General Introduction

Infectious diseases and their high mortality rates worldwide are a major cause of health concerns. Such diseases may either be new emergent infections or they may be common infections which increase in importance owing to issues like resistance development to used medicines. Of all infectious diseases, influenza deserves particular attention as it undergoes a high rate of antigenic change, giving rise to new types of influenza strains for which there is no known treatment¹.

The influenza virus was first isolated in 1933. This event was the culmination of many years of research attempting to find the causative agent of the influenza pandemic of 1918, which resulted in 20 million deaths in less than 4 months, and ever since remained as one of the most feared acute threats to human health².

Among the three most common genera of Influenza viral strains (A, B and C), influenza A mutates more rapidly and hence is more virulent and lethal than the other two types. The influenza A virus causes a respiratory disease and it is transmitted by droplets of body fluids, e.g. tears and saliva. It causes mild to severe symptoms, especially among the young and elderly, and may result in serious illness, leading eventually to death. The symptoms associated to this type of infection are: fever, cough, sore throat, muscle or body aches and headaches. The World Health Organization (WHO) keeps constant surveillance of influenza outbreaks worldwide and recommends annual vaccination of high-risk groups, such as children and the elderly. However, vaccination only grants temporary immunity. Therefore, the discovery of new antiviral agents against influenza A virus assumes paramount relevance, since according to WHO annual outbreaks result in 3–5 million severe cases and between 250,000 and 500,000 deaths worldwide³.

The search for novel therapeutic interventions for influenza A viral infection is a challenging pursuit. The targeting of viral proteins, such as hemagglutinin and neuraminidase, has the inextricable obstacle of giving rise to resistance and, thus, new approaches are required to address this unmet medical need.

1.1. Influenza A Virus

1.1.1. Classification

Influenza, commonly referred to as “flu”, is caused by RNA viruses belonging to the *Orthomixoviridae* family, which at present consists of seven genera: Influenza A, Influenza B, Influenza C, Influenza D, Thogotovirus, Isavirus and Quaranjavirus (International Committee on Taxonomy of Viruses, 2017)⁴. The three genera of Influenza virus (A, B and C), which are identified by antigenic variations in their nucleoprotein (NP) and matrix protein (M1), show different tropism for vertebrates. While influenza B and C circulate almost exclusively in humans, influenza A viruses are established also in different animal species including horses, swine and wild birds⁵. In humans, influenza A and B viruses are the predominant cause of significant disease, as acute febrile respiratory tract infection, whereas influenza C virus infects primarily young children – usually resulting in a mild respiratory illness⁶ – and it is not included in the seasonal influenza vaccine.

1.1.2. Genome and Structure

The influenza A virus genome is composed of eight negative-sense single stranded RNA gene segments (ssRNA) that encode 11 major proteins and several auxiliary peptides (Table 1). Production of infectious progeny virus requires incorporation of all eight viral RNA segments and occurs at the apical membrane of infected cells⁷.

Table 1. Influenza A virus genomic segments and major proteins.

Segment	Protein	Protein Function
1	Polymerase basic protein 2 (PB2)	mRNA cap recognition
2	Polymerase basic protein 1 (PB1)	RNA elongation, endonuclease activity
	PB1-F2	Pro-apoptotic activity
	N40	
3	Polymerase acidic protein (PA)	Protease activity
	PA-X	Modulates host response
4	Hemagglutinin (HA)	Major antigen, receptor binding and fusion activities
5	Nucleoprotein (NP)	Nuclear import regulation
6	Neuraminidase (NA)	Sialidase activity, virus release
7	Matrix Protein (M1)	viral Ribonucleic Protein (vRNP) interaction, RNA nuclear export, viral budding
	Matrix Protein (M2)	Virus uncoating and assembly
8	Non-structural protein (NS1)	Regulation of host gene expression
	Nuclear Export Protein (NEP)	Nuclear export of RNA

Influenza A virus is an encapsulated virus. The viral envelope contains a lipid-bilayer obtained from the host cell that covers the capsid containing the virus genome. On the surface of influenza A virus, there are two major glycoproteins: hemagglutinin and neuraminidase. Figure 1 shows the structure of influenza A virus.

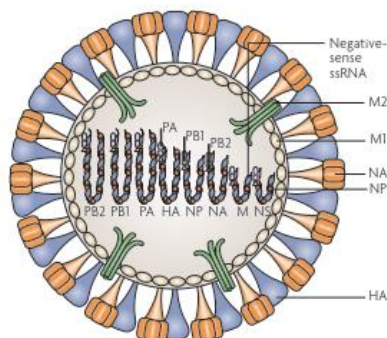


Figure 1. Schematic structure of influenza A virus⁸. Influenza A viruses are enveloped, single-stranded, negative-sense RNA viruses that contain eight gene segments that encode 16 proteins.

There are 18 known subtypes of HA and 11 of NA. The combination of HA and NA subtypes encodes the individual classifications of each influenza A virus strain. Generally, an influenza A virus is classified as H_xN_y, where *x* stands for the HA subtype number and *y* stands for the NA subtype.

Non-structural protein 1 (NS1) is encoded by viral segment 8, which also encodes the viral nuclear export protein, NEP. (NS1) of influenza A virus has attracted much attention for its role in modifying the host innate immune response and controlling virus replication. NS1 is as a potential target for antiviral drug discovery based on its structure, activities, genetics, and overall importance in virus replication and pathogenesis. It is a highly conserved protein of 230-237 amino acids that is produced in abundant levels throughout infection.

1.1.3. Replication Cycle

A host is mandatory for the replication of influenza A virus, and its infection is a multistage process that includes attachment, entry, fusion and uncoating, genome transcription, viral protein synthesis, assembly and finally budding of progeny virions (Figure 2)⁹.

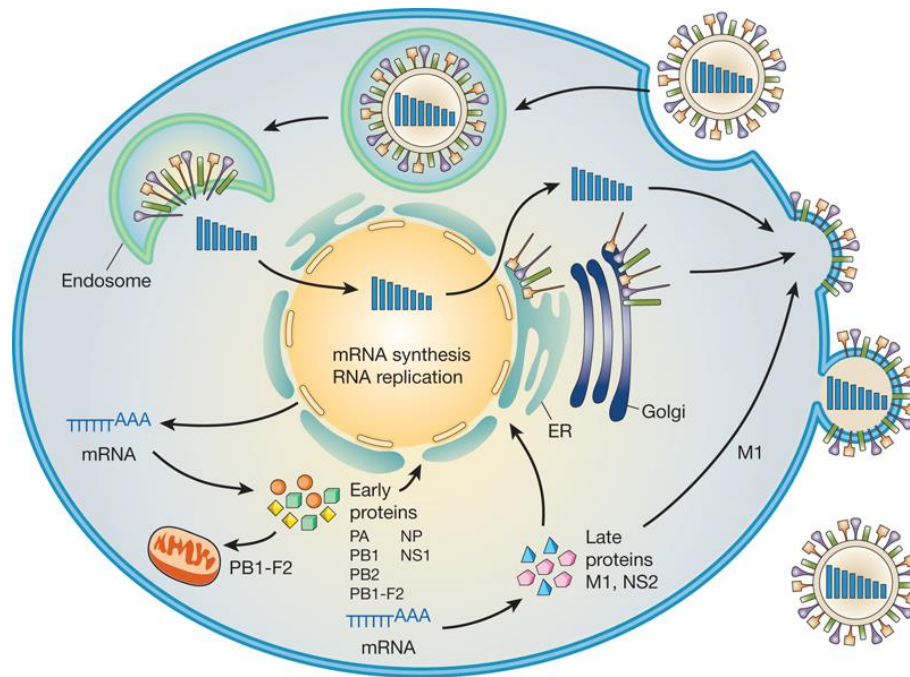


Figure 2. Schematic diagram of the influenza viral life cycle¹⁰. Virus infection is initiated by binding of the virus to sialylated host cell-surface receptors, and entry is mediated by endocytosis. In the host cell, fusion of viral and endosomal membranes occurs at low pH, which enables the release of the segmented viral genome into the cytoplasm. The viral genome is subsequently translocated to the nucleus, where it is transcribed and replicated. Following synthesis in the cytoplasm, viral proteins are assembled into viral ribonucleoproteins (vRNPs) in the nucleus. Export of vRNPs to the cytoplasm is mediated by M1 and NS2. Virus particles are assembled at the cell membrane, and the newly generated progeny virus buds into extracellular fluid.

1.1.3.1. Virus entry into the host cell

Influenza A viruses predominantly enter the host cell by endocytosis after the viral haemagglutinin (HA) protein, which is a homotrimer that forms spikes on the viral lipid membrane, binds to sialic acid found in the surface of the host's cell membrane¹¹. After the binding of hemagglutinin with the host cell's sialic acid residues, receptor-mediated endocytosis occurs and the virus enters the host cells in an endosome. The endosomal acidification triggers the fusion of the viral and endosomal membranes. The acidic environment of the endosome is not only important for the fusion of the viral and endosomal membranes but also opens the M2 ion channel. M2 is a type III transmembrane protein that forms tetramers, whose transmembrane domains form a channel that acts as a proton-selective channel. The viral M2 ion channel concurrently promotes acidification of the virion interior, which dissociates the M1 matrix protein from the viral genome. vRNPs that are released from endosomes are transported into the nucleus through the nuclear pore complex (NPC).

1.1.3.2. Entry of vRNPs into the nucleus

RNP play an important role during the virus infection cycle. Influenza A virus replication occurs in the host cells' nucleus.

The viral proteins that make up the vRNP are NP, Pa, PB1, and PB2 (Figure 3). These proteins have known nuclear localization signals (NLSs) that can bind to the cellular nuclear import machinery and, thus, enter the nucleus.

During infection, the influenza A virus enters the host cell by clathrin-mediated endocytosis, and after viral membrane fusion occurs in the endosome, it releases viral RNPs into the cytosol. Viral RNPs enter the host nucleus by active transport. In the nucleus, the RNPs from the infecting virus serve as active templates for the synthesis of viral mRNA¹².

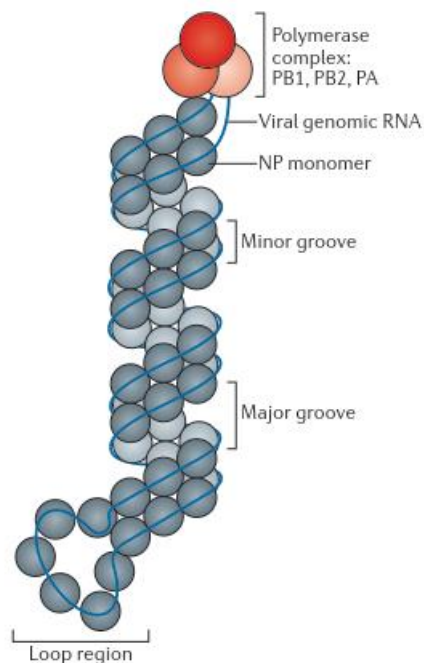


Figure 3. Influenza Ribonucleoprotein Particle (RNP)¹³. Each influenza viral ribonucleoprotein (vRNP) consists of one single-stranded, negative-sense genomic RNA associated with multiple nucleoprotein (NP) monomers and a single trimeric polymerase complex (composed of PB1, PB2 and PA).

1.1.3.3. Transcription and replication of the viral genome

The influenza viral genome is made up of negative sense strands of RNA. For the genome to be transcribed, it first must be converted into a positive sense RNA to serve as template to produce viral RNAs. The negative-sense viral RNAs are transcribed to positive-sense messenger RNAs (mRNA) by the transcriptase, consisting of PB1, PB2 and PA (Figure 3) which are part of the RNPs. In a process referred to as “cap-snatching”, the

transcriptase steals short cap regions from cellular mRNAs as primers for initiation of viral mRNA synthesis. Cap snatching by the viral transcriptase inhibits the synthesis of cellular proteins in favour of production of viral components⁹.

1.1.3.4. Assembly and budding of the viral proteins

After the vRNPs have left the nucleus, all that is left for the virus to do is to form viral particles and leave the cell. Since influenza is an enveloped virus, it uses the host cell's plasma membrane to form the viral particles that leave the cell and go on to infect neighbouring cells.

Virus particles bud from the apical side of polarized cells. Because of this, HA, NA, and M2 must be transported to the apical plasma membrane. M1, which is present underneath the lipid bilayer, is important in the final step of closing and budding off the viral particle.

The death of the host cell occurs after the release of newly replicated virions⁹.

1.2. Antiviral Drugs

The first anti-influenza drugs were identified either using large-scale screening methods or by chance and their modes of action were not completely understood. On the other hand, the current development of new antivirals is based on X-ray crystallographic structures of influenza proteins, an approach usually called structure-based drug discovery.

The anti-influenza drugs are classified according to their target in the viral life cycle. Such antivirals are used as inhibitors of the following steps in the replication cycle: attachment of the virus to host cell receptors, endocytosis and fusion of viral and cell membranes, replication and transcription of the viral genome, synthesis of viral proteins, assembly of the viral progeny and the release of new virions into the outside environment¹⁴.

1.2.1. Inhibitors of haemagglutinin

Haemagglutinin is a glycoprotein located on the surface of influenza virions. During the initial step of infection, haemagglutinin binds specifically to the host cell sialic receptors and enable entry of the virus into the cell cytoplasm by fusion of viral and cell membranes.

Inhibitors of haemagglutinin are not commercialized as antivirals, but, currently, a haemagglutinin inhibitor is in phase II human trials. The sialidase fusion protein DAS181

(Fludase[®]) is a novel broad-spectrum haemagglutinin inhibitor that enzymatically removes sialic acid receptors from respiratory epithelium cells preventing virus attachment. This antiviral is active against both A and B influenza strains at nanomolar concentrations and causes minimal cytopathic effects.

1.2.2. M2 ion channel inhibitors

M2 ion channel is a transmembrane viral protein that mediates the selective transport of protons into the interior of the influenza virion. Conductance of protons acidifies the internal space of the viral particle and facilitates the haemagglutinin-mediated membrane fusion which consequently results in the uncoating of the influenza nucleocapsid and import of the viral genome into the nucleus. Adamantanes are potent M2 channel inhibitors. Two adamantane derivatives, amantadine (Figure 4A) and rimantadine (Figure 4B), are commercially available under the name of Symmetrel[®] and Flumadine[®], respectively.

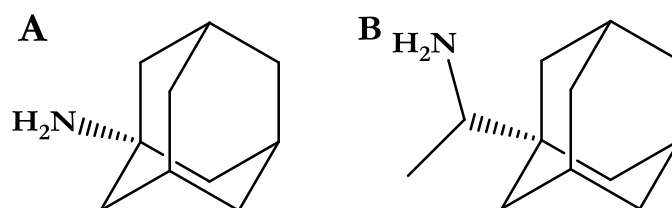


Figure 4. Chemical structures of derivatives of adamantane, known as inhibitors of M2 ion channels. A. Amantadine. B. Rimantadine.

The adamantanes show strong anti-influenza activity at micromolar concentrations. At present, the application of adamantanes for treatment of influenza infections is not recommended because of the rapid emergence of drug-resistance virus variants. Another disadvantage of adamantanes is their activity against influenza A virus strains only.

1.2.3. Inhibitors of viral RNA polymerase

Transcription and replication of the influenza virus genome is carried out by the influenza RNA polymerase. Polymerase activity is needed for the elongation of nascent RNA chains, whereas endonuclease activity is essential for cleavage of 5'-capped primer sequence of the host mRNA. This "cap snatching" process is important for the initiation of viral RNA transcription. Influenza RNA polymerase is a very appropriate target for the development of new broad-specific antivirals because of its highly-conserved structure

among influenza strain and it is thought that the influenza polymerase plays an important role in virus adaptation and human-to-human transmission.

Two classes of RNA polymerase inhibitors have been described based on different mechanisms of action. The first class is represented by nucleoside analogues for the blocking of viral RNA chain elongation. An inhibitor that belongs to this group is favipiravir (T-705) (Figure 5). Favipiravir is an inhibitor of influenza A, B and C, including variants resistant to amantadine or oseltamivir.

The second class is represented by the compounds which block the endonuclease and cap-binding domains of the polymerase. These antivirals include cap analogues, short capped oligonucleotides and small organic compounds, such as 4-substituted 2,4-phenylbutanoic acid and flutamide.

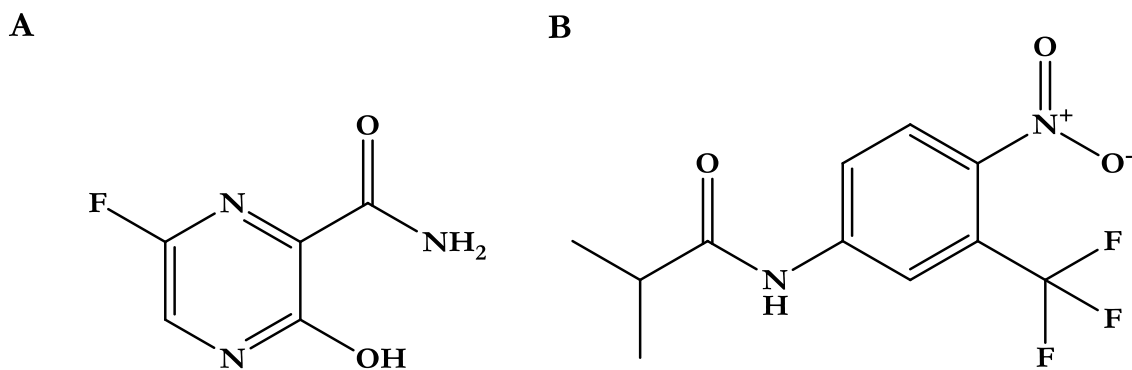


Figure 5. Commercially available inhibitors of viral RNA polymerase. A. Favipiravir. B. Flutamide.

1.2.4. Inhibitors of neuraminidase

Neuraminidase is an antigenic glycoprotein present in the surface envelope of the influenza viral particles, which hydrolytically cleaves the terminal sialic acid from the host cell receptors. Thus, it plays an important role in the release of viral progeny from the membranes of infected cells and facilitates the movement of the infectious viral particles in the mucus of the respiratory epithelia.

Neuraminidase has been established as an important target for the prophylaxis and treatment of influenza infections, for the following reasons; the structure of the neuraminidase active site is highly conserved between influenza A and B strains; resistance to neuraminidase inhibitors has been less reported than to other anti-influenza drugs, nevertheless, the intensive application of neuraminidase inhibitors for influenza treatment

results in an increasing number of drug-resistant strains; and finally, neuraminidase is an easily accessible target for antiviral drugs, with an extracellular mode of action.

At present, several anti-influenza drugs are commercially available, especially the inhalant zanamivir (Figure 6A) with the trademark Releza[®], and the orally administered oseltamivir (Figure 6B) (Tamiflu[®]). In response to the emergence of some influenza strains resistant to oseltamivir, peramivir and laninamivir have been recently developed.

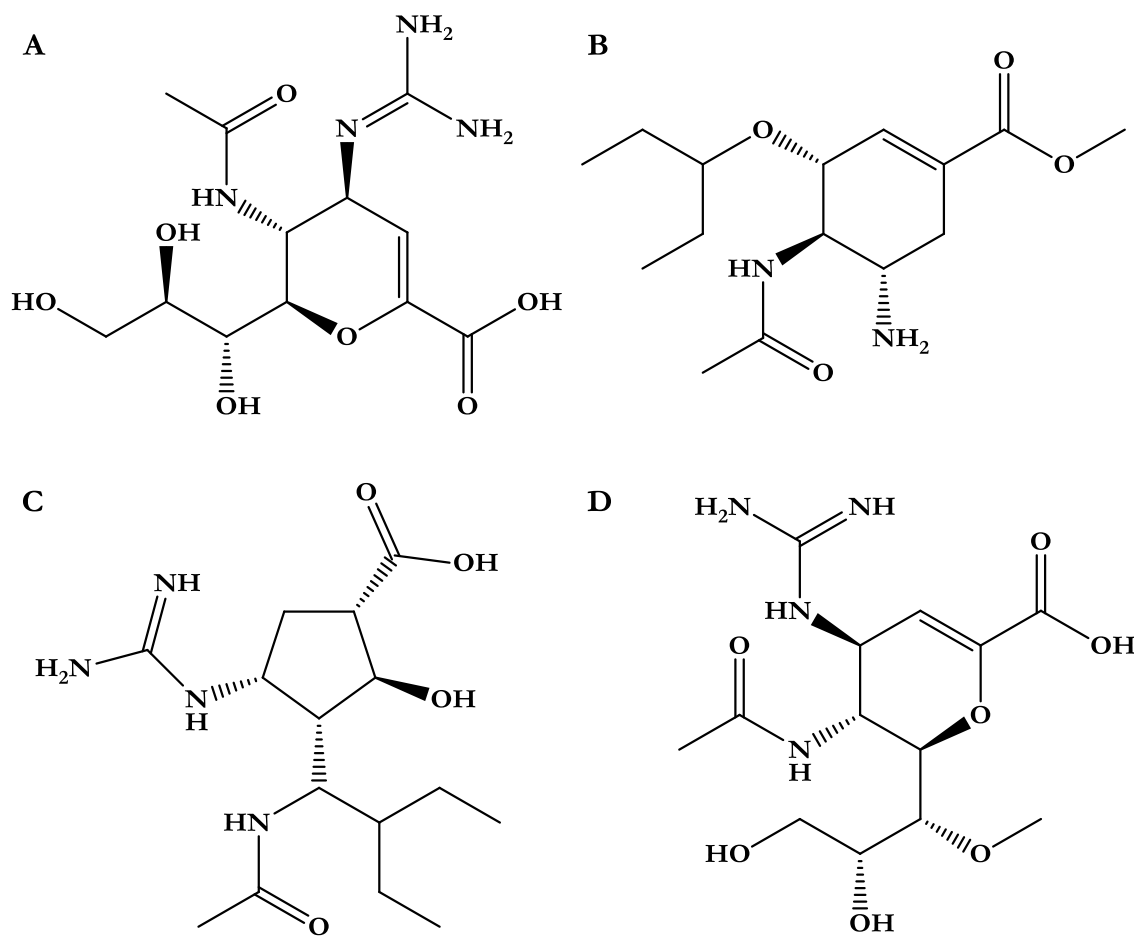


Figure 6. Commercially available inhibitors of Neuraminidase. A. Zanamivir. B. Oseltamivir. C. Peramivir. D. Laninamivir.

