Exploration structurale des protéines pour leur validation en tant que cibles thérapeutiques

Sarah Naceri

Université Paris Cité

École doctorale Pierre Louis de Santé Publique :
Épidémiologie et Sciences de l'Information Biomédicale (ED393)

Unité de Biologie Fonctionnelle et Adaptative (BFA) (UMR 8251)
Équipe "Modélisation Computationnelle des Interactions Protéine-Ligand"

Exploration structurale des protéines pour leur validation en tant que cibles thérapeutiques

Par Sarah NACERI

Thèse de doctorat en Informatique médicale

Dirigée par Anne-Claude CAMPROUX
et co-encadrée par Delphine FLATTERS

Présentée et soutenue publiquement le 02 Juin 2023
Devant un jury composé de:

Alexandre VARNEK	PU	Université de Strasbourg	Rapporteur
Isabelle CALLEBAUT	DR	Sorbonne Université	Rapportrice
Arnaud BLONDEL	CR-HDR	Institut Pasteur	Examinateur
Isabelle THERET	CR	Institut de recherche SERVIER	Examinatrice
Annick MÉJEAN	PU	Université Paris Cité	Examinatrice
Anne-Claude CAMPROUX	PU	Université Paris Cité	Directrice de thèse
Delphine FLATTERS	MCF	Université Paris Cité	Membre invité
Daniel MARC	CR-HDR	Université de Tours	Membre invité

RÉSUMÉ

Exploration structurale des protéines pour leur validation en tant que cibles thérapeutiques

L'identification et la validation d'une cible thérapeutique à l'aide d'approches *in silico* basées sur la structure, reposent notamment sur des techniques de modélisation moléculaire. Parmi ces techniques, on peut citer celles qui permettent la détection de leurs sites de liaison par l'estimation des cavités à la surface de la structure de la cible et la caractérisation de leurs propriétés de druggabilité, c'est-à-dire de leur capacité à accepter un ligand de type médicament.

Malgré l'augmentation du nombre de structures tridimensionnelles des protéines disponibles, les cibles thérapeutiques restent difficiles à identifier en raison de la flexibilité des protéines dans l'organisme. Les changements conformationnels des protéines peuvent induire des modifications locales au niveau des résidus qui composent les sites de liaison et par conséquent modifier les propriétés de druggabilité des protéines. Différents outils permettent l'identification des poches, mais seulement une partie d'entre eux proposent une prédiction de leur potentiel de druggabilité, et la quantification de leur déformabilité n'est pas proposée.

Mon projet de thèse consiste à développer un protocole d'exploration structurale des protéines pour leur validation en tant que cibles thérapeutiques en intégrant des approches de détection des poches, de prédiction de leur druggabilité tout en prenant en compte la flexibilité des protéines. Cette stratégie permet d'étudier le polymorphisme structural des protéines; de caractériser les sites de liaison et leur évolution en termes de résidus ainsi que leur druggabilité.

Ce protocole repose sur deux étapes principales. Une étape de recherche et de prédiction des poches druggables à l'aide d'un outil développé au sein de l'équipe appelé PockDrug. Cet outil utilise des descripteurs physico-chimiques et géométriques ainsi que des techniques statistiques supervisées pour estimer et caractériser les cavités protéiques. Une étape d'étude de la dynamique des protéines afin d'évaluer leur polymorphisme structural et leur flexibilité. Finalement, les données issues de ces étapes sont intégrées et analysées à l'aide de techniques d'analyse statistique multivariée non supervisée telle que la classification pour suivre l'évolution et la flexibilité des poches en étudiant les résidus qui les composent au cours des dynamiques. Ce protocole permet d'évaluer la déformabilité des poches en termes de résidus, leur évolution au cours du temps, leur fréquence d'apparition ainsi que leur potentiel druggable. De plus, cette méthode permet de tester l'impact des mutations ou des partenaires sur la structure des protéines et leur site de liaison et leur flexibilité, ainsi que sur leur druggabilité.

Le travail effectué a abouti à la publication de deux articles portant sur la cible thérapeutique Non Structurale 1 du virus Influenza A. L'article [1] concerne l'analyse du polymorphisme structural de différents sous-types de la protéine Non Structurale 1 (NS1) à l'aide de simulations de dynamique moléculaire afin de comprendre la flexibilité de la protéine NS1. Cette analyse a permis de démontrer l'indépendance du polymorphisme de cette protéine vis-à-vis du sous-type. En outre, ces simulations ont permis de mettre en évidence la stabilité du domaine de liaison à l'ARN, qui joue un rôle essentiel dans la réplication du virus. La deuxième publication [2] a été menée pour rechercher au niveau de cette région stable du domaine de liaison à l'ARN une poche druggable fréquente et commune aux différents sous-type étudiés de NS1 ainsi qu'une analyse de la déformabilité de cette poche pour identifier les résidus clés à cibler par une molécule thérapeutique .

Un troisième article [3], porte sur l'analyse du domaine de liaison au récepteur de la protéine Spike du virus SARS-CoV-2. Pour cette étude, en tenant compte de la flexibilité des protéines, le protocole a été étendu à la recherche des poches sur l'ensemble de la surface de la protéine Spike et a permis d'identifier trois sites de liaison d'intérêt, pouvant potentiellement entraîner l'inactivation de la protéine Spike.

Mots-clés: sites de liaison, dynamique moléculaire, cible thérpeutique, Influenza A, SARS-CoV-2, apprentissage automatique

Articles:

[1] **Naceri, S.**, Marc, D., Camproux, A. C.*, & Flatters, D.* (2022). Influenza A Virus NS1 Protein Structural Flexibility Analysis According to Its Structural Polymorphism Using Computational Approaches. *International Journal of Molecular Sciences, 23*, 1805. https://doi.org/10.3390/ijms23031805

[2] **Naceri, S.**, Marc, D., Blot, R., Flatters, D.*, & Camproux, A. C.* (2022). Druggable Pockets at the RNA Interface Region of Influenza A Virus NS1 Protein Are Conserved across Sequence Variants from Distinct Subtypes. *Biomolecules, 13*, 64. https://doi.org/10.3390/biom13010064

[3] Ghoula, M., **Naceri, S.**, Sitruk, S., Flatters, D., Moroy, G., & Camproux, A. (2023). Identifying promising druggable binding sites and their flexibility to target the receptor-binding domain of SARS-CoV-2 spike protein. *Computational and structural biotechnology journal*. https://doi.org/10.1016/j.csbj.2023.03.029

ABSTRACT

Structural exploration of proteins for their validation as therapeutic targets

The identification and validation of therapeutic targets using structure-bases *in silico* approaches relies, in particular, on molecular modeling techniques. Among these techniques, the detection of binding sites involves estimating cavities on the target's surface and characterizing their druggability properties, which refers to their ability to accept a drug-like ligand.

Despite the increasing number of available protein structures, identifying therapeutic targets remains challenging due to the flexibility of proteins in the body. Conformational changes in proteins can induce local modifications at the level of residues that make up the binding sites, altering the druggability properties of the proteins. Different tools allow for pocket identification, but only a few predict their druggability potential, and none provide a quantification of their deformability.

This protocol consists of two main steps: a search of pockets and prediction of their druggability using a team-developed tool called PockDrug. This tool utilizes physicochemical and geometric descriptors, as well as supervised statistical techniques, to estimate and characterize protein cavities. The second step involves studying protein dynamics to evaluate their structural polymorphism and flexibility. Finally, data from these steps are integrated and analyzed using unsupervised multivariate statistical analysis techniques, such as classification, to track the evolution and flexibility of pockets by studying the residues that compose them during dynamics. This protocol allows for evaluating the deformability of pockets in terms of residues, their temporal evolution, frequency of occurrence, as well as their druggable potential. Additionally, this method enables testing the impact of mutations or partners on protein structure, binding site flexibility, and druggability.

The work conducted resulted in the publication of two articles regarding the Non Structural 1 therapeutic target of the Influenza A virus. The first article [1] concerns the analysis of the structural polymorphism of different NS1 protein subtypes using molecular dynamics simulations to understand the flexibility of the NS1 protein. This analysis demonstrated the independence of the polymorphism of this protein regarding the subtype. Additionally, these simulations allowed highlighting the stability of the RNA binding domain, which plays an essential role in virus replication. The second publication [2] aimed to search for a frequent and common druggable pocket within the stable region of the RNA binding domain of NS1 studied in different subtypes. It also involved analyzing the deformability of this pocket to identify the key residues to target with a therapeutic molecule.

A third article [3] focuses on the analysis of the receptor binding domain of the Spike protein of the SARS-CoV-2 virus. For this study, the protocol was extended to the search for pockets on the entire surface of the Spike protein, taking into account protein flexibility. This approach allowed identifying three interesting binding sites, which could potentially lead to the inactivation of the Spike protein.

Key words: binding sites, molecular dynamic, therapeutic target, Influenza A, SARS-CoV-2, machine learning

Articles:

[1] **Naceri, S.**, Marc, D., Camproux, A. C., & Flatters, D. (2022). Influenza A Virus NS1 Protein Structural Flexibility Analysis According to Its Structural Polymorphism Using Computational Approaches. *International Journal of Molecular Sciences*, *23*, 1805. https://doi.org/10.3390/ijms23031805

[2] **Naceri, S.**, Marc, D., Blot, R., Flatters, D., & Camproux, A. C. (2022). Druggable Pockets at the RNA Interface Region of Influenza A Virus NS1 Protein Are Conserved across Sequence Variants from Distinct Subtypes. *Biomolecules*, *13*, 64. https://doi.org/10.3390/biom13010064

[3] Ghoula, M., **Naceri, S.**, Sitruk, S., Flatters, D., Moroy, G., & Camproux, A. (2023). Identifying promising druggable binding sites and their flexibility to target the receptor-binding domain of SARS-CoV-2 spike protein. *Computational and structural biotechnology journal*. https://doi.org/10.1016/j.csbj.2023.03.029

REMERCIEMENTS

Je souhaiterais exprimer ma gratitude sincère envers tous ceux qui m'ont apporté leur soutien durant ces trois années de thèse et qui, par leur bienveillance, m'ont permis de mener à bien mes travaux de recherche.

Je tiens tout particulièrement à remercier le Pr. Anne Claude Camproux, ma directrice de thèse, qui m'a offert l'opportunité de réaliser cette thèse dans un domaine correspondant à mes compétences antérieures. Je suis extrêmement reconnaissante de votre sagesse et de votre patience, ainsi que pour votre accompagnement et vos encouragements tout au long de cette aventure.

Je remercie chaleureusement mon encadrante, Dr Delphine Flatters, pour son soutien constant, ses précieux conseils, sa bienveillance et son accompagnement tout au long de cette période. Même lorsque je perdais totalement espoir, vous avez toujours su m'aider à trouver des solutions à mes problèmes. Vous êtes une source inépuisable de connaissance et j'ai eu la chance de travailler avec vous et d'apprendre de vous.

J'adresse également mes remerciements à Pierre Tuffery, Gautier Moroy, Samuel Murail et à tous les membres de l'équipe CMPLI pour leur sympathie et leur accueil au cours de ces trois années. Je souhaite exprimer ma reconnaissance envers le Pr Jean-Marie Dupret, directeur du laboratoire BFA; l'université Paris Cité et l'École Doctorale 393 Pierre Louis de Santé Publique pour le financement de ma thèse.

Je voudrais également remercier mes collègues, Natacha Cerisier, Pierre Laville, Bryan Dafniet, Mariem Ghoula, Vanille Lejal, Guillaume Ollitrault, Rachel Blot et Ines Rahali, pour leur amitié et leur soutien sans faille. Les petits gestes tels que les chocolats et les viennoiseries déposés sur mon bureau ont rendu les moments de fatigue plus supportables et m'ont aidé à rester motivée. Un grand merci Pr Olivier Taboureau, considéré comme le Père Noël des doctorants en raison de sa grande générosité en matière de partage de friandises.

Je souhaite également exprimer toute ma gratitude envers mes sœurs de cœur Celia et Nouara ainsi qu'à ma famille qui, malgré la distance, ont toujours été à mes côtés, m'ont soutenu, et ont cru en mes capacités. Une pensée particulière à mon oncle et ma tante, Alexios et Anna Naceri, ainsi qu'à mes cousins et petits cousins qui m'ont toujours encouragé à poursuivre mes études supérieures. Leur présence a été précieuse dans les moments les plus importants de ma vie et je leur en suis reconnaissante.

Je tiens à rendre hommage en particulier à mes parents, qui m'ont apporté un soutien indéfectible et m'ont permis d'atteindre le grade de docteur aujourd'hui, c'est à vous que je dois cette réussite. Je remercie mes deux grands frères qui ont été pour moi des exemples tout au long de mes études. Leur assiduité et leur persévérance ont été une source d'inspiration pour moi, et j'ai été chanceuse de pouvoir compter sur leur appui et leurs conseils. Je tiens également à remercier mon époux, qui a été un soutien inestimable tout au long de ma thèse. Sa patience, son encouragement et son amour ont été un pilier essentiel dans les moments difficiles et m'ont apporté un réconfort précieux.

Je souhaite exprimer ma profonde reconnaissance envers toutes les personnes qui ont contribué, de près ou de loin, à la réalisation de mon projet. Je tiens à vous remercier sincèrement pour votre soutien et votre précieuse aide, car sans vous, je n'aurais pas pu y arriver.

ABRÉVIATIONS

ACE2 : *Angiotensin-Converting Enzyme 2*

CASTp : *Computed Atlas of Surface Topography of Proteins*

DM : Dynamique Moléculaire

ED : *Effector Domain*

MM-PBSA : *Molecular Mechanics Poisson Boltzmann Surface Area*

NIH : *National Institutes of Health*

NS1 : *Non Structural 1*

PDB : *Protein Data Bank*

RBD : *Receptor Binding Domain*

RBM: *Receptor Binding Motif*

RMN : Résonance Magnétique Nucléaire

RMSD : *Root Mean Square Deviation*

RMSF : *Root Mean Square Fluctuation*

RNA-BD : *RNA Binding Domain*

3D : tri-dimensions

TABLE DES MATIÈRES

LISTE DES FIGURES ET TABLES

traite le suivi des poches de surface du RBD et la prédiction de la druggabilité par échantillonnage de 1 000 conformations du RBD extraites des simulations de dynamique moléculaire, l'estimation des poches et la prédiction de la druggabilité à l'aide de PockDrug. L'étape 3 concerne le regroupement hiérarchique de toutes les poches RBD regroupées en terme de similarité de résidus. Cette classification permet l'identification de clusters de poches avec des résidus communs, correspondant à des sites de liaison fréquemment observés le long des simulations de dynamique moléculaire. La variabilité des résidus au sein d'un cluster de poches illustre la flexibilité des résidus du site de liaison correspondant. Enfin, ce protocole a identifié avec succès trois sites de liaison et leurs résidus clés susceptibles d'être ciblés par des inhibiteurs pour prévenir l'interaction avec ACE2.

Figure 7: Illustration de la structure du complexe RBD-ACE, constitué de la protéine Spike trimérique (représentée en rouge, bleu et vert) et de la protéine ACE2 (en violet). Dans cette structure, les trois chaînes de la protéine Spike sont dans un état ouvert et actif. Chaque monomère de la protéine Spike possède une région de liaison au récepteur (en rouge) qui peut interagir avec le récepteur ACE2 de la cellule hôte. Cette interaction se produit via la région du motif de liaison au récépteur (RBM) encadrée en noir.

Tableau 1: Nombre total de poches, localisées au niveau de l'interface RBM, druggable, et hautement druggables identifiées sur les 1000 conformations issues des simulations de dynamique moléculaire du RBD dans l'état non lié à l'ACE2.

INTRODUCTION

Pour comprendre le fonctionnement des cellules, il est impératif de comprendre le fonctionnement des protéines car elles accomplissent la plupart des tâches des cellules vivantes (Alberts et al., 2002).

Une protéine cible est une entité biochimique dont l'activité peut être modulée par un principe actif (par exemple une petite molécule ou une enzyme) afin de stimuler l'activité thérapeutique pour agir sur l'état pathologique étudié. Les protéines, comme les enzymes, jouent un rôle clé dans l'organisme et sont responsables, de manière directe ou indirecte, de son bon fonctionnement. Dans le cas d'un dysfonctionnement cellulaire ou d'une pathologie, il est indispensable d'identifier les protéines cibles afin de développer des molécules médicamenteuses capables de remédier à la défaillance (Chan et al., 2010).

La localisation des protéines peut être soit à l'interface des cellules soit à l'intérieur de celles-ci et leur interaction avec les médicaments ou les enzymes peut se faire de façon directe en se liant à leur site actif afin d'inhiber ou d'activer une fonction, ou se lier à une protéine pour en inhiber une autre, cela peut déclencher une cascade de signalisation par exemple. Les protéines virales peuvent également être des cibles potentielles. Cela s'explique par le fait que les protéines virales sont souvent spécifiques aux virus et ne sont pas présentes dans les cellules de l'organisme. En inhibant leur fonction on peut donc cibler spécifiquement le virus sans affecter le bon fonctionnement des cellules de l'organisme. Cependant, il est important de noter que les protéines virales peuvent également muter et évoluer rapidement, ce qui peut rendre inefficaces certaines thérapies telles que les anticorps. C'est pour cette raison que les scientifiques sont constamment à la recherche de nouvelles stratégies efficaces à long terme et pouvant s'affranchir de l'effet des mutations.

Afin de contribuer efficacement à la validation des protéines en tant que cibles thérapeutiques *in silico,* il est primordial de caractériser leur « druggabilité » c'est-à-dire leur capacité à lier une molécule médicamenteuse. De ce fait, une mauvaise sélection de ces cibles thérapeutiques et le manque d'information sur leur druggabilité, sont une des principales causes du grand taux d'échec lors de l'étape préliminaire de la découverte des médicaments. Cependant, l'identification de ces sites de liaison potentiels est grandement facilitée par la disponibilité de la structure tridimensionnelle (3D) de la protéine cible.

La Protein Data Bank (PDB) (Berman et al., 2000) recense désormais plus de 190,000 structures de protéines déterminées expérimentalement, ce qui est une réussite considérable. Bien que cela ne représente qu'une fraction des milliards de séquences protéiques connues (UniProt : https://www.uniprot.org/).

Le nombre de protéines cibles identifiées chez l'homme est encore relativement faible, avec environ 893 biomolécules d'origine humaine et pathogène identifiées sur lesquelles agissent 1578 médicaments approuvés par la Food Drug Administration en 2016 (Santos et al., 2017). Des approches informatiques ont été développées pour permettre une exploration bioinformatique structurale à grande échelle. Ces approches sont de plus en plus importantes pour identifier des cibles thérapeutiques, même si leurs structures ne sont pas encore résolues expérimentalement. Les algorithmes de modélisation par homologie, qui sont une approche courante de la bioinformatique, permettent de prédire la structure d'une protéine en se basant sur la structure connue d'une protéine homologue. Un accès ouvert à plus de 200 millions de prédictions de structures de protéines est disponible pour accélérer la recherche scientifique (Jumper et al., 2021; Eswar et al., 2006). Avec les dernières avancées technologiques, l'identification de cibles thérapeutiques à partir de l'information de leur structure devient plus accessible.

Ainsi, le champ de la bioinformatique structurale connaît une croissance remarquable. De nombreux outils en bioinformatique structurale ont été développés depuis 2012 permettant d'identifier et de caractériser les cavités médicamenteuses à la surface des protéines appelées "poches protéiques" et d'en prédire leur druggabilité. Les poches les plus prometteuses pour le développement de médicaments sont celles dont les propriétés physico-chimiques et géométriques sont les plus favorables en termes d'affinité et de spécificité permettant une interaction optimale avec le ligand recherché. Toutefois, les protéines ne sont pas des structures rigides, mais plutôt des structures dynamiques et flexibles, qui peuvent subir des changements de conformation en réponse à des stimuli environnementaux et à des interactions avec d'autres molécules.

En bioinformatique structurale, il est possible de simuler *in silico* la dynamique des biomolécules par des simulations de dynamique moléculaire (DM). La variabilité structurale des protéines au cours de leur dynamique conduit à des variations des poches situées à la surface des différentes conformations. Ainsi, en suivant ces poches, il est possible de caractériser plus efficacement les sites éventuels de liaison (Alberts et al., 2002).

Dans ce contexte, l'objectif de ma thèse a été de développer un protocole de recherche de poches protéiques qui prend en compte les propriétés dynamiques des protéines et par conséquent la flexibilité de leurs sites actifs qui seraient susceptibles d'accepter une molécule médicamenteuse dite "*Drug-like*", c'est-à-dire qui possède des propriétés chimiques et physiques qui la rendent potentiellement appropriée pour être utilisée comme médicament. La prise en compte de l'impact de certaines mutations sur les sites de liaison est aussi un point important car des mutations peuvent entraîner des changements conformationnels et induire une perte d'interaction avec les ligands ou encore des effets secondaires.

Dans ma thèse, je me suis particulièrement intéressée à l'analyse structurale de deux protéines virales pour identifier des sites de liaison potentiels et les résidus clés à cibler par une molécule thérapeutique pour inhiber l'interaction de la protéine. Mon protocole a d'abord été appliqué sur la protéine Non-structurale 1 (NS1) du virus Influenza A, responsable de l'infection des cellules de l'organisme par le virus Influenza A en contrant les défenses antivirales de la cellule infectée, aidant ainsi le virus à échapper au système immunitaire inné (Hale et al., 2008; Zhirnov et al., 2002). Cette protéine est connue pour son polymorphisme structural et selon les sous-types et les formes dans lesquelles elle peut exister, elle présente des niveaux de virulence différents. Pour cela nous avons procédé à l'étude de la flexibilité de la protéine NS1 par des simulations de dynamique moléculaire afin d'établir le lien entre les formes de NS1 et leur sous-type. Ces simulations ont permis de montrer la forte stabilité du domaine de liaison à l'ARN (RNA-BD) et prouver que son polymorphisme structural n'était pas dépendant de sa séquence. Ce travail a donné lieu à une publication en premier auteur dans "*International Journal of Molecular Sciences*".

Après avoir démontré la stabilité du RNA-BD de la protéine NS1 pour les différents sous-types étudiés, j'ai poursuivi mes travaux avec une recherche et un suivi des poches localisées au niveau du sillon du RNA-BD, situé entre deux hélices alpha, sur des conformations issues des trajectoires de dynamique moléculaire, afin d'identifier les cavités principales de l'interface entre la protéine NS1 et l'ARN viral (région RNA-BD). En utilisant les DM obtenues lors de la précédente étude, un site de liaison stable, fréquent et commun à tous les sous-types a été identifié et caractérisé. Ce site de liaison présente des propriétés de druggabilité qui en font une cible thérapeutique intéressante pour la protéine NS1. Les résultats de cette recherche ont été publiés dans le journal "*Biomolecules*".

J'ai, par la suite, mis en place un protocole pour la contribution à la validation du domaine de liaison au récepteur de la protéine Spike du virus SARS-CoV2 en tant que cible thérapeutique. L'interaction entre l'enzyme angiotensine de conversion 2 (ACE2) et le RBD de la protéine Spike est la voie d'entrée principale du virus dans la cellule hôte humaine, ce

qui conduit à l'infection. Pour étudier la flexibilité du RBD, j'ai effectué des simulations de dynamique moléculaire. Cependant, cette fois-ci, j'ai élargi la recherche et le suivi des poches à l'ensemble de la surface de la protéine, contrairement à la recherche ciblée et limitée au RNA-BD effectuée sur la protéine NS1. Cette approche plus large m'a permis d'explorer les poches à une plus grande échelle et d'identifier trois sites de liaison d'intérêt situés à différentes localisations de la protéine. Parmi ceux-ci, deux ont déjà été cités dans la littérature, tandis que le troisième n'a jamais été identifié auparavant. Ces sites d'intérêt pourraient être ciblés pour empêcher l'interaction entre le RBD et ACE2. Les résultats de cette recherche ont été publiés dans un article dans “Computational and Structural Biotechnology Journal”.

CHAPITRE 1. État de l’art

1.1 Exploration des protéines

1.1.1 Définition des protéines

Les protéines sont des macromolécules biologiques composées d'acides aminés, qui sont liés les uns aux autres par des liaisons peptidiques pour former des chaînes polypeptidiques. Elles jouent un rôle crucial dans la plupart des processus biologiques, notamment la régulation de la croissance et de la division cellulaire, la réponse immunitaire, la détection et la réponse aux signaux extérieurs, et la catalyse enzymatique. Les protéines ont été décrites pour la première fois par Jöns Jacob Berzelius en 1838. Depuis lors, de nombreuses études ont été menées pour comprendre leur structure, leur fonction et leur rôle dans la biologie cellulaire.

La fonction de chaque protéine est liée à sa structure unique. La séquence d'acides aminés d'une protéine détermine sa structure secondaire et tertiaire, qui détermine à son tour sa fonction. La fonction des protéines est, par conséquent, intimement liée à leur structure. C’est pour cette raison que des banques de données, mises à disposition de toute la communauté scientifique, et qui recensent les protéines déterminées expérimentalement, ont été créées.

1.1.2 Les banques de données protéiques

L'histoire des bases de données de protéines remonte aux années 1960 et 1970, lorsque les premiers efforts ont été faits pour compiler les séquences de protéines à partir de publications scientifiques. Les premières bases de données de séquences comprenaient PIR (Protein Information Resource), qui était maintenue par l'Université de Georgetown (Barker, 1998), GenBank (une base de données de séquences nucléiques), qui était maintenue par le National Institutes of Health (NIH) aux États-Unis (Benson et al., 2009), Swiss-Prot (Bairoch & Boeckmann, 1993), créée par le Swiss Institute of Bioinformatics (devenu par la suite UniProt, base de données de séquences protéiques). Ces premières bases de données ont été principalement construites manuellement et ont nécessité beaucoup de travail pour collecter les informations. Avec l'avancement de la technologie de séquençage, il y avait de plus en plus de données disponibles et les bases de données se sont développées pour inclure des séquences de protéines de divers organismes, des domaines de protéines et des données fonctionnelles. Les bases de données de séquences nucléiques et protéiques ont été des outils clés dans la compréhension de la biologie moléculaire, mais elles ont leurs limites en termes de compréhension des fonctions et des interactions des molécules biologiques. Pour cela, la connaissance de la structure tridimensionnelle des molécules biologiques est cruciale. Les premières études de cristallographie aux rayons X sur la structure des protéines ont commencé bien avant les années 1970. En fait, les premières structures de protéines ont été résolues dans les années 1950, notamment la structure de la myoglobine en 1958 (Kendrew et al., 1958), pour laquelle John Kendrew et Max Perutz ont reçu le prix Nobel de chimie en 1962. En ce qui concerne les bases de données de structures des macromolécules biologiques, une base de données a été mise en place dans les années 1970, la Protein Data Bank. Cette banque de données recense à la fois les informations sur la structure tridimensionnelle des protéines (en majorité), des acides nucléiques ou encore de leurs complexes avec d’autres molécules biologiques. En 2022, le nombre de structures de protéines humaines disponibles dans la PDB a été estimé à environ 198.000.

1.1.3. La structure des protéines

La détermination des protéines expérimentalement passe par des étapes d'Extraction de la protéine à partir d'échantillons biologiques (par exemple, les cellules, les tissus ou les fluides corporels) dans lequel elle se trouve. Elles sont ensuite purifiées pour éliminer les contaminants et obtenir une protéine pure. Une fois la protéine purifiée, sa structure est analysée par les méthodes de cristallographie aux rayons X, la résonance magnétique nucléaire (RMN), la spectrométrie de masse, etc… (Matthews, 1976; Wüthrich, 1990).

Ce processus de détermination des structures protéiques expérimentalement peut être très long et très coûteux. C'est pour cette raison que des chercheurs en bioinformatique ont mis à disposition de la communauté scientifique plusieurs outils de prédiction de la structure des protéines, considérée comme un modèle structural de la protéine. Ces modèles 3D de protéine permettent d'aider à comprendre la fonction de ces protéines et à développer des thérapies sur la base de ces modèles structuraux en un temps plus court.

Le développement de tels outils a révolutionné le domaine de la biologie structurale. En effet, le nombre de structures tridimensionnelles prédites a considérablement augmenté ce qui a contribué à l'accroissement des stratégies thérapeutiques envisagées.

Parmi les outils permettant de modéliser des structures de protéines à partir de leur séquence, on peut citer SWISS-MODEL (Waterhouse et al., 2018) et Modeller (Eswar et al., 2006) qui sont des outils de modélisation par homologie. Le processus de modélisation commence par l'alignement de la séquence de la protéine cible avec une séquence similaire déjà résolue expérimentalement. I-Tasser (Iterative Threading Assembly Refinement) (Yang et al., 2015) est une approche hiérarchique de la prédiction de la structure des protéines, basée sur des méthodes d'enfilage. Alphafold (Jumper et al., 2021) est un système d'intelligence artificielle développé par DeepMind. La dernière version d'AlphaFold repose sur une nouvelle approche d'apprentissage automatique qui inclut les connaissances physiques et biologiques à partir de la structure des protéines, et en s'appuyant sur les alignements multi-séquences, dans un algorithme de deep learning. Le développement d'Alphafold a été considéré comme une avancée majeure dans le domaine de la prédiction de la structure des protéines, car il a permis d'obtenir des résultats de haute qualité à une vitesse beaucoup plus rapide que les méthodes de prédiction précédentes.

La modélisation de structures 3D est donc considérée comme une bonne alternative lors de l'identification d'une cible thérapeutique non disponible sur les banques de données de structure de protéines.

Il convient de noter que la prédiction de la structure de protéines à partir de la séquence est une tâche complexe et que les résultats peuvent varier en fonction de la qualité des données d'entrée et de la complexité de la protéine étudiée. Les modèles générés par ces outils doivent donc être validés expérimentalement pour confirmer leur précision. Bien que ces outils de prédiction de la structure des protéines présentent des limites, ils sont extrêmement utiles pour l'aide à la découverte de nouvelles cibles thérapeutiques et pour comprendre les fonctions des protéines. Ils permettent également d'accélérer le développement de médicaments en fournissant des informations précieuses sur la structure des protéines cibles.

1.2 Identification des cibles thérapeutiques

1.2.1 Processus d'identification d'une cible thérapeutique

La première étape de l'identification d'une cible thérapeutique consiste à identifier la maladie à traiter. Cette étape peut être réalisée en analysant les symptômes, les signes cliniques et les données épidémiologiques. Une fois la maladie identifiée, il est important de

comprendre les mécanismes biologiques sous-jacents qui sont responsables de la maladie en utilisant des techniques d'analyse biochimique, de biologie cellulaire et de génétique pour étudier les voies de signalisation, les processus métaboliques et les interactions moléculaires impliquées dans la pathologie. Lors de cette exploration, les protéines impliquées dans la maladie sont identifiées puis validées pour leur pertinence biologique. Une fois les protéines validées comme des cibles potentielles pour le traitement de la maladie, il est important d'évaluer leur faisabilité thérapeutique en analysant leur localisation, leur structure, leur stabilité et leur accessibilité pour les médicaments. La dernière étape consiste à concevoir des médicaments qui peuvent interagir avec ces cibles de manière spécifique et efficace.

Toutefois, le processus d'identification d'une cible thérapeutique peut prendre plusieurs années et nécessite une collaboration entre des scientifiques de différentes disciplines, notamment la biochimie, la biologie cellulaire, la génétique, la pharmacologie et la chimie médicinale.

Aujourd'hui, la bioinformatique offre une approche plus rapide et efficace pour l'analyse de grandes quantités de données biologiques, facilitant ainsi l'identification rapide de cibles thérapeutiques potentielles pour les maladies. Les informations générées par ces approches peuvent être utilisées pour concevoir des médicaments spécifiques qui ciblent les protéines responsables d'une maladie donnée, rationalisant ainsi la conception de médicaments et réduisant les coûts de développement. En outre, en ciblant spécifiquement une protéine, cette approche *in silico* réduit les risques d'effets secondaires indésirables associés aux médicaments qui ciblent plusieurs protéines.

1.2.2 Identification des sites de liaison

Les interactions entre les protéines et leurs ligands jouent un rôle essentiel dans la compréhension des processus biologiques au niveau moléculaire. Les sites de liaison représentent les zones spécifiques de la protéine qui interagissent avec le ligand. Pour caractériser ces interactions, il est possible de mesurer expérimentalement les constantes de dissociation entre la protéine et le ligand, ou de provoquer des mutations qui altèrent l'affinité du complexe protéine-ligand. En identifiant les résidus clés du site de liaison, on peut mieux comprendre le mécanisme d'interaction et concevoir des médicaments et des thérapies ciblant ces interactions. Dans les années 1990, de nouvelles méthodes informatiques ont été développées pour identifier les sites de liaison protéiques. Les algorithmes de surface moléculaire ont été utilisés pour générer des cartes de surface des protéines et pour identifier les cavités internes (Wallace et al., 1995). Les algorithmes d'amarrage moléculaire (Docking) ont également été développés pour prédire les interactions entre des ligands connus et la surface des protéines. Au cours des années 2000, des techniques expérimentales ont été développées pour identifier les sites de liaison, notamment grâce à l'utilisation de la résonance magnétique nucléaire (RMN) et de la spectroscopie de masse pour étudier les interactions entre les protéines et les ligands en solution. De nouvelles techniques expérimentales ont permis d'obtenir des images détaillées de la structure tridimensionnelle des protéines en solution telle que la cryo-microscopie électronique (Cryo_EM).

Cependant, l'identification des sites de liaison avec ces méthodes est limitée par le fait qu'un ligand doit nécessairement être présent pour cette détection, c'est pour cela que des techniques plus prometteuses en bioinformatique ont émergé afin de permettre une identification et une caractérisation des poches à la surface des structures protéiques non-complexées à des ligands présentant les propriétés de druggabilité nécessaire à la fixation d'une petite molécule médicamenteuse. Aujourd'hui, une combinaison de techniques informatiques et expérimentales est utilisée pour identifier et caractériser les poches protéiques. Cette approche est particulièrement utile pour la découverte de nouveaux médicaments et la conception de thérapies ciblées.

1.3 La dynamique des protéines

1.3.1 La flexibilité des protéines et des sites de liaison

Les protéines ne sont pas des structures statiques, mais plutôt des ensembles de molécules en constante agitation thermique. Cette flexibilité leur permet d'une part de stabiliser leur structure tridimensionnelle et d'autre part d'interagir avec d'autres molécules, comme les ligands, les enzymes et autres protéines (Carlson & McCammon, 2000). Il a été démontré que dans certains cas la présence d'un ligand pouvait non seulement affecter la fonction mais aussi la conformation des protéines. Par exemple, une protéine peut changer de forme lorsqu'elle se lie à un ligand, ce qui peut avoir des conséquences importantes sur la fonction de la protéine et pour l'ensemble de la cellule ou de l'organisme. L'α-lactalbumine est un exemple de protéine qui exerce différentes fonctions biologiques en fonction de son état conformationnel. Dans un pH acide et en présence de cofacteur de type "acide gras C18:1" elle adopte une forme qui peut induire l'apoptose dans les cellules tumorales et immatures tout en épargnant les cellules saines (Håkansson et al., 1995; Svensson et al., 2000; Köhler et al., 2001). Une autre fonction de l'α-lactalbumine est liée à son effet protecteur contre les ulcères gastriques. Il a été découvert que l'α-lactalbumine était seule responsable de cet effet (Matsumoto et al., 2001). Ainsi, l'α-lactalbumine a plusieurs fonctions biologiques qui dépendent des conformations qu'elle adopte et des cofacteurs qu'elle lie.

Au niveau local des protéines, les cavités susceptibles d'accueillir un ligand ont souvent été étudiées sur des structures rigides, c'est-à-dire en supposant que la structure tridimensionnelle de la protéine cible est fixe et ne subit pas de changements conformationnels significatifs. Cependant, il est maintenant bien établi que la structure tridimensionnelle de la protéine n'est pas rigide, et que la protéine peut subir des fluctuations. Ces changements conformationnels peuvent modifier la géométrie et la flexibilité des sites de liaison, affectant ainsi la capacité des protéines à interagir avec différents ligands. Une étude récente a mis en évidence l'importance de détecter les sites de liaison cryptiques qui ne sont pas visibles dans les structures cristallographiques en l'absence de ligand, mais qui deviennent visibles lors d'événements de liaison. Ces sites présentent des propriétés de druggabilité favorables à la liaison des ligands et pourraient offrir une opportunité de cibler des protéines difficiles à atteindre. Toutefois, leur nature cachée les rend difficiles à repérer par les méthodes de screening expérimental. Les simulations moléculaires figurent parmi les méthodes les plus efficaces pour identifier et caractériser ces sites de liaison cryptiques (Kuzmanic et al., 2020).

Pour tenir compte de cette flexibilité, des techniques telles que la dynamique moléculaire (Karplus & Petsko, 1990; Parrinello & Rahman, 1981) ont été développées en bioinformatique pour étudier l'espace des conformations possibles des protéines. La DM utilise l'équation de l'énergie potentielle et un champ de forces spécifiques, composée de termes liés correspondant aux interactions entre les atomes et de termes non liés correspondant aux interactions de van der waals, des liaisons hydrogènes ainsi que des interactions entre charges. Il s'agit de simulations numériques consistant à calculer l'évolution temporelle des positions et des vitesses d'un système composé de N atomes en interaction, en intégrant numériquement les équations de la mécanique classique newtonienne. En introduisant des vitesses sur chaque coordonnée atomique, on peut calculer l'énergie totale du système, qui est la somme de l'énergie cinétique (associée au mouvement des atomes) et de l'énergie potentielle (associée aux interactions entre les atomes). La dynamique moléculaire permet ainsi d'observer l'état d'une protéine et les changements qui se produisent au cours du temps, en modélisant les effets de la température, de la pression et des forces électromagnétiques. Les trajectoires issues de ces simulations sont par la suite analysées par

des calculs de Root Mean Square Deviation (RMSD), cette mesure est utilisée pour comparer les positions d'atomes sur les différentes conformations au cours de la DM pour suivre l'évolution du changement conformationnel des protéines. Au niveau local des protéines, il est possible de mesurer la flexibilité des acides aminés dans une protéine donnée, en calculant l'écart de position de chaque atome dans une conformation donnée par rapport à sa position moyenne sur l'ensemble des conformations de la trajectoire grâce au Root Mean Square Fluctuation (RMSF). Ces approches permettent d'obtenir des informations sur les différents états conformationnels de la protéine et de déterminer comment ces états sont affectés par l'environnement ou par l'interaction avec d'autres molécules biologiques.

1.4 Les différents outils de recherche de poche

Les cavités (poches) à la surface des protéines jouent un rôle crucial dans leur fonctionnement et sont les principales régions à cibler par de petites molécules médicamenteuses. Il existe plusieurs outils de détection pour identifier ces poches, chacun basé sur ses propres techniques et algorithmes. Certains outils se basent sur le calcul de la surface moléculaire, comme CASTp (Computed Atlas of Surface Topography of Proteins), pour repérer ces poches (Binkowski et al., 2003). D'autres outils, tels que LigSite, utilisent une analyse détaillée des interactions entre la protéine et le ligand en combinant des informations géométriques et énergétiques pour identifier les sites de liaison les plus favorables (Hendlich et al., 1997).

Les outils de détection de poches protéiques peuvent être limités par l'absence d'informations sur la capacité d'un site de liaison à accueillir un médicament, ce qu'on appelle la druggabilité. Pour pallier cette limitation, de nouveaux outils ont été développés pour estimer les poches, même en l'absence de ligand, et fournir un score de druggabilité. Par exemple, l'outil DoGSitescorer utilise un algorithme d'apprentissage automatique entraîné sur un ensemble de données de sites de liaison protéiques connus comme étant "druggable" et "non-druggable" (Volkamer et al., 2012). De même, PockDrug (Borrel et al., 2015; Hussein et al., 2015) est un outil de prédiction de la druggabilité de poches qui se base sur des descripteurs physico-chimiques et géométriques après avoir estimé les cavités à l'aide de Fpocket (Guilloux et al., 2009) donnant un score de druggabilité allant de 0 à 1, à savoir que les poches dont le score de druggabilité est inférieur 0.5 sont considérées comme non druggables et les poches avec un score de plus de 0.5 sont dites druggables. Cette combinaison 'estimation des poches et prédiction de la druggabilité' permet de mettre en évidence des poches putatives pour estimer les cavités médicamenteuses.

Bien que l'estimation des poches et de la prédiction de leur druggabilité soit une grande avancée pour l'identification des sites de liaison aux médicaments, de nouveaux outils ont été développés il existe très peu d'outils permettant d'estimer les poches sur des protéines dynamiques.

Parmi les rares outils disponibles, on peut citer POVME (Wagner et al., 2017), qui utilise une méthode basée sur la DM pour identifier et caractériser les cavités dans les protéines et de mesurer leur volume pour aider à la conception de ligands ou encore D3Pocket (Chen et al., 2019) qui est également un outil de prédiction de poches protéiques qui utilise une méthode de calcul basée sur la DM pour prendre en compte la forme, la taille, la profondeur, les propriétés physico-chimiques et la flexibilité de la protéine. Cet outil fournit trois principales informations sur les caractéristiques des poches: la stabilité des résidus dans les poches au cours des simulations, la corrélation entre des groupes de poches en termes de taille et l'estimation de la continuité des poches au cours du temps en termes de déformabilité. Sa principale limite est qu'il ne permet pas le traitement de fichiers de trajectoire excédant un volume de 50MB. De plus les deux outils précédemment cités ne sont

pas conçus pour calculer un score de druggabilité, pour obtenir des informations sur la druggabilité des poches identifiées, il est nécessaire de les combiner avec d'autres outils tels que DoGSitescorer ou Pockdrug.
A ce jour, aucun outil complet ne permet d'estimer les sites de liaison à la surface des protéines, de les caractériser et de prédire leur druggabilité tout en tenant compte de leur flexibilité.

1.5 Les propriétés des poches et leur druggabilité

Un potentiel site de liaison est considéré comme druggable s'il est capable d'accepter une molécule "drug-like" dont les propriétés incluent notamment une taille et une forme appropriées pour s'adapter à des récepteurs spécifiques dans le corps, une solubilité suffisante pour permettre une distribution efficace dans le corps, une stabilité métabolique, une biodisponibilité adéquate et une toxicité acceptable. Les molécules "drug-like" sont souvent identifiées à l'aide de critères physico-chimiques décrits par Lipinski (Lipinski et al. 2001) tels que la masse moléculaire (inférieur à 500 Daltons), le coefficient de partition octanol-eau (logP) inférieur à 5, le nombre de liaisons hydrogène-donneur (inférieur à 5) et hydrogène-accepteur (inférieur à 10). D'autres critères sont le nombre d'atomes de carbone et la présence de groupes fonctionnels spécifiques (Ealy et al., 2017; Benet et al., 2016). Toutefois, ces critères ne garantissent pas une efficacité pour une utilisation chez l'homme. Une série d'essais et d'études sont nécessaires pour évaluer la sécurité et l'efficacité d'une molécule avant qu'elle ne soit approuvée comme médicament.

Pour qu'une poche protéique puisse être utilisée comme un potentiel site de liaison aux médicaments efficaces, elle doit remplir plusieurs critères importants dont la complémentarité géométrique et physico-chimique. Cela signifie que la forme du ligand doit s'adapter à celle de la poche et que les interactions intermoléculaires entre le ligand et la poche doivent être optimisées. La poche doit contenir des groupes fonctionnels appropriés tels que des groupes hydrophobes, des groupes électrostatiques ou des groupes aromatiques, qui peuvent interagir avec la molécule médicamenteuse de manière sélective. La flexibilité des sites de liaison est également un critère important (section 1.3.1) car elle doit être suffisamment flexible pour s'adapter à la forme du ligand et être capable de s'en dissocier lorsque la fonction est accomplie. Enfin, la poche doit être suffisamment accessible pour que le médicament puisse y accéder. Cela dépend de la localisation de la poche dans la protéine et de l'accessibilité de la poche à partir de la surface de la protéine.

CHAPITRE 2. Protocole d’identification des sites de liaison

2.1 Sélection de la protéine cible et étude de sa flexibilité

La sélection d’une cible thérapeutique pertinente nécessite une compréhension approfondie de la pathologie impliquée. Il est essentiel de rechercher les informations importantes sur la maladie, les symptômes, les mécanismes sous-jacents et les voies métaboliques impliquées pour identifier les protéines qui jouent un rôle dans la pathologie. Pour ce faire, il faut parcourir des bases de données en ligne telles que Uniprot ou encore effectuer des recherches bibliographiques pour rassembler les connaissances existantes.

Une fois la protéine d'intérêt identifiée, il est important d'évaluer sa pertinence biologique en examinant ses interactions avec d'autres molécules pour comprendre sa fonction biologique. Avant de valider le choix de la cible à étudier, il est important de s'assurer que des données structurales de bonne résolution sont disponibles sur la protéine dans la PDB pour une analyse précise. Dans le cas où ces données ne sont pas disponibles, il peut être nécessaire de construire un modèle de la protéine, par exemple en utilisant des méthodes de modélisation par homologie en utilisant des protéines structurées similaires comme modèles.

Une fois que nous disposons de la structure de la protéine d'intérêt, une étape importante consiste à évaluer sa flexibilité en réalisant des simulations de dynamique moléculaire (Karplus & Petsko, 1990; Parrinello & Rahman, 1981). Dans ce travail, l'outil GROMACS (Abraham et al., 2015; Berendsen et al., 1995) a été utilisé . C'est un outil bien fourni en paramètres et options qui permet une grande souplesse pour s'adapter aux besoins spécifiques de chaque simulation. Différents paramètres permettent d'ajuster les simulations selon le système étudié. Par exemple, le fichier de topologie contient les informations sur les atomes, les liaisons, les angles, et autres détails. Le fichier de coordonnées donne les positions exactes des atomes dans l'espace tridimensionnel. La périodicité est mise en place à l'aide d'une boîte spéciale qui permet de simuler des systèmes infinis. Les potentiels de force, comme Lennard-Jones et Coulomb, décrivent les interactions (attractives et répulsives) entre les atomes. Les intégrateurs temporels, tels que Verlet et leapfrog, calculent les trajectoires des particules au cours du temps et permettent de résoudre les équations différentielles de mouvement des atomes dans le système. Des ensembles thermostatiques et barostatiques permettent de contrôler la température et la pression du système simulé afin de conserver l'équilibre dynamique du système. Enfin, les paramètres de simulation tels que la durée, le pas de temps, la température initiale, la pression initiale, sont définis dans les fichiers de configuration.

En examinant les résultats de ces simulations, nous pouvons obtenir par des calculs de RMSD, de RMSF et de distance entre les différentes régions de la protéine, des informations précieuses sur sa flexibilité et sur la façon dont elle interagit avec d'autres molécules dans son environnement biologique.

2.2 Echantillonnage des conformations au cours des dynamiques moléculaires et recherche de poches druggables

Au cours de la simulation de dynamique moléculaire, une grande quantité de données est générée. Pour explorer les mouvements de la protéine à différents moments, on extrait des conformations à partir des trajectoires de DM. L'extraction de conformations est une étape importante dans l'analyse des simulations de DM, car elle permet de comprendre la flexibilité de la protéine, ainsi que les changements structurels qui se produisent au cours du temps. Ces conformations peuvent être extraites aléatoirement ou à des intervalles de temps réguliers afin

de parcourir le maximum d'états conformationnels que la protéine peut adopter au cours des simulations de DM, et pour explorer de manière efficace l'espace des conformations possibles et de capturer les transitions conformationnelles importantes.

Une fois que les conformations sont extraites, les poches à leur surface sont identifiées avec l'outil PocketDrug qui est un logiciel de bio-informatique, développé en interne au sein de notre laboratoire (Borrel et al., 2015). Ce logiciel utilise 52 descripteurs pour fournir une description détaillée des poches estimées à la surface de chaque conformation. Un ensemble de 36 descripteurs physico-chimiques et 16 descripteurs géométriques qui permettent de prédire la druggabilité de ces cavités qui peuvent servir de sites potentiels de liaison pour des petites molécules médicamenteuses. Parmi ces propriétés physico-chimiques et géométriques on retrouve la polarité, l'hydrophobicité, l'accessibilité de surface, la présence de résidus chargés ou d'acides aminés aromatiques, le volume ainsi que la forme et la profondeur des poches.

Cet outil est basé sur des méthodes d'apprentissage automatique supervisé pour prédire un score de druggabilité, en utilisant la méthode d'estimation de poches Fpocket. Ces méthodes sont basées sur des modèles préalablement entraînés à partir de données d'entraînement contenant des informations sur les caractéristiques physico-chimiques et géométriques des poches ainsi que leur druggabilité correspondante. En utilisant ces modèles, PockDrug est en mesure d'évaluer de manière prédictive la druggabilité des poches dans de nouvelles conformations. Les descripteurs extraits des poches sont utilisés comme entrées pour le modèle, qui génère ensuite un score de druggabilité en se basant sur les relations apprises lors de la phase d'entraînement.

Après avoir analysé les sites de liaison dans chaque conformation, différents filtres peuvent être appliqués aux poches afin de les sélectionner de manière plus spécifique. Parmi ces filtres, on trouve la druggabilité, qui permet de ne garder que les poches susceptibles de se lier à des ligands en se basant sur le score de durggabilité. On peut également utiliser un filtre basé sur la taille des poches, en sélectionnant par exemple les grandes poches ou en éliminant les poches ayant moins de résidus qu'un seuil défini. De plus, il est possible d'appliquer un filtre basé sur la localisation des poches en croisant les résidus clés d'intérêt avec les poches identifiées. Cela permet de cibler spécifiquement les poches situées dans une région de la protéine qui est supposée être favorable à l'interaction avec un partenaire, cette région peut être définie par exemple en repérant les résidus de la protéine impliqués dans les interactions avec d'autres protéines ou des ligands. En utilisant ces filtres, on peut affiner la sélection des poches et se concentrer sur celles qui sont les plus pertinentes pour l'étude en cours.

2.3 Analyse et classification des poches obtenues sur l'échantillon de conformations

Une fois que les poches ont été identifiées, une analyse statistique multivariée non supervisée de l'ensemble des poches est effectuée sur l'échantillon de conformations 3D. La classification hiérarchique est utilisée pour regrouper les poches en fonction de la similarité de leurs acides aminés constitutifs. Ainsi, les poches qui partagent les mêmes acides aminés se retrouvent dans les mêmes classes, offrant ainsi une vue plus précise des motifs structuraux communs. Contrairement aux approches précédentes qui se limitaient à classifier les poches en fonction de leurs propriétés physico-chimiques et géométriques, cette nouvelle méthode de classification repose sur le suivi des résidus constitutifs des poches. Elle permet d'obtenir une représentation structurée des différentes classes de poches, facilitant ainsi l'interprétation des résultats et l'identification de motifs récurrents ainsi que des résidus clés favorisant la druggabilité. En suivant les résidus, il devient possible de détecter les variations subtiles dans la composition des acides aminés au sein des classes de poches, ce qui contribue à une meilleure compréhension des relations structure-fonction dans le cadre de notre étude.

Pour faciliter la visualisation et l'analyse de cette classification, des scripts R ont été développés. Ces scripts permettent de générer des représentations graphiques telles que des heatmaps, qui affichent la similarité entre les poches et mettent en évidence les regroupements. De plus, ces scripts permettent de réaliser des tableaux de comptage qui indiquent le nombre de poches dans chaque classe, ainsi que des histogrammes qui représentent la fréquence des résidus présents dans les poches, le score de druggabilité associé à chaque poche, ainsi que la proportion des classes par rapport au nombre total de poches identifiées.

