Université de Montréal

Développement et optimisation de potentiels simplifiés de la famille OPEP et étude de molécules thérapeutiques contre la COVID-19

par

Vincent Binette

Département de physique
Faculté des arts et des sciences

Thèse présentée en vue de l'obtention du grade de
Philosophiæ Doctor (Ph.D.)
en Physique

June 12, 2022

Université de Montréal

Faculté des arts et des sciences

Cette thèse intitulée

Développement et optimisation de potentiels simplifiés de la famille OPEP et étude de molécules thérapeutiques contre la COVID-19

présentée par

Vincent Binette

a été évaluée par un jury composé des personnes suivantes :

Rikard Blunck
(président-rapporteur)

Normand Mousseau
(directeur de recherche)

Rafael Najmonovich
(membre du jury)

Patrick Lagüe
(examinateur externe)

(représentant du doyen de la FESP)

Résumé

La bio-modélisation numérique est un domaine hautement multidisciplinaire à la frontière entre la biologie, la physique, les mathématiques et l'informatique. Il s'agit d'un domaine en pleine effervescence grâce à une habile exploitation des avancées informatiques et algorithmiques. Parmi ses sujets d'étude, on retrouve les protéines, des molécules de grand intérêt. En effet, elles sont des nanomachines jouant des fonctions primordiales pour la survie de tout organisme. En plus de leurs fonctions naturelles, certaines protéines sont associées au développement de diverses maladies ou pourraient servir de molécules thérapeutiques. La vision traditionnelle de la biologie moléculaire stipule que les fonctions des protéines sont étroitement liées à leur structure tri-dimensionnelle elle-même déterminée par les propriétés physico-chimiques de la séquence en acides aminés. Ainsi, l'étude de la structure est indispensable. Les méthodes de la bio-modélisation numérique, en partenariat avec l'expérience, sont particulièrement appropriées pour l'étude des protéines. Cette thèse s'articulera donc autour de trois classes de méthodes qui permettent d'étudier divers aspects des protéines.

Le premier chapitre présentera les améliorations apportées à PEP-FOLD, une méthode simplifiée pour la prédiction structurelle *de novo* des petits peptides. Deux des trois éléments-clés de PEP-FOLD ont été peaufinés, l'alphabet structurel et le potentiel gros-grain sOPEP, avec comme résultat une amélioration de la qualité des prédictions. Cette nouvelle version est comparée aux méthodes de prédictions utilisant les plus récents développements de l'apprentissage machine. Le second chapitre présentera les résultats de simulations numériques sut deux petites molécules thérapeutiques contre la COVID-19, grâce à des méthodes basées sur la physique. En collaboration avec les résultats expérimentaux, nos simulations montrent que nos deux molécules pourraient prévenir des interactions cruciales pour l'émergence de la maladie. Finalement, le dernier chapitre présentera quelques résultats préliminaires au développement du potentiel simplifié aaOPEP, qui permettra d'étudier les processus d'agrégation et de fibrillation de la protéine β-amyloïde, associés à l'apparition de la maladie d'Alzheimer. Ce processus étant fondamentalement multi-échelle, au niveau spatial et temporel, le développement de méthodes simplifiées est essentiel pour obtenir le portrait global du phénomène.

Mots-clés: protéine, prédiction structurelle, simulation numérique, COVID-19, aaOPEP

Abstract

Molecular modeling is a multidisciplinary enterprise combining the fields of biology, physics, mathematics and informatics. By utilizing improvements in both computer hardware and algorithms, the field is experiencing a spectacular growth in the past two decades. Proteins are nanomachines and play multiple essential functions in the life of every organisms. Additionally, proteins are also associated with the emergence of many diseases and could also be used as therapeutic molecules. In the classical view of molecular biology, protein's functions are closely related to its tri-dimensional structure, which is encoded by the chemical properties of the amino acid sequence. Therefore, the study of protein's structure is of fundamental importance. Tools from molecular modeling are, in partnership with experimental techniques, very well suited to study proteins. The following thesis will be divided into three main classes of techniques, each able to study a wide range of protein's characteristics.

The first chapter present improvements made to PEP-FOLD, a simplified, freely-available online and successful technique for *de novo* peptide-structure prediction. Two of the three key components of PEP-FOLD were revisited in this work; the structural alphabet and the coarse-grained potential sOPEP. These modifications lead to an important increase in the quality of PEP-FOLD's predictions. A thorough comparison with state-of-the-art machine learning techniques is made and we highlight key successes and possible future improvements. The second chapter present the study of potential therapeutic molecules against COVID-19 using physics-based techniques. These results, combined with experimental data from immunobinding assay and SPR microscopy, showed that our two small molecules could prevent key interactions between the wild-type/mutant SARS-CoV-2 and the cells of the host and therefore could potentially be potent therapeutic molecules against COVID-19. Finally, the last chapter present preliminary results about the development of the new aaOPEP forcefield designed to study the multi-scale process of amyloid-β aggregation and fibrillation, associated with the Alzheimer disease. This new potential will take the core ideas of the coarse-grained potential OPEP into the all-atom regime and will allow to study bigger systems over longer time-scale.

Keywords: protein, structure prediction, molecular dynamics, COVID-19, aaOPEP

Résumé 5

Abstract 7

Liste des sigles et des abréviations 33

Remerciements 35

Introduction 37

Bio-modélisation numérique: en effervescence 37

Thermodynamique vs dynamique/Photo vs film 38

Rapidité vs complexité 39

Aperçu 40

Chapitre 1. Biologie pour l'étude de SARS-CoV-2 43

1.1. Maladie 43

1.2. Fonctionnement du virus 44

1.3. Développement thérapeutique 46

1.4. Variants 47

1.5. Contexte 48

Chapitre 2. Méthode pour l'étude de SARS-CoV-2 51

2.1. Mécanique moléculaire 51

2.1.1. Champ de force 52

2.1.1.1. Interactions liées 52

2.1.1.2. Interactions non-liées 53

2.1.1.3. Protéine et Ligand 54

2.1.1.4. Solvant 56

2.2. Dynamique moléculaire 57

2.2.1. Intégration 58

2.2.2. Thermostat et Barostat 58

2.2.3. Contraintes 60

2.3. Minimisation de l'énergie 62

2.4. Interactions protéines/ligands 63

2.4.1. Amarrage moléculaire 64

2.4.2. MM/PBSA 66

2.4.3. Méthodes expérimentales ... 70
2.4.3.1. Essais de liaison de ligand ... 70
2.4.3.2. Biosenseur par résonance des plasmons de surface ... 72

Chapitre 3. Corilagin and 1,3,6-Tri-O-galloy-β-D-glucose: potential inhibitors of SARS-CoV-2 variants ... 77

3.1. Introduction ... 79

3.2. Materials and Methods ... 81
3.2.1. MD simulations ... 81
Identification of the interactions between ACE2 and RBD in the complex ... 82
Assessing the structural flexibility of the starting ACE2 and RBD structures ... 83
3.2.2. Molecular docking ... 84
3.2.3. Protein-ligand simulations ... 84
3.2.4. Analysis ... 85
3.2.5. Binding free-energy ... 85
3.2.6. Products ... 86
3.2.7. Surface Plasmon Resonance (SPR) ... 86
3.2.8. SARS CoV-2 RBD Spike Protein and Human ACE2 Binding Inhibitor Assay 87
3.2.9. TGG, corilagin and human ACE2 binding assay ... 87
3.2.10. Statistical analysis for binding assays ... 88

3.3. Results ... 88
3.3.1. The impact of Corilagin and TGG on the ACE/RBD Wild-Type interactions 88
3.3.1.1. Molecular Docking ... 88
3.3.1.2. Molecular Dynamics ... 88
3.3.1.3. Interactions and Binding Energies ... 89
3.3.1.4. Corilagin/TGG ability to disrupt the ACE2/wild-type RBD interactions 90
3.3.1.5. Surface Plasmon Resonance ... 93
3.3.1.6. Binding assays ... 94
3.3.2. The impact of Corilagin and TGG on mutant RBDs ... 96
The impact of the mutations on RBD's stability ... 97
3.3.3. Corilagin and TGG interactions with ACE2 and RBD ... 98
3.3.3.1. Molecular Docking ... 98
3.3.3.2. Molecular Dynamics ... 99
3.3.3.3. Interactions and Binding Energies ... 100
3.3.3.4. Corilagin ability to disrupt the ACE2/mutated RBD interactions ... 101
3.3.3.5. TGG ability to disrupt the ACE2/mutated RBD interactions ... 102

3.4. Discussion ... 103

3.4.1. Best ligand targeting ACE2 and the wild-type RBD . . . 103
3.4.2. The impact of the mutation of the RBD and therapeutic potential of the ligands . . . 104
3.4.2.1. The impacts of RBD mutations on its structural ensemble . . . 104
3.4.2.2. Therapeutic potential of the ligands . . . 105
Conclusion . . . 106
Author Contributions . . . 107
Acknowledgments . . . 107
3.5. Addendum: On the study's limitations . . . 108
Chapitre 4. Repliement des protéines: une introduction . . . 109
4.1. Structures des protéines . . . 109
4.2. Repliement des protéines . . . 112
4.3. Similitude structurelle . . . 114
4.3.1. RMSD . . . 114
4.3.2. BC-score . . . 115
4.3.3. CAD-score . . . 116
4.4. Prédiction structurelle . . . 118
Chapitre 5. Méthode pour l'amélioration de PEP-FOLD . . . 121
5.1. Quelques notions théoriques . . . 121
5.1.1. Potentiel gros-grain . . . 121
5.1.2. Fonction de distribution radiale . . . 122
5.1.3. Potentiel de champ moyen . . . 123
5.2. Historique des potentiels OPEP . . . 123
5.2.1. OPEPv1 . . . 124
5.2.2. OPEPv3 . . . 125
5.2.3. OPEPv4 . . . 128
5.2.4. OPEPv5 . . . 129
5.2.5. OPEPv6 . . . 131
5.3. PEP-FOLD . . . 131
5.3.1. Acides aminés à alphabet structurel . . . 132
5.3.2. Reconstruction tridimensionnelle . . . 134
5.3.3. sOPEP . . . 135
5.4. Objectifs des travaux . . . 137

Chapitre 6. A generalized attraction-repulsion potential and revisited fragment library improves PEP-FOLD peptide structure prediction 139

6.1. Introduction 141

6.2. Materials and Methods 144
6.2.1. PEP-FOLD 144
6.2.2. Library of fragments 145
6.2.3. sOPEP 146
6.2.3.1. Bonded Potential 146
6.2.3.2. Non-Bonded Potential 147
6.2.3.3. Explicit Hydrogen Bond 147
6.2.4. Optimization Protocol 149
6.2.4.1. Decoys Classification 149
6.2.4.2. Selection of protein targets 150
6.2.4.3. Decoys generation 152
6.2.4.4. Parameters optimization 152
6.2.4.5. Iterated optimization procedure 154