Propagation of viruses in the presence of antiviral drugs increases the selection pressure for mutations increasing the resistance to the antivirals. Adamantane resistant strains are typically characterized by a single substitution in the transmembrane region of the M2 ion channel. On the other hand, resistance to neuraminidase inhibitors can result from mutations in the neuraminidase active site, but also from amino acid substitutions on the molecular surface of the neuraminidase protein.

Resistance to derivatives of adamantane is acquired rapidly and by a high number of virus strains, on the contrary resistance to neuraminidase inhibitors take longer and occurs with a relatively low frequency. The increasing emergence of drug-resistant influenza strains highlights the need to search continuously for novel targets and design of new inhibitors. Lately, the protein NS1 has been considered to be a potential target to the discovery of new antivirals¹⁴.

1.3. NS1 as a therapeutic target

All viruses need to deal with the host antiviral response. Cells have been equipped with multiple sensors of viral infection that trigger potent antiviral pathways, including the induction of interferons (IFNs) and IFN-stimulated genes that inhibit viral replication. Conversely, viruses have developed strategies to counteract the host antiviral pathways by hiding from cellular sensors, by actively inhibiting antiviral pathways, or both. In the case of the Influenza virus, the non-structural protein 1 (NS1) plays an important role in virus replication and in the counteracting of the host innate immune response. This protein has been proposed as a potential therapeutic target¹⁵. NS1 is a multifunction protein and it binds and sequesters dsRNA interfering with host mRNA processing, controls viral RNA replication, and facilitates preferential viral mRNA translation. NS1 disables the host immune response primarily via interactions with interferon production and action and also by inhibition of activation of sentinel dendritic cells. Targeting NS1 seems to be a promising novel strategy for influenza therapy since it offers the possibility of inhibiting the disease progression on several levels with only one compound. NS1 antagonists could prevent disabling of the host immune defence, as well as decrease production of viral progeny.

2. Aims and Project Workflow

The main objective of this work is the identification of molecular fragments holding affinity for the NS1 protein of the influenza virus, which may be used as building blocks for potential tool compounds and/or potent NS1 inhibitors. This main goal will be achieved through deployment of an innovative screening protocol combining *in silico* and experimental components.

In greater depth, the aims of this work are:

- ⌘ Structural analysis of the RNA-binding domain (RBD) and effector domain (ED) of NS1;
- ⌘ Design of a new computational protocol for the discovery of fragments to be used as building blocks for potent NS1 inhibitors;
- ⌘ Experimental validation of the most promising fragments using Nuclear Magnetic Resonance (NMR) and chemical shift perturbation analysis.

In order to offer the reader an overall grasp of this project, here I provide a brief description of the project workflow that leads to the experimental confirmation of active fragments interfering with the pre-defined biological target – the NS1 protein –, and thus to new advances in the design of novel inhibitors of NS1 function (see 7). Because some of the concepts and methodologies explored in this work have not yet been introduced in this document, each step of the workflow is described in a non-technical and conceptual manner.

In a preliminary step, the RNA-binding domain of NS1 was chosen after cloning, production and purification of ^{15}N -labelled protein, and NMR analysis of both RBD and ED domains. Unfortunately, no optimal experimental conditions were obtained for the analysis of the ED by NMR, and thus in this work only RBD will be used. The ^1H - ^{15}N HSQC NMR spectrum of RBD showed large chemical shift dispersion, suggesting the possibility of subjecting this domain to *in silico* studies and then conduct experimental validation using NMR spectroscopy.

The **first step** of the workflow comprises the detection of potential binding sites and pockets, or more broadly, any favourable regions of the NS1 surface for interaction with small organic molecules (*hot spots*) in the RNA-Binding domain. RBD interacts with dsRNA to inhibit the nuclear export of mRNA and is also involved in modulating pre-mRNA splicing. Our *hot spot* analysis extends the concept of binding site analysis to the assessment of what is generally referred to by *druggability*, i.e. an attempt to predict whether the detected pockets and subpockets are likely to bind molecules that hold the typical characteristics of known drugs¹⁶.

Despite the latest advances in the characterization of NS1 as a potential target for drug discovery, there is little to no information on known ligands of both domains (RBD and ED) that could be used as templates for virtual screening. The **second step** of this workflow represents the main *screening* step – and is arguably the most innovative aspect of this work. It involves two main components:

- 1) The assembly of a library of protein structures in complex with fragments, through filtering of protein-ligand complexes of the database PDBbind, following the so-called “rule of three” (Ro3) – a set of empirical rules encoding for fragment-like molecules; the PDBbind database contains experimental affinity values for protein-ligand interactions of structures deposited in the Protein Data Bank (PDB), such as K_i , K_d and IC_{50} ; the resulting set of fragment-bound protein structures were subjected to more exhaustive binding site analyses and calculation of molecular descriptors based on Molecular Interaction Fields (MIFs);
- 2) An all-to-one MIFs-based alignment and comparison between the cavities comprising the fragments in PDBbind-derived complexes and RBD’s binding pockets – via an approach called IsoMIF – allowing a ranking of fragments whose respective enclosing sites better matched the binding pockets in RBD.

Fragments belonging to protein structures whose binding sites received high similarity scores and holding high ligand efficiency (calculated from PDBbind data) are selected to be progressed to an *in silico* confirmatory step by molecular docking.

In **step three**, molecular docking is performed using a docking packaged specifically designed for docking of fragments – SEED (Solvation Energy for Exhaustive Docking). SEED docks polar fragments so that at least one hydrogen bond with optimal distance to a protein polar group is made, while hydrophobicity maps are used for the docking of apolar fragments. By performing docking of the selected fragments within the binding sites of RBD, the optimal orientations of the fragments is determined and visually analysed.

In **step four**, the commercial availability of the fragments passing the previous steps is then surveyed in multiple catalogues of worldwide chemical vendors. Compounds that are commercially available may be acquired for experimental validation using NMR spectroscopy.

Finally, in **step five** ^1H - ^{15}N HSQC NMR spectra, in the absence and in the presence of the fragments, are collected. Large chemical shift dispersion of RBD’s ^{15}N - ^1H HSQC spectrum allows the experimental confirmation of fragment binding by chemical shift perturbation methods.

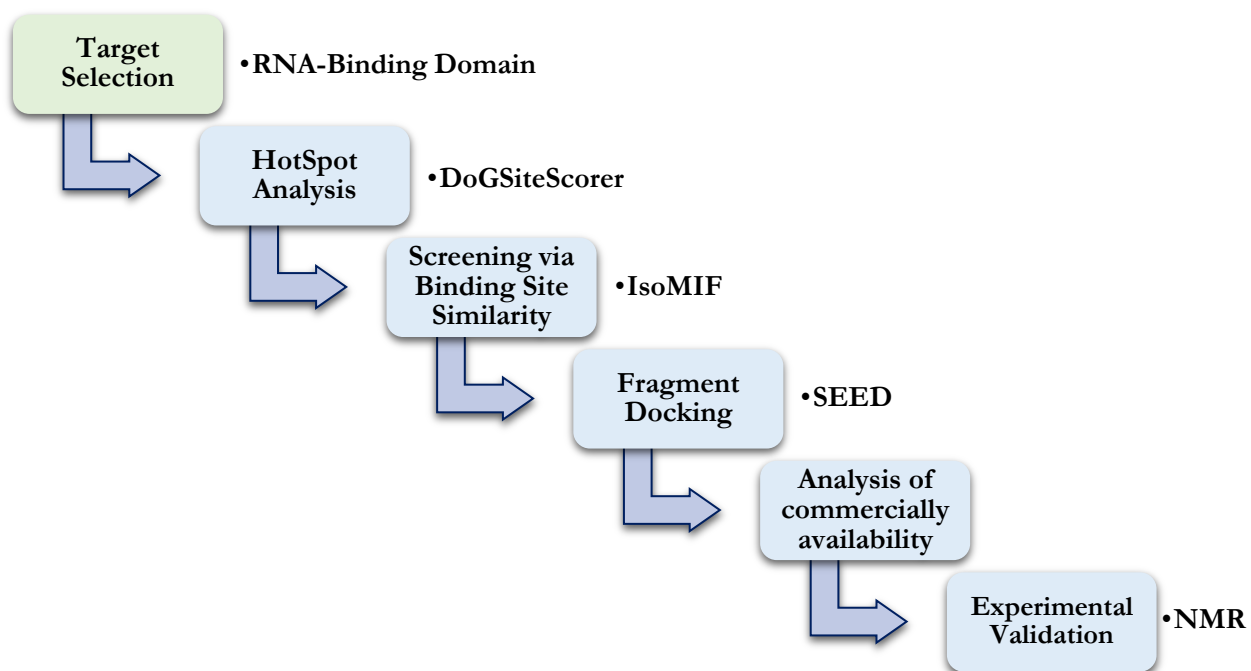


Figure 7. Workflow of the project presented in this dissertation, with every main step highlighted in blue. The preliminary step of target selection is shown in green.

3. Target Identification and Characterization

Targeting NS1 seems to be a promising novel strategy for influenza therapy since it offers the possibility of inhibiting the disease progression on several levels with only one compound. NS1 antagonists could prevent disabling of the host immune defence, as well as decrease production of viral progeny.

3.1. Introduction

3.1.1. NS1 structural description

NS1, a highly-conserved protein, is encoded by segment 8 of the viral genome and ranges in amino acid chain length from 215 to 237 amino acids. NS1 comprises two functional domains: RNA-binding Domain (RBD) (amino acids 1-73) (Figure 8) and Effector Domain (ED) (amino acids 86-215/230) (Figure 9). These two domains are connected by a short linker of 13 amino acids.

The isolated RBD forms a six-helical head to tail homodimer, which binds to dsRNA (Figure 8B). Remarkably, only one amino acid per subunit (R38/R38') in RBD is absolutely required for dsRNA binding. The R38/R38' residues from the two monomers form hydrogen bonds with each other, as well as with the two RNA strands, thereby anchoring the dsRNA. In addition to R38, positively charged residues in the middle of the dsRNA binding surface, such as R35, R37, and K41 establish hydrogen bonds and electrostatic interactions with both strands of the dsRNA.

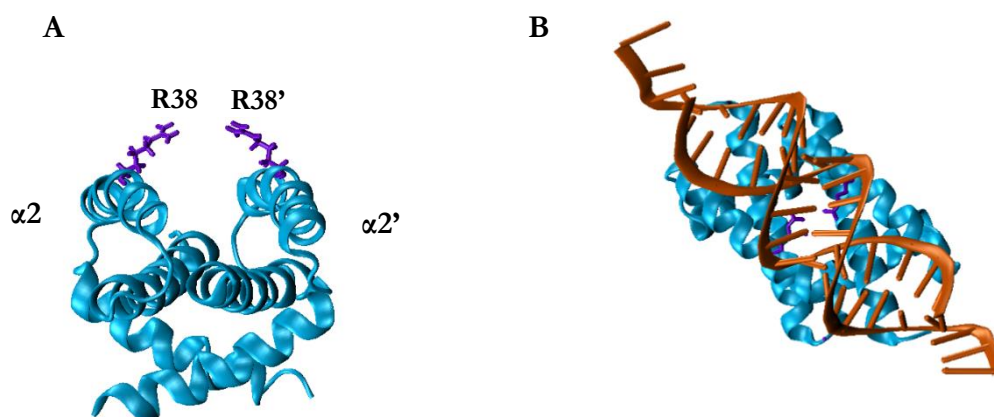


Figure 8. A. NMR structure of the RNA-binding domain with arginine 38 highlighted (PDB code 2N74). B. Crystallographic structure of the RNA-binding domain bound with dsRNA (PDB code 2ZKO¹⁶).

The isolated ED of NS1 comprises seven β -strands and three α -helices and can also dimerize (Figure 9). The mode of RBD dimerization does not vary between strains and is not affected by the interaction with dsRNA¹⁷. On the other hand, several possible ED homodimer interfaces have been proposed based upon crystal structures obtained for NS1 from various strains.



Figure 9. Crystallography structure of NS1's ED domain. (PDB code 2GX9¹⁸).

The number of cellular proteins reported to interact with NS1 is very large¹⁹. One of the many functions of NS1 is to inhibit interferon response. NS1 subverts the development of immune responses by counteracting the signalling of the Pattern Recognition Receptor (PRR), co- and post-transcriptionally inhibiting host gene expression and post-transcriptionally inactivating interferon-stimulated gene products²⁰.

NS1 contributes to Influenza A's variability by displaying a substantial number of sequence, length, structure, modification and functional polymorphisms. Structural polymorphisms of NS1 are mainly dependent on the length of the interdomain linker (Figure 10). The linker length and the identity of residue 71 determine the preference of RBD orientation toward ED and provide the structural basis for strain-dependent NS1 functions²¹.

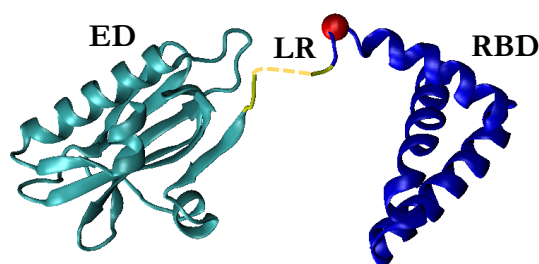


Figure 10. Schematic representation of the monomeric form of the structure of NS1 highlighting the RNA-binding domain (RBD, in blue), the Effector domain (ED, in cyan) and the linker region (LR, in yellow). The position of residue 71 is represented by de red bead.

While the interdomain linker of NS1 is 12 amino acid long in most protein variants, its length in highly pathogenic avian influenza A viruses is only 7 amino acids. This fact indicates that, although there are no known interaction motifs within the linker, it can still influence virulence²².

Although functional NS1 is necessary for replication in immune-competent systems^{23,24}, different subtypes of NS1 may vary in their ability to modulate certain cellular responses. For instance, despite different NS1s being capable of binding to the components of the retinoic acid-inducible gene I (RIG-I) pathway, which is essential for detecting viral RNA and initiating the innate immune response, the extent to which they inhibit RIG-I signalling varies, and some H1N1, H2N2 and human H3N2 viruses fail to pre-transcriptionally block IFN production²⁵⁻²⁸. These viruses compensate for their inefficiency in pre-transcriptionally control the innate immune response with strong CPSF30 binding²⁹. NS1 binds the 30 kDa subunit of CPSF30 that is required for the 3'end processing of cellular pre-mRNAs. As a consequence of the sequestering of CPSF30 by the NS1 protein, unprocessed cellular pre-mRNA accumulate in the nucleus, and cellular mRNA production in the cytoplasm is inhibited, including interferon mRNAs and other antiviral mRNAs.

3.1.2. Structure-based drug design

Structure-based drug discovery (SBDD) uses both the knowledge of the three-dimensional structure of a protein and *in silico* techniques to identify putative small molecules with biological activity against a desired target.

As the availability of crystal structures increased in the early 1990s, several computational methods were developed to use the structure of the protein target to discover novel hit compounds. Such methods include molecular docking, *de novo* design and molecular dynamics (MD)-based techniques.

Receptor-based virtual screening methods are typically based on molecular docking and try to score compounds stored in large databases to identify which ones will interact with a target protein. Docking methods can be quite useful at predicting the binding position and orientation of ligands that are known to bind to a protein.

De novo design attempts to use the structure of the protein to generate novel chemical structures that can bind to the protein. There is a myriad of *de novo* design

algorithms, some of them depend on identifying initial hot spots of interactions, which results in fragments that are then grown into complete ligands.

Fragment-based lead discovery is based on the premise that most ligands that bind strongly to a protein active site can be considered as a combination of several smaller molecular fragments. Fragments can be identified by screening a relatively small library of molecules (400–20,000) by X-ray crystallography, NMR spectroscopy or functional assays. The structures of the fragments that bind to the protein can then be used to design new ligands by adding new functionalities to the fragment, by merging together or linking various fragments or by grafting features of the fragments onto existing ligands. Critical challenges facing fragment screening include the design of fragment libraries containing sufficient diversity, and the synthetic difficulties associated with fragment evolution.

3.1.3. Hot Spot Analysis

A very important step in structure-based drug design is locating ligand-binding sites. The interactions of a protein with ligands are critical to its biochemical function. Usually, only a few residues at defined locations on a protein's surface participate in these interactions, the so-called protein binding sites. Proteins have pockets that evolved to bind small molecules. Within these pockets, areas that make a large contribution to the binding affinity may be found – typically referred to as *hot spots*.

There are many different programs that can perform the search for hot spots, a well-established one is GRID. This program is able to locate favourable positions for atomic probes on a protein surface. GRID places several atomic probes parameterized in a force field (GRUB force field) at each point on a three-dimensional (3D) grid, positioned over selected regions of the target protein and then computes favourable molecular interaction potentials for the used probes³⁰.

Some hot spot detection methods attempt to further classify potential binding sites as “druggable”, i.e. capable of interacting with chemical compounds that resemble drugs (in terms of physicochemical properties) and thus trigger a biological response. In fact, various aspects of protein *druggability* can be assessed *in silico*³¹⁻³³. One first aspect is the presence or absence of a protein cavity that is large enough and with enough depth to accommodate a small molecule. A second aspect may be whether the physicochemical properties of this cavity can complement the properties of a drug-like molecule through the existence of a

suitable pharmacophore arrangement (mostly, stereoelectronic features) at the binding site. The shape and existence of micro-cavities or narrow clefts that can provide strong interacting hot spots³⁴⁻³⁷ for the ligand is the third aspect of the assessment of the binding pocket.

DoGSiteScorer is a well-established webserver and can be used to spot potential binding pockets and subpockets in a protein structure. It analyses the geometric and physicochemical properties of these pockets and estimates their *druggability* with assistance of a machine learning technique called Support Vector Machines (SVMs). The first step of the DogSiteScorer procedure consists of predicting potential pockets on the protein based exclusively on the protein heavy atoms coordinates. This method is based on grid methods (e.g. GRID) and uses a Difference of Gaussian filter to detect potential binding pockets merely based on the tridimensional structure of the protein. With this operation, positions on the protein surface are identified where the location of a sphere-like object is favourable. Based on a density threshold, these positions are clustered to potential subpockets. Finally, neighbouring subpockets are merged to pockets. Numerous geometric and physico-chemical properties are then automatically calculated for the predicted pockets and respective subpockets, like pocket volume, surface, shape and enclosure. Pocket atoms counts, number of functional groups and amino acid composition describe the physico-chemical features of the pocket. Moreover, the lipophilic character of the pockets is addressed by calculating their lipophilic surface and the overall hydrophobicity ratio. Pocket volume and surface are calculated by counting the grid points constituting the pocket volume or its surface and multiplying this number with the grid box volume or surface, respectively. A breadth-first search is used for pocket depth computation, starting from the solvent exposed pocket parts toward the most deeply buried regions. Ellipsoids fitted into the pocket volume reflect the overall pocket shape. The pocket enclosure is derived from the ratio between pocket hull and surface grid points. Each atom within 4 Å of any pocket point is considered a pocket atom. Pocket atom counts or functional groups and amino acid compositions describe the physico-chemical features of the pocket.

For druggability predictions, a supervised machine learning technique, more precisely SVM, is incorporated. Based on a discriminate analysis, a subset of descriptors best suited to separate druggable from undruggable pockets has been selected. The model has been trained and tested on the non-redundant version of the druggable dataset³⁸. External cross validation, randomly taking one half of the data as training and the other half as test set, showed a mean accuracy of 90%.

For each input structure, the method predicts potential pockets, describes them through descriptors and queries the SVM model for druggability estimations. A druggability score between 0 and 1 is returned. The higher the score the more druggable the pocket is estimated to be³⁹.

3.1.4. Experimental techniques for protein characterization

Protein structure characterization may be carried out using a variety of experimental techniques, including X-ray diffraction, nuclear magnetic resonance (NMR), circular dichroism (CD), light scattering (LS), size exclusion chromatography coupled with multi-angle light scattering (SEC-MALS), and fluorescence-based methods. Calorimetric methods like differential scanning calorimetry (DSC) are useful for extracting thermodynamic properties related, for example, with protein folding and stability.

3.1.4.1. DSC

Proteins can undergo thermally-induced conformational changes. These structural rearrangements result in the absorption of heat caused by the redistribution of non-covalent interactions. Differential scanning calorimeters measure this heat uptake. Indeed, differential scanning calorimetry (DSC) is a thermodynamic technique that measures heat capacity as a function of temperature. DSC has been widely used to study protein thermodynamics, folding, and interactions⁴⁰.

A macromolecule in solution is in equilibrium between its native (folded) and denatured (unfolded) conformations. In general, the higher the thermal transition midpoint (T_m), the more stable the molecule is. DSC measures the enthalpy (ΔH) of unfolding that results from heat-induced denaturation. It is also used to determine the change in heat capacity (ΔC_p) of denaturation upon the temperature increase. DSC can help elucidating the factors that contribute to the folding and stability of native macromolecules. These include the hydrophobic effect, electrostatic interactions, Van der Waals interactions, hydrogen bonding, conformational entropy and the physical environment.

3.1.4.2. CD

Circular dichroism (CD) refers to the differential absorption of the left and right circularly polarised components of plane-polarised radiation (Equation 1). This effect occurs when a chromophore is chiral.

$$\Delta A(\lambda) = A(\lambda)_{LCPL} - A(\lambda)_{RCPL} \quad (\text{Equation 1})$$

A chiral chromophore is optically active yielding a CD signal for one of the following reasons: it is intrinsically chiral because of its structure; it is covalently linked to a chiral centre in the molecule or it is placed in an asymmetric environment by the three-dimensional structure adopted by the molecule.

Protein secondary structure can be evaluated by CD spectroscopy in the far-UV spectral region (190-250 nm) and it is based on the excitation of electronic transitions of the peptide amide groups. The peptide backbone forms characteristic secondary structures, such as α -helices, β -sheets and random coil, with specific Φ , Ψ dihedral angles and H-bond patterns affecting the CD spectrum⁴¹. (Figure 11)

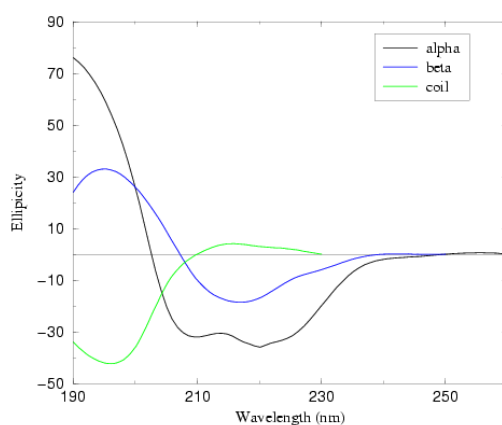


Figure 11. Circular Dichroism spectra of proteins with different representative secondary structures.

CD data are presented in terms of ellipticity $[\theta]$ (degrees) or differential absorbance (ΔA).