Cette classification repose sur une méthodologie rigoureuse, en utilisant la méthode Ward D2 (Murtagh & Legendre, 2014) qui est une méthode de classification hiérarchique couramment employée en statistique et en apprentissage automatique pour regrouper des observations en fonction de leur similarité. Dans le contexte de notre étude, la similarité entre les poches est évaluée en se basant sur les résidus communs qu'elles contiennent. Plus précisément, la méthode de Ward D2 calcule la distance entre deux groupes en mesurant la somme des carrés des différences entre les observations de chaque groupe. En d'autres termes, elle cherche à minimiser la variance intra-groupes, en regroupant les poches qui partagent des caractéristiques similaires. Ainsi, plus la somme des carrés des différences entre les observations de chaque groupe est faible, plus la similarité entre les poches est élevée.

L'utilisation de la méthode de Ward D2 permet d'obtenir une classification hiérarchique précise et structurée des poches, en tenant compte de leur similarité en termes de résidus communs. Cette approche fournit une vue d'ensemble des regroupements et des relations entre les poches, facilitant ainsi l'interprétation des résultats et la compréhension des motifs structuraux partagés.

2. 4 Identification des sites de liaison

Les poches identifiées et classifiées font l'objet d'une analyse approfondie pour évaluer leur druggabilité et leur localisation dans la protéine. La classification hiérarchique permet de suivre l'évolution des poches ainsi que l’identification des sites de liaison au cours des différentes conformations échantillonnées lors des simulations de dynamique moléculaire. Cette analyse permet d'identifier les sites de liaison les plus fréquemment observés, caractérisés par des propriétés physico-chimiques et géométriques favorables à l'interaction avec des ligands médicamenteux.

En associant les résidus identifiés aux poches classifiées, on obtient une représentation claire des sites potentiels de liaison dans la protéine, ce qui facilite l'interprétation des résultats et la visualisation des régions clés impliquées dans l'interaction avec les ligands. Ces résidus clés peuvent être utilisés comme cibles potentielles pour la conception de ligands médicamenteux spécifiques et efficaces vis-à-vis de la protéine cible.

De plus, le suivi des résidus au cours des simulations permet de détecter les variations conformationnelles des sites de liaison, en identifiant les résidus flexibles ou adoptant différentes conformations. Ces variations peuvent avoir un impact sur l'accessibilité et l'affinité des sites de liaison pour les ligands, ce qui est essentiel pour comprendre la dynamique des interactions ligand-protéine et guider la conception de molécules modulatrices de l’activité.

Enfin, les sites de liaison potentiels sont sélectionnés en utilisant des critères tels que la taille de la poche, la similarité structurale avec des sites de liaison connus, la proximité avec d'autres résidus importants et la stabilité conformationnelle. Cette étape de sélection permet de retenir les sites les plus pertinents pour le développement de médicaments spécifiques aux propriétés des sites de liaison, en mettant l'accent sur les résidus les plus fréquents et les caractéristiques structurales significatives.

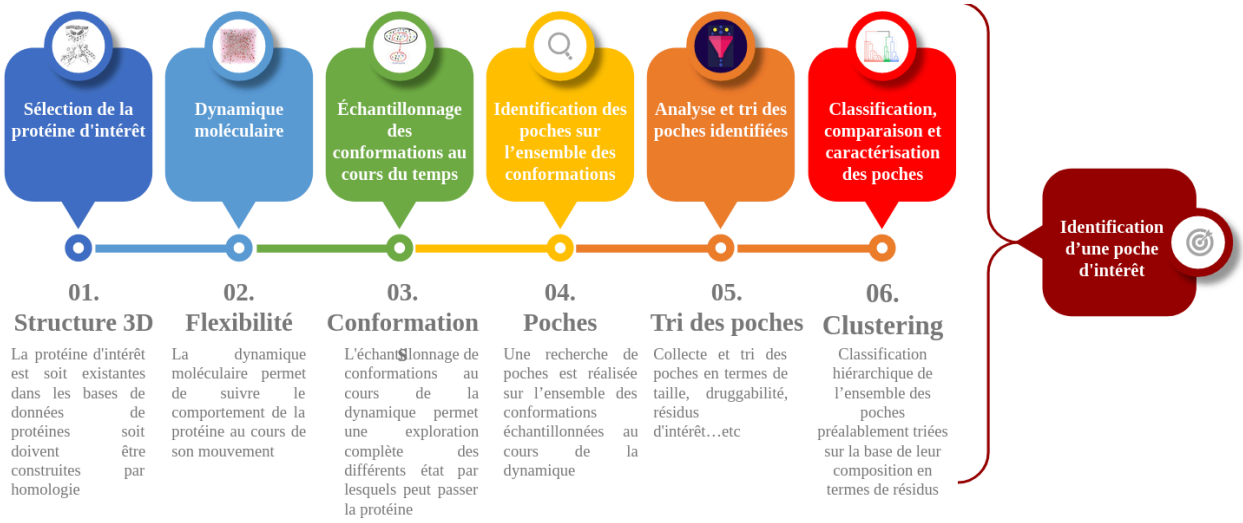


Figure 1: Schéma décrivant les différentes étapes du protocole utilisé pour identifier les sites de liaison sur des structures dynamiques. Ce protocole comporte les étapes suivantes: Sélection de la protéine d'intérêt, étude de la flexibilité de la protéine par simulation de dynamique moléculaire, échantillonnage des conformations au cours des dynamiques, détection des poches sur chacune des conformations, évaluation des poches en termes de propriété de druggabilité et de localisation, classification des poches en fonction de leur propriété de druggabilité et d'emplacement, enfin sélection des sites de liaison potentiels en utilisant des critères de sélection appropriés

Le protocole présenté dans le schéma de la figure 1 a été appliqué à deux protéines virales: la protéine NS1 du virus Influenza A et le domaine RBD de la protéine Spike impliqué dans l'infection par le virus SARS-CoV-2.

CHAPITRE 3. Analyse structurale de la protéine NS1

3.1 Etude du polymorphisme de la protéine NS1

Cette partie du travail est consacrée à l'étude du polymorphisme structural des sous-types de la protéine NS1 du virus Influenza A. L'analyse de la flexibilité des protéines est une étape importante dans la recherche de cibles thérapeutiques. En effet, la capacité des protéines à adopter différentes conformations est essentielle pour leur fonctionnement biologique. Cet aspect dynamique des protéines est donc un critère indispensable à prendre en compte dans la validation des cibles thérapeutiques. Dans cette partie, la flexibilité de plusieurs sous-types de la protéine NS1 du virus Influenza A connue pour son polymorphisme structural a été étudiée et a donnée lieu à la publication suivante:
Naceri, S., Marc, D., Camproux, A. C.*, & Flatters, D.* (2022). Influenza A Virus NS1 Protein Structural Flexibility Analysis According to Its Structural Polymorphism Using Computational Approaches. *International Journal of Molecular Sciences*, *23*, 1805.

3.1.1 Présentation de la protéine NS1

La protéine NS1 du virus Influenza A joue un rôle crucial dans le blocage de la réponse immunitaire chez les virus de la grippe. Bien qu'elle ne soit pas incluse dans la particule virale, NS1 est hautement exprimée dans les cellules infectées. En effet, le virus Influenza A pénètre à l'intérieur de la cellule de l'hôte par attachement via les membranes, le virus est ensuite endocyté, ses particules virales sont libérés dans la cellule et fusionnent avec le matériel génétique de la cellule hôte pour produire les protéines virales (Figure 2). Parmi ces protéines on retrouve la protéine NS1 dont l'action principale consiste à inhiber la synthèse cellulaire et à empêcher l'activation de facteurs clés du système interféron. En outre, NS1 favorise la synthèse de protéines virales et régule potentiellement la synthèse d'ARN viral (Garaigorta & Ortín, 2007; Hale et al., 2008).

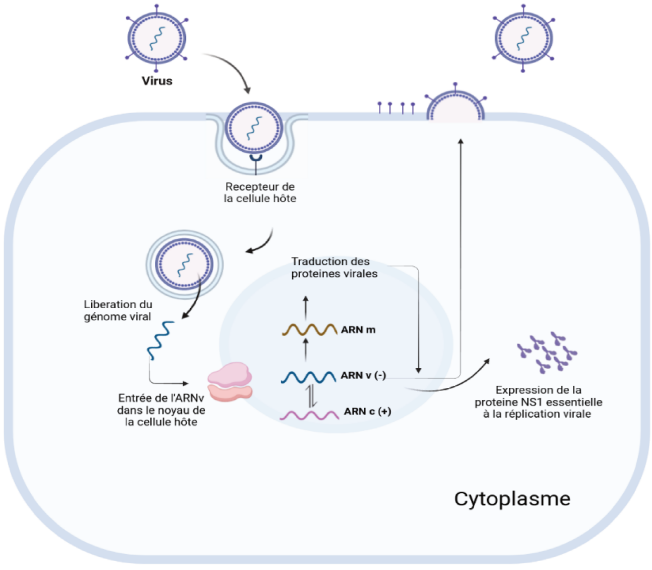


Figure 2: Cycle de réplication du virus de la grippe A dans les cellules hôtes. Au début, le virus se lie aux récepteurs de surface de la cellule hôte par adsorption, puis pénètre dans la cellule en fusionnant avec la membrane cellulaire et libère son contenu. Ensuite, l'ARN se réplique en utilisant les ressources de la cellule hôte, tandis que la protéine NS1 est synthétisée à partir de cet ARN grâce

aux ribosomes de la cellule hôte. Les différentes parties du virus de la grippe A sont assemblées dans la cellule hôte lors de l'étape d'assemblage, après quoi les nouveaux virus quittent la cellule hôte en bourgeonnant à travers la membrane cellulaire et en emportant une partie de celle-ci avec eux pendant l'étape de libération.

La protéine NS1 est un homodimère composé de deux domaines: Le domaine de liaison à l'ARN, noté RNA-BD qui se lie à l'ARN via la région du sillon définie par les deux hélices α2 et α2' situées à la surface du domaine, joue un rôle crucial dans la liaison au virus, la réplication virale et dans la déficience immunitaire; le domaine effecteur (ED) connu pour interagir avec de nombreux autres partenaires protéiques. Chaque monomère ED est relié au RNA-BD (dimère obligatoire) par une région non structurée hautement flexible d'une dizaine d'acides aminés appelée "*linker*" (Qian et al., 1994). Dans certaines structures cristallographiques de NS1, le linker est affecté par la délétion de cinq acides aminés ("Linker court") (Figure 3).

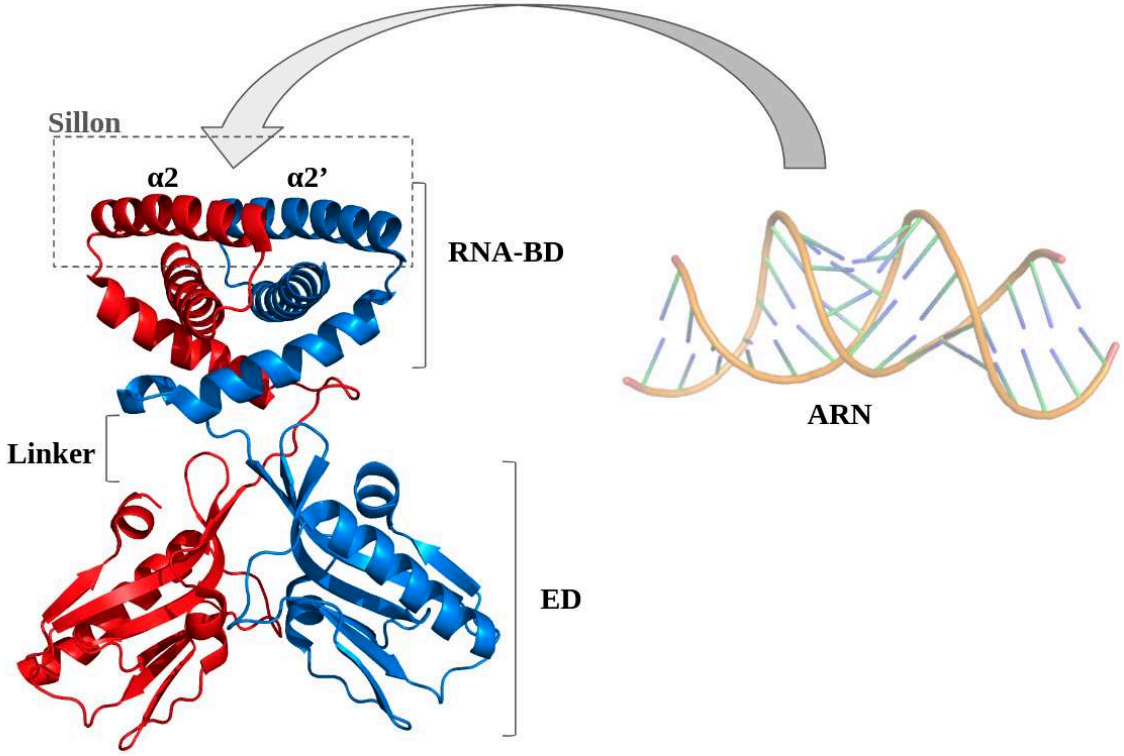


Figure 3: Représentation structurale du sous-type H6N6 (code PDB: 4OPA) de la protéine NS1. NS1 est un homodimère composé de deux chaînes colorées en bleu et rouge, elle est composée du domaine de liaison à l'ARN (RNA-BD) en dimère obligatoire ainsi que du domaine effecteur (ED) en deux monomères indépendants. Les deux domaines de la protéine NS1 sont liés par la région linker composée d'une dizaine d'acides aminés non structurés. L'ARN viral est connu pour interagir au niveau de la région du sillon, encadrée en pointillées gris, définie par les deux hélices α2 et α2'.

Les différentes structures cristallographiques disponibles dans la PDB illustrent un mécanisme d'ouverture et de fermeture des monomères ED, selon leur positions relatives au dimère RNA-BD. NS1 existe sous plusieurs formes (fermée, semi-ouverte et ouverte) et se caractérise donc par un polymorphisme structural. Toutefois, il n'existe que quatre structures cristallographiques de NS1 dans la PDB: deux structures du sous-type H6N6 de forme fermée et semi-ouverte, une structure du sous-type H1N1 de forme semi-ouverte extraite du

complexe avec son partenaire TRIM25 et une structure du sous-type H5N1 de forme ouverte (Carrillo et al., 2014; Koliopoulos et al., 2018; Bornholdt & Nagabhushana, 2008). Les séquences des différents sous-types se distinguent les unes des autres par des mutations pouvant affecter la structure et les fonctions de la protéine.

3.1.2. Structures disponibles et structures modélisées par homologie

Le nombre de structures complètes (“full-length”) de la protéine NS1 disponibles dans la PDB est limité. Il a été nécessaire de reconstruire les dimères, en appliquant une opération de symétrie sur la structure cristallographique, lorsque seul le monomère était disponible. Puis, une modélisation par homologie a été réalisée, pour chacun des sous-types, pour obtenir les deux formes complémentaires à celle déjà disponible dans la PDB. L'objectif de l'étude était de mener une analyse complète sur les trois sous-types de la protéine NS1 dans leurs trois formes possibles (fermée, semi-ouverte et ouverte). De plus, NS1 présente une variabilité de longueur de linker : court ou long, selon le sous-type.

Au total 18 structures 3D de la protéine NS1 ont été générées (Figure 4), le schéma ci-dessous représente les structures cristallographiques, leur sous-type d’origine, la longueur des linkers associés ainsi que les modèles complémentaires de formes construits par homologie pour chacun des sous-types. Dans la suite de ce travail, seuls les sous-types dont la structure cristallographique était connue ainsi que les modèles par homologie de leurs formes complémentaires ont été retenus.

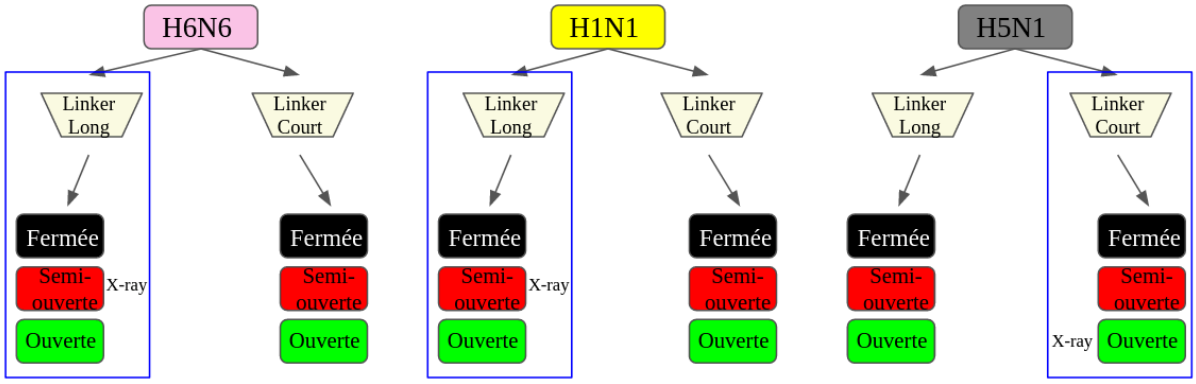


Figure 4: Structures de la protéine NS1 utilisées dans le cadre de l'étude de son polymorphisme. Quatre structures cristallographiques sont disponibles sur la PDB : H6N6 et H1N1 à linker long et de formes semi-ouverte ainsi que H6N6 et H5N1 à linker court et de forme fermée et ouverte respectivement. Les autres structures, qui présentent des formes complémentaires, ont été obtenues grâce à la modélisation par homologie. Les structures encadrées en bleu sont celles qui ont été retenues pour la suite de cette étude.

3.1.3 Flexibilité de la protéine NS1

Pour étudier la flexibilité de la protéine NS1 dans différents sous-types et formes, des simulations de dynamique moléculaire de 150ns ont été effectuées sur neuf structures initiales sélectionnées pour l'étude. L’ampleur des mouvements de la protéine au cours des trajectoires de DM a été estimée par des calculs de RMSD et de RMSF pour observer et quantifier les mouvements de l’ensemble des systèmes globalement sur toute la protéine et localement au niveau du RNA-BD.

Le polymorphisme structural de la protéine NS1 est caractérisé par des mouvements

d'ouverture et de fermeture des monomères ED par rapport au dimère RNA-BD, via la région hautement flexible du linker. Pour évaluer les mouvements de la protéine au cours des simulations sur les neuf structures sélectionnées dans cette étude, les distances entre les centres de masse des monomères ED et entre chaque monomère ED et le dimère RNA-BD ont été mesurées tout au long des trajectoires. Ces mesures ont permis de donner un aperçu quantitatif précis sur le mouvement de chaque région de la protéine NS1.

3.1.3.1 Comportement du domaine ED dans les différents sous-types:

Au cours de l'analyse des trajectoires de DM des variations comportementales observées au niveau du domaine ED de la protéine NS1 de divers sous-types du virus de la grippe. La forme fermée de la protéine est la plus stable, tandis que les formes ouvertes présentent de légers mouvements des domaines ED, avec une torsion plus prononcée de l'un des monomères, surtout lorsque le linker est court. La forme semi-ouverte a subi les changements conformationnels les plus importants, allant d'une ouverture maximale à une fermeture presque totale, probablement dus à une relaxation de la protéine NS1. Les structures semi-ouvertes de sous-types H1N1 et H5N1 ont subi des mouvements considérables, contrairement à la structure semi-ouverte de sous-type H6N6, qui est restée relativement stable dans une conformation tendant vers l'état ouvert.

Les résultats suggèrent que la structure de la protéine NS1 est sensible aux variations de la séquence des sous-types du virus de la grippe , ce qui peut entraîner des changements conformationnels, qui à leur tour, peuvent impacter la fonction de la protéine.

3.1.3.2 Stabilité du domaine de liaison à l'ARN (RNA-BD)

Contrairement au domaine ED qui présente des variations conformationnelles importantes en fonction de la forme initiale de la protéine et de la longueur du linker, le RNA-BD a montré une grande stabilité. Les courbes de RMSD ont confirmé cette stabilité, indiquant que la conformation de ce domaine reste similaire tout au long de la simulation. Dans une étude précédente, le RNA-BD de la protéine NS1 du sous-type H6N6 de forme fermée avait été étudié (Figure 5). Les résultats avaient montré la grande stabilité de ce domaine au cours des simulations de DM et avaient permis d'identifier grâce aux descripteurs de PockDrug une poche fréquemment druggable localisée dans le sillon susceptible d'accueillir un médicament capable d'inhiber l'interaction entre NS1 et l'ARN (Hussein et al., 2020). Cependant, les formes fermées ont été observées comme étant les plus stables lors de notre analyse des DM des différents sous-types et formes. Les résultats indiquent que la présence de cette poche druggable doit être confirmée sur toutes les formes et sous-types. Pour éviter de dépendre d'un sous-type spécifique lors du développement d'une thérapie, il s'est avéré nécessaire d'étudier la flexibilité des trois sous-types dans leurs trois formes possibles.

3.1.4. Conclusion

Dans cette étude, neuf conformations de la protéine NS1 représentant différents sous-types et formes ont été analysées. Pour observer la stabilité et les propriétés intrinsèques de la protéine en fonction de son sous-type, une combinaison d'approches de modélisation par homologie et de simulations de DM a été utilisée. Les simulations ont été effectuées en imposant la structure initiale de la protéine NS1 dans l'une des trois formes structurales possibles: fermée, semi-ouverte ou ouverte.

Les résultats ont montré que les trois formes structurales étudiées sont possibles avec des différences de stabilité et de comportement. Les formes fermées sont très stables, les formes ouvertes sont globalement stables mais présentent une asymétrie de comportement, et la forme semi-ouverte est la plus instable et a plusieurs états intermédiaires selon l'état semi-ouvert initial. Toutefois, cette instabilité du domaine ED n'affecte en rien la stabilité du RNA-BD dans les différents sous-types et formes, ce qui fait de NS1 une bonne cible thérapeutique. Des sites de liaison potentiels ont été identifiés, lors d'une précédente étude, à l'interface de la structure de H6N6 via des simulations de DM **(Hussein et al., 2020)**, ce qui suggère que le fait de cibler le RNA-BD pourrait être une stratégie efficace pour développer un traitement indépendant du sous-type à condition d'étendre l'étude aux différents sous-types et formes. La prochaine étape consiste à identifier les poches communes à tous les sous-types pour confirmer le potentiel thérapeutique de la protéine NS1.

International Journal of
Molecular Sciences

MDPI

Article

Influenza A Virus NS1 Protein Structural Flexibility Analysis According to Its Structural Polymorphism Using Computational Approaches

Sarah Naceri [1], Daniel Marc [2,3], Anne-Claude Camproux [1,†] and Delphine Flatters [1,*,†]

[1] Université de Paris, CNRS, INSERM, Unité de Biologie Fonctionnelle et Adaptative, 75013 Paris, France; sarah.naceri@etu.u-paris.fr (S.N.); anne-claude.camproux@u-paris.fr (A.-C.C.)
[2] Equipe 3IMo, UMR1282 Infectiologie et Santé Publique, INRAE, 37380 Nouzilly, France; daniel.marc@inrae.fr
[3] UMR1282, Infectiologie et Santé Publique, Université de Tours, 37000 Tours, France
* Correspondence: delphine.flatters@u-paris.fr
† These authors contributed equally to this work.

Abstract: Influenza A viruses are highly contagious RNA viruses that cause respiratory tract infections in humans and animals. Their non-structural protein NS1, a homodimer of two 230-residue chains, is the main viral factor in counteracting the antiviral defenses of the host cell. Its RNA-binding domain is an obligate dimer that is connected to each of the two effector domains by a highly flexible unstructured linker region of ten amino acids. The flexibility of NS1 is a key property that allows its effector domains and its RNA binding domain to interact with several protein partners or RNAs. The three-dimensional structures of full-length NS1 dimers revealed that the effector domains could adopt three distinct conformations as regards their mutual interactions and their orientation relative to the RNA binding domain (closed, semi-open and open). The origin of this structural polymorphism is currently being investigated and several hypotheses are proposed, among which one posits that it is a strain-specific property. In the present study, we explored through computational molecular modeling the dynamic and flexibility properties of NS1 from three important influenza virus A strains belonging to three distinct subtypes (H1N1, H6N6, H5N1), for which at least one conformation is available in the Protein Data Bank. In order to verify whether NS1 is stable in three forms for the three strains, we constructed homology models if the corresponding forms were not available in the Protein Data Bank. Molecular dynamics simulations were performed in order to predict the stability over time of the three distinct sequence variants of NS1, in each of their three distinct conformations. Our results favor the co-existence of three stable structural forms, regardless of the strain, but also suggest that the length of the linker, along with the presence of specific amino acids, modulate the dynamic properties and the flexibility of NS1.

Keywords: molecular dynamics; influenza a virus; non-structural protein 1 polymorphism

Citation: Naceri, S.; Marc, D.; Camproux, A.-C.; Flatters, D. Influenza A Virus NS1 Protein Structural Flexibility Analysis According to Its Structural Polymorphism Using Computational Approaches. *Int. J. Mol. Sci.* **2022**, *23*, 1805. https://doi.org/10.3390/ijms23031805

Academic Editor: Paulino Gómez-Puertas

Received: 31 December 2021
Accepted: 1 February 2022
Published: 4 February 2022

1. Introduction

Influenza A viruses (IAVs) are highly contagious RNA viruses that cause respiratory tract infections in humans and animals. Being responsible for epizootics in birds and several species of mammals, they can transmit to humans, either as sporadic cases of human infection with swine or avian viruses, or as pandemic viruses, which eventually will infect a large part of the human population. The two latest pandemic influenza A viruses (H3N2 in 1968 and H1N1 in 2009) have since been responsible for seasonal influenza for more than 50 years and twelve years, respectively. Annually, seasonal influenza causes three to five million severe cases and 290,000 to 650,000 deaths due to respiratory complications [1,2].

The eight-segment viral genome encodes more than 10 viral proteins, among which is the non-structural protein 1 (NS1), which is absent from the virion but highly expressed in the infected cells. NS1 plays an important role in countering the antiviral defenses of

the infected cell, thereby helping the virus to escape the innate immune system [3,4]. For this reason, novel therapeutic strategies have recently been explored that antagonize NS1's activities [5,6]. NS1 is known to bind non-specifically to double-stranded RNA (dsRNA) as well as to viral and cellular RNAs [7]. However, the effector domain is also proving to be an important therapeutic target given its multiple protein partners and its role in the structural polymorphism of NS1 [6,8,9].

NS1 is a homodimer of two 230-residue chains. It is comprised of two domains that are connected by an unstructured and highly flexible region of ten amino acids, which can be noted as a long or short linker (respectively referred as LL and SL in this study). The RNA-binding domain (RBD) is an obligate dimer involving residues 1–73 of each chain, arranged as three pairs of symmetrically positioned anti-parallel alpha helices. The two antiparallel helices 2 and 2′ form the RNA-binding interface, thanks to several basic amino acids including arginine 38 and lysine 41. The RBD binds several double-stranded RNAs (dsRNAs) as well as viral and cellular RNA, with no obvious sequence-specificity. The RBD has long been identified as an independent domain; it has been expressed as a recombinant protein and its structure has been established, both by crystallography and by nuclear magnetic resonance [10,11]. The different three-dimensional (3D) structures available in the Protein Data Bank (PDB) [12] indicate that this folding is well conserved across the diversity of influenza A virus strains. Residues 86–203 of each chain fold autonomously in an effector domain (ED), which is composed of three helices and seven beta-strands; the latter are organized into a broad sheet of antiparallel strands enveloping a long helix [13] and have been described as stable as a monomer or as a dimer. The different structures available on the ED dimer showed several dimerization interfaces, either via the helix 5 (residues 170–188) to form a helix–helix interface, or via a short strand (residues 88–91) to form a strand–strand interface [14]. According to the position of the two ED domains relative to the RBD dimer, three distinct conformations or forms (open, closed, semi-open) have been described in the literature [15–17]. Through this structural polymorphism, ED domains play an important role in the interaction with different protein partners [18]. These interactions are responsible for disrupting the defense function of the host cell [8,19].

The first full-length NS1 structure, published in 2008, was that from an H5N1 strain (pdb id: 3F5T) [20]. This structure corresponds to an open conformation, with each ED facing the RBD dimer through its short strand at position 88–91 while the surface of the long helix 5 is exposed to the solvent. Subsequently, two additional complete NS1 structures from two other strains, H6N6 and H1N1, were available in the PDB (pdb id: 4OPH in 2014 and 5NT2 in 2018, respectively) [15,21]. These two structures describe an intermediate form where both sides of the ED are more or less exposed and correspond to semi-open conformations. These 3D structures highlight a structural polymorphism and emphasize its conformational plasticity. The hypothesis of the co-existence of these three forms for the NS1 protein would explain the numerous possibilities of interactions with its partners [22]. However, each of the three full-length NS1s that was experimentally crystallized yielded a unique structure, giving no experimental support to the putative flexibility of NS1 in vivo. This lack of data leads to several hypotheses to explain this structural polymorphism including dependence on strain, the length of the RBD-ED linker or specific NS1 sequence residues. Additionally, the possibility that NS1 adopts these three forms regardless of the strain but conditional to physiological conditions and localization (nuclear or cytoplasmic) in the cell was also proposed [14].

In this study, we focused on the structural polymorphism of NS1 for three strains with full length structures available in the PDB. These are the PR8 strain, of the H1N1 subtype, and the H6N6 and H5N1 strains, which are avian strains. Structurally, the H5N1 protein is available with a short linker in an open form [20]. The H1N1 protein is characterized by a long linker and a semi-open form. This protein was resolved as a complex with a coiled coil domain of the TRIM25 protein and in the presence of a mutation at positions 187. In this complex, only the ED domains interact with TRIM25 (via sheets face). The authors show that the refolding of each ED is only mildly affected by the interaction compared

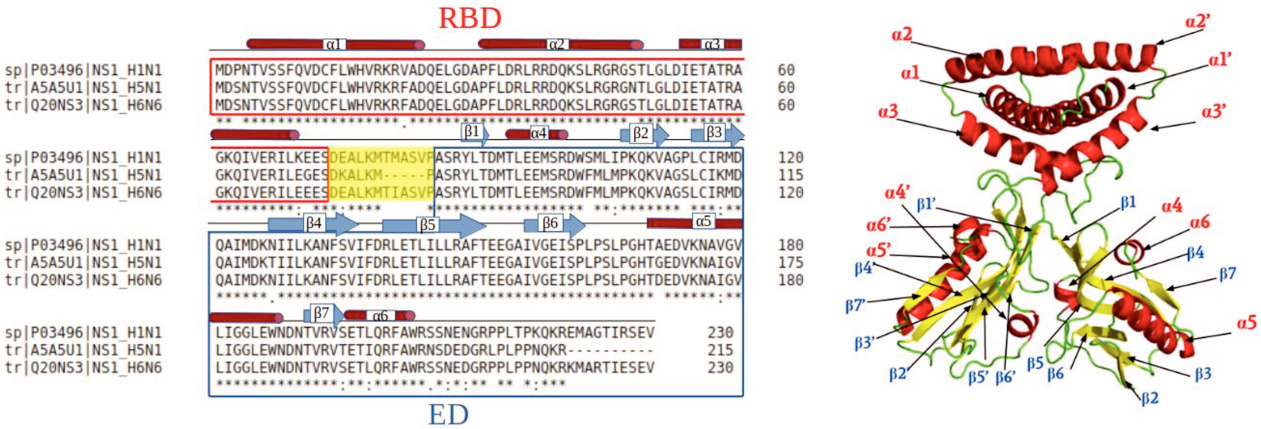
to the free-form ED. However, the authors showed that the relative position of the two ED domains is conditioned by the presence of TRIM25 in a semi-open form [21,23]. The structure of the NS1 protein of H6N6 is characterized by a long linker and a semi-open form without any partner or mutation. In this computational study, in order to verify whether NS1 is stable in all three forms and for the three strains, we constructed one model by homology for each of the two complementary forms not known experimentally by crystallography for each strain studied. In a second step, molecular dynamics simulations were performed for three forms obtained on the three strains to study the stability of the structures relative to the form and the strain. The structural and dynamics study of NS1 and the characterization of RBD-ED movements allowed us to better understand the dynamic properties of NS1 in its different forms for different strains in order to develop a strain-independent therapy.

2. Results

2.1. NS1 Protein Properties Depending on Three Strains

2.1.1. Strain Specific Properties of NS1

The multiple alignment of the NS1 protein sequences of three H1N1, H6N6 and H5N1 strains (Uniprot identifiers P03496, Q20NS3 and A5A5U1 respectively), which all belong to the A allele of NS1, emphasizes the high degree of conservation (Figure 1a). The RBD region is particularly well conserved especially at the level of the α2 helix, which forms the RNA-interaction surface (Figure 1b). Only a few differences are observed, especially for residues located in a loop or at the end of a helix (3, 22, 48, 70, 71, 75). The H5N1 NS1 differs by its characteristic deletion of residues 80–84 in the linker region (short linker, SL). Some variations are also observed in the ED: residues 103, 106 and 114 in connecting loops, residues 118, 127, 170, 178, 195 and 198 in secondary structures, as well as residues 205, 207, 214, 216, 218, 221 in the C-terminal region (residue numbering is that of the H1N1 sequence). The C-terminal extension is not considered in this study because of its disordered behavior.



(**a**) (**b**)

Figure 1. Multiple sequence alignment (Clustal Ω) of the three strains (H1N1, H5N1, H6N6) and secondary structure of the NS1 protein. (**a**) The residues boxed in red correspond to the RBD dimer (composed mainly of helices), the residues boxed in blue correspond to the ED domain (composed of α helices and β sheets) and the twelve residues boxed in yellow compose the linker region, which connects the RBD and ED domains. In the H5N1 sequence, the linker region is shorter due to five missing residues at positions (80–84). (**b**) Illustration in cartoon of the secondary structure (α helices in red, β strands in yellow and loop in green) of the NS1 protein (Pymol), pdb code: 4OPA.

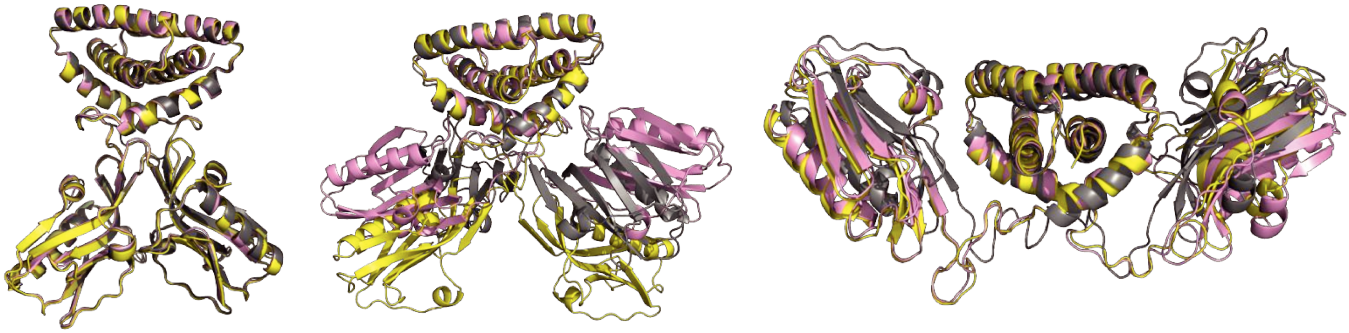
2.1.2. NS1 Structural Properties According to the Three Forms

The crystal structures of the H1N1, H6N6 and H5N1 variants (PDB codes 5NT2, 4OPH and 3F5T respectively) are representative of the semi-open (5NT2, 4OPH) and open (3F5T) conformations, while the closed conformation is represented by the structure (PDB 4OPA)

of another variant of the H6N6 sequence, which is characterized by the engineered deletion of residues 80–84, along with the presence of a glutamic acid at position 71 [15,21].

In addition to its crystal structure corresponding to a given conformation, each sequence was fitted, through a homology modelling (HM) approach, into the crystal structure representing the two remaining conformations, resulting in a set of three conformations for each sequence. Thus for the H1N1 and H6N6 variants, one closed and one open conformation were generated by homology, using the corresponding crystal structures of the H6N6 variant in the closed (pdb id: 4OPA) or open (pdb id: 6OQE) conformation as templates. Of note, the two latter structures are those of the H6N6 variant harboring the 80–84 deletion, a deletion that we reversed when building the HM models. The closed and semi-open conformations of the H5N1 variant were modeled using as templates the corresponding crystal structures of the H6N6 (pdb id: 4OPA) and H1N1 (pdb id: 5NT2) variants, respectively. The three crystal structures were named XR ($H1N1^{XR}$, $H5N1^{XR}$, $H6N6^{XR}$), as opposed to HM, for the six homology models (for more details, see Section 4).

In Figure 2, for each conformation, we compared the three structures through alignment and superimposition using the Pymol software.



(**a**) (**b**) (**c**)

Figure 2. Illustration of the superposition of the structures of the H1N1 strains in yellow, H5N1 in grey and H6N6 in pink with Pymol. (**a**) Superposition of the closed forms of the three strains ($H1N1^{HM}$, $H5N1^{HM}$, $H6N6^{HM}$); (**b**) superposition of the semi-open form structures of the three strains $H1N1^{XR}$, $H5N1^{HM}$, $H6N6^{XR}$); (**c**) superposition of the open form structures ($H1N1^{HM}$, $H5N1^{XR}$, $H6N6^{HM}$).

The NS1 structures show a faithful superimposition in the closed ($H1N1^{HM}$, $H5N1^{HM}$, $H6N6^{HM}$) and open forms ($H1N1^{HM}$, $H5N1^{XR}$, $H6N6^{HM}$) (Figure 2a,c), regardless of the linker length. In the semi-open form, three strain conformations ($H1N1^{XR}$, $H5N1^{HM}$ and $H6N6^{XR}$) show a different orientation of the ED with respect to the RBD, specifically for H1N1 (Figure 2b). These variable conformations may suggest that the semi-open form is an intermediate state between the closed and open forms. The $H6N6^{XR}$ conformation is in semi-open form with an open tendency and the $H1N1^{XR}$ conformation is in semi-open form with a closed tendency just like the $H5N1^{HM}$ conformation, which was constructed by homology from the $H1N1^{XR}$ structure as a template. The structures of the same three forms overlap very well at the level of the dimeric RBD domain. Depending on the form, the main structural differences are observed in the position of each ED domain relative to the RBD domain or relative to each other.

In order to characterize the orientation of these ED monomers for the three forms and three strains, we calculated three distances (see Section 4). The distances of the geometric center (see method) for each initial static structure between the two monomeric ED (d1) domains and then between each of the monomers (chains A and B) of the ED domain and the RBD dimer (d2 and d3) (Figure 3 and Table S1) were calculated. These initial static structures correspond to time 0 nanosecond (t = 0 ns) and were subsequently used as the initial structure to run 150 ns molecular dynamics simulations. The results are shown in Figure 3 and detailed in Table S1.

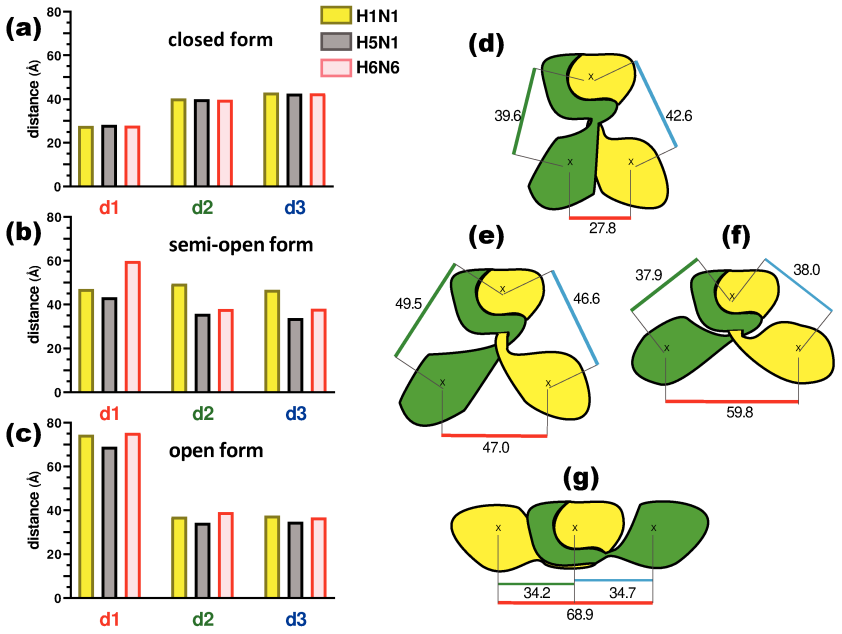


Figure 3. On the left are histograms of the three distances (d1, d2, d3) for the different strains (H1N1 in yellow, H5N1 in grey and H6N6 in pink): (**a**) in the closed form, (**b**) in the semi-open form and (**c**) in the open form. On the right, schematic representation of the dimeric NS1 protein (each chain is respectively in green and yellow) with the values of the distances d1, d2 and d3 indicated in Angstroms (underlined in red, green and blue respectively): (**d**) closed form, (**e**) semi-open form for H1N1, (**f**) semi-open form for H6N6 and (**g**) open form.

The distance (d1) between the two ED monomers A and B in the closed conformation is consistently close to 28 Å, while it reaches its maximal values (69–75 Å) in the open conformation (Figure 3a,c, Table S1). This distance is more variable in the semi-open form depending on the strain (Figure 3b). The H5N1 strain shows the smallest d1 in the semi-open form, which can be explained by the fact that it is a model built on a semi-open template with a closed tendency (H1N1 template) but also with a short linker. This brings the ED monomers closer together with a distance of 43.3 Å. The H1N1 and H6N6 strains in the semi-open form with a closed and open tendency result in a d1 distance of 47 Å and 59.8 Å respectively (Figure 3e,f). These differences in distance for the same form can be related to the crystallographic structure of the two strains, particularly H1N1, which was crystallized in a complex with TRIM25 via the ED domain, which was constrained by these interactions. Despite the variations in the d1 distance observed in particular for the semi-open forms, we note that d1 characterizes well the three possible structural forms for the NS1 protein (Figure 3d–g). The distances d2 and d3 are more or less symmetrical regardless of the strain or form, being close to 40 Å in the closed form and between 34 and 39 Å in the open form. These distances show more variations in the case of the semi-open form with the largest values for $H1N1^{XR}$ (46.6 Å and 49.5 Å) and shortest values for $H5N1^{HM}$ (35.6 Å and 33.7 Å). The H5N1 form is the unique form with a short linker compared to H1N1 or H6N6 forms.

These results observed on the initial structures show characteristics that are specific to each crystallographic structure associated with a structural form of NS1. These characteristics are also found in the model structures depending on the template structure used. Moreover, these homology models are simulated in different forms that are not yet experimentally solved (by crystallography or NMR). Molecular dynamics simulations allow us to verify the stability of such models. In order to get rid of the properties of the template structures and to highlight the intrinsic structural properties linked to each strain, we performed molecular dynamics simulations on these 9 initial structures (i.e., the 3 crystal structures and the 6 models built by homology).

2.2. Dynamic Properties of NS1 Structures

Each initial NS1 structure (t = 0) (in three forms crossed with the three strains) were submitted as a starting structure to molecular dynamics (MD) simulations. This MD approach allows one to explore the conformational space of each structure and to study the structural and dynamic properties, intrinsic to the protein sequence. From simulations of 150 ns duration, 150,000 snapshots (one conformation per picosecond) were extracted and analyzed (Figure 4).

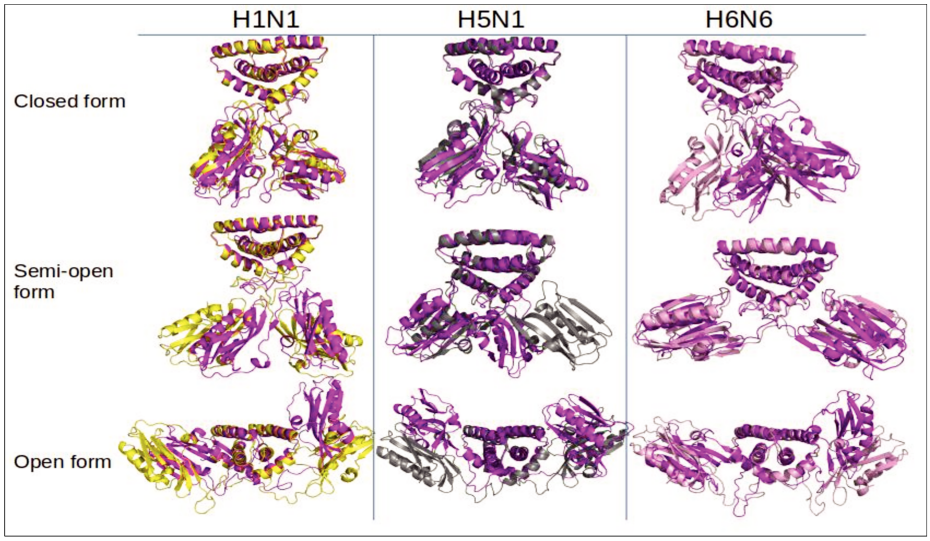


Figure 4. Superimposition of structures according to the three strains represented in the cartoon with Pymol. For each of the nine structures, the conformation at the end of the simulation (t = 150 ns) was extracted and superimposed on the initial structure (t = 0 ns). The initial H1N1, H5N1 and H6N6 conformations are colored respectively in yellow, grey and pink. The final conformations are shown in magenta.

2.2.1. Stability of NS1 Structures from the Three Strains in Different Forms

The Cα Root Mean Square Deviation (RMSD) curves were calculated on the 150,000 snapshots of each of the nine initial structures (Figures 5, S1–S4). The RMSD values (average and standard deviation) of each monomer are represented in Table 1 and the corresponding histograms in Figure S5.

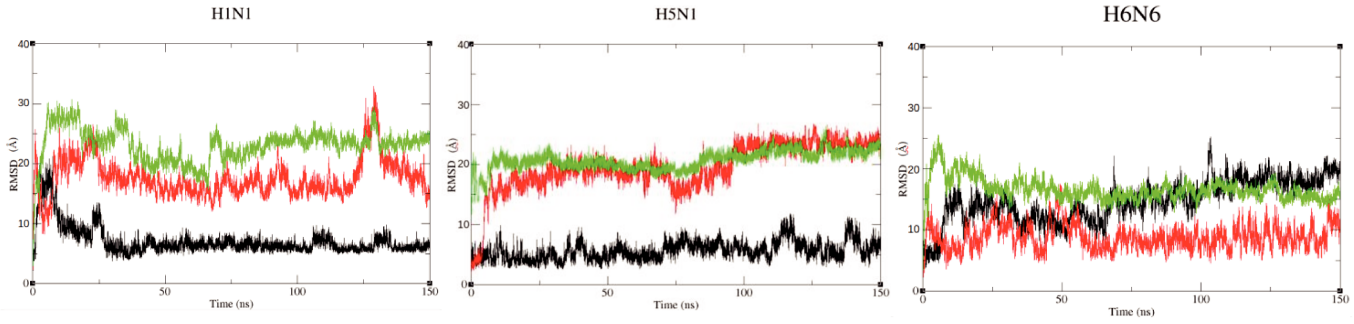


Figure 5. Cα RMSD calculated on the ED domain fitting to the RBD for the three forms (closed, semi-open, open) respectively colored in black, red and green.

This shows that the structures are very stable in the closed form regardless of the strain, especially for H1N1^{HM} and H5N1^{HM} with average RMSD values ranging from 2 to 5 Å (Table 1). For the open forms, the RMSD curves still show a plateau but with higher values ranging from 9 to 12 Å for the three strains. These results show that the protein undergoes greater conformational changes from its initial structure at t = 0 ns for the open

form (Figure 4). It is confirmed that the semi-open form is more variable depending on the strain. The RMSD are around 10 Å for H1N1^{XR} and H5N1^{HM}. These two structures move away from the starting conformation over time and RMSDs reach values similar to those of the open form. On the other hand, in H6N6^{XR}, this semi-open form seems to be quite stable throughout the trajectory with an RMSD average value not exceeding 5 Å (Table 1, Figure S5). Of note, the initial semi-open structure of H1N1 and H5N1 was derived from a crystallographic structure complexed with TRIM25 and molecular dynamics are performed on the non-complexed NS1 protein. During the simulation, the initial structure moves away from its complexed conformation and shows the characteristics of the non-complexed semi-open form.

Table 1. Average Cα RMSD and distances for each strain (H1N1, H5N1, H6N6) in the three forms (closed, semi-open, open) during the trajectory. RMSD_All is the average RMSD calculated over the whole protein, RMSD_RBD is the RMSD of the RBD dimer after fitting to the RBD, RMSD ED_A and ED_B are the RMSD of each monomer A and B of the ED domain calculated independently and RMSD ED fit on RBD corresponds to the RMSD of the EDs relative to the reference frame attached to the RBD. The distances (d1, d2, d3) correspond respectively to the average distances over time between the two ED monomers, between ED chain A and RBD and between ED chain B and RBD. The average values are indicated with their standard deviation.

Strains/Forms	**Closed Form**			**Semi-Open Form**			**Open Form**		
Strains	**H1N1^{HM}**	**H5N1^{HM}**	**H6N6^{HM}**	**H1N1^{XR}**	**H5N1^{HM}**	**H6N6^{XR}**	**H1N1^{HM}**	**H5N1^{XR}**	**H6N6^{HM}**
RMSD Values (in Å)									
RMSD All	3.5 ± 0.5	2.3 ± 0.4	5.0 ± 1.1	10.2 ± 1.0	8.6 ± 1.4	4.2 ± 0.7	10.3 ± 1.6	11.9 ± 2.1	8.9 ± 0.7.9
RMSD RBD	1.9 ± 0.1	1.9 ± 0.1	2.4 ± 0.2	3.3 ± 0.1	2.3 ± 0.1	2.2 ± 0.2	3.1 ± 0.01	3.0 ± 0.1	2.7 ± 0.1
RMSD ED_A	1.5 ± 0.1	1.2 ± 0.1	1.1 ± 0.1	1.2 ± 1	1.6 ± 0.2	1.3 ± 0.2	1.8 ± 0.2	2.5 ± 0.2	2.1 ± 0.1
RMSD ED_B	1.6 ± 0.1	1.2 ± 0.2	1.3 ± 0.1	1.2 ± 0.1	1.1 ± 0.1	1.1 ± 0.1	1.4 ± 0.1	1.9 ± 0.2	2.1 ± 0.1
RMSD ED fit to RBD	7.1 ± 2.4	5.6 ± 1.5	14.7 ± 3.4	17.6 ± 3.1	19.5 ± 4	8.8 ± 2.0	23.1 ± 2.7	20.4 ± 2.9	16.6 ± 1.8
Distances values									
Distance d1	25.4 ± 0.7	28.8 ± 0.6	27.8 ± 0.8	32.4 ± 2.3	36.4 ± 1.5	60.0 ± 2.0	66.3 ± 3.4	66.5 ± 1.7	64.5 ± 1.7
Distance d2	42.4 ± 1.0	40.9 ± 0.5	42.7 ± 0.7	42.0 ± 2.0	41.1 ± 1.6	38.5 ± 0.6	34.3 ± 1.9	36.3 ± 1.0	34.7 ± 0.6
Distance d3	41.9 ± 0.7	42.6 ± 0.8	42.4 ± 1.0	44.9 ± 1.0	35.3 ± 0.4	38.3 ± 0.5	36.1 ± 1.3	34.5 ± 1.3	32.0 ± 1.0

In order to identify if these conformational changes, inducing high RMSD values, are mainly due to movements of the ED domain observed between the beginning and the end of the simulation or to intrinsic deformations of the ED domain fold or of the RBD dimer fold, we calculated the RMSD-RBD dimer (RMSD^{RBD}) and each ED domain separately.