6.3. Comparison with state-of-the-art techniques 155

6.4. Results 155
6.4.1. Updated library of fragments 155
6.4.2. Optimization of the sOPEPv2 parameters 156
6.4.3. Impact on structure Prediction 158
6.4.3.1. Improvements 162
6.4.4. Comparison 167

6.5. Discussion 169
6.5.1. Dependence on target size and secondary structure 170
6.5.2. PEP-FOLD's limitations 171

6.6. Conclusion 173

6.7. Acknowledgement 174

6.8. Supplementary material 174

6.9. Addendum 174

Chapitre 7. Préambule au développement de aaOPEP 177

7.1. Maladie d'Alzheimer 178
7.1.1. Amyloïde-β 178

7.1.2. Protéine intrinsèquement désordonnée ... 178
7.2. Échange de répliques Hamiltonien ... 179
Chapitre 8. Développements préliminaires de aaOPEP ... 181
8.1. Conception du potentiel ... 181
8.2. Paramétrisation initiale ... 182
8.3. Simulation d'Amyloïde-β ... 183
8.3.1. Protocole des simulations ... 183
8.3.2. Monomère ... 185
8.3.3. Dimère ... 186
8.3.4. Clusterisation ... 186
8.4. Outil: simulateur OPEP ... 188
8.5. Perspective ... 190
Conclusion ... 193
8.6. Développement thérapeutique contre la COVID-19 ... 193
8.7. Prédictions structurelles par PEP-FOLD ... 194
8.8. Méthode simplifiée pour l'étude de l'Alzheimer ... 195
Références bibliographiques ... 197
Annexe A. Supporting Figures: A generalized attraction-repulsion potential and revisited fragment library improves PEP-FOLD peptide structure prediction ... 211
A.1. Tested Targets ... 211
Annexe B. Supporting Figures: Corilagin and 1,3,6-Tri-O-galloy-β-D-glucose: potential inhibitors of SARS-CoV-2 variants ... 237

1.1 **Variants Préoccupants: Informations générales.** Le tableau présente, de gauche à droite, la nouvelle et l'ancienne appellation, le pays de première détection et certaines mutations cruciales au niveau du RBD pour chacun des variants préoccupants. 48

1.2 **Variants Préoccupants: Impact sur l'épidémiologie.** Chacune des colonnes présente les résultats pour un variant préoccupant en termes de transmission (ligne 1), de sévérité des symptômes (ligne 2), de risque de réinfection (ligne 3) et d'impact sur la vaccination (ligne 4). Le tout est tiré du site de l'OMS [**13**]. 48

2.1 **Poids d'AutoDock VINA.** La colonne de gauche présente le terme du potentiel et la colonne de droite présente le poids associé à celui-ci. 65

3.1 **ACE2-RBD contacts blocked by the ligands.** RBD(RBM) residues blocked by corilagin (Cor.) and TGG. The RBM residues showed are those that are specifically involved in a contact pair with ACE2 that is formed with a probability of at least a 60% during the ACE2-RBD complex MD simulation. Nonpolar, polar, positively charged and negatively charged residues are shown respectively in gray, green, red and blue. The formation of a contact with the ligand is shown in gray. The presence of a H-bond or a salt-bridge is indicated by *HB* and *SB*, respectively. The star (*) indicates that a H-bond is present in the experimental structure. The dagger (†) beside SB for Lys458 indicates that this residue was added to the table even if its contact probability with ACE2 is less than 60% (45%) because it forms a salt-bridge with E23 of ACE2. 94

3.2 **The binding affinity between corilagin/TGG and ACE2/RBD.** The second column shows the VINA binding affinity for the best pose found during docking. The third column shows the average VINA binding affinity computed over the interval of convergence (90-100 ns for RBD/E484K-N501Y with Corilagin and 75-100 ns for the rest) of the ligand-protein MD simulations. The fourth column shows the MMPBSA binding free energy computed over the same interval using the g_mmpbsa tools [**57**]. The average and standard deviation are computed using a 500-steps of bootstrap analysis and 40 ps snapshots on the interval of convergence (90-100 ns for RBD/E484K-N501Y with corilagin and 75-100 ns for the rest). For RBD/E484K-N501Y with corilagin, we compare the results for the two binding sites. Site 1 is the corilagin's localization before its disassociation (from 20-70 ns). Site 2 is the corilagin's localization after its reassociation (from 90-100 ns) 101

6.1 **Energy ranking for targets in the *NG* ensemble.** LowE and Native Energy: energy of the lowest energy model generated using sOPEP2/Lib2, and

experimental structure using sOPEPv2, respectively. Native Rank: ranking of the experimental structure compared to models generated using sOPEP2/Lib2. PEP-FOLD's predictions are ordered from 1 to 500 in order of increasing energy; rank 0 means that the experimental structure has a lower energy than all predictions while a rank of 501 means the experimental structure has a higher energy than all predictions. 163

6.2 **Proteins for which the classification of the lowest energy prediction (TOP1) is improved.** The notations are identical to that of table 6.1. Columns two and three present the results for the lowest energy prediction and the best prediction in the TOP5 for sOPEPv1/Lib1, while columns four and five present the same results for sOPEPv2/Lib2. Each column present the quality assessment in terms of CAD-CG and, in parenthesis, BC-WDC. Color coding: CAD-CG scores corresponding to the native, near-native and non-native classification are shown respectively in green, yellow and red. 164

6.3 **Targets for which the classification of the lowest energy prediction (TOP1) is deteriorated.** The notations are identical to that of table 6.2. 165

6.4 **Ranking of incorrectly predicted targets.** For each target incorrectly predicted within the five lowest energy, the ranking of the experimental structure and the quality assessment in terms of CAD-CG(BC-WDC) and ranking of the first non Non-native prediction are presented respectively for column 2 to 4. PEP-FOLD's predictions are ordered from 1 to 500 in order of increasing energy; rank 0 means that the experimental structure has a lower energy than all predictions while a rank of 501 means the experimental structure has a higher energy than all predictions. 172

6.5 **Statistical significance of PEP-FOLD's improvements.** The statistical significance was computed on the CAD-CG score of the TOP1, TOP5 and best in TOP5 predictions (row 1 to 3), as presented in Figure 6.5. Results that are significantly better for PEP-FOLD with Lib2 and sOPEPv2 are presented in green. 175

6.6 **Statistical significance of the comparison between PEP-FOLD and other popular prediction techniques.** The statistical significance was computed on the CAD-CG score of the TOP1 predictions, as presented in Figure 6.7. The top and bottom parts of the table present the results in term of secondary structure and sequence length respectively. Results that are significantly better for PEP-FOLD are presented in green. 175

A.1 **sOPEP2 non bonded potential parameters.** 215

A.2 **Comparison of PEP-FOLD's predictions for the parametrization (G/IC) ensemble.** The parametrization (G/IC) ensemble contains 25 protein targets. The table is split into three parts. From top to bottom, we show the results for the lowest energy prediction, the TOP5 lowest energy predictions and the best prediction inside the TOP5. The first column describes the library/potential pairing, the next three columns are the number of decoys generated in the Native, Near-Native and Non-Native class respectively and the last column shows the average CAD-CG score and in parenthesis, the BC score computed on the WDC defined on the *Protein Data Bank*. 216

A.3 **Comparison of PEP-FOLD's predictions for the validation G/IC ensemble.** The validation G/IC ensemble contains 40 protein targets. The table is split into three parts. From top to bottom, we show the results for the lowest energy prediction, the TOP5 lowest energy predictions and the best prediction inside the TOP5. The first column describe the library/potential pairing, the next three columns are the number of decoys generated in the Native, Near-Native and Non-Native class respectively and the last column shows the average CAD-CG score and in parenthesis, the BC score computed on the WDC defined on the *Protein Data Bank*. 217

A.4 **Comparison of PEP-FOLD's predictions for the validation G/CC ensemble.** The validation G/CC ensemble contains 50 protein targets. The table is split into three parts. From top to bottom, we show the results for the lowest energy prediction, the TOP5 lowest energy predictions and the best prediction inside the TOP5. The first column describe the library/potential pairing, the next three columns are the number of decoys generated in the Native, Near-Native and Non-Native class respectively and the last column shows the average CAD-CG score and in parenthesis, the BC score computed on the WDC defined on the *Protein Data Bank*. 218

A.5 **Comparison of PEP-FOLD's lowest energy predictions by peptide's length.** The table is split into two parts. From top to bottom, we show the results for peptides of length 0 to 50 amino acids and of length 51 to 70 amino acids respectively. The first column describe the library/potential pairing, the next three columns are the number of decoys generated in the Native, Near-Native and Non-Native class respectively and the last column shows the average CAD-CG score and in parenthesis, the BC score computed on the WDC defined on the *Protein Data Bank*. 219

A.6 **Comparison of PEP-FOLD's lowest energy predictions by peptide's structural class.** The table is split into three parts. From top to bottom,

we show the results for α-targets, β-targets and α/β-targets respectively. The first column describe the library/potential pairing, the next three columns are the number of decoys generated in the Native, Near-Native and Non-Native class respectively and the last column shows the average CAD-CG score and in parenthesis, the BC score computed on the WDC defined on the *Protein Data Bank*. 220

A.7 **Comparison for targets in the parametrization G/IC set - TOP1.** Each column show the CAD-CG (and BC-WDC in parenthesis). For each target, we recall its secondary structure class and size in amino-acids. The next two columns show the results of the lowest energy prediction (TOP1) PEP-FOLDv1 and PEP-FOLDv2 respectively. The next three column show the results using APPTest, RaptorX and AlphaFold2 respectively. Color coding: CAD-CG scores corresponding to the native, near-native and non-native classification are shown respectively in green, yellow and red. 221

A.8 **Comparison for targets in the validation G/IC set - TOP1.** Notations and color coding are similar to that of Table A.7 222

A.9 **Comparison for targets in the validation G/CC set - TOP1.** Notations and color coding are similar to that of Table A.7 223

A.10 **Comparison for targets in the validation NG set - TOP1.** Notations and color coding are similar to that of Table A.7 224

A.11 **Comparison for targets in the parametrization G/IC set - Best in TOP5 (five lowest energy models).** Notations and color coding are similar to that of Table A.7 225

A.12 **Comparison for targets in the validation G/IC set - Best in TOP5.** Notations and color coding are similar to that of Table A.7 226

A.13 **Comparison for targets in the validation G/CC set - Best ni TOP5.** Notations and color coding are similar to that of Table A.7 227

A.14 **Comparison for targets in the validation NG set - Best in TOP5.** Notations and color coding are similar to that of Table A.7 228

B.1 **Experimental and numerical comparison of H-bonds between ACE2-RBD.** The first and second columns indicate the residues (and atoms) of respectively RBD and ACE2, involved in the formation of a H-bond, while the third column indicates the donor/acceptor length. The first three columns were computed on the minimized crystal structure [**21**]. Bold donor/acceptor pairs were also identified in the experimental paper [**21**]. The last two columns show the

occurrence and the donor/acceptor length computed on the convergence interval (250-500 ns) of our MD of the ACE2/RBD complex. 239