Differential absorbance can be converted in ellipticity using the following expression:

$$\theta_{\text{obs}} = 32.98 \Delta A$$

For protein far-UV CD the ellipticity may be converted in mean residue ellipticity using the following expression

$$[\theta]_{\text{MRW}} = \frac{\theta_{\text{obs}} \times 100 \times \text{MM}}{c \times l \times n_A}$$

where $[\theta]_{\text{MRW}}$ is the mean residue ellipticity ($\text{deg.cm}^2.\text{dmol}^{-1}$), θ_{obs} is the observed ellipticity in degrees, MM is the molecular weight in Da, c is the concentration in g.ml^{-1} , l is the pathlength in cm and n_A is the number of aminoacid residues in the protein.

3.1.4.3. SEC-MALS

Size-exclusion chromatography is a technique for separating macromolecules based on their size, more precisely their hydrodynamic volume, and shape. Separation is achieved by differential exclusion and inclusion of solutes as they pass through a stationary phase consisting of cross-linked beads with pores of defined size. The process is based on the difference in permeation rates of each solute molecule into the interior of gel particles.

A column of gel particles or porous matrix is in equilibrium with a suitable mobile phase for the molecules to be separated by size. The elution time is dependent on an individual protein's ability to access the pores of the matrix. Large molecules remain in the volume external to the beads, as they are unable to enter the pores. The resulting shorter flow path means that they pass through the column faster, thus emerging early. Proteins that are excluded from the pores completely elute in what is designated the void volume, V_0 . Small molecules that can access the liquid within the pores of the beads are retained longer and therefore pass more slowly through the column. The elution volume for molecules completely included in the pores is designated the total volume, V_t . The elution volume (V_e) for a given protein will lie between V_0 e V_t .

By adding a detector to the chromatographic system, it is possible to follow the separation in SEC. The elution of proteins is detected by UV absorption, at 280 nm, that is the absorption region of the aromatic amino acids, like tryptophan and tyrosin.

The combination of SEC with a multi-angle light scattering (MALS) detector offers the possibility of measuring the molecular weight of a protein and its oligomeric state in solution, independently of the elution volume.

The principle of Static Light Scattering is that a beam of polarized light is focused onto the sample molecule and the scattered light is detected with a photodetector. In Multi Angle Light Scattering the scattered light is detected at various angles at the same time. The intensity of the scattered light at each angle is proportional to the molar mass and the concentration of the molecules under investigation. The relation between light dispersion and the molecular weight is given by the following equation:

$$\frac{K^*c}{R(\theta)} = \frac{1}{P(\theta)MM} + 2A_2C$$

where $R(\theta)$ is the excess intensity of scattered light at a given angle (θ), c is the sample concentration in $\text{g}\cdot\text{ml}^{-1}$, MM is the molecular weight, A_2 is the second virial coefficient that results from solute-solvent interaction and usually the value used for proteins is $10^{-5} \text{ mL}\cdot\text{mol}/\text{g}^2$, $P(\theta)$ is the complex function describing the angular dependence of the scattered light and K is the optical constant equal to:

$$K = \frac{4\pi^2 n_0^2}{N_A \lambda_0^4} \left(\frac{dn}{dc} \right)^2$$

where n is the solvent refractive index and $(dn/dc)^2$ is the refractive index increment, N_A is the Avogadro's number and λ_0 is the wavelength of the scattered light⁴².

3.1.4.4. NMR

NMR spectroscopy is a powerful tool for the analysis of macromolecular structure and dynamics. NMR is based on the magnetic properties of the nucleus, which are sensitive to the chemical environment. This gives the opportunity to measure properties such as dynamics, chemical exchange or relaxation.

The NMR phenomenon is based on the fact that nuclei of atoms have magnetic properties that can be utilized to yield chemical information. Quantum mechanically subatomic particles (electrons, protons and neutrons) can be imagined as spinning on their axes. In many atoms (such as ^{12}C) these spins are paired against each other, such that the nucleus of the atom has no overall spin. However, in many atoms (such as ^1H , ^{13}C and ^{15}N) the nucleus does possess an overall spin. The rules for determining the net spin of a nucleus are as follows:

- If the number of neutrons and the number of protons are both even, then the nucleus has NO spin.
- If the number of neutrons plus the number of protons is odd, then the nucleus has $I=1/2, 3/2, 5/2$.
- If the number of neutrons and the number of protons are both odd, then the nucleus has $I=1, 2, 3$.

The magnetic moment μ is proportional to the spin angular momentum vector, I , with a factor that is called the gyromagnetic ratio, γ .

$$\mu = \gamma \cdot I$$

Different atoms have different gyromagnetic ratio which is important when performing NMR experiments. The magnetic moment can be oriented only in $2I+1$ different orientations, which in presence of an external magnetic field have different energy. Nucleus with $I=1/2$, such as ^1H , ^{13}C or ^{15}N have two different possible orientations of the magnetic moment, and therefore, when placed in a magnetic field, they have two energy levels. The energy level corresponding to the orientation along the magnetic field is lower and more favourable and thus more populated, such as shown in Figure 12. It is however possible to excite the nucleus to the higher energy level using radio frequency (RF) pulses. The frequency of the electromagnetic radiation must match the energy gap between the two energy levels – the resonance condition.

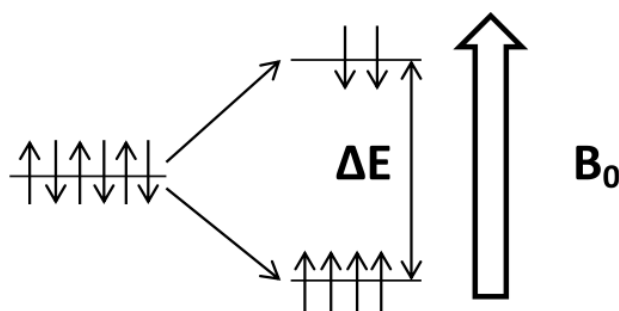


Figure 12. The effect of an external magnetic field B_0 on the orientation of the magnetic moment for a nucleus with $I=1/2$. The nuclei with orientation along the external field have lower energy. The difference in energy between the two energy fields is given by ΔE

The difference in energy between the two energy levels is given by the equation

$$\Delta E = \hbar \cdot \gamma \cdot B_0$$

To induce transitions between the energy levels the applied RF pulse needs to be of the correct energy or frequency:

$$\Delta E = h \cdot \nu$$

The sensitivity (the signal-to-noise ratio) depends on the difference in spin population of the energy levels, making the NMR signal weaker when compared to other spectroscopic methods. It is therefore important to optimize the signal strength, with the use of higher frequency spectrometers and nuclei with high gyromagnetic ratio and high natural abundance, thus the importance of ^1H -NMR. Other ways to optimize the signal is

to measure at higher protein concentration, thus increasing the number of nuclei observed. Additionally, the NMR experiments are executed several times and the results are added to improve the signal-to-noise ratio.

Chemical shifts

The chemical environment of the nucleus influences the magnetic properties, so the resonance frequency is not the same for all nuclei of the same element. In fact, the effective magnetic field acting on each nucleus, and thus the resonance frequency of that nucleus depends also on the local electron distribution. The electrons movement generates a magnetic field of its own, which may add or oppose to the externally applied magnetic field (B_0). This phenomenon is called shielding. In a molecule, not only the atom itself, but the neighbouring atoms and groups in the molecule have an influence that can be either shielding or deshielding to a given nucleus, due to the structure of the molecule. Thus, the chemical environment of a nucleus largely influences its resonance frequency. The most shielded nuclei are at the lowest ppm (right of the spectrum), and more deshielded nuclei have a higher chemical shift (left of the spectrum). This effect is what makes NMR so valuable for chemists, as it allows distinguishing between different atoms of the same element in a molecule.

2D-NMR Experiments

To solve the problem of overlapping NMR signals in a 1D-NMR spectrum, it is possible to record 2D-NMR experiments. This is usually needed for proteins, since even small proteins often have hundreds of protons. Since the magnetization of a nucleus is coupled to the magnetic moment of the nucleus next to it, it is possible to excite one nucleus, for example ^1H , and transfer the magnetization along the spin-spin coupling to the neighbouring magnetic nucleus up to three chemical bonds away, or by dipolar cross relaxation which allows the magnetization to transfer through space for internuclear distances lower than 5 Å. When processing, a double Fourier transform is used, which gives the spectra its common appearance in 2D^{43,44}. A common 2D-NMR experiment is HSQC (heteronuclear single quantum coherence). HSQC correlates a proton to a neighbouring heteronucleus, usually ^{15}N or ^{13}C . The ^1H - ^{15}N HSQC is a useful spectrum that gives information on the state of the protein and if it is properly folded. For most aminoacid residues, the only nitrogen atom is present in the peptide bond. Thus, the ^1H - ^{15}N HSQC experiment gives one peak in the amide region per amino acid in the protein. If the protein is unfolded, all the ^1H - ^{15}N amide peaks have almost the same chemical

environment and are then positioned approximately in the same region of the spectrum. If the protein is folded, however, the amino acids have different chemical environments and the peaks in the spectrum are spread out in a way that is specific for every protein, as a consequence of its particular sequence and structure. The ^1H - ^{15}N HSQC spectrum is thus often used as a fingerprint for a particular protein, and changes in the surrounding elements, such as the addition of a ligand, are detected by changes in the position of the peak in the spectrum (Chemical Shift Perturbation).

3.1.4.5. Fluorescence spectroscopy

Fluorescence is the result of a three-stage process that occurs in certain chemical groups or molecules called fluorophores. The process leading to fluorescence is illustrated by the Jablonski diagram shown in figure 13.

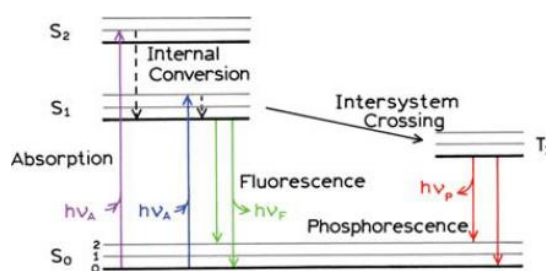


Figure 13. Jablonski diagram and simplified representation of the fluorescence process. S_0 represents the ground singlet electronic state; S_1 and S_2 are the successively higher energy excited singlet electronic states. T_1 is the lowest energy triplet state.

The process of fluorescence begins with the excitation of the fluorophore from a ground singlet electronic state (S_0) to a higher energy singlet state (S_1). The fluorescence emission occurs as the fluorophore decays from the singlet electronic excited state to an allowable vibrational level in the electronic ground state.

The biochemical applications of fluorescence often utilize intrinsic protein fluorescence. In proteins, the three aromatic amino acids, phenylalanine, tyrosine, and tryptophan, are all fluorescent but only tyrosine and tryptophan are used experimentally because their absorption and fluorescence quantum yield are high enough to give a good fluorescence signal. A valuable feature of intrinsic protein fluorescence is the high sensitivity of tryptophan to its local environment. Changes in the emission spectra of tryptophan often occur in response to protein conformational changes or protein-ligand

interactions. Protein fluorescence is generally excited at the absorption maximum of tryptophan, near 280 nm, and emission collected between 330 and 360 nm⁴⁵.

3.2. Methods

3.2.1. Structural information retrieval and quality assessment

In the early stages of drug discovery, it is necessary to access the availability of 3D structures of a chosen target. The Protein Data Bank (PDB) is a well-known repository of the 3D structures of proteins⁴⁶.

Three-dimensional (3D) structures of proteins are typically obtained using one of two main techniques: X-ray crystallography and NMR. Structures obtained from both techniques can be found in the PDB.

In this work, 45 NS1 structures were found in PDB, of which 42 were determined by X-ray crystallography and 3 were determined by NMR. These structures were downloaded and subjected to quality assessment using Ramachandran plots.

In this work, protein structures were visually analysed using Chimera⁴⁷ and PyMOL⁴⁸. Chimera is developed and supported by the “Resource for Biocomputing, Visualization, and Informatics”, University of California, San Francisco, USA, and is an extensible program for interactive visualization and analysis of molecular structures and related data, including density maps, supramolecular assemblies, sequence alignments, docking results, trajectories, and conformational ensembles. PyMOL is a molecular visualization software distributed by the company Schrödinger, LLC, New York, USA. PyMOL is a molecular graphics tool that has been widely used for 3D visualization of proteins, nucleic acids, small molecules, electron densities, surfaces and trajectories. It is also capable of editing molecules and making movies.

3.2.2. Binding site identification and druggability assessment

Hot spot analysis and *druggability* assessment was performed using the web tool DoGSiteScorer, (<http://proteinsplus.zbh.uni-hamburg.de/>)⁴⁹. The structure of RBD (PDB code 2N74) was uploaded into DoGSiteScorer’s webserver. Next, the settings required for the *druggability* calculation were chosen: herein, we decided to calculate both pockets and subpockets in the RBD surface. A second setting corresponding to the binding site prediction granularity was defined: here, we opted for calculating not only the pocket’s

geometric properties but also their predicted *druggability*. In a final setting, we opted for performing all calculations on both RBD chains.

3.2.3. Experimental characterization of NS1 cloned domains (ED and RBD)

3.2.3.1. DSC

To study the thermal stability of the cloned domains of NS1, differential scanning calorimetry (DSC) was used. DSC measurements of the RBD and ED were performed in a Malvern MicroCal VP-Capillary microcalorimeter. DSC studies of RBD were carried out in a range of temperature between 10 °C and 120 °C, at 2 atm. RBD and ED concentrations were 1.21 mg/mL in PBS buffer, at pH 7.4.

3.2.3.2. CD

Far-UV CD data were acquired on a Olis DSM 20 Circular Dichroism spectropolarimeter continuously purged with nitrogen, equipped with a Quantum Northwest CD 150 temperature-controlled system, and controlled by the Globalworks software. Scans were collected between 195-260 nm at 1 nm intervals. Two spectral scans with an integration of 5 seconds per nm were averaged at 20 °C.

RBD concentration was 12.9 μM and ED concentration was at 7.24 μM . The concentration of RBD labelled with ^{15}N was 16 μM and the concentration of ED labelled with ^{15}N was 8.2 μM . Protein samples were prepared in phosphate sodium buffer 10 mM and 140 mM of NaCl, pH 7.4. The cuvette pathlength was 1 mm. Baselines with buffer were also acquired and subtracted from the raw data.

3.2.3.3. SEC-MALS

Size-exclusion chromatography samples of ED, with an injection volume of 100 μl , contained 1.33 mg/mL of ED in phosphate sodium buffer 20 mM and 100 mM of NaCl, pH 6.9). SEC samples of RBD with an injection volume of 100 μl , contained 1.26 mg/mL of RBD phosphate sodium buffer 20 mM and 100 mM of NaCl, pH 6.9.

All samples were analysed at room temperature. The samples were injected into a WTC-030S5 SEC column (Wyatt), previously equilibrated with running buffer (20 mM NaPi, 100 mM NaCl, pH 6.9) and eluted at 0.5 $\text{ml}\cdot\text{min}^{-1}$. The running buffer was previously degassed and filtered through membranes of 0.2 μM porous.

3.2.3.4. NMR

¹H-¹⁵N HSQC Experiments

The ¹H-¹⁵N HSQC experiments on RBD were performed at 7 °C on a Varian VNMRS 600 MHz spectrometer and the ¹H-¹⁵N HSQC experiments on ED were performed at 25 °C on a Bruker Avance III HD 500 MHz spectrometer. NMR samples of RBD, with a final volume of 400 µl, contained 0.4 mM of RBD in PBS (20 mM NaPi, 100 mM NaCl, pH 6.9), 5% of D₂O, 25 µM of DSS and 0.025% of Sodium Azide. NMR samples of ED, with a final volume of 400 µl, contained 0.26 mM of ED in the corresponding buffer, 25 µM DSS, 0.025% of Sodium Azide, 1% Glycerol and 5% of D₂O.

Diffusion Coefficient Experiments

The diffusion coefficient of RBD was obtained at 25 °C on a Bruker Avance III 400 MHz spectrometer with DOSY experiments. NMR samples, with a final volume of 450 µl, contained 120 µM of RBD, in the corresponding buffer, PBS (20 mM NaPi, 100 mM NaCl, pH 6.9), and 50 µl of D₂O.

3.2.3.5. Fluorescence

Fluorescence experiments were performed on a Varian Cary Eclipse spectrofluorometer controlled by the Varian Cary Eclipse software version 1.1. The fluorescence spectra of RBD in solution with different percentages of DMSO was collected between 300 and 400 nm with an excitation wavelength of 280 nm. Assays were carried out in 5 x 5 mm pathlength cuvettes. Baselines without protein were also acquired and subtracted from the corresponding raw data.

The samples used to study the stability of RBD in solution with a percentage of DMSO between 2 and 10% contained 15 µM of RBD. The spectra were collected 10 minutes after DMSO was added.

3.3. Results and Discussion

3.3.1. Structural information retrieval and quality assessment

Quality of the NMR structure (PDB code 2N74) was assessed by WHAT_CHECK, that is a functionality of the stand-alone program WHAT_IF, and it is represented by Ramachandran plots (Figure 14). According to the plot statistics, 96.5% of

all residues were in favourable regions and 99.3% of all residues were in allowed regions. Most of the amino acids are in the allowed region for right-handed α -helix structure.

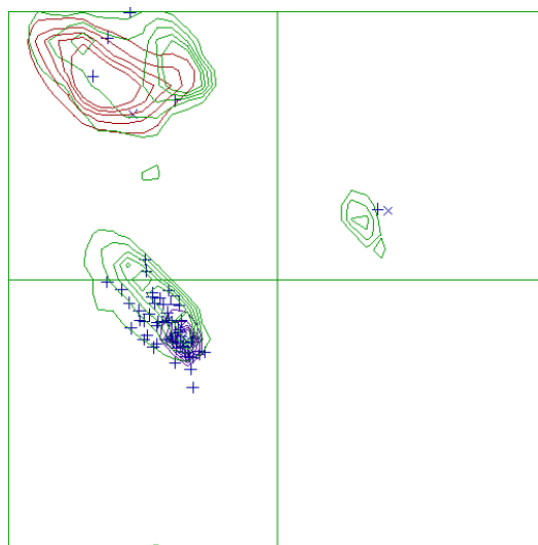


Figure 14. Ramachandran plot for the NMR structure of NS1-RBD (PDB code 2N74; chain A) validated by the WHAT_CHECK/WHAT_IF program.

3.3.2. Pocket and subpocket identification and *druggability* assessment

Target *druggability* encompasses not only the ability of protein binding sites to be complementary with small molecules in terms of physicochemical properties (like size, shape, electrostatics and hydrophobicity) – in order to successfully bind them with high affinity – but also the ability to bind small molecules holding certain physicochemical properties that place them in the so-called “drug-like” property space, implying that a binding site is suitable for interactions with molecules that may be optimized into a therapeutic drug candidate.

In this work, the *druggability* of the RNA-binding domain of NS1 was studied using DoGSiteScorer⁴⁹. The aim was to extract a *druggability* value, as well as more common metrics such as pocket and subpocket volumes. The amino acid composition of the identified pockets was compared to data present in the literature on the most critical amino acids for the interaction of RBD with dsRNA, in order to elaborate strategies to inhibit protein-dsRNA interactions.

DoGSiteScorer was able to identify 5 major cavities in the surface of RBD using the protein structure with PDB accession code 2N74. The predicted pockets and respective characteristics are represented in Table 2 and in Annex B.

Table 2. The five major pockets of the RBD (PDB ID 2N74) predicted by DoGSiteScorer.

Pocket	Volume (Å ³)	Surface (Å ²)	Drug Score
P_0	1022.34	1507.46	0.82
P_1	667.07	1142.65	0.79
P_2	432.64	697.38	0.81
P_3	315.39	536.71	0.54
P_4	120.32	382.36	0.3

In Figure 15, three of the five pockets predicted by DoGSiteScorer are highlighted on the NMR structure of RBD (PDB code 2N74). The largest pocket, represented in blue, encompasses the key residues on the dsRNA-binding surface, arginine 38, and arginine 35, which are important residues for the RBD-dsRNA interaction⁵⁰. This pocket has a depth value of 22.81 Å and a ratio of apolar amino acids of 0.53. A druggable pocket generally is characterized by large pocket volume, high depth as well as a high apolar amino acid ratio, meaning that this pocket is considered druggable, and can be used to discover new small-molecules with the ability to inhibit the RBD-dsRNA interaction. The type of fragment preferable to interact with this pocket is an acceptor and/or donor of hydrogen bonds, since only in this pocket 76 amino acids are hydrogen bond acceptors and 32 are hydrogen bond donors. The pocket represented in yellow consists of serine 42, threonine 5, glycine 45 and arginine 46. This pocket has a depth value of 20.07 Å and a high apolar amino acid ratio of 0.53, meaning this pocket can be considered druggable. In this pocket, 38 amino acids are hydrogen bond acceptors and 11 are hydrogen bond donors, proving that small-molecules with both polar and apolar areas can interact with the amino acids within this pocket. The pocket represented in orange consists of arginine 35, proline 31 and aspartate 34, and has a depth value of 11.95 Å and an apolar amino acid ratio of 0.43. Despite the low value of depth, when compared to the other two pockets, the high apolar amino acid ratio indicates that this pocket can also be used for the discovery of small-molecules with the ability to interact within this pocket, resulting in the impairment of the interaction of the RBD with the dsRNA.

These three pockets may be valuable targets for developing small-molecule inhibitors of RBD-dsRNA interactions.

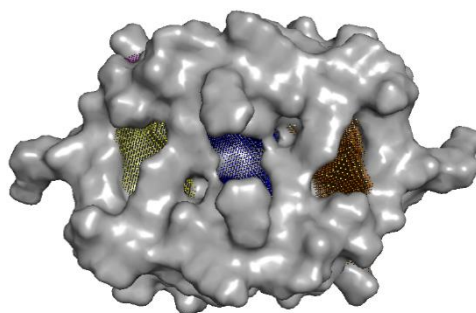


Figure 15. Druggable pockets of NS1-RBD (PDB code 2N74) identified and analysed by DoGSiteScorer.

3.3.3. Experimental characterization of both domains of NS (ED and RBD)

3.3.3.1. DSC

The T_m value for ED was 49.6 °C (Figure 16A) and for RBD was 63.8 °C (Figure 16B), indicating that RBD is more stable to thermal denaturation than ED. The DSC profile for both domains clearly demonstrate that these are well folded proteins with a sharp, cooperative transition between the folded and unfolded states.