The Cα RMSD of the RBD dimer conformations during the simulation compared to its initial structure (t = 0 ns) was calculated (Figure S1). The RMSD values vary between 1.9 and 3.3 Å for all nine trajectories. This result shows that the intrinsic fold of the RBD dimer is very stable during the trajectories regardless of the sequence variant and of the initial conformation (Table 1, line RMSD RBD, Figure S5).

Similarly, we calculated the Cα RMSD of each chain of the ED domain during the simulation relative to their conformation in the initial structure (Figure S3). The results show a very good stability of each monomer of the ED domain with RMSD average values not exceeding 2.5 Å (Table 1, lines RMSD ED_A and RMSD ED_B).

The relatively high RMSD values that we observed on the whole protein are therefore not due to a deformation of the 3D structure of these two ED chains or to a change in the RBD dimer fold (Figure S3). Rather, these mainly correspond to the motions of the two EDs relative to the RBD, as shown by the consistently high RMSD values reported in Table 1, line RMSD ED fit to RBD.

To quantify the movement of each monomeric ED allowed by its flexible linker region, we calculated the Cα RMSD of the ED domain (including the linker region) relative to

each other. We first verified that the presence or absence of the linker did not alter the RMSD curve by overlaying the results. The RMSD curves of the ED monomers with and without linker were superimposed, showing that the movement of the ED and the linker are not dissociated.

The H1N1^{HM} and H5N1^{HM} structures are very stable in the closed form, with RMSD values of 7.1 and 5.6, respectively. In contrast, the H6N6^{HM} strain has a RMSD of 14.7 Å. These high RMSD values are not due to the ED monomers moving away from each other (the closed form is well maintained) but to the linker moving, causing the dimeric ED to move horizontally away from its initial position (Figure 4). The semi-open form seems to be less stable compared to these two strains. It should be remembered that H1N1^{XR} (5NT2) is a semi-open form with a closed tendency and that H5N1^{HM} is a short-linker strain, which is not experimentally defined in this form; we have modelled it with an H1N1 type template (semi-open with closed tendency). By superimposing the structures at the beginning and at the end of the simulation, we can observe an approximation of the ED domains in the semi-open form. The movements of the two EDs are even more considerable in the open conformations, with average RMSD values (ED fit to RBD) of 23.1 and 20.4 Å for H1N1^{HM} and H5N1^{XR}, respectively (Table 1, line RMSD ED fit to RBD, and Figure 5, green curves). While all three structures consistently remain in the open conformation, these motions include rotations about several alpha carbons of the linker region (Figures 4 and 5).

The RMSD of each of the RBD and ED domains shows that the folding is very stable throughout the dynamics simulation (Table 1, Figures S1 and S3). This stability is confirmed on extended dynamics (340 ns, 298 ns, 180 ns) of three forms of H6N6 (closed, semi-open and open respectively) (Table S2, Figure S6).

The RMSD results suggest that the NS1 protein is able to adopt open and closed forms regardless of the strain or the length of the linker. The semi-open form seems to be particular because it depends on the strain and the intermediate states of its initial structure (semi-open with closed or open tendency). Depending on the tendency of its initial semi-open conformation, it converges towards a more open conformation (H6N6^{XR}) or more closed conformation (H1N1^{XR} or H5N1^{HM}).

2.2.2. Identification of Flexible Regions in the NS1 Structures

For each of the RBD and ED domains (including the linker region), we plotted the root mean square fluctuation of the alpha carbons (CαRMSF) in the A and B chains in order to assess the flexibility regions. Our analysis revealed a region of high flexibility around residue 30 on the RBD domain, with RMSF values that can be higher than 1 Å. This region corresponds to the region of the α1-α2 connecting loop and to the N-terminal part of the α2 helix (Figures 1b and S2). This flexibility is even more pronounced in open forms. In contrast, the most stable region is observed at the α1 helix, which is buried in the center of the RBD and constrained by its interactions with the α2, 2′ and α3, 3′ helices.

Beyond the linker region, which is known to be highly flexible, the ED domain also shows some flexible regions, particularly around residue 175, a flexible region corresponding to the N-terminal half of the α5 helix. This region is exposed to the solvent in both the open and the closed form (Figure S4).

2.3. Dynamic Motions of the ED Domains during Simulations

The RMSD results reflect the dynamic motions of the two EDs during the simulation. However, these movements of the EDs can be both related to opening or closing over time (translational motion) or reorientation (rotational motion). To support this result, the distances between the geometrical centers of the ED and RBD were computed during each of the nine simulations (Table 1) and compared to their initial values (Table S1). These distances were plotted in Figure 6.

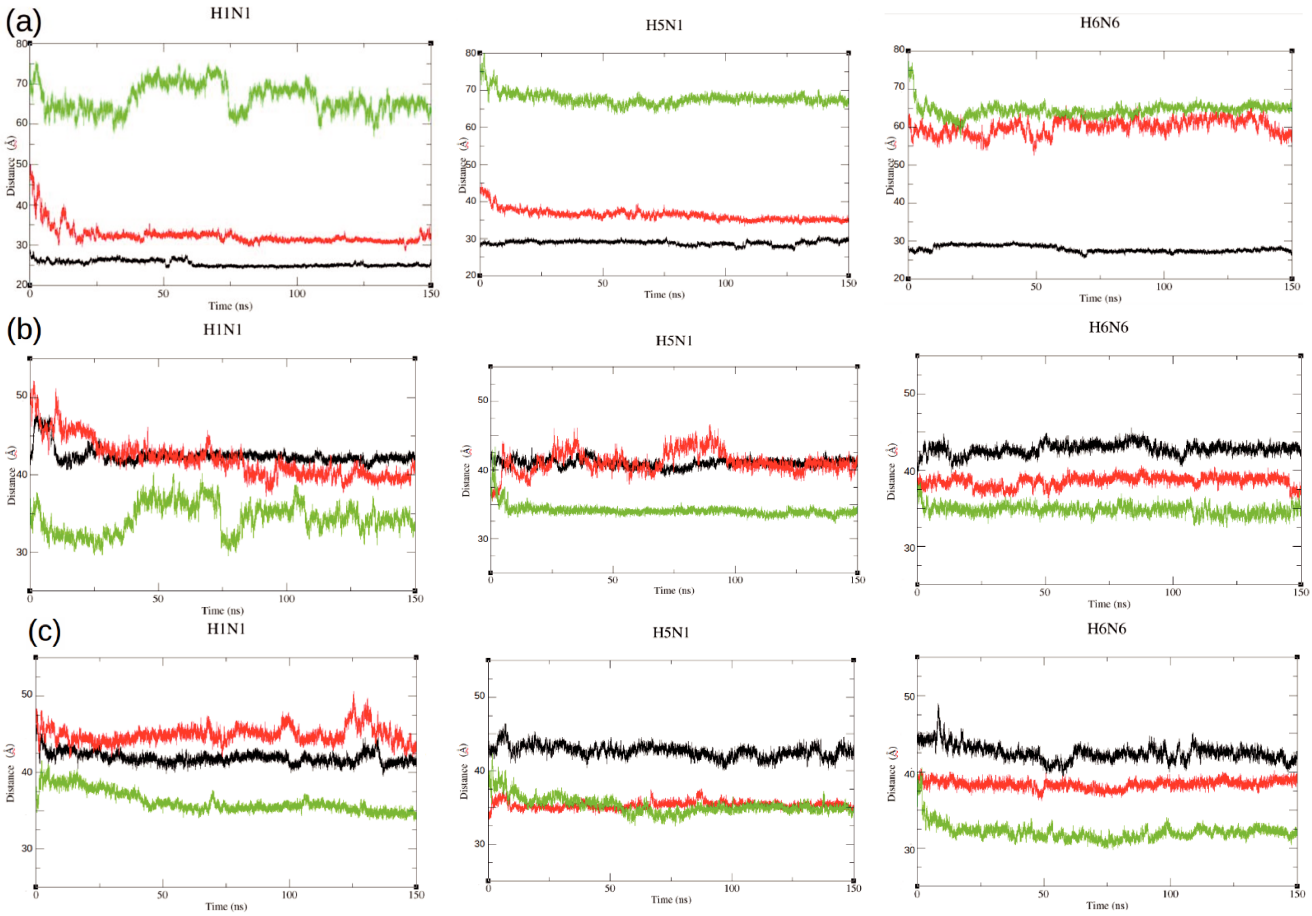


Figure 6. Average distances between geometric centers of ED domains for the three strains (H1N1, H5N1, H6N6) in the three forms (closed, semi-open and open), respectively, colored in black, red and green. (**a**) Distances (d1) between geometric centers of ED domains; (**b**) Distances (d2) between the geometric center of ED domain chain A and the geometric center of RBD dimer; (**c**) Distances (d3) between the geometric center of ED domain chain B and the geometric center of RBD dimer.

In the closed form, the d1 distance between the two ED domains is stable throughout the dynamics regardless of the strain, with the value ranging from 25 to 29 Å on average (Figure 6a, black curves). The EDs in the open form get closer, moving from 69–75 Å for the initial structure to 64–66 Å on average during the simulation (they are also closer to the RBD domain) (Figure 6a, green curves). Similarly for the semi-open forms, for the H1N1^{XR} and H5N1^{HM} structures (semi-open forms with a closed tendency), we observed a closing of the ED domains moving from 43–47 Å to 32–36 Å on average during the simulation and thus a closing, like an attraction, between the two domains (Figure 6a, red curves). For the H6N6^{XR} (semi-open form with an open tendency), the EDs maintain a constant distance throughout the trajectory. The distance value is around 60 Å on average and the EDs seem to be too far away to be able to approach each other and no force is exerted to promote this.

In addition to the distance d1 between the two ED monomers, we calculated the distance between each ED monomer and the RBD domain dimer during molecular dynamics (d2 and d3) (Figure 6b,c).

This analysis allowed us to identify an asymmetry in the movement of the ED domain monomers A and B in relation to the RBD, as in the case of the semi-open H1N1^{XR} structure where we could observe an approach of the ED monomers towards the RBD. The values d2 and d3 decreased from 49 and 46 Å (Table 1) during the molecular dynamics to an average distance of 42 and 45 Å, respectively. This asymmetry was also observed for the semi-open H5N1^{HM}, where a distance of ED chain A and RBD (d2) increased from 35 Å at t = 0

ns to an average of 41.1 Å during the trajectory. This difference was not observed in d3 (Figure 6c, Table 1).

3. Discussion

This dynamic in silico work is the first to be performed on full length structures of the NS1 protein. Our objective in this study was to characterize the structural polymorphism by exploring three different strains in terms of sequence and shape. Based on the results of this study, we showed that NS1 in the closed form exhibits high stability in molecular dynamics simulations with a stable ED–ED dimerization interface that involves the strands of the beta-sheets. These closed-form structures are all derived from a homology-based model constructed from the only complete closed-form structure of NS1 available by crystallography. This structure corresponds to a closed-form obtained from the H6N6 strain for which a deletion of linker residues 80–84 was made (H6N6 short linker closed form, pdb id 4OPA) [15]. The authors hypothesized that this closed form may be characteristic of short linker strains. Our results suggest that the closed form is also possible and stable for long linker NS1 structures, as in the case of H1N1 or H6N6 strains. The closed form is characterized by a distance of about 25–28 Å between the geometric centers of the ED domains and a distance around 42 Å between the EDs and the RBD dimer. The folding of each ED domain and RBD dimers are very stable as shown by the RMSD and RMSF values. The dimerization interface is essentially maintained by contacts between the short β1 strand (residues 88–91) and linker residues and between the α4 helix and the loop connecting strand β6 and α5 helix. The ED–ED dimerization interface is a strand–strand interface as described by Kerry and colleagues [14]. The presence of a long linker shows that the dimeric ED in closed form can be moved from its initial position without affecting the closed form (maintaining the d1 distance throughout the dynamics) as observed for the H6N6 strain.

Structures of NS1 in the open conformation were all obtained from crystals formed by NS1 variants harboring a short linker that was either characteristic of the strain (H5N1 isolates, structure 3F5T in pdb) or genetically engineered (H6N6, pdb id 6OQE). In our study, we constructed long linker open-form homology models for the H1N1 and H6N6 variant based on these SL crystallographic structures as templates. The three NS1 strains with either LL or SL also appear to be stable in open form. Indeed, the RMSD curves reached a plateau, showing that the conformation remains globally stable after the initial structure relaxes in an aqueous environment. Similarly the distance d1, which measures the distance between the two EDs, was maintained throughout the dynamics. Thus, the three structures remain stable in open form, even if each ED can drift away from its initial position during the simulation. In this open form, the face of the beta-sheet with the short β1 strand of chain A ED domain is oriented toward the α3 helix of chain B RBD domain (and vice versa). For the H5N1 strain, we could observe in the superposition of structures at the beginning and at the end of the simulation (Figure 4) that one of the EDs showed a slight translation with respect to its initial position but also a significantly more twisted folding of the beta-sandwich. This was also described in the work of Bornholdt, et al., 2008 [20]. The authors suggested that this is due to either a constraint imposed by the presence of the RBD dimer or to the length of the linker in H5N1 (SL). Indeed, supporting this view, the twisting of the β-sandwich is less pronounced in the case of H6N6 and H1N1, both variants harboring a long linker.

In both the open and closed forms, we could observe that the face containing the short strand β1 (residues 88–91) is consistently less exposed to the solvent because it is either at the ED–ED interface (closed form), or at the ED–RBD interface (open form). This allows the α5 helix of the ED to be completely exposed to the solvent, including W187 residue that is localized at the C-terminal extremity of the helix. Since this helix (residues 170–188) and its residue W187 play a central role in another mode of ED-dimerization [14,17,18], these conformations are both expected to be prone to a multimeric state of the NS1 protein in solution [22,24].

The semi-open form is comprised of several states from very open to almost closed. In this conformation, both interaction sides (short strand β1 side or α5 helix side) are solvent-exposed, allowing several modes of interactions with various protein partners. Thus, it is a semi-open conformation that was crystal-captured in the interaction of H1N1 NS1 with TRIM25 [21]. In our molecular dynamic simulations, this is the form that underwent the largest conformational changes, exploring various degrees of the opening states (from very open to almost closed). While the $H6N6^{XR}$ structure remained relatively stable in a conformation that tended toward the open state with a consistently high inter-ED distance, the EDs of the $H1N1^{XR}$ structure underwent large motions, and the same was true for $H5N1^{HM}$ that was built on the $H1N1^{RX}$ template, even in spite of its short linker region. We hypothesize that this large conformational change corresponds to a relaxation of NS1 in freeing itself from the constraint imposed by its interaction with TRIM25.

Finally, our results show that the dynamics of the RBD dimer are extremely stable regardless of sequence variations, as indicated by the RMSD values calculated on this domain. The RMSF show a region that fluctuates more around position 30 (Figure S2). These results are in agreement with the work of Abi Hussein et al., 2020 [25]. This region coincides with the loop that connects the α1 and α2 helices. This loop is exposed to the solvent and, in the open form, faces an ED domain. Each ED has a very stable fold in all nine simulations, with slight variations, among which is twisting in the open form of the H5N1, which has a short linker. The RMSF values reveal several regions that present wider fluctuations, such as in the α5 helix region or in the β4–β5 strands region of the ED domain. These regions were also highlighted from RMSF calculations on a molecular dynamic simulation of an ED dimer from H1N1 strain [26]. In this study, the structure of the ED dimer was resolved by NMR experimental technique.

This in silico work confirms the compatibility of the three forms for the three strains, in agreement with the current full-length crystallography data [14,18,21,24]. Approaches such as NMR on full-length structures will allow in vitro confirmation of the NS1′s flexibility.

4. Materials and Methods

The three selected strains were H1N1, H5N1 and H6N6. The H1N1 strain was available in the PDB (Protein Data Bank) in full length and homodimeric structure in semi-open form with a long linker (pdb code: 5NT2) [21], H6N6 in full length and monomer structure in semi open form with long linker (pdb code: 4OPH) [13], and H5N1 structure in open form with short linker (del 80–84) (pdb code: 3F5T) [20]. The NS1 Protein sequences were extracted from the UniProt database and compared by a multiple alignment algorithm using Clustal Ω [27].

4.1. Preparation of the Crystal Structures

The structures of NS1 from H5N1 and from H6N6 recovered in the PDB were completed in term of the missing residues and dimers were rebuilt with the Protein Interfaces, Surfaces and Assemblies tool (PDBePISA v1.52, 2014) (https://www.ebi.ac.uk/pdbe/pisa/ accessed on 30 August 2021).

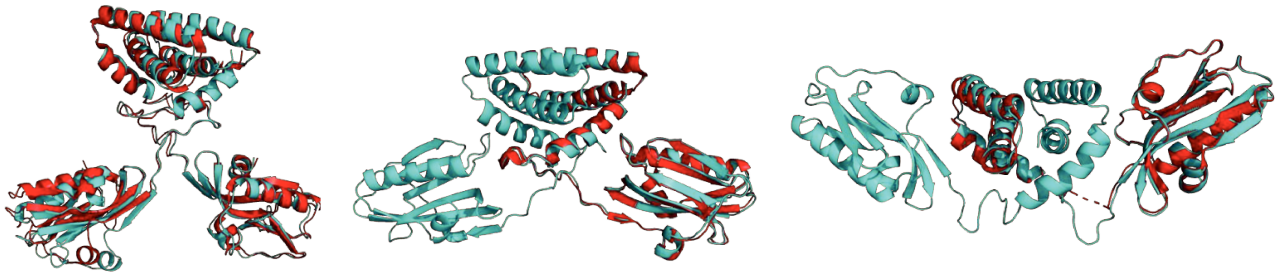
The structure of 5NT2 was already complete and in dimer form, we just dissociated it from its complex with TRIM25. The R38A/K41A engineered mutations for the three strains and also the W187A mutation specific to 5NT2 were reversed. The overlaid before and after reconstruction structures are illustrated with Pymol software (Figure 7).

4.2. Building Structural Models by Homology

For each strain, we built models by homology for the two complementary forms not known experimentally by crystallography, in order to study the behavior of the protein in the three strains and the three forms (closed, semi-open, open).

The closed form was modelled for the three strains on the 4OPA Xray template, which is a H6N6 structure where a deletion in the linker was induced (del 80–84). The open form was modeled for H6N6 and H1N1 strains on the 6OQE Xray template, which is also

a H6N6 structure with a short linker. For these two forms, the missing residues in the linker were reversed in H6N6 and H1N1 models. The semi-open form was modeled for the H5N1 strain on the 5NT2 Xray template, reversing the deleted residues in the linker. As for crystal structures, the R38A/K41A mutations and also the W187A mutation specific to 5NT2 were reversed in all the models.



(**a**) (**b**) (**c**)

Figure 7. Alignment of structures before and after reconstruction with the PISA tool. The red representations are the crystallographic structures before reconstruction and the cyan representations are the reconstructed structures: (**a**) structure of the NS1 protein of the H1N1 strain in the semi-open form with closed tendency (5NT2); (**b**) structure of the NS1 protein of the H6N6 strain in the semi-open form with open tendency (4OPH); (**c**) structure of the NS1 protein of the H5N1 strain in the open form (3F5T).

Table 2 summarizes the homology modeling steps performed with the MODELLER tool (v 9.23) [28].

Table 2. The nine initial structures from crystallography experiment or homology modeling. For each structure, the sequence Uniprot identifier, the pdb id of the crystallographic structure or of the template structure used for homology modeling and the reversed mutations are listed.

Strains/Forms	Closed Form	Semi-Open Form	Open Form
H1N1	**H1N1^{HM}** Uniprot identifier: **P03496** Xray structure template: 4OPA Reverse mutations: R38A/K41A/W187A	**H1N1^{XR}** Uniprot identifier: **P03496** Xray structure: 5NT2 Reverse mutations: R38A/K41A/W187A	**H1N1^{HM}** Uniprot identifier: **P03496** Xray structure template: 6OQE Reverse mutations: R38A/K41A/W187A
H5N1	**H5N1^{HM}** Uniprot identifier: **A5A5U1** Xray structure template: 4OPA Reverse mutations: R38A/K41A	**H5N1^{HM}** Uniprot identifier: **A5A5U1** Xray structure template: 5NT2 Reverse mutations: R38A/K41A	**H5N1^{XR}** Uniprot identifier: **A5A5U1** Xray structure: 3F5T Reverse mutations: R38A/K41A
H6N6	**H6N6^{HM}** Uniprot identifier: **Q20NS3** Xray structure template: 4OPA Reverse mutations: R38A/K41A	**H6N6^{XR}** Uniprot identifier: **Q20NS3** Xray structure: 4OPH Reverse mutations: R38A/K41A	**H6N6^{HM}** Uniprot identifier: **Q20NS3** Xray structure template: 6OQE Reverse mutations: R38A/K41A

4.3. Molecular Dynamics (MD) Simulation

The resulting nine structures described in Table 2, were first processed with ProPka to assign protonation states at pH 7 and to construct the missing side chains, ensuring that the new atoms were not reconstructed too close to existing atoms. MD simulations were performed with Gromacs 2019.5 [29], using the Amber99SB force field [30] under periodic boundary conditions. All structures were simulated as immersed in a cubic water box of the TIP3P water molecule model. Non-bonded interactions were truncated in a cut-off distance of 14 Å (for the closed and semi-open forms) and 18 Å (for the open form) for the electrostatic twin-range cut-off and the Van der Waals cut-off. The energy of the system was minimized over 50,000 cycles, using the steepest descent algorithm for energy minimization. Then, counter-ions were added to neutralize the system.

Each MD simulation was preceded by a 1ns heating/equilibration period during which harmonic constraints were imposed on the atomic positions of the protein and counter-ions. Each simulation was performed at constant temperature (300 K) and pressure (1 atm), the isothermal-isobaric ensemble (NPT), coupling the system to a heat bath, using the Berendsen algorithm. The LINCS algorithm was applied to all bond lengths to constrain them, allowing for a 2 fs integration time step. 150 ns simulations were carried out for each of the nine structures (3 forms * 3 strains). Dynamics of three forms of H6N6 were extended to check stability. The MD trajectories were visualized using Visualized Molecular Dynamics (VMD 1.9.2) [31] and the figures of the different snapshots were made with Pymol software [32].

Using the GROMACS tools, several properties were analyzed throughout the simulations to validate their quality and stability, including Root Mean Square Deviation (RMSD) and Root Mean Square Fluctuation (RMSF) calculated on the atomic coordinates of the Cα atoms.

The Cα RMSD analyses allowed us to study the stability of the whole protein but also of the RBD domain as a first step, and then the behavior of the ED domain in different strains and forms. The (Cα RSMF) was analyzed to identify flexible residuals along the NS1 sequence.

4.4. Distance Calculations between Domains

In order to quantify the opening and closing movements of the ED domains, we decided to measure three distances between the geometric centers of the different domains (see Figure 8). We measured the distance between the geometric centers of the two monomers of the ED domain (composed of residues 86–203 for each chain); this distance was called d1. Then, we measured the distances between the geometric center of the RBD (composed of residues 1–72 for each chain A and B) and each geometric center of the ED monomers; these distances were respectively called d2 and d3. To perform these calculations, we used the "distance" module of the GROMACS software by creating indexes that include the atoms of each group (ED chain A, ED chain B, RBD dimer) and plotted the distance between the different groups during trajectory.

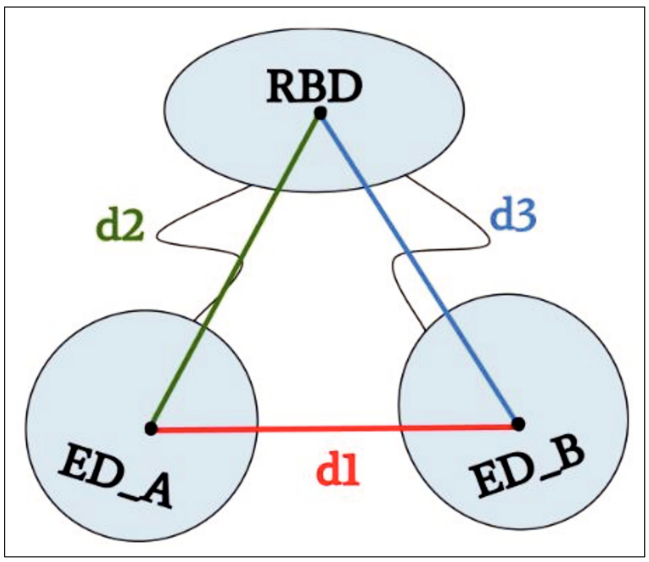


Figure 8. Distance diagrams (d1, d2, d3). The distance d1 is the distance between geometric centers of the ED monomers. The distances d2 and d3 are the distances between geometric centers of each monomer chain A and B of the ED domain and the RBD dimer.

5. Conclusions

In this paper, we studied the dynamic properties of nine structural conformations of the NS1 protein. These nine conformations are both representative of different viral strains (H1N1, H6N6 and H5N1) and of the structural polymorphism described in the literature. To carry out this study, we combined both homology modeling approaches to complete the structural forms not available in the Protein Data Bank and molecular dynamics simulations to observe the stability and intrinsic properties of the NS1 protein as a function of its strain when its initial structure is imposed in one of the three structural forms (closed, semi-open, open).

Our results are consistent with the hypothesis that all three forms would be possible regardless of strain. However, we note differences in stability depending on the shape and degree of openness of the initial conformation. Closed forms are very stable. In the open forms, an asymmetry in the behavior of the EDs can be observed (particularly in the H5N1 strain) despite the maintenance of the form in its open state. This would be caused by a translation of the ED domain with respect to the RBD dimer and by a deformation of the β-sheet, which could be more twisted when the structure contains a short linker. The semi-open form is more unstable and has several intermediate states. Depending on the initial semi-open state, the shape tends to close or remain stable with the same degree of openness as the initial structure.

Homology and computational molecular dynamics approaches allow us to characterize the movements of the ED-RBD during simulations and give us a better understanding of the dynamic properties of the NS1 protein structure in its different forms and strains. This allowed us to conclude that the forms adopted by NS1 are not strain-dependent as NS1 is able to be in all three states regardless of the strain. This information is crucial in the development of an effective therapeutic strategy on the different strains, for instance, to limit interactions at the level of the RBD–RNA interface in order to prevent viral RNA replication, but also at the level of the ED domain interacting with different protein partners depending on the orientation of the ED monomers and the interface exposed to the solvent.

From the perspective of targeting NS1 for novel antiviral strategies, a first step towards the identification of small compounds targeting NS1 RBD was proposed by Abi Hussein et al., 2020 [25]. They explored the druggability of RBD using molecular dynamics simulations of one H6N6 crystal structure. They confirmed the remarkable stability of the H6N6 RBD structure and were able to identify potential binding pockets in the groove delimited by the antiparallel α2-helices that make up its RNA-binding interface. They highlighted the druggability of some of these pockets and the strict conservation of the residues involved, across the large sequence diversity of NS1, thus emphasizing the robustness of such an approach to identify broadly active RBD NS1-targeting compounds. In the present study, the molecular dynamics allowed us to study the stability of the RBD and characterize the ED-RBD movements during the simulations for different strains, thus providing a better understanding of the dynamic properties of the NS1 structure in its different forms. Our data confirm the remarkable stability of the RBD for different strains regardless of the conformation and of the linker length, thus emphasizing the interest of targeting the RBD by drug design approaches to develop a strain-independent therapy. Towards this aim, the next step will consist in extracting RBD druggable pockets, common to the different strains and then to use docking approaches to search for candidate compounds that bind these pockets.

Moreover, a study will be carried out to simulate models of these three strains built with a short linker versus a long linker in order to better understand and confirm certain hypotheses on the impact of the length of the linker region on the polymorphism of the NS1 protein, or to study other strains.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/ijms23031805/s1.

Author Contributions: S.N.: conceptualization, methodology, formal analysis, investigation, writing—original draft preparation, writing—review and editing. D.M.: writing—review and editing, project administration, funding acquisition. A.-C.C. conceptualization, methodology, investigation, writing—original draft preparation, writing—review and editing, supervision, project administration, funding acquisition. D.F.: conceptualization, methodology, investigation, writing—original draft preparation, writing—review and editing, supervision, project administration, funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding: The authors gratefully acknowledge the financial support of the Université de Paris, the CNRS institute, and the INSERM institute.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is contained within the article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. WHO. Influenza (Seasonal). Available online: https://www.who.int/en/news-room/fact-sheets/detail/influenza-(seasonal) (accessed on 20 December 2021).
2. Iuliano, A.D.; Roguski, K.M.; Chang, H.H.; Muscatello, D.J.; Palekar, R.; Tempia, S.; Cohen, C.; Gran, J.M.; Schanzer, D.; Cowling, B.J.; et al. Estimates of global seasonal influenza-associated respiratory mortality: A modelling study. *Lancet* **2018**, *391*, 1285–1300. [CrossRef]
3. Hale, B.G.; Randall, R.E.; Ortín, J.; Jackson, D. The multifunctional NS1 protein of influenza A viruses. *J. Gen. Virol.* **2008**, *89*, 2359–2376. [CrossRef] [PubMed]
4. Zhirnov, O.P.; Konakova, T.E.; Wolff, T.; Klenk, H.D. NS1 Protein of Influenza A Virus Down-Regulates Apoptosis. *J. Virol.* **2002**, *76*, 1617–1625. [CrossRef] [PubMed]
5. Engel, D.A. The influenza virus NS1 protein as a therapeutic target. *Antivir. Res.* **2013**, *99*, 409–416. [CrossRef]
6. Rosário-Ferreira, N.; Preto, A.J.; Melo, R.; Moreira, I.S.; Brito, R.M.M. The Central Role of Non-Structural Protein 1 (NS1) in Influenza Biology and Infection. *Int. J. Mol. Sci.* **2020**, *21*, 1511. [CrossRef] [PubMed]
7. Marc, D.; Barbachou, S.; Soubieux, D. The RNA-binding domain of influenza virus non-structural protein-1 cooperatively binds to virus-specific RNA sequences in a structure-dependent manner. *Nucleic Acids Res.* **2012**, *41*, 434–449. [CrossRef] [PubMed]
8. Xia, S.; Robertus, J.D. X-ray structures of NS1 effector domain mutants. *Arch. Biochem. Biophys.* **2010**, *494*, 198–204. [CrossRef]
9. Krug, R.M. Functions of the influenza A virus NS1 protein in antiviral defense. *Curr. Opin. Virol.* **2015**, *12*, 1–6. [CrossRef]
10. Cheng, A.; Wong, S.M.; Yuan, Y.A. Structural basis for dsRNA recognition by NS1 protein of influenza A virus. *Cell Res.* **2008**, *19*, 187–195. [CrossRef]
11. Trapp, S.; Soubieux, D.; Lidove, A.; Esnault, E.; Lion, A.; Guillory, V.; Wacquiez, A.; Kut, E.; Quéré, P.; Larcher, T.; et al. Major contribution of the RNA-binding domain of NS1 in the pathogenicity and replication potential of an avian H7N1 influenza virus in chickens. *Virol. J.* **2018**, *15*, 55. [CrossRef] [PubMed]
12. Berman, H.M. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242. [CrossRef] [PubMed]
13. Marc, D. NS1 des Virus Influenza: Une Protéine Très « Influente ». *Virologie (Montrouge)* **2012**, *16*, 95–106. [CrossRef] [PubMed]
14. Kerry, P.S.; Ayllon, J.; Taylor, M.A.; Hass, C.; Lewis, A.; García-Sastre, A.; Randall, R.E.; Hale, B.G.; Russell, R.J. A Transient Homotypic Interaction Model for the Influenza A Virus NS1 Protein Effector Domain. *PLoS ONE* **2011**, *6*, e17946. [CrossRef] [PubMed]
15. Carrillo, B.; Choi, J.M.; Bornholdt, Z.A.; Sankaran, B.; Rice, A.P.; Prasad, B.V.V. The Influenza A Virus Protein NS1 Displays Structural Polymorphism. *J. Virol.* **2014**, *88*, 4113–4122. [CrossRef] [PubMed]
16. Hale, B.G. Conformational plasticity of the influenza A virus NS1 protein. *J. Gen. Virol.* **2014**, *95*, 2099–2105. [CrossRef] [PubMed]
17. Aramini, J.M.; Ma, L.C.; Zhou, L.; Schauder, C.M.; Hamilton, K.; Amer, B.R.; Mack, T.R.; Lee, H.W.; Ciccosanti, C.T.; Zhao, L.; et al. Dimer Interface of the Effector Domain of Non-structural Protein 1 from Influenza A Virus. *J. Biol. Chem.* **2011**, *286*, 26050–26060. [CrossRef]
18. Shen, Q.; Cho, J.H. The structure and conformational plasticity of the nonstructural protein 1 of the 1918 influenza A virus. *Biochem. Biophys. Res. Commun.* **2019**, *518*, 178–182. [CrossRef]
19. Jureka, A.; Kleinpeter, A.; Cornilescu, G.; Cornilescu, C.; Petit, C. Structural Basis for a Novel Interaction between the NS1 Protein Derived from the 1918 Influenza Virus and RIG-I. *Structure* **2015**, *23*, 2001–2010. [CrossRef]
20. Bornholdt, Z.A.; Prasad, B.V.V. X-ray structure of NS1 from a highly pathogenic H5N1 influenza virus. *Nature* **2008**, *456*, 985–988. [CrossRef]

21. Koliopoulos, M.G.; Lethier, M.; van der Veen, A.G.; Haubrich, K.; Hennig, J.; Kowalinski, E.; Stevens, R.V.; Martin, S.R.; Reis, E.; Sousa, C.; et al. Molecular mechanism of influenza A NS1-mediated TRIM25 recognition and inhibition. *Nat. Commun.* **2018**, *9*, 1820. [CrossRef]
22. Mitra, S.; Kumar, D.; Hu, L.; Sankaran, B.; Moosa, M.M.; Rice, A.P.; Ferreon, J.C.; Ferreon, A.C.M.; Prasad, B.V.V. Influenza A Virus Protein NS1 Exhibits Strain-Independent Conformational Plasticity. *J. Virol.* **2019**, *93*. [CrossRef] [PubMed]
23. Gack, M.U.; Albrecht, R.A.; Urano, T.; Inn, K.S.; Huang, I.C.; Carnero, E.; Farzan, M.; Inoue, S.; Jung, J.U.; García-Sastre, A. Influenza A Virus NS1 Targets the Ubiquitin Ligase TRIM25 to Evade Recognition by the Host Viral RNA Sensor RIG-I. *Cell Host Microbe* **2009**, *5*, 439–449. [CrossRef] [PubMed]
24. Kleinpeter, A.B.; Jureka, A.S.; Falahat, S.M.; Green, T.J.; Petit, C.M. Structural analyses reveal the mechanism of inhibition of influenza virus NS1 by two antiviral compounds. *J. Biol. Chem.* **2018**, *293*, 14659–14668. [CrossRef] [PubMed]
25. Abi Hussein, H.; Geneix, C.; Cauvin, C.; Marc, D.; Flatters, D.; Camproux, A.C. Molecular Dynamics Simulations of Influenza A Virus NS1 Reveal a Remarkably Stable RNA-Binding Domain Harboring Promising Druggable Pockets. *Viruses* **2020**, *12*, 537. [CrossRef]
26. Cho, J.H.; Zhao, B.; Shi, J.; Savage, N.; Shen, Q.; Byrnes, J.; Yang, L.; Hwang, W.; Li, P. Molecular recognition of a host protein by NS1 of pandemic and seasonal influenza A viruses. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 6550–6558. [CrossRef] [PubMed]
27. Sievers, F.; Wilm, A.; Dineen, D.; Gibson, T.J.; Karplus, K.; Li, W.; Lopez, R.; McWilliam, H.; Remmert, M.; Söding, J.; et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **2011**, *7*, 539. [CrossRef]
28. Webb, B.; Sali, A. Comparative Protein Structure Modeling Using Modeller. *Curr. Protoc. Bioinform.* **2016**, *54*. [CrossRef]
29. Abraham, M.J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J.C.; Hess, B.; Lindahl, E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *1–2*, 19–25. [CrossRef]
30. Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J.L.; Dror, R.O.; Shaw, D.E. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins Struct. Funct. Bioinform.* **2010**, *78*, 1950–1958. [CrossRef]
31. Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graph.* **1996**, *14*, 33–38. [CrossRef]
32. De Lano, W.L. The PyMOL Molecular Graphics System. 2002. Available online: www.pymol.org (accessed on 15 December 2021).

International Journal of
Molecular Sciences

MDPI

Supplementary Materials:

The following are available online at https://www.mdpi.com/article/10.3390/ijms23031805/s1.

Figure S1:

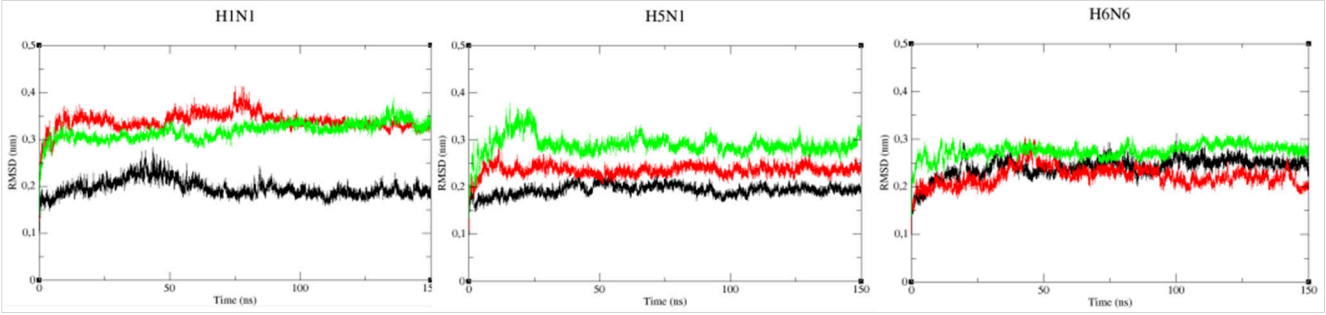


Figure S1. RMSD curves of the RBD domain for the three strains in the three forms (closed in black, semi-open in red and open in green).

Figure S2:

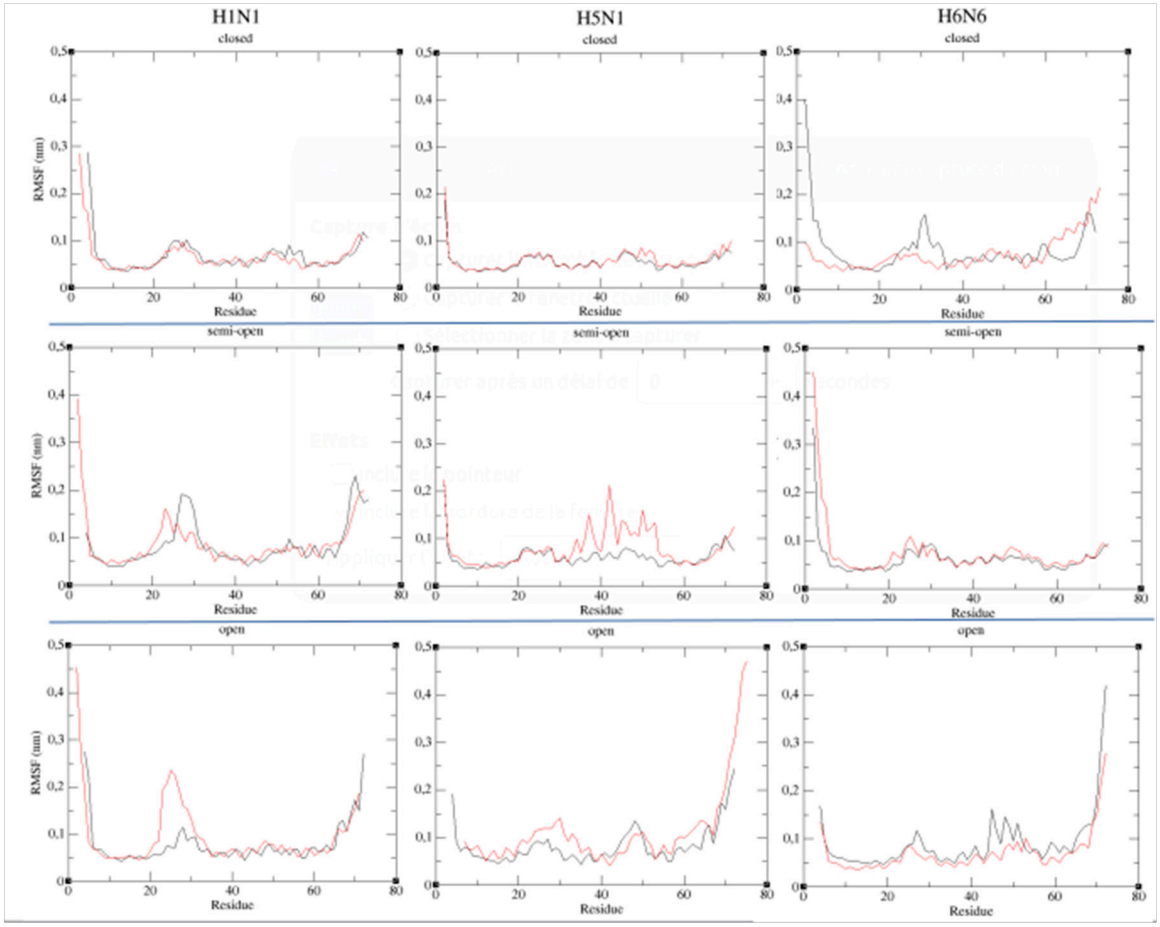


Figure S2. (Cα RMSF) curves of the two chains of the RBD domain (residues 1 to 72) for the three forms. The chains A and B colored in black and red respectively.

Figure S3:

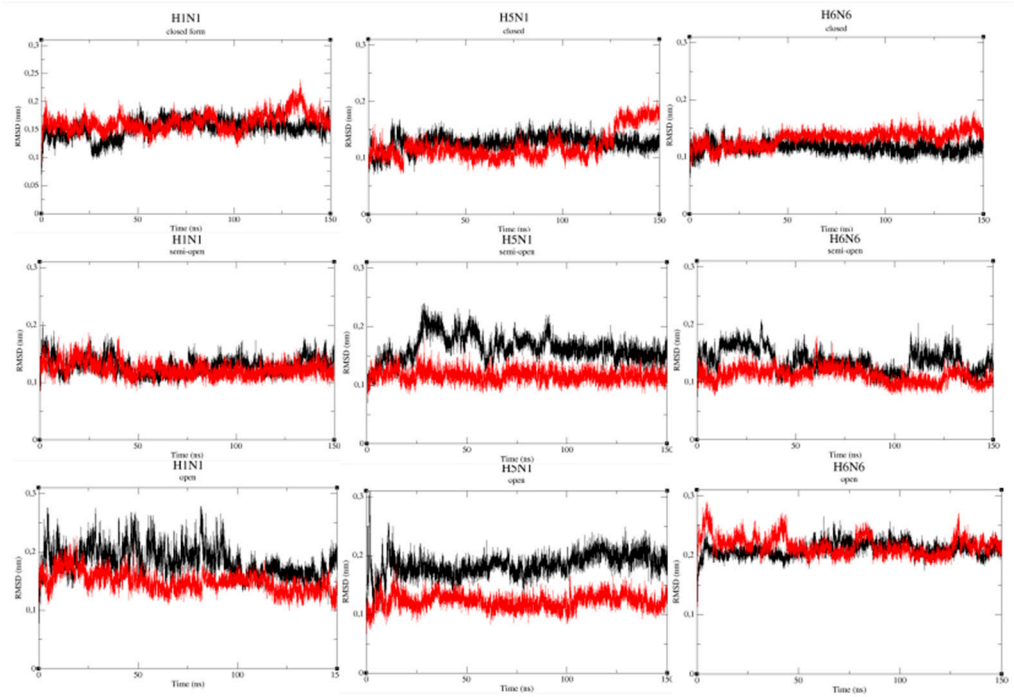


Figure S3. (C*α* RMSD) curves of each ED monomers A (colored in black) and B (colored in red) fittest on themselves for the three strains and the three forms.

Figure S4:

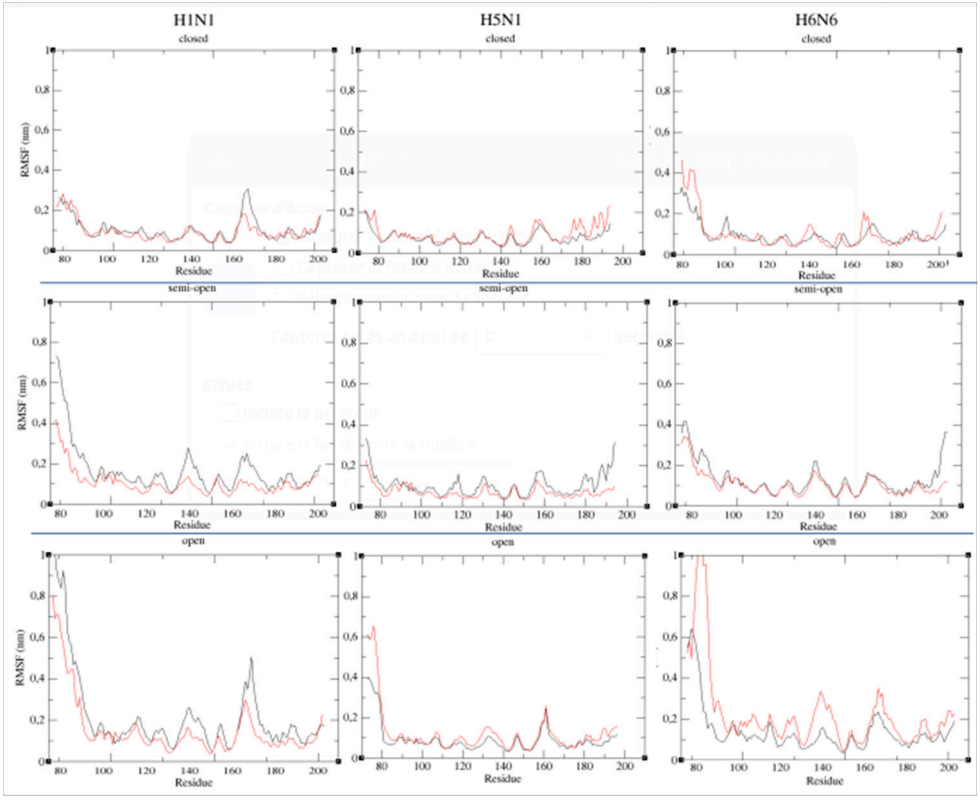
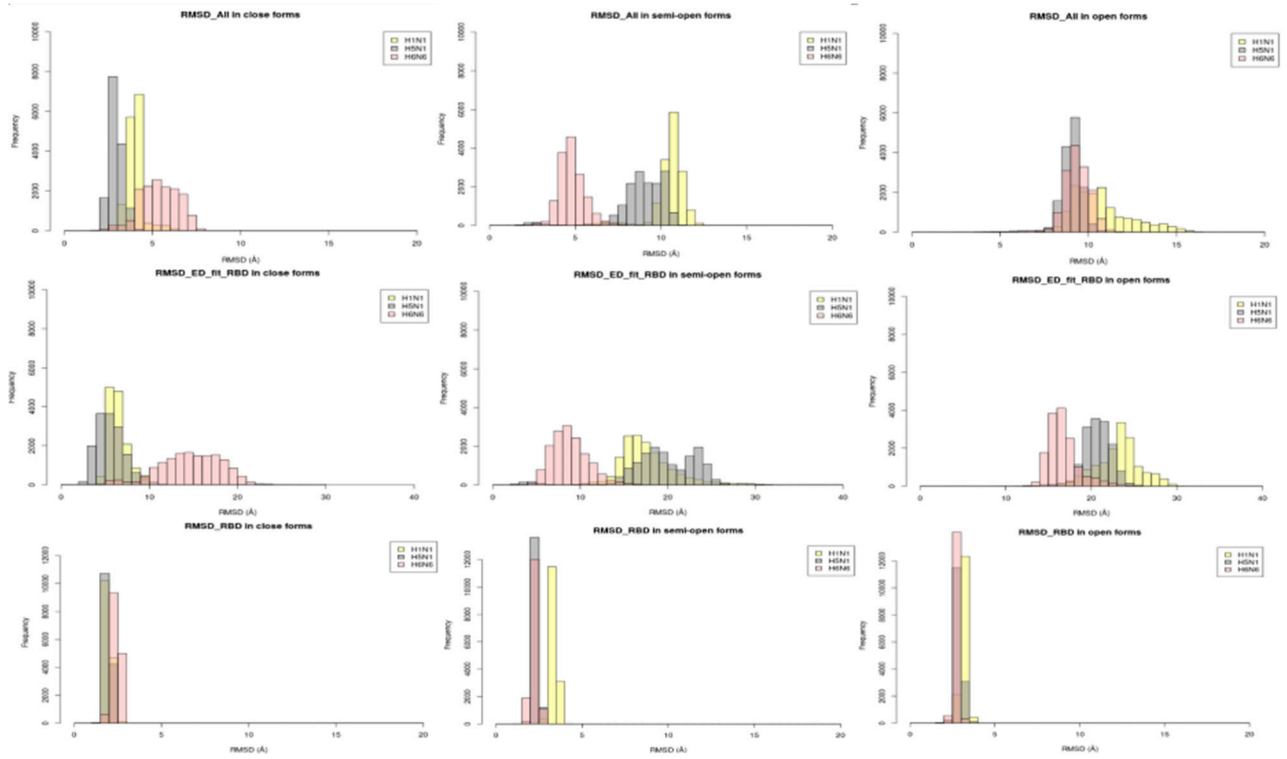
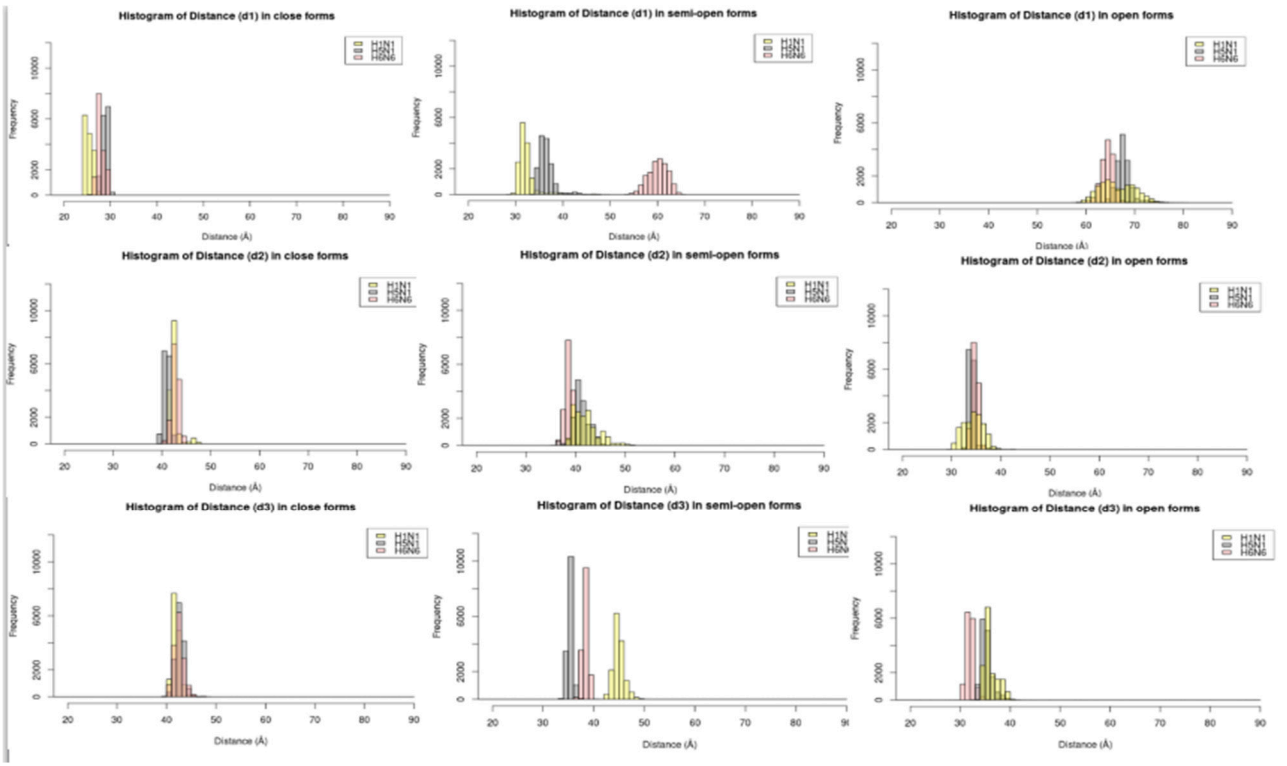


Figure S4. (C*α*RMSF) curves of the two chains of ED domain including the linker region for the three strains in the three forms. The chains A and B colored in black and red respectively.