Liste des tableaux

0.1 **RAPIDITÉ -vs- COMPLEXITÉ.** Ce schéma représente le positionnement de certaines méthodes en termes de rapidité (axe horizontal) et de complexité (axe vertical). La zone (A), en bleu, correspond aux méthodes complexes et lentes, la zone (B), en vert, correspond aux méthodes intermédiaires et la zone (C), en orange, correspond aux méthodes rapides, mais simplifiées. La zone rouge représente la zone interdite: des méthodes simplifiées et lentes ou des méthodes complexes et rapides. 40

1.1 **Mécanisme d'action du virus SARS-CoV-2.** La figure est tirée de l'article de Philip V'kovski *et coll.* [**3**]. La zone (A) inclut les étapes de reconnaissance de la cellule hôte. La zone (B) inclut les étapes de réplication virale. Finalement, la zone (C) inclut les étapes, d'expulsion de nouveaux virus. 45

2.1 **Interactions liées et le potentiel associé.** De haut en bas, on retrouve les liens atomiques, les angles de valence et les angles dièdres. À gauche, on retrouve une représentation visuelle de l'interaction et à droite le potentiel associé à cette interaction. Tous les paramètres sont tirés de AMBER99SB*-ILDN [**29**]. 53

2.2 **Interactions non-liées et le potentiel associé.** De haut en bas, on retrouve les interactions répulsives/attractives et les interactions électrostatiques. À gauche, on retrouve une représentation visuelle de l'interaction et à droite le potentiel associé à cette interaction. Tous les paramètres sont tirés de AMBER99SB*-ILDN [**29**]. 54

2.3 **Géométrie du modèle d'eau TIP3P.** L'atome d'oxygène et les deux atomes d'hydrogène sont présentés respectivement en rouge et en gris. δ indique la charge partielle de chacun des atomes. 57

2.4 **Protocole d'ELISA utilisé dans l'étude COVID.** Les molécules de RBD, les bloqueurs, ACE2, la biotine, la streptavidine, la peroxydase de raifort et les substrats chromogènes sont présentés respectivement en turquoise, noir, vert, rouge, mauve, bleu et gris. L'absorbance est calculée à 450 nm grâce au ratio de l'intensité lumineuse initiale (I_0) et l'intensité lumineuse finale (I). 71

2.5 **Schéma d'un montage SPR typique.** La surface métallique, cruciale pour la résonance des plasmons, est présentée en couleur or et les récepteurs liés à celle-ci sont présentés en vert. Le ligand est présenté en orange. 73

2.6 **Profil classique de réponse de biosenseur SPR pour l'étude des interactions récepteurs/ligands.** Réponse SPR (en unité de réponse) en fonction du temps. Les étapes d'association, d'équilibre et de dissociation sont présentées respectivement en bleu, orange et vert. Les paramètres utilisés sont $k_a = 1 \times 10^5$ $\mathrm{M^{-1}s^{-1}}$, $k_d = 1.8 \times 10^{-4}$ $\mathrm{s^{-1}}$ ($K_D = 1.8 \times 10^{-9}$ M) et $[L] = 80$ nM.

Ces paramètres furent choisis afin de produire une courbe comparable à celle entre corilagin/RBD présentée à la Figure 3.5. 75

3.1 **The docked position of corilagin on RBD and the three mutants**. The RBM segment and the rest of the RBD are shown respectively in red and teal. Residues 484 and 501, both the location of tested mutation, are shown in pink and purple respectively. The ligand in black and gold is respectively the conformation after docking and the center of the biggest cluster sampled during the converged part of the MD simulation respectively. 90

3.2 **The docked position of TGG on RBD and the three mutants**. The RBM segment and the rest of the RBD are shown respectively in red and teal. Residues 484 and 501, both the location of tested mutation, are shown in pink and purple respectively. The ligand in black and gold is respectively the conformation after docking and the center of the biggest cluster sampled during the converged part of the MD simulation respectively. 91

3.3 **Corilagin interaction maps.** The interaction maps of corilagin with wildtype RBD (top left), RBD/E484K (top right), RBD/N501Y (bottom left) and RBD/E484K-N501Y (bottom right) are shown for the center of the biggest cluster computed on the convergence interval using the protein backbone atoms and ligand non-hydrogen atoms. The nonpolar contacts, defined by a distance smaller than 0.40 nm, between the ligand and the protein are shown as red arcs. H-bonds and their donor/acceptor distance are shown in green. All figures were generated using LigPlot [**103**, **104**]. 92

3.4 **TGG interaction maps.** The interaction maps of TGG with wildtype RBD (top left), RBD/E484K (top right), RBD/N501Y (bottom left) and RBD/E484K-N501Y (bottom right) for the center of the biggest cluster computed on the convergence interval using the protein backbone atoms and ligand non-hydrogen atoms. The nonpolar contacts, defined by a distance smaller than 0.40 nm, between the ligand and the protein are shown as red arcs. H-bonds and their donor/acceptor distance are shown in green. All figures were generated using LigPlot [**103**, **104**]. 93

3.5 **Characterization of molecular interactions by surface plasmon resonance. A, B**) Binding kinetics of corilagin and TGG on immobilized (**A**) RBD and (**B**) ACE2. The recombinant proteins RBD and ACE2 are respectively immobilized on a CM5 sensor chip and increasing concentrations of polyphenols are injected to evaluate binding kinetics. **C**) Kinetics of ACE2 binding to immobilized RBD (left panel) and kinetics of RBD binding to immobilized ACE2 (right panel). **D**) Pre-incubation of RBD (50 nM) for 30 minutes with increasing

concentrations of corilagin or TGG inhibit the binding of RBD to immobilized ACE2. 95

3.6 **Inhibitory effects of TGG, corilagin and their mixture on the interaction between SARS CoV-2 Spike protein and human ACE2.** TGG (**A**) and corilagin (**B**) are tested at different concentrations (0.1, 1, 5 and 10 μM) and their mixture (**C**) (0,1, 1, 5 μM) to evaluate their ability to inhibit the binding of immobilized Spike protein (0.5 μg/ml) to human biotin labeled ACE2 (0.5 μg/ml), by using the ELISA assay. The absorbance values at 450 nm of human ACE2 (0.5 μg/ml) are set to 100%. Results are expressed as mean ± standard error of the mean (SEM) of two (combined effect) or three independent assays. Statistical analysis is performed using the One-way ANOVA followed by the Dunnett's post hoc test with $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$ compared to human ACE2 (0.5 μg/ml). 96

3.7 **Inhibitory effects of TGG, corilagin and their mixture on the interaction between human ACE2 and ACE2 antibody (18-740 AA).** TGG (**A**) and corilagin (**B**) are tested at different concentrations (0.1, 1, 5 and 10 μM) and their mixture (C) (0,1, 1, 5 μM) to study their ability to inhibit the binding of immobilized ACE2 antibody (0.5 μg/ml) to human biotin labeled ACE2 (0.5 μg/ml), by using the ELISA assay. The absorbance values at 450 nm of human ACE2 (0.5 μg/ml) were set to 100%. Results are expressed as mean ± standard error of the mean (SEM) of two (combined effect) or three independent assays. Statistical analysis was performed using the One-way ANOVA followed by the Dunnett's post hoc test with $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$ compared to human ACE2 (0.5 μg/ml). 96

3.8 **Mutations effect on the solvent accessibility of RBD alone.** RBD's per residue solvent accessible surface area (SASA) difference between the MD and the experimental structure. Only the residues of RBD interacting with ACE2 (contact probability greater than 60% during the MD simulation) in the complex structure are shown. The red residue number indicates the position of a mutation. The SASA of wildtype (blue), E484K (teal), N501Y (yellow) and E484K/N501Y (red) are compared. The error bars correspond to the standard deviation over the 250-500 ns interval. 98

4.1 **Structure de la chaîne principale et lien peptidique.** L'azote, les carbones (C, C_α), l'oxygène et l'hydrogène sont présentés respectivement en bleu, vert, rouge et gris. De haut en bas, on voit la formation du lien peptidique via la libération d'une molécule d'eau. 110

4.2 **Diagramme de Ramachandran.** La figure est adaptée du livre de Tamar Schlick [**1**]. Les zones appelées α et β correspondent respectivement aux paires d'angles dièdres ϕ/ψ associées aux structures secondaires d'hélice-α et de feuillet β respectivement. 111

4.3 **Structure des chaînes latérales.** La chaîne latérale (et le carbone-α) de chacun des acides aminés. Les atomes de carbones, d'oxygène, d'azote et d'hydrogène sont présentés respectivement en vert, rouge, bleu et blanc. Les images furent réalisées avec PYMOL [**95**]. 112

4.4 **Schéma des principales structures secondaires.** De haut en bas, on retrouve les hélices α, les feuillets β antiparallèles et les feuillets β parallèles. À gauche, on retrouve un schéma de la structure, tiré de la protéine 1PGB, et réalisé avec PYMOL [**95**]. À droite, on retrouve les ponts-H, représentés par des lignes pointillées, caractéristiques de chacune de ces structures. Seuls les atomes de la chaîne principale sont présentés. 113

4.5 **Paysage d'énergie libre en forme d'entonnoir.** L'image est adaptée de l'article de Ken A Dill *et coll.* [**131**]. Paysage énergétique caractéristique du repliement des protéines. Les niveaux d'entropie et d'énergie sont présentés respectivement horizontalement et verticalement. 114

4.6 **Exemple d'un tétraèdre à la base du calcul du BCscore.** Le tétraèdre est formé par trois carbones-α et le centre géométrique de tous les carbones-α. Cette figure présente, pour la protéine 1PGB, un tétraèdre formé par le centre géométrique des carbones α (C) et les carbones α des résidus LEU5, THR18 et TYR32. 116

5.1 **Représentation gros-grain des potentiels OPEP.** La chaîne principale est représentée en tout-atome. Les atomes d'azote, d'oxygène, d'hydrogène et de carbone sont présentés respectivement en bleu, en rouge, en gris et en vert. La chaîne latérale est représentée par un seul centre d'interaction, présenté ici en orange et en magenta. 125

5.2 **Représentation des quatre distances définissant une lettre de l'alphabet structurel.** Pour la lettre A de l'alphabet structurel, les quatre distances d_1, d_2, d_3 et d_4 sont présentées. Le plan dans lequel se situe les trois premiers carbones-α est représenté par le parallélogramme vert. 132

5.3 **Représentation des lettres de l'alphabet structurel de PEP-FOLD.** Un fragment représentatif (fragment 0) est présenté pour chacune des lettres de l'alphabet structurel (a,A-Z) utilisé par PEP-FOLD. 133