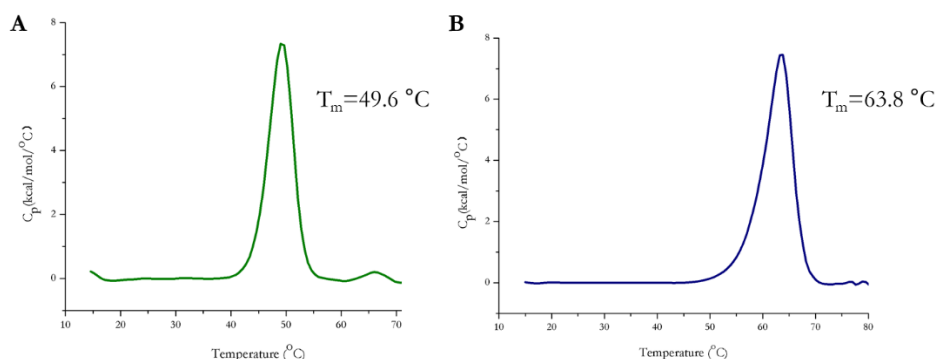


Figure 16. Differential Scanning Calorimetry of both structural domains of NS1. A - DSC heating curve of the NS1 Effector Domain. The T_m of unfolding of the ED is 49.6°C. B - DSC heating curve of NS1 RNA-binding domain. The T_m of unfolding of RBD is 63.8°C.

3.3.3.2. CD

The X-ray crystallographic structure of ED, with a resolution of 1.8 Å, is shown in Figure 17A (PDB code 3RVC). The proportion of different types of secondary structure in the ED three-dimensional structure was calculated with the algorithm DSSP (Definition of Secondary Structure of Proteins)⁵¹. This algorithm indicates that, from the total of 152 amino acids, 28 are present in three α -helices and 50 are present in seven strands of β -

sheet. Thus, the percentages obtained from the DSSP algorithm are 18% for α -helix, 32% for β -sheet and 50% for another type of structures.

In its native state, the effector domain of NS1 presents a mixture of secondary structure in β -sheet and helix- α (Figure 17A). The corresponding circular dichroism in far-UV spectrum is presented in Figure 17B and was collected at pH 6.9 and a temperature of 25°C. The spectra of ED show a minimum of ellipticity at 215 nm, characteristic of proteins with a high percentage of β -sheet, thus suggesting that the cloned ED as native-like structure in solution.

Concerning the stability of ED against the lyophilisation process, there seem to exist no structural differences captured by the CD spectra before (blue) and after (green) lyophilisation (data shown in 17B), suggesting that this process does not seem to affect the structure of the ED.

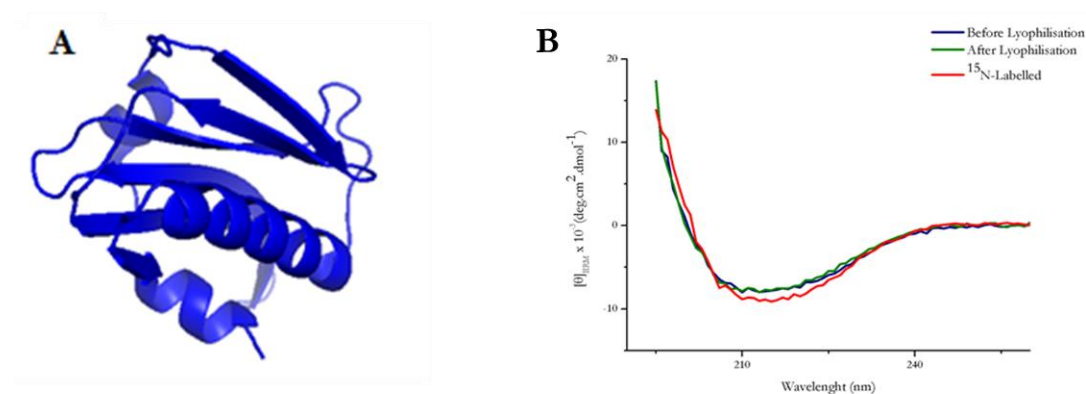


Figure 17. A. Cartoon representation of the 3D structure of NS1-ED (PDB ID 3RVC). B. Far-UV CD spectra of NS1-ED in PBS buffer, at pH 7.4.

In order to study the structure of RBD in solution, CD experiments were performed. The CD spectra of RBD before lyophilisation and labelled with ^{15}N (Figure 18B - light blue and orange, respectively) shows that RBD presents native-like structure composed mostly by α -helix, which agrees with the NMR structure of RBD available in PDB (code 2N74). In this case, the algorithm DSSP considers that from the 73 amino acids, 58 residues are in α -helix. Thus, the equivalent percentage is 79% of α -helix and 21% corresponds to other type of secondary structure.

In its native form, the RBD of NS1 is mainly α -helix (Figure 18A). The far-UV circular dichroism spectrum was collected at pH 6.9 and at 25 °C. The spectrum presents minima of ellipticity at 210 and 220 nm, which is characteristic of proteins with high

percentages of α -helix, thus demonstrating that the cloned RBD as native-like structure in solution.

Concerning the stability of RBD to the process of lyophilisation, like with ED there seem to exist no structural differences in CD spectra before (blue) and after (green) lyophilisation (data shown in Figure 18B), showing that this process does not seem to affect the structure of RBD.

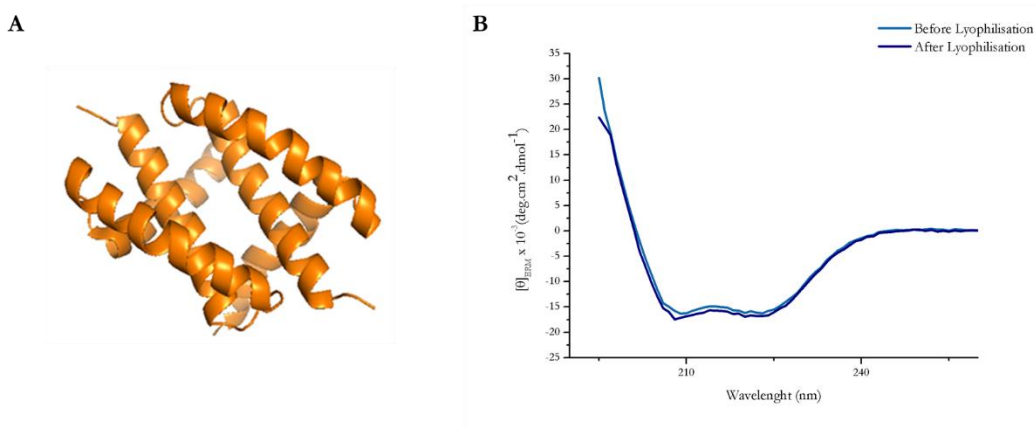


Figure 18. A. Cartoon representation of the 3D structure of NS1-RBD (PDB ID 2N74). B. Far-UV CD spectra of NS1-RBD, in PBS buffer, at pH 7.4.

3.3.3.3. SEC-MALS

SEC-MALS was used to determine if the NS1-RBD, in solution, was in monomeric or dimeric form, since in order to interact with the hosts' protein is necessary the dimerization of this domain. This system of size-exclusion chromatography coupled with multi-angle laser light scattering is proven to be very useful in determining the molecular weight of protein⁵².

In the RBD chromatogram (Figure 19) the peak with elution time of approximately 23 minutes corresponds to a dimeric state, with molecular weight equal to 22.2 kDa. The peak with elution time of 25 minutes corresponds to a monomeric state of 10.6 kDa.

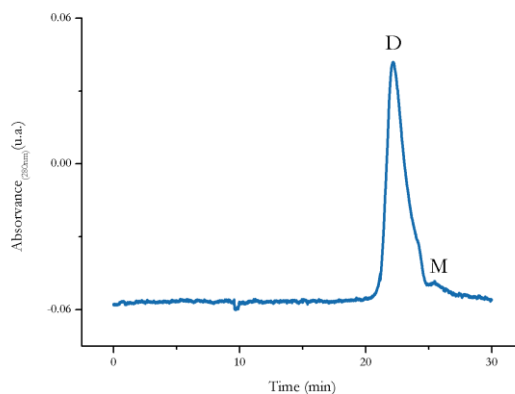


Figure 19. Size-exclusion chromatography coupled to a multi-angle laser light scattering (SEC-MALLS) instrument of the RNA-binding domain of NS1. The chromatogram was run at a flow rate of 0.5 mL/min, in 20 mM sodium phosphate buffer, 100 mM sodium chloride, pH 6.9. Peak D, with an elution of approximately 21 minutes, corresponds to a molecular species a molecular weight of 22.2 kDa, coinciding with RBD in dimeric state and peak M corresponds to a molecular weight of 10.6 kDa, coinciding with RBD in a monomeric form.

3.3.3.4. NMR

The ^1H - ^{15}N HSQC experiment is one of the most frequently recorded experiments in protein NMR. The HSQC experiment can be performed using isotopically labelled proteins. Such labelled proteins are usually produced by expressing the protein in cells grown in ^{15}N -labelled media.

The ^1H - ^{15}N HSQC is normally the first heteronuclear spectrum acquired for the assignment of resonances where each amide peak is assigned to a residue in the protein. If the protein is folded, the peaks are usually well-dispersed, and most of the individual peaks can be distinguished. If there is a large cluster of severely overlapped peaks around the middle of the spectrum, that would indicate the presence of significant unstructured elements in the protein.

The ^1H - ^{15}N HSQC is often used to screen candidates for their suitability for structure determination by NMR, as well as optimization of sample conditions. The HSQC experiment is also useful for detecting the binding interface in protein-protein interactions, as well as to probe interactions with small-molecule ligands.

In the case of the NS1-ED, the large line widths of the NMR peaks in the ^1H - ^{15}N HSQC spectrum may indicate exchange between multiple conformations or a monomer-

dimer equilibrium at higher protein concentrations (Figure 21). The ED sample prepared for the NMR experiments tended to precipitate with time, despite the use of the experimental conditions presented in the literature for the preparation of ED in solution. Therefore, there is the need to optimize the experimental protocol for NMR sample preparation of the Effector Domain of NS1, before following to in silico approaches.

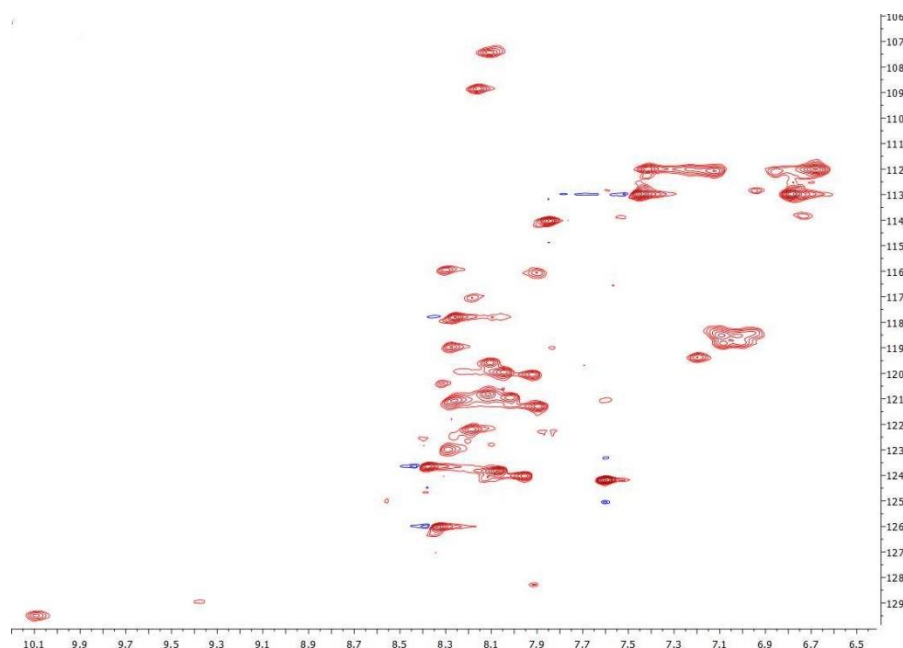


Figure 20. ^1H - ^{15}N HSQC spectrum of ^{15}N -labelled NS1-ED. The spectrum was collected at a ^1H frequency of 500 MHz with a 0.26 mM protein sample in PBS, pH 6.9.

The ^1H - ^{15}N HSQC of NS1-RBD shows large chemical shift dispersion and thus presents high potential to be used in compound screening based on chemical shift perturbation experiments (Figure 21). A table with most of the chemical shift assignments for ^1H and ^{15}N amides of the RBD sequence is shown in Annex A.

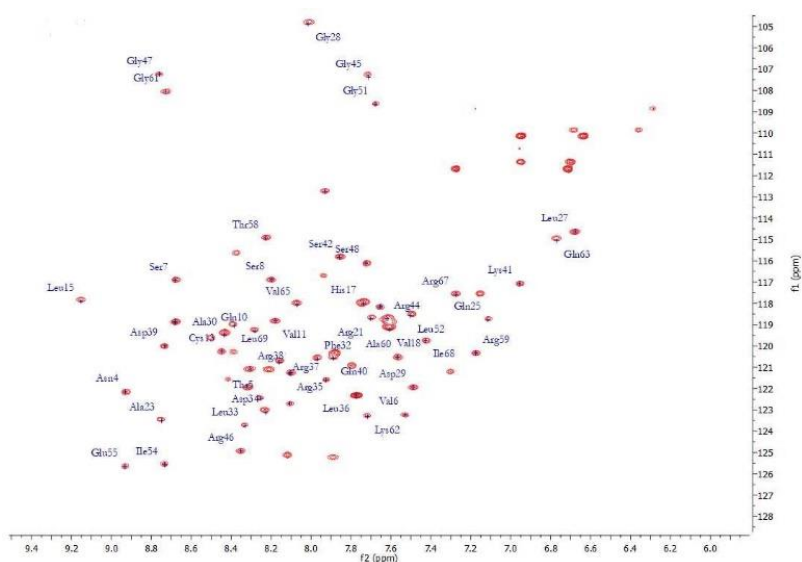


Figure 21. ^1H - ^{15}N HSQC spectrum of the RBD of NS1 with residue-specific backbone assignments indicated and detailed in Annex A. The spectrum was collected at a ^1H frequency of 600 MHz with a 0.40 mM protein sample in PBS, pH 6.9.

Diffusion Coefficient experiments

Diffusion-ordered spectroscopy (DOSY) seeks to separate the NMR signals of different molecular species in solution according to their diffusion coefficient.

In order to confirm the results obtained for NS1-RBD by SEC-MALLS, we determined its diffusion coefficient by NMR (Figure 22).

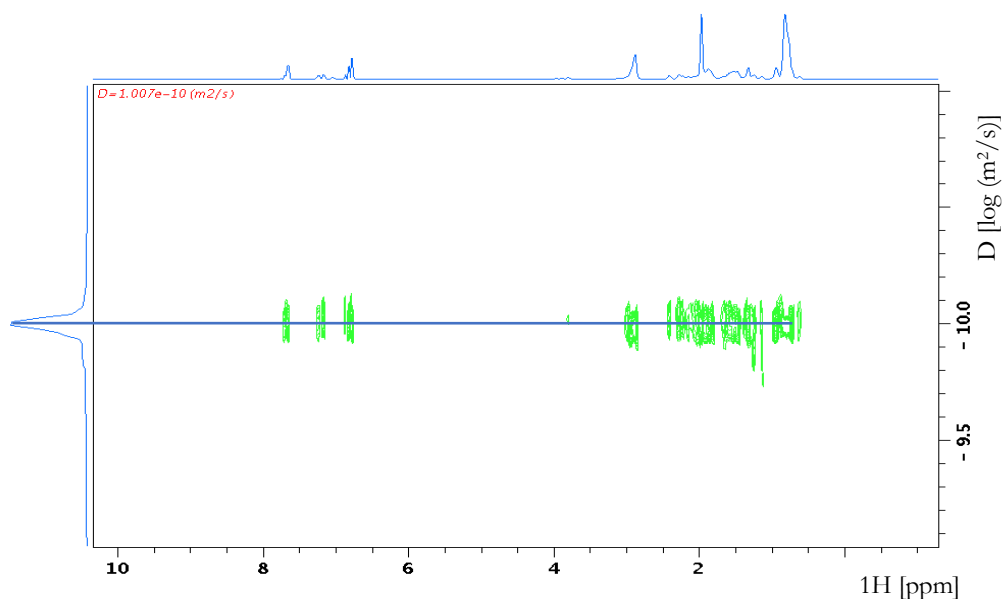


Figure 22. ^1H DOSY spectrum of the RBD-NS1 at 25°C . Spectrum was collected at a ^1H frequency of 400 MHz with a protein sample at a 0.12 mM concentration.

The value obtained from the diffusion coefficients experiments was $1.007 \times 10^{-6} \text{ cm}^2/\text{s}$. It is possible to determine the value of the diffusion coefficient from the Stokes-Einstein Law:

$$D = 8.34 \times 10^{-8} \left(\frac{T}{\eta M^{\frac{1}{3}}} \right)$$

where the value of M is 21.3 kDa, that corresponds to the molecular weight of the dimeric form of the RBD. The value obtained from the equation was $1.007 \times 10^{-6} \text{ cm}^2/\text{s}$, what is in complete agreement with the experimental value.

These results are in agreement with the results in SEC-MALLS proving that the RBD, in aqueous solution, is in dimeric form.

3.3.3.5. Fluorescence

One of the aims of this project is to experimentally validate with NMR the *in silico* approach followed. Many of the organic compounds identified and obtained are soluble in DMSO. So to access the stability of RBD in solution with different percentages of DMSO, studies of fluorescence were performed (Figure 23), since the presence of DMSO hinders the collection of far UV CD spectra.

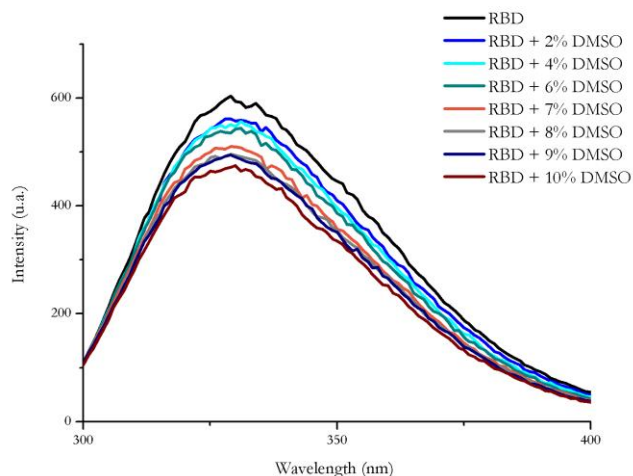


Figure 23. Fluorescence spectra of NS1-RBD in the presence of different percentages of DMSO at pH 6.9. Spectra were collected at 25°C. The RBD concentration was 15 μM , the excitation wavelength was 280 nm and cuvette pathlength was 5x5 mm. Baseline spectra with buffer and various percentages of DMSO were also acquired and subtracted to the raw data.

The fluorescence spectra of RBD in the absence and in the presence of different percentages of DMSO are very similar in shape and have the same emission maxima, which indicates that no significant changes in protein structure occur and that RBD is stable in solution containing DMSO up to 10%. This provides the possibility of using various percentages of DMSO to solubilize the molecular fragments without compromising the structure of the protein, in the NMR validation experiments.

4. Virtual Fragment Screening

4.1. Introduction

4.1.1. Fragment-based lead discovery

Fragment-based lead discovery is becoming a powerful tool to aid the discovery of new small-molecule therapeutics, establish the druggability of a biological target, and discover alternative inhibition sites on already established targets⁵³.

A “fragment” is a particularly small organic molecule, presenting relatively weak affinity for certain protein region, and that may become part of a larger, high-affinity molecule. In some cases, these fragments are part of known drugs that have been used as starting points to find new inhibitors for different biological targets⁵⁴. Typically, fragments are aromatic or ring heterocyclic compounds with a molecular weight inferior to 300 Da, soluble and chemically stable. The success of fragment-based lead discovery is based on two rationales. First the chemical space can be probed much better when a smaller threshold for the maximal molecular weight is chosen. A better sampling of the chemical space also leads to the improvement of the hit rates in fragment-based screening. Second, fragments hits are weak binders, so they must form high quality interactions with the protein in order to bind sufficiently for detection with biophysical techniques.

FBLD is based on the notion that a fragment can be linked or grown with another fragment to improve potency.

Researchers from Astex Pharmaceuticals proposed a set of empirical rules to define a “fragment”, known as “Rule of 3”⁵⁵, which demands a molecular weight lower than 300 Da, number of hydrogens bond donors and acceptors under 3, calculated log P under 3, number of rotatable bonds under 3 and polar surface area under 60.

The optimization of fragment hits to develop candidates typically includes a significant increase in affinity that is accompanied by important changes in other molecular properties relevant in drug discovery. The primary objective of early-phase fragment optimizations is to increase affinity toward the target, and the optimization almost inevitably results in an increase in both molecular weight and lipophilicity. The excessive increase of these latter properties, however, leads to suboptimal pharmacokinetic profiles. Therefore, their control during optimization is highly desirable to ensure balanced optimization of affinity, molecular weight, and lipophilicity. Various efficiency indices have been proposed for the simultaneous monitoring of these properties. Efficiency indices are

typically composite measures that can characterize the balance of the above properties. They are especially useful at decision points like hit and lead selection and in the early phases of optimizations when compound properties from various series must be compared and the result of significant structural modifications has to be assessed.

Ligand Efficiency (LE) is encoded by a ratio between an affinity metric (e.g. pIC₅₀) and the number of non-hydrogen atoms:

$$LE = \frac{\Delta G_{\text{binding}}}{HA} \approx \frac{pIC_{50}}{HA}$$

where IC₅₀ is the measure of the effectiveness of a substance in inhibiting a specific biological or biochemical function and HA is the number of heavy atoms, meaning the non-hydrogen atoms.