Figure S5:



(**a**)



(**b**)

Figure S5: (a) Histograms of RMSD values for the different strains (H1N1 in yellow, H5N1 in grey and H6N6 in pink): RMSDs are calculated on the whole protein, on the ED after fitted on RBD and on RBD respectively. Histograms are represented respectively for NS1 in the closed form (on the left), in the semi-open form (in the middle) and in the open form (on the right). (b) Histograms of the three distance values for the different strains (H1N1 in yellow, H5N1 in grey and H6N6 in pink): distances are calculated between the ED domains (d1), between the RBD dimer and the ED A chain (d1) and

between the RBD dimer and the ED B chain (d3) respectively. Histograms are represented respectively for NS1 in the closed form (on the left), in the semi-open form (in the middle) and in the open form (on the right).

Table S1:

Distance (Å)	**Closed form**			**Semi-open form**			**Open form**		
Strains	**H1N1^{HM}**	**H5N1^{HM}**	**H6N6^{HM}**	**H1N1^{XR}**	**H5N1^{HM}**	**H6N6^{XR}**	**H1N1^{HM}**	**H5N1^{XR}**	**H6N6^{HM}**
Distance d1	27.7	28.1	27.8	47.0	43.3	59.8	74.4	68.9	75.3
Distance d2	40.2	39.9	39.6	49.5	35.6	37.9	37	34.2	39
Distance d3	42.9	42.4	42.6	46.6	33.7	38	37.5	34.7	36.6

Table S1: Table of distances between the different domains of NS1 for three strains in the three forms on the initial structures. The distance (d1) corresponds to the distance between the geometric center of the two ED monomers. The distances (d2, d3) correspond respectively to the distance between each of the geometric center of the monomer chains A and B of the ED domain and the geometric center of the RBD dimer.

Table S2:

Strains/ Forms	**Closed form**		**Semi-open form**		**Open form**	
strains	**H6N6^{HM}**		**H6N6^{XR}**		**H6N6^{HM}**	
RMSD values (in Å)						
Time of simulation (in ns)	150	340 (extended)	150	298 (extended)	150	180 (extended)
RMSD All	5.0 ± 1.1	6.5 ± 1.3	4.2 ± 0.7	4.7 ± 0.7	8.9 ± 7.9	9.3 ± 0.7
RMSD RBD	2.4 ± 0.2	2.3 ± 0.1	2.2 ± 0.2	2.1 ± 0.1	2.7± 0.1	2.7 ± 0.1
RMSD ED fit to RBD	14.7 ± 3.4	17.5 ± 4.1	8.8 ± 2.0	8.8 ± 0.2	16.6 ± 1.8	16.6 ± 1.7

Table S2: Average C*α* RMSD for extended simulations of H6N6 strain in the three forms (closed, semi-open, open) during the trajectory. RMSD_All is the average RMSD calculated over the whole protein, RMSD_RBD is the RMSD of the RBD dimer after fitting to the RBD, RMSD ED_A and ED_B are the RMSD of each monomer A and B of the ED domain calculated independently and RMSD ED fit on RBD corresponds to the RMSD of the ED domains fitting to the RBD dimer.

Figure S6:

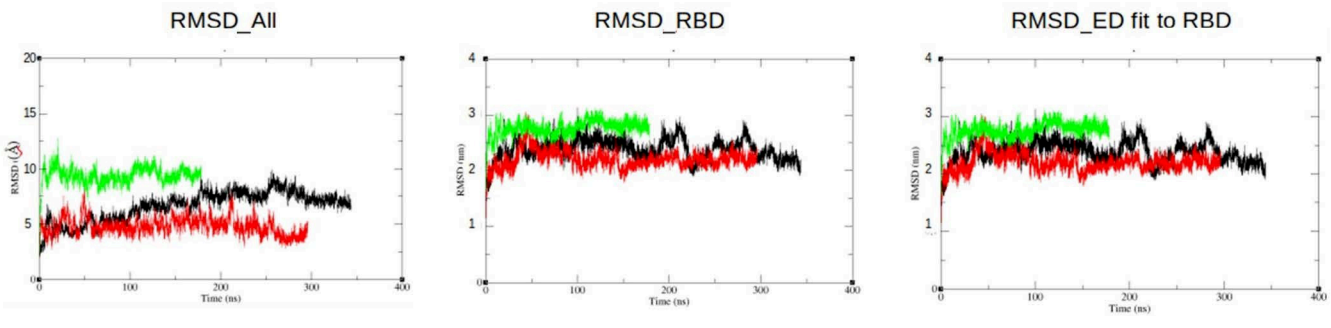


Figure S6. C*α* RMSD for extended simulations of H6N6 strain in the three forms (closed, semi-open, open) during 340ns, 298ns and 180ns respectively. RMSD_All is the average RMSD calculated over the whole protein, RMSD_RBD is the RMSD of the RBD dimer after fitting to the RBD, RMSD ED_A and ED_B are the RMSD of each monomer A and B of the ED domain calculated independently and RMSD ED fit on RBD corresponds to the RMSD of the ED domains fitting to the RBD dimer. .

3.2 Détection d’un site de liaison druggable du domaine RNA-BD commun aux différents sous-types de NS1

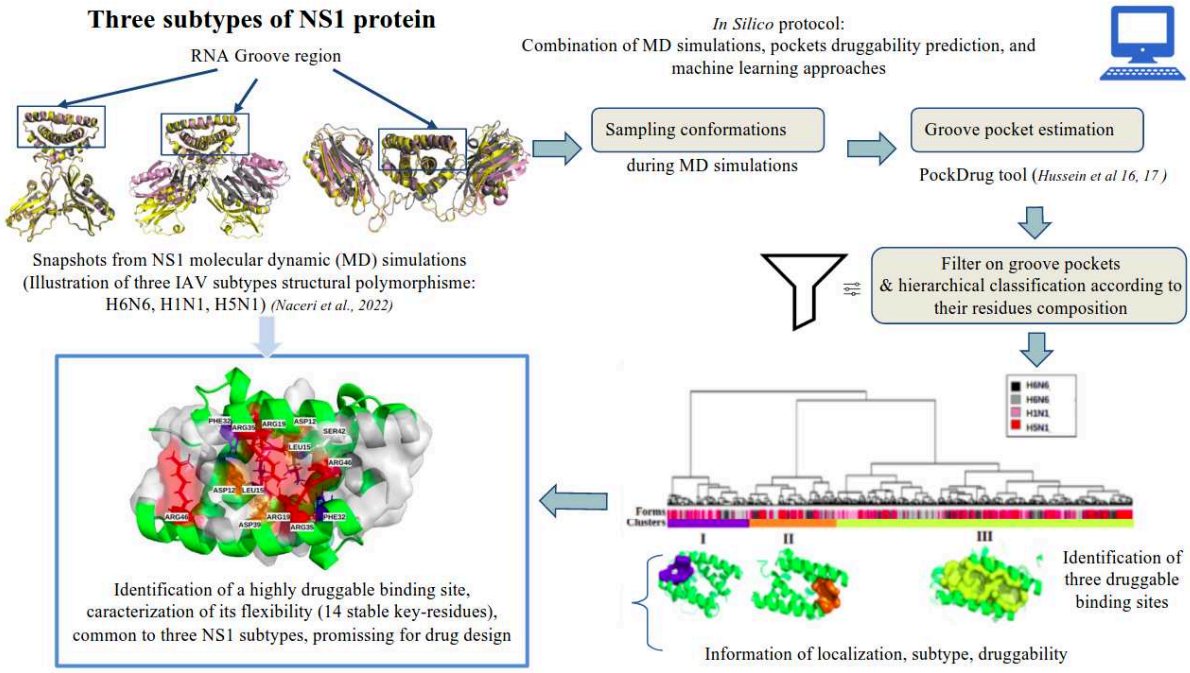


Figure 5: Protocole d’identification d’un site de liaison druggable commun aux différents sous-types de la protéine NS1 du virus Influenza A au niveau du RNA-BD. Trois sous-types ont été étudiés à savoir H6N6, H1N1 et H5N1. Après étude de la stabilité du domaine RNA-BD au cours des simulations de dynamique moléculaire un échantillonnage à intervalle régulier a été réalisé sur les trajectoires de dynamique moléculaire pour ensuite rechercher, à l’aide de l’outil PockDrug, les poches au niveau de la région du sillon. L’ensemble des poches identifiées a été classifié en terme de composition en résidus et a permis de mettre en évidence trois classes de poches druggables dont une grande poche fréquente, hautement druggable partageant un socle commun de 14 résidus et présente chez les trois sous-types.

Les simulations de DM ont permis d'étudier la structure de la protéine NS1 et ont mis en évidence la grande stabilité du RNA-BD au cours des trajectoires. Cette région étant essentielle à la réplication du virus, elle représente la cible privilégiée pour inhiber son interaction avec l'ARN. Ainsi, une analyse des cavités présentes dans la région du sillon au niveau du RNA-BD a été effectuée afin de mieux caractériser ses propriétés. Dans cette étude, seules les structures cristallographiques disponibles dans la base de données PDB (validées expérimentalement) ont été utilisées, notamment les sous-types H6N6 de forme fermée (linker court), H6N6 semi-ouverte (linker long), H1N1 semi-ouverte (linker long) et H5N1 ouverte (linker court) correspondant respectivement aux structures PDB: 4OPA, 4OPH, 5NT2 et 3F5T (Carrillo et al., 2014; Bornholdt & Prasad, 2008; Koliopoulos et al., 2018).

L'objectif de cette étude est d'identifier les poches qui apparaissent au niveau du sillon du RNA-BD au cours des changements conformationnels de la protéine sur les quatre trajectoires de DM, afin de suivre leur évolution mesurer leur déformabilité, leur taille, les résidus les plus fréquemment impliqués ainsi que leur druggabilité. Ce protocole vise à combiner les approches de la DM et de la recherche de poches afin de vérifier si la poche druggable identifiée sur le sous-type H6N6 de forme fermée (linker court) (Hussein et al.,

2020) est également identifiée suivant notre approche dans les quatre structures examinées. Cette étude a pour objectif de confirmer que NS1 est une cible thérapeutique prometteuse et qu'elle serait capable d'accueillir une molécule médicamenteuse efficace et indépendante de la forme ou du sous-type. Cette analyse est schématisée par la Figure 5. Ce travail a donné lieu à la publication suivante:
Naceri, S., Marc, D., Blot, R., Flatters, D.*, & Camproux, A. C.* (2022). Druggable Pockets at the RNA Interface Region of Influenza A Virus NS1 Protein Are Conserved across Sequence Variants from Distinct Subtypes. *Biomolecules*, *13*, 64.

3.2.1 Exploration conformationnelle au cours du temps

Les simulations de DM ont permis de suivre au niveau atomique les changements conformationnels et les fluctuations de la protéine NS1. En étudiant la flexibilité de la protéine NS1, il est également possible d'observer comment les poches druggables situées à sa surface évoluent au cours du temps. Suivant le protocole présenté dans le chapitre 2, un échantillonnage des conformations au cours des simulations de DM a été effectué toutes les 300 ps sur les trajectoires simulées à l'aide de GROMACS (Abraham et al., 2015; Berendsen et al., 1995). Cette fréquence d'échantillonnage permet de collecter un grand nombre de conformations (500 pour chacune des trajectoires de 150 ns), soit un total de 2000 conformations pour les quatre simulations de DM des différents sous-types étudiés. L'ensemble de ces conformations a servi à la recherche de poches susceptibles d'apparaître au cours des simulations.

3.2.2 Recherche de poches

La recherche de poche est effectuée sur les 2000 conformations collectées au cours des trajectoires grâce à l'outil PockDrug (Borrel et al., 2015). Près de 34,000 poches ont été identifiées sur les 2000 conformations échantillonnées. Un filtre sur la localisation des poches a été appliqué pour ne garder que les poches comprenant au moins un résidu de chacune des hélices α2 et α2' constituant la région du sillon localisée au niveau du RNA-BD; les poches dont la taille était inférieure à 8 résidus et supérieure à 50 résidus ont été retirés car la taille de la poche ne doit pas être trop grande autrement le ligand risque de ne pas être suffisamment bien fixé et ne pas former des interactions spécifiques avec la protéine, tandis qu'une poche trop petite peut limiter les options de conception de ligands potentiels; enfin un filtre sur la druggabilité, où seules les poches druggable ont été prises en compte car ce sont les poches d'intérêt à suivre.

Ces différents filtres ont réduit le nombre de poches à étudier à 1272 poches druggables. Une étape d'extraction de données sur ces poches a ensuite été effectuée pour rassembler les informations nécessaires à la suite de l'étude, notamment les numéros de conformations dans l'ordre d'enregistrement des trajectoires de DM ainsi que les numéros de poches associés à chaque conformation, les résidus qui constituent les poches, le nombre de résidus par poche ainsi que le score de druggabilité. De plus, l'information sur le sous-type a été indexée pour permettre un suivi des poches en fonction de la structure d'origine pour chacune des simulations.

3.2.3 Analyse statistique multivariée de l'ensemble des poches retenues

Afin de suivre l'évolution des poches au cours des simulations de DM selon les sous-types auxquels elles sont liées, une classification des poches en fonction de leur composition en termes d'acides aminés a été réalisée. Le protocole mis en place permet de réaliser une classification hiérarchique à partir des données collectées en comparant les

résidus qui composent l’ensemble des poches druggables identifiées. Les clusters identifiés lors de la classification ont permis de regrouper les poches ayant les mêmes propriétés en termes de résidus ce qui permet de donner un aperçu des différentes régions de localisation des poches au niveau du sillon au cours du mouvement de la protéine ainsi que l’information sur leur stabilité, c'est-à-dire leur fréquence d'apparition au cours de la DM chez les différents sous-types.

3.2.4 Sites de liaisons identifiées

A l’issue de la classification, trois régions ont été identifiées au niveau de la région du sillon du RNA-BD. Deux d’entre elles sont localisées aux extrémités du sillon, caractérisées par une petite taille de poches. La troisième poche est localisée au centre du sillon incluant davantage de résidus et caractérisée par un socle commun aux différents sous-types de 14 résidus. Le score de druggabilité de cette poche était beaucoup plus élevé que pour les deux poches précédentes. Cette poche commune a, pour cette étude, été retenue comme le site de liaison d'intérêt à cibler lors de la recherche de molécules thérapeutique. En effet, la présence d’un ligand inhibiteur au niveau de cette poche commune, dont les propriétés physico-chimiques et géométriques seraient complémentaires au site de liaison, serait susceptible d'empêcher l’interaction avec l’ARN, par conséquent bloquer la réplication du virus influenza A.

3.2.5 Conclusion

Plusieurs approches ont été utilisées pour déterminer la présence d'un site de liaison propice à la fixation d'une petite molécule inhibitrice sur la surface de la protéine NS1 du virus de la grippe A. Ces approches comprennent les simulations de DM, la recherche de poches druggable et des méthodes d’apprentissage non supervisées. Nous avons examiné quatre structures de formes différentes des trois sous-types de la protéine NS1 pour identifier un site cible potentiel qui pourrait empêcher l'interaction entre le RNA-BD de la protéine NS1 et l'ARN viral.

Trois sites de liaison druggables ont été identifiés dans la région du sillon sur les quatre structures. La combinaison de ces approches a permis d’identifier un grand site de liaison commun hautement druggable localisé au centre du sillon impliquant 14 résidus clés très fréquents au cours des simulations dont certains sont susceptibles d’interagir avec l'ARN de par leurs propriétés physico-chimiques. Ce site commun présente un score de druggabilité élevé dans plus de 20% des conformations échantillonnées dans les simulations de MD. Cela confirme les conclusions des deux études antérieures sur le polymorphisme de la protéine NS1 et la recherche de poches sur le sous-type H6N6, qui ont suggéré que NS1 est une cible thérapeutique prometteuse pour le virus Influenza A.

biomolecules

MDPI

Article

Druggable Pockets at the RNA Interface Region of Influenza A Virus NS1 Protein Are Conserved across Sequence Variants from Distinct Subtypes

Sarah Naceri [1], Daniel Marc [2,3], Rachel Blot [1], Delphine Flatters [1,*,†] and Anne-Claude Camproux [1,*,†]

[1] Unité de Biologie Fonctionnelle et Adaptative, CNRS, INSERM, Université Paris Cité, F-75013 Paris, France
[2] Equipe 3IMo, UMR1282 Infectiologie et Santé Publique, INRAE, F-37380 Nouzilly, France
[3] UMR1282 Infectiologie et Santé Publique, Université de Tours, F-37000 Tours, France
* Correspondence: delphine.flatters@u-paris.fr (D.F.); anne-claude.camproux@u-paris.fr (A.-C.C.)
† These authors contributed equally to this work.

Abstract: Influenza A viruses still represent a major health issue, for both humans and animals. One of the main viral proteins of interest to target is the NS1 protein, which counters the host immune response and promotes viral replication. NS1 is a homodimer composed of a dimeric RNA-binding domain (RBD), which is structurally stable and conserved in sequence, and two effector domains that are tethered to the RBD by linker regions. This linker flexibility leads to NS1 polymorphism and can therefore exhibit different forms. Previously, we identified a putative drug-binding site, located in the RBD interface in a crystal structure of NS1. This pocket could be targeted to block RNA binding and inhibit NS1 activities. The objective of the present study is to confirm the presence of this druggable site, whatever the sequence variants, in order to develop a universal therapeutic compound that is insensitive to sequence variations and structural flexibility. Using a set of four NS1 full-length structures, we combined different bioinformatics approaches such as pocket tracking along molecular dynamics simulations, druggability prediction and classification. This protocol successfully confirmed a frequent large binding-site that is highly druggable and shared by different NS1 forms, which is promising for developing a robust NS1-targeted therapy.

Keywords: binding site; influenza A virus; non-structural protein 1; groove-pocket; drug design; structural polymorphism

Citation: Naceri, S.; Marc, D.; Blot, R.; Flatters, D.; Camproux, A.-C. Druggable Pockets at the RNA Interface Region of Influenza A Virus NS1 Protein Are Conserved across Sequence Variants from Distinct Subtypes. *Biomolecules* **2023**, *13*, 64. https://doi.org/10.3390/biom13010064

Academic Editor: Alessandro Paiardini

Received: 2 December 2022
Revised: 24 December 2022
Accepted: 25 December 2022
Published: 29 December 2022

1. Introduction

Influenza A viruses are enveloped viruses belonging to the family Orthomyxoviridae [1]. They infect a wide spectrum of bird and mammal species, and in humans they are responsible for seasonal epidemics and global pandemics, resulting in severe pneumonia and death in affected patients. The 1918 influenza pandemic virus caused severe pneumonia, resulting in an estimated 50 million deaths worldwide. The morbidity and mortality rate of the influenza A virus reaches three to five million severe cases and 290,000 to 650,000 deaths each year [2–4]. Currently, there are few alternatives to the influenza vaccine, the efficacy of which depends on the subtype of influenza A virus and can be jeopardized by amino-acid substitutions in hemagglutinin, the major viral antigen.

Small compound therapy can be useful when prevention measures do not suffice. Currently, the available anti-influenza drugs comprise the neuraminidase and the viral polymerase inhibitors, both of which were designed based on the knowledge of target proteins structures. Non-structural protein 1 (NS1) stands as a possible additional target for novel antiviral compounds. NS1 has been extensively studied, because of its multifunctional character. Absent from the viral particle, it is highly expressed in the cytoplasm and nucleus of infected cells, where it is able to interact with various components involved in the interferon system, inhibiting this response. NS1 relies on several mechanisms to

attenuate the interferon response during influenza-virus infection, and stands as the main interferon antagonist encoded by the virus [5]. This prominent role and its requirement to promote viral replication [6] make it a promising target for novel antiviral therapies. Thus, targeting the NS1 protein may provide a good outcome in the treatment of influenza infection [7–9].

NS1, a homodimer of two 230 residue polypeptides, is comprised of three structural domains, where only residues 1–203 have been resolved in crystallography. The RNA-binding domain (RBD) is an obligate dimer of residues 1–73, which is connected to the two effector domains (ED, residues 86–203) by the short, unstructured, linker regions [10]. The RBD is arranged as three pairs of symmetrically positioned antiparallel alpha-helices. Residues 86–203 of each chain fold autonomously into an ED. To explore the NS1 protein as a target for drug design, three dimensional (3D) structures of the protein are needed. However, only four full-length NS1 proteins (from subtypes H6N6, H1N1 and H5N1) have been crystallized and are available as 3D structures in the Protein Data Bank (PDB) [11]. The flexible linker enables the quaternary structure to adopt polymorphic shapes, of which three main forms have been observed: closed, semi-open and open, according to the position of each ED relative to its counterpart and to the dimeric RBD [12–14].

B.G Hale reviewed the conformational plasticity of NS1 [13]. This study underlined the importance of considering the dynamic properties of NS1, essential for its multiple functions. Therefore, NS1 conformational plasticity can impact the structure. Using computational molecular-dynamics (MD) simulations, our team explored the dynamic and flexibility properties of the four above-mentioned full-length NS1 structures [15]. Molecular-dynamics simulation is a computational approach classically used to characterize the variability of the global domain or 3D structure. In order to explore to what extent the structure as well as its flexible conformation were robust to variations in the amino-acid sequences, the four structures were submitted to MD simulation, allowing us to follow the dynamic of the conformational changes over time. The gradual switch from closed to open form along the MD simulations was studied by computing the three distances between the centers of mass of the domains and, notably, between the ED monomers, to characterize the opening state of the conformation. Our results confirmed the co-existence, regardless of the subtypes, of three structural forms in a dynamic equilibrium. Our data also suggested that the linker length, as well as the presence of specific amino acids, modulate the dynamic properties and the flexibility of NS1. Finally we confirmed the remarkable stability of the RBD for different subtypes, regardless of the forms and linker size, contrary to the ED, thus emphasizing the interest of targeting the RBD using drug-design approaches to develop a strain-independent therapy [15].

Other studies [16] have performed pocket extraction and druggability prediction on several PDB-registered crystal structures, focusing on ED monomers or ED dimers in different forms. Three main druggable pockets were thus identified in the ED, corresponding to regions that are highly conserved across different virus subtypes. Combining sequence analysis and druggability-prediction algorithms, the same authors [17] systematically probed in silico the druggability of the NS1 effector domain. Given the flexibility of the ED domain and its interaction with many partners, it proved more relevant to target the RBD for possible therapy because, beyond its stability, it is the main partner of the RNA leading to viral replication through its interaction.

To explore NS1 as a therapeutic target, studies have been carried out by researchers who have hypothesized that inhibiting NS1-RBD interaction could alter the functionality of the RBD and significantly reduce the replication potential of the virus and its pathogenicity [18,19]. This work showed that the RBD interface region is rather well conserved among influenza A viruses, based on PDB-registered crystal structures. They described a deep pocket, localized between the antiparallel alpha-helices 2 and 2′ of the RBD interface, which could be explored as a potential drug-binding site of NS1. However, many studies have been developed and have confirmed the importance of predicting the druggability

of proteins and binding sites in order to prioritize the targets for the development of therapies [20].

Although the RBD was confirmed as a very stable dimer [15], there is an inherent structural flexibility that could impact the pocket druggability. Indeed, the importance of taking into account the cryptic or hidden, transient and flexible pocket to identify reliable binding sites for drug design is increasingly highlighted [21]. Recently, our group preliminarily confirmed interest in considering the flexibility of the 3D structure using molecular dynamics simulations to extract druggable pockets from the RBD in the case of the H6N6 NS1 subtype with a structure in a closed form [22]. We performed a pocket analysis by considering the intrinsic flexibility of NS1, using H6N6 NS1 full-length protein MD simulations. Our main result was the identification of a druggable binding site located between the two antiparallel helices α2 and α2′ forming the RNA-binding interface, which we defined as the "RBD-groove". While this RBD-groove was not detected as a druggable pocket in the rigid crystal structure, its druggable properties were consistently observed in a substantial fraction of the conformations that were sampled during the MD simulations. This frequently observed druggable pocket confirmed the potentiality to prevent the binding of NS1 to RNA.

In this study, our aim is to confirm that such a druggable RBD-groove is stable, by considering NS1 structural polymorphism, to validate the interest of targeting the RBD for developing a robust and subtype-independent small compound targeting NS1. To that end, we study the existence of a druggable binding site located at the RBD-groove by considering its structural flexibility and NS1 structural polymorphism (shape and linker size). Our approach consists in estimating pockets conformation sampled from MD simulations on four available full-length NS1 structures from three subtypes (H6N6, H1N1, H5N1). We focused on pockets located at the RBD-groove, and characterized their druggability properties. A classification of these pockets based on residue-composition similarity allowed us to identify three main druggable clusters of groove-pockets describing putative druggable binding sites. These latter were analyzed in terms of frequency, residue composition, physico-chemical and druggability properties, variability and representativeness of different subtypes of the influenza A virus. One of these sites is of particular interest to target by the drug-design approach, as it is a large, druggable site, stable to structural polymorphism, and therefore common to the different subtypes.

2. Materials and Methods

2.1. Selection of NS1 Full-Length Structures and Study of Their Flexibility

2.1.1. Selection of Four Different Forms of the NS1 Protein

Three NS1 proteins from subtypes (H6N6, H1N1, H5N1) are considered, co-crystallized in full-length and available as 3D structures in the PDB. Although RBD has been shown to be stable, it is important to perform the full-length protein study to assess the impact of the NS1 polymorphism during pockets tracking.

Figure 1A illustrates a multiple sequence alignment, calculated with the ClustalΩ algorithm [23], of H6N6, H1N1 and H5N1 sequences from the UniProt database (Table S1), using the EMBOSS tools [24]. It can be noted that, despite sequence mutability, there is strong sequence conservation at the α2 and α2′ helices in the groove region. The RBD (1–73 amino acids) and ED (86–203 amino acids) domains connected by short (SL) or long (LL) linker regions (deletion or not of five residues "Δ80–84") are indicated in Figure 1B.

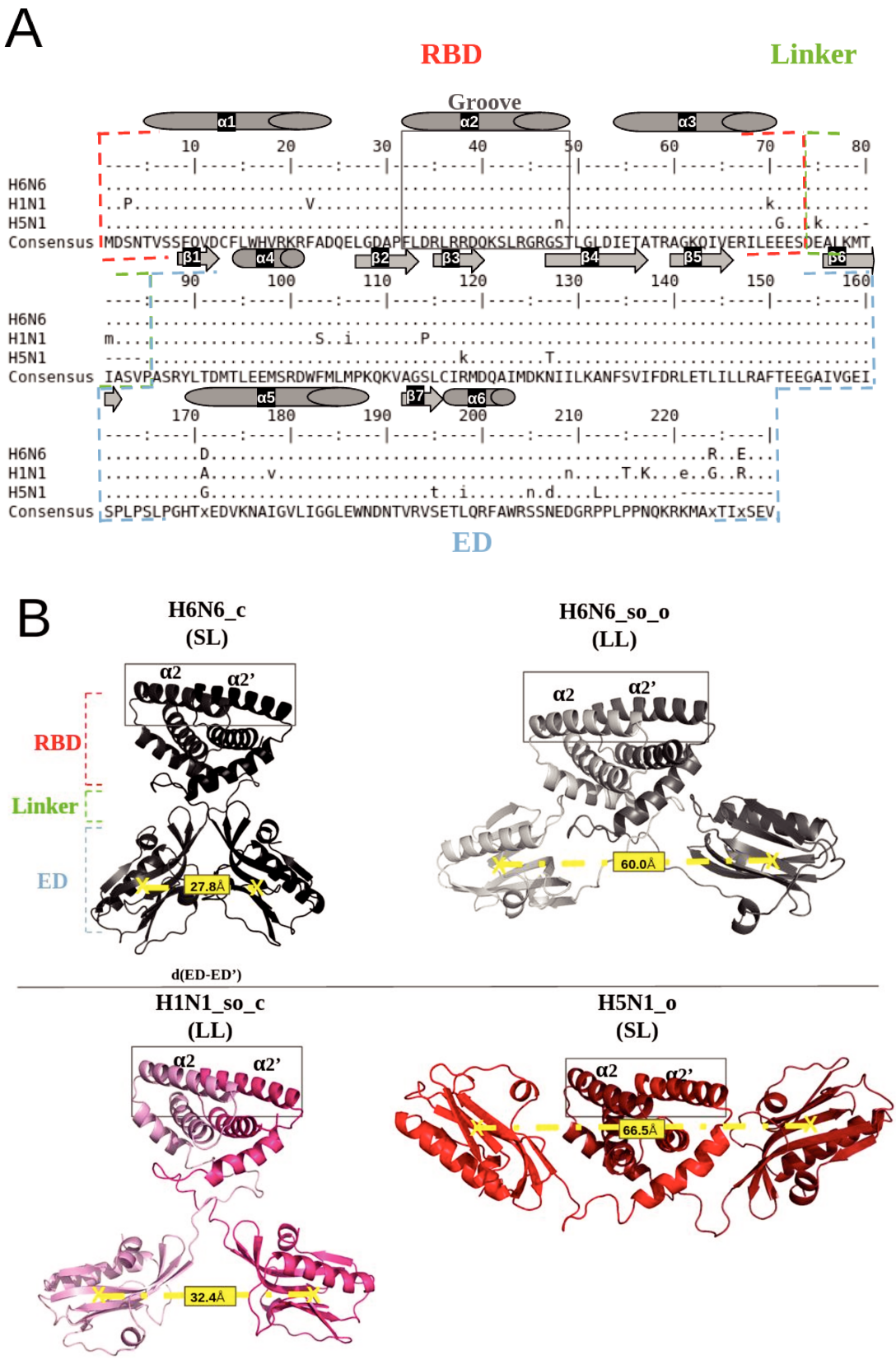


Figure 1. Illustration of the NS1 sequence conservation and the structural polymorphism of the NS1 protein: (**A**) Alignment of H6N6, H1N1 and H5N1 sequences extracted from the UniProt database (accession number: P03496, A5A5U1, Q20NS3 respectively) [25] using the "showalign" tool of software suite EMBOSS [24], showing conservation of the RBD region and particularly the "RBD-groove", where

the sequence is boxed in full gray lines. The RBD, ED and linker (long (LL) or short (SL)) are bordered by red, blue and green dotted lines, respectively. The α-helices are represented by cylinders and the β-sheets by arrows. Only differences relative to the consensus sequence (above and underneath) are shown, and the consensus line can be displayed in a mixture of upper- and lower-case symbols. Upper case indicates strong consensus and lower case indicates weak consensus; hyphens in H5N1 sequence are deletions (Δ80–84). (**B**) Representation of the four NS1 initial structures (4OPA, 4OPH, 5NT2 and 3F5T), colored in black, gray, pink and red, respectively, using the PyMol visualization software [26]. The "RBD-groove" is bordered by gray, full lines enclosing the two α2 and α2′ helices. The boundaries of the RBD, ED and the linker are shown in red, blue and green, respectively. The d(ED-ED') distance between the EDs monomers centers of mass quantifies the structural polymorphism of the different structures, which is on average 27.8 Å for H6N6_c, 60.0 Å for H6N6_so_o, 32.4 Å for H1N1_so_c and 66.5 Å for H5N1_o.

Four full-length NS1 crystallographic structures associated with three different subtypes of H6N6 (pdb codes: 4OPA and 4OPH), H1N1 (pdb code: 5NT2) and H5N1 (pdb code: 3F5T) are currently available in the PDB and were considered in this study [12,27,28]. After the rebuilding and reversion of the engineered mutations, i.e., R38A/K41A, and specifically the W187A mutation for 5NT2, the complete conformations of the initial structures' templates were obtained. The corresponding four initial structures are described in Table S1. The PDB structures 4OPA and 4OPH correspond to the H6N6 subtype but are in different forms and linker sizes: 4OPA is a closed form (its ED are separated by a distance of 28 Å) associated with a short linker (five amino acids Δ80–84 are deleted), whereas 4OPH is a semi-open form with open tendency (its two ED are separated by a distance of 60 Å) (Table S1), associated with a long linker (for distance measurement between two EDs, see Naceri et al., 2022 [15]. They are named H6N6_c and H6N6_so_o. The 5NT2 structure corresponds to the H1N1 subtype in a homodimeric structure and is classified as a semi-open form with closed tendency (its two ED are separated by a distance of 47 Å,) associated with a long linker, named H1N1_so_c. The 3F5T corresponds to the H5N1 subtype in an open form (its EDs are separated by a distance of 69 Å), associated with a short linker (Δ80–84), named H5N1_o.

These four initial NS1 structures (H6N6_c, H6N6_so_o, H1N1_so_c and H5N1_o) are illustrated in Figure 1B, with the RBD-groove indicated, as defined by Abi Hussein et al., 2020, and corresponding to the region between the two antiparallel helices α2 and α2′ forming the RNA-binding interface [22].

2.1.2. Sampling of NS1 Conformations during MD Simulation

MD simulations were run for 150 ns, in an isothermal–isobaric (NPT) ensemble at 300 K and 1 bar, at pH 7 and using the AMBER-99SB force field with the GROMACS software package V2019.5 [29]. The detailed protocol is described in Naceri et al., 2022 [15].

The sample of 500 frames of the NS1 full-length conformations were stored at regular intervals from each of the four MD simulations generated from the crystallographic structures (H6N6_c, H6N6_so_o, H1N1_so_c and H5N1_o). A resulting set of 2000 NS1 full-length conformations was obtained, with 500 conformations extracted from MD simulations obtained on each structure.

2.2. Machine-Learning Analysis of the Groove-Pockets to Identify Main Pocket Cluster

2.2.1. Estimation of Groove-Pockets

From these 2000 NS1 conformations, pocket extraction and characterization were performed using the PockDrug tool preliminary developed by the team [20]. Pockets were estimated using an automated geometry-based Fpocket [30]. Estimated pockets were characterized in terms of 19 physicochemical, geometrical properties and a druggability score prediction [31]. The list of residues involved in each pocket was also additionally extracted.

In this study, we applied three different filters (size of pocket $\geq$ 8 residues, pockets located on the groove and only druggable ones) to select pockets. The pockets considered in the groove are those including at least one residue from each of the two α2 and α2′ helices located on each RBD monomer of the NS1 protein (residues 30–50) that delimit the groove, as proposed by Abi Hussein et al., 2020 [22]. The corresponding pockets are called "groove-pockets". The last filter to select druggable pockets using the PockDrug is a threshold druggability score ($\geq$ 50%) indicating a druggable pocket.

2.2.2. Identification of Main Clusters of Pockets Located at the Groove Region

In Abi Hussein et al., 2020, we studied the evolution and similarity of the pockets along MD simulations in terms of the Euclidean distance between the pockets from PockDrug, based on the physico-chemical and geometrical descriptors [22]. In this work, we aim to study the evolution and similarity of pockets for three NS1 distinct sequence variants, H1N1, H6N1 and H5N1, and the impact of structural polymorphism on the occurrence of pockets and their druggability. Thus, we decided to quantify the pockets similarity in terms of common residues, using binary distance. A binary distance of 0 corresponds to two pockets including identical residues, and a distance of 1 corresponds to two pockets without common residue. We performed a hierarchical classification of the pockets using the binary distance using the Ward metric method (ward.D2) [32,33] and the R Hclust package [34].

Dendrogram visualizations were performed using the Heatmaps2 package in R (v 3.9) [35] to illustrate pocket similarity in terms of residues. The difference in terms of residues between pockets and pocket clusters increases with the dendogram branch-lengths (representative of the average binary distances). The resulting classification visualizes the similarity of the pockets in terms of common residues, and allows the identification of different main clusters of similar pockets.

2.3. Analysis of Druggable Binding Sites Identified by Main Pocket Clusters

Pockets with a similar residue composition are clustered within the same cluster, while pockets with different residue compositions are divided among different clusters. In this way, the set of similar pockets in a cluster can be considered as describing the flexibility of an associated binding site during the MD simulations for the four NS1 structures (H6N6_c, H6N6_so_o, H1N1_so_c and H5N1_o). The number of pockets in a cluster indicates the frequency of appearance of the binding site on the 2000 RBD-groove conformations. It provides quantification of the persistence of this binding site when considering groove evolution over time. The flexibility of the binding site is described by the residue variability of its associated pocket cluster. Highly frequent residues of a binding site (observed in more than 75% of the pocket cluster) are considered as a base, and named key-residues.

The analysis of main pocket clusters allows the detection of recurrent druggable binding sites and their characterization in terms of appearance frequency, key-residue composition, variability and druggability properties. Each binding site can be described by the analysis of its cluster of pockets. The size of the cluster provides the occurrence of a binding site along the MD simulations. The variability analysis of the pocket residues allows for the studying of the deformability of its corresponding site. The most frequent residues of the pocket cluster are considered as key residues of the binding site.

The main pocket clusters can also be analyzed in terms of the composition of the four structures (H6N6_c, H6N6_so_o H1N1_so_c and H5N1_o). Thus, we can determine if a binding site is specific, or common to the four considered structures.

3. Results

3.1. Extraction of Pockets Located at the RBD-Groove and Analysis of Its Druggability

The druggability of the RBD-groove for the four initial structures (H6N6_c, H6N6_so_o, H1N1_so_c and H5N1_o) was first analyzed. A groove was considered druggable if it included at least one druggable pocket. We searched the pockets located in the groove

region, i.e., between helices α2 and α2′, and characterized them in terms of druggability, using PockDrug. For three out of the four initial structures, we observed between one and three pockets in the RBD-groove that are predominantly non-druggable; see Table 1. However, two pockets in the RBD-groove of H1N1_so_c are predicted to be druggable by PockDrug (with druggability scores of 0.57 and 0.68), indicating that the H1N1_so_c RBD-groove is druggable. Therefore, considering the four initial crystal structures of NS1, only a quarter of the RBD-grooves can be identified as druggable.

Table 1. Number of groove-pockets (total or druggable) and groove conformations, according to the four forms of NS1 (H6N6_c, H6N6_so_o, H1N1_so_c, H5N1_o) obtained in the initial static PDB structures and on a set of 500 conformations sampled during the MD simulations.

Subtype_Forms	H6N6_c	H6N6_so_o	H1N1_so_c	H5N1_o	Total of Four Structures
Initial (static) structure/MD sampling conformations	1/ 500	1/ 500	1/ 500	1/ 500	4/ 2000
Number of groove-pockets [Initial structure/ MD conformations (%)]	2/ 933 (27.7)	1/ 861 (25.6)	3/ 740 (22.0)	2/ 826 (24.5)	8/ 3360
Number of druggable groove-pockets [Initial structure/MD conformations (%)]	0/ 276 (29.6)	0/ 329 (38.2)	2/ 257 (34.7)	0/ 410 (49.6)	2/ 1272 (37.9)
Occurrence of druggable groove conformations [Initial structure/ MD conformations (%)]	0/ 239 (47.8)	0/ 285 (57.0)	1/ 233 (46.6)	0/ 345 (69.0)	1/ 1102 (55.1)

The evolution of the RBD-groove in terms of druggability was then explored by molecular dynamic simulations, to consider the flexibility of the structures. We analyzed the 2000 conformations, corresponding to four sets of 500 conformations extracted from each MD simulation of H6N6_c, H6N6_so_o, H1N1_so_c and H5N1_o (see method section).

We observed a total of 3360 pockets located at the RBD-groove and extracted from the 2000 conformations. This results in more than one pocket per RBD-groove (1.68 on average) for each conformation. Furthermore, a homogeneous distribution of these pockets was observed, ranging from 740 (22%) to 933 (28%) on the four samples of 500 conformations obtained from H6N6_c, H6N6_so_o, H1N1_so_c and H5N1_o (Table 1).

Overall, we observed that almost 40% (1272/3360) of the groove-pockets were druggable. Although the average number of pockets in the groove was almost equivalent for the four simulations, a more marked difference was observed with respect to their druggable properties: only 30% and 35% of the observed pockets were druggable for H6N6_c and H1N1_so_c, respectively, compared to almost 40% and 50% for H6N6_so_o and H5N1_o, respectively.

We then focused our analysis on the druggable RBD-grooves, i.e., on the 55.1% of the conformations exhibiting an RBD-groove that had at least one druggable pocket. This consistently observed druggability of the RBD-groove during the MD simulations is confirmed here for the four considered structures. Almost half of the RBD-grooves are druggable on the closed form (H6N6_c) and the semi-open form with closed tendency (H1N1_so_c); see Table 1. Concerning H6N6_c, the frequency of the druggable RBD-groove (47.8%) is very close to the one (~44%) obtained by Abi Hussein et al., 2020, which used three short MD simulations of 50 ns each, and only took into account pockets with more than 14 residues [22]. The higher frequency of druggable RBD-grooves (69.0%) is observed on the H5N1_o, leading to the suggestion that the open form may be more suitable for the presence of druggable pockets. Indeed, the second highest frequency of druggable RBD-groove (57%) is observed on the H6N6_so_o, in the semi-open form close to an open form.

Our results confirmed the importance of MD simulations in studying pockets druggability. Indeed, the groove-pockets are detected as druggable in more than half of the conformations over time, for the NS1 protein from different subtypes.

Consequently, the RBD-groove region meets the druggability criteria in 55.1% of the conformations for the different NS1 subtypes. We can note that the number of druggable pockets increases with the opening of the structure; the structure which presents the fewest druggable pockets is the H6N6_c, and the highest one is the H5N1_o, wherease semi-open forms presents a different number of druggable pockets depending on their degree of opening and closing.

This confirms that the RBD-groove region is able to form potential binding pockets, which, in more than half of the conformations, meet the druggability criteria, during MD simulations of the four initial structures. This groove region is an interesting candidate for targeting by a drug-like ligand for the different structures, regardless of its polymorphism.

3.2. Identification of Main Clusters of Druggable Groove-Pockets on Different NS1 Subtypes

We performed a hierarchical clustering of these 1272 druggable pockets, based on their residues composition, to identify main clusters and quantify their frequency of occurrence in the four different NS1 structures. Figure 2A illustrates the classification, resulting in three main clusters, I, II and III. These three most frequently observed clusters correspond to three distinct binding sites, noted as I, II and III, with a frequency of occurrence of 17.3%, 18.5% and 64.1%, respectively, of the 1272 druggable groove-pockets.

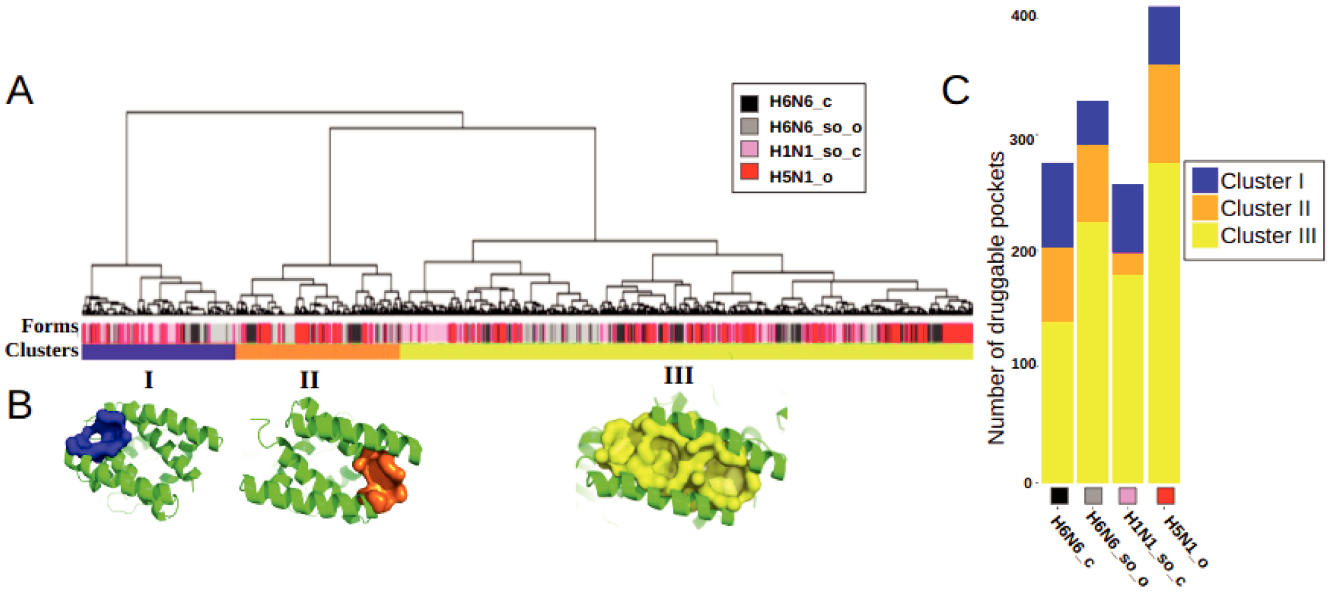


Figure 2. Classification and representation of the three main identified druggable groove-pocket clusters: (**A**) Hierarchical classification of druggable groove-pockets extracted from the MD simulations of the four crystallographic NS1 structures (H6N6_c, H6N6_so_o, H1N1_so_c and H5N1_o), indicated by the colors black, gray, pink and red, respectively, on the Forms line. Three clusters (I, II and III) were identified in the *Clusters* line, colored blue, orange and yellow, and comprising 220, 236 and 816 druggable groove-pockets respectively. (**B**) Representative pockets of three binding sites are illustrated on one representative groove conformation, in blue, orange and yellow for clusters I, II and III, respectively. (**C**) Histogram of the distribution of druggable pockets per cluster for each of the H6N6_c, H6N6_so_o, H1N1_so_c and H5N1_o structures.

The clusters are well differentiated, as illustrated by the long length of the branches (proportional to the distance between them), indicating the absence or scarcity of common residues. On the other hand, pockets within a cluster are separated by very small distances, indicating that they are composed mainly of similar residues. This proximity of intra-cluster pockets confirms that the pockets of a cluster correspond to a same binding site.

The variability of the pockets of a cluster in terms of residues allow the description of the flexibility of the binding site along the dynamics. The heatmap in Figure S1 presents the classification of the pockets and their associated residues. It shows, for instance, no common residue between clusters I and II. One pocket of each of these three clusters is represented in Figure 2B. This figure illustrates how pockets associated with clusters I and II are located at the opposite side of the groove; however, the three clusters are observed in the four structures. The histogram in Figure 2C shows the equilibrium representation of the three binding sites in a quantified way on the four distinct structures. However, we can observe a higher frequency of binding sites I and II in the closed form H6N6_c and a higher frequency of the binding site III in the open or semi-open-tending-to-open form (H5N1_o and H6N6_so_o).

As these three binding sites share few common key residues, some site co-occurrence is possible within a groove conformation. However, we found that the majority (85.4%) of the 1102 druggable RBD-grooves included only one of these druggable binding sites (on the same sampled conformation), with a similar frequency for the four NS1 structures. Two concomitant binding sites were observed in only 13.9% of the time-sampled druggable RBD-grooves, while 0.7% of them simultaneously exhibited one druggable pocket in each of the three sites (for an example, see Figure S2).

The three binding sites are observed on the four structures, as illustrated on Figure 2B. Thus, it can be concluded that three identified druggable binding sites are similarly observed on the four NS1 subtypes. These three sites appear repeatedly and frequently throughout the 150ns MD simulations obtained from the four different NS1 structures considered. Pockets are regularly observed along the MD simulations, as illustrated in Figure S3, on a sample of 100 conformations of each trajectory.

3.3. Characterization of Druggable Binding Sites of the Groove

Each binding site can be described through the analysis of its associated cluster of pockets. The size of the cluster provides the occurrence of a binding site on the 2000 conformations. The binding sites I and II are observed in at least 17.3%, while the most frequent binding site, III, is observed in 64.1% of the 1272 druggable groove-pockets, i.e., in 40% of the time-sampled conformations (Table 1). Binding sites I and II correspond to rather small pockets, and they include, on average, 11.7 $\pm$ 2.6 and 13.3 $\pm$ 4.9 residues per pocket, associated with an average druggable score of 66% $\pm$ 12 and 68% $\pm$ 13, respectively. The most frequent binding site, III, corresponds to larger pockets of 31 $\pm$ 7.1 residues on average, associated with an average druggability score of 71% $\pm$ 13.

The variability analysis of the residue of the pocket cluster allows for the study of the deformability of the site. The most frequent residues of the pocket cluster characterized the key residues of the binding site. The residue frequency of the three pocket clusters is summarized in histograms, with the proportion of the four initial structures indicated in Figure 3. Each binding site includes stable and frequent residues, observed in 75% of the pocket cluster. These frequent residues are considered as key residues that critically determine the druggability of the binding site. Each pocket cluster also includes some residues that are much less frequently observed (<25% of the pockets cluster). These residues appear anecdotally in the binding site, when we consider its deformability along the simulations. The three sites were regularly observed in the MD simulations of the four different structures (Figure S3). These analyses confirm that not only the key residues, but also most of the associated residues, are observed on the four different structures. The key-residues of the binding sites are listed in Table 2.

Binding sites I and II include height key-residues. The height key-residues of site I are shown in Figure 4A. Binding site II exhibits eights similar key-residues, but on the opposite chains (Figure 4B). These two sites are located at the two opposite extremities of the groove. The presence of three aromatic or aliphatic key-residues (PHE, PRO and LEU) likely accounts for the high druggability score of their relatively small pockets. We

also notice that these sites involve two polar amino acids (THR, SER) and two charged key-residues (ARG and ASP) (Table 2).