5.4 **PEP-FOLD.** Présentation du protocole PEP-FOLD. À partir de la séquence d'acides aminés (1), un SVM est utilisé afin de générer les probabilités que chacune des positions soit décrite par chacune des lettres de l'alphabet structure (2). Par la suite, un algorithme de Foward-Backtrack est appliqué afin de générer un nombre désiré de séquences dans l'espace des lettres de l'alphabet structurel (3). Par la suite, ces séquences sont assemblées (4) via la superposition des fragments et un algorithme greedy (5). Une fois la reconstruction terminée, on obtient les prédictions tridimensionnelles (6). L'exemple présenté ici est pour la protéine 1b03. 136

6.1 **Coarse-grained representation in sOPEP.** The backbone is represented in a all-atoms format with all backbone atoms but HA — N, H_N, C_α, C and O — present while side-chains are represented by a single interaction center. 146

6.2 **Target classification based on PEP-FOLD predictions.** Each protein target is classified in one of three ensemble: *Generated/Correctly Classified* (G/CC), *Generated/Incorrectly Classified* (G/IC) and *Not Generated* (NG). Targets for which the best predicted structure is in the non-native class are placed in ensemble NG (no predicted structure in the Native or Near-Native class). Targets for which the lowest energy structure is in a worst class than the best predicted structure are placed in ensemble (G/IC). Finally, targets for which the lowest energy structure is in the same class as the best predicted structure are placed in ensemble G/CC. 152

6.3 **Overview of the optimization protocol.** In the initial step of the optimization (top green frame), the parametrization ensemble composed of 500 PEP-FOLD predictions for each targets. The sOPEP parameters are then optimized utilizing an iterative procedure (bottom blue frame), in which the parametrization ensemble is improved by adding newly generated PEP-FOLD predictions. 154

6.4 **Improvements in the number of unsolved inequalities during the optimization protocol.** For each optimization step (x-axis), the improvements are presented as the additional fraction in the number of unsolved inequalities over the previous optimization step. 157

6.5 **Classification of PEP-FOLD's predictions and average CAD-CG of PEP-FOLD's predictions.** x-axis: name of the proteins' sets. *Param. G/IC*, *Vali. G/IC* and *Vali. G/CC* refer to the parametrization, the validation G/IC and the validation G/CC set containing 25, 39 and 48 proteins, respectively. Bar width are proportional to the number of proteins of the set. **Left** side: classification of PEP-FOLD's predictions. The native, near-native and non-native classification are shown in green, yellow and red, respectively. For each set, the four columns represent from left to right, the original library/original potential, the original

library/re-optimized potential, the new library/original potential and the new library/re-optimized potential respectively. Panels (A), (B) and (C): fraction of proteins per class considering the lowest energy model only (TOP1), the five lowest energy models (TOP5), the best CAD-CG in the TOP5, respectively. **Right** side: average CAD-CG of 3D predictions. For each set, the four columns represent from left to right, the original library/original potential (light orange), the original library/re-optimized potential (light blue), the new library/original potential (orange) and new library/re-optimized potential (blue) respectively. The CAD-CG associated with the near-native and native classification are shown respectively in yellow and green (y-axis). 159

6.6 **Lowest energy predictions for proteins going from non-native to native predictions.** The left and right columns show the results for the original library with the original potential, in orange, and the new library with the new potential, in blue, respectively. The experimental structure is shown in gray. The structures are aligned on C_α of residues of the well-defined core as presented on the *Protein Data Bank*. Pictures were generated using Pymol [**95**] and secondary structure elements were determined using STRIDE [**179**]. 166

6.7 **Average CAD-CG score for the TOP1 prediction using five prediction approaches.** Panel (A): results when targets are classified by structural class, with respectively 60, 32 and 21 proteins in the α, β and α/β categories. Panel (B): results when targets are classified by length, with respectively 17, 48 and 50 targets with less than 26 amino acids, between 26 and 50 amino acids and between 51 and 70 amino acids. Of note: PEP-FOLD is usually limited to up to only 50 amino-acids. The RaptorX-server minimum accepted length is 26 amino acids. The APPTest does not consider sequences with more than 40 amino acids. The CAD-CG associated with the near-native and native classification are shown respectively in yellow and green (y-axis). Protein targets from the *NG* ensemble are excluded for this figure. 169

8.1 **Exemples de potentiel de MIE de aaOPEP.** Les fonctions de distributions radiales obtenues via 300 ns de dynamique moléculaire sont présentées à gauche (en bleu). Les lignes verticales noire et rouge correspondent respectivement à la valeur du zéro et à la valeur du minimum du potentiel pour cette paire d'atomes dans AMBERff99SB*-ILDN. Le PMF, obtenu de la courbe de gauche, et le potentiel de MIE optimisé sont présentés à droite respectivement en bleu et en orange. ... 184

8.2 **Convergence du monomère d'Aβ_{40}.** Le panneau (A) présente le configuration initiale. Le panneau (B) présente l'index de la réplique présente à la réplique sans échelonnage de l'énergie potentielle (réplique 0) en fonction du temps. Les

panneaux (C) et (D) présentent respectivement le rayon de giration en fonction du temps et la distribution des rayons de giration dans les deux dernières tranches de 100 ns de la simulation (600-700 ns en bleu et 700-800 ns en orange). Finalement, les panneaux (E) et (F) présentent respectivement la surface accessible au solvant en fonction du temps et la distribution des surfaces accessibles au solvant pour les deux dernières tranches de 100 ns de la simulation (600-700 ns en bleu et 700-800 ns en orange). 187

8.3 **Convergence du dimère d'Aβ_{40}.** Le panneau (A) présente la configuration initiale. Le panneau (B) présente l'index de la réplique sans échelonnage de l'énergie potentielle (réplique 0) en fonction du temps. Le panneau (C) présente le rayon de giration en fonction du temps; la chaîne A, la chaîne B et les deux chaînes sont présentées respectivement en orange, vert et bleu, Le panneau (D) présente la distribution des rayons de giration des deux chaînes dans les deux dernières tranches de 100 ns de la simulation (350-450 ns en bleu et 450-550 ns en orange). Le panneau (E) présente respectivement la surface accessible au solvant en fonction du temps; la chaîne A, la chaîne B et les deux chaînes sont présentées respectivement en orange, vert et bleu. Finalement, le panneau (F) présente la distribution des surfaces accessibles au solvant des deux chaînes pour les deux dernières tranches de 100 ns de la simulation (350-450 ns en bleu et 450-550 ns en orange). 188

8.4 **Clusterisation du monomère et dimère d'Aβ.** Les résultats au niveau du monomère et du dimère sont présentés respectivement à gauche et à droite. La population cumulative en fonction des clusters est présentée dans le haut des colonnes. En dessous, on retrouve les centres des clusters que nous avons classifiés comme natif. La clusterisation a été réalisée à l'aide de l'algorithme de l'article de Daura *et coll.* [**94**]. 189

A.1 **Parametrization set all-atom contact map.** A contact between two side-chains is considered if the distance between at least one heavy atom pairs follows $r_{ij} < R_i + R_j + d$, where R_x is the van der Waals radius of atom type X and d is a cutoff here fixed at 1.5Å. Neighbouring residues is the amino acid sequence were not considered. 229

A.2 **PEP-FOLD' models average BC-WDC.** The x-axis is the name of our proteins' sets. In this case, *Param.*, *To Improve* and *Good* refers to the parametrization, the validation "To Improve" and the validation "Good" set containing respectively 25, 40 and 50 proteins. The bar width is proportional to the number of proteins in each set. For each set, the four columns represent from left to right, the original library/original potential (light orange), the original

library/re-optimized potential (light blue), the new library/original potential (orange) and new library/re-optimized potential (blue) respectively. Panel (A) shows the average BC-WDC [**137**] for the lowest energy (TOP1) predictions. Panel (B) shows the the average BC-WDC for the five lowest energy (TOP5) predictions. Panel (C) shows the average BC-WDC for the prediction with the best CAD-CG in the TOP5. The WDC is taken from the *Protein Data Bank* validation report. 230

A.3 **PEP-FOLD' models average scores by target size** The x-axis is the name of our proteins' sets. In this case, *0-50aa* and *50-70aa* refers to the proteins of length between 0 and 50 amino acids (inclusively) and to the proteins of length between 50 (exclusively) and 70 amino acids set containing respectively 65 and 50 proteins. The bar width is proportional to the number of proteins in each set. For each set, the four columns represent from left to right, the original library/original potential (light orange), the original library/re-optimized potential (light blue), the new library/original potential (orange) and new library/re-optimized potential (blue) respectively. Panel (A) shows the average CAD-CG for the lowest energy (TOP1) predictions. Panel (B) shows the the average BC-WDC for the lowest energy (TOP1) predictions. The WDC is taken from the *Protein Data Bank* validation report. 231

A.4 **Updated attractive/repulsive potential of sOPEPv1 compared to sOPEPv2.** The updated sOPEPv2 potential (in blue) compared to the sOPEPv1 potential (in orange). For these interactions, sOPEPv2 is more permissive at short distances than sOPEPv1. 232

A.5 **Updated attractive/repulsive potential of sOPEPv1 compared to sOPEPv2.** The updated sOPEPv2 potential (in blue) compared to the sOPEPv1 potential (in orange). For these interactions, sOPEPv2 is less permissive at short distances than sOPEPv1. 233

A.6 **Updated repulsive potential of sOPEPv1 compared to sOPEPv2.** For these interactions, sOPEPv2 is less permissive at short distances than sOPEPv1. 234

A.7 **PEP-FOLD' models average scores by secondary structure class** The x-axis is the name of our proteins' sets with α, β, α/β referring to the proteins of α, β and α/β secondary structure set containing respectively 52, 31 and 21 proteins. The bar width is proportional to the number of proteins in each set. For each set, the four columns represent from left to right, the original library/original potential (light orange), the original library/re-optimized potential (light blue), the new library/original potential (orange) and new library/re-optimized potential (blue) respectively. Panel (A) shows the average CAD-CG for the lowest energy

(TOP1) predictions. Panel (B) shows the the average BC-WDC for the lowest energy (TOP1) predictions. The WDC is taken from the *Protein Data Bank* validation report. 235

A.8 **Native secondary structure reproduction by secondary structure class.** The x-axis is the name of our proteins' sets with α, β, α/β referring to the proteins of α, β and α/β secondary structure set containing respectively 60, 32 and 21 proteins. The bar width is proportional to the number of proteins in each set. For each set, the four columns represent from left to right, the original library/original potential (light orange), the original library/re-optimized potential (light blue), the new library/original potential (orange) and new library/re-optimized potential (blue) respectively. Bars above and below the zero line shows α-helix and β-sheet native secondary structure reproduction respectively. Panel (A) shows the average native secondary structure reproduction for the lowest energy (TOP1) predictions. Panel (B) shows the the average native secondary structure reproduction for the five lowest energy (TOP5) predictions. Secondary structure assignments were done using STRIDE [**179**]. 236