As non-hydrogen atoms can be of many different types and a key property of a compound is its molecular weight (MW), an extension of the concept of ligand efficiency has been introduced recently that is based on expressing the binding affinity as pK_i/pK_d/pIC₅₀ and using MW as reference expressed in kDa. Four related efficiency indices have been proposed: binding efficiency index (BEI), surface efficiency index (SEI), lipophilic ligand efficiency (LLE) and ligand efficiency lipophilicity (LELP)⁵⁶. BEI is the binding efficiency index relating potency to molecular weight on a per kDa scale:

$$BEI = \frac{pIC_{50}}{MW}$$

where MW is the molecular weight in kDa. SEI is the surface efficiency index monitoring the potency gains as related to the increase in polar surface area (PSA) referred to 100 Å:

$$SEI = \frac{pIC_{50}}{\left(\frac{PSA}{100\text{\AA}}\right)}$$

where PSA is the polar surface area. Lipophilic ligand efficiency (LLE) defined as the difference of the negative logarithm of a potency measure (pK_d, pK_i, or pIC₅₀) and log P (or log D):

$$LLE = pK_{d,i} \text{ or } pIC_{50} - \text{clogP}$$

LLE describes the contribution of lipophilicity to potency. Compounds with reduced complexity, like fragments are typically polar compounds often with limited potency that makes their LLE less desirable. As a consequence, comparative evaluation of these compounds, that are otherwise considered to be promising, is challenging. This

limitation of LLE is due to the neglected effect of ligand size that calls for alternative metrics, especially in the case of fragments. The concept of lipophilicity-corrected ligand efficiency is defined by

$$\text{LELP} = \frac{\log P}{\text{LE}}$$

i.e. as the ratio of log P and ligand efficiency (LE), therefore depicting the “price” of ligand efficiency “paid” in lipophilicity (encoded by the log P). LELP is meaningful for log P values typical in most of the discovery programs and allows the evaluation of both fragments, lead-like and drug-like compounds⁵⁷.

4.1.2. Computational methods for FBLD

In silico fragment-based approaches are loosely defined and may involve the use of a range of techniques spanning from cheminformatics analysis, such as in the construction of virtual fragment library for screening, through the use of more traditional molecular modelling methods, like (fragment) docking, and solvent-mapping techniques to identify *hot spots* for fragment binding, all the way to the use of structural bioinformatics to study protein structures bound to fragments.

A successful fragment-based lead discovery campaign requires a fragment library possessing several important characteristics, including proper collection size, specific physicochemical properties and chemical diversity. A fragment library range in size from 10^2 to 10^4 . Physicochemical factors to be considered in building a fragment library are molecular weight, number of rotatable bonds, lipophilicity, number of hydrogen bond donors and acceptors and polar surface area. The physicochemical factors should obey the “Rule of Three” described in the section 4.1.1..⁵⁸

Solvent mapping is a fragment-based active site mapping method that consists on the identification and characterization of important regions, often called *hot spots*, that are characterized by the substantially contribution to the binding free energy in the binding pocket of the drug target⁵⁹. There are two major classes of methods for active site mapping: geometric algorithms and probe mapping algorithms. For the latter, fragments are used as molecular probes to detect hot spots and consists in moving small organic functional groups around the protein surface and determining their most energetically favourable binding positions.

FTMap is a well-established webserver for active site mapping by probe mapping. It attempts to identify low energy clusters of probe molecules on protein surfaces. A set of 16 small molecules are used to probe the target protein surface and hotspots are then identified as sites where most probe molecules likely bind⁶⁰.

One important step in a fragment-based lead discovery campaign consists in docking the fragments⁶¹. In this step, a docking program is used to place computer-generated representations of a small molecule into the active site of the target, in a variety of positions, conformations and orientations. Each predicted binding mode is often called a *pose*. To identify the energetically most favourable pose, each pose is typically scored based on its complementarity to the target in terms of shape and properties such as electrostatics⁶². A “good” score for a given molecule indicates that it is potentially a “good” binder.

Fragment poses extracted from docking experiments can be used during fragment linking methods, since the best poses can be grown to take advantage of nearby pockets. Docking presents some benefits over experimental fragment screening: it is a low cost, fast technique, which escapes solubility concerns daunting experimental screening. It also allows one to screen compounds that have not been purchased or yet synthesized.

Despite its benefits, fragment docking presents big challenges. Naturally, one of them, is that fragments are more difficult to dock than drug-like molecules because fragments are weak binders and it is harder to differentiate between native and low-energy poses. Moreover, fragments are more expected to occupy binding sites on a protein surface that can accommodate low molecular weight compounds with limited specificity. The biggest challenge in fragment docking, however, is related to the choice of scoring function, since most docking scoring functions have been parameterized using drug-like compounds with molecular weight and other chemical properties significantly different from those of fragments. While it can be argued that docking scoring functions are, in general, poor predictors of ligand binding affinity, even when dealing with drug-like compounds, this may still be one of the reasons for their poor performance in predicting the binding affinity of fragments for proteins.

4.1.3. An innovative approach to fragment screening based on IsoMIF

In this work, a new fragment-based screening protocol is proposed. The protocol is grounded on an analysis of similarity between binding sites predicted for the NS1 protein and many fragment-bound protein structures deposited in online repositories – such as PDB and PDBbind – using molecular interaction fields (MIFs).

Traditionally, MIF analysis have been mostly used to identify energetically favourable interaction sites on a macromolecular target^{63,64}. MIFs depict the physicochemical environment of a protein surface, which is often characterized by three-dimensional descriptors such as hydrophobicity, hydrogen bond donors and acceptors and positive and negative charge isocontour maps. MIFs can be calculated with distinct tools, such as the software GRID and IsoMIF.

The number of protein structures deposited in the protein data bank (PDB) is gradually growing. Following the hypothesis that proteins with similar binding sites and/or biochemical functions may bind similar ligands, local similarities within binding sites can be identified by comparing molecular recognition features in the binding sites, like size, shape, and physicochemical properties.

In general, binding site comparisons are carried out in three main steps: 1) encoding of the molecular recognition features of the binding site of two proteins (e.g. the target protein and one database protein), i.e. the “binding site representation”, 2) alignment between the two binding site representations, and 3) quantification or scoring of similarity between the two sites. The first step determines the molecular features of the binding site. In the subsequent step, the optimal superposition of two binding site representations is determined, often by methods like geometric hashing⁶⁵ or graph-based clique detection methods⁶⁶. In the final similarity scoring step, the degree of similarity between the superimposed representations is quantified.

In this work, we explore a new methodology called IsoMIF, where MIFs are calculated using six chemical probes representing hydrophobic, aromatic, hydrogen bond donor/acceptor, and positively/negatively. Similarities can be identified using an approximation of the Bron and Kerbosch⁶⁷ graph-matching algorithm and scored with a Tanimoto coefficient of the matched probes in the largest clique. The Bron and Kerbosch graph-matching algorithm is used to detect the maximum common subgraph (MSC) isomorphisms, by measuring the largest ensemble of vertices between two cavities that

have corresponding interaction types and are in geometrically equivalent positions. A node in the association graph is a pair of vertices, one from each of the two MIFs being compared that have at least one energetically significant common interaction. An edge is drawn between two nodes in the association graph if the distances between the two corresponding vertices in each MIF are within 3.0 Å. This distance threshold allows accounting implicitly for geometric variability between similar binding sites that is the result of conformational flexibility.

MIF similarities score (MSS) of a clique is calculated as a Tanimoto score:

$$\text{MSS} = \frac{N_c}{N_a + N_b - N_c}$$

where N_c is the sum of common probes in vertices belonging to the clique and represent the set of potential intermolecular interactions in equivalent geometric position in the two MIFs. N_a and N_b represent the sum of energetically significant probes in each of the two MIFs under comparison.

The IsoMIF method was used to compare the binding site of the RNA-binding domain with the binding site of the proteins from the protein-fragments complexes in the PDBbind database, and then the fragments are extracted from the complexes protein-fragments and used in the posterior phases of the computational protocol: docking, so the computational protocol is based on a reverse screening strategy.

4.1.4. NMR for FBLD

The relative weakness of the fragment-target interaction requires biophysical methods very sensitive to detect binding. Examples of such techniques include Nuclear Magnetic Resonance (NMR) (ligand- and protein-based), X-ray crystallography, Surface Plasmon Resonance (SPR) and Isothermal Titration Calorimetry (ITC). X-ray crystallography is commonly used in structural biology and plays an important role in the identification of fragment hits. Fragments can be soaked into the crystallized protein and after diffraction the fragment will be visible in the resulting electron density map. This method allows for the most detailed analysis, at the atomic level, of the mode of binding of a fragment to a protein structure.

SPR can be used to study biomolecular interactions, providing information about kinetics as well binding affinity. SPR-based biosensors are sufficiently sensitive and high throughput to provide complete fragment screens on libraries of several thousand

compounds in just a few weeks per target. Biosensors provide quantitative binding information for ranking fragments by affinity and ligand efficiency and can support ongoing quantitative structure–activity efforts during fragment hit-to-lead development.

ITC is a thermodynamic technique that measure the heat absorbed or released during a binding event. ITC has been used as a fragment screening tool and allows the determination of binding affinity⁶⁸.

NMR spectroscopy can be used to study the structure, function and dynamics of proteins. In fragment-based lead discovery, NMR spectroscopy is one of the biophysical techniques used to detect protein-ligand interactions and was the technique chosen to detect protein-fragment binding in this work. Direct assessment of protein resonances unveils where in the protein an interaction occurs. To do so, it is necessary to assign the resonance of backbone amides, in order to map the specific residues in the protein that are involved in the interaction – which in turn requires isotope-labelling of the protein.

The most common protein-based NMR technique is bidimensional (2D) ¹H-¹⁵N heteronuclear single quantum correlation (HSQC)⁶⁹ of ¹⁵N-labelled proteins. In this experiment, the amides' NH group is observed, which allows the monitoring of binding events. 2D HSQC spectra in the absence and in the presence of small-molecule compounds can be collected. This simple approach allows obtaining chemical shift perturbation (CSP)⁷⁰ data, exposing binding events and sites of interaction in the protein.

The HSQC spectra provides the correlation between the nitrogen and amide proton and each amide yields a peak in the HSQC spectra. In general, each residue would produce an observable peak in the spectra, with the exception of proline which lacks an amide proton, and normally the N-terminal residue, which has a free NH₃⁺ group attached).. In addition to the backbone amide resonances, sidechains with nitrogen-bound protons will also produce peaks.

In a typical HSQC spectrum, the NH₂ peaks from the sidechain of asparagine and glutamine appear as doublets on the top right corner, and a smaller peak may appear on top of each peak due to deuterium exchange from the D₂O normally added to an NMR sample, conferring these sidechain peaks a distinctive appearance. The sidechain amide peaks from tryptophan are usually shifted downfield and appear near the bottom-left corner of the spectra. The backbone amide peaks of glycine normally appear near the top of the spectrum.

The chemical shift perturbation (CSP) assay has been used extensively to identify binding sites of small molecules. This approach gained popularity when SAR by NMR was introduced in 1996 by Shucker⁷¹. SAR by NMR (Figure 24) first uses CSP data from weakly interacting compounds in order to optimize them for a given site in the protein. The second step is to find adjacent sites in the protein where another fragment is binding and optimize it as far as possible. The third step is to disclose the orientation of the bound ligands in order to guide their linkage and elaboration and maintain this orientation in the final compound, thereby achieving high specificity to that target. This technique allows high-affinity ligand elaboration and reduces the laborious chemical synthesis necessary to achieve high potency. The SAR by NMR method has facilitated the development of highly potent and specific compounds and it continues to be one of the most popular and successful NMR techniques for FBLD^{72,73}.

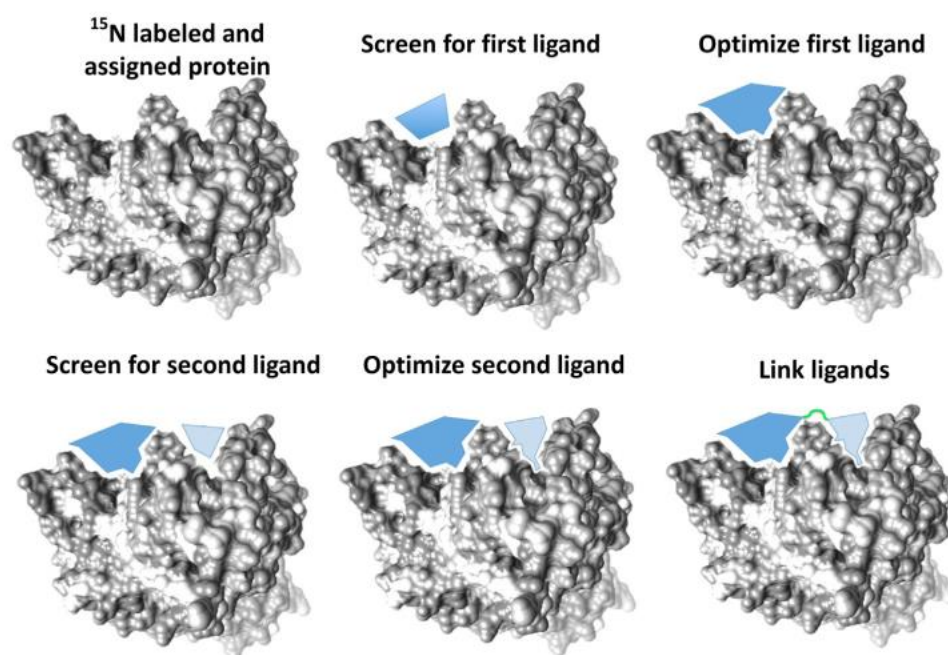


Figure 24. Schematic representation of SAR by NMR⁷⁵.

4.2. Methods

4.2.1. Assembly of fragment-bound protein screening library

The database chosen to search for protein-bound molecular fragments was PDBbind⁷⁶. The PDBbind database is updated annually in to keep up with the growth of PDB. In this project, the PDBbind database, version 2016 (downloaded on 15/12/2016),

was used because, aside the X-ray structure itself, it also provides the experimental values of protein-ligand interaction.

Filter-itTM ⁷⁷ is a program for filtering out database molecules with unwanted molecular properties – or retaining molecules with desired molecular properties. In this project, Filter-it was used to filter the PDBbind according to the “Rule of Three” for fragment-like molecules (Table 3).

Table 3. Parameters derived from “Rule of three” and used to construct the fragment library.

Molecular Weight	100-300 Da
Number of Rotatable bonds	0 - 3
Hydrogen bond donors	0 - 3
Hydrogen bond acceptors	0 - 3
CLogP	-3 - 3
TPSA	0 - 60

4.2.2. Fragment screening via binding-site similarity analysis

In this work, we envisaged and deployed a computational fragment screening protocol based on the high-throughput comparison of protein binding site Molecular Interaction Field (MIF) similarities. Our protocol is not a fragment-centric virtual screening protocol, since the actual screening is grounded on the comparison of MIF similarities between NS1 binding sites and all binding sites detected in our fragment-bound protein library. IsoMIF was used as the main engine of our protocol ⁷⁸. IsoMIF is a stand-alone software package divided into three distinct programs: GetCleft, MIF and a program actually called IsoMIF. GetCleft searches for cavities in the protein surface that may be relevant for protein-ligand interactions. MIF computes molecular interaction fields (MIFs) for multiple probes within the cavity identified by GetCleft. In this work, the six chemical probes implemented in IsoMIF were used:

- hydrophobic
- aromatic
- hydrogen bond donor
- hydrogen bond acceptor

- positive charge
- negative charge

Finally and most importantly, the IsoMIF program computes MIF similarities via alignment of the various MIF fields for all the different probes. The measurement of similarity between binding sites is calculated using Tanimoto coefficients, defined by

$$MSS = \frac{N_c}{N_a + N_b - N_c}$$

where N_c is the sum of common probes in vertices belonging to the clique and represent the set of potential intermolecular interactions in equivalent geometric position in the two MIFs. N_a and N_b represent the sum of energetically significant probes in each of the two MIFs under comparison.

The software IsoMIF requires several parameters to proceed to the calculations of the cavities in the protein surface, the molecular interaction fields and, finally, the binding site similarity. The GetCleft tool to proceed to the calculation of the cavity in contact with the ligand, in the case of the protein-fragment complexes from PDBbind, requires the definition of the ligand code, pose and the chain where it interacts. The IsoMIF tool requires the following parameters to calculate the similarity between the binding site of RBD and the binding site of the proteins from PDBbind: the coarse-grain step sequence 1, that corresponds to 1.5 Å, and the max cliques of 1000.

Input protein PDB files were processed with the program Reduce, in order to add the hydrogens atoms⁷⁷. In this step, OH, SH, NH₃⁺ groups and methionine methyl groups, and the side-chains of asparagine, glutamine, and histidine were reoriented to optimize hydrogens bonds and eschew van der Waals overlaps.

All calculations were performed in a Fedora 22 64-bit linux box with the following specifications: 24 Gb of memory, Intel Xeon model and CPU running at 3.4 GHz. Copies of the scripts are provided in Annex C. The computational experiences were performed in duplicate.

4.2.3. Fragment hit confirmation via docking

In this project, SEED (Solvation Energy for Exhaustive Docking) was chosen for post-screening analysis via docking of selected fragments into NS1 binding pockets⁷⁹. The docking approach implemented in the program SEED determines the optimal positions and orientations of small-sized molecular fragments in the binding site of the protein.

SEED is classified as an exhaustive search method for fragment docking. SEED uses polar and apolar vectors to describe the fragment and the binding site. Polar vectors are defined as originating from a polar atom. The length and the orientation of the vectors are then based on all the favourable angles and distances to establish an H-bond based on the involved atom types. Vectors that point towards occupied regions of space (i.e. receptor) are discarded. The sampling phase consists of matching the polar vectors of the fragment with the ones from the binding site, so that the distance between the H-bond donor and the H-bond acceptor is favourable with respect to the atom types. The fragment is then rotated around the H-bond axis and the user has control over the number of rotations around each axis. Apolar vectors are defined in a two-steps procedure. First, points are distributed uniformly on the solvent-accessible surface (SAS) of the receptor binding site and the ligand. Secondly, a low dielectric sphere (probe) is run over the aforementioned points in order to evaluate the desolvation energy, and the van der Waals (vdW) interaction with the receptor. Only the best points according to the two energetic terms are kept. Apolar vectors for the fragment and the receptor are then defined by joining each point on the SAS with its corresponding atom centre. The sampling consists of matching apolar vectors of the fragment with the ones of the binding site, so their van der Waals distance is optimal. As in the case of polar vectors, the fragment is then rotated around the axis defined by the fragment atom and its receptor counterpart.

System Set-up

The structure of RNA-binding domain of NS1 was taken from the PDB database (code 2N74). The water molecules were removed. Hydrogen atoms were added with the molecular modelling program WITNOTP. Partial charges were assigned to the protein and to the fragments with the MPEOE method implemented in WITNOTP.

SEED uses an input file with the amino acids that form the binding site of the target assigned (input file shown in Annex D) and the fragments to use in docking indicated. SEED also requires the identification of the docking type, apolar, polar or both.

4.2.4. Fragment hit validation via NMR

NMR spectroscopy is a powerful tool for fragment hit validation. As the ligand is added to a protein, some chemical shifts are perturbed. Usually, these belong to amino acids close to the interaction surface. However, it is important to remember that a change in chemical shifts merely implies that there is a change in the magnetic environment of a nucleus, not a direct interaction with a binding partner. Thus, if a protein undergoes substantial structural rearrangement upon complex formation, widespread chemical shift perturbations may be observed, including in residues which are far from the interaction site (but which nonetheless experience a change in their structural environment). Chemical shift mapping is a very straight forward method and can provide information about both the location and strength of a binding event.

4.2.4.1. Experimental procedure for hit validation via NMR

The chemical shift perturbation experiments of RBD in the presence of fragments were performed at 25 °C on a Varian VNMRs 600 MHz spectrometer. NMR samples contained 0.40 mM of RBD in PBS (100 mM NaCl and 20 mM NaPi) , 5% of D₂O, 25 mM of DSS and 0.025% of Sodium Azide in a final volume of 400 μ L. All fragment stock were prepared in DMSO-d₆ and the NMR experiments were performed at 1:1 RBD-ligand stoichiometry in 2% of DMSO, and also at 1:2 in 5% DMSO, in the final sample. The NMR characterization of the fragments is shown in annex E and F.

4.3. Results and discussion

4.3.1. Assembly of fragment-bound protein screening library

Because no reference ligands of NS1-RBD are currently known, we have decided to follow a fragment-based approach with the aim of probing RBD's propensity to interact with specific chemistry – and thus identify chemical moieties that may be merged to form high affinity tool compounds. With this in mind, we have assembled a library of proteins bound to compounds that obey the fragment's "Rule of Three" (Ro3). The source of such complexes was a web resource called PDBbind⁷⁵. This repository was chosen because besides the X-ray structures deposited therein, it also stores experimental affinity and/or bioactivity values relating protein-ligand interactions, such as K_d , K_i or IC_{50} .

Astex's "Rule of Three" was employed to filter a set of 13285 structures downloaded from PDBbind, which involved calculating the number of rotatable bonds, of

hydrogen bond donor/acceptors, molecular weight, cLogP and TPSA. Of the 13285 structures, 509 structures contained the compounds that obey the imposed Ro3 filters (Figure 25).

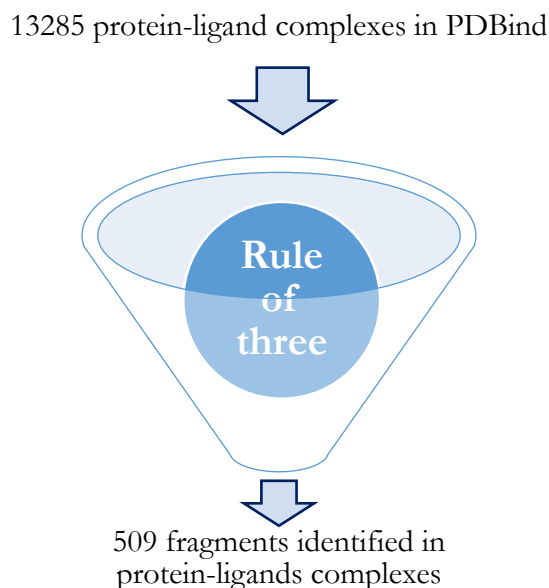


Figure 25. Scheme of the filtration step in the assembly of the fragment-bound library.

Physicochemical properties distribution

According to the literature⁸⁰, fragments must have less or equal to 16 heavy atoms. Figure 26A shows the distribution of the number of heavy atoms of the fragments retained in our screening library. As can be seen from the histogram, the vast majority of fragments in the library is below the 16 heavy atoms threshold. In total, only 79 fragments violate the 16 HA threshold.

Equally, the distribution of the molecular weight of the 509 fragments shows that the majority of fragments have less than 200 Da (Figure 26B) and more than 1 of cLogP (Figure 26C). In respect to TPSA, most fragments fall within the 50 Å² range, and no fragments are above 60 Å² threshold (Figure 26D). Conforming to hydrogen bond type, the largest clusters of fragments present two (2) hydrogen bond acceptor atoms and one (1) hydrogen bond donor atom (Figure 26E/F). The distribution of the number of rotatable bonds is approximately similar for all the fragments from the collection (Figure 26G).