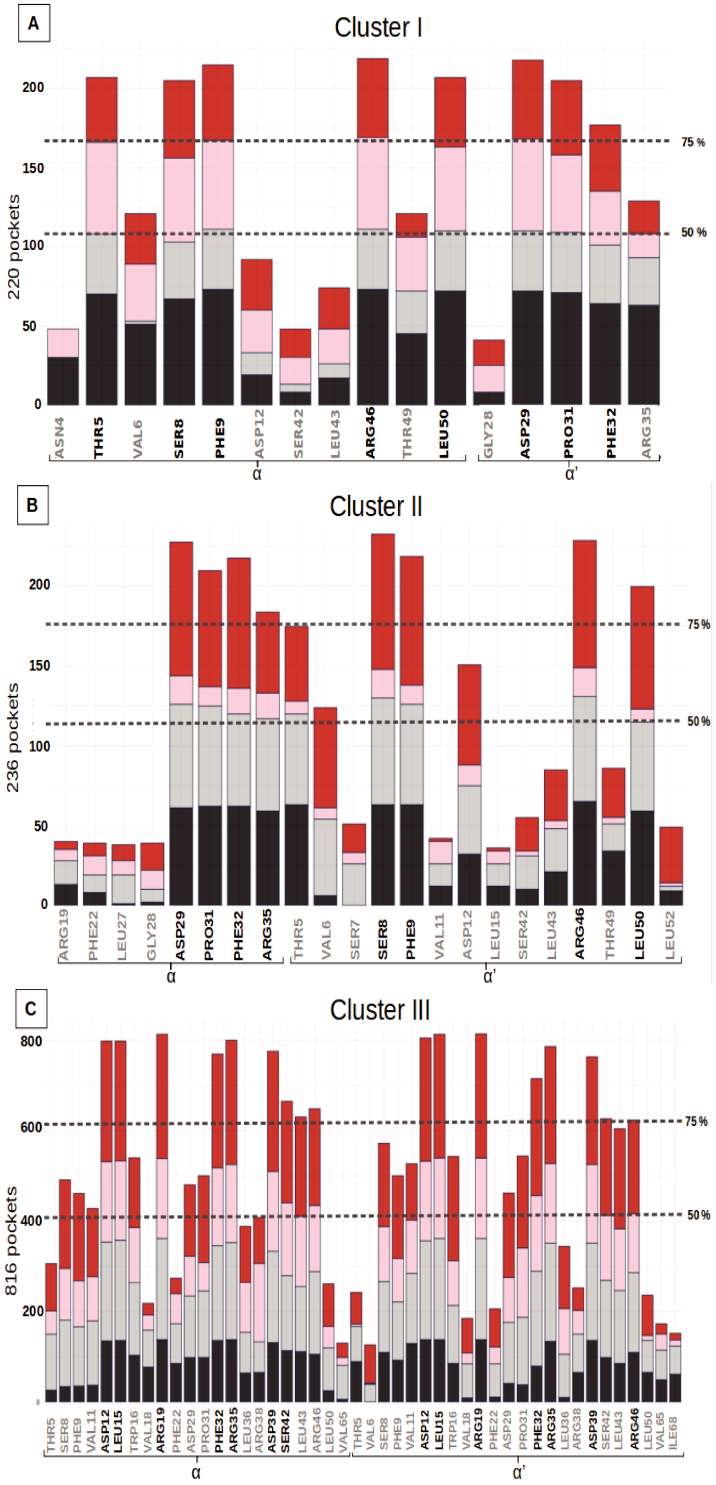


Figure 3. Residue frequency histograms of the druggable groove-pockets (for residues in more than 15% of the pockets) for the three clusters: (**A**) Cluster I, (**B**) Cluster II, (**C**) Cluster III. The bars in the histograms represent the distribution of the number of pockets, including a specific residue for each of the structures H6N6_c, H6N6_so_o, H1N1_so_c, H5N1_o shown in black, gray, pink and red respectively. Two dotted lines indicate the residues seen in more than 50% and 75% of the pockets in the class studied, and the residues selected as key residues in each class are shown in black on the X axis.

Table 2. Table of the three groove-druggable binding sites (I, II, III). Total pocket number observed in different clusters or binding sites. Average druggability score of pockets associated with each binding site and its corresponding standard deviation (sd). Key residues observed in more than 75% of the cluster of druggable pockets for each binding site are listed.

Pocket Clusters	Pocket Number	Druggability Score Mean (sd)	Key Residues of the Clusters
I	220	0.66 (0.12)	8 residues: α: THR5, SER8, PHE9, ARG46, LEU50 α': ASP29, PRO31, PHE32
II	236	0.68 (0.13)	8 residues: α: ASP29, PRO31, PHE32, ARG35, α': SER8, PHE9, ARG46, LEU50
III	816	0.71 (0.13)	14 residues: α: ASP12, LEU15, ARG19, PHE32, ARG35, ASP39, SER42 α': ASP12, LEU15, ARG19, PHE32, ARG35, ASP39, ARG46
Total	1272	0.68 (0.12)	

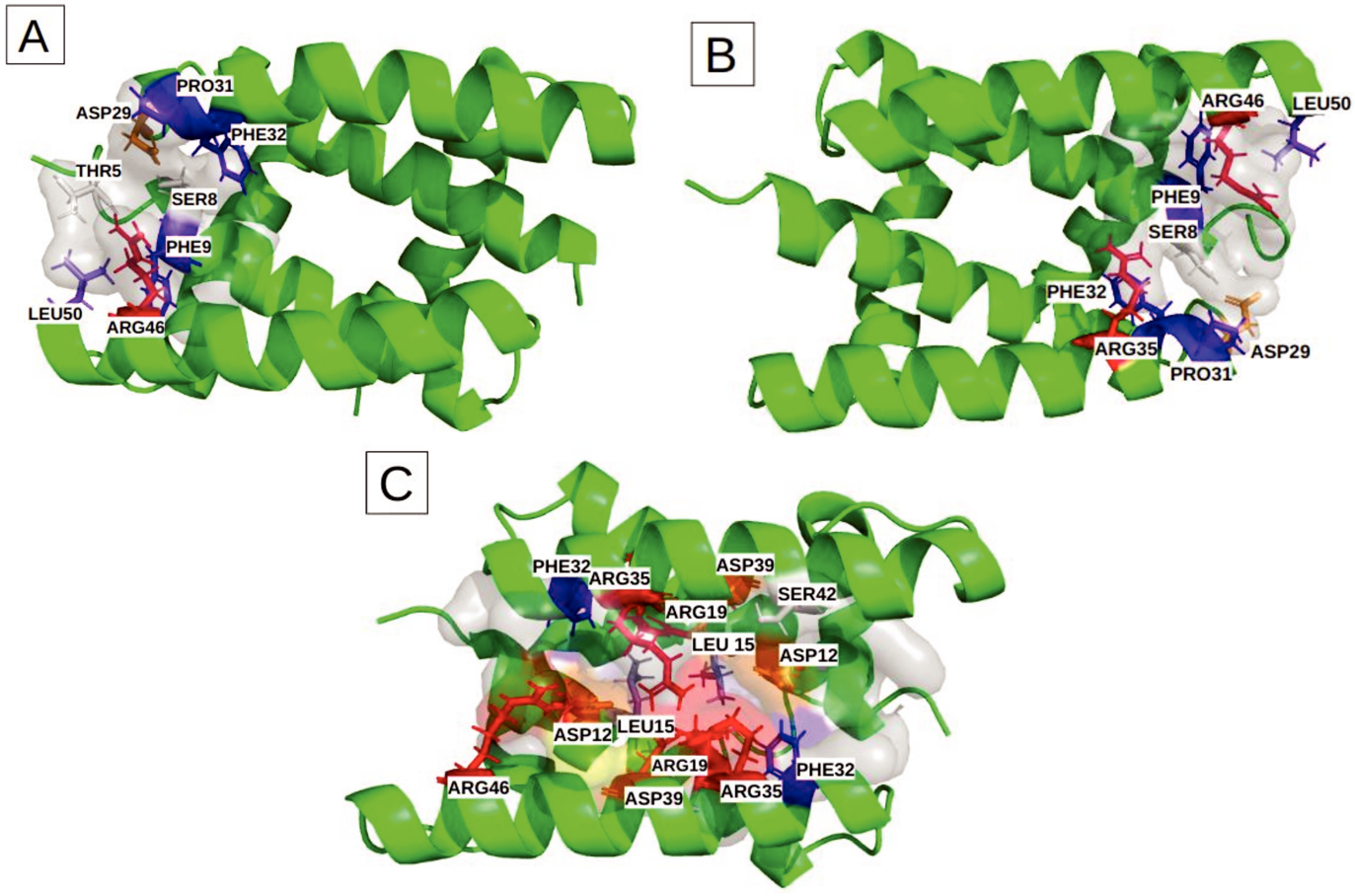


Figure 4. Characterization of the physico-chemical properties of the three druggable groove-pockets identified according to the classification into three clusters: the most frequent residues are colored in red and orange for the positively and negatively charged ones, in blue and slate-blue for the aromatic and aliphatic ones, and in white for the polar. (**A**) is cluster I, which is a small pocket at the end of the groove, (**B**) is cluster II, which is a small pocket located at the opposite side of cluster I. (**C**) is a large pocket identified on cluster III, located in the center of the groove, recovering a large groove interface.

The most frequent binding site, III, corresponds to larger pockets (31 residues on average) that involve 14 key residues. Their corresponding pockets are symmetrical, according to the alpha helices 2 and 2′, and they are located in the center of the groove (Figure 4C). The high druggability score of 71% ± 13 likely relies on (i) its large size, (ii) the presence of two aromatic (PHE) and two aliphatic (LEU) residues and (iii) one polar amino acid (SER). We noticed also the presence of nine charged key-residues (ARG and ASP) bordering the pockets.

3.4. Identification of a Large Common Binding Site

We then set out to further analyze binding site III. Because of its large size, its accessibility and its central position in the middle of the groove, along with its high druggability, this binding site appears the most suitable one to target with a drug-design approach. A small compound occupying this site is likely to prevent or to disrupt the NS1s interaction with RNAs. This site III very frequently adopts a druggable conformation, and involves a large number of residues, including the 14 key residues listed in Table 2.

It is generally recognized that pockets with at least 14 residues, and exceeding the volume of 600 $Å^3$, are the most suitable for binding drug-like molecules [36]. Based on this rationale, we decided to deepen our analysis on the pockets of cluster III, which contain simultaneously 14 key residues. This results in a sub-cluster of 424 pockets, observed in more than 38.3% of the 1102 druggable RBD-grooves (i.e., in 21.2% of the 2000 time-sampled conformations). This sub-cluster of pockets is found in the different subtypes with a similar distribution: 21.8%, 27.5%, 26.4% and 33.7% for the four forms, respectively. The corresponding binding site is called a "common binding site" (Figure 5).

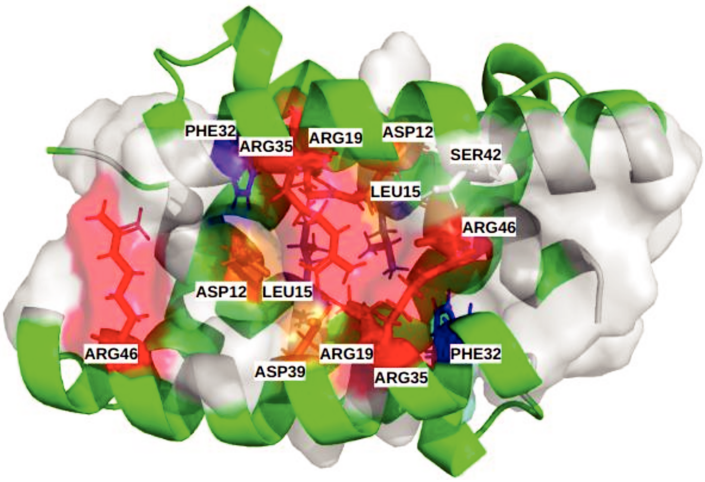


Figure 5. Illustration of the common binding site of the different subtypes, including the 14 most frequent residues (α [ASP12, LEU15, ARG19, PHE32, ARG35, ASP39, SER42); α′ [ASP12, LEU15, ARG19, PHE32, ARG35, ASP39, ARG46]). Physico-chemical properties of the corresponding residues colored red and orange for the positively and negatively charged ones, respectively, blue and slate-blue for the aromatic and aliphatic ones, respectively, and white for polar.

Figure S1 shows the hierarchical classification of the heatmap of the 424 "common binding site" pockets, in the MD simulations with information on the residues involved and the respective extracted groove-pocket numbers. The supplementary band with four colors (black, grey, pink and red) indicates the structures (H6N6_c, H6N6_so_o, H1N1_so_c and H5N1_o, respectively), from which the pockets are extracted.

The preference of some moderately frequent residues for one of the four structures can be noted. For example, residues ASN4, PHE22, ASP24, GLN25, GLU26, LEU27, GLY29 are predominantly observed in pockets associated with the semi-open form with close tendency (H1N1_so_c). The residues PHE14, ASP34, ILE68, LEU69 are mainly observed for pockets associated with the H6N6_so_o and H5N1_o forms.

Finally, this analysis confirms the existence of a broad-consensus druggable binding site (with a common base of 14 key-residues), present in all viral subtypes and whatever the NS1 forms and the length of the linker. Its associated pockets tend to be polar, charged, moderately hydrophobic and composed of a large number of residues.

4. Discussion

In this study, we focused on the selection and characterization of pockets on the RNA-binding domain of the influenza A protein NS1, using conformations that were time-sampled during four parallel 150 ns molecular dynamics simulations from four distinct initial structures of full-length NS1. The application of unsupervised methods such as the hierarchical clustering of pockets allowed for the identification of the main druggable binding sites of the groove region. The study of binding-site flexibility and its variability in terms of residues allow for the description of its deformability during MD simulations for the different forms of NS1. It confirms that, despite the stability of the RBD, the MD simulation induces local movements of the side chains of the residues involved in the groove region, and impacts the RNA binding site. Indeed, MD simulations, by probing the flexibility of the initial crystal structures, have revealed the druggability of the groove-pockets which were below the druggability threshold in three of the four static crystallographic structures used in this study [37].

We identified, through the hierarchical classification of the pockets according to their composition in terms of residues, three main pocket clusters. Two of these (binding sites I and II) are rather small, and less frequently in a druggable conformation. Sites I and II are located at opposite extremities of the groove formed by the α2 and α2′ helices. In contrast, binding site III is much larger, and much more frequently in a druggable conformation. This pocket, at the center of the groove, could be used as a drug binding site that is able to host a wider variety of ligands, and would be more likely to prevent RNA-RBD interaction. MD simulations of the four structures (H6N6_c, H6N6_so_o, H1N1_so_c, H5N1_o) revealed that this central pocket frequently adopted a druggable conformation, thanks to the involvement of its 14 most frequent residues. This accessible central pocket, which is shared by the four structures, has a high therapeutic potential. This common pocket could be targeted by drug-design approaches in order to identify small compounds that may prevent the interaction of NS1 with RNAs, and thereby inhibit its activities, regardless of sequence variations of the protein.

In some protein–ligand complexes, the variation has been shown to correlate with the electrochemical characteristic of the ligand molecules. To achieve binding, proteins have been observed to engage in subtle balances between electrostatic and hydrophobic interactions, to generate stabilizing binding-free-energies [38]. Physico-chemical complementarity is generally considered to be the driving force behind molecular bonding. Complementarity of electrostatic potentials, for example, is considered to be the force that attracts the ligand from the solvent to the binding site [39]. The hydrophobicity of the binding pocket is generally correlated with the properties of the ligands in protein–ligand complexes. The hydrophobic parts of the ligand often interact with the hydrophobic parts of the protein [40]. Looking more closely at the physico-chemical properties of the groove-pockets identified at the RBD–RNA interface, we noticed the presence of several positively and negatively charged residues, such as (ARG19, ARG35, ARG46) and (ASP12, ASP29, ASP39). The presence of charged residues promotes the drug-like molecules binding in the pockets, mainly through the formation of hydrogen bonds or salt bridges. For example, arginine residues play a role in partially neutralizing the ligand charges in the binding pocket, allowing ligand to enter. Positive electrostatic patches above and below the binding entry also contribute to the main attractive forces that lead the drug to the protein surface near the binding site [41].

Hydrophobic residues are also identified (PHE9, LEU15, PRO31, PHE32, LEU50) as well as a few polar residues (THR5, SER8, SER42). Studies have shown that there is a high correlation between the hydrophobicity of the molecular environments and

the experimentally determined desolvation energies. Similar to the cores of proteins, the binding sites of small molecules are often made of hydrophobic residues that could positively contribute to the binding of organic molecules in aqueous environments, and are involved in maintaining a stable 3D arrangement of the ligand in the binding pocket [42–44].

The polar and positively charged amino acids such as SER and ARG, which form strong ionic hydrogen bonds (salt bridges), are known to be the most favorable contacts for protein–RNA interaction [45–47]. Recently, Wacquiez et al., 2020, in their study of the NS1 structure and sequence determinants governing the interactions of RNAs in the case of H7N1 (PDB: 6SX2), described how the RBD-groove aligns itself almost perfectly in parallel to the RNA, to bind it. These authors underlined several RBD-groove residues involved in the RNA interaction [48]. Our three identified binding-sites include six residues (THR5, ASP29, PRO31, ARG35, SER42, ARG46) of the ten described by Cheng et al., 2008, (THR5, ASP29, ASP34, ARG35, ARG38, SER42, ARG46, THR49), as directly involved in the interaction with RNA; the limitation of their study lies in the fact that they did not take into account the NS1 polymorphism [49]. Our study confirms the presence of this binding site in the different strains, while taking into account the NS1 polymorphism. Five other contacts were described by Wacquiez et al., 2020, (PRO31, ARG35, ARG37, ARG38, THR49). Two of these residues, ARG38 and THR49, also detected in the RNA interactions, are partly observed (in approximately 50% of cases) in the pocket clusters of binding sites I and III.

Although the three binding sites are generally identified separately on the conformations, we were able to observe the presence of the three pockets simultaneously on a single conformation (Figure S2). The three pockets can therefore coexist, but in a rarer way (0.7%), revealing a much larger interaction interface with RNA.

The results of our work, as well as the literature elements, confirm our hypothesis that the RNA binding domain of the NS1 protein is a promising therapeutic target, firstly because the frequent druggable pockets identified at the RBD interface show favorable physico-chemical and geometrical properties for hosting a drug candidate, and secondly because this druggable region is identified in the different subtypes (H6N6, H1N1 and H5N1) of the influenza A virus.

5. Conclusions

Our study shows the relevance of combining molecular dynamics simulations with the monitoring of pockets and the prediction of their druggability. Indeed, with this approach, we were able to identify three well-characterized binding sites in the groove region of the NS1 RBD and describe their variability. Unsupervised classification allowed us to confirm the presence of these binding sites across the four distinct structures representing three sequence variants of NS1. We also noticed that the key residues involved in the high druggability scores were highly or strictly conserved across the four NS1s from the subtypes H6N6, H1N1 and H5N1.

In this paper we worked on four full-length structures of the NS1 protein of the influenza A virus. These four structures were co-crystallized in different forms (closed, semi-open with closed tendency, semi-open with open tendency, and open shape). In our previous study, we were able to demonstrate that the RBD of the NS1 protein was highly stable during molecular dynamics simulations, regardless of the initial form of the structure. Our work therefore consists of searching for protein cavities likely to host a drug candidate capable of inhibiting the interaction between the RBD domain of the NS1 protein and RNA on these different structures and subtypes. In order to verify whether there is a druggable binding site on their surface, and what its properties would be, we coupled molecular dynamics simulations, pocket search, druggability prediction, and statistical approaches.

According to our previous study in Naceri et al., 2022, NS1 can adopt its various forms, regardless of its sequence that characterize its plasticity. Here, we were also able to extract three main druggable binding sites located at the groove (between α-helices 2 and 2′) on all four structures. More than 55% of the groove is druggable, whatever the structural polymorphism. The four structures we have identified have a large binding site

in the center of the groove that is highly druggable and involves 14 key residues, some of them being directly involved in RNA interaction. In more than 20% of the time-sampled conformations in the MD simulations, this site has a high druggability score. This confirms the NS1 RBD groove as a good therapeutic target, as it would be able to host a universal therapy effective on the different subtypes.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/biom13010064/s1. Table S1. Summary of the four structures (H6N6_c, H6N6_so_o, H1N1_so_c, H5N1_o), corresponding Uniprot (Q20NS3, P03496, A5A5U1) and PDB Identifiers (4OPA, 4OPH, 5NT2, 3F5T) together with information of forms (close, semi-open with close tendency/open tendency and open), linker size, and details of reversed mutations in the homology modeling Figure S1. Heatmap of the 424 common binding-site pockets of the four structures H6N6_c, H6N6_so_o, H1N1_so_c and H5N1_o colored in black, gray, pink and red respectively. The Y axis corresponds to the list of residues that compose the pockets in order of appearance from top to bottom on the two chains, in succession. Their 14 key residues are indicated with cyan arrows on the Y axis enclosed by the chains α and α' (α [ASP12, LEU15, ARG19, PHE32, ARG35, ASP39, SER42); α' [ASP12, LEU15, ARG19, PHE32, ARG35, ASP39, ARG46]). The X axis lists the druggable groove-pockets clustered according to their similarity in terms of residue composition. Each column represents a pocket and the residues it contains. Figure S2. Representation of one rare case of NS1 conformation (conformation 133 of 4OPA MD simulations) that visits simultaneously the three identified groove binding sites. The blue pocket corresponds to binding site I, the orange to binding site II and the yellow to binding site III. (A) Full-length representation (B) RBD representation, (C) Coloring of one pocket representative of each of the three clusters colored according to the chemical nature of its amino acids: positively charged, negatively charged, polar and hydrophobic amino acids are respectively indicated in in red, blue, white and purple. Figure S3. Dotplot of three identified clusters (Y axis) observed per conformation for a set of 100 conformations (X axis) sample on molecular dynamics simulations of the four structures, H6N6_c, H6N6_so_o, H1N1_so_c and H5N1_o. For each conformation (X axis), the presence of a pocket belonging to clusters I, II and III is indicated by a blue, orange and yellow square, respectively.

Author Contributions: S.N.: conceptualization, methodology, formal analysis, investigation, writing—original draft preparation, writing—review and editing. D.M.: writing—review and editing, project administration, funding acquisition. R.B.: methodology, investigation. D.F.: conceptualization, methodology, investigation, writing—original draft preparation, writing—review and editing, supervision, project administration, funding acquisition. A.-C.C.: conceptualization, methodology, investigation, writing—original draft preparation, writing—review and editing, supervision, project administration, funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding: The authors gratefully acknowledge the financial support of the doctoral school "Pierre Louis de santé publique", the Université Paris Cité, the CNRS institute, and the INSERM institute.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

Acknowledgments: We would like to thank Ines-Sabine RAHALI and Hanane ARKOUB for their useful review and the doctoral school "Pierre Louis de santé publique".

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhao, M.; Wang, L.; Li, S. Influenza A Virus–Host Protein Interactions Control Viral Pathogenesis. *Int. J. Mol. Sci.* **2017**, *18*, 1673. [CrossRef] [PubMed]
2. Influenza (Seasonal). 2018. Available online: https://www.who.int/news-room/fact-sheets/detail/influenza-(seasonal) (accessed on 20 December 2021).

3. Thompson, W.W.; Weintraub, E.; Dhankhar, P.; Cheng, P.Y.; Brammer, L.; Meltzer, M.I.; Bresee, J.S.; Shay, D.K. Estimates of US influenza-associated deaths made using four different methods. *Influenza Other Respir. Viruses* **2009**, *3*, 37–49. [CrossRef] [PubMed]
4. Fukuyama, S.; Kawaoka, Y. The pathogenesis of influenza virus infections: The contributions of virus and host factors. *Curr. Opin. Immunol.* **2011**, *23*, 481–486. [CrossRef] [PubMed]
5. Hale, B.G.; Steel, J.; Medina, R.A.; Manicassamy, B.; Ye, J.; Hickman, D.; Hai, R.; Schmolke, M.; Lowen, A.C.; Perez, D.R.; et al. Inefficient Control of Host Gene Expression by the 2009 Pandemic H1N1 Influenza A Virus NS1 Protein. *J. Virol.* **2010**, *84*, 6909–6922. [CrossRef] [PubMed]
6. Engel, D.A. The influenza virus NS1 protein as a therapeutic target. *Antivir. Res.* **2013**, *99*, 409–416. [CrossRef]
7. Rosário-Ferreira, N.; Preto, A.J.; Melo, R.; Moreira, I.S.; Brito, R.M.M. The Central Role of Non-Structural Protein 1 (NS1) in Influenza Biology and Infection. *Int. J. Mol. Sci.* **2020**, *21*, 1511. [CrossRef] [PubMed]
8. Talon, J.; Horvath, C.M.; Polley, R.; Basler, C.F.; Muster, T.; Palese, P.; García-Sastre, A. Activation of Interferon Regulatory Factor 3 Is Inhibited by the Influenza A Virus NS1 Protein. *J. Virol.* **2000**, *74*, 7989–7996. [CrossRef]
9. García-Sastre, A. Induction and evasion of type I interferon responses by influenza viruses. *Virus Res.* **2011**, *162*, 12–18. [CrossRef]
10. Marc, D.; Barbachou, S.; Soubieux, D. 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.* **2012**, *41*, 434–449. [CrossRef]
11. Berman, H.M. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242. [CrossRef]
12. Carrillo, B.; Choi, J.M.; Bornholdt, Z.A.; Sankaran, B.; Rice, A.P.; Prasad, B.V.V. The Influenza A Virus Protein NS1 Displays Structural Polymorphism. *J. Virol.* **2014**, *88*, 4113–4122. [CrossRef] [PubMed]
13. Hale, B.G. Conformational plasticity of the influenza A virus NS1 protein. *J. Gen. Virol.* **2014**, *95*, 2099–2105. [CrossRef] [PubMed]
14. Aramini, J.M.; Ma, L.C.; Zhou, L.; Schauder, C.M.; Hamilton, K.; Amer, B.R.; Mack, T.R.; Lee, H.W.; Ciccosanti, C.T.; Zhao, L.; et al. Dimer Interface of the Effector Domain of Non-structural Protein 1 from Influenza A Virus. *J. Biol. Chem.* **2011**, *286*, 26050–26060. [CrossRef]
15. Naceri, S.; Marc, D.; Camproux, A.C.; Flatters, D. Influenza A Virus NS1 Protein Structural Flexibility Analysis According to Its Structural Polymorphism Using Computational Approaches. *Int. J. Mol. Sci.* **2022**, *23*, 1805. [CrossRef] [PubMed]
16. Trigueiro-Louro, J.; Santos, L.A.; Almeida, F.; Correia, V.; Brito, R.M.; Rebelo-de-Andrade, H. NS1 protein as a novel anti-influenza target: Map-and-mutate antiviral rationale reveals new putative druggable hot spots with an important role on viral replication. *Virology* **2022**, *565*, 106–116. [CrossRef]
17. Trigueiro-Louro, J.M.; Correia, V.; Santos, L.A.; Guedes, R.C.; Brito, R.M.; Rebelo-de-Andrade, H. To hit or not to hit: Large-scale sequence analysis and structure characterization of influenza A NS1 unlocks new antiviral target potential. *Virology* **2019**, *535*, 297–307. [CrossRef]
18. Darapaneni, V.; Prabhaker, V.K.; Kukol, A. Large-scale analysis of influenza A virus sequences reveals potential drug target sites of non-structural proteins. *J. Gen. Virol.* **2009**, *90 Pt 9*, 2124–2133. [CrossRef] [PubMed]
19. Yin, C.; Khan, J.A.; Swapna, G.V.; Ertekin, A.; Krug, R.M.; Tong, L.; Montelione, G.T. Conserved surface features form the double-stranded RNA binding site of non-structural protein 1 (NS1) from influenza A and B viruses. *J. Biol. Chem.* **2007**, *282*, 20584–20592. [CrossRef]
20. Abi Hussein, H.; Geneix, C.; Petitjean, M.; Borrel, A.; Flatters, D.; Camproux, A.C. Global vision of druggability issues: Applications and perspectives. *Drug Discov. Today* **2017**, *22*, 404–415. [CrossRef]
21. Kuzmanic, A.; Bowman, G.R.; Juarez-Jimenez, J.; Michel, J.; Gervasio, F.L. Investigating Cryptic Binding Sites by Molecular Dynamics Simulations. *Acc. Chem. Res.* **2020**, *53*, 654–661. [CrossRef]
22. Abi Hussein, H.; Geneix, C.; Cauvin, C.; Marc, D.; Flatters, D.; Camproux, A.C. Molecular Dynamics Simulations of Influenza A Virus NS1 Reveal a Remarkably Stable RNA-Binding Domain Harboring Promising Druggable Pockets. *Viruses* **2020**, *12*, 537. [CrossRef] [PubMed]
23. Sievers, F.; Wilm, A.; Dineen, D.; Gibson, T.J.; Karplus, K.; Li, W.; Lopez, R.; McWilliam, H.; Remmert, M.; Söding, J.; et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **2011**, *7*, 539. [CrossRef] [PubMed]
24. Rice, P.; Longden, I.; Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* **2000**, *16*, 276–277. [CrossRef] [PubMed]
25. UniProt. (n.d.). Available online: https://www.uniprot.org/ (accessed on 28 September 2021).
26. De Lano, W.L. The PyMOL Molecular Graphics System. 2002. Available online: www.pymol.org (accessed on 16 October 2022).
27. Koliopoulos, M.G.; Lethier, M.; van der Veen, A.G.; Haubrich, K.; Hennig, J.; Kowalinski, E.; Stevens, R.V.; Martin, S.R.; Reis e Sousa, C.; Cusack, S.; et al. Molecular mechanism of influenza A NS1-mediated TRIM25 recognition and inhibition. *Nat. Commun.* **2018**, *9*, 1820. [CrossRef] [PubMed]
28. Bornholdt, Z.A.; Prasad, B.V.V. X-ray structure of NS1 from a highly pathogenic H5N1 influenza virus. *Nature* **2008**, *456*, 985–988. [CrossRef] [PubMed]
29. Abraham, M.J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J.C.; Hess, B.; Lindahl, E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *1–2*, 19–25. [CrossRef]
30. Schmidtke, P.; Le Guilloux, V.; Maupetit, J.; Tuffery, P. fpocket: Online tools for protein ensemble pocket detection and tracking. *Nucleic Acids Res.* **2010**, *38*, W582–W589. [CrossRef]

31. Borrel, A.; Regad, L.; Xhaard, H.; Petitjean, M.; Camproux, A.C. PockDrug: A Model for Predicting Pocket Druggability That Overcomes Pocket Estimation Uncertainties. *J. Chem. Inf. Model.* **2015**, *55*, 882–895. [CrossRef]
32. Murtagh, F.; Legendre, P. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *J. Classif.* **2014**, *31*, 274–295. [CrossRef]
33. Ward, J.H. Hierarchical Grouping to Optimize an Objective Function. *J. Am. Stat. Assoc.* **1963**, *58*, 236–244. [CrossRef]
34. Hahsler, M.; Hornik, K.; Buchta, C. Getting Things in Order: An Introduction to the *R* Packageseriation. *J. Stat. Softw.* **2008**, *25*, 1–34. [CrossRef]
35. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2020; Available online: https://www.R-project.org (accessed on 16 October 2022).
36. Bromley, D.; Bauer, M.R.; Fersht, A.R.; Daggett, V. An in silico algorithm for identifying stabilizing pockets in proteins: Test case, the Y220C mutant of the p53 tumor suppressor protein. *Protein Eng. Des. Sel.* **2016**, *29*, 377–390. [CrossRef] [PubMed]
37. Stank, A.; Kokh, D.B.; Fuller, J.C.; Wade, R.C. Protein Binding Pocket Dynamics. *Acc. Chem. Res.* **2016**, *49*, 809–815. [CrossRef] [PubMed]
38. Kahraman, A.; Morris, R.J.; Laskowski, R.A.; Favia, A.D.; Thornton, J.M. On the diversity of physicochemical environments experienced by identical ligands in binding pockets of unrelated proteins. *Proteins Struct. Funct. Bioinform.* **2009**, *78*, 1120–1136. [CrossRef] [PubMed]
39. Livesay, D.R.; Jambeck, P.; Rojnuckarin, A.; Subramaniam, S. Conservation of Electrostatic Properties within Enzyme Families and Superfamilies. *Biochemistry* **2003**, *42*, 3464–3473. [CrossRef]
40. Kahraman, A.; Morris, R.J.; Laskowski, R.A.; Thornton, J.M. Variation of geometrical and physicochemical properties in protein binding pockets and their ligands. *BMC Bioinform.* **2007**, *8* (Suppl. 8), S1. [CrossRef]
41. Huang, H.C.; Briggs, J.M. The association between a negatively charged ligand and the electronegative binding pocket of its receptor. *Biopolymers* **2002**, *63*, 247–260. [CrossRef]
42. Guo, Z.; Li, B.; Cheng, L.T.; Zhou, S.; McCammon, J.A.; Che, J. Identification of Protein–Ligand Binding Sites by the Level-Set Variational Implicit-Solvent Approach. *J. Chem. Theory Comput.* **2015**, *11*, 753–765. [CrossRef]
43. Schmidtke, P.; Barril, X. Understanding and predicting druggability. A high-throughput method for detection of drug binding sites. *J. Med. Chem.* **2010**, *53*, 5858–5867. [CrossRef]
44. Nisius, B.; Sha, F.; Gohlke, H. Structure-based computational analysis of protein binding sites for function and druggability prediction. *J. Biotechnol.* **2012**, *159*, 123–134. [CrossRef]
45. Gupta, A.; Gribskov, M. The Role of RNA Sequence and Structure in RNA–Protein Interactions. *J. Mol. Biol.* **2011**, *409*, 574–587. [CrossRef] [PubMed]
46. Han, K.; Nepal, C. PRI-Modeler: Extracting RNA structural elements from PDB files of protein-RNA complexes. *FEBS Lett.* **2007**, *581*, 1881–1890. [CrossRef] [PubMed]
47. Treger, M.; Westhof, E. Statistical analysis of atomic contacts at RNA-protein interfaces. *J. Mol. Recognit.* **2001**, *14*, 199–214. [CrossRef] [PubMed]
48. Wacquiez, A.; Coste, F.; Kut, E.; Gaudon, V.; Trapp, S.; Castaing, B.; Marc, D. Structure and Sequence Determinants Governing the Interactions of RNAs with Influenza a Virus Non-Structural Protein NS1. *Viruses* **2020**, *12*, 947. [CrossRef]
49. Cheng, A.; Wong, S.M.; Yuan, Y.A. Structural basis for dsRNA recognition by NS1 protein of influenza A virus. *Cell Res.* **2008**, *19*, 187–195. [CrossRef]

Supplementary Materials:

Table S1. Summary of the four structures (H6N6_c, H6N6_so_o, H1N1_so_c, H5N1_o), corresponding Uniprot (Q20NS3, P03496, A5A5U1) and PDB Identifiers (4OPA, 4OPH, 5NT2, 3F5T) together with information of forms (Close, Semi-open with close tendency/open tendency and Open), linker size, and details of reversed mutations in the homology modeling

Structures	IDs	Forms	Reverse mutations details
H6N6_c	Uniprot: **Q20NS3** Xray PDB: 4OPA	Close (dED-ED= 27.8Å) Short linker (Δ 80-84)	R38A/K41A
H6N6_so_o	Uniprot: **Q20NS3** Xray PDB: 4OPH	Semi-open with open tendency (dED-ED=60.0Å) Long linker	R38A/K41A
H1N1_so_c	Uniprot: **P03496** Xray PDB: 5NT2	Semi-open with close tendency (dED-ED= 32.4Å) Long linker	R38A/K41A/W187A
H5N1_o	Uniprot: **A5A5U1** Xray PDB: 3F5T	Close (dED-ED= 66.5Å) Short linker (Δ 80-84)	R38A/K41A

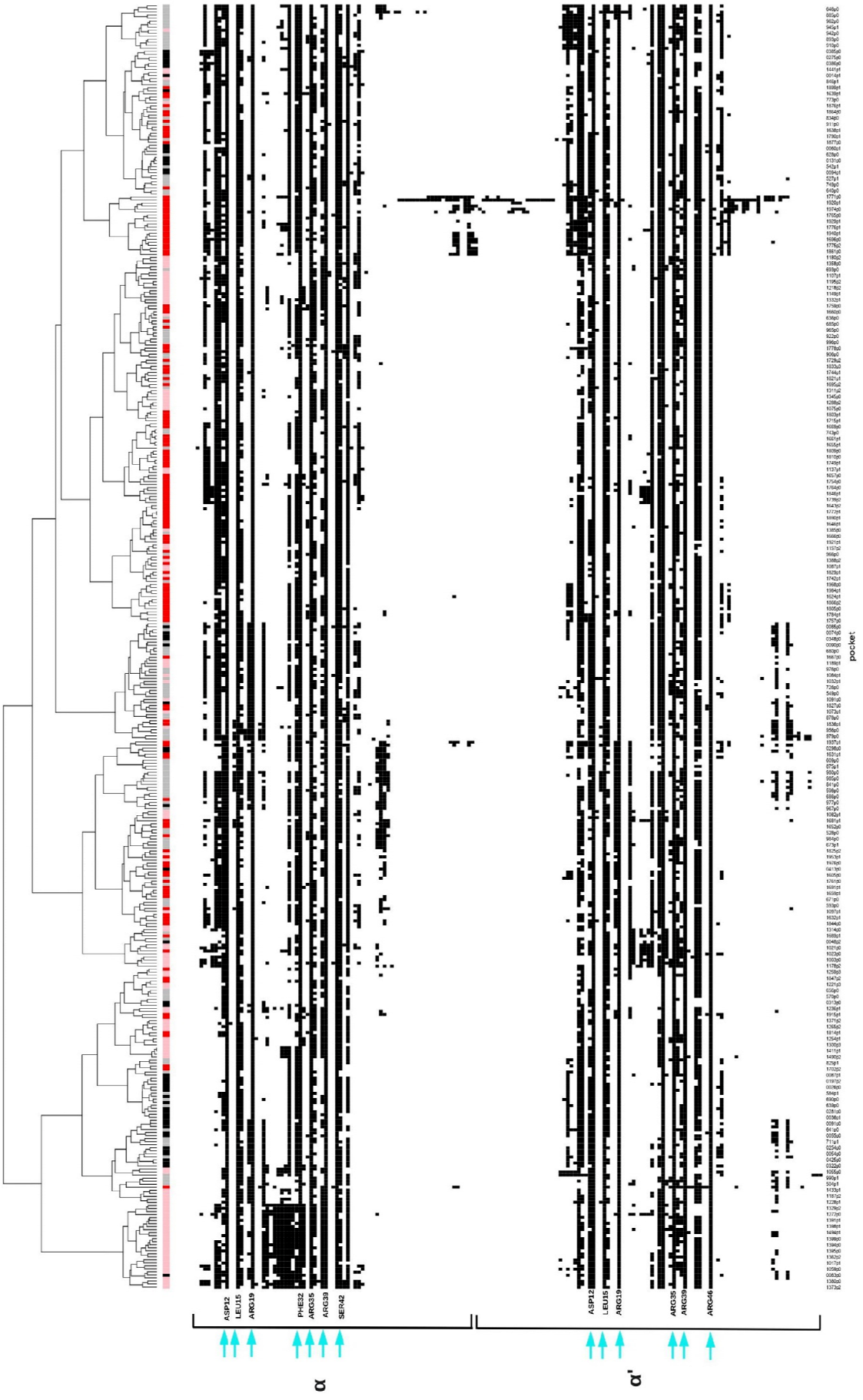

ASP12
LEU15
ARG19
PHE32
ARG35
ARG39
SER42
α
ASP12
LEU35
ARG19
ARG35
ARG39
ARG46
α'
pocket

Figure S1. Heatmap of the 424 common binding site pockets of the four structures H6N6_c, H6N6_so_o, H1N1_so_c and H5N1_o colored in black, gray, pink and red respectively. The Y axis corresponds to the list of residues that compose the pockets in order of appearance from top to bottom on the two chains in succession. Their 14 key-residues are indicated with cyan arrows on the Y axis enclosed by the chains α and α' (α [ASP12, LEU15, ARG19, PHE32, ARG35, ASP39, SER42); α' [ASP12, LEU15, ARG19, PHE32, ARG35, ASP39, ARG46]). The X axis lists the druggable groove-pockets clustered according to their similarity in terms of residue composition. Each column represents a pocket and the residues it contains.

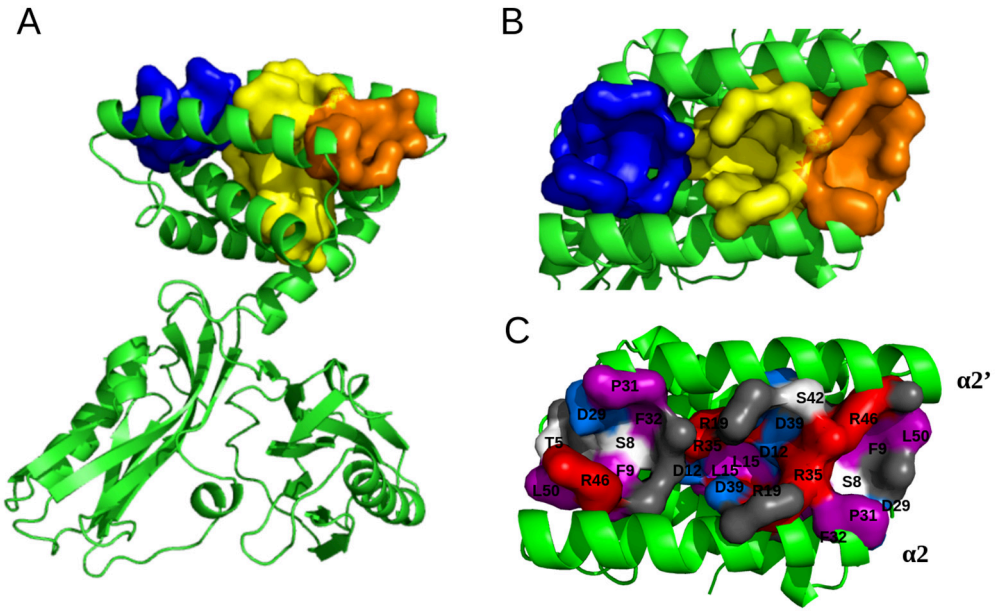


Figure S2. Representation of one rare case of NS1 conformation (conformation 133 of 4OPA MD simulations) that visits simultaneously the three identified groove binding sites. The blue pocket corresponds to the binding site I, the orange to the binding site II and the yellow to the binding site III. (A) Full length representation (B) RBD representation **(C)** Coloring of one pocket representative of each of the three clusters colored by chemical nature of its amino acids: positively charged, negatively charged , polar and hydrophobic amino acids are respectively indicated in in red, blue, white and purple.

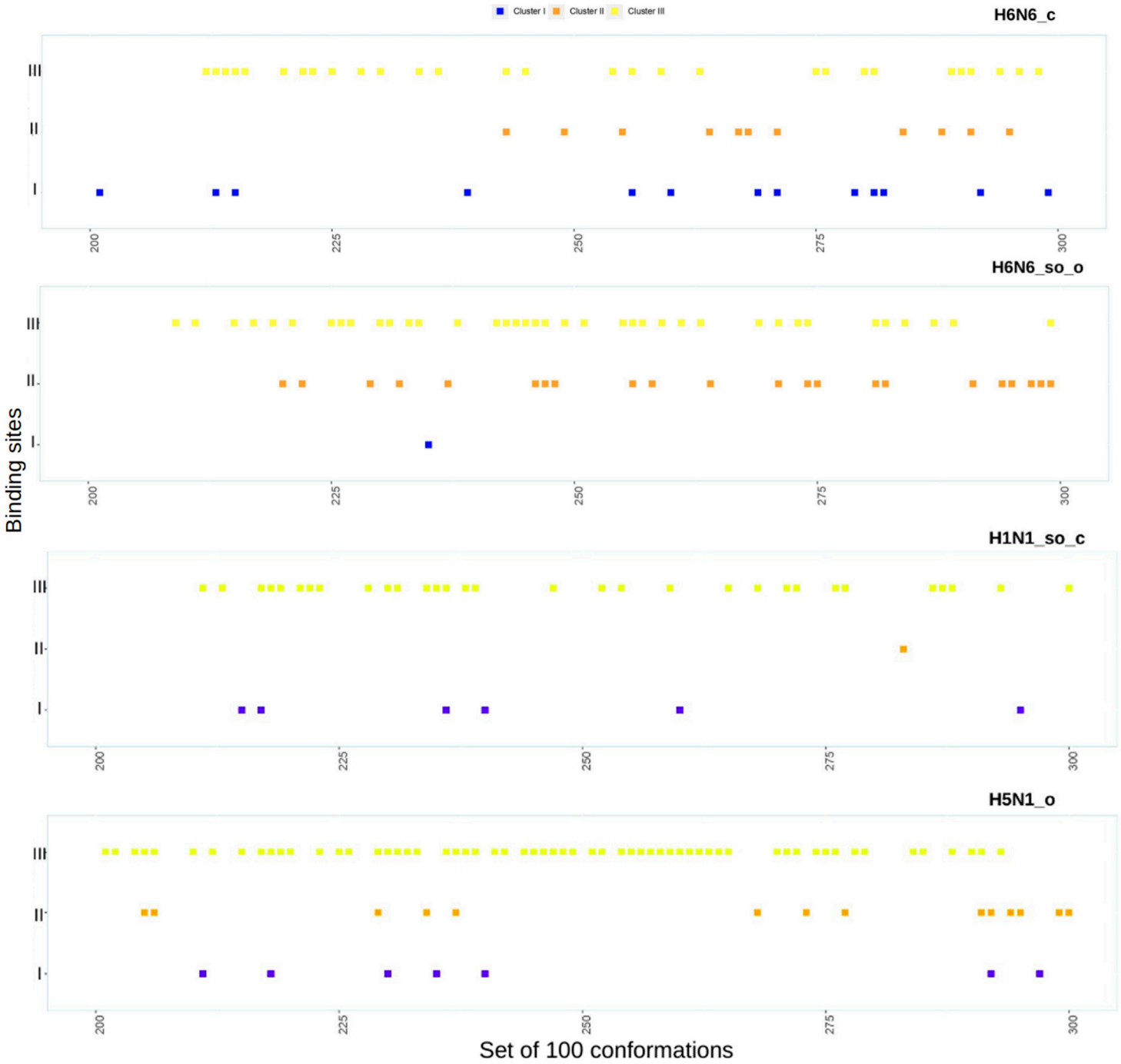


Figure S3. Dotplot of three identified clusters (Y axis) observed per conformation for a set of 100 conformations (X axis) sample on molecular dynamics simulations of the four structures H6N6_c, H6N6_so_o, H1N1_so_c and H5N1_o. For each conformation (X axis), the presence of a pocket belonging to clusters I, II and III is indicated by a respectively blue, orange and yellow square.

CHAPITRE 4. Analyse structurale de la protéine Spike du virus SARS-CoV-2

Découvert en Chine en décembre 2019, le virus SARS-CoV-2 a entraîné près de 760 millions de cas confirmés et 6,8 millions de décès, provoquant une pandémie qualifiée comme telle par l'OMS en mars 2020 (https://covid19.who.int/). Malgré leur caractère révolutionnaire dans le domaine thérapeutique, les anticorps monoclonaux souffrent d'une fabrication coûteuse et d'une méthode d'administration peu agréable pour les patients, nécessitant une injection intraveineuse. Actuellement, il existe quatre médicaments disponibles pour traiter le COVID-19 : le Remdesivir, le Molnupiravir, le Baricitinib et le Paxlovid. Cependant, la recherche se poursuit pour découvrir des petites molécules qui pourraient agir efficacement malgré les mutations du virus. Le domaine trimérique RBD se trouve dans deux états: Il peut être fermé, dans ce cas le virus est inactif, comme il peut être ouvert et permettre au récepteur ACE2 d'interagir avec un de ses monomères, dans ce cas le virus est actif et peut pénétrer dans la cellule hôte.

L'objectif de ce travail est de: i) comprendre la flexibilité du RBD à l'état libre et à l'état complexé avec ACE2. ii) identifier les résidus clés de l'interaction entre les deux partenaires pour proposer des stratégies thérapeutiques ciblées. iii) identifier les cavités à la surface du RBD qui sont les moins impactées par les mutations et qui possèdent les propriétés physico-chimiques, géométriques et de druggabilité les plus favorables à l'accueil de petites molécules thérapeutiques. Cette analyse est schématisée par la Figure 6. Ce travail a abouti à la publication en co-premier auteur suivante:

Ghoula, M., **Naceri, S.**, Sitruk, S., Flatters, D., Moroy, G., & Camproux, A. (2023). Identifying promising druggable binding sites and their flexibility to target the receptor-binding domain of SARS-CoV-2 spike protein. *Computational and structural biotechnology journal.*

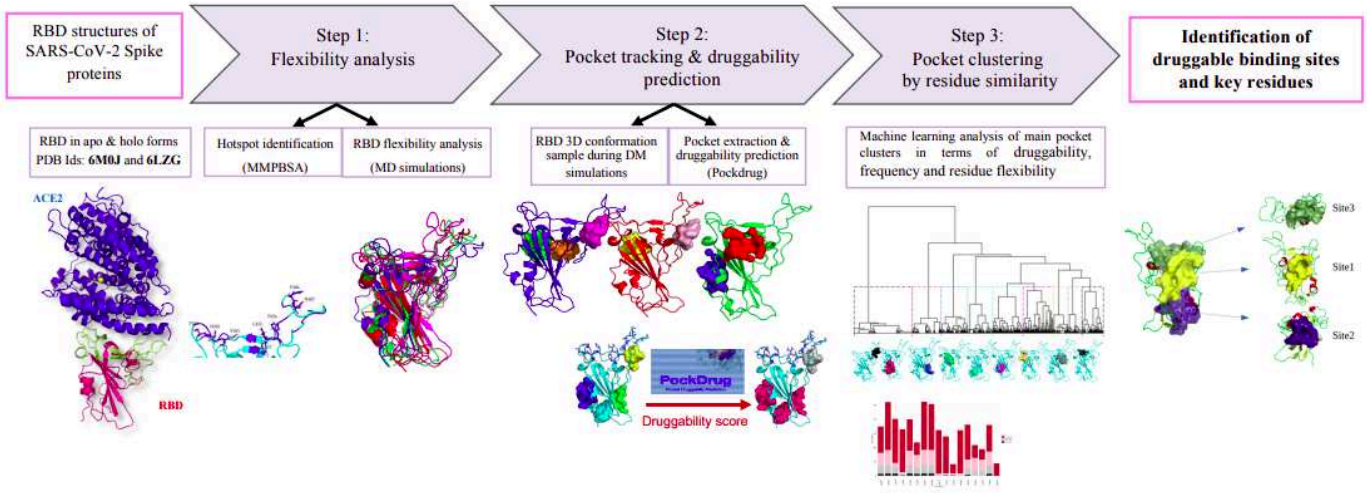


Figure 6: Résumé du protocole d'identification des sites de liaison druggables du domaine de liaison au récepteur du virus SARS-CoV-2 en trois étapes. L'étape 1 consiste en l'analyse de la flexibilité de la protéine à l'aide de simulations de dynamique moléculaire et l'identification des résidus hotspots des interactions ACE2-RBD à l'aide de la méthode MMPBSA. L'étape 2 traite le suivi des poches de surface du RBD et la prédiction de la druggabilité par échantillonnage de 1 000 conformations du RBD extraites des simulations de dynamique moléculaire, l'estimation des poches et la prédiction de la druggabilité à l'aide de PockDrug. L'étape 3 concerne le regroupement hiérarchique de toutes les poches RBD regroupées en terme de similarité de résidus. Cette classification permet l'identification de clusters de poches avec des résidus communs, correspondant

à des sites de liaison fréquemment observés le long des simulations de dynamique moléculaire. La variabilité des résidus au sein d'un cluster de poches illustre la flexibilité des résidus du site de liaison correspondant. Enfin, ce protocole a identifié avec succès trois sites de liaison et leurs résidus clés susceptibles d'être ciblés par des inhibiteurs pour prévenir l'interaction avec ACE2.