B.1 **Convergence of the ACE2-RBD complex MD simulation. (A)** The simulated system with ACE2 in magenta (A1A2, residues 19-83), orange (HS, residues 322-362) and green (else) as well as RBD in red (RBM, residues 438-506) and teal (else). **(B)** Root mean square deviation (RMSD) on the backbone atoms (N, Cα, C and O) from the experimental crystal structure of the complex (PDB:6M0J) as a function of time for the regions at the interface (A1A2 and HS for ACE2 and RBM for RBD). The average and the standard deviation of the RMSD on the converged interval (250-500 ns) are shown in the inset. **(C)** The DSSP secondary structure as a function of time for ACE2 and RBD. The average and the standard deviation of the secondary structure on the converged interval are compared to the values computed on the experimental structure (EXP) on the right side. Only the α-helix, β-sheet, turn and coil content are shown. The difference with 100% is associated rest of the DSSP secondary structure classes (β-bridge, bend, 310-helix and π-helix). **(B-C)** The figures depict the running average using a 5-ns time window. The $\pm 1\sigma$ interval is shown by the semitransparent region around the curve. 238

B.2 **Contacts between ACE2 and RBD during the MD simulation. (A)** Probability contact maps between A1A2 (residues 19-83 of ACE2) and RBM (residues 438-506 of RBD) as well as between HS (residues 322-362 of ACE2) and RBM (residues 438-506 of RBD). A contact is considered between two residues if the distance between any pair of atoms is smaller than 0.40 nm. The presence of

a contact in the experimental crystal complex (PDB:6M0J) is shown by a red dot. **(B)** H-bonds probability of ACE2 residues with RBD (top) and of RBD residues with ACE2 (bottom). A H-bond is considered when the donor-acceptor distance is less than 0.35 nm and when the hydrogen-donor-acceptor angle is less than 35 degrees. The involvement of each residue in a H-bond in the experimental crystal structure is indicated by the white star. **(C)** Salt-bridges probability of ACE2 with RBD. A salt-bridge is considered when the distance between two oppositely charged groups is less than 0.40 nm. The presence of the salt-bridges in the crystal structure is shown by the white star. **(A-B-C)** All probabilities are computed on the converged interval (250-500 ns). 240

B.3 **Convergence of the ACE2 MD simulation. (A)** The simulated system with ACE2 in magenta (A1A2, residues 19-83), orange (HS, residues 322-362) and green (else). **(B)** Root mean square deviation (RMSD) on the backbone atoms (N, Cα, C and O) from the structure of ACE2 in the experimental crystal of the complex (PDB:6M0J) as a function of time for the whole protein (ACE2) and the regions at the interface (A1A2 and HS). The average and the standard deviation of the RMSD on the converged interval (250-500 ns) are shown in the inset. **(C)** The DSSP secondary structure as a function of time for ACE2 (left) and A1A2+HS (right). **(B-C)** The figures depict the running average using a 5-ns time window. The $\pm 1\sigma$ interval is shown by the semitransparent region around the curve. **(D)** The DSSP per residue secondary structure for A1A2 (left) and HS (right) on the converged interval (250-500 ns). The experimental secondary structure for each residue is illustrated by the square below the 0 mark. White means other secondary structure. 241

B.4 **Convergence of the RBD MD simulation. (A)** The simulated system with RBD in red (RBM, residues 438-506) and teal (else). **(B)** Root mean square deviation (RMSD) on the backbone atoms (N, Cα, C and O) from the structure of RBD in the experimental crystal of the complex (PDB:6M0J) as a function of time for the whole protein (RBD) and the region at the interface (RBM). The average and the standard deviation of the RMSD on the converged interval (250-500 ns) are shown in the inset. **(C)** The DSSP secondary structure as a function of time for RBD (left) and RBM (right). **(B-C)** The figures depict the running average using a 5-ns time window. The $\pm 1\sigma$ interval is shown by the semitransparent region around the curve. **(D)** The DSSP per residue secondary structure for RBM on the converged interval (250-500 ns). The experimental secondary structure for each residue is illustrated by the square below the 0 mark. White means other secondary structure. 242

B.5 **The initial structure of each ligand.** (TOP) Corilagin (BOTTOM) TGG. The figures were generated using PyMOL [**95**]. 243

B.6 **Configuration ensembles and box used for the molecular docking. (Top)** Box used for the docking on ACE2. The box is centered around the point (6.5427, 8.3338, 6.5367) nm with a size of 2.2640 nm, 5.2072 nm, 1.4633 nm respectively in x, y and z. The A1A2 region (residues 19-83) and the HS region (residues 322-362) are shown respectively in magenta and orange. **(Bottom)** Box used for docking on the RBD. The box is centered around the point (6.13835, 6.8906, 1.1736) nm with a size of 2.7625 nm, 4.3358 nm, 2.6026 nm respectively in x, y and z. The RBM segment (residues 438-506) is shown in red. 244

B.7 **A) Human ACE2 protein binding to immobilized SARS-CoV-2 RBD Spike protein** (0.5 μg/ml) using an increasing dose of human ACE2 protein (0,015 to 2 μg/ml). **B) Human ACE2 protein binding to immobilized ACE2 antibody** (0.5 μg/ml) using an increasing dose of human ACE2 protein (0,015 to 2 μg/ml). Results are expressed as mean ± standard error. 245

B.8 **Corilagin's ability of the generated docking conformations to block the ACE2-RBD complex formation.** The docking conformations generated by AutoDock VINA [**55**] as a function of its binding energy (x-axis) and fraction of ACE2/RBD contacts such conformation is able to black (y-axis). The occurrence of such conformation is shown as the z-axis. 245

B.9 **TGG's ability of the generated docking conformations to block the ACE2-RBD complex formation.** The docking conformations generated by AutoDock VINA [**55**] as a function of its binding energy (x-axis) and fraction of ACE2/RBD contacts such conformation is able to black (y-axis). The occurrence of such conformation is shown as the z-axis. 246

B.10 **Ligands poses on ACE2.** The docked position of corilagin (right) and TGG (left) on the ACE2 protein. The A1A2 segment, the HS segment and the rest of the RBD is shown in purple, orange and green respectively. The ligand in black and gold is respectively the conformation after docking and the center of the biggest cluster sampled during MD simulation respectively. 246

B.11 **Ligands interaction map with ACE2.** The interaction maps of corilagin (right) and TGG (left) with ACE2 are shown for the center of the biggest cluster computed on the convergence interval using the protein backbone atoms and ligand non-hydrogen atoms. The nonpolar contacts, defined by a distance smaller than 0.40 nm, between the ligand and the protein are shown as red arcs. H-bonds and their donor/acceptor distance are shown in green. All figures were generated using LIGPLOT [**103**, **104**]. 247

B.12 **Convergence of the RBD mutants MD simulations.** Backbone RMSD on the N, Cα, C and O atoms from the wild-type experimental structure and DSSP secondary structure as a function of time for (A) wildtype RBD, (B) E484K, (C) N501Y and (D) E484K/N501Y mutations. 248

B.13 **Convergence of the ligands–RBD MD simulations.** Convergence is assessed by monitoring the RMSD on the backbone atoms of RBD (N, Cα, C and O) and the non-hydrogen atoms of the ligands from the initial structure as a function of time for (A) wildtype RBD, (B) E484K, (C) N501Y and (D) E484K/N501Y mutations. To the left, corilagin–RBD. To the right, TGG–RBD. 249

B.14 **Ligand-protein contact network on A1A2 and RBM.** The contact probability map between the A1A2 segment of ACE2 (vertical axis) and the RBM segment of RBD (horizontal axis) is shown in blue. These probabilities were computed on the converged interval of the MD simulation on the ACE2-RBD complex. The red disks indicate the residues blocked by corilagin (top row) and TGG (bottom row) during our ligand-protein MD simulations on ACE2, RBD, RBD/E484K, RBD/N501Y and RBD/E484K-N501Y. The size of the circles is proportional to the interaction probability with the residues. 250

B.15 **Ligand-protein contact network on HS and RBM.** The contact probability map between the HS segment of ACE2 (vertical axis) and the RBM segment of RBD (horizontal axis) is shown in blue. These probabilities were computed on the converged interval of the MD simulation on the ACE2-RBD complex. The red disks indicate the residues blocked by corilagin (top row) and TGG (bottom row) during our ligand-protein MD simulations on ACE2, RBD, RBD/E484K, RBD/N501Y and RBD/E484K-N501Y. 250

Liste des figures

Liste des sigles et des abréviations

Aβ	Amyloïde-β
CASP	"Critical Assessment of protein Structure Prediction"
HMM	Modèle de Markov caché("Hidden Markov Model")
MM/PBSA	Molecular Mechanics Poisson-Boltzmann Surface Area
OPEP	"Optimized Potential for Efficient peptide-structure Prediction"
PID	Protéine intrinsèquement désordonnée
PMF	Potentiel de champ moyen ("Potential Mean Force")
PSO	Optimisation par essaim particulaire ("Particle Swarm Optimization")
RBD	"Receptor Binding Domain"
RDF	Fonction de distribution radiale ("Radial Distribution Function")

RMSD	Racine de l'erreur quadratique moyenne ("Root Mean Square Deviation")
SA	Alphabet Structurel ("Structural Alphabet")
SPR	Résonance Plasmon Surface ("Surface Plasmon Resonance")
SVM	Machine à vecteurs de support ("support vector machine")
VOC	Variant préoccupant ("Variant of concern")
VOI	Variants d'intérêt ("Variant of interest")

Remerciements

Je ne peux commencer cette section autrement qu'en remerciant mon directeur de thèse, le professeur Normand Mousseau. Cela fait maintenant de nombreuses années que je travaille avec Normand, des stages au baccalauréat au doctorat, et rien de tout cela n'aurait été possible sans la confiance et le soutien qu'il m'a accordés. Malgré le soulagement de terminer (enfin) mes études, c'est avec émotions que je quitte son groupe de recherche.

J'aimerais par la suite remercier tous les collaborateurs de mes différents travaux; professeur Pierre Tuffery pour les travaux sur PEP-FOLD et sOPEP, Sébastien Côté, Sébastien Bélanger et Roger Gaudreault et tout particulièrement, Phuong Trang Nguyen, Mohamed Haddad, professeur Charles Ramassamy et professeur Steve Bourgault pour l'apport des résultats expérimentaux, qui sont, à mon avis, une partie clé de l'article sur la Covid-19. L'expertise et la contribution de tous ont été essentielles à tout ce qui est présenté dans ce document.

J'aimerais aussi remercier tous mes collègues. Ceux du passé: Oscar Restrepo, Sami Mahmoud et Mickaël Trochet. Ceux du présent: Mijanur Rahman, Aynour Koshravi, Jeffrey De Lile, Renaud Girard, Eugène Sanscartier, Joseph Lefebvre et Carl Lévesque, dont je n'ai malheureusement pas eu le plaisir de partager beaucoup le quotidien, pandémie oblige... J'aimerais remercier tout particulièrement Simon Gelin pour l'organisation des fameux festivals de bouffe et les pauses café remplies de discussion sur la bière, la littérature et le cinéma! Je dois aussi donner un remerciement spécial à Sébastien Côté, qui m'a formé lors de mon tout premier stage d'été, et que je côtoie encore aujourd'hui sur une base régulière. Beaucoup de choses ont changé depuis ce tout premier stage, mais pas le plaisir de collaborer avec lui en recherche ou de partager nos heures de diner.