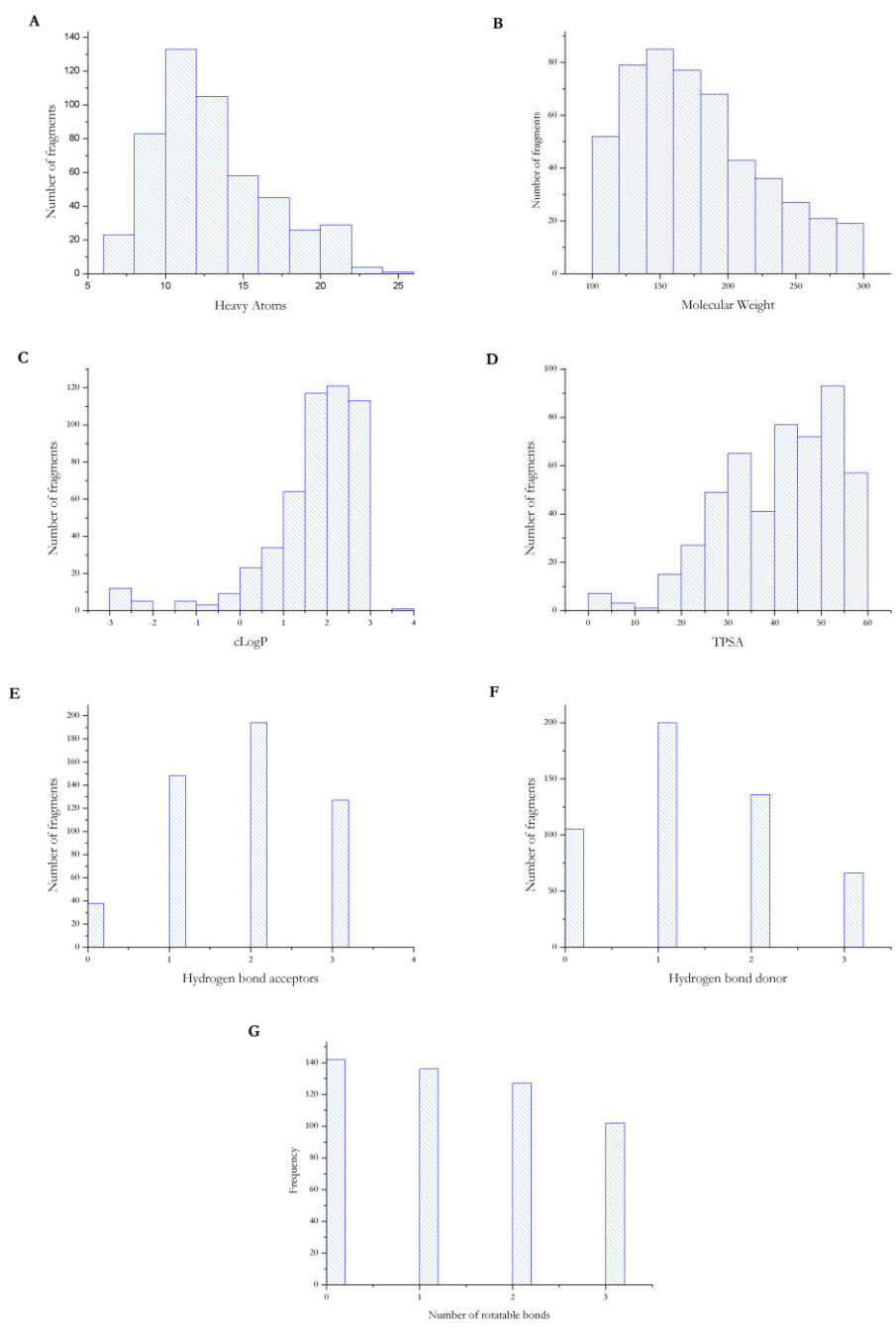


Figure 26. Physicochemical property distribution of the fragments contained in the assembled fragment-bound library.

Ligand efficiency indices

Typically, fragments have weak binding affinity for proteins. Affinity is increased by growing new functional groups or by linking two hit fragments bound in adjacent pockets. Ligand efficiency indices are widely-used for benchmarking different hit fragments and thus guide the lead generation and optimization.

In the literature⁸¹, various ligand efficiency indexes have been proposed, such as LEI (ligand efficiency index), BEI (binding efficiency index), SEI (surface efficiency index) and LLE (ligand lipophilicity efficiency).

Figure 27 shows the relation between BEI and LEI for the 509 fragments assembled. It is possible to conclude there is a linear correlation between these two ligand efficiency indices. According to the literature⁸², BEI values should be higher than 19.5 and LEI should be above 0.37. Figure 27 shows that the majority of fragments bound to PDBbind complexes follows the guideline previous described.

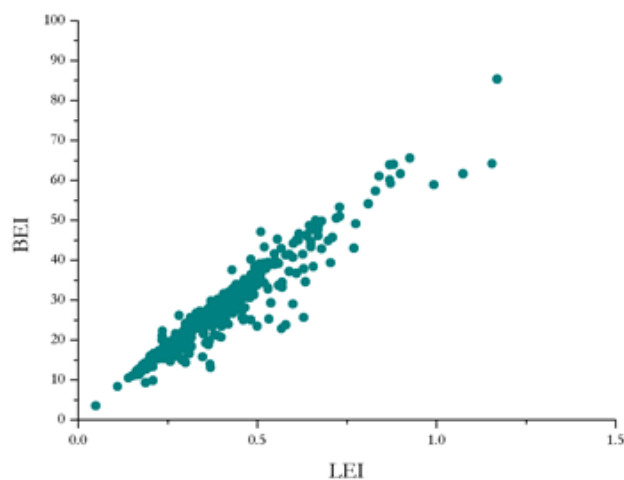


Figure 27. Plot of BEI vs LEI for the 509 fragments in the assembled virtual library.

In respect to the surface efficiency index (SEI) (Figure 28), which tries to capture ligand efficiency using the molecule's polar surface area (PSA) as a descriptor of molecular size, the highest number of fragments is concentrated in the SEI interval 5-20, suggesting a prospect of good permeability.

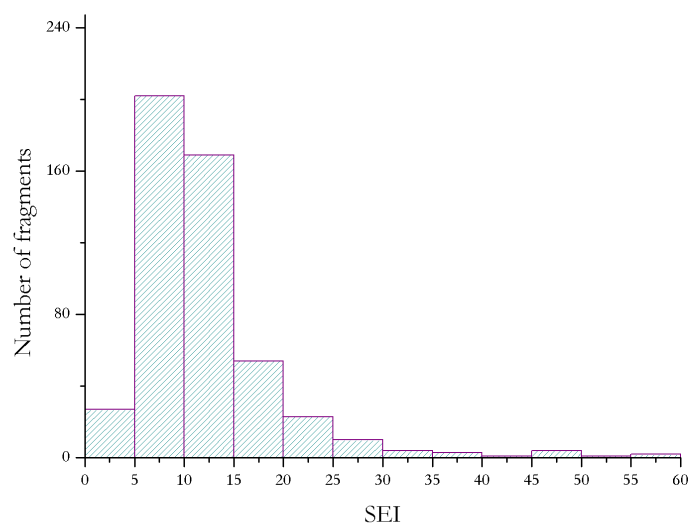


Figure 28. Distribution of Surface Efficiency Index across the fragments retained in our screening library.

Ligand lipophilicity efficiency (LLE)⁸³ has been proposed to assess ‘medchem-friendliness’ of small organic molecules and provides a way to evaluate a ligand’s efficiency in binding (affinity) as a proportion of its lipophilicity. The challenge is to increase potency without increasing lipophilicity, for lipophilicity is one of the major factor for compound promiscuity.

According to the literature⁸³, the preferable values for LLE are higher than 3. LLE values calculated for the retained fragments from PDBbind follow this guideline, since there is a large number of fragments holding LLE values over 3 (Figure 29).

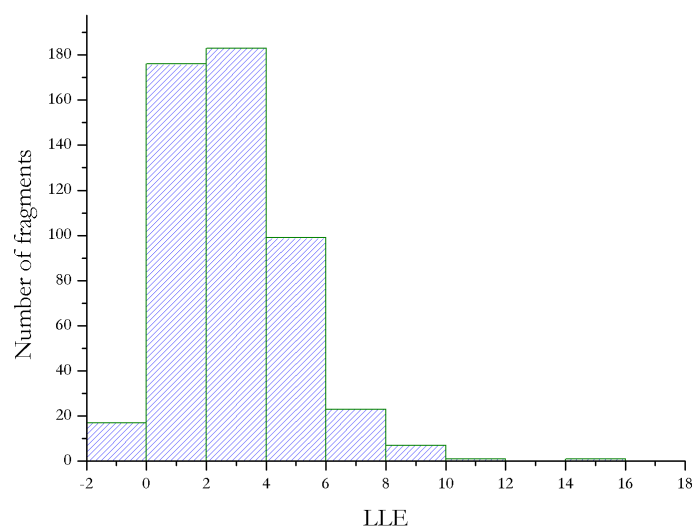


Figure 29. Distribution of Ligand Lipophilicity Efficiency across the fragments retained in our screening library.

4.3.2. Fragment Screening via binding-site similarity

An exemplary representation of MIF probe points calculated within the main binding cavity in the RBD surface is provided in Figure 30.

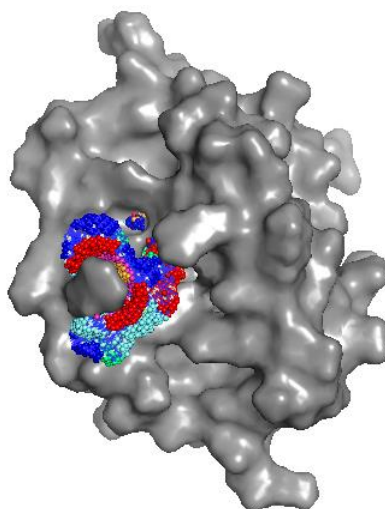
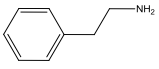
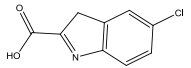
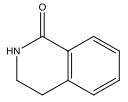
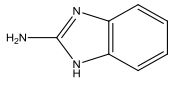
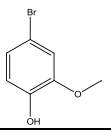


Figure 30. Surface representation of RBD and molecular interaction fields calculated using MIF software. The colour cyan corresponds to the hydrophobic probe; orange corresponds to the aromatic probe; blue corresponds to the hydrogen bond donor probe; red corresponds to the hydrogen bond acceptor probe; the positive charge probe is represented by the colour green and the negative charge probe is represented by the colour magenta.

MIFs on binding cavities of the fragment-bound proteins in our screening library (509 structures) were calculated using six probe types: H-bond donor/acceptor, aromatic,

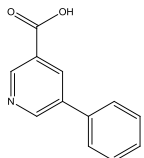
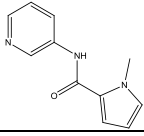
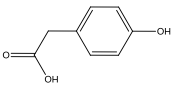
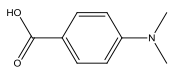
hydrophobic and positively/negatively charged probes. These MIFs were compared with our query MIF, which is the binding pocket of NS1-RBD with dsRNA. All protein structures analysed were ranked and arranged by binding site similarity – quantified via Tanimoto scores. The top-5 fragments belonging to the binding sites with highest values of Tanimoto Score were chosen to perform the next phase of the computational protocol: docking (Table 4).

Table 4. Summary of calculated ligand efficiency indexes, LE and BEI, for the top-5 fragments extracted from protein binding sites holding the highest similarity (Tanimoto score) with the main cavity in NS1-RBD.

Fragment	Tanimoto Score	LE	BEI
	0.5625	0.33	24.65
	0.5476	0.188	12.62
	0.5250	0.46	32.20
	0.5172	0.37	27.78
	0.50	0.19	9.28

Fragments belonging to protein binding sites with lower Tanimoto score were also chosen to validate the method (Table 5).

Table 5. Summary of the calculated ligand efficiency indexes, LE and BEI, for the 5 fragments extracted from protein binding sites holding the lowest similarity (Tanimoto score) with the main cavity in NS1-RBD.

Fragment	Tanimoto Score	LE	BEI
	0.186	0.41	22.4
	0.222	0.49	27.17
	0.2273	0.29	15.22
	0.2063	0.53	39.28

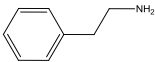
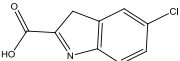
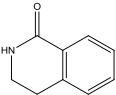
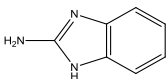
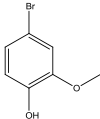
4.3.3. Fragment hit confirmation via docking

For screening of fragment libraries, Caflisch and colleagues have developed a suite of programs - called SEED that dock and score fragment poses. In this project, fragment docking was used to predict the positioning and orientation of the chosen fragments within the identified cavities of NS1-RBD.

To perform fragment docking, SEED requires the definition of the binding site. In this step of the protocol, the binding site was defined by the amino acids in the proximity of the pocket/cavity identified by DoGSiteScorer and GetCleft (Chapter 3), as well as by the amino acids considered important for RBD-dsRNA interactions that are not in the proximity of the cavity identified by GetCleft and DoGSite Scorer.

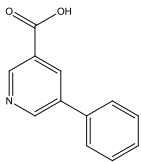
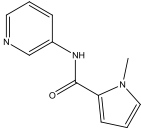
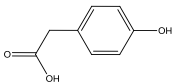
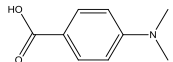
The SEED score of the fragments extracted from the binding sites with higher Tanimoto score are summarized in Table 6.

Table 6. Docking scores for the top-5 fragments extracted from protein binding sites holding the highest Tanimoto Scores.

Name	Fragment	Tanimoto Score	Docking Score
2-phenylethylamine		0.5625	-4.57 kcal/mol
5-chloro-1H-indole-2-carboxylic acid		0.5476	-9.76 kcal/mol
7-chloro-3,4-dihydroisoquinolin-1(2H)-one		0.5250	-12.30 kcal/mol
2H-benzimidazol-2-amine		0.5172	-10.25 kcal/mol
4-bromo-2-methoxyphenol		0.50	-10.05 kcal/mol

Equally, docking was performed for the fragments belonging to the binding sites with lowest Tanimoto score. These results are summarized in table 7.

Table 7. Docking scores for fragments extracted from protein binding sites holding the lowest Tanimoto Scores.

Name	Fragment	Tanimoto Score	Docking Score
5-phenylpyridine-3-carboxylic acid		0.186	-7.53 kcal/mol
1-methyl-N-(pyridin-3-yl)-1H-pyrazole-5-carboxamide		0.222	-5.45 kcal/mol
4-Hydroxyphenylacetate		0.2273	-8.64 kcal/mol
4-(dimethylamino)benzoic acid		0.2063	-7.38 kcal/mol

4.3.3.1. Fragment hit validation via NMR

The ultimate goal of this project was the experimental validation of the computational predictions by Chemical Shift Perturbation experiments. In Figure 31 is shown the HSQC of free RBD, represented in red, the first addition of 4-hydroxyphenylacetate, represented in green, and the second addition, represented in grey.



Figure 31. ^1H - ^{15}N HSQC of free RBD and first (1:1) and second additions (1:2) of 4-hydroxyphenylacetate. Free RBD is represented in red, first addition in green and second addition in grey.

The amino acids that show highest chemical shift perturbation are Glu55, Ile54, Ala57, Arg59 and Asp29.

The results obtained during the experimental validation step were then compared with the results obtained during the step of druggability assessment. DoGSiteScorer identified a subpocket that is composed by the amino acids Glu55 and Ile54 that has a volume of 102.98 \AA^3 , depth values of 7.64 \AA^2 and a *drugscore* value of 0.2 (Figure 32). The pocket composed by this subpocket presents a *druggability* value of 0.82. The pocket has a volume of 1022.34 \AA^3 and a depth value of 22.81 \AA^2 .

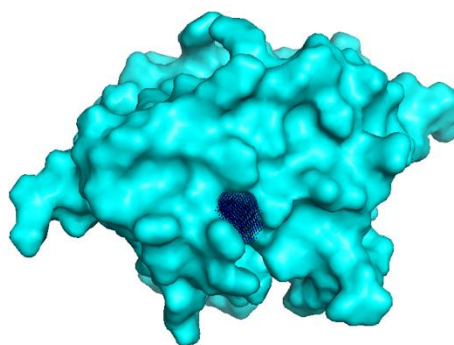


Figure 32. Representation of the RBD surface with the subpocket composed by the amino acids Ile 54 and Glu55 represented in dark-blue.

Other amino acids, such as Ala57 and Arg59, are near of a pocket that has a volume of 667.07 \AA^3 , a value of depth of 17.10 \AA^2 and a druggability value of 0.79 (Figure 33).

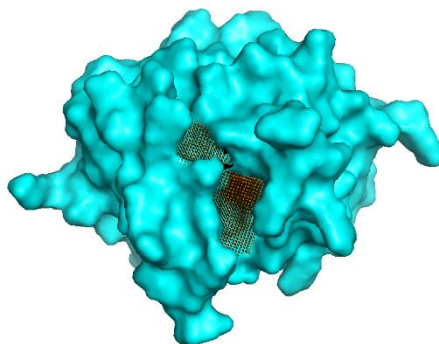


Figure 33. Representation of the RBD surface with the subpocket composed by amino acids Ala57 and Arg59 represented in orange.

Chemical shift perturbation experiments were also performed with 7-chloro-3,4-dihydroisoquinolin-1-(2H)-one, since the binding site of the protein containing this fragment has high similarity with the binding site of RBD.

In Figure 34 is shown the ^{15}N - ^1H HSQC spectra of free RBD, represented in red, and the ligand titrations for 7-chloro-3,4-dihydroisoquinolin-1(2H)-one. The first addition, corresponding to a stoichiometry of 1:1, is represented in green, and the second addition, corresponding to a stoichiometry of 1:2, is represented in grey.

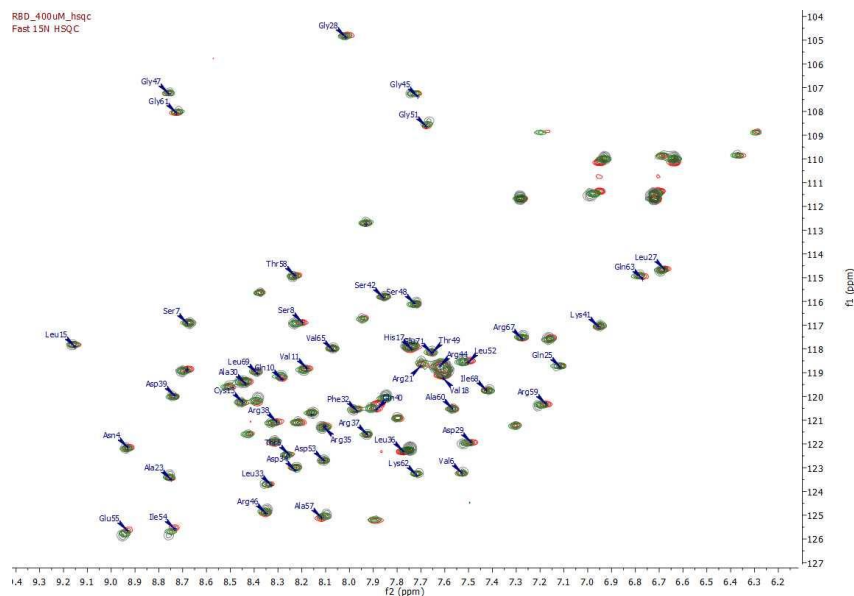


Figure 34. ^1H - ^{15}N HSQC of free RBD and first (1:1) and second additions (1:2) of 7-chloro-3,4-dihydroisoquinolin-1(2H)-one. Free RBD is represented in red, first addition in green and second addition in grey.

The amino acids, such as Trp16, Arg35, Leu36, Gln40 and Leu43, that shown chemical shift perturbation are in contact with the two distinct subpockets. One has a volume of 304.83 \AA^3 , a depth value of 525.73 \AA^2 and a druggability value of 0.27. The other pocket has a volume of 293.63 \AA^3 , a depth value of 571.15 \AA^2 and a druggability value of 0.22. Both subpockets belong to the same pocket that has the following characteristics: volume equals to 1022.34 \AA^3 , depth equals to 1507.46 \AA^2 and a druggability value of 0.82 (Figure 35).

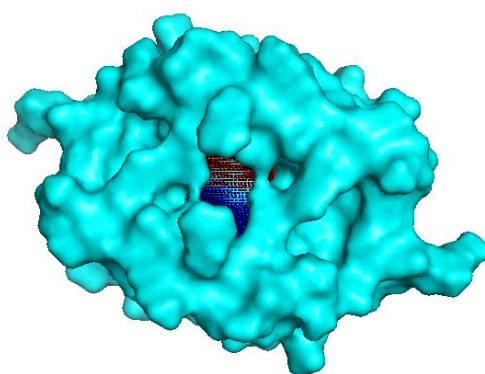


Figure 35. Representation of the RBD surface with the subpockets in the proximity of amino acids Trp16, Arg35, Leu36, Asp39, Gln40 and Leu43 represented in red and blue.

The results previously presented show the success of computational protocol designed, since the fragment extracted from the protein binding site with lowest Tanimoto

score doesn't interact with the binding site of RBD. The fragment extracted from the protein binding site with the highest Tanimoto score interacts with the amino acids from the binding site of RBD. These results are also in agreement with the druggability assessment of the RBD. Both fragments interact with *hot spots* in the protein surface with high druggability values, i.e. putative pockets on the protein surface capable of binding high-affinity drug-like molecules.

5. Conclusions

The aim of the present work was the identification of tool compounds with potential to be developed into inhibitors of influenza's NS1 protein. To achieve this, an innovative computational protocol was developed for fragment-based discovery of NS1 inhibitors, followed by experimental validation using NMR. Experimental validation requires the cloning, expression and purification of each one of the domains of NS1 (RBD and ED) and their structural characterization using several techniques (NMR, CD, SEC-MALS).

The ^1H - ^{15}N HSQC spectrum of RBD-NS1 shows large chemical shift dispersion and thus presents high potential to be used in compound screening based on chemical shift perturbation methods. The results of SEC-MALS demonstrate that, in solution, the RBD is in a dimeric state with a molecular weight of 22.2 kDa. By contrast, in solution, the ED is a monomer with a molecular weight of 16.7 kDa. The diffusion coefficient of RBD was calculated using the value of the molecular weight of RBD and the Einstein-Stokes Law. The diffusion coefficient calculated for the RBD dimer is $1.007 \times 10^{-6} \text{ cm}^2/\text{s}$ and the experimental result obtained by DOSY NMR experiments is exactly the same ($1.007 \times 10^{-6} \text{ cm}^2/\text{s}$), showing that the results obtained with SEC-MALS for RBD are in agreement with the diffusion coefficient determined by NMR.