4.1 Présentation du domaine RBD de la protéine Spike

La protéine Spike, également connue sous le nom de protéine S, est la protéine de surface majeure du virus SARS-CoV-2, responsable de la maladie COVID-19. Elle est constituée de deux sous-unités reliées par une région de liaison et de plusieurs sous-domaines, dont le plus important est le domaine de liaison au récepteur RBD (Figure 7). Au début de nos travaux très peu de structures 3D de la protéine étaient disponibles, nous avons fait le choix de travailler sur deux structures cristallographiques du complexe RBD-ACE2 (code PDB: 6M0J et 6LZG), ce qui a permis de comprendre comment le virus interagit avec les cellules hôtes.

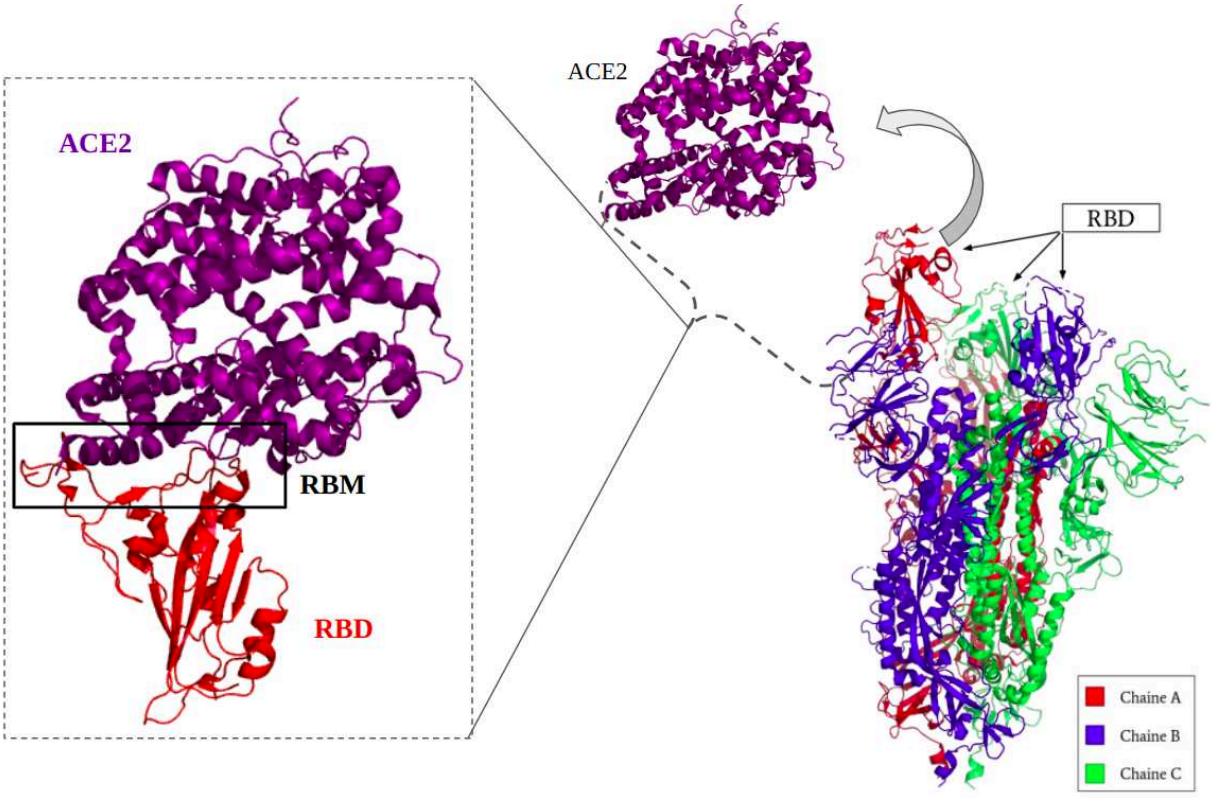


Figure 7: Illustration de la structure du complexe RBD-ACE, constitué de la protéine Spike trimérique (représentée en rouge, bleu et vert) et de la protéine ACE2 (en violet). Dans cette structure, les trois chaînes de la protéine Spike sont dans un état ouvert et actif. Chaque monomère de la protéine Spike possède une région de liaison au récepteur (en rouge) qui peut interagir avec le récepteur ACE2 de la cellule hôte. Cette interaction se produit via la région du motif de liaison au récépteur (RBM) encadrée en noir.

4.2 Flexibilité des structures liées et non liées du RBD (Figure 7, step 1)

Dans cette première étape, des simulations de dynamique moléculaire ont été réalisées sur les structures 6M0J et 6LZG pour étudier le comportement dynamique et la stabilité du RBD dans son état lié et non lié à l'ACE2. Pour cela, 10 simulations indépendantes de 100ns chacune ont été effectuées pour chaque structure, puis les trajectoires ont été concaténées pour former une longue trajectoire de 1µs. Les simulations MD ont permis d'étudier la

flexibilité des RBD avec et sans ACE2, et les deux structures se sont montrées très stables pour le RBD lié et non lié tout au long des trajectoires. Cependant, l'analyse du RMSF a montré que les structures non liées ont des fluctuations plus importantes au niveau du motif de liaison au récepteur (RBM). Cette observation s'explique par le fait que la contrainte exercée par l'ACE2 sur les résidus à l'interface du complexe RBD-ACE2 est relâchée dans les structures non liées.

Une identification des résidus clés de l'interaction entre le RBD et ACE2 a été réalisée par ma collaboratrice en utilisant les méthodes MM-PBSA qui a permis d'identifier 9 résidus hotspot (le protocole n'est pas décrit dans ce manuscrit).

4.3 Suivi des poches du RBD au cours des simulations de DM (Figure 6, step 2)

Dans cette deuxième étape, les poches à la surface du RBD non lié ont été étudiées en utilisant les structures statiques 6M0J et 6LZG disponibles sur la PDB. À l'aide du programme PockDrug, quatre et sept poches ont été estimées respectivement sur les deux structures initiales. Cette estimation concorde avec les résultats obtenus par Trigueiro-Louro et al. sur les mêmes structures PDB, qui ont également observé un nombre variable de poches (Trigueiro-Louro et al., 2020).

Une fois la flexibilité du RBD étudiée, une extraction de 1000 conformations au cours des trajectoires a été réalisée à intervalle régulier, sur lesquelles les poches ont été recherchées. L'objectif de la recherche de poches est d'identifier les régions les plus favorables à la liaison de molécules thérapeutique pour empêcher l'interaction entre le RBD et ACE2 et par conséquent l'entrée du virus dans la cellule hôte. Les résultats obtenus pour les deux structures 6M0J et 6LZG étaient très similaires, c'est pourquoi la suite des travaux a été effectuée uniquement sur 6M0J.

En croisant les résultats obtenus de l'étude MM-PBSA réalisée par ma collaboratrice pour ce travail et ceux de la recherche de poches nous avons constaté que 98,5 % des poches d'interface contenaient au moins l'un des neuf résidus hotspots, ce qui concorde parfaitement avec le fait que ces résidus impliqués dans l'interaction avec ACE2 aient des propriétés physico-chimiques et des emplacements favorables à la formation des poches. Le nombre de poches identifiées sur les 1000 conformations issues de DM du RBD dans l'état non lié à l'ACE2 est représenté dans le tableau 1.

Nombre de poches identifiées sur 1000 conformations (6M0J)	
Nombre de poches total	5065
Nombre de poches à l'interface RBM	1570
Nombre de poches druggable (score de druggabilité ≥ 0.5)	2553
Nombre de poches hautement druggable (score de druggabilité ≥ 0.75)	1037

Tableau 1: Nombre total de poches, localisées au niveau de l'interface RBM, druggable, et hautement druggables identifiées sur les 1000 conformations issues des simulations de dynamique moléculaire du RBD dans l'état non lié à l'ACE2.

Pour étudier l’impact d’ACE2 sur l’interface d’interaction du RBD, une analyse des poches extraites du complexe ACE2-RBD lié a également été effectuée et plus d’un tiers des poches dans la région RBM ont été perdus en raison de la présence d’ACE2.

4.4 Classification des poches du RBD et sélection des sites de liaison druggables (Figure 6, step 3)

Une fois les poches identifiées sur le RBD non lié au cours des simulations de DM, une classification en fonction de la similarité des poches en termes de résidus qui les composent a été réalisée. Cette classification a permis d'identifier des groupes de poches avec des résidus communs correspondant à des sites de liaison fréquemment observés durant les simulations MD. Pendant les simulations de DM, les huit groupes de poches ont été identifiés sur les 1 000 conformations. Quatre des groupes (II, III, V, VII) étaient principalement druggables, tandis que les quatre autres groupes (I, IV, VI, VIII) contenaient un nombre plus faible de poches druggables.

Après analyse de la classification, trois sites de liaison (Site 1, 2, 3) ont été retenus en raison de leur fréquence, de la similarité de leurs résidus, de leur score de druggabilité et de leur emplacement. La classification permet l'identification de clusters de poches avec des résidus communs, correspondant à des sites de liaison fréquemment observés le long des simulations de DM. La variabilité des résidus au sein d'un cluster de poches illustre la flexibilité des résidus du site de liaison correspondant. Enfin, ce protocole a identifié avec succès trois sites de liaison et leurs résidus clés susceptibles d’être ciblés par des inhibiteurs pour prévenir l'interaction avec ACE2.

4.5 Impact des mutations sur le RBD les hotspots et les poches

Les cinq variants (Alpha, Beta, Gamma, Delta et Omicron) ont subi des mutations au niveau du RBD qui augmente leur transmissibilité et leur virulence. Les mutations en question sont les suivantes:

- Alpha: N501Y
- Beta: K417N, E484K, N501Y
- Gamma: K417T, E484K, N501Y
- Delta: L452R et T478K.
- Omicron est concerné par le plus grand nombre de mutations 12 au total (G339D, S371L, S373P, S375F, K417N, N440K, G446S, E448A, S477N, T478K, Q493R, G496S, Q498R, N501Y et Y505H).

Les mutations affectent les résidus hotspots, en particulier les sites 1 et 3, dans les différents variants mis à part les mutations du variant Delta qui ne concerne aucun des sites que nous avons identifiés.

Le variant Omicron est hautement muté, avec 15 résidus mutés dans la région RBD. Le site 1 n’est impacté que par la mutation S375F d'Omicron. Ce résidu n’est impliqué dans aucun contact avec ACE2, ou avec les anticorps, ni impliqué dans la trimérisation du RBD. Sa mutation en F n’a que légèrement modifié la forme et les propriétés hydrophobes. Le site 1 semble être donc relativement bien conservé.

Le site 3 situé à l'interface d'interaction RBD-ACE2 est composé de plusieurs résidus hot spots ; il peut donc être intéressant à cibler par des molécules inhibitrices. Cependant, en raison des mutations induites par les différents variants principalement les mutations K417N, N501Y et E484, le site 3, favorisant l’affinité RBD-ACE2 et l'échappement aux anticorps monoclonaux, demeure difficile à bloquer, contrairement au site 1 qui est faiblement affecté par la seule mutation S375F et le site 2 qui n'est actuellement affecté par aucune mutation.

Ces deux sites peuvent donc être ciblés par des inhibiteurs qui seraient efficaces pour tous les variants actuellement connus.

4.6 Conclusion

Dans cette étude, nous avons examiné la flexibilité du RBD lorsqu'il est non lié ou en interaction avec ACE2 afin d'identifier les sites de liaison susceptibles d'être ciblés par des inhibiteurs. Nous avons mis en place un protocole qui combine les simulations de dynamique moléculaire, l'identification des résidus hot spots de l'interaction RBD-ACE2, la prédiction de la druggabilité des poches et d'apprentissage automatique non supervisée pour regrouper les poches RBD en fonction de leur similarité en termes de résidus. Ce protocole permet d'identifier les sites les plus favorables à l'accueil des molécules médicamenteuses, de les caractériser et de vérifier leur stabilité.

Trois sites de liaison ont été retenus comme prometteurs. Le site 3 dans la région d'interaction RBD-ACE2 est contenant quatre résidus hot spots, ce qui le rend intéressant à cibler pour empêcher l'entrée du virus SARS-CoV-2, mais des mutations limitent l'efficacité des inhibiteurs. Le site 2 est un site connu pour être verrouillé par un acide linoléique, qui maintient le RBD dans sa forme inactive, et il est très prometteur à cibler, car il subit peu de mutations. Quant au site 1, il est hautement druggable et très accessible et n'est affecté par aucune mutation. De plus, cette étude le décrit pour la première fois comme potentiel site de liaison aux médicaments. Ce site est observé sur les monomere RBD à des distances proches les unes aux autres cela suggère que la conception de composés capables de se lier à deux sites 1 dans deux monomères RBD pourrait empêcher le passage de la protéine spike de la forme fermée inactive à la forme ouverte active.

Computational and Structural Biotechnology Journal 21 (2023) 2339–2351

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Identifying promising druggable binding sites and their flexibility to target the receptor-binding domain of SARS-CoV-2 spike protein

M. Ghoula [1], S. Naceri [1], S. Sitruk, D. Flatters, G. Moroy [*,1], A.C. Camproux [*,1]

Université Paris Cité, CNRS, INSERM, Unité de Biologie Fonctionnelle et Adaptative, F-75013 Paris, France

ARTICLE INFO

Article history:
Received 25 October 2022
Received in revised form 16 March 2023
Accepted 16 March 2023
Available online 18 March 2023

Keywords:
SARS-CoV-2
Spike protein
Structural flexibility
Hot spot residues
Pocket tracking
Binding site flexibility
Druggable binding sites
Key-residues
Molecular dynamics simulation
COVID-19 variants

ABSTRACT

The spike protein of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is crucial for viral infection. The interaction of its receptor-binding domain (RBD) with the human angiotensin-converting enzyme 2 (ACE2) protein is required for the virus to enter the host cell. We identified RBD binding sites to block its function with inhibitors by combining the protein structural flexibility with machine learning analysis. Molecular dynamics simulations were performed on unbound or ACE2-bound RBD conformations. Pockets estimation, tracking and druggability prediction were performed on a large sample of simulated RBD conformations. Recurrent druggable binding sites and their key residues were identified by clustering pockets based on their residue similarity. This protocol successfully identified three druggable sites and their key residues, aiming to target with inhibitors for preventing ACE2 interaction. One site features key residues for direct ACE2 interaction, highlighted using energetic computations, but can be affected by several mutations of the variants of concern. Two highly druggable sites, located between the spike protein monomers interface are promising. One weakly impacted by only one Omicron mutation, could contribute to stabilizing the spike protein in its closed state. The other, currently not affected by mutations, could avoid the activation of the spike protein trimer.

1. Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is responsible for the COVID-19 outbreak [1,2]. It was originally discovered in Wuhan, China, in late December 2019. The World Health Organization (WHO) labeled this epidemic a pandemic in March 2020 and reported nearly 760 million confirmed cases of COVID-19 and 6.8 million deaths by the end of March 2023 (https://covid19.who.int/). Important variants of the SARS-CoV-2 virus have emerged in the UK, Brazil, India, and South Africa from December 2020 to the end of November 2021. Five variants have been recognized by the WHO as variants of concern and are labeled as coronavirus variants: Alpha, Beta, Gamma, Delta, and Omicron [3–5]. SARS-CoV-2 is an extremely unstable virus [6,7], which is favorable for the development of new variants.

COVID-19 vaccines have enabled the reduction of the spread, severity, and death caused worldwide. Eight drugs were approved in May, 2022. Three of these drugs are biologics, which aim to block viral attachment and entry into human cells [8–10]. Two are a combination of two monoclonal antibodies intended to prevent mutational escape: casirivimab/imdevimab (commercialized under the name Ronapreve) [8] and bamlanivimab/etesevimab [9]. The third drug, sotrovimab (Xevudy) [10], is also a monoclonal antibody. Although monoclonal antibodies are an important therapeutic advancement, their manufacturing costs are high, and they are not convenient for patients because they are administered intravenously. Small molecules are often cheaper and easier to produce than protein- or peptide-based drugs [11]. They can withstand a wide range of delivery modalities, including oral administration, making them a preferred choice among pharmaceutical chemists. Currently, four small-molecule drugs are available for the treatment of COVID-19. Remdesivir and molnupiravir exert their antiviral action by perturbing viral RNA replication. Baricitinib attenuates the uncontrollable inflammatory response by the immune system owing to SARS-CoV-2 infection, referred to as a cytokine storm, by specifically inhibiting Janus kinases. From the end of December 2021, a fourth drug, named Paxlovid, has become available for people who are at high risk of developing severe COVID-19 symptoms [12]. Paxlovid is a combination of two antiviral medications, nirmatrelvir

* Corresponding authors.
E-mail addresses: gautier.moroy@u-paris.fr (G. Moroy), anne-claude.camproux@u-paris.fr (A.C. Camproux).
[1] These authors contributed equally

https://doi.org/10.1016/j.csbj.2023.03.029

and ritonavir, administered orally. Several repurposed drugs and new drug candidates are currently in phase III and IV trials. However, it is still crucial to develop drugs that can alleviate the severity of COVID-19 in individuals who are at high risk for progression to severe COVID-19.

The instability of the SARS-CoV-2 genome and the high possibility of new variants emerging make the discovery of new treatments and effective maintenance of already discovered drugs challenging [13]. Therefore, it is crucial to understand the interaction mechanisms of SARS-CoV-2 at the molecular level and the impact of its variants on these interactions.

The SARS-CoV-2 genome encodes four structural proteins: the spike, envelope, membrane, and nucleocapsid proteins, and 16 nonstructural proteins. Spike, envelope, and membrane proteins form the viral envelope, and the nucleocapsid protein binds to the RNA genome [14–16]. The spike protein is a homotrimeric glycoprotein and each monomer is composed of 1273 residues [17]. In coronaviruses, the spike protein can interact with human angiotensin-converting enzyme 2 (ACE2) to initiate fusion with host cells [18]. The receptor-binding domain (RBD) of the spike protein is responsible for the interaction with ACE2 [19]. Therefore, the RBD is a crucial protein target for the development of COVID-19 drugs. The 17 residues on the RBD that interact directly with ACE2 have been grouped under the name of the receptor-binding motif (RBM). Targeting the RBM using small molecules, therapeutic peptides, and neutralizing antibodies was determined to be an attractive method to inhibit the ability of the spike protein to bind ACE2, owing to its low glycosylation [20–22].

Advancements in structural biology and structural bioinformatics methods have enabled the elucidation of molecular and dynamic mechanisms of protein-protein or protein-molecule interactions. It is possible to design molecules capable of disrupting protein-protein interactions using structure-guided approaches. For this purpose, it is important to first understand the flexibility of the structures and to identify the residues essential for stabilizing the interaction, commonly named “hot spots”.

However, the interaction between a drug and a target protein depends on a few key residues as well as on a larger protein cavity or pocket, referred to as the binding site, which must have physicochemical and geometrical properties in agreement with those of the ligand [23]. Therefore, it is also crucial to determine which regions of the protein surface have the suitable druggability profile that can be targeted by therapeutic molecules.

The first SARS-CoV-2 spike protein structure was resolved in February 2020 [24]. The number of available spike protein structures has increased rapidly, explaining the structural mechanisms that allow SARS-CoV-2 entry into the host cell. According to cryo-electron microscopy (cryo-EM) structures of the spike protein, RBD exists in two states: a closed (or “down”) and an open (or “up”) state. In the latter state, the RBD is less buried and can interact directly with ACE2 [24]. At the beginning of this study, three crystal structures of the wild-type ACE2/RBD complex were available: PDB IDs 6M0J [25], 6LZG [26], and 6VW1 [27].

With the release of the experimental structures of the SARS-CoV-2 RBD-ACE2 complex, numerous studies have been performed to elucidate the molecular and dynamic mechanisms involved in this protein-protein interaction.

Considering the flexibility of the protein partners improved the identification of hot spot residues. Spinello et al. identified 12 hot spot residues using SARS-CoV-1 and SARS-CoV-2 PDB structures and the molecular mechanics Poisson Boltzmann surface area (MM-PBSA) method [28,29]. Using structural comparisons between the SARS-CoV-1 RBD-ACE2 X-ray structure and a SARS-CoV2 RBD-ACE2 model built by homology and the MM-PBSA method, the authors identified ten hot spots for the affinity of the SARS-CoV-2 RBD for ACE2 [30]. Recently, a combination of molecular dynamics (MD) simulations and the Molecular mechanics-generalized Born surface area (MM-GBSA) method enabled the identification of 13 hot spot residues [31]. Overall, the hot spot residues assessed using these different methods showed coherency but clear dependency on the employed experimental structures and the protocol used.

Identification of these hot spot residues is a key step in the design of drug molecules capable of disrupting protein-protein interactions. However, the protein surface region or binding site to be targeted by therapeutic compounds also requires characterization in terms of the physicochemical and geometrical properties inherent to its flexibility. The binding sites must be druggable and possess physicochemical and geometrical properties that allow them to bind to drug-like molecules [23].

Some studies have focused on the determination of druggable sites on SARS-CoV-2 spike RBD, in which occupancy could reduce, directly or through an allosteric mechanism, the interaction between the RBD and ACE2 [32–35]. Trigueiro-Louro et al. identified druggable pockets from different spike protein structures using consensus from pocket estimations [32]. They concluded that the RBD is one of the two most promising conserved druggable regions. They identified four to seven RBD pockets from three RBD-ACE2 complex experimental structures (PDB IDs 6VW1, 6LZG, and 6M0J) without considering protein flexibility. They selected two pockets characterized as druggable, observed in the three RBD structures. Carino et al. [33] identified 300 putative pockets in the whole trimeric structure of the spike protein using the Fpocket estimation program [36]. They selected six pockets on the RBD structure based on their potential druggability, structural rigidity, and sequence conservation between the SARS-CoV-2 and SARS-CoV-1 RBD. They identified two continuous pockets in the central β-sheet core of the spike RBD that were targetable by steroidal molecules based on virtual screening of FDA-approved drugs on these six RBD pockets. Using in vitro assays, they confirmed that several compounds highlighted by virtual screening can reduce the ability of the RBD to bind to ACE2.

Although these studies are promising, they have been performed only on static protein structures. The dynamic aspect of the structure, as well as the flexibility of the pockets, were not considered. To identify reliable druggable binding sites for drug design, the importance to consider the presence of cryptic, transitory, and flexible pockets, have been underlined by Stank et al. [37], Abi Hussein et al. [38], and Kuzmanic et al. [39]. Recent research conducted by Dokainish et al. [40] stressed the importance of considering the intrinsic flexibility of the SARS-CoV-2 spike protein receptor domains. They identified cryptic pockets from RBD intermediate states determined using MD simulations and concluded that the intrinsic flexibility of the spike protein must be taken into consideration when developing vaccines or antiviral medications.

In this study, we proposed a protocol considering the structural flexibility of RBD while combining machine learning approaches to identify promising RBD druggable binding sites for drug design development.

For this purpose, we performed a structural flexibility analysis of both unbound RBD and RBD-ACE2 complex using MD simulations. We also identified hot spots (crucial residues involved in the RBD-ACE2 interaction) using MM-PBSA energetics computation. Using a large number of conformations extracted from the MD simulations, we estimated the pockets on the protein surface during the MD simulations and characterized them in terms of the druggability score using a supervised machine learning method developed in the laboratory [41,42]. We then performed unsupervised hierarchical classification on the large set of estimated pockets using pocket similarity in terms of residue composition. Main pocket clusters regrouping pockets with a similar residue composition correspond to binding sites, frequently observed along DM simulations. Analysis of

these pocket clusters in terms of residue localization, frequency, variability, and druggability allowed the identification of promising druggable pockets and of their key residue in terms of contribution to the binding site and its druggability. Additionally, the potential mechanistic impact of mutations of SARS-CoV-2 variants of concern on the selected druggable binding sites located in key regions of the RBD surface and hot spots is discussed.

2. Materials and methods

2.1. Protein preparation

Three crystal structures of the SARS-CoV-2 RBD complexed with the human ACE2 receptor were considered (PDB IDs 6VW1 [25], 6M0J [26], and 6LZG [27]). Overall, the three structures are very similar, with a maximum backbone root mean square deviation (RMSD) of 1.25 Å. However, 6VW1 is based on a chimeric SARS-CoV-2 RBD protein and displays poorer resolution than the 6M0J and 6LZG structures. Coordinates of 6M0J and 6LZG crystal structures were determined at 2.45 Å and 2.50 Å, respectively. Therefore, our study focused only on the 6M0J and 6LZG structures. Each complex was protonated at physiological pH (7.4) using the PROPKA tool [43].

2.2. MD simulations

MD simulations were performed on these two crystallographic structures to study the dynamic behavior and stability of the ACE2-RBD complex. They were also performed to study the dynamic behavior of the unbound state of the RBD initially extracted from the ACE2-RBD complex.

MD simulations were performed using the GROMACS software package [44] with the CHARMM36m all-atom additive protein force field [45] under periodic boundary conditions. A dodecahedron water box of TIP3P water molecules was used to run the simulations. The full simulation system consisted approximately of 12,500 atoms. Non-bonded interactions were truncated at a cut-off distance of ten Å for electrostatic twin-range cut-off and van der Waals cut-off. Neighbor searching was performed every 10 steps. The energy of the system was minimized over approximately 1000 steps using the steepest descent algorithm for energy minimization after the ion addition and neutralization of the systems. Each system was equilibrated with an NVT (Number of particles, Volume, and Temperature) ensemble during 1 ns at a temperature of 300 K and a coupling constant of 0.1 ps. Subsequently, each simulation was performed under number of particles, pressure, and temperature conditions, coupling the system to a heat bath using the Berendsen algorithm and setting the temperature at 300 K and the pressure at 1 bar during 1 ns. For the production step, ten independent runs of 100 ns with different initial velocities were performed. The LINCS algorithm [46] was applied to all the bond lengths to constrain them, and an integration time step of 2 fs was used. The electrostatic interactions were cut off at 1.2 nm and were calculated using the particle mesh Ewald method. As for Lennard-Jones interactions, they were cut off by 1.2 nm by using the potential shift Verlet method [47]. Periodic boundaries conditions were also applied to all systems. The temperature was maintained at 300 K with a V-rescale thermostat [48] and τ_T coupling constant of 0.1 ps. The pressure was maintained at 1 bar with a τ_P constant of 2.0 ps and a compressibility of 4.5×10^{-5} bar^{-1}using the Parrinello-Rahman barostat [49].

For each PDB structure, the trajectories were merged for the analyses, resulting in a total of 1 μs. The RMSD analysis measures the average distances between the starting structure and each structure obtained over time. The root-mean-square fluctuation (RMS Fluctuation) was analyzed to identify flexible residues.

Trajectory analysis was performed using the GROMACS tool (v 2019.5). They were visualized using VMD (v. 1.9.4a38) [50]. The plots were generated using R (v. 3.1.1) [51] and GNUPLOT (v. 5.2) [52]. Figures were generated using PyMOL [53].

2.3. Identification of hot spot residues

The MM-PBSA method is widely used for binding free energy calculations from conformations extracted from the MD trajectory [54]. In this study, MM-PBSA calculations were performed on MD simulations of the RBD-ACE2 complexes (6M0J and 6LZG). The representative conformations of each independent MD simulation were extracted through cluster analysis using the GROMACS gmx cluster tool. The frames were selected in such a way as to cover a wide range of trajectories and to sample different conformational spaces of the complex. Accordingly, representative frames of each of the determined clusters were extracted for every MD simulation using the Gromos algorithm [55] and the best optimal backbone RMSD cutoff was selected. To choose a reasonable cut-off value for each trajectory, we varied the backbone RMSD cutoff between 0.9 and 1.5 Å. Thereby, we found dominant clusters that captured > 70% of the trajectory for each MD simulation. Thus, for every dominant cluster (seven clusters/MD), we extracted one representative conformation, which was subjected to MM-PBSA calculations.

A free energy decomposition analysis was performed using MM-PBSA residue decomposition to retrieve the contribution energy of each amino acid represented on the binding interface of both RBD and ACE2. The total free energy and its individual components for each individual system (6M0J and 6LZG) were averaged and weighted based on the cluster populations, that is, a higher weight was assigned to the conformations extracted from more populated clusters.

We used the *g_mmpbsa* package for MM-PBSA computations [56]. The dielectric relative constant ε was set to eight for proteins and 80 for water [57]. In this approach, the binding free energy ΔG_{bind} between the protein and ligand/protein includes different energy terms and can be calculated as follows:

$$\begin{aligned}\Delta\text{Gbind} &= \text{Gcomplex} - (\text{Gprotein} + \text{Gligand}) \\ &= \Delta\text{EMM} - \text{T}\Delta\text{S} + \Delta\text{Gsol} \\ &= \Delta\text{Evdw} + \Delta\text{Eelec} - \text{T}\Delta\text{S} + \Delta\text{GPB} + \Delta\text{GSA}\end{aligned}$$

where ΔE_{MM} is the gas-phase interaction energy, which is the sum of the van der Waals energy ΔE_{vdw} and the electrostatic energy ΔE_{elec}. ΔG_{sol} is the sum of the polar solvation energy ΔG_{PB} and nonpolar solvation energy ΔG_{SA}. The polar solvation energy was calculated using the Poisson Boltzmann (PB) approximation model, whereas the nonpolar solvation energy was estimated using the solvent-accessible surface area (SASA). The entropy contribution (− TΔS) was ignored in this study because of its high computational cost. After the calculation, the binding free energies were decomposed into each residue. We considered hot spots as residues with binding energy below − 1 (favorable) or above + 1 kcal.mol^{-1} (unfavorable). It is important to note that the more negative the energy, the more favorable is the contribution. In contrast, positive energy values indicate unfavorable interactions and a poor contribution to the complex.

2.4. Pocket estimation, druggability prediction and tracking along the MD simulations

The estimation of the RBD pockets was first performed on the static structures from the PDB.

Pockets were estimated with PockDrug tool based on the automated geometry-based method from Fpocket [36] independent of ligand proximity information. The ensemble of pockets was then characterized in terms of 19 physicochemical and geometrical

properties and druggability score using PockDrug [41,42]. PockDrug provides a prediction of the druggability score which, if greater than 50%, indicates a druggable pocket.

Secondly, pocket tracking was run on the sample of MD generated conformations of the RBD unbound and bound states. This allows the identification of new pockets during dynamics or changes in pocket properties between the PDB conformations or conformations observed during MD. This approach can also detect some separation or fusion of pockets resulting from both local and global alterations, including those occurring in transient and allosteric pockets.

To consider the flexibility of the RBD, we estimated the pockets from conformations obtained from the RBD MD simulations. A total of 1000 RBD conformations were sampled from MD simulations at regular intervals. RBD pockets estimated from this series of 1000 conformations were merged into a pocket set.

2.5. Clustering of pockets and identification of druggable binding sites

We performed unsupervised hierarchical classification using residue pocket similarity to identify binding sites frequently observed along MD simulations, corresponding to main pocket clusters.

The similarity of the RBD pockets can be quantified using binary distance based on common residues. A binary distance of 0 corresponds to two identical pockets in terms of residues. Hierarchical classification of the pockets was performed using the binary distance, Ward metric (ward.D2) [58], and R Hclust package [59]. Dendrogram visualizations were performed using the Heatmaps2 package in R39 [51] to illustrate pocket similarity in terms of residues. The dendrogram lengths between the pockets and/or pocket clusters are proportional to their binary distances.

The resulting pocket classification allows the identification of the main clusters of pockets similar in terms of residues composition, corresponding to binding sites frequently observed along MD simulations. The dissimilarity between the pockets or pocket clusters increased with dendrogram lengths between pockets or clusters. Main pocket clusters were compared i) between pocket sets extracted from two PDB IDs (6LZG and 6M0J) to assess the impact of the initial PDB structure and ii) extracted from bound and unbound RBD conformations to assess the influence of ACE2 interaction.

The number of similar pockets within one cluster indicates the frequency of the appearance of its corresponding binding site. The flexibility of the binding site is described by the residue variability of its corresponding pockets. Analysis of pocket clusters was performed in terms of frequency, residue contributions and variability, RBD localization, and druggability scores. Finally, we combined these statistical and flexibility analyses to select binding sites based on the criteria of frequency and residue stability, localization in key regions of the RBD, accessibility, and druggability scores.

We also crossed the binding site analysis with the hot spot information obtained from the MM-PBSA analysis and evaluated the potential structural impact of mutations observed in the five SARS-CoV-2 variants of concern (Alpha, Beta, Gamma, Delta, and Omicron) on these selected hot spots and binding sites to identify the druggable binding sites that can also be targeted in mutated systems.

2.6. Protocol of identification of druggable binding sites and their flexibility

Combining supervised and unsupervised machine learning techniques with a traditional flexibility study through MD simulations and MM-PBSA energetic computation analysis yields a comprehensive protocol. This three-step protocol identify druggable binding sites frequently observed on the RBD protein and their key residues in terms of druggability and contribution to stability (Fig. S1).

3. Results and discussion

3.1. RBD-ACE2 structures

Several PDB structures have elucidated the interaction network between the ACE2 protein and RBD spike protein [25–27]. The RBM is composed of 17 residues (K417, G446, Y449, Y453, L455, F456, A475, F486, N487, Y489, Q493, G496, Q498, T500, N501, G502, and Y505) that are in contact with ACE2 residues to stabilize the complex interaction [60,61]. With a cutoff distance of 4 Å, the complex was maintained with 13–18 hydrogen bonds and a salt bridge (Fig. 1). SARS-CoV-2 forms distinct patches of residues that spread along the interaction surface. Two hydrogen bonds were formed between Y489 and Y83, and between N487 and Q24 (Fig. 1 A). Q493 interacts with both K31 and E35 ACE2 residues through hydrogen bonding and its two crystallized side chains (Fig. 1B). A single salt bridge formed between K417 and D30 (Fig. 1B). On the other side of the surface, two large patches of residues establish strong intra- and intermolecular interaction networks [62]. This network includes hydrogen bonding, hydrophobic, and π–π interactions [63]. The complex is stabilized by hydrogen bonds between G446-Q42, Y449-D38, Q498-Q42, T500-Y41, N501-Y41, N501-N330, G496-K353, Q498-K353, G502-K353, Y505-E37, D355-R357, and Y505-R393 (Fig. 1 C, Fig. 1D, and Fig. 1E). Residues involved in hydrophobic and van der Waals interactions played an important role in the affinity of the complex (Fig. S2). T27 interacted with Y473, A475, and F456 through hydrophobic packing. F28 interacts with Y489, and H34 with Y453 and Q493 through van der Waals interactions. Other hydrophobic patches were found on both protein surfaces, including L45 with V445, G446, and Q98, L79 with F486 and Y489, and Y505 with K353.

3.2. RBD-ACE2 flexibility

Proteins are dynamic molecules with intrinsic flexibility and often undergo conformational changes upon partner binding. Therefore, it is essential to consider their dynamic behavior to predict which surface regions may be of interest.

RBD flexibility was studied with and without ACE2 by MD simulations. Mean Cα RMSDs have been found to stabilize for both bound proteins systems (6M0J and 6LZG) around 2.5 Å with fewer fluctuations (Fig. S3 and Fig. S4). These results highlight the stability of the SARS-CoV-2 RBD and ACE2 complex throughout all the simulations (Fig. S3A and Fig. S4A). The low mean Cα RMSD values indicated that the unbound RBD was stable (Fig. S3D and S4D). No major flexibility variation was observed between the bound and unbound RBD (Fig. S3C, Fig. S3D, Fig. S4C and Fig. S4D), considering both unbound and bound RBD simulations, have not been reported in other studies. To examine the flexibility and local changes in the complex, Cα RMSF versus the residue number of both RBD systems were investigated (Fig. S5). RMSF analysis revealed that the RBM region flexibility increased to a greater extent in the unbound RBD structure than in the bound RBD and increased from approximately 1.5 to 2.0 Å. This was not surprising, as the RBM mediates contact with ACE2, which tends to be more stable when the complex is formed (Fig. S5A and Fig. S5C) than in unbound RBD (Fig. S5B and Fig. S5D).

3.3. Hot spot residue analysis

MM-PBSA calculations were performed to assess the binding free energies of the ACE2-RBD complexes and to estimate the contribution of residues in this interaction. The MM-PBSA calculations were applied to specific frames of representative states that were extracted from the MD trajectories after clustering analysis. Clustering analysis was based on a specific series of 23 residues for the RBD and

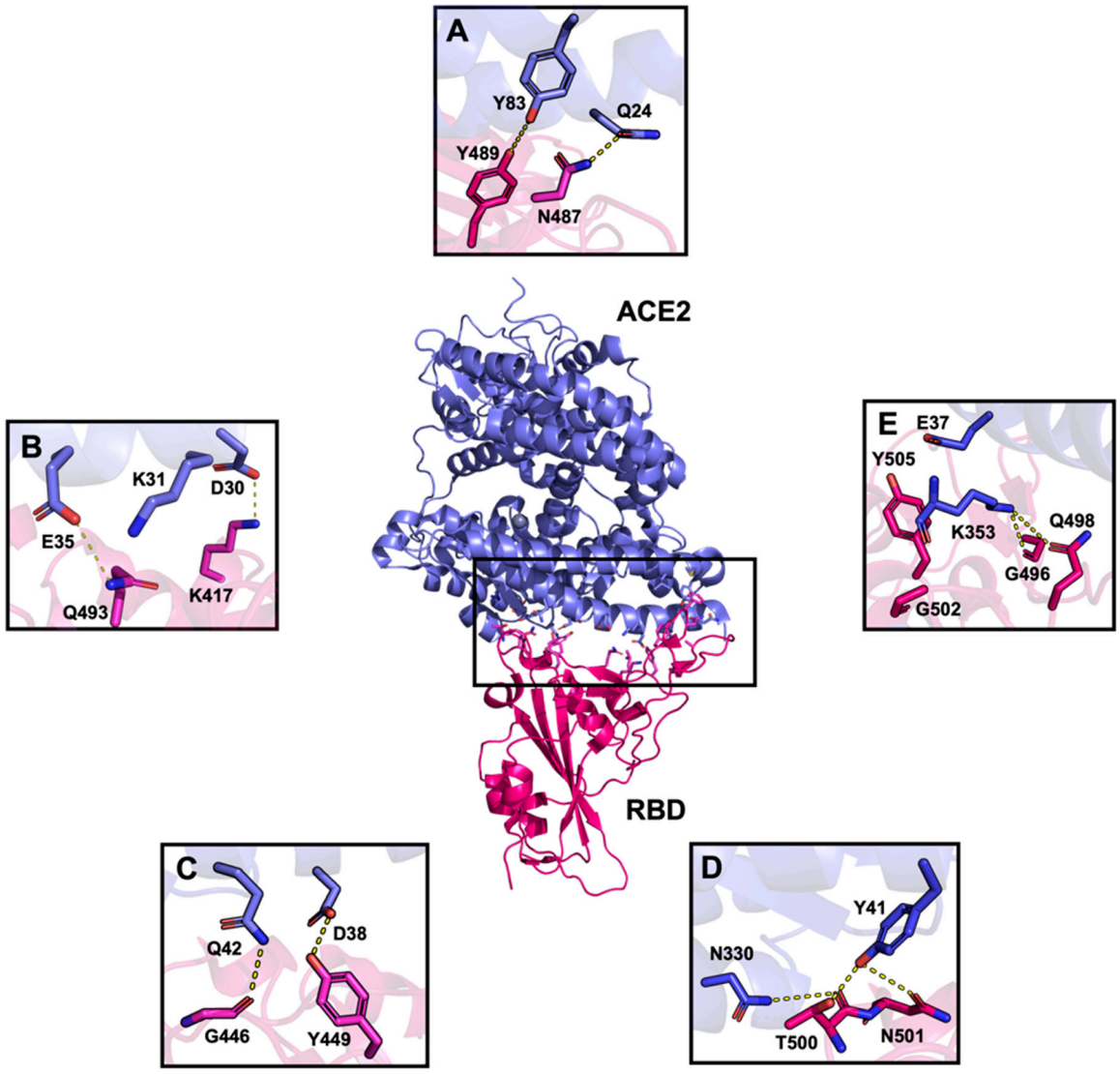


Fig. 1. Interaction networks of the SARS-CoV-2 RBD and ACE2 complex (PDB: 6M0J). (A) Hydrogen bonds with Q24-N487 and Y83-N487. (B) Salt bridge between D30 and K417 and hydrogen bonds with K31-Q493 and E35-Q493. (C) Hydrogen bonds between Q42-G446 and D38-Y449. (D) Hydrogen bonds between Y41-T500, N330-T500, and Y41-N501. (E) Hydrogen bonds between K353-G496, K353-Q498, E37-Y505, and G502-K353. ACE2 is indicated in blue and RBD in pink. Interactions were configured using PyMOL software. Only the side chains of the residues are represented, excluding for glycine residues.(For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

22 residues for ACE2. The residues corresponded to the amino acids spanning the binding interface with a cut-off of 4.0 Å. These residues are K417, G446, G447, Y449, Y453, L455, F456, Y473, A475, G476, S477, E484, F486, N487, Y489, F490, Q493, G496, F497, Q498, T500, N501, and G502 for the RBD and S19, Q24, T27, F28, D30, K31, H34, E35, E37, D38, Y41, Q42, L45, L79, M82, Y83, N330, K353, G354, D355, R357, and R393 for ACE2.

Based on the MM-PBSA calculations, SARS-CoV-2 RBD binds to ACE2 with a ΔG_{bind} of − 41.2 ± 20.2 and − 47.1 ± 21.8 kcal.mol^{-1}, for 6M0J and 6LZG structures, respectively (Table 1). The binding free energies of the two complexes are similar.

Table 1
Average binding free energy and standard deviations in the MD simulations of the RBD-ACE2 complex for PDB IDs 6M0J and 6LZG.

PDB ID	Binding energy components (kcal.mol^{-1})			
	ΔE_{MM}	ΔE_{Polar}	ΔE_{Apolar}	ΔG_{bind}
6M0J	-59.3 ± 3.7	85.1 ± 21.6	-7.7 ± 0.6	-41.2 ± 20.2
6LZG	-59.2 ± 3.9	82.8 ± 22.9	-7.7 ± 0.6	-47.1 ± 21.8

Free energy decomposition analysis was also performed to obtain a detailed insight into the interactions between each residue in the binding interface of the two proteins. The binding interaction of each residue included four terms: molecular mechanics contribution, polar contribution, apolar contribution, and total energy contribution. The individual components of each residue were averaged for each system (6M0J and 6LZG). The hot spot residues on the RBD surface were essentially the same for the core interactions in 6M0J and 6LZG (Fig. 2).

Notably, residues F486 and Y489 inserted into a hydrophobic pocket on the surface of ACE2 formed by residues including T27, F28, L79, and M82. Thus, the presence of aromatic amino acids in the pocket may provide an additional binding force through π-stacking interactions (Fig. 3A) and explain the favorable ΔE_{MM}, which is mainly due to the favorable contribution of ΔE_{vdw}. The residue K417 showed a significant ΔE_{polar} contribution that reaches 11.0 kcal.mol^{-1}. However, the fact that this residue formed a salt bridge and a hydrogen bond with residue D30 led to a more negative ΔE_{MM} contribution of − 14.0 kcal.mol^{-1} (Fig. 3B) and, consequently, it counterbalanced the unfavorable polar solvation energy. Moreover,

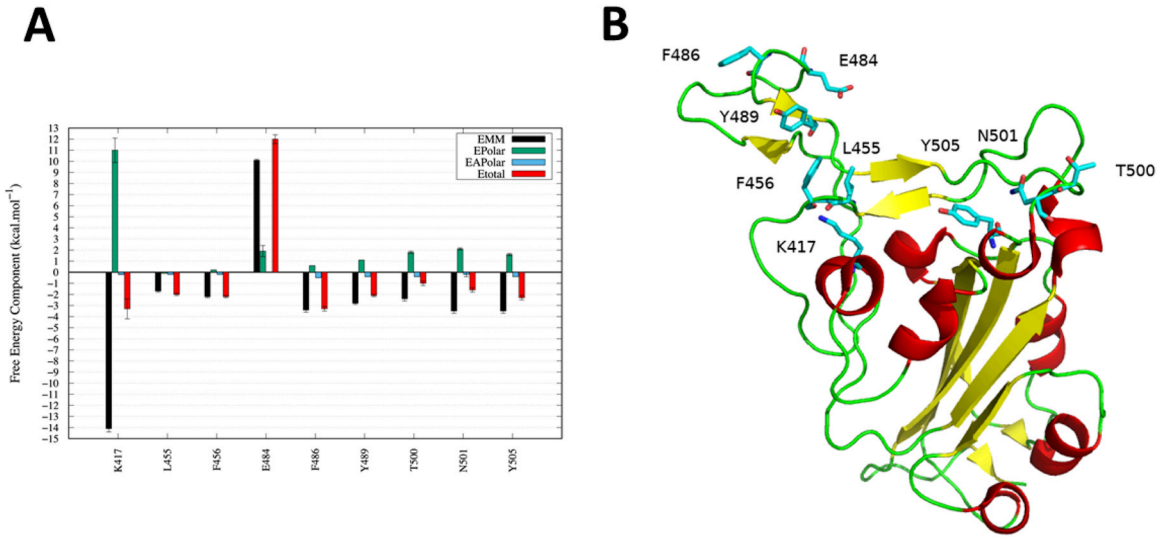


Fig. 2. (A) The energy components of the nine residues on the RBD that contribute significantly to the ΔG of binding with ACE2. The average binding energy components and standard deviations are calculated from the MD simulations of the RBD-ACE2 complex (PDB IDs 6M0J and 6LZG). (B) RBD displayed in cartoon representation. The hot spot residues are indicated by cyan sticks.

ACE2 residue Y41 interacted with Q498 and Y505 through van der Waals interactions and formed two hydrogen bonds with residues T500 and N501 through its side chain (Fig. 3C). Although ACE2 residue H34 was not directly involved in any intermolecular hydrogen bond with RBD, it nonetheless provided favorable polar contacts with the side chain of L455 of the virus protein (Fig. 3D).

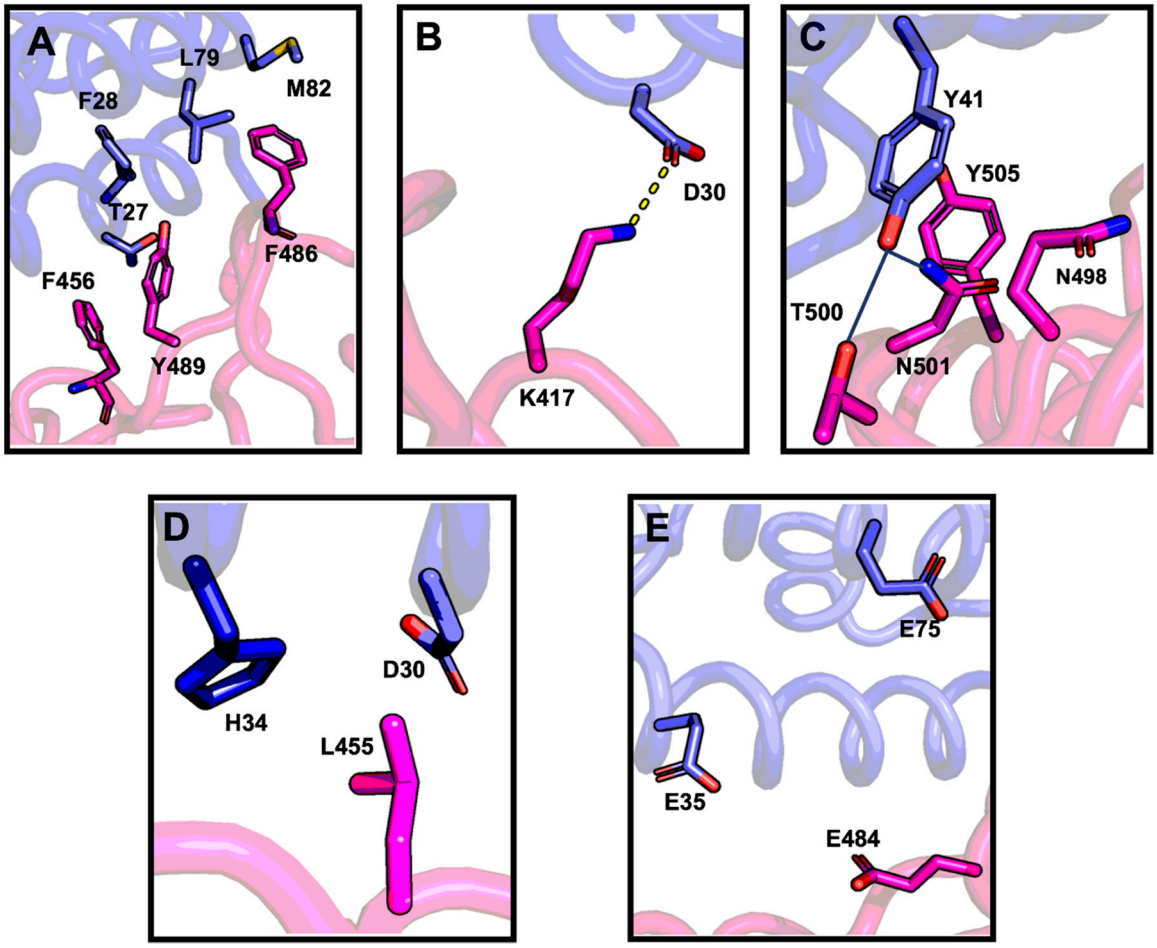


Fig. 3. Contact residues of ACE2-RBD interface issued from MM-PBSA analysis. ACE2 is colored in blue and RBD is indicated in pink. Amino acids are represented by sticks and colored based on their respective proteins. A) Hydrophobic pockets, including F456, F486, and Y489, B) K417 and D30 salt-bridge, and C) ACE2 Y41 interacts with T500 and N501, D) L455 and H34 interactions, and E) E484 negative repulsive charges with E35 and E75. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

L455 also formed intermolecular van der Waals interaction with D30 (Fig. 3D). Interestingly, E484 displayed a large positive E_{MM} contribution that reaches 10.0 kcal.mol^{-1} (Fig. 2). Thus, E484 free energy contribution (E_{TOT}) was found to be disfavoring complex binding and formation. Two negatively charged residues, E35 and E75, surrounded E484, which creates long-range opposing repulsive charges and explains the unfavorable contribution of this residue to the complex (Fig. 3E).

Altogether, our analysis shows that among the residues comprising the ACE2-RBD interface, some critical amino acids are known as the evolutionary signature of RBD, and major hot spots of both proteins contribute favorably to complex stability and formation [27]. In conclusion, eight favorable hot spot residues (K417, L455, F456, F486, Y489, T500, N501, and Y505) and one unfavorable hot spot residue (E484) have been identified as important SARS-CoV-2 residues that are critical for ACE2 binding.