Le plus grand défi du doctorat est qu'il s'agit d'un projet exigeant, stressant, obsédant et qui affecte lentement mais surement tous les aspects d'une vie. Peut-être sans s'en douter, toutes les personnes partageant mon quotidien ces dernières années ont joué des rôles importants dans tous mes accomplissements.

Ainsi, j'aimerais remercier dans un premier temps tous mes amis; les anciens du secondaire ou du CÉGEP, les grimpeurs et les joueurs de go. Je souhaite remercier spécialement Vincent Dumont, colocataire depuis sept ans et ami depuis 15 ans. Je ne peux exprimer

l'utilité d'avoir sous son toit le support de quelqu'un comprenant les hauts et (surtout) les bas du doctorat. L'aide de tous fut fortement apprécié.

Je dois aussi remercier mes parents, ma soeur et toute ma famille pour leur soutien indéfectible. Sans hyperbole, ils croyaient très certainement en moi plus que je ne le faisais moi-même.

Finalement, un grand merci à l'amour de ma vie Kim-Dan Nguyen pour avoir toujours été à mes côtés, dans les bons et mauvais moments, et me permet de donner le meilleur de moi-même. C'est peut-être la fin du doctorat, mais c'est le début de nouvelles aventures.

Introduction

À l'intérieur de tout organisme biologique, on retrouve des processus microscopiques complexes essentiels pour la survie. Ces processus moléculaires, même pour la plus simple cellule, mettent en scène des milliers de molécules aux différentes natures, structures, fonctions, partenaires d'interaction, etc. Parmi ces molécules, on retrouve les protéines qui jouent une panoplie de rôles cruciaux, tant au niveau de la structure que des processus cellulaires [**1**]. En effet, les protéines font partie de nombreux éléments structuraux comme les muscles (actine et myosine), les tendons (glycoprotéines), les os (collagène), et bien d'autres [**1**]. En plus de ces rôles structuraux, les protéines sont aussi de complexes nanomachines. Les fonctions des protéines vont du transport de l'oxygène (hémoglobine) à la catalyse de réactions chimiques (enzymes) en passant par les échanges membranaires (canaux ioniques, aquaporines pour le transport ou les récepteurs couplés aux protéines G pour la communication) et la réparation/modification/réplication du matériel génétique [**1**].

Si leur bon fonctionnement est nécessaire pour la survie cellulaire, les protéines peuvent aussi jouer des rôles délétères. Par exemple, le virus SARS-CoV-2 utilise sa protéine *Spike* pour initier le processus d'infection de l'hôte [**2**,**3**], avec comme conséquence la pandémie de COVID-19 dont les répercussions sont toujours bien présentes. Un autre exemple est celle de la protéine amyloïde-β, dont le mauvais repliement, l'agrégation et la formation de fibres amyloïdes sont des caractéristiques clés de la maladie d'Alzheimer [**4**]. Toutes ces raisons, et bien d'autres encore, font de l'étude des protéines un domaine de recherche incontournable.

Bio-modélisation numérique: en effervescence

Le domaine de la bio-modélisation numérique étudie les systèmes biologiques via des approches hautement multidisciplinaires à la frontière de la biologie, des mathématiques, de la physique, de l'ingénierie et de l'informatique [**5**]. Les nombreuses avancées importantes des dernières années font de la bio-modélisation numérique un domaine en pleine effervescence.

Deux éléments-clés expliquent la popularité grandissante de la bio-modélisation numérique: les avancées matérielles et les avancées méthodologiques. Au niveau des avancées matérielles, la bio-modélisation numérique utilise avec succès les plus récents développements

informatiques, comme l'efficacité grandissante des superordinateurs ou l'utilisation des processeurs graphiques (GPU), afin de repousser les limites du domaine [**6**]. Notablement, le développement de nouveaux superordinateurs conçus et adaptés pour la bio-modélisation, comme ANTON2 [**7**], a permis de repousser les limites temporelles au-delà de la milliseconde. En combinaison avec les avancées matérielles, les avancées méthodologiques permettent habilement de lever certaines entraves sans changement au niveau informatique [**6**]. L'échantillonnage avancé, comme l'échange de répliques ou la métadynamique permettant d'accélérer l'échantillonnage des simulations, les nouvelles percées de l'apprentissage machine, par exemple dans la prédiction *de novo* de la structure des protéines, ou l'utilisation de chaînes de Markov pour l'analyse des données, ne sont que quelques exemples d'algorithmes récents qui propulsent le domaine vers de nouveaux horizons [**5**, **6**, **8**]. Ainsi, dans les 50 dernières années, l'impact des avancées matérielles et des avancées méthodologiques en bio-modélisation numérique surpasse même les prédictions de la loi de Moore [**8**].

La bio-modélisation numérique est aujourd'hui un complément incontournable aux méthodes expérimentales [**5**] et les chercheurs combinent maintenant de façon routinière ces deux types de méthodes [**6**] dans le but d'étudier une multitude de phénomènes biologiques d'intérêt: améliorer le long et coûteux processus de développement de nouveaux médicaments [**9**], étudier les protéines amyloïdes associées à la maladie d'Alzheimer [**4**] ou étudier d'autres types de molécules comme les acides nucléiques [**10**].

Thermodynamique vs dynamique/Photo vs film

Grâce à d'importants efforts au niveau expérimental, incluant la cristallographie aux rayons-X, la résonnance magnétique nucléaire et la cryo-microscopie électronique, les structures de plus de 150 000 protéines [**8**] sont disponibles gratuitement en ligne sur la *Protein Data Bank* [**11**].

Ces gigantesques bases de données sont à la base des méthodes "knowledge-based" qui utilisent des algorithmes d'apprentissage afin de découvrir les notions pertinentes contenues dans celles-ci et d'extrapoler ces résultats à des problèmes similaires. Ces méthodes sont parmi les plus populaires pour faire la prédiction structurelle *de novo* pour les protéines, c'est-à-dire, de prédire la structure native (tridimensionnelle) uniquement à partir de la séquence en acide aminé. Tout récemment, la méthode d'apprentissage machine AlphaFold, développée par Google, a obtenu des résultats d'une qualité sans précédent à la dernière rencontre CASP ("Critical Assessment of protein Structure Prediction") [**12**]. Par contre, l'applicabilité des méthodes "knowledge-based" est plus limitée lorsque la quantité/qualité de données n'est pas suffisante, comme pour l'étude de l'ARN ou des protéines intrinsèquement désordonnées [**6**, **8**].

La structure native d'une protéine, bien qu'elle soit d'un grand intérêt, ne représente qu'une des conformations d'intérêt; la structure native est une photo de l'état final. Or, une multitude de paramètres importants pour un système ne peut se calculer uniquement à partir de l'état natif.

Comparativement aux méthodes "knowledge-based", les méthodes physiques sont, quant à elles, dérivées des principes fondamentaux des lois de la physique. Comme elles sont basées uniquement sur des principes premiers, elles restent pertinentes pour étudier des systèmes mal représentés par les bases de données [**8**]. Ces méthodes sont derrière une partie importante du développement et de la paramétrisation des champs de force utilisés pour les simulations numériques. Les simulations numériques permettent d'obtenir l'évolution temporelle d'un système, un peu à la manière d'un film. Elles permettent de dériver de l'information complémentaire à celle obtenue via la structure native. Distributions des conformations, énergies libres, taux de transition, constantes d'équilibre ne sont que quelques paramètres pouvant être dérivés des simulations numériques [**8**].

En bref, c'est en combinant la "photo" de la structure native, dérivée de méthodes "knowledge-based", et le "film", dérivé de méthodes physiques, que l'on obtient le portrait global du phénomène d'intérêt.

Rapidité vs complexité

De nombreux phénomènes biologiques d'intérêt sont des processus multi-échelle, tant au niveau temporel qu'au niveau spatial. Par exemple, le temps de repliement d'une petite épingle à cheveux d'ARN est de l'ordre de la microseconde à la milliseconde tandis que le temps de repliement d'un riboswitch avec ligand est de l'ordre de la seconde, de la minute ou même plus [**10**]. Un autre exemple est le processus de formation des fibres amyloïdes par la protéine amyloïde-β; le monomère, de quelques centaines d'atomes, à la formation de la fibre amyloïde qui peut contenir quelques milliers de protéines d'β-amyloïde. Ainsi, la taille et la complexité des systèmes d'intérêt étudiés à l'aide des méthodes numériques augmentent continuellement [**6**].

L'étude de systèmes plus grands sur des échelles de temps plus longues nécessite un certain nombre de simplifications. La Figure 0.1 présente différentes méthodes numériques le long des axes rapidité et complexité. Dans le coin inférieur gauche, zone (A), on retrouve les méthodes qui sont les plus rigoureuses au niveau théorique, mais qui requièrent la plus grande intensité de ressources comme les calculs *Ab initio* de mécanique quantique ou les calculs d'énergie libre par perturbations alchimiques. Au centre, zone (B), on retrouve les méthodes offrant une balance entre la rapidité et la complexité; plus rapides que les méthodes de la zone (A) au prix de certaines approximations. Dans cette zone, on retrouve les simulations de mécanique moléculaire ou l'estimation de l'énergie libre comme la MMPB/SA. Finalement, la zone

(C) regroupe les méthodes focalisant essentiellement sur la rapidité avec une simplification importante du modèle, comme les modèles gros-grain ou l'amarrage moléculaire. Toutes ces méthodes ont leurs forces et leurs faiblesses et le choix d'une (ou plusieurs) d'entre elle(s) dépend de la nature du problème étudié.

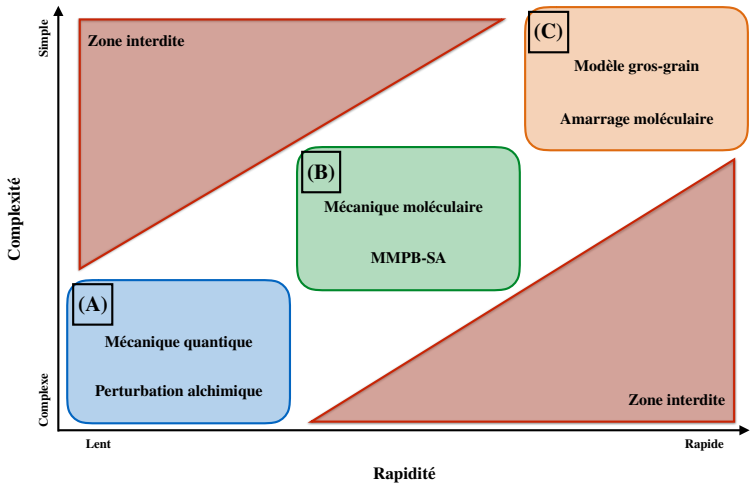


Fig. 0.1. RAPIDITÉ -vs- COMPLEXITÉ. Ce schéma représente le positionnement de certaines méthodes en termes de rapidité (axe horizontal) et de complexité (axe vertical). La zone (A), en bleu, correspond aux méthodes complexes et lentes, la zone (B), en vert, correspond aux méthodes intermédiaires et la zone (C), en orange, correspond aux méthodes rapides, mais simplifiées. La zone rouge représente la zone interdite: des méthodes simplifiées et lentes ou des méthodes complexes et rapides.