The computational protocol consists of comparing the chemical features of the binding site of the RBD with proteins from protein-fragments complexes presented in a database known as PDBbind. The fragments extracted from the proteins with high and low binding site similarity, were "docked" against the RBD. Finally, the computational results were validated using Chemical Shift Perturbation methods. Fragment hit validation by NMR is very important at this stage of the fragment-based lead discovery approach, because NMR is a very sensitive technique to identify protein-ligand interactions and allow validation of the computational results obtained from the innovative computational protocol designed during this project.

Residues in NS1 that mediate the RBD-dsRNA interaction, either directly or via improving the complex stability, include Thr5, Pro31, Asp34, Arg35, Arg38, Lys41, Gly45, Arg46 and Thr49. So fragments with the possibility to interact with RBD should have complementary features, such as hydrophobicity and hydrophilicity and charged groups. The structure of the fragments obtained from the computational protocol are characterized for both apolar and polar regions and both positive and negative charges, which validates the computational protocol. The innovative computational protocol designed in this

project is a new approach that can be used to rank and determine fragments to be used in posterior phases of the fragment-based lead design.

The results obtained using chemical shift perturbation methods shows that the fragment extracted from the binding site with low similarity with the binding site of RBD interacts in a different region within the protein surface. The fragment extracted from the binding site with high similarity with the binding site of RBD interacts with the amino acids that composes the interface of RBD-dsRNA. These results show the success of the computational protocol designed.

6. Future perspectives

In the future, this computational protocol can be applied to the effector domain of NS1 or other proteins.

The experimental validation of the computational results for RBD should be extended to the remaining fragments, to assess the quality of the computational protocol designed during this project.

The results obtained for the RBD can be used for the next phases of fragment-based drug discovery by merging or linking the identified fragments with the help of structure-guided strategies. Merging starts with identification of two overlapping reference fragments, and taking the decoration of one molecule and substitutes it with the core of the other. The output molecule is, therefore, a combination of both fragments. Linking starts with two non-overlapping reference fragments, where the first molecule is grown, then chemically linked to replacements that match the second. The challenge is to build successful bridging chemistry between the two fragments.

The computational protocol can also be used to study potential inhibitors of protein-protein interactions, meaning the dimerization of RBD, which is also critical for the dsRNA-RBD interaction.

7. Bibliography

1. Khanna, M., Kumar, P., Choudhary, K., Kumar, B., and Vijayan, V. (2008) Emerging influenza virus: A global threat, *J Bioscience* 33, 475–482.
2. Hannoun, C. (2014) The evolving history of influenza viruses and influenza vaccines, *Expert Rev Vaccines*, taylorfrancis 12, 1085–1094.
3. <http://www.who.int/mediacentre/factsheets/fs211/en/> visited in 08/08/2017
4. https://talk.ictvonline.org/ictv-reports/ictv_9th_report/negative-sense-rna-viruses-2011/w/negrna_viruses/209/orthomyxoviridae visited in 08/08/2017
5. viralzone.expasy.org/all_by_protein/223.html visited in 21/11/2016
6. Gatherer, D. The 2009 H1N1 influenza outbreak in its historical context. *J. Clin. Virol.* 45, 174–178 (2009).
7. Fodor, E. (2013) The RNA polymerase of influenza A virus: mechanisms of viral transcription and replication., *Acta Virol.* 57, 113–22.
8. Nelson, M. I., and Holmes, E. C. (2007) The evolution of epidemic influenza, *Nature reviews genetics*, Nature Publishing Group 8, 196–205.
9. Kapoor S., Dhama K. (2014) Replication Cycle of Influenza Viruses. In: *Insight into Influenza Viruses of Animals and Humans*. Springer, Cham.
10. Neumann, G., Noda, T., and Kawaoka, Y. (2009) Emergence and pandemic potential of swine-origin H1N1 influenza virus, *Nature* 459, 931–939.
11. Marsh, M., and Helenius, A. (2006) Virus Entry: Open Sesame, *Cell* 124, 729–740.
12. Zheng, W., and Tao, Y. (2013) Structure and assembly of the influenza A virus ribonucleoprotein complex, *Febs Lett*, wiley 587, 1206–1214.
13. Eisfeld, A., Neumann, G., and Kawaoka, Y. (2014) At the centre: influenza A virus ribonucleoproteins, *Nature Reviews Microbiology* 13, 28–41.
14. Eyer, L., and Hruska, K. (2013) Antiviral agents targeting the influenza virus: a review and publication analysis, *Veterinari Medicina* 58, 113–185.
15. Engel, D. (2013) The influenza virus NS1 protein as a therapeutic target, *Antivir Res*, sciencedirect 99, 409–416.
16. Cheng, A., Wong, S. M., and Yuan, Y. A. (2009) Structural basis for dsRNA recognition by NS1 protein of influenza A virus., *Cell Res.* 19, 187–95.
17. Zhou, Y., and Huang, N. (2015) Binding site druggability assessment in fragment-based drug design., *Methods Mol. Biol.* 1289, 13–21.
18. Bornholdt, Z., and Prasad, V. (2006) X-ray structure of influenza virus NS1 effector domain, *Nat Struct Mol Biology*, nature 13, 559–560.
19. Marc, D., Barbachou, S., and Soubieux, D. (2013) The RNA-binding domain of influenzavirus non-structural protein-1 cooperatively binds to virus-specific RNA sequences in a structure-dependent manner, *Nucleic Acids Res*, oup 41, 434–449.
20. Hale, Randall, Ortin, and Jackson. (2008) The multifunctional NS1 protein of influenza A viruses, *J Gen Virol*, highwire 89, 2359–2376.
21. Ayllon, J., and García-Sastre, A. (2015) *Influenza Pathogenesis and Control - Volume II*
22. Carrillo, Choi, J.-M., Bornholdt, Sankaran, Rice, and Prasad. (2014) The Influenza A Virus Protein NS1 Displays Structural Polymorphism, *Journal of Virology* 88, 4113–4122.
23. Long, J.-X., Peng, D.-X., Liu, Y.-L., Wu, Y.-T., and Liu, X.-F. (2008) Virulence of H5N1 avian influenza virus enhanced by a 15-nucleotide deletion in the viral nonstructural gene, *Virus Genes* 36, 471–478.

24. García-Sastre, A., Egorov, A., Matasov, D., Brandt, S., Levy, D. E., Durbin, J. E., Palese, P., and Muster, T. (1998) Influenza A virus lacking the NS1 gene replicates in interferon-deficient systems., *Virology* 252, 324–30.
25. Donelan, N. R., Basler, C. F., and García-Sastre, A. (2003) A recombinant influenza A virus expressing an RNA-binding-defective NS1 protein induces high levels of beta interferon and is attenuated in mice., *J. Virol.* 77, 13257–66
26. Kochs, G., García-Sastre, A., and Martínez-Sobrido, L. (2007) Multiple anti-interferon actions of the influenza A virus NS1 protein., *J. Virol.* 81, 7011–21
27. Haye, Burmakina, Moran, García-Sastre, and Fernandez-Sesma. (2009) The NS1 Protein of a Human Influenza Virus Inhibits Type I Interferon Production and the Induction of Antiviral Responses in Primary Human Dendritic and Respiratory Epithelial Cells, *Journal of Virology* 83, 6849–6862.
28. Munir, Zohari, Metreveli, Baule, Belak, and Berg. (2011) Alleles A and B of non-structural protein 1 of avian influenza A viruses differentially inhibit beta interferon production in human and mink lung cells, *Journal of General Virology* 92, 2111–2121.
29. Munir, M., Zohari, S., Belák, S., and Berg, M. (2012) Double-Stranded RNA-Induced Activation of Activating Protein-1 Promoter Is Differentially Regulated by the Non-structural Protein 1 of Avian Influenza A Viruses, *Viral Immunology*.
30. Krug, R. (2015) Functions of the influenza A virus NS1 protein in antiviral defense, *Curr Opin Virology, sciencedirect* 12, 1–6.
31. Goodford. (1985) A computational procedure for determining energetically favorable binding sites on biologically important macromolecules, *Journal of Medicinal Chemistry* 28, 849–857.
32. Fauman, E., Rai, B., and Huang, E. (2011) Structure-based druggability assessment—identifying suitable targets for small molecule therapeutics, *Current Opinion in Chemical Biology* 15, 463–468.
33. Schmidtke, P., and Barril, X. (2010) Understanding and predicting druggability. A high-throughput method for detection of drug binding sites., *J. Med. Chem.* 53, 5858–67.
34. Huang, N., and Jacobson, M. (2010) Binding-Site Assessment by Virtual Fragment Screening, *PLoS ONE* 5, e10109.
35. Fuller, J., Burgoyne, N., and Jackson, R. (2009) Predicting druggable binding sites at the protein–protein interface, *Drug Discovery Today* 14, 155–161.
36. Sugaya, N., and Furuya, T. (2011) Dr. PIAS: an integrative system for assessing the druggability of protein-protein interactions., *BMC Bioinformatics* 12, 50.
37. Villoutreix, B., Bastard, K., Sperandio, O., Fahraeus, R., Poyet, J.-L., Calvo, F., Deprez, B., and Miteva, M. (2008) In Silico-In Vitro Screening of Protein-Protein Interactions: Towards the Next Generation of Therapeutics, *Curr Pharm Biotechno, bentham* 9, 103–122.
38. Panjkovich, A., and Daura, X. (2010) Assessing the structural conservation of protein pockets to study functional and allosteric sites: implications for drug discovery, *Bmc Struct Biol, springer* 10, 1–14.
39. Schmidtke,P. and Barril,X. (2010) Understanding and predicting druggability. a highthroughput method for detection of drug binding sites. *J. Med. Chem.*, 53, 5858–5867.
40. Volkamer, A., Kuhn, D., Rippmann, F., and Rarey, M. (2012) DoGSiteScorer: a web server for automatic binding site prediction, analysis and druggability assessment., *Bioinformatics* 28, 2074–5.
41. Johnson, C. M. (2013) Differential scanning calorimetry as a tool for protein folding and stability., *Arch. Biochem. Biophys.* 531, 100–9.
42. Kelly, SM, Jess, TJ, and Price, NC. (2005) How to study proteins by circular dichroism, *Biochimica et Biophysica Acta (BBA)-Proteins*.

43. Sahin E., Roberts C.J. (2012) Size-Exclusion Chromatography with Multi-angle Light Scattering for Elucidating Protein Aggregation Mechanisms. In: Voynov V., Caravella J. (eds) Therapeutic Proteins. Methods in Molecular Biology (Methods and Protocols), vol 899. Humana Press, Totowa, NJ.
44. Hore PJ. Nuclear magnetic resonance. 1st ed. New York: Oxford University Press Inc.; 1995.
45. Bodenhausen G, Ruben DJ. Natural abundance nitrogen-15 NMR by enhanced heteronuclear spectroscopy. *Chem Phys Lett.* 1980;69(1):185-189.
46. Weiss. (1999) Fluorescence Spectroscopy of Single Biomolecules, *Science* 283, 1676–1683.
47. <https://www.rcsb.org/pdb/home/home.do> visited in 08/08/2017
48. Pettersen, E., Goddard, T., Huang, C., Couch, G., Greenblatt, D., Meng, E., and Ferrin, T. (2004) UCSF Chimera?A visualization system for exploratory research and analysis, *Journal of Computational Chemistry* 25, 1605–1612.
49. Yuan, S., Chan, H. C., and Hu, Z. (2017) Using PyMOL as a platform for computational drug design, *Wiley Interdisciplinary Reviews: Computational Molecular Science* 7, e1298.
50. <http://proteinsplus.zbh.uni-hamburg.de/> visited in 02/03/2017
51. Das, K., Aramini, J., Ma, L.-C., Krug, R., and Arnold, E. (2010) Structures of influenza A proteins and insights into antiviral drug targets, *Nat Struct Mol Biology*, nature 17, 530–538.
52. Lees, J. G., Miles, A. J., Wien, F., and Wallace, B. A. (2006) A reference database for circular dichroism spectroscopy covering fold and secondary structure space., *Bioinformatics* 22, 1955–62.
53. Hong, P., Koza, S., and Bouvier, E. S. (2012) Size-Exclusion Chromatography for the Analysis of Protein Biotherapeutics and their Aggregates., *J. Liq. Chromatogr. Relat. Technol.* 35, 2923–2950.
54. Hubbard, R. E. (2016) The Role of Fragment-based Discovery in Lead Finding, in *Fragment-based Drug Discovery Lessons and Outlook* (eds D. A. Erlanson and W. Jahnke), Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany.
55. Erlanson, D. (2012) Topics in Current Chemistry, pp 1–32.
56. Congreve, M., Carr, R., Murray, C., and Jhoti, H. (2003) A “rule of three” for fragment-based lead discovery?, *Drug Discov. Today* 8, 876–7.
57. Abad-Zapatero, C. (2007) Ligand efficiency indices for effective drug discovery., *Expert Opin Drug Discov* 2, 469–88.
58. Tarcsay, A., Nyíri, K., and Keseru, G. M. M. (2012) Impact of lipophilic efficiency on compound quality., *J. Med. Chem.* 55, 1252–60.
59. Rognan D. (2011) Fragment-Based Approaches and Computer-Aided Drug Discovery. In: Davies T., Hyvönen M. (eds) *Fragment-Based Drug Discovery and X-Ray Crystallography*. Topics in Current Chemistry, vol 317. Springer, Berlin, Heidelberg
60. Hall, D., Kozakov, D., and Vajda, S. (2012) Analysis of Protein Binding Sites by Computational Solvent Mapping, pp 13–27, springer.
61. Ngan, C., Bohnuud, T., Mottarella, S., Beglov, D., Villar, E., Hall, D., Kozakov, D., and Vajda, S. (2012) FTMAP: extended protein mapping with user-selected probe molecules, *Nucleic Acids Res*, oup 40, W271–W275.
62. Chen, Y., and Shoichet, B. (2009) Molecular docking and ligand specificity in fragment-based inhibitor discovery, *Nat Chem Biol*, nature 5, 358–364.
63. Kitchen, D., Decornez, H., Furr, J., and Bajorath, J. (2004) Docking and scoring in virtual screening for drug discovery: methods and applications, *Nat Rev Drug Discov*, nature 3, 935–949.

64. P. J. Goodford (1985) A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *Journal of Medicinal Chemistry* 28 (7), 849-857
65. Artese, A., Cross, S., Costa, G., Distinto, S., Parrotta, L., Alcaro, S., Ortuso, F., and Cruciani, G. (2013) Molecular interaction fields in drug discovery: recent advances and future perspectives, *Wiley Interdiscip Rev Comput Mol Sci*, wiley 3, 594–613.
66. Shulman-Peleg, A., Nussinov, R., and Wolfson, H. (2004) Recognition of Functional Sites in Protein Structures, *J Mol Biol*, sciencedirect 339, 607–633.
67. Kinoshita, K., and Nakamura, H. (2003) Identification of protein biochemical functions by similarity search using the molecular surface database eF- site, *Protein Sci*, wiley 12, 1589–1595.
68. Bron, C., and Kerbosch, J. (1973) Algorithm 457: finding all cliques of an undirected graph, *CommunAcm*, acm 16, 575–577
69. Winter, A., Higuero, A., Marsh, M., Sigurdardottir, A., Pitt, W., and Blundell, T. (2012) Biophysical and computational fragment-based approaches to targeting protein–protein interactions: applications in structure-guided drug discovery, *Q Rev Biophys*, cambridge 45, 383–426.
70. Bodenhausen, G., and Ruben, D. (1980) Natural abundance nitrogen-15 NMR by enhanced heteronuclear spectroscopy, *Chemical Physics Letters* 69, 185–189.
71. Williamson, M. (2013) Using chemical shift perturbation to characterise ligand binding, *Progress in Nuclear Magnetic Resonance Spectroscopy* 73, 1–1.
72. Shuker, S. B., Hajduk, P. J., Meadows, R. P., and Fesik, S. W. (1996) Discovering high-affinity ligands for proteins: SAR by NMR., *Science* 274, 1531–4.
73. Hajduk, P., Meadows, R., & Fesik, S. (1999). NMR-based screening in drug discovery. *Quarterly Reviews of Biophysics*, 32(3), 211-240.
74. Hajduk, P., Bures, M., Praestgaard, J., and Fesik, S. (2000) Privileged Molecules for Protein Binding Identified from NMR-Based Screening, *Journal of Medicinal Chemistry* 43, 3443–3447.
75. Dias, D., and Ciulli, A. (2014) NMR approaches in structure-based lead discovery: Recent developments and new frontiers for targeting multi-protein complexes, *Progress in Biophysics and Molecular Biology* 116, 101–112.
76. <http://www.pdbbind.org.cn/>
77. <http://silicos-it.be.s3-website-eu-west-1.amazonaws.com/software/filter-it/1.0.2/filter-it.html>
78. <http://biophys.umontreal.ca/nrg/NRG/IsoMIF.html>
79. <http://kinemage.biochem.duke.edu/software/reduce.php>
80. Majeux, N., Scarsi, M., Apostolakis, J., Ehrhardt, C., and Caflisch, A. (1999) Exhaustive docking of molecular fragments with electrostatic solvation., *Proteins* 37, 88–105.
81. Hopkins, A., Keserü, G., Leeson, P., Rees, D., and Reynolds, C. (2014) The role of ligand efficiency metrics in drug discovery, *Nat Rev Drug Discov*, nature 13, 105–121
82. Ferenczy, G., and Keserü, G. (2016) *Fragment-based Drug Discovery Lessons and Outlook*, pp 75–98, wiley.
83. Mortenson, P. N., and Murray, C. W. (2011) Assessing the lipophilicity of fragments and early hits, *Journal of computer-aided molecular design*, Springer 25, 663–667.

8. Annexes

Annex A

Table A. Assignments of the ^1H (NH) and ^{15}N (NH) chemical shifts for each amino acid of the RNA-binding domain (RBD) of NS1.

Sequence ID	Residue	^1H Chemical Shift (ppm)	^{15}N Chemical Shift (ppm)
4	Asn	8.92	122.21
5	Thr	8.25	122.47
6	Val	7.52	123.30
7	Ser	8.67	116.96
8	Ser	8.19	116.96
9	Phe		
10	Gln	8.281	119.222
11	Val	8.070	117.977
12	Asp	8.101	121.256
13	Cys	8.503	119.712
14	Phe		
15	Leu	9.147	117.865
16	Trp	8.159	120.758
17	His	7.739	118.010
18	Val	7.667	119.234
19	Arg		
20	Lys	8.252	122.505
21	Arg		
22	Val		
23	Ala	8.748	123.507
24	Asp		
25	Gln	7.115	118.766
26	Glu	7.930	112.746
27	Leu	6.678	114.637
28	Gly	8.023	104.880
29	Asp	7.489	121.938
30	Ala	8.435	119.479
31	Pro		
32	Phe	7.968	120.614
33	Leu	8.328	123.740
34	Asp	8.226	123.128
35	Arg	8.104	121.281
36	Leu	7.773	122.376
37	Arg	7.923	121.604
38	Arg	8.307	121.082
39	Asp	8.733	119.984
40	Gln	7.888	120.566
41	Lys	6.954	117.070
42	Ser	7.859	115.829
43	Leu	8.733	125.545
44	Arg	7.614	118.772
45	Gly	7.713	107.383
46	Arg	8.350	124.925
47	Gly	8.759	107.249
48	Ser	7.722	116.109
49	Thr	7.652	118.199
50	Leu		
51	Gly	7.678	108.634
52	Leu	7.503	118.527

53	Asp	8.105	122.704
54	Ile	8.733	125.588
55	Glu	8.932	125.647
56	Thr		
57	Ala	8.120	125.165
58	Thr	8.218	114.927
59	Arg	7.174	120.323
60	Ala	7.566	120.527
61	Gly	8.726	108.068
62	Lys	7.717	123.318
63	Gln	6.677	114.641
64	Ile		
65	Val	8.069	118.054
66	Glu		
67	Arg	7.275	117.576
68	Ile	7.425	119.746
69	Leu	8.385	119.045
70	Lys		
71	Glu		
72	Glu		
73	Ser		

Annex B

Output DoGSiteScorer

name	lig_cov	poc_cov	lig_name		volume	enclosure	surface	depth	surf/vol		
	lid/hull	ellVol	ell c/a	ell b/a	siteAtms	accept	donor				
	hydrophobic_interactions		hydrophobicity		metal	Cs	Ns	Os	Ss		
	Xs	negAA	posAA	polarAA	apolarAA	ALA	ARG	ASN	ASP		
	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO
	SER	THR	TRP	TYR	VAL	simpleScore	drugScore				
P_0	0.00	0.00	""	1022.34	0.04	1507.46	22.81	1.474519	2401745015	-	
	-	0.42	0.50	289	76	32	29	0.21	0	204	38
46	1	0	0.12	0.18	0.16	0.53	0	7	0	4	1
3	2	0	1	5	8	1	0	5	0	3	1
2	0	6	0.57	0.816223							
P_0_0	0.00	0.00	""	304.83	0.04	542.76	10.13	1.780533	4120657417	-	
	-	0.41	0.49	82	24	17	7	0.15	0	53	15
14	0	0	0.20	0.27	0.07	0.47	0	4	0	3	0
1	0	0	0	0	4	0	0	1	0	0	0
1	0	1	0.29	0.271738							
P_0_1	0.00	0.00	""	293.63	0.06	517.15	10.15	1.761230	1195381943	-	
	-	0.38	0.46	77	24	17	8	0.16	0	50	14
13	0	0	0.21	0.29	0.07	0.43	0	4	0	3	0
1	0	0	0	0	4	0	0	1	0	0	0
1	0	0	0.27	0.224325							
P_0_2	0.00	0.00	""	102.98	0.10	332.92	7.64	3.232860	7496601283	-	
	-	0.43	0.87	34	8	7	7	0.32	0	24	5
5	0	0	0.12	0.38	0.25	0.25	0	1	0	0	1
0	1	0	1	1	0	1	0	0	0	0	1
1	0	0	0.01	0.203629							
P_0_3	0.00	0.00	""	78.21	0.00	163.64	0.00	2.092315	560669991	-	
	-	0.46	0.90	52	12	4	3	0.16	0	38	6
8	0	0	0.10	0.10	0.20	0.60	0	1	0	1	0
0	0	0	0	0	1	0	0	3	0	2	0
0	0	2	0.00	0.370244							
P_0_4	0.00	0.00	""	76.35	0.00	171.02	0.69	2.239947	6096922073	-	
	-	0.29	0.66	47	12	4	4	0.20	0	36	4
7	0	0	0.12	0.12	0.12	0.62	0	1	0	1	0
0	0	0	0	0	1	0	0	2	0	1	0
0	0	2	0.00	0.301424							
P_0_5	0.00	0.00	""	60.35	0.00	110.93	3.22	1.838111	10190555097	-	
	-	0.35	0.40	35	6	1	0	0.00	0	28	2
5	0	0	0.09	0.09	0.09	0.73	0	1	0	0	0
1	1	0	0	2	2	0	0	1	0	0	0
0	0	3	0.00	0.5							
P_0_6	0.00	0.00	""	55.30	0.00	89.78	1.44	1.623508	1374321882	-	
	-	0.36	0.44	34	6	0	0	0.00	0	28	2
4	0	0	0.00	0.11	0.00	0.89	0	1	0	0	0