These results are in good agreement with other studies in which the authors identified 10–13 hot spot residues using similar bioinformatics methods on different RBD structures, that is, X-ray structures or models built by homology modelling, from SARS-CoV-1 or SARS-CoV-2 [29–31,64–66].

3.4. Tracking of RBD pockets during the MD simulations

We first examined pockets that were extracted from the unbound RBD system from static PDB. From the initial 6M0J and 6LZG PDB IDs, we estimated four and seven pockets on the unbound RBD, respectively, using the PockDrug program [41,42]. This is consistent with the observation of a variable number of pockets estimated by Trigueiro-Louro et al. on the same PDB structures [32].

We secondly examined pockets that were extracted from the unbound RBD systems considering their flexibility. We extracted two sets of pockets from two datasets of 1000 conformations sampled from each of the unbound RBD structures 6M0J and 6LZG during the MD simulations. The average number of RBD pockets by conformation in these two sets were highly similar (5.1 versus 5.8) despite the different number of pockets estimated on the two static PDBs. The frequency of druggable pockets (50.4% versus 54.0%) and the frequency of pockets located at the interface (31.0% versus 25.0%) were also highly similar within both 6M0J and 6LZG MD trajectories. The consistency of the results obtained from the MD simulations, regardless of the initial structure, shows the importance of integrating the protein and pocket structural flexibility for binding site identification and selection.

The results were consistent for both the 6M0J and 6LZG PDBs; therefore, we decided to focus only on the 6M0J MD trajectory. The average number of 5.1 pockets in the RBD conformation (from one to a maximum of nine pockets) results in a total set of 5065 pockets. Of these pockets, 50.4% were predicted to be druggable (druggability score ≥ 50%), and 40.6% of these pockets were highly druggable (druggability score ≥ 75%). Nearly one-third of the pockets (31.0%) included at least one of the 17 main RBM interacting residues and were located at the RBD-ACE2 interface. We also combined pocket analysis with the hot spot results obtained from the MM-PBSA study to identify the druggable binding sites containing these hot spots. Interestingly, almost all of these interface pockets (98.5%) included at least one of the nine hot spot residues (identified in Section 3.3). This is consistent with the fact that hot spot residues involved in energetic interactions have physicochemical properties and locations that promote pocket accessibility.

To determine how ACE2 affects RBD conformation and binding regions, we subsequently examined pockets extracted from the bound RBD system. The same pocket tracking protocol was applied to 1000 RBD conformations extracted from MD simulations of 6M0J in the RBD-ACE2 complex form. Therefore, a set of 4428 pockets with a pocket druggability frequency of 50.1% was extracted from the 6M0J RBD-ACE2 complex trajectory. This is close to the pocket druggability frequency of 50.4% of the 5065 pockets extracted from the unbound RBD trajectory. Fewer pockets (approximately 640 pockets) were observed in the RBM region of bound RBD. This can be explained by the presence of ACE2, which limits pocket formation in the RBM region. These results clearly show that consideration of the flexibility of the structure leads to the identification of comparable and replicable RBD pockets when utilizing isolated or complex forms, with the exception of the area in contact with ACE2. Comparable steady results were also obtained using 6LZG MD simulations (not shown).

3.5. RBD pockets clustering based on pocket residue similarity

After quantifying the druggable pockets on the two different systems, we performed a hierarchical classification of the pockets extracted from the 6M0J unbound RBD trajectory based on residue similarity. This classification of pockets based on residue composition allows the identification of clusters of pockets with common residues, corresponding to binding sites frequently observed along the MD simulations. Furthermore, the variability in residues within a pocket cluster illustrates the residue flexibility of the corresponding binding site. Highly similar classification results were obtained for pocket sets extracted from MD simulations of the 6M0J and 6LZG PDB. Here, the classification obtained for the 5065 pockets extracted from the 6M0J unbound RBD trajectory is presented. This classification resulted in the identification of eight main pocket clusters, enumerated as clusters I to VIII (Fig. 4). We analyzed these main clusters in terms of frequency, residue localization, stability, and druggability scores.

These eight clusters were observed in at least one-third of the 1000 conformations, and did not correspond to rare pockets (Table S2). They were regularly observed during the MD simulations. The least frequent (cluster VI) was observed in 32.3% of the conformations, whereas the most frequent (cluster I) was observed in approximately 90% of the conformations.

Moreover, these eight clusters were well-characterized in terms of druggability (Fig. 4 and Table S2). Four clusters were mainly druggable (II, III, V, VII, including 68.9–99.2% of druggable pockets), whereas four others identified clusters included a weak or moderate part of druggable pockets (I, IV, VI, VIII including 0.2–27.9% of druggable pockets). Interestingly, each cluster was defined by specific residues that constituted pockets that were well differentiated and localized in different regions of the RBD (Fig. 4). Clusters I, II, III, IV, and VI regrouped very similar pockets, as indicated by the small distances between the pockets within each cluster. Clusters V and VIII showed less pocket similarity in terms of residues, and each could be split into two main homogeneous sub-clusters. Cluster VII was the most variable and included three main subclusters (Fig. 4).

3.6. Selection of RBD druggable binding sites

A more detailed analysis was performed to study the properties and localization of the eight main clusters resulting from RBD pocket classification (Fig. 4) in order to extract druggable binding sites of interest. Six clusters (II to VII) were distant from the RBD-ACE2 interface, and two clusters (I and VIII) were located close to the RDB-ACE2 interface.

Two clusters (IV and VI) were moderately observed in the 1000 RBD conformations (48.8% and 32.3%) and were weakly druggable (27.9% and 7.4%, respectively). The cluster pockets corresponded to two exposed protein cavities in both the closed and open states of the spike protein trimer. These cavities did not establish contact with the ACE2 interface or another RBD monomer. Consequently, they did not form valuable target sites. The other four clusters (II, III, V, and VII) were observed in approximately half or more of the 1000

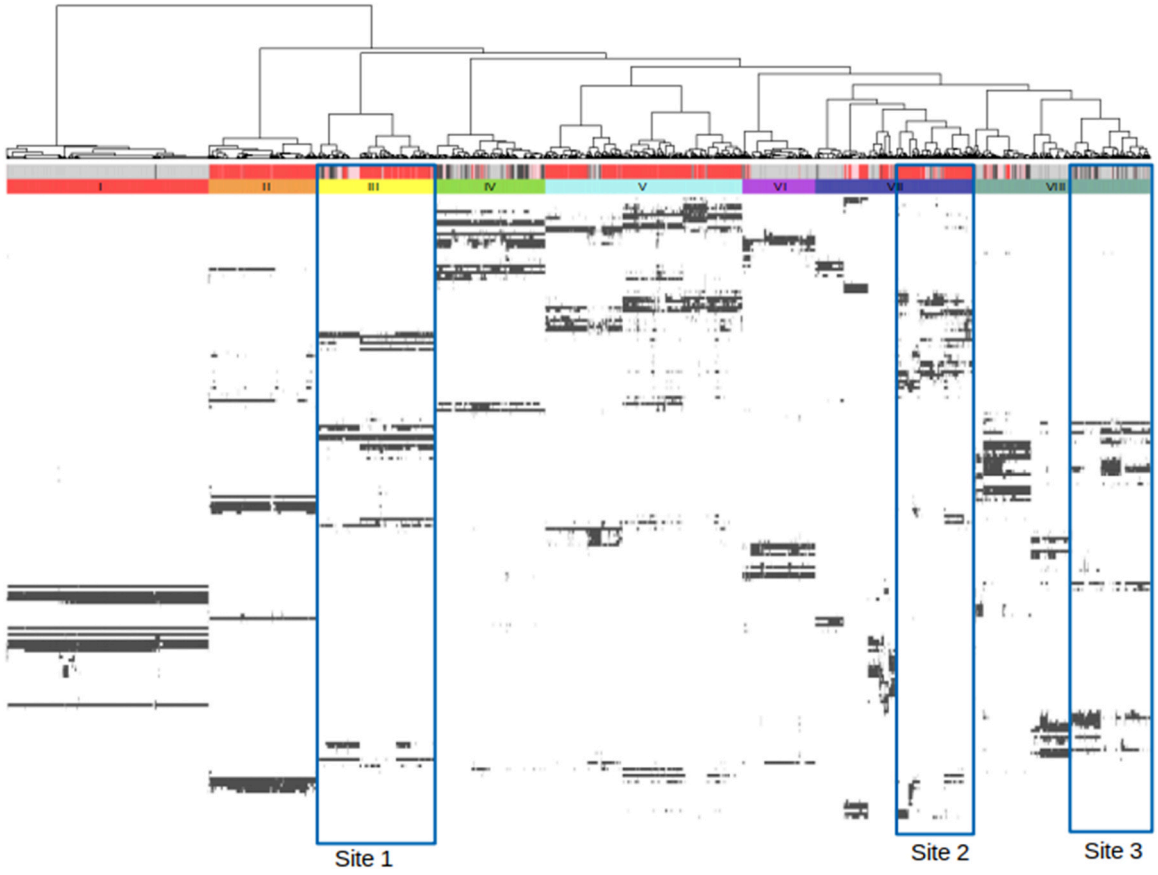


Fig. 4. Hierarchical classification of 5065 RBD pockets extracted from 1000 conformations of unbound RBD (6M0J). Each column corresponds to a pocket. First-color bars are colored according to four levels of druggability scores: respectively in black, grey, pink, and garnet for the non-druggable [0.00, 0.25], less druggable [0.25, 0.50], moderately druggable [0.50, 0.75] and highly druggable [0.75, 1.00] pockets, as predicted using PockDrug [41]. Second-color bars are colored based on the eight main clusters of pockets (noted I to VIII). At the bottom, the residues are ordered according to the numbering of the protein sequence and indicated in black when involved in the pocket. The three blue rectangles indicate the three sub-cluster of pockets corresponding to selected sites of interest.(For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

conformations and included mainly druggable pockets (> 68%). Cluster II was frequent (47.9%) and mainly druggable (99.2%). However, its pockets were located between the two spike protein interfaces in both states and are thus buried. Thus, this cluster cannot be targeted by therapeutic molecules. Cluster V was highly frequent (87.3%) and most of its pockets were druggable (93.0%). The corresponding pockets were highlighted as sites of interest in a previous study [33]. However, this cluster included N343, which is covalently linked to N-acetylglucosamine in the RBD. Notably, N-acetylglucosamines were not used in the MD simulations in this study. Therefore, even without this information, our protocol could identify a region matching an N-linked glycosylation site. We do not consider this cluster of pockets a priority to be targeted by therapeutic molecules due to the presence of N-acetylglucosamine.

Cluster III was observed in half of the conformations (52.4%) and mainly included druggable pockets (85.9%) (Table S2). The pockets were located on the interface region between two RBD proteins in the closed state of the spike protein trimer (Fig. S6). This cluster regroups similar pockets (with 17 residues observed in more than 15% of its pockets) (Table S2). It is noteworthy that Carino et al. highlighted one potential druggable pocket from the static spike protein trimer (PDB ID:6VSB), but it corresponded only to a partial sub-pocket of ten residues of our cluster of pockets [33]. Therefore, it may be of interest to design molecules that bind this cluster of 17 residue pockets to stabilize the closed state and avoid the activation of the spike protein trimer. This frequent and druggable cluster, named site 1 (Table 2), is a potential site of interest for the design of new therapeutic molecules.

Table 2
The three selected potential binding sites with their residue composition (only residue observed in more than 15% of their associated pockets), their occurrence on 1000 RBD conformations and the percentage of druggable pockets within each site. Hot spot residues are in bold.

Selected sites (sub-Cluster)	Residues	occurrence	Druggable pockets frequency (%)
Site 1 (Cluster III)	S375, T376, K378, Y380, G404, E405, V407, R408, I410, A411, Q414, V433, A435, V503, G504, Y508, V510	52.4	85.9
Site 2 (sub-cluster VII)	A363, D364, Y365, S366, L368, Y368, N370, F374, F377, C379, V382, P384, T385, L387, N388, D389, L390, C391, F392, C432, I434, L513, F515, V524, G526	35.0	87.7
Site 3 (sub-cluster VIII)	R403, D405, E406, R408, Q409, T415, G416, **K417**, I418, Y453, **L455**, S494, Y495, G496, F497, Q498, **N501, Y505**	35.0	13.1

Cluster VII was frequent (70.4%) and included pockets mainly structurally located at the bottom end of the RBD and partially buried in the spike protein trimer. However, this cluster regroups pockets that vary in terms of residues; thus, cluster VII can be decomposed into four main pocket sub-clusters. Interestingly, the most frequent sub-cluster, named site 2, which was also observed in 35% of the conformations, corresponded to homogeneous and druggable pockets (87.8%) (Table 2). The pockets point toward another spike protein monomer and are consequently partially buried. This subcluster coincides with the pocket identified by Toelzer et al., in which linoleic acid is bound [67]. Linoleic acid appears to stabilize the spike protein in its closed state and induces a reduced ACE2 interaction in vitro. Therefore, our protocol was successfully identified and characterized in terms of druggability and residue stability in this specific region, called site 2, which corresponds to the experimentally known binding site entrance of a small molecule.

We also extracted two frequent clusters, I and VIII, which were located close to the RBD-ACE2 interface. Cluster I was quasi permanent (86.6%), although it displayed no druggability (0.2%). It contains nearby pockets located on the RBD loop between residues T470 and P491 and thus excluded the association of every RBM region. Moreover, cluster I pockets were exposed to the solvent in both the closed and open states of the spike protein. Due to its location and undruggable properties (Table S2), cluster I is not suitable for targeting therapeutic molecules.

Our results showed that cluster VIII was frequent (78.0%) and partially druggable (21.7%). It is relatively variable and regroups three main subclusters located at the ACE2 interface. The most frequent sub-cluster, named site 3, was observed in 35.0% of the conformations and included 13.1% of druggable pockets (Table 2). It is centralized at the interface and includes three hot spots, K417, N501, and Y505, among the 18 residues constituting this pocket. Consequently, site 3 is particularly appealing as a targeted therapeutic molecule.

Considering their occurrence, residue similarity, druggability score, and localization, three sites of interest (sites 1, 2, and 3) were selected for further study (Fig. 5).

These sites are observed in more than 30% of the conformations and are associated with an average druggability score range of 13.1–87.7%. The pockets of these three sites included 20 residues on average (Table 2). The contribution of residues to their pocket cluster, as well as the druggability score, was made available for the three selected sites in Fig. S7. Along with these results, we have shown that the two most frequent sites, 1 and 2, are distant from the RBM, even though they display significant druggable scores. In contrast, site 3 has been classified as less frequent yet druggable.

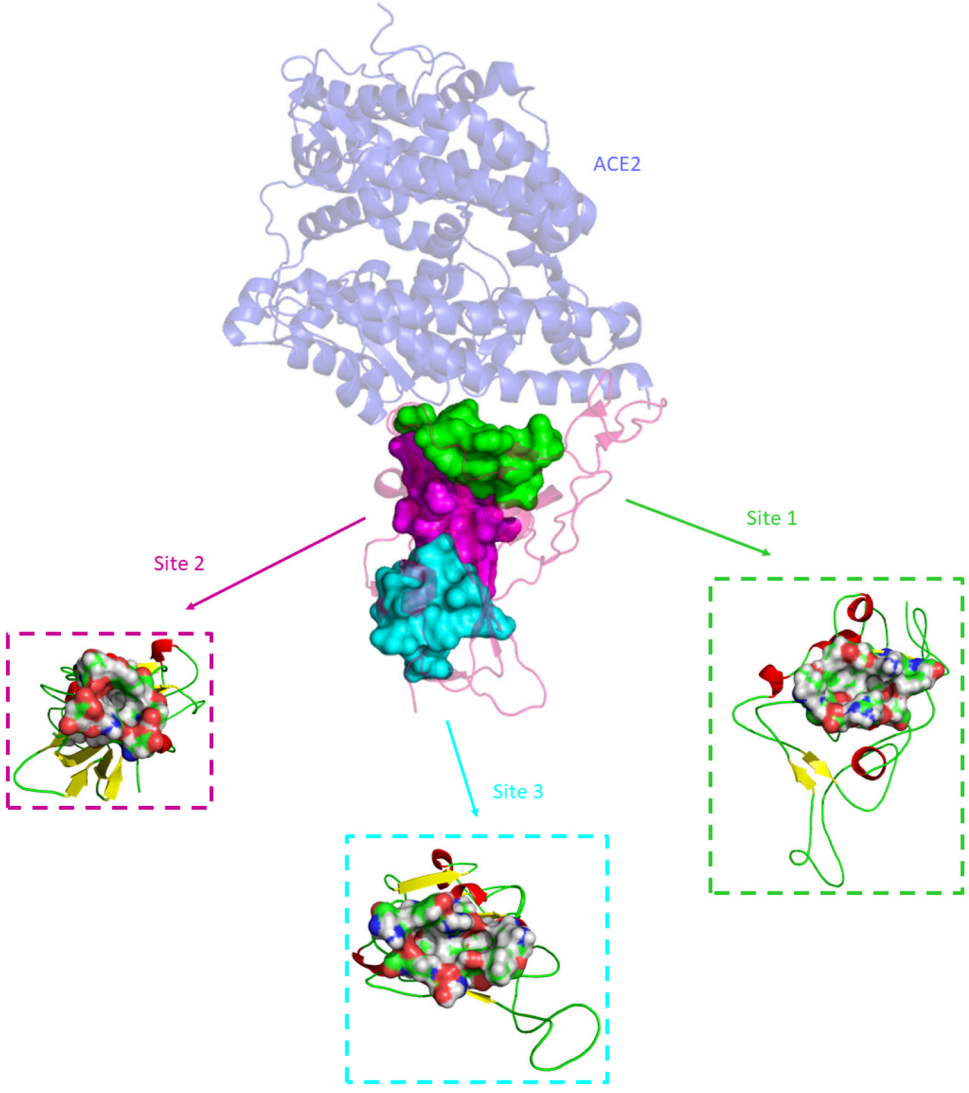


Fig. 5. Representation of the three sites of interest selected on the RBD surface (site 1 represents the trimerization pocket, site 2 indicates a highly druggable pocket capable of holding a drug candidate, and site 3 corresponds to the interface region pocket).

Compared to the other sites, an important feature of site 3 is its location, which is close to the RBD-ACE2 interface, and it includes three highly contributing hot spots: K417, N501, and Y505.

Dokainish et al. identified several pockets with residues overlapping with those of binding site 3 [40]. This site was targeted by nilotinib. The latter study, as many others, validates our protocol because we identified the same site. This site includes four hot spot residues, suggesting the significance of targeting it to disrupt the interaction between RDD and ACE2.

Ten representative pockets were selected for each of the three sites to illustrate the binding site and their flexibility (pocket PDB, residues, and global properties) that could be used for the design of novel compounds that target the spike protein. Table S3 provides descriptions of physicochemical, geometrical, and druggability scores. The ten selected pockets and their corresponding RBD conformations are freely available in the PDB format (https://data.mendeley.com/datasets/xhnjgfbgzr/1).

3.7. Potential impact of mutations on RBD hot spot and binding sites

It is known that the mutations associated with five variants of concern, Alpha, Beta, Gamma, Delta, and Omicron, cause a substantial increase in their transmissibility, virulence, and antigenic escape ability. During the evolution of the SARS-CoV-2 genome, RBD mutations had an impact on spike protein affinity for ACE2 to improve virus transmissibility and/or decrease its affinity for antibodies to improve its antigenic escape ability. The Alpha variant includes only one mutated residue, N501Y, whereas Beta, Gamma, and Omicron variants include both N501Y and K417N/T and E484K/A mutated residues. The Omicron variant also has 12 supplementary mutated residues (G339D, S371L, S373P, S375F, N440K, G446S, S477N, T478K, Q493R, G496S, Q498R, and Y505H). The Delta variant includes two specific mutations (L452R and T478K).

It is of interest to analyze the potential structural impact of RBD mutations of these variants of concern, notably, relative to the nine hot spots we have identified. These mutations directly impacted sites 1 and 3 and three hot spot residues, which correspond to the most frequent mutations, from site 3 (Table 3).

Two mutations (N501 and K417) corresponded to hot spots and played an important role in the stabilization of the interaction between RDB and ACE2. This confirms that these mutations can directly affect the affinity of the RBD for ACE2 and the interactions with antibodies targeting the RBM. The mutation N501Y is observed in Alpha, Beta, Gamma, and Omicron variants, suggesting that it provides SARS-CoV-2 with a selective advantage. E484 and K417 mutations are present in the N501Y mutation in the Beta, Gamma, and Omicron variants. The N501Y mutation counterbalances the decrease in ACE2 affinity due to mutations in K417 and E484, whereas the latter tends to enhance the ability of the variant to escape from neutralizing antibodies. This explains why these three mutations were selected simultaneously during the SARS-CoV-2 evolution. More specifically, the N501 hot spot is stabilized by a hydrogen bond with residue Y41 and contributes favorably to the binding free energy of the complex (Fig. 1D and Figs. 2A and 3C). It is surrounded by a K353 hydrophobic alkyl chain and Y41 benzene ring. Understandably, a tyrosine switch is a better choice because a more favorable interaction can be made with the hydrophobic pocket, particularly π-π stacking with residue Y41. Thus, the mutation of N501 in Y501 can explain the enhanced affinity of the RBD to the ACE2 receptor [68].

Table 3
Mutations in RBD observed in the SARS-CoV-2 variants of concern. The hot spot residues identified by the MM-PBSA analyze are in bolt.

WHO label	Mutations in RBD	Impact on three selected sites
Alpha	**N501Y**	Site 3 (**N501Y**)
Beta	**K417N**, **E484K**, **N501Y**	Site 3 (**K417N**, **N501Y**)
Gamma	**K417T**, **E484K**, **N501Y**	Site 3 (**K417N**, **N501Y**)
Delta	L452R, T478K	No site
Omicron	G339D, S371L, S373P, S375F, **K417N**, N440K, G446S, S477N, T478K, **E484A**, Q493R, G496S, Q498R, **N501Y**, Y505H	Site 3 (**K417N**, **N501Y**, G496S, Q498R, Y505H) Site 1 (S375F)

E484 mutations are present with N501Y and K417 mutations in the Beta, Gamma, and Omicron variants. In the Beta and Gamma variants, the E484K mutation swaps the charge of the side chain. It is a considerable switch from negatively charged glutamic acid (−) to a positively charged lysine (+). In the Omicron variant, the E484A mutation also induces loss of the negative charge carried by the E residue. It is important to note that E484 is located on the RBD loop (amino acids 470–490) and is enclosed by several negatively charged residues, E35 and E75, from the ACE2 receptor. The position of three neighboring glutamic acids could be unfavorable due to repulsive forces; thus, introducing a positively charged residue or a small neutral residue might create a more favorable interaction between the two proteins. This may explain the unfavorable free energy decomposition of E484 during MM-PBSA analysis (Fig. 2 A). Similar to N501, E484 is a critical epitope residue for SARS-CoV-2 neutralizing antibodies. Therefore, charge change can also be a method to alter the electrostatic complementarity of known antibodies binding to this region, leading to better virus adaptation [69].

Concerning the K417 mutations, some studies showed that they may be responsible for increased binding with ACE2 and a decreased affinity for SARS-CoV-2 antibodies when combined with N501Y and E484K [70]. K417 forms a steady salt-bridge with residue D30 and has a favorable energy contribution to the complex (Figs. 1B, 2 A, and 3B). Consequently, abolishing this strong interaction decreases the binding affinity of the RBD and ACE2 complexes. However, the K417 mutation is only present with the N501 and E484 mutations, which may compensate for the loss of affinity with the ACE2 receptor by forming new favorable interactions. Additionally, a recent study showed that K417 is another critical epitope that forms strong salt bridges with SARS-CoV-2 antibody residues [68]. Thus, the reason behind the K417 abrupt change to asparagine is that viral adaptation is vital and overrides the binding affinity. In fact, with this triple mutation, SARS-CoV-2 may be harder to handle and can easily escape antibodies. The explanation for the K417 change to N or T may be a viral adaptation to decrease the affinity of antibodies for spike protein and evade the immune system. Regardless of the reasons for the K417 mutation, the replacement of a residue with a positive charge by residues with a hydrophilic side chain should be considered for drug design.

These mutations directly impacted three hot spot residues, which corresponded to the most frequent mutations. Thus, we analyzed the potential structural impact of RBD mutations of these variants on the three selected binding sites. These mutations directly affected sites 1 and 3. Additionally, site 3 would be greatly impacted by three hot spot residues, which correspond to the most common mutations in the variants of concern (Table 3).

For the Delta variant, L452 and T478 mutations did not occur in residues forming the highlighted sites and were far from the hot spot residues, indicating that inhibitors targeting these sites would not require adaptation to treat this variant.

The Omicron variant was found to be highly mutated. For example, 15 mutated residues were found in the RBD region alone (G339D, S371L, S373P, S375F, K417N, N440K, G446S, S477N, T478K, E484A, Q493R, G496S, Q498R, N501Y, and Y505H). Only the Omicron S375F mutation has been observed to impact site 1 (Fig. 6A). Currently, there is no available scientific information on how this mutation affects the SARS-CoV-2 life cycle. S375 was exposed to the solvent in both the open and closed states of the spike protein trimer

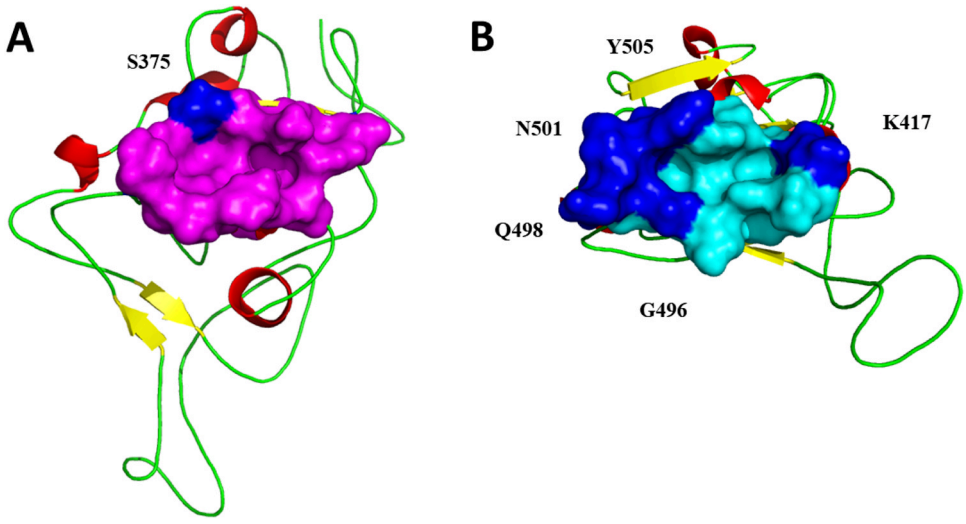


Fig. 6. Mutations, indicated in blue, observed in the variants of concern in A) site 1 and B) site 3. S375 mutation occurs in the Omicron variant. N501 mutation is observed in the Alpha, Beta, Gamma, and Omicron variants; K417 mutation is observed in the Beta, Gamma, and Omicron variants. G496, Q498, and Y505 mutations are specific to the Omicron variant.(For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

and did not form a non-covalent bond with any other residue. As a result, it does not appear to play a special role in the spike protein trimer action. Because the alcohol group in the spike residue was replaced by an aromatic cycle in the F residue, the S375F mutation only slightly altered the shape and hydrophobic properties of site 1. Therefore, site 1 seems to be relatively well-conserved among the variants of concern.

In summary, site 3 is located on the RBD-ACE2 interaction interface and is formed by several hot spot residues; therefore, it may be of interest to focus on this region with the aim of targeting it with inhibitory molecules.to focus on this region with the aim of targeting it with inhibitory molecules. However, because of several mutations that occur in different variants of concern, site 3 is difficult to block, in contrast to sites 1 and 2. Even though site 1 seems to be weakly affected by only one mutation, site 2 is currently not affected by any mutation. These two sites can be targeted by inhibitors that would be effective for all currently known variants.

Sites 1 and 2 are contiguous on the RBD surface. Interestingly, a neutralizing antibody isolated from convalescent Covid-19 patients and named CR3022, can bind to the RBD surface corresponding to sites 2 and 3 [71]. This interaction occurs only when two of the three spike proteins in the spike protein trimer are in an open state. Therefore, these sites are accessible to the inhibitory molecules. In the case of mutations that prevent the binding of neutralizing antibodies, targeting these sites by inhibitory molecules seems to be a promising therapeutic approach; notably, site 2 is currently not affected by any mutation.

4. Conclusions

In this study, we analyzed the dynamic behavior of the unbound RBD in complex with the ACE2 protein to identify binding sites that can be targeted by inhibitory molecules. We proposed a protocol combining MD simulations, hot spot identification, pocket tracking along the simulations, and pocket druggability prediction using a supervised method. Then, an unsupervised machine learning analysis was applied to cluster similar RBD pockets in terms of residue composition. This protocol allows the identification of druggable binding sites frequently observed along the MD simulations, the characterization of their residue flexibility and of key residues contributing to the druggability.

The stability of these pocket clusters was verified on different PDBs, in apo and holo forms, which allowed us to select the most pertinent druggable binding sites. Based on their RBD localization and druggability assessments, these three sites seem to be particularly promising. The potential effect of mutations in the variants of concern on these three binding sites was investigated.

Site 3 is located at the RBD-ACE2 protein interaction region and is formed by four hot spot residues. It is therefore an interesting target to disrupt the interaction between the spike protein and ACE2 protein and, consequently, to prohibit SARS-CoV-2 entry into the cell. However, several mutations of residues forming this site have been observed in SARS-CoV-2 variants. This suggests that if an inhibitory molecule can be designed against this site, it should be efficient for only a limited number of variants. Site 2, in which linoleic acid interacts to lock the spike protein in closed form, is relevant to the target. It is highly druggable, undergoes only one mutation in the Omicron variant, and is consequently an interesting site for targeting. Site 1 is highly druggable, accessible and no mutations were observed at this site. To our knowledge, this is the first time it has been characterized using in silico methods. In the spike protein trimer, we can observe that three sites 1 observed on each spike protein monomer are near in space (Fig. S6). This is also a promising target region. This suggests that the design of molecules able to bind to at least two sites 1 located within two RBD monomers may prohibit the transition between the inactive closed form and the active open form of the spike protein.

In summary, our combined protocol provides new insights and highlight opportunities on three binding sites for the development of inhibitors of the RBD of the spike protein.

Ethical approval

Not applicable.

CRediT authorship contribution statement

Mariem Ghoula: Conceptualization, Methodology, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualization. **Sarah Naceri:** Conceptualization, Methodology, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualization. **Samuel Sitruk:** Formal analysis, Investigation. **Delphine Flatters:** Conceptualization,

Methodology, Validation, Formal analysis, Writing – original draft, Writing – review & editing, Visualization. **Gautier Moroy:** Conceptualization, Methodology, Validation, Formal analysis, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition. **Anne-Claude Camproux:** Conceptualization, Methodology, Validation, Formal analysis, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

All authors declare no conflicts of interest.

Acknowledgments

This work was supported by the Agence Nationale de la Recherche (PIF21 Project no. ANR-19-CE18–0023). The authors gratefully acknowledge the financial support of the Université Paris Cité, CNRS Institute, and INSERM Institute. This study was performed using HPC resources from GENCI-CINES. The authors are thankful to Patrick Fuchs, Audrey Deyawe Kongmeneck, and Rachel Blot for their helpful comments on this manuscript.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2023.03.029.

References

[1] Coronaviridae Study Group of the International Committee on Taxonomy of VirusesThe species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2 Nat Microbiol 5 2020 536 544 doi: 10.1038/s41564-020-0695-z.

[2] Chen L, Liu W, Zhang Q, Xu K, Ye G, Wu W, Sun Z, Liu F, Wu K, Zhong B, Mei Y, Zhang W, Chen Y, Li Y, Shi M, Lan K, Liu Y. RNA based mNGS approach identifies a novel human coronavirus from two individual pneumonia cases in 2019 Wuhan outbreak. Emerg Microbes Infect 2020;9:313–9. https://doi.org/10.1080/22221751.2020.1725399

[3] Xie X, Liu Y, Liu J, Zhang X, Zou J, Fontes-Garfias CR, Xia H, Swanson KA, Cutler M, Cooper D, Menachery VD, Weaver SC, Dormitzer PR, Shi PY. Neutralization of SARS-CoV-2 spike 69/70 deletion, E484K and N501Y variants by BNT162b2 vaccine-elicited sera. Nat Med 2021;27:620–1. https://doi.org/10.1038/s41591-021-01270-4

[4] Tegally H, Wilkinson E, Giovanetti M, Iranzadeh A, Fonseca V, Giandhari J, Doolabh D, Pillay S, San EJ, Msomi N, Mlisana K, von Gottberg A, Walaza S, Allam M, Ismail A, Mohale T, Glass AJ, Engelbrecht S, Van Zyl G, Preiser W, Petruccione F, Sigal A, Hardie D, Marais G, Hsiao NY, Korsman S, Davies MA, Tyers L, Mudau I, York D, Maslo C, Goedhals D, Abrahams S, Laguda-Akingba O, Alisoltani-Dehkordi A, Godzik A, Wibmer CK, Sewell BT, Lourenço J, Alcantara LCJ, Kosakovsky Pond SL, Weaver S, Martin D, Lessells RJ, Bhiman JN, Williamson C, de Oliveira T. Detection of a SARS-CoV-2 variant of concern in South Africa. Nature 2021;592:438–43. https://doi.org/10.1038/s41586-021-03402-9

[5] Starr TN, Greaney AJ, Dingens AS, Bloom JD. Complete map of SARS-CoV-2 RBD mutations that escape the monoclonal antibody LY-CoV555 and its cocktail with LY-CoV016. Cell Rep Med 2021;2:100255. https://doi.org/10.1016/j.xcrm.2021.100255

[6] Giovanetti M, Benvenuto D, Angeletti S, Ciccozzi M. The first two cases of 2019–nCoV in Italy: Where they come from? J Med Virol 2020;92:518–21. https://doi.org/10.1002/jmv.25699

[7] Sallam M, Ababneh NA, Dababseh D, Bakri FG, Mahafzah A. Temporal increase in D614G mutation of SARS-CoV-2 in the Middle East and North Africa. Heliyon 2021;7:e06035. https://doi.org/10.1016/j.heliyon.2021.e06035

[8] Taylor PC, Adams AC, Hufford MM, de la Torre I, Winthrop K, Gottlieb RL. Neutralizing monoclonal antibodies for treatment of COVID-19. Nat Rev Immunol 2021;21:382–93. https://doi.org/10.1038/s41577-021-00542-x

[9] Hwang YC, Lu RM, Su SC, Chiang PY, Ko SH, Ke FY, Liang KH, Hsieh TY, Wu HC. Monoclonal antibodies for COVID-19 therapy and SARS-CoV-2 detection. J Biomed Sci 2022;29:1. https://doi.org/10.1186/s12929-021-00784-w

[10] Iketani S, Liu L, Guo Y, Liu L, Chan JF, Huang Y, Wang M, Luo Y, Yu J, Chu H, Chik KK, Yuen TT, Yin MT, Sobieszczyk ME, Huang Y, Yuen KY, Wang HH, Sheng Z, Ho DD. Antibody evasion properties of SARS-CoV-2 Omicron sublineages. Nature 2022;604:553–6. https://doi.org/10.1038/s41586-022-04594-4

[11] Choi S, Choi KY. Screening-based approaches to identify small molecules that inhibit protein–protein interactions. Expert Opin Drug Disco 2017;12:293–303. https://doi.org/10.1080/17460441.2017.1280456

[12] Hammond J, Leister-Tebbe H, Gardner A, Abreu P, Bao W, Wisemandle W, Baniecki M, Hendrick VM, Damle B, Simón-Campos A, Pypstra R, Rusnak JM. EPIC-HR investigators, oral nirmatrelvir for high-risk, nonhospitalized adults with covid-19. N Engl J Med 2022;386:1397–408. https://doi.org/10.1056/nejmoa2118542

[13] Duarte CM, Ketcheson DI, Eguíluz VM, Agustí S, Fernández-Gracia J, Jamil T, Laiolo E, Gojobori T, Alam I. Rapid evolution of SARS-CoV-2 challenges human defenses. Sci Rep 2022;12:6457. https://doi.org/10.1038/s41598-022-10097-z

[14] Mariano G, Farthing RJ, Lale-Farjat SLM, Bergeron JRC. Structural Characterization of SARS-CoV-2: Where We Are, and Where We Need to Be. Front Mol Biosci 2020;7:605236. https://doi.org/10.3389/fmolb.2020.605236

[15] Yang H, Rao Z. Structural biology of SARS-CoV-2 and implications for therapeutic development. Nat Rev Microbiol 2021;19:685–700. https://doi.org/10.1038/s41579-021-00630-8

[16] Walls AC, Park YJ, Tortorici MA, Wall A, McGuire AT, Veesler D. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. e6 Cell 2020;181:281–92. https://doi.org/10.1016/j.cell.2020.02.058

[17] Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, Wang W, Song H, Huang B, Zhu N, Bi Y, Ma X, Zhan F, Wang L, Hu T, Zhou H, Hu Z, Zhou W, Zhao L, Chen J, Meng Y, Wang J, Lin Y, Yuan J, Xie Z, Ma J, Liu WJ, Wang D, Xu W, Holmes EC, Gao GF, Wu G, Chen W, Shi W, Tan W. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet 2020;395:565–74. https://doi.org/10.1016/s0140-6736(20)30251-8

[18] Li W, Moore MJ, Vasilieva N, Sui J, Wong SK, Berne MA, Somasundaran M, Sullivan JL, Luzuriaga K, Greenough TC, Choe H, Farzan M. Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus. Nature 2003;426:450–4. https://doi.org/10.1038/nature02145

[19] Tai W, He L, Zhang X, Pu J, Voronin D, Jiang S, Zhou Y, Du L. Characterization of the receptor-binding domain (RBD) of 2019 novel coronavirus: implication for development of RBD protein as a viral attachment inhibitor and vaccine. Cell Mol Immunol 2020;17:613–20. https://doi.org/10.1038/s41423-020-0400-4

[20] Ling R, Dai Y, Huang B, Huang W, Yu J, Lu X, Jiang Y. In silico design of antiviral peptides targeting the spike protein of SARS-CoV-2. Peptides 2020;130:170328. https://doi.org/10.1016/j.peptides.2020.170328

[21] Pandey P, Rane JS, Chatterjee A, Kumar A, Khan R, Prakash A, Ray S. Targeting SARS-CoV-2 spike protein of COVID-19 with naturally occurring phytochemicals: an *in silico* study for drug development. J Biomol Struct Dyn 2021;39:6306–16. https://doi.org/10.1080/07391102.2020.1796811

[22] Hussain A, Hasan A, Nejadi Babadaei MM, Bloukh SH, Chowdhury MEH, Sharifi M, Haghighat S, Falahati M. Targeting SARS-CoV2 Spike Protein Receptor Binding Domain by Therapeutic Antibodies. Biomed Pharmacother 2020;130:110559. https://doi.org/10.1016/j.biopha.2020.110559

[23] Abi Hussein H, Geneix C, Petitjean M, Borrel A, Flatters D, Camproux AC. Global vision of druggability issues: applications and perspectives. Drug Discov Today 2017;22:404–15. https://doi.org/10.1016/j.drudis.2016.11.021

[24] Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh CL, Abiona O, Graham BS, McLellan JS. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science 2020;367:1260–3. https://doi.org/10.1126/science.abb2507

[25] Lan J, Ge J, Yu J, Shan S, Zhou H, Fan S, Zhang Q, Shi X, Wang Q, Zhang L, Wang X. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. Nature 2020;581:215–20. https://doi.org/10.1038/s41586-020-2180-5

[26] Wang Q, Zhang Y, Wu L, Niu S, Song C, Zhang Z, Lu G, Qiao C, Hu Y, Yuen KY, Wang Q, Zhou H, Yan J, Qi J. Structural and Functional Basis of SARS-CoV-2 Entry by Using Human ACE2. e9 Cell 2020;181:894–904. https://doi.org/10.1016/j.cell.2020.03.045

[27] Shang J, Ye G, Shi K, Wan Y, Luo C, Aihara H, Geng Q, Auerbach A, Li F. Structural basis of receptor recognition by SARS-CoV-2. Nature 2020;581:221–4. https://doi.org/10.1038/s41586-020-2179-y

[28] Genheden S, Ryde U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. Expert Opin Drug Disco 2015;10:449–61. https://doi.org/10.1517/17460441.2015.1032936

[29] Spinello A, Saltalamacchia A, Magistrato A. Is the Rigidity of SARS-CoV-2 Spike Receptor-Binding Motif the Hallmark for Its Enhanced Infectivity? Insights from All-Atom Simulations. J Phys Chem Lett 2020;11:4785–90. https://doi.org/10.1021/acs.jpclett.0c01148

[30] Delgado JM, Duro N, Rogers DM, Tkatchenko A, Pandit SA, Varma S. Molecular basis for higher affinity of SARS-CoV -2 spike RBD for human ACE2 receptor. Proteins 2021;89:1134–44. https://doi.org/10.1002/prot.26086

[31] Jawad B, Adhikari P, Podgornik R, Ching WY. Key interacting residues between RBD of SARS-CoV-2 and ACE2 receptor: combination of molecular dynamics simulation and density functional calculation. J Chem Inf Model 2021;61:4425–41. https://doi.org/10.1021/acs.jcim.1c00560

[32] Trigueiro-Louro J, Correia V, Figueiredo-Nunes I, Gíria M, Rebelo-de-Andrade H. Unlocking COVID therapeutic targets: a structure-based rationale against SARS-CoV-2, SARS-CoV and MERS-CoV Spike. Comput Struct Biotechnol J 2020;18:2117–31. https://doi.org/10.1016/j.csbj.2020.07.017

[33] Carino A, Moraca F, Fiorillo B, Marchianò S, Sepe V, Biagioli M, Finamore C, Bozza S, Francisci D, Distrutti E, Catalanotti B, Zampella A, Fiorucci S. Hijacking SARS-CoV-2/ACE2 receptor interaction by natural and semi-synthetic steroidal agents

acting on functional pockets on the receptor binding domain. Front Chem 2020;8:572885. https://doi.org/10.3389/fchem.2020.572885
[34] Gervasoni S, Vistoli G, Talarico C, Manelfi C, Beccari AR, Studer G, Tauriello G, Waterhouse AM, Schwede T, Pedretti A. A comprehensive mapping of the druggable cavities within the SARS-CoV-2 therapeutically relevant proteins by combining pocket and docking searches as implemented in pockets 2.0. Int J Mol Sci 2020;21:5152. https://doi.org/10.3390/ijms21145152
[35] Olotu FA, Omolabi KF, Soliman MES. Leaving no stone unturned: allosteric targeting of SARS-CoV-2 spike protein at putative druggable sites disrupts human angiotensin-converting enzyme interactions at the receptor binding domain. Inform Med Unlocked 2020;21:100451. https://doi.org/10.1016/j.imu.2020.100451
[36] Le Guilloux V, Schmidtke P, Tuffery P. Fpocket: an open source platform for ligand pocket detection. BMC Bioinforma 2009;10:168. https://doi.org/10.1186/1471-2105-10-168
[37] Stank A, Kokh DB, Fuller JC, Wade RC. Protein binding pocket dynamics. Acc Chem Res 2016;49:809–15. https://doi.org/10.1021/acs.accounts.5b00516
[38] Abi Hussein H, Geneix C, Cauvin C, Marc D, Flatters D, Camproux AC. Molecular dynamics simulations of influenza A virus NS1 reveal a remarkably stable RNA-binding domain harboring promising druggable pockets. Viruses 2020;12:537. https://doi.org/10.3390/v12050537
[39] Kuzmanic A, Bowman GR, Juarez-Jimenez J, Michel J, Gervasio FL. Investigating cryptic binding sites by molecular dynamics simulations. Acc Chem Res 2020;53:654–61. https://doi.org/10.1021/acs.accounts.9b00613
[40] Dokainish HM, Re S, Mori T, Kobayashi C, Jung J, Sugita Y. The inherent flexibility of receptor binding domains in SARS-CoV-2 spike protein. ELife 2022;11:e75720. https://doi.org/10.7554/elife.75720
[41] Borrel A, Regad L, Xhaard H, Petitjean M, Camproux AC. PockDrug: a model for predicting pocket druggability that overcomes pocket estimation uncertainties. J Chem Inf Model 2015;55:882–95. https://doi.org/10.1021/ci5006004
[42] Hussein HA, Borrel A, Geneix C, Petitjean M, Regad L, Camproux AC. PockDrug-Server: a new web server for predicting pocket druggability on holo and apo proteins. Nucleic Acids Res 2015;43:W436–42. https://doi.org/10.1093/nar/gkv462
[43] Li H, Robertson AD, Jensen JH. Very fast empirical prediction and rationalization of protein pKa values. Proteins 2005;61:704–21. https://doi.org/10.1002/prot.20660
[44] Abraham MJ, Murtola T, Schulz R, Páll S, Smith JC, Hess B, Lindahl E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX 2015;1:19–25. https://doi.org/10.1016/j.softx.2015.06.001
[45] Huang J, Rauscher S, Nawrocki G, Ran T, Feig M, de Groot BL, Grubmüller H, MacKerell AD. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. Nat Methods 2017;14:71–3. https://doi.org/10.1038/nmeth.4067
[46] Hess B, Bekker H, Berendsen HJC, Fraaije JGEM. LINCS: a linear constraint solver for molecular simulations. J Comput Chem 1997;18:1463–72. https://doi.org/10.1002/(SICI)1096-987X(199709)18:12%3C1463::AID-JCC4%3E3.0.CO;2-H
[47] Grubmüller H, Heller H, Windemuth A, Schulten K. Generalized verlet algorithm for efficient molecular dynamics simulations with long-range interactions. Mol Simul) 1991;6(1–3):121–42. https://doi.org/10.1080/08927029108022142
[48] Bussi G, Donadio D, Parrinello M. Canonical sampling through velocity rescaling. J Chem Phys 2007;126(1):014101. https://doi.org/10.1063/1.2408420
[49] Parrinello M, Rahman A. Polymorphic transitions in single crystals: a new molecular dynamics method. J Appl Phys 1981;52:7182–90. https://doi.org/10.1063/1.328693
[50] Humphrey W, Dalke A, Schulten K. VMD: Visual Molecular Dynamics. J Mol Graph 1996;14:33–8. https://doi.org/10.1016/0263-7855(96)00018-5
[51] R Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, 2020. https://www.R-project.org.
[52] T. Williams, C. Kelley, many others, GNUPLOT 5.2. An Interactive Plotting Program, 2019. ⟨http://www.gnuplot.info/⟩.
[53] Schrödinger, L.L.C. The PyMOL Molecular Graphics System, Version 2.0, 2015. https://pymol.org/2/.
[54] Vorobjev YN, Almagro JC, Hermans J. Discrimination between native and intentionally misfolded conformations of proteins: ES/IS, a new method for calculating conformational free energy that uses both dynamics simulations with an explicit solvent and an implicit solvent continuum model. Proteins 1998;32:399–413. https://doi.org/10.1002/(SICI)1097-0134(19980901)32:4%3C399::AID-PROT1%3E3.0.CO;2-C
[55] Daura X, Gademann K, Jaun B, van Gunsteren WF, Mark AE. Peptide folding: when simulation meets experiment. Angew Chem Int Ed 1999;38:236–40. https://doi.org/10.1002/(SICI)1521-3773(19990115)38:1/2%3C236::AID-ANIE236%3E3.0.CO;2-M
[56] Kumari R, Kumar R. Open source drug discovery consortium, A. Lynn, *g_mmpbsa* - A GROMACS Tool for High-Throughput MM-PBSA Calculations. J Chem Inf Model 2014;54:1951–62. https://doi.org/10.1021/ci500020m
[57] Kukic P, Farrell D, McIntosh LP, García-Moreno EB, Jensen KS, Toleikis Z, Teilum K, Nielsen JE. Protein dielectric constants determined from NMR chemical shift perturbations. J Am Chem Soc 2013;135:16968–76. https://doi.org/10.1021/ja406995j
[58] Murtagh F, Legendre P. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? J Classif 2014;31:274–95. https://doi.org/10.1007/s00357-014-9161-z
[59] Hahsler M, Hornik K, Buchta C. Getting things in order: an introduction to the R package seriation. J Stat Soft 2008;25:1–34. https://doi.org/10.18637/jss.v025.i03
[60] Mittal A, Manjunath K, Ranjan RK, Kaushik S, Kumar S, Verma V. COVID-19 pandemic: insights into structure, function, and hACE2 receptor recognition by SARS-CoV-2. PLoS Pathog 2020;16:e1008762. https://doi.org/10.1371/journal.ppat.1008762
[61] Yi C, Sun X, Ye J, Ding L, Liu M, Yang Z, Lu X, Zhang Y, Ma L, Gu W, Qu A, Xu J, Shi Z, Ling Z, Sun B. Key residues of the receptor binding motif in the spike protein of SARS-CoV-2 that interact with ACE2 and neutralizing antibodies. Cell Mol Immunol 2020;17:621–30. https://doi.org/10.1038/s41423-020-0458-z
[62] Veeramachaneni GK, Thunuguntla VBSC, Bobbillapati J, Bondili JS. Structural and simulation analysis of hotspot residues interactions of SARS-CoV 2 with human ACE2 receptor. J Biomol Struct Dyn 2021;39:4015–25. https://doi.org/10.1080/07391102.2020.1773318
[63] Delgado Blanco J, Hernandez-Alias X, Cianferoni D, Serrano L. In silico mutagenesis of human ACE2 with S protein and translational efficiency explain SARS-CoV-2 infectivity in different species. PLoS Comput Biol 2020;16:e1008450. https://doi.org/10.1371/journal.pcbi.1008450
[64] Othman H, Bouslama Z, Brandenburg JT, da Rocha J, Hamdi Y, Ghedira K, Srairi-Abid N, Hazelhurst S. Interaction of the spike protein RBD from SARS-CoV-2 with ACE2: similarity with SARS-CoV, hot-spot analysis and effect of the receptor polymorphism. Biochem Biophys Res Commun 2020;527:702–8. https://doi.org/10.1016/j.bbrc.2020.05.028
[65] Chakraborty S. Evolutionary and structural analysis elucidates mutations on SARS-CoV2 spike protein with altered human ACE2 binding affinity. Biochem Biophys Res Commun 2021;534:374–80. https://doi.org/10.1016/j.bbrc.2021.01.035
[66] Williams-Noonan BJ, Todorova N, Kulkarni K, Aguilar MI, Yarovsky I. An active site inhibitor induces conformational penalties for ACE2 recognition by the spike protein of SARS-CoV-2. J Phys Chem B 2021;125(10):2533–50. https://doi.org/10.1021/acs.jpcb.0c11321
[67] Toelzer C, Gupta K, Yadav SKN, Borucu U, Davidson AD, Kavanagh Williamson M, Shoemark DK, Garzoni F, Staufer O, Milligan R, Capin J, Mulholland AJ, Spatz J, Fitzgerald D, Berger I, Schaffitzel C. Free fatty acid binding pocket in the locked structure of SARS-CoV-2 spike protein. Science 2020;370:725–30. https://doi.org/10.1126/science.abd3255
[68] Luan B, Huynh T. Molecular mechanism of the N501Y mutation for enhanced binding between SARS-CoV-2′s spike protein and human ACE2 receptor. J Med Chem 2021;65:2820–6. https://doi.org/10.1021/acs.jmedchem.1c00311
[69] Weisblum Y, Schmidt F, Zhang F, DaSilva J, Poston D, Lorenzi JC, Muecksch F, Rutkowska M, Hoffmann HH, Michailidis E, Gaebler C, Agudelo M, Cho A, Wang Z, Gazumyan A, Cipolla M, Luchsinger L, Hillyer CD, Caskey M, Robbiani DF, Rice CM, Nussenzweig MC, Hatziioannou T, Bieniasz PD. Escape from neutralizing antibodies by SARS-CoV-2 spike protein variants. eLife 2020;9:e61312. https://doi.org/10.7554/elife.61312
[70] Gobeil SM, Janowska K, McDowell S, Mansouri K, Parks R, Stalls V, Kopp MF, Manne K, Li D, Wiehe K, Saunders KO, Edwards RJ, Korber B, Haynes BF, Henderson R, Acharya P. Effect of natural mutations of SARS-CoV-2 on spike structure, conformation and antigenicity. eabi6226 Science 2021;373. https://doi.org/10.1126/science.abi6226
[71] Yuan M, Wu NC, Zhu X, Lee CD, So RTY, Lv H, Mok CKP, Wilson IA. A highly conserved cryptic epitope in the receptor binding domains of SARS-CoV-2 and SARS-CoV. Science 2020;368:630–3. https://doi.org/10.1126/science.abb7269

Supplementary information

Table S1.
Summary table representing the hydrogen bonds and salt bridges interactions network in the 6M0J PDB structure. Backbone (bb) and side chain (sd) interactions were specified.