Aperçu

Dans cette thèse, je présenterai les divers travaux effectués au cours de mon doctorat. Ces travaux s'orientent principalement le long de deux axes sur lesquels reposent les trois sections de cette thèse. Le premier axe est l'utilisation des méthodes numériques traditionnelles: simulations numériques tout-atome et méthodes basées sur la physique. Ces méthodes ont servi à l'étude du potentiel thérapeutique de petites molécules contre la COVID-19. Ces simulations permettent de générer l'évolution temporelle du système et d'obtenir le "film" et non uniquement la "photo". De plus, notre étude a été complémentée de résultats expérimentaux et démontre la véritable synergie des approches.

Le second axe de cette thèse est l'utilisation et le développement de méthodes simplifiées pour l'étude des protéines. Comme mentionné précédemment, les systèmes d'intérêt en biologie sont complexes et mettent en relation une multitude de partenaires d'interaction, d'environnements, de ligands, etc., comme ce fut le cas pour notre étude des molécules thérapeutiques contre la COVID-19. De ce fait, leur étude nécessite le développement de méthodes simples, rapides et efficaces. C'est sur ce second axe que reposeront les deux

dernières sections de cette thèse. Dans un premier temps, je discuterai des améliorations apportées à la méthode PEP-FOLD, une approche simplifiée pour la prédiction *de novo* de la structure des peptides et des petites protéines. Si cette méthode est principalement "knowledge-based", elle utilise aussi un potentiel gros-grain, sOPEP, pour lequel des paramètres optimisés sur de larges bases de données viennent modifier les interactions physiques. Ce potentiel gros-grain est dérivé de la famille des potentiels OPEP et une des parties du projet visait son amélioration et la ré-optimisation de ses paramètres. Dans un second temps, je décrirai les résultats préliminaires au niveau du développement du potentiel simplifié aaOPEP. Comme pour sOPEP utilisé dans PEP-FOLD, aaOPEP est développé à l'aide de la philosophie des potentiels OPEP, mais en régime tout-atome. Ce potentiel est développé avec l'objectif d'étudier les processus d'agrégation et de fibrillation d'amyloïde-β associés à la maladie d'Alzheimer, un système complexe, multi-échelle et difficile à étudier à l'aide des méthodes numériques traditionnelles.

Ensemble, ces trois chapitres forment un bon aperçu des outils de la bio-modélisation numérique; de l'utilisation des méthodes complexes de simulation tout-atome au développement de méthodes simplifiées gros-grain, de la prédiction de la structure native ("photo") à l'ensemble des conformations de la mécanique statistique ("film").

Chapitre 1

Biologie pour l'étude de SARS-CoV-2

Détectée pour la première fois en décembre 2019 en Chine, dans la région de Wuhan, la maladie à coronavirus 2019 (COVID-19) est à l'origine d'une pandémie mondiale avec, en date du 21 septembre 2021, un cumulatif de 228 millions de cas et de 4.6 millions de décès [**13**]. Cette maladie est causée par un virus de la famille des β-coronavirus, baptisé SARS-CoV-2, dont l'origine, si elle reste incertaine, serait zoonotique [**2**, **14**]. Le développement rapide de vaccins efficaces contre la SARS-CoV-2 permet un meilleur contrôle de la pandémie dans de nombreux pays. Par contre, les mutations continuelles du virus mènent à l'émergence de multiples variants qui pourrait fragiliser le retour à la normale. Ces nouveaux *variants préoccupants* (VOC) et *variants d'intérêt* (VOI) pourraient modifier le taux de transmission, la sévérité et le taux de réinfection à la maladie, en plus du risque d'évasion immunitaire.

Dans ce chapitre, je ferai un bref aperçu du contexte dans lequel s'insèrent nos travaux présentés au chapitre suivant. Dans l'ordre, je discuterai de la maladie de la COVID-19, du fonctionnement du virus SARS-CoV-2, des différents variants et du développement thérapeutique. Il est crucial de mentionner que la pandémie et la recherche sur la COVID-19 évoluent rapidement. Les sections de cette thèse en lien avec la COVID-19 sont donc limitées aux informations disponibles lors de leur rédaction (septembre 2021). Ainsi, des parties cruciales de la suite de la pandémie, notamment, l'émergence du variant Omicron et le développement du médicament Paxlovid de Pfizer, ne seront pas discutées.

1.1. Maladie

La COVID-19 est une maladie respiratoire dont les principaux symptômes sont la fièvre, la toux et la dyspnée (gêne respiratoire). Une combinaison de ces trois symptômes a été observée pour près de 70% des malades [**15**]. Une panoplie d'autres symptômes a aussi été observée incluant les douleurs musculaires (36%), les maux de tête (34%), les maux de gorge (20%), la diarrhée (19%), la nausée (11%) et la perte de goût/odorat (8%) [**15**].

En plus de la nature des symptômes, leur sévérité varie grandement: des patients infectés asymptomatiques, c'est-à-dire qui ne présentent aucun symptôme, jusqu'aux patients infectés ayant des symptômes sévères de détresse respiratoire et de malfonctionnement de multiples organes pouvant mener à la mort, en passant par les symptômes mineurs et majeurs. On estime qu'entre 17% et 33% des patients infectés seront asymptomatiques [**2**]. La présence de comorbidités (diabète, maladie cardio-vasculaire, etc.) augmente d'un facteur six le risque d'hospitalisation et d'un facteur 12 le risque de décès [**15**].

1.2. Fonctionnement du virus

À l'origine de la COVID-19, on retrouve le virus baptisé SARS-CoV-2, de la classe des coronavirus. Du latin corona, signifiant couronne, les coronavirus sont caractérisés par une enveloppe virale recouverte de glycoprotéines, rappelant la forme de couronne à l'origine du nom [**2**, **3**]. Cette enveloppe contient un simple brin d'ARN à polarité positive encodant le génome du virus [**3**]. Le génome des coronavirus est parmi les plus grands des virus à ARN avec entre 26 et 32kb [**3**, **14**]. SARS-CoV-2 est proche parent ($\sim$96% d'identité de séquence) avec certains coronavirus-β que l'on retrouve chez les chauve-souris [**2**, **16**]. Bien qu'elle reste nébuleuse, l'origine du virus serait zoonotique; il se serait transmis à l'humain directement de la chauve-souris ou via une espèce intermédiaire [**2**, **16**].

Comprendre les mécanismes moléculaires du virus à l'origine du développement de la COVID-19 est crucial pour le développement d'un traitement thérapeutique efficace [**17**]. Pour se reproduire, les virus doivent pénétrer le cytoplasme d'une cellule hôte afin d'utiliser son appareil de traduction, tout en évitant la réponse immunitaire. La Figure 1.1 présente un schéma simplifié des mécanismes d'action du virus.

Pour SARS-CoV-2, la première étape de l'infection est la reconnaissance des cellules hôtes, présentée dans la zone (A) de la Figure 1.1. Pour ce faire, SARS-CoV-2 utilise la protéine *Spike* qui se situe sur son enveloppe. La protéine *Spike* est sous-divisée en deux sections: S1 et S2. S1 est exposé au solvant et contient le "receptor-binding domain", RBD, qui permet l'interaction avec les cellules de l'hôte et donnera la spécificité aux cellules touchées. S2, quant à lui, est une région transmembranaire impliquée dans les réarrangements structuraux menant à la fusion de la membrane du virus et de la cellule hôte [**3**]. Dans le cas de SARS-CoV-2, la protéine *Spike* se lie au niveau de l'enzyme de conversion de l'angiotensine 2 (ACE2). En effet, la surexpression de ACE2 de plusieurs espèces (humain, cochon, civette) en cellules HeLa a montré que SARS-CoV-2 peut pénétrer les cellules, ce qui n'arrive pas lorsqu'il n'y a pas d'expression de ACE2 [**18**]. ACE2 se retrouve sur les cellules à la surface de multiples organes comme le coeur, le foie, les reins et les poumons, tout particulièrement au niveau des cellules épithéliales de type II que l'on retrouve dans les alvéoles [**19**, **20**]. ACE2 joue un rôle crucial au niveau du système rénine-angiotensine (RAS). Le RAS est un

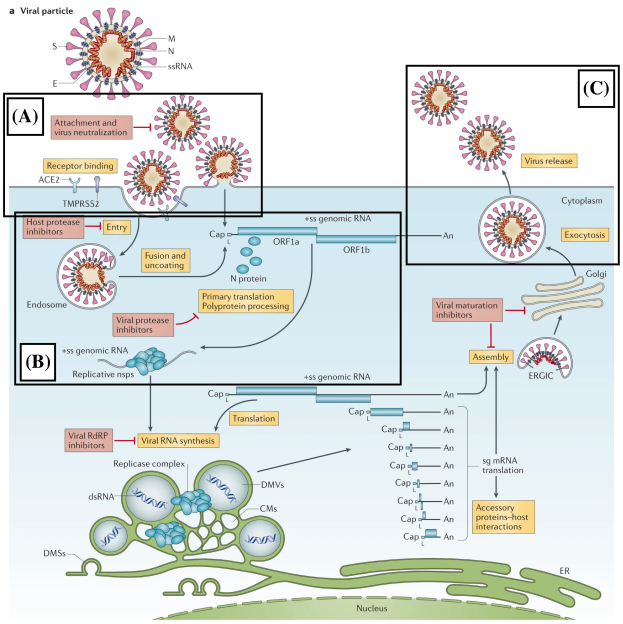


Fig. 1.1. Mécanisme d'action du virus SARS-CoV-2. La figure est tirée de l'article de Philip V'kovski *et coll.* [**3**]. La zone (A) inclut les étapes de reconnaissance de la cellule hôte. La zone (B) inclut les étapes de réplication virale. Finalement, la zone (C) inclut les étapes, d'expulsion de nouveaux virus.

système complexe d'événements moléculaires qui permet de réguler, entre autres, la pression sanguine et l'homéostasie des électrolytes, crucial pour de nombreux organes comme le coeur, les vaisseaux sanguins et les reins. ACE2 fait l'hydrolyse de l'angiotensine II, un vasoconstricteur, en angiotensine (1-7), un vasodilatateur [**19**]. Plus spécifiquement, c'est le "receptor-binding motif", RBM, du RBD qui interagit avec ACE2 [**21**].

Une fois l'étape de reconnaissance réalisée, SARS-CoV-2 doit maintenant pénétrer le cytoplasme de la cellule. Pour ce faire, la protéine *Spike* est clivée par la protéase transmembranaire à sérine 2 (TMPRSS2), une peptidase présente sur la membrane de la cellule hôte [**3**]. Cette peptidase est essentielle et son inhibition est suffisante pour prévenir l'entrée du virus dans la cellule [**22**]. Ce clivage par TMPRSS2 mène ultimement à la fusion des membranes du virus et de la cellule hôte et à la libération du génome du virus.

Une fois le génome à l'intérieur du cytoplasme de la cellule hôte, une série de processus menant à la prise de contrôle du système de traduction de la cellule hôte et à la multiplication virale est enclenchée. Ces étapes sont présentées dans la zone (B) de la Figure 1.1. À l'extrémité 5' du génome du virus, on retrouve deux cadres de lecture ouverts: ORF1a et ORF1b. Ceux-ci encodent deux polyprotéines, de longues protéines produites par le virus qui seront clivées via des peptidases (du virus ou de la cellule hôte) et produiront les protéines virales. Le clivage des deux polyprotéines, pp1a et pp1b, mène à l'obtention de 16 protéines non-structurelles (nsp) (encodées par le virus, mais ne faisant pas partie du

virus). 15 d'entre elles forment le centre de transcription et de réplication virale (RTC) contenant, entre autres, des enzymes pour le traitement, la modification et la correction du génome du virus [**3**]. Notamment, le nsp3 et nsp5 contiennent des peptidases à cystéine. La peptidase du nsp5, désignée par $3CL^{pro}$ est responsable de la majorité des clivages des polyprotéines. $3CL^{pro}$ s'auto-clive directement des polyprotéines et par la suite clivera ces dernières à 11 emplacements distincts [**16**]. Le clivage de nsp1 se fait aussi rapidement puisqu'il est impliqué dans la prise de contrôle du système de traduction de la cellule hôte; il permet de favoriser la traduction de l'ARN viral plutôt que de l'ARN cellulaire par la cellule hôte [**3**]. Finalement, les protéines nsp2 à nsp16 composent le RTC. Tout particulièrement, nsp12 à nsp16 contiennent les enzymes permettant la synthèse, la modification et la correction de l'ARN tandis que nsp2-nsp11 ont un rôle de support dans la modulation de la membrane intracellulaire, l'évitement de la réponse immunitaire de la cellule hôte et la production de cofacteurs. La synthèse de l'ARN est faite par l'ARN polymérase ARN-dépendante (RdRP) de nsp12. nsp14 joue le rôle d'exonucléase 3'-5' et corrige les erreurs sur le brin d'ARN. Une fois le génome du virus répliqué, il peut alors être utilisé pour former de nouveau RTC ou bien être assemblé en de nouveaux virions au niveau du réticulum endoplasmique/appareil de Golgi. Ces virions seront ensuite expulsés de la cellule hôte par exocytose et pourront infecter d'autres cellules [**3**], comme présenté à la zone (C) de la Figure 1.1.

1.3. Développement thérapeutique

Une de ces approches thérapeutiques est le développement de médicaments [**17**,**23**]. Trois étapes du mécanisme d'action du virus, présentées à la section précédente, sont principalement étudiées pour le développement d'un médicament:

- **L'interface ACE2/Protéine-S**: La protéine-S de SARS-CoV-2 située à la surface du virus permet de reconnaître le récepteur ACE2 de la cellule de l'hôte [**3**].
- **La protéase TMPRSS2**: La protéase transmembranaire TMPRSS2 des cellules hôtes clive la protéine-S du virus permettant la fusion de leur membrane respective. Il a été montré que l'inhibition de TMPRSS2 est suffisante pour prévenir l'entrée de SARS-CoV-2 dans les cellules [**22**].
- **La protéase $3CL^{pro}$**: La protéase $3CL^{pro}$ est la principale protéase encodée dans le génome du virus et clive les poly-protéines pp1a et pp1ab en onze sites. Cette étape est essentielle pour la formation du RTC qui permet la réplication du génome du virus [**3**, **16**].

Ainsi, le développement de molécules thérapeutiques pouvant interférer avec une (ou plusieurs) des étapes mentionnées ci-dessus pourrait mener à l'obtention d'un médicament contre la COVID-19. Malgré la recherche intensive de molécules thérapeutiques depuis le début de la pandémie, il n'y a, à ce jour (septembre 2021), toujours aucun médicament

disponible contre la COVID-19 [**2**]. Les résultats positifs de certaines molécules (Remdesivir, colchicine, etc.), démontrés par certaines études au début de la pandémie, n'ont pu être reproduits [**2**].

D'autres approches sont aussi présentement utilisées/étudiées. Des vaccins efficaces furent obtenus à la fin de 2020 et les campagnes de vaccination se poursuivent encore partout sur la planète. En date du 21 septembre 2021, plus de 5.5 milliards de doses ont été administrées dans le monde [**24**]. Des ensembles d'anticorps, comme le Bamlanivimab, le Casirivimab, etc. peuvent aussi être prescrits pour les patients infectés avec symptômes mineurs présentant des risques de complications [**2**, **14**]. Finalement, des traitements immunomodulateurs, comme des corticostéroïdes, peuvent être utilisés pour contrer une réponse excessive du système immunitaire et des chocs cytokiniques [**2**, **14**].

1.4. Variants

À cause des pressions évolutives, SARS-CoV-2 est en constante mutation ce qui permet l'émergence de nouveaux variants. L'étude de ces variants est d'une importance capitale, puisqu'ils pourraient mener à un changement important de l'épidémiologie de la maladie ou réduire l'efficacité des traitements actuels et des vaccins. Par exemple, certaines données semblent démontrer que certains variants ont une plus grande résistance aux traitements avec anticorps et aux vaccins [**25**, **26**].

Depuis le début de la pandémie, l'Organisation Mondiale de la Santé (OMS) répertorie l'émergence des nouveaux variants. Les variants dits *préoccupants* (VOCs) possèdent une ou plusieurs des caractéristiques suivantes: (1) une augmentation du taux de transmission ou de l'épidémiologie, (2) une augmentation de la virulence ou du diagnostic clinique et (3) une perte d'efficacité des vaccins, des mesures de santé publique, etc. [**27**]. Jusqu'à présent, quatre variants de SARS-CoV-2 répondent à ces critères. L'appellation de ces variants a passablement changé depuis leur émergence. Le Tableau 1.1 présente un récapitulatif des différentes appellations des VOCs, de leur lieu d'émergence et des mutations au niveau du RBD de la protéine-S. Ces variants se sont rapidement propagés hors de leur lieu d'émergence et les variants Alpha, Beta, Gamma et Delta ont été détectés respectivement dans 193, 142, 96 et 185 pays différents [**13**].

De nombreuses études sont présentement en cours afin de déterminer l'impact de ces variants au niveau de la transmission, de la sévérité, de la réinfection et de l'efficacité du vaccin. Les résultats restent préliminaires et ne font pas toujours consensus. L'agrégation des résultats par l'OMS est présentée au Tableau 1.2. Globalement, tous les VOCs présentent une augmentation du taux de transmission. Les variants Alpha, Gamma et Delta présentent une augmentation de la sévérité de la maladie menant à une augmentation des risques d'hospitalisation. Les effets sur la sévérité du variant Beta restent inconnus. Les variants Beta,

VOC	Ancienne Appellation	Première détection	Mutation du RBD
Alpha	B.1.1.7	Royaume-Uni	N501Y, A570D
Beta	B.1.351	Afrique du Sud	K417N, E484K, N501Y
Gamma	P.1	Brésil	K417N, E484K, N501Y
Delta	B.1.617.2	Inde	L452R, T478K, E484Q

Tableau 1.1. Variants Préoccupants: Informations générales. Le tableau présente, de gauche à droite, la nouvelle et l'ancienne appellation, le pays de première détection et certaines mutations cruciales au niveau du RBD pour chacun des variants préoccupants.

	Alpha	Beta	Gamma	Delta
Transmission	↑	↑	↑	↑
Sévérité	↑	Inconnu	↑	↑
Réinfection	=	↑	↑	↑
Vaccination	=	=: Sévère / ↓: symptômatique	Inconnu	=: Sévère / ↓: symptômatique

Tableau 1.2. Variants Préoccupants: Impact sur l'épidémiologie. Chacune des colonnes présente les résultats pour un variant préoccupant en termes de transmission (ligne 1), de sévérité des symptômes (ligne 2), de risque de réinfection (ligne 3) et d'impact sur la vaccination (ligne 4). Le tout est tiré du site de l'OMS [**13**].

Gamma et Delta présentent une augmentation des risques de réinfection. Finalement, le facteur le plus important concerne l'efficacité des vaccins. Pour le moment, seuls les variants Beta et Delta semblent avoir un impact sur l'efficacité du vaccin, uniquement pour les cas symptômatiques. L'efficacité de la prévention des cas graves de la maladie est préservée pour tous les VOCs.

Finalement, en plus des VOCs, de nombreux autres variants émergent en permanence. L'OMS utilise la classification de *variants d'intérêt* (VOI) lorsqu'un variant possède: (1) des modifications génétiques associées à de possibles modifications de l'épidémiologie du virus (transmission, sévérité, réinfection, vaccination, etc.) et (2) la présence de transmission communautaire et une augmentation du nombre de cas du variant dans le temps. Présentement, les variants Mu et Lambda sont classés comme VOIs.

1.5. Contexte

Bien que des vaccins efficaces contre la COVID-19 aient été développés, ils ne sont pas infaillibles et les risques de réinfection sont bien réels. De plus, l'émergence de nombreux variants pourrait compromettre l'efficacité des vaccins et la fin de la pandémie. Le développement et l'amélioration de méthodes thérapeutiques, vaccins, médicaments, anticorps, demeure pertinent. Dans ce contexte, nous avons étudié le potentiel thérapeutique de deux petites molécules, la corilagine et la 1,3,6-Tri-O-galloy-β-D-glucose (TGG) sur certaines des mutations clés des variants de SARS-CoV-2; E484K qui est présente au niveau des variants Beta et Gamma et N501Y qui est présente au niveau des variants Alpha, Beta et Gamma.

Nous nous sommes intéressés à l'interface entre ACE2 et le RBD de la protéine *Spike*, cruciale pour la reconnaissance des cellules hôtes par le virus.

Chapitre 2

Méthode pour l'étude de SARS-CoV-2

Notre étude sur l'impact de petites molécules thérapeutiques, Corilagin et TGG, sur les variants du SARS-CoV-2, s'est effectuée en deux parties. La première partie est numérique. Nous avons combiné des simulations d'amarrage moléculaire (docking) et de dynamique moléculaire (MD). L'affinité des ligands pour ACE2 et la protéine *Spike* fut estimée par des calculs de MMPB-SA. Cette première partie offre un portrait au niveau atomique des interactions protéines/ligands.

La seconde partie est expérimentale. Des expériences immuno-enzymatiques (ELISA) et de résonance plasmon de surface (SPR) vont complémenter et renforcer les conclusions