0	0	0	0	2	2	0	0	1	0	0	0
0	0	3	0.00	0.554566							
P_0_7	0.00	0.00	""	50.69	0.00	243.57	2.43	4.805089761294141		-	
	-	0.20	0.55	42	8	5	5	0.28	0	31	6
4	1	0	0.25	0.12	0.25	0.38	0	1	0	2	1
1	0	0	0	1	1	0	0	0	0	0	0
1	0	0	0.00	0.277742							
P_1	0.00	0.00	""	667.07	0.14	1142.65	17.10	1.7129386721033775		-	
	-	0.10	0.17	135	20	16	48	0.57	0	105	19
11	0	0	0.08	0.42	0.08	0.42	0	4	0	0	0
0	2	0	2	0	4	4	0	2	0	0	2
0	0	4	0.51	0.792978							
P_1_0	0.00	0.00	""	405.57	0.19	612.32	13.11	1.509776364129497		-	
	-	0.22	0.78	68	8	3	25	0.69	0	58	4
6	0	0	0.12	0.25	0.00	0.62	0	0	0	0	0
0	2	0	0	0	4	4	0	2	0	0	0
0	0	4	0.44	0.75822							
P_1_1	0.00	0.00	""	139.20	0.08	448.14	8.86	3.219396551724138		-	
	-	0.23	0.31	44	8	9	16	0.48	0	32	9
3	0	0	0.00	0.50	0.12	0.38	0	2	0	0	0
0	0	0	1	0	1	1	0	1	0	0	1
0	0	1	0.10	0.254559							
P_1_2	0.00	0.00	""	122.30	0.08	385.43	7.98	3.1515126737530665		-	
	-	0.24	0.33	45	6	7	14	0.52	0	35	7
3	0	0	0.00	0.50	0.12	0.38	0	2	0	0	0
0	0	0	1	0	1	1	0	1	0	0	1
0	0	1	0.08	0.354956							
P_2	0.00	0.00	""	432.64	0.10	697.38	20.07	1.6119175295857988		-	
	-	0.11	0.19	106	38	11	23	0.32	0	69	16
21	0	0	0.14	0.14	0.19	0.52	2	3	1	2	0
0	1	1	0	0	5	0	0	3	1	1	1
0	0	0	0.23	0.80613							
P_2_0	0.00	0.00	""	326.02	0.09	476.74	12.14	1.4623029262008467		-	
	-	0.31	0.36	69	24	10	18	0.35	0	46	10
13	0	0	0.15	0.15	0.31	0.38	0	2	1	2	0
0	0	1	0	0	2	0	0	2	1	1	1
0	0	0	0.33	0.267797							
P_2_1	0.00	0.00	""	106.62	0.10	290.21	10.16	2.7219095854436315		-	
	-	0.19	0.23	52	18	3	9	0.30	0	35	7
10	0	0	0.17	0.08	0.17	0.58	2	1	0	1	0
0	1	1	0	0	3	0	0	2	0	1	0
0	0	0	0.01	0.230182							
P_3	0.00	0.00	""	315.39	0.08	536.71	11.95	1.7017343606328674		-	
	-	0.30	0.39	66	24	11	25	0.42	0	45	10
11	0	0	0.14	0.14	0.29	0.43	1	2	1	2	0
0	0	1	0	0	2	0	0	2	1	1	1
0	0	0	0.17	0.538228							

P_4	0.00	0.00	""	120.32	0.07	382.36	9.54	3.177859042553192	-		
	-	0.29	0.30	49	16	3	12	0.39	0	32	8
9	0	0	0.18	0.09	0.09	0.64	2	1	0	1	0
0	1	1	0	0	3	0	0	2	0	0	0
0	0	0	0.00	0.304673							

Annex C

Scripts for Multiple Structural Alignment

```
#!/bin/bash

###Definition of Environment variables

isomifdir=/applic/IsoMif/IsoMif150311

workdir=/home/acunha/RBD-NS1/ISOMIF

scriptdir=/home/acunha/RBD-NS1/ISOMIF

outputcleft=/home/acunha/RBD-NS1/ISOMIF/GetCleft

outputmif=/home/acunha/RBD-NS1/ISOMIF/Mif

outputisomif=/home/acunha/RBD-NS1/ISOMIF/IsoMif

cd $workdir

for pdbfile in $(ls $workdir/complex_PL/*.pdb) ; do

    pdbcode=`grep DBREF $pdbfile | awk '{print $2}' | sort | uniq`

    ligandcode=`grep $pdbcode $workdir/ligands | awk '{print $2}'`

    ligandpose=`grep $pdbcode $workdir/ligands | awk '{print $3}'`

    ligandchain=`grep $pdbcode $workdir/ligands | awk '{print $4}'`

    $isomifdir/getcleft_linux_x86_64 -p $pdbfile -o $outputcleft/$pdbcode -s -a
    ${ligandcode}${ligandpose}${ligandchain}-

done

cd $workdir

for pdbhfile in $(ls $workdir/complex_PL_with_h/*.pdb) ; do

    pdbbid=`grep DBREF $pdbhfile | awk '{print $2}' | sort | uniq`

    sph=`ls $workdir/getcleft/$pdbbid*sph*`

    $isomifdir/mif_linux_x86_64 -p $pdbhfile -g $sph -o $outputmif/ -t $pdbbid

done
```

```
cd $workdir

for mif in $(ls $workdir/mif/*.mif) ; do

    $isomifdir/isomif_linux_x86_64 -p1 $workdir/2n74h.mif -p2 $mif -o $outputisomif/ -c 1
-w -wc -a 1000

done

cd $workdir
```

Annex D

SEED input file

```
#Parameter filename
seed.par
#Dielectric constant of the solute (receptor and fragment)
2.0
#Ratio of kept vectors for docking: polar / apolar
1 1
#Number of cluster members saved in output files
10
#The docked fragments are saved in the dir ./outputs
#Filename for output log file
./seed.out
#write (w) or read (r) Coulombic grid / grid filename
r./scratch/coulombic_20residues.grid
#write (w) or read (r) van der Waals grid / grid filename
r./scratch/vdwaals_20residues.grid
#write (w) or read (r) receptor desolvation grid / grid filename
r./scratch/receptor_desolv_20residues.grid
#Receptor coordinates (in mol2 format) filename
./2n74.mol2
#Binding site residue list
#First line: number of residues
11
12
15
16
19
32
35
36
38
39
40
41
43
#Modification of February 2002:
#List of points (e.g. ligand heavy atoms of a known ligand-
receptor complex structure) in the binding site used to select
polar and apolar rec. vectors which satisfy the angle criterion
(see parameters file)
#First line: number of points (0: no removal of vectors using
the angle criterion)
#Following lines: coordinates of the points
3
-4.553   -2.308   -22.326
 0.427   -1.012   -25.607
-1.388   -4.127   -28.99
#           Metals in the binding site
#           Make sure that the residue number of the metal is in
the
#           binding site residue list.
#           First line: total number of coordination points
```

```

#Following lines:  atom number of metal / x y z of coordination
point
0
#Spherical cutoff for docking (y,n / sphere center / sphere
radius)
y   -2.133 -1.359 -25.539   10.0
#Fragment library specifications
#First line: Number of fragments / dock+energy (n), only energy
(y)
#Following lines: Fragment filename /
#apolar docking, polar docking, or both (a,p,b) /
#energy cutoff in kcal/mol / 2nd clustering cutoff in kcal/mol
6 n
./ligands_docking/2fx6.mol2          b      0.0   0.0
./ligands_docking/1utm.mol2         b      0.0   0.0
./ligands_docking/3ad7.mol2         b      0.0   0.0
./ligands_docking/4x8u.mol2         b      0.0   0.0
./ligands_docking/4x8t.mol2         b      0.0   0.0
./ligands_docking/4x8s.mol2         b      0.0   0.0
end

```

Annex E

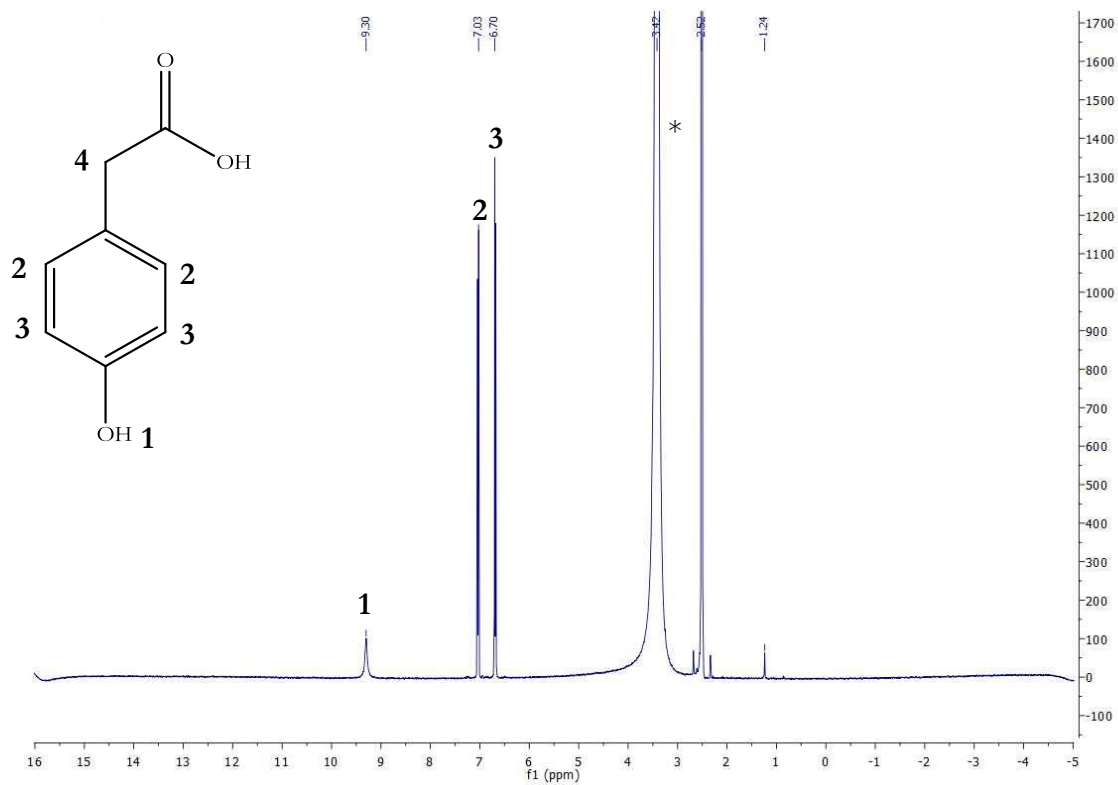


Figure A. ^1H NMR spectrum of 4-hydroxyphenylacetate (10 mM) in DMSO. The asterisk corresponds to DMSO- d_6 .

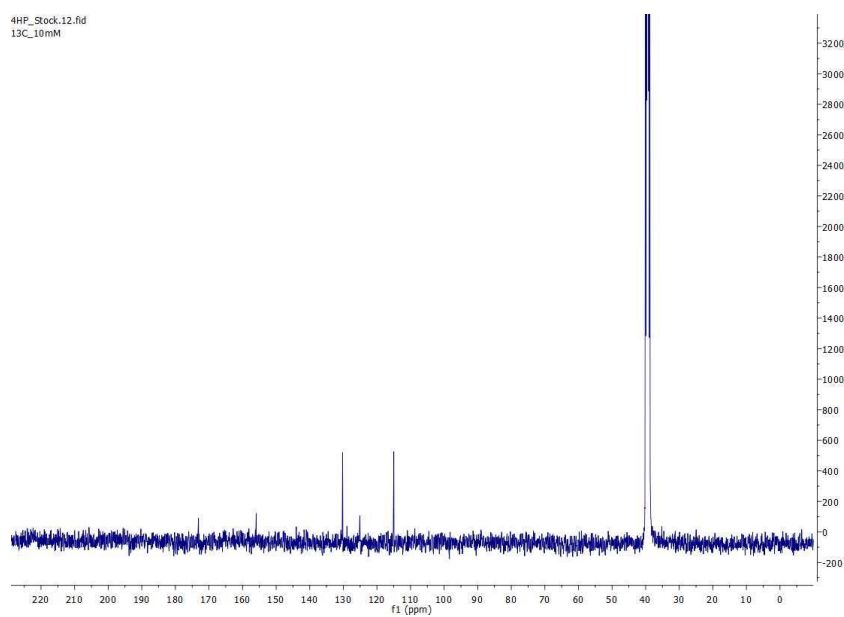


Figure B. ^{13}C NMR spectrum of 4-hydroxyphenylacetate (10 mM) in DMSO.

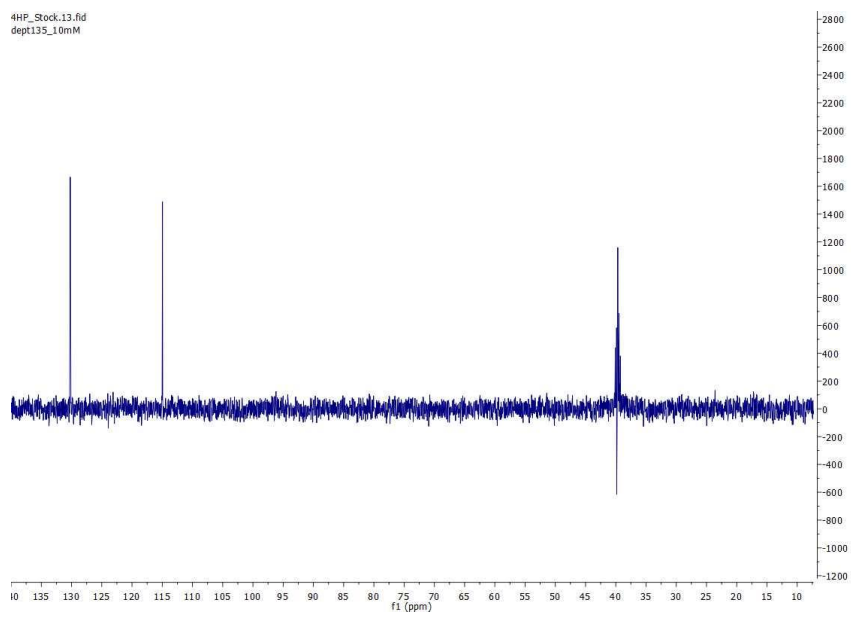


Figure C. DEPT ^{13}C -NMR spectrum of 4-hydroxyphenylacetate (10 mM) in DMSO.

Annex F

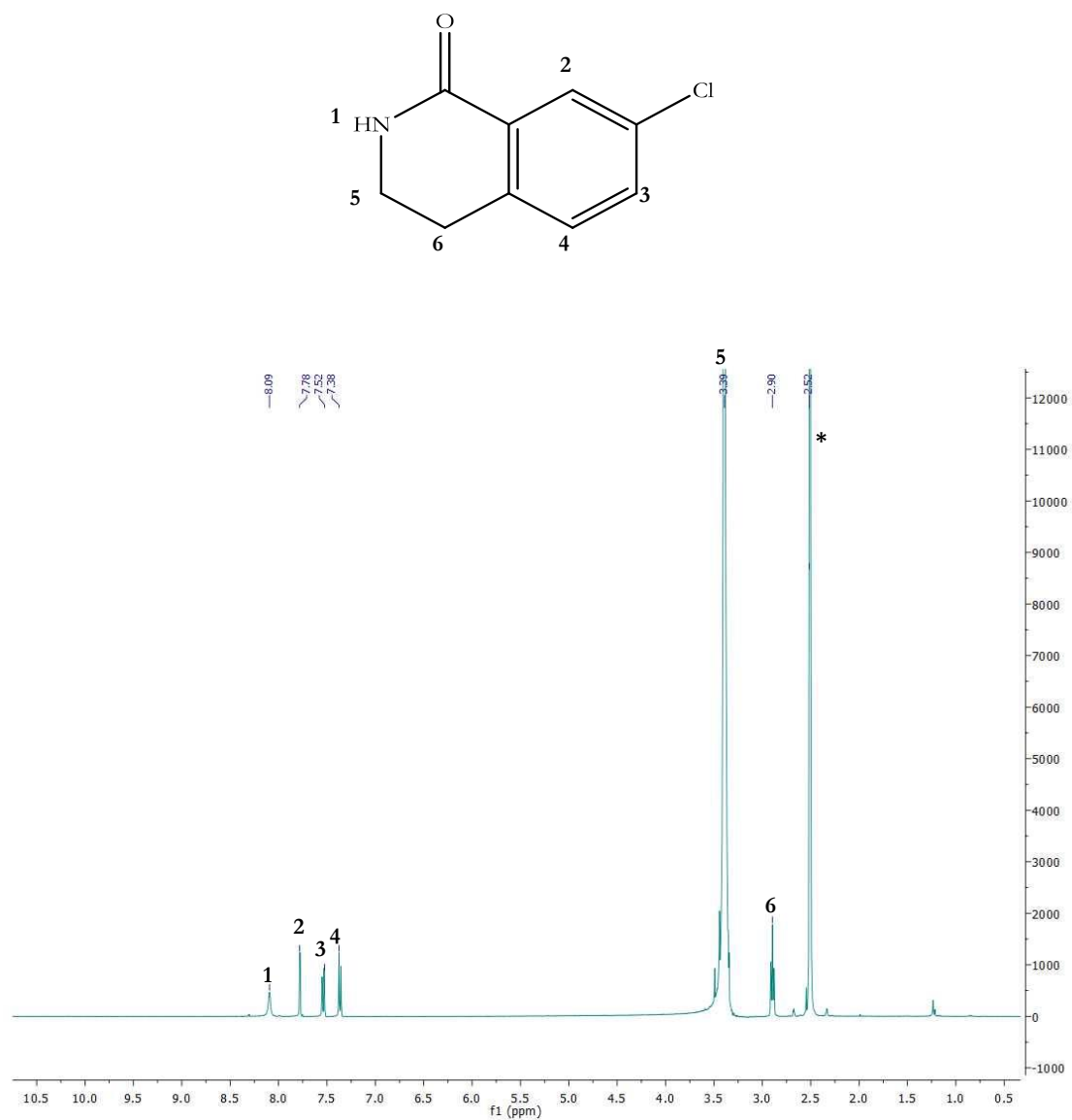


Figure A. ¹H NMR spectrum of 7-chloro-3,4-dihydroisoquinolin-1(2H)-one (10 mM) in DMSO. The asterisk corresponds to DMSO-d₆.

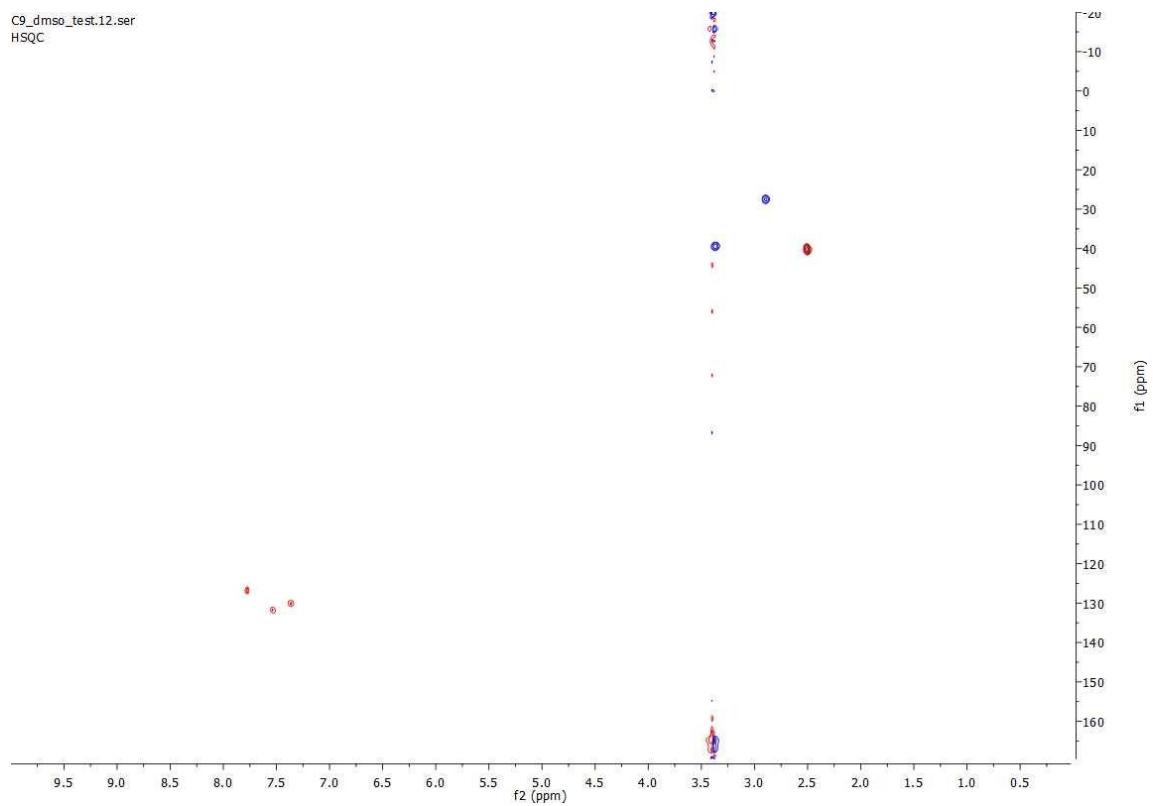


Figure B. ^1H - ^{13}C HSQC of 7-chloro-3,4-dihydroisoquinolin-1(2H)-one (10 mM) in DMSO- d_6 .