Donor	Acceptor	Distance	Type of interaction
K417(sd)	D30 (sd)	2.9 Å	Salt bridge
Q42 (sd)	G446(bb)	3.2 Å	Hydrogen bond
Y449(sd)	D38 (sd)	2.7 Å	Hydrogen bond
	Q42 (sd)	2.8 Å	Hydrogen bond
N487(sd)	Q24 (sd)	2.7 Å	Hydrogen bond
Y83 (sd)	N487(sd)	2.8 Å	Hydrogen bond
Y83 (sd)	Y489(sd)	3.5 Å	Hydrogen bond
K31 (sd)	Q493(sd)	3.6 Å	Hydrogen bond
Q493(sd)	E35 (sd)	3.5 Å	Salt bridge
K353(sd)	G496(bb)	3.1 Å	Hydrogen bond
Q42 (sd)	Q498(sd)	3.4 Å	Hydrogen bond
Y41 (sd)	T500(sd)	2.6 Å	Hydrogen bond
T500(sd)	D355(sd)	3.3 Å	Hydrogen bond
N501(sd)	Y41 (sd)	3.4 Å	Hydrogen bond
G502(sd)	K353(bb)	2.8 Å	Hydrogen bond
Y505(sd)	E37 (sd)	3.5 Å	Hydrogen bond
R393(sd)	Y505(sd)	3.7 Å	Hydrogen bond

Table S2.
The eight main clusters of pockets with (the average number of residues by pockets of the cluster), their residue composition (only residues observed in more than 15% of the clusters), their pocket occurrence on 1,000 RBD 6M0J conformations and their percentage of druggable pockets.

Cluster	(#Number of residue) Residues constituting the pocket	Occurrence	Druggable pocket frequency (%)
I	(12) R454, F456, R457, K458, S459, D467, S469, E471, I472, Y473, Q474, P491	894	0.2
II	(14) R355, V382, Y396, P426, D428, F429, T430, G431, F464, S514, F515, E516, L517, L518	479	99.
III Site 1	(17) S375, T376, K378, Y380, G404, D405, V407, R408, I410, A411, Q414, V433, A435, V503, G504, Y508, V510	524	85.9
IV	(17) P337, E340, V341, A344, T345, R346, F347, A348, N354, R355, K356, R357, I358, A397, D398, S399, V511	488	27.9
V	(26) L335, C336, P337, F338, G339, E340, V341, F342, N343, I358, D364, Y365, S366, V367, L368, S371, A372, S373, F374, V395, A397, I434, W436, VAL511, L513	873	93.0
VI	(13) T345, R346, F347, A348, S349, L441, D442, S443, K444, N448, N450, Y451, R509	323	7.4
VII Site 2	(24) W353, R355, N360, C361, V362, D364, Y365, S366, L368, Y369, F377, P384, T385, L387, N388, L390, F392, F464, R466, F515, A522, C525, G526	704	68.9
VIII Site 3	(30) R403, D405, E406, Q409, I410, A411, G413, Q414, T415, G416, K417, I418, A419, Y423, K424, L425, D427, N439, S443, Y453, S494, Y495, G496, F497, Q498, P499, N501, Y505, Q506, P507	780	21.7

Table S3.

The 16 pocket descriptors (5 geometrical and 9 physicochemical), the druggability score (and its standard deviation), as proposed by the PockDrug server **[37]**, for ten representative pockets by selected sites.

Conformation	Diameter hull	Polar residues	Smallest size	Nlys atom	Aromatic residues	Volume hull	Otyr atom	Number of residues	Surface hull	Ooh atom	Hydrophobic kyte	Radius cylinder	Aliphatic residues	Hydrophobic residues	Score Drugg	Confidence
Site 1																
a	20.5	0.5	7.3	0.0	0.1	777.3	0.05	15	490.3	0.05	0.26	10.2	0.26	0.8	0.9	0.04
b	18.1	0.5	7.4	0.0	0.1	746.0	0.05	14	460.2	0.05	0.30	8.7	0.28	0.7	0.9	0.04
c	19.3	0.5	7.4	0.0	0.1	763.2	0.04	15	486.4	0.02	0.26	9.4	0.26	0.8	0.9	0.04
d	18.2	0.5	8.6	0.0	0.1	998.9	0.04	16	555.6	0.04	0.02	9.0	0.25	0.7	0.8	0.06
e	20.4	0.5	7.4	0.0	0.1	858.0	0.04	15	513.9	0.04	0.26	10.2	0.26	0.8	0.9	0.03
f	17.7	0.5	8.6	0.0	0.1	806.5	0.04	15	484.9	0.04	0.26	8.8	0.26	0.8	0.9	0.04
g	19.1	0.5	9.0	0.0	0.2	933.2	0.04	17	541.4	0.02	0.18	9.7	0.23	0.7	0.9	0.03
h	19.8	0.5	9.3	0.0	0.1	864.4	0.04	15	516.5	0.04	0.26	9.8	0.26	0.8	0.9	0.04
i	20.5	0.5	8.4	0.0	0.1	876.4	0.04	15	523.8	0.04	0.26	10.2	0.26	0.8	0.9	0.03
j	19.1	0.5	8.0	0.0	0.1	767.6	0.04	15	474.3	0.02	0.26	9.4	0.26	0.8	0.9	0.04
Site 2																
a	15.7	0.6	9.1	0.0	0.2	635.1	0.02	14	406.7	0.025	0.78	7.5	0.28	0.7	0.9	0.00
b	19.7	0.4	10.6	0.0	0.2	1277.6	0.02	20	642.9	0.017	1.025	9.9	0.25	0.8	0.9	0.00
c	19.2	0.5	11.2	0.0	0.1	1207.5	0.00	19	633.4	0.016	1.03	9.1	0.31	0.7	0.9	0.00
d	17.3	0.5	9.5	0.0	0.2	921.4	0.00	17	509.7	0.0	1.01	8.1	0.29	0.7	0.9	0.01
e	14.4	0.6	8.2	0.0	0.2	547.3	0.00	14	367.1	0.057	0.78	6.9	0.28	0.7	0.9	0.01
f	13.6	0.5	7.9	0.0	0.2	508.5	0.00	13	344.5	0.0	1.11	6.8	0.30	0.7	0.9	0.01
g	17.2	0.5	9.9	0.0	0.2	975.8	0.00	18	540.7	0.04	0.73	8.7	0.27	0.7	0.9	0.01
h	14.2	0.4	6.8	0.0	0.2	423.8	0.03	11	313.8	0.0	0.91	7.0	0.36	0.7	0.9	0.00
i	14.4	0.5	9.2	0.0	0.2	662.7	0.03	13	415.9	0.03	0.65	7.3	0.31	0.6	0.9	0.00
j	17.8	0.6	8.6	0.0	0.2	728.1	0.00	16	448.3	0.0	0.64	8.8	0.25	0.6	0.9	0.01
Site 3																
a	15.8	0.7	6.3	0.0	0.3	360.4	0.08	10	302.8	0.0	-0.87	7.5	0.20	0.7	0.6	0.10
b	16.8	0.6	7.2	0.0	0.3	461.5	0.06	12	349.9	0.0	-0.78	8.4	0.10	0.6	0.7	0.06
c	15.2	0.7	6.4	0.0	0.3	337.5	0.08	10	302.1	0.0	-0.87	7.5	0.20	0.7	0.6	0.14
d	16.4	0.6	8.0	0.0	0.3	518.8	0.07	12	375.8	0.0	-0.78	8.2	0.10	0.6	0.7	0.08
e	13.4	0.7	7.3	0.0	0.4	425.2	0.11	10	312.4	0.0	-0.97	6.5	0.10	0.7	0.7	0.12
f	15.0	0.7	6.7	0.0	0.4	410.5	0.03	10	318.0	0.0	-0.97	7.5	0.10	0.7	0.6	0.08
g	18.2	0.6	8.7	0.0	0.3	626.8	0.07	12	420.5	0.0	-0.78	8.9	0.10	0.6	0.7	0.06
h	18.9	0.6	7.0	0.0	0.3	532.3	0.05	11	395.5	0.0	-0.78	9.4	0.10	0.7	0.7	0.02
i	17.2	0.6	7.9	0.0	0.3	486.2	0.06	12	359.1	0.0	-0.78	8.4	0.16	0.6	0.7	0.04
j	17.8	0.7	8.3	0.0	0.3	603.4	0.06	13	409.3	0.0	-0.99	8.7	0.15	0.6	0.5	0.09

Fig. S1.

Summary diagram of protocol of identification of druggable binding site of the SARS-CoV-2 receptor-binding domain in three steps. Step 1 concerns the protein three-dimensional flexibility analysis using molecular dynamics simulations and the identification of hotspot residues of the ACE2-RBD interactions using MMPBSA methods. Step 2 concerns RBD surface pocket tracking and druggability prediction by sample of 1,000 RBD conformations extracted from molecular dynamics simulations, pocket estimation and druggability prediction using PockDrug. Step 3 concerns the hierarchical clustering of all estimated RBD pockets in terms of residue similarity. This classification allows the identification of clusters of pockets with common residues, corresponding to binding sites frequently observed along the MD simulations. The variability in residues within a pocket cluster illustrates the residue flexibility of the corresponding binding site. Finally, this protocol successfully identified three druggable sites and their key residues aiming to target with inhibitors for preventing ACE2 interaction.

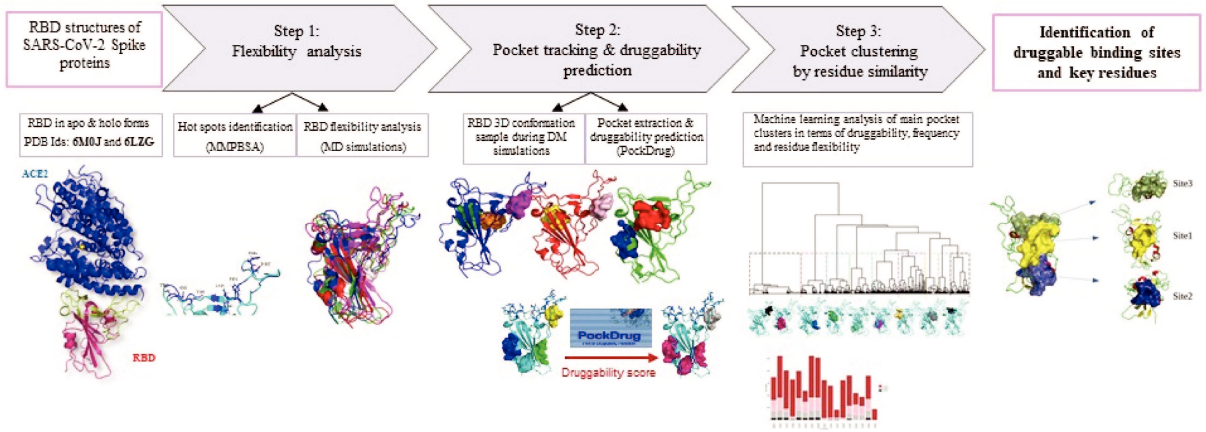


Fig. S2.

Hydrophobic, van der Waals, and π-π stacking interactions in the SARS-CoV-2 RBD and ACE2 complex (PDB: 6M0J): A) van der Waals interactions between F28, H34, Y489, Y453, and Q493, B) hydrophobic interactions with L45 and V445, G446, Q498 and Y505 with K353, C) hydrophobic interactions between T27, Y473, A475 and F476, D) van der Waals interaction with F28 and Y489, hydrophobic interactions between L79, F486 and Y489 and π-π stacking with Y83 and F486. ACE2 protein is colored in blue and RBD is colored in pink. Interactions were displayed using PyMOL. Only the sidechains of the residues are represented except for glycine residues.

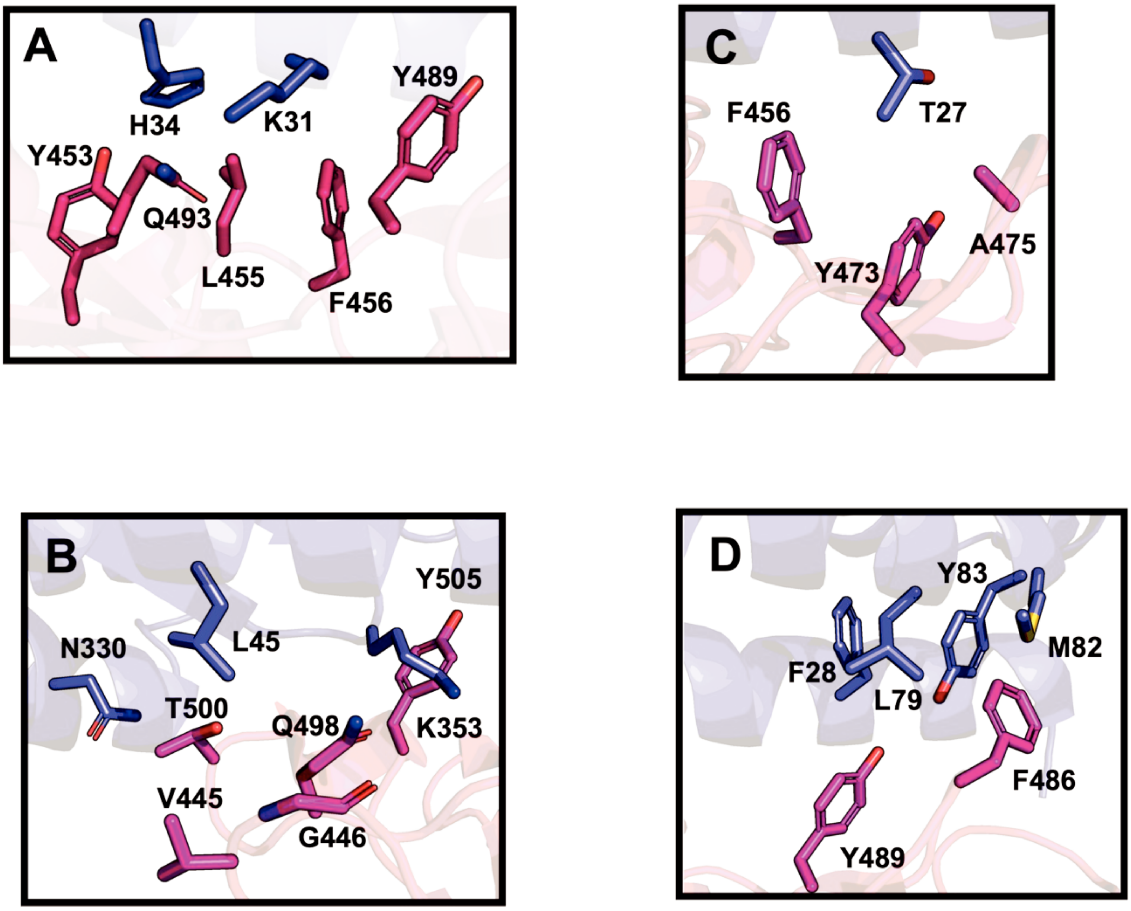


Fig. S3.

Evaluation of the C RMSD of A) 6M0J complex, B) bound ACE2, C) bound RBD and D) unbound RBD. RMSD fluctuations are colored in black and standard deviations are colored in grey.

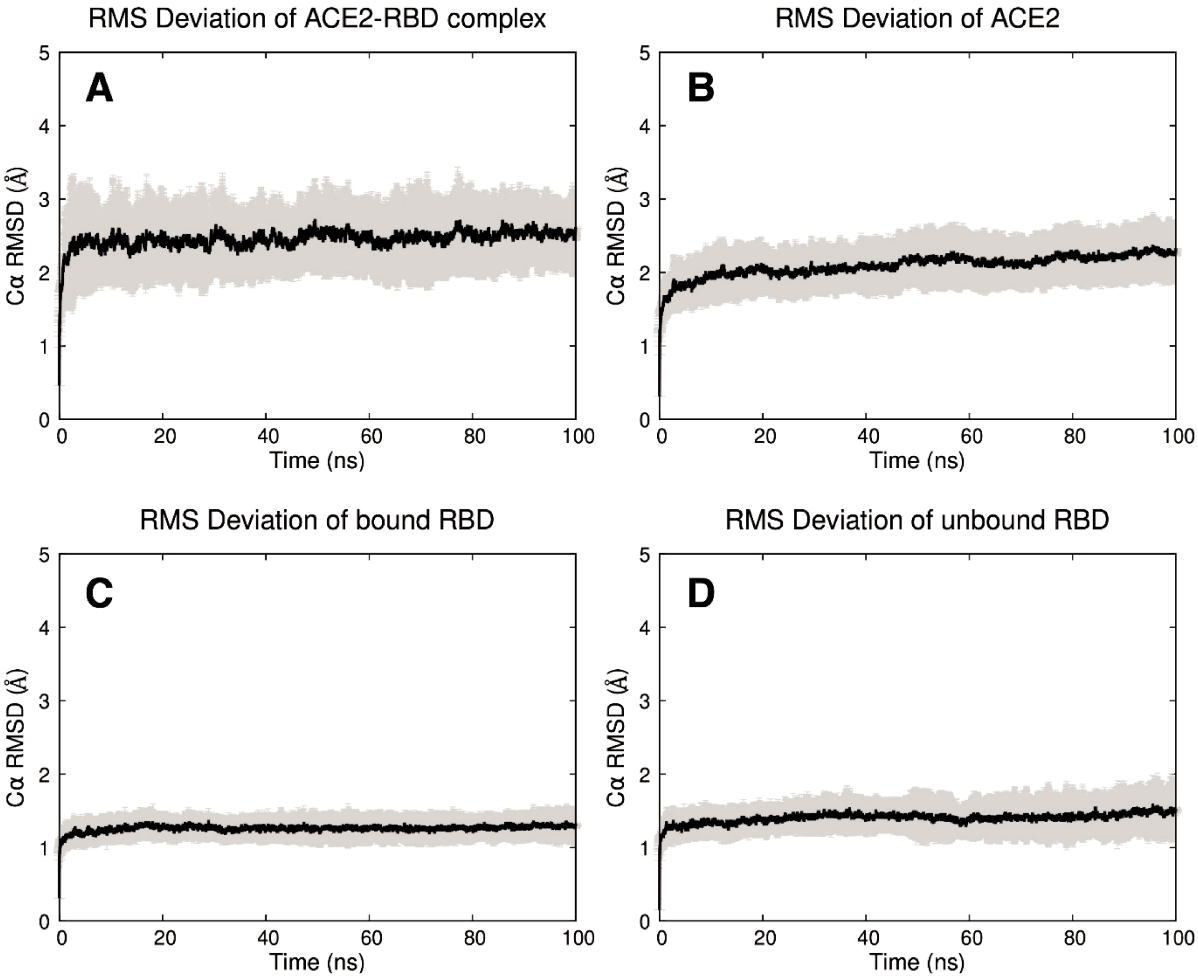


Fig. S4.

Evaluation of the C RMSD of A) 6LZG complex, B) bound ACE2, C) bound RBD and D) unbound RBD. RMSD fluctuations are colored in black and standard deviations in grey.

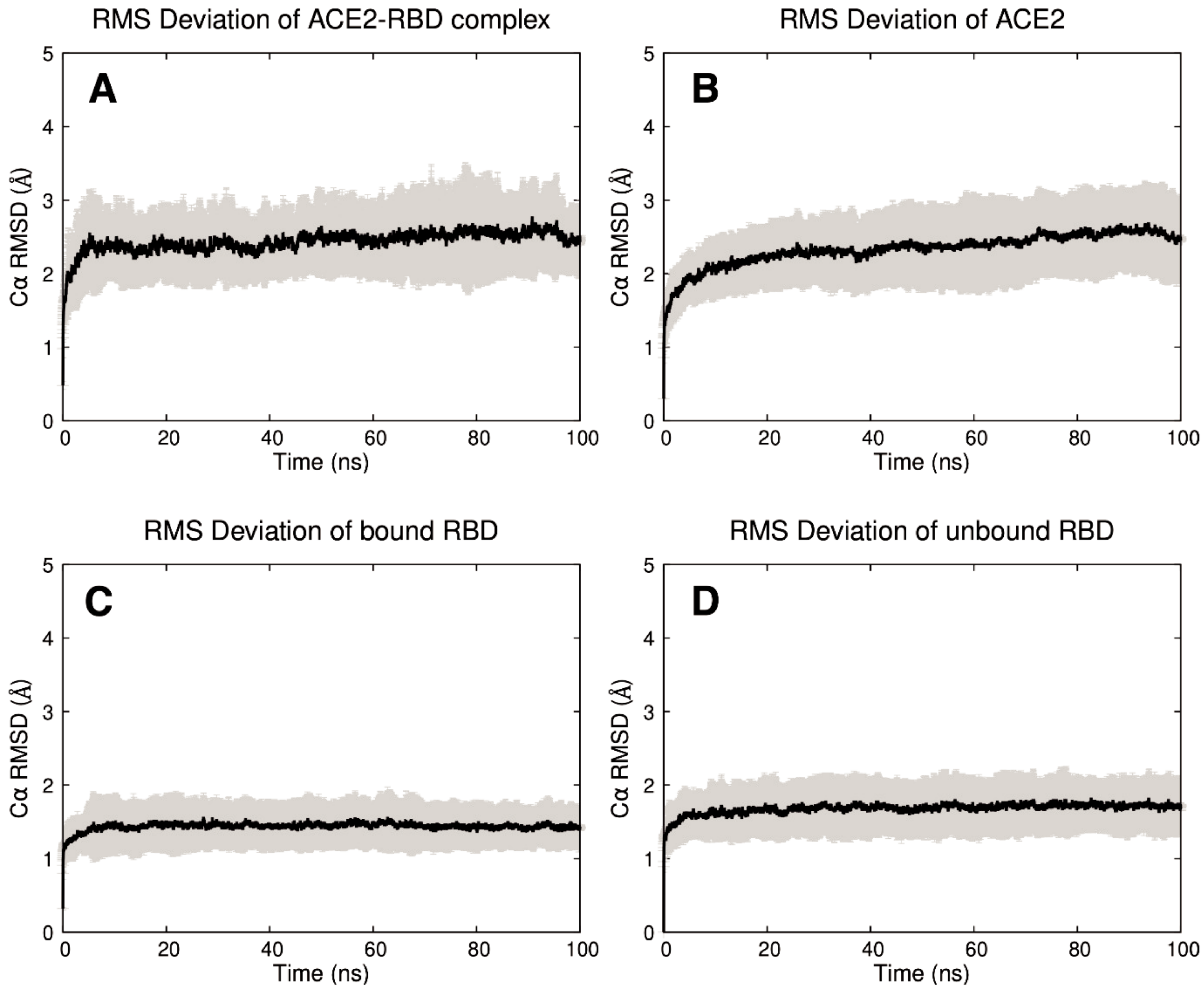


Fig. S5.

The RMS fluctuation of the 6M0J RBD in A) unbound and B) bound state and 6LZG RBD in C) unbound and D) bound state. RBM region is marked in black. Average RMSF is colored in black. Standard deviation is colored in red. B-factor is colored in blue.

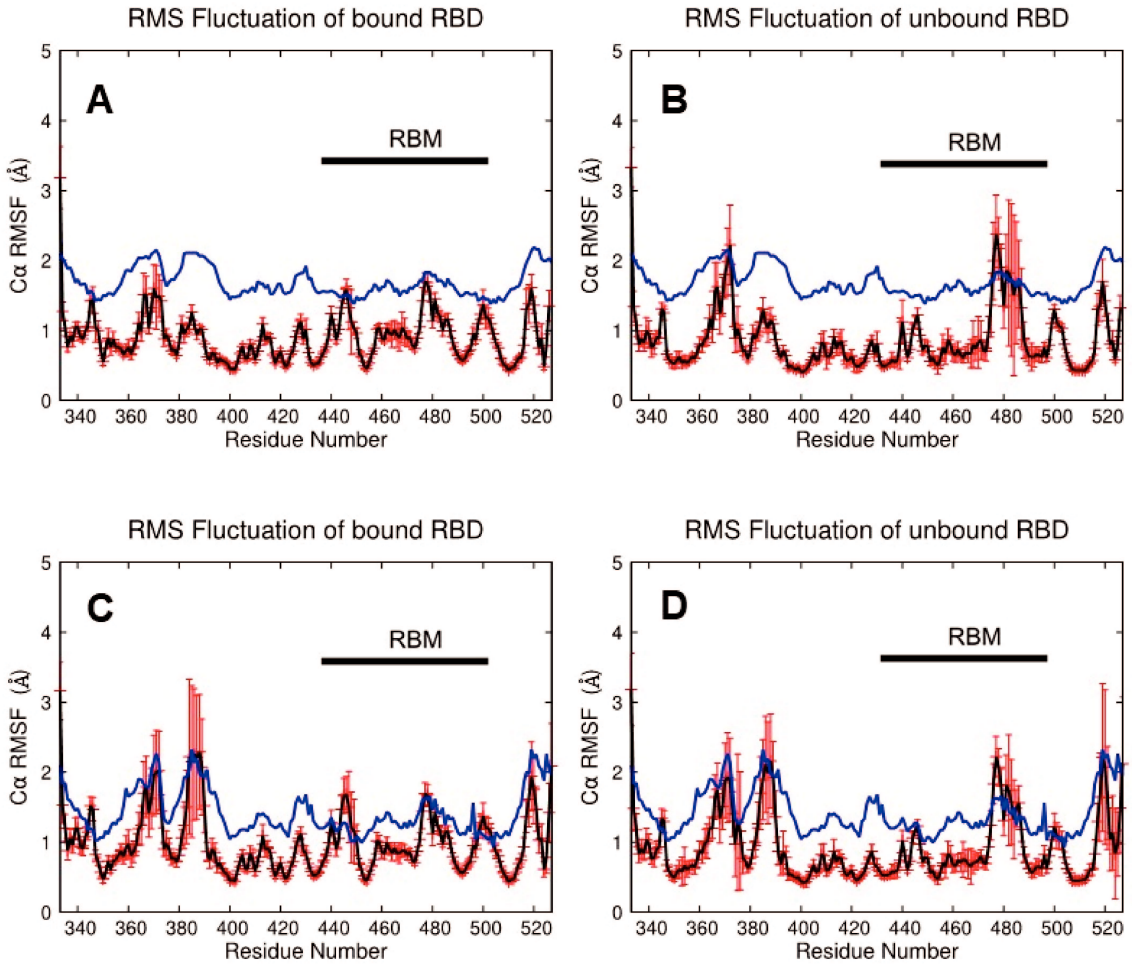


Fig. S6.
Surface representation of Spike trimer, each monomer is displayed in yellow, cyan, and green. Site 3 on each monomer is colored in red. The Spike trimer is A) in the closed state (PDB ID: 6VXX) and B) in the open state (PDB ID: 6VYB).

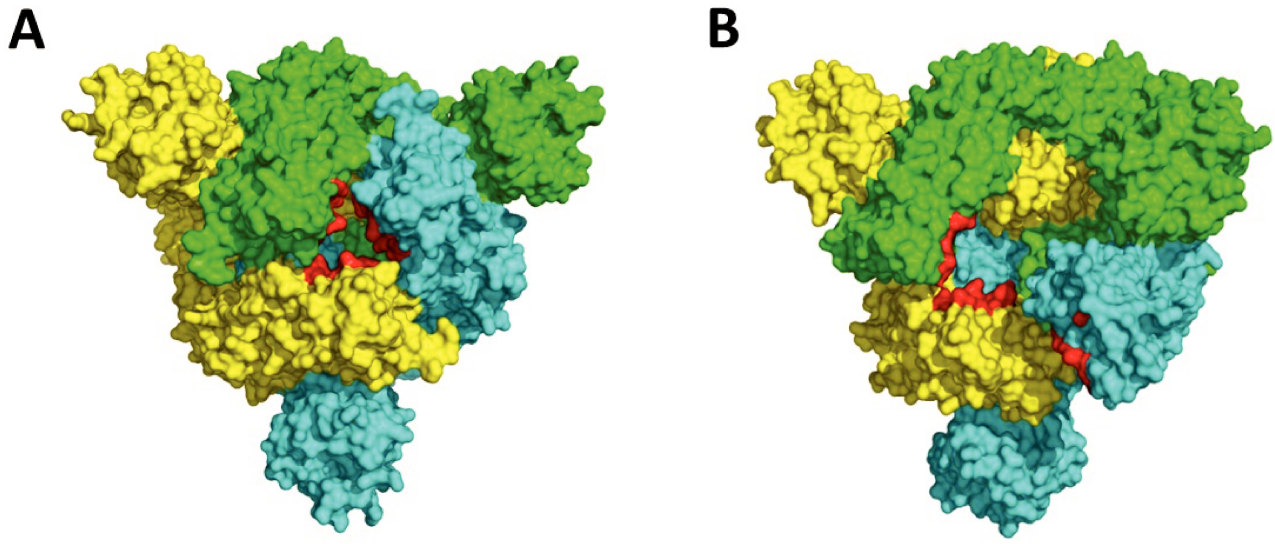
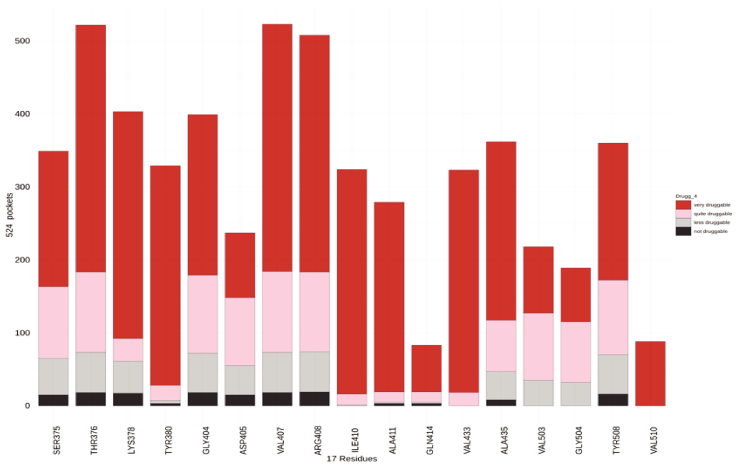


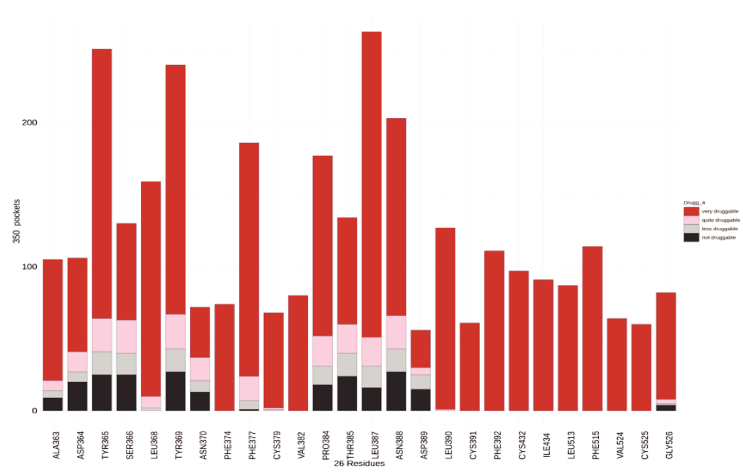
Fig. S7.

Histogram of residue composition of the pockets associated to each of the three sites. Only residues observed in more than 15% of their associated pockets are indicated. The bars of each residue are colored in proportion to the part of the pocket with 4 scores of druggability predicted using PockDrug **[37]**: in black, grey, pink and garnet for the non-druggable [0.00, 0.25], less druggable [0.25, 0.50], quite druggable [0.50, 0.75] and highly druggable [0.75, 1.00] pockets respectively, as

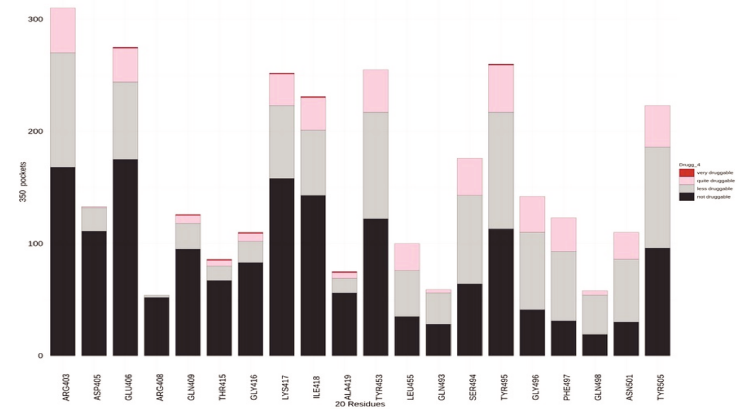
Site 1



Site 2



Site 3



CONCLUSION ET PERSPECTIVES

L'identification et la validation des cibles thérapeutiques sont des étapes cruciales pour le développement de thérapies efficaces. Les médicaments sont conçus pour interagir avec des cibles thérapeutiques au niveau des sites de liaison de manière spécifique afin de modifier leur activité et d'améliorer les symptômes des pathologies dans lesquelles elles sont impliquées. L'identification des sites de liaison expérimentalement peut être longue, coûteuse et dépendante de la présence de ligands. C'est pour cette raison que des approches en bioinformatique ont révolutionné les stratégies thérapeutiques en offrant des méthodes plus rapides et moins coûteuses pour l'identification des sites de liaison des cibles. De plus, il est possible de générer des modèles de prédiction à partir des séquences lorsque les structures de protéines définies expérimentalement ne sont pas disponibles. Les modèles prédictifs peuvent, par la suite, être utilisés pour prédire les propriétés des cavités à la surface des protéines, telles que la taille, la forme et l'accessibilité des poches avec une grande fiabilité.

Toutefois, pour caractériser avec précision des sites de liaison potentiels et suivre leur évolution au cours du temps en fonction de la flexibilité des protéines, il est primordial d'effectuer une recherche de poches en prenant en compte les propriétés dynamiques de ces structures car les protéines sont souvent soumises à des changements conformationnels importants pour remplir leur fonction biologique, telles que l'interaction avec d'autres protéines ou la liaison de ligands. Ces changements conformationnels peuvent modifier la forme et l'accessibilité des poches à la surface des protéines, ce qui peut avoir un impact important pour la liaison des ligands et pour le développement de médicaments.

Pour atteindre cet objectif, il a été nécessaire de développer une méthode qui combine les approches de recherche de poches et de flexibilité des protéines, afin de permettre un suivi efficace des résidus impliqués dans les poches identifiées à la surface des différents états conformationnels d'une protéine au cours de son mouvement.

Le protocole développé a été appliqué à la protéine NS1 du virus Influenza A. Son implication dans la réplication virale en fait une cible thérapeutique d'intérêt. Les travaux sur l'étude du polymorphisme structural de la protéine NS1 dans différents sous-types (H6N6, H1N1, H5N1) ont montré que les trois formes connues de la protéine NS1 (fermée, semi-ouverte, ouverte) étaient possibles et ont révélé une grande stabilité de la région RNA-BD quelle que soit la forme ou le sous-type. Le RNA-BD de l'un des sous-types (H6N6) de la protéine NS1 a fait l'objet d'une étude antérieure qui a montré une stabilité de sa structure au cours des trajectoires de DM. La présence d'un site de liaison potentiel au niveau de la région du sillon a également été mise en évidence grâce aux descripteurs physico-chimiques et géométriques.

Dans la continuité de ces travaux, une recherche de poches a été étendue aux autres sous-types de la protéine NS1, en utilisant des conformations obtenues à partir de simulations de DM précédemment réalisées pour étudier le polymorphisme de cette protéine. Cette approche a permis d'identifier, de façon ciblée, trois sites de liaison au niveau du sillon, dont deux de petites tailles conservés entre les différents sous-types de la protéine NS1. Ces résultats renforcent l'hypothèse selon laquelle cette protéine pourrait être une cible efficace pour le développement de médicaments antiviraux.

Une deuxième étude a été menée sur le RBD de la protéine Spike du virus responsable de la pandémie Covid-19, le SARS-CoV-2. Cette étude a examiné la flexibilité du RBD grâce à des simulations de DM. Une recherche de poches a été réalisée sur l'ensemble de la surface de la protéine, contrairement à l'étude précédente sur la protéine NS1, qui s'était concentrée sur une région spécifique de la protéine. Trois sites de liaison druggables ont été identifiés à différents endroits de la protéine Spike. Deux d'entre eux

avaient déjà été identifiés dans des études antérieures : le premier se trouve au niveau de l'interface d'interaction avec l'ACE2 et le deuxième est capable de se lier à un acide linoléique, ce qui maintient la protéine Spike en trimère dans sa forme inactive. Le troisième site de liaison, situé à l'interface de trimérisation du RBD, n'avait jamais été identifié jusqu'à présent.

En conclusion, pour mener à bien l'identification des sites de liaison des protéines *in silico*, il est crucial de disposer de structures 3D de haute qualité. Cela est dû au fait que la fonction des protéines dépend étroitement de leur conformation spatiale, qui est déterminée par leur structure 3D. La prise en compte de la flexibilité des protéines en bioinformatique nécessite des simulations de dynamique moléculaire, qui sont basées sur la résolution numérique des équations de mouvement pour chaque atome de la protéine. Cependant, la simulation de DM peut être une tâche complexe et nécessite des ressources informatiques importantes, en particulier pour les systèmes de grande taille. De plus, l'analyse des résultats de DM peut être difficile car elle produit une quantité importante de données.

Le choix de la méthode d'échantillonnage et le nombre de conformations extraites sont aussi des facteurs clés qui peuvent influencer les résultats de la recherche de poches. Le nombre de conformations extraites dépend de l'objectif de la simulation. Dans certains cas, quelques conformations suffisent pour obtenir des résultats significatifs, alors que dans d'autres cas, des milliers de conformations sont nécessaires pour couvrir l'espace conformationnel de manière adéquate. Il est donc important de prendre en compte ces facteurs lors de la planification et de l'exécution de simulations de dynamique moléculaire afin d'obtenir des résultats fiables et précis.

Il est également important de garder à l'esprit que les méthodes d’estimation des poches sont basées sur des approximations et des hypothèses, qui peuvent introduire des erreurs dans les prédictions. En effet, ces méthodes peuvent ne pas prendre en compte tous les facteurs pertinents pour la formation de poches ce qui peut conduire à des prédictions erronées. Il est donc important d'utiliser ces méthodes avec précaution et de les valider expérimentalement autant que possible. L'intégration des simulations de dynamique moléculaire permet de considérer un plus grand nombre de conformations de la protéine et sa flexibilité, ce qui peut aider à identifier et caractériser des liaisons non détectables dans des formes statiques.

Cette méthode permet de suivre les poches de liaison transitoires et allostériques qui peuvent se former lors des fluctuations de la protéine, ce qui permet de détecter des sites de liaison potentiels qui peuvent être ignorés par d'autres approches. En outre, la classification des poches en clusters avec des résidus communs peut aider à identifier les sites de liaison les plus pertinents, car ils sont observés fréquemment au cours des simulations de dynamique moléculaire. En utilisant cette méthode, les chercheurs peuvent également observer la flexibilité des résidus du site de liaison, ce qui peut aider à concevoir des médicaments qui peuvent s'adapter à des structures flexibles, permettant ainsi une meilleure efficacité du médicament. De plus, en combinant la recherche de poches avec l'analyse de la flexibilité des protéines et la druggabilité des sites de liaison, le protocole développé permet d’identifier des résidus clés des sites de liaison en terme d’implication dans la druggabilité des poches, leur fréquence d’apparition au cours du temps, l’impact des mutations, etc… Cette connaissance des résidus clés des poches peut ensuite guider les approches de docking sur les résidus clés et donc la recherche des molécules partenaires. À plus long terme, cette meilleure caractérisation des sites de liaison potentiels, de leur flexibilité et leur résidus clés peut contribuer à améliorer la recherche de médicaments.

En perspectives d’application, les sites de liaison potentiels et leur flexibilité qui ont été identifiés à des régions ciblées (tel qu'au niveau du RNA-BD de la protéine NS1) ou sur l'ensemble de la surface protéique (tel que pour le RBD de la protéine Spike), peuvent faire

l'objet d'études plus poussées, notamment par criblage virtuel de bibliothèques de molécules existantes ou de molécules nouvellement conçues à l'aide de la modélisation moléculaire.

Enfin, le protocole développé sera optimisé et compilé sous la forme d'un site web et mis à la disposition de la communauté scientifique, ce qui permettra une identification plus efficace des sites de liaison pour la conception de médicaments.

BIBLIOGRAPHIE

Abraham, M., Murtola, T. J., Schulz, R., Páll, S., Smith, J. C., Hess, B., & Lindahl, E. (2015). GROMACS : High performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX, 1-2, 19-25.

Alberts, B. (2002). Protein Function. Molecular Biology of the Cell - NCBI Bookshelf.

Bairoch, A. M., & Boeckmann, B. (1993). The SWISS-PROT protein sequence data bank, recent developments. Nucleic Acids Research, 21(13), 3093-3096.

Barker, W. W. (1998). The PIR-International Protein Sequence Database. Nucleic Acids Research, 26(1), 27-32.

Benet, L. Z., Hosey, C. M., Ursu, O., & Oprea, T. I. (2016). BDDCS, the Rule of 5 and drugability. Advanced Drug Delivery Reviews, 101, 89-98.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2009). GenBank. Nucleic Acids Research, 38(suppl_1), D46-D51.

Berendsen, H. J. C., Van Der Spoel, D., & Van Drunen, R. (1995). GROMACS : A message-passing parallel molecular dynamics implementation. Computer Physics Communications, 91(1-3), 43-56.

Berman, H. M., Westbrook, J. D., Feng, Z., Gilliland, G. L., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The Protein Data Bank. Nucleic Acids Research, 28(1), 235-242.

Binkowski, T. A., Naghibzadeh, S. E., & Liang, J. (2003). CASTp : Computed Atlas of Surface Topography of proteins. Nucleic Acids Research, 31(13), 3352-3355.

Bornholdt, Z. A., & Nagabhushana, H. (2008). X-ray structure of NS1 from a highly pathogenic H5N1 influenza virus. Nature, 456(7224), 985-988.

Carrillo, B., Choi, J. Y., Bornholdt, Z. A., Sankaran, B., Rice, A. S., & Nagabhushana, H. (2014). The Influenza A Virus Protein NS1 Displays Structural Polymorphism. Journal of Virology, 88(8), 4113-4122.

Chan, J., Nislow, C., & Emili, A. (2010). Recent advances and method development for drug target identification. Trends in Pharmacological Sciences, 31(2), 82-88.

Chen, Z., Zhang, X., Peng, C., Wang, J., Xu, Z., Chen, K., Shi, J., & Zhu, W. (2019). D3Pockets : A Method and Web Server for Systematic Analysis of Protein Pocket Dynamics. Journal of Chemical Information and Modeling, 59(8), 3353-3358.

Ealy, J. B., Abouomar, N., Cogan, J. W., Flauta, P., Nassar, L., Mekolochik, M., Ramzy, S. R., Shannon, C., & Yazgi, H. (2017). Estimated Binding Energies of Drug-Like and Nondrug-Like Molecules in the Active Site of HIV-1 Integrase, 1BIS.pdb, and Two Mutant Models : Y143R and N155H. Advances in Bioscience and Biotechnology.

Garaigorta, U., & Ortín, J. (2007). Mutation analysis of a recombinant NS replicon shows that influenza virus NS1 protein blocks the splicing and nucleo-cytoplasmic transport of its own viral mRNA. Nucleic Acids Research, 35(14), 4573-4582.

Guilloux, V. L., Schmidtke, P., & Tufféry, P. (2009). Fpocket : An open source platform for ligand pocket detection. BMC Bioinformatics, 10(1).

Håkansson, A., Zhivotovsky, B., Orrenius, S., Sabharwal, H., & Svanborg, C. (1995). Apoptosis induced by a human milk protein. Proceedings of the National Academy of Sciences of the United States of America, 92(17), 8064-8068.

Hale, B. G., Randall, R. E., Ortín, J., & Jackson, D. A. (2008). The multifunctional NS1 protein of influenza A viruses. Journal of General Virology, 89(10), 2359-2376.

Hendlich, M., Rippmann, F., & Barnickel, G. (1997). LIGSITE : automatic and efficient detection of potential small molecule-binding sites in proteins. Journal of Molecular Graphics & Modelling, 15(6), 359-363.

Hussein, H. A., Borrel, A., Geneix, C., Petitjean, M., Regad, L., & Camproux, A. (2015). PockDrug-Server : a new web server for predicting pocket druggability on holo and apo proteins. Nucleic Acids Research, 43(W1), W436-W442.

Hussein, H. A., Geneix, C., Cauvin, C., Marc, D., Flatters, D., & Camproux, A. (2020). Molecular Dynamics Simulations of Influenza A Virus NS1 Reveal a Remarkably Stable RNA-Binding Domain Harboring Promising Druggable Pockets. Viruses, 12(5), 537.

Jumper, J. M., Evans, R., Pritzel, A., Green, T. J., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A. M., Romera-Paredes, B., Nikolov, S., Jain, R. D., Adler, J.,. . . Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583-589.

Karplus, M., & Petsko, G. A. (1990). Molecular dynamics simulations in biology. Nature, 347(6294), 631-639.

Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckoff, H., & Phillips, D. (1958). A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis. Nature, 181(4610), 662-666.

Köhler, C., Gogvadze, V., Håkansson, A., Svanborg, C., Orrenius, S., & Zhivotovsky, B. (2001). A folding variant of human α-lactalbumin induces mitochondrial permeability transition in isolated mitochondria. European journal of biochemistry, 268(1), 186-191.

Koliopoulos, M. G., Lethier, M., Van Der Veen, A. G., Haubrich, K., Jung, K., Kowalinski, E., Stevens, R., Martin, S. F., Sousa, C. R. E., Cusack, S., & Rittinger, K. (2018). Molecular mechanism of influenza A NS1-mediated TRIM25 recognition and inhibition. Nature Communications, 9(1).

Kuzmanic, A., Bowman, G. R., Juárez-Jiménez, J., Michel, J., & Gervasio, F. L. (2020). Investigating Cryptic Binding Sites by Molecular Dynamics Simulations. Accounts of Chemical Research, 53(3), 654-661.

Lipinski, C. A., Lombardo, F., Dominy, B. W., & Feeney, P. W. (1997). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Advanced Drug Delivery Reviews, 23(1-3), 3-25.

Matsumoto, H., Shimokawa, Y., Ushida, Y., Toida, T., & Hayasawa, H. (2001). New Biological Function of Bovine α-Lactalbumin : Protective Effect against Ethanol- and Stress-induced Gastric Mucosal Injury in Rats. Bioscience, Biotechnology, and Biochemistry, 65(5), 1104-1111.

Matthews, B. C. (1976). X-ray Crystallographic Studies of Proteins. Annual Review of Physical Chemistry, 27(1), 493.

Murtagh, F., & Legendre, P. (2014). Ward's Hierarchical Agglomerative Clustering Method : Which Algorithms Implement Ward's Criterion ? Journal of Classification, 31(3), 274-295.

Parrinello, M., & Rahman, A. (1981). Polymorphic transitions in single crystals : A new molecular dynamics method. Journal of Applied Physics, 52(12), 7182-7190.

Qian, X., Alonso-Caplen, F. V., & Krug, R. M. (1994). Two functional domains of the influenza virus NS1 protein are required for regulation of nuclear export of mRNA. Journal of Virology, 68(4), 2433-2441.

Santos, R., Ursu, O., Gaulton, A., Bento, A. P., Donadi, R. S., Bologa, C., Karlsson, A., Al-Lazikani, B., Hersey, A., Oprea, T. I., & Overington, J. P. (2017). A comprehensive map of molecular drug targets. Nature Reviews Drug Discovery, 16(1), 19-34.

Svensson, M., Håkansson, A., Mossberg, A., Linse, S., & Svanborg, C. (2000). Conversion of α-lactalbumin to a protein inducing apoptosis. Proceedings of the National Academy of Sciences of the United States of America, 97(8), 4221-4226.

Trigueiro-Louro, J. M., Correia, V. G., Figueiredo-Nunes, I., Gíria, M., & Rebelo-De-Andrade, H. (2020). Unlocking COVID therapeutic targets : A structure-based rationale against SARS-CoV-2, SARS-CoV and MERS-CoV Spike. Computational and structural biotechnology journal, 18, 2117-2131.

Volkamer, A., Kuhn, D., Rippmann, F., & Rarey, M. (2012). DoGSiteScorer : a web server for automatic binding site prediction, analysis and druggability assessment. Bioinformatics, 28(15), 2074-2075.

Wagner, J., Sørensen, J., Hensley, N., Wong, C. C. Y., Zhu, C., Perison, T., & Amaro, R. E. (2017). POVME 3.0 : Software for Mapping Binding Pocket Flexibility. Journal of Chemical Theory and Computation, 13(9), 4584-4592.

Wallace, A. M., Laskowski, R. A., & Thornton, J. M. (1995). LIGPLOT : a program to generate schematic diagrams of protein-ligand interactions. Protein Engineering Design & Selection, 8(2), 127-134.

Waterhouse, A. L., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F. T., De Beer, T. A. P., Rempfer, C., Bordoli, L., Lepore, R., & Schwede, T. (2018). SWISS-MODEL : homology modelling of protein structures and complexes. Nucleic Acids Research, 46(W1), W296-W303.

Wüthrich, K. (1990). Protein structure determination in solution by NMR spectroscopy. Journal of Biological Chemistry, 265(36), 22059-22062.

Yang, J., & Zhang, Y. (2015). I-TASSER server : new development for protein structure and function predictions. Nucleic Acids Research, 43(W1), W174-W181.

Zhirnov, O. P., Konakova, T. E., Wolff, T., & Klenk, H. (2002). NS1 Protein of Influenza A Virus Down-Regulates Apoptosis. Journal of Virology, 76(4), 1617-1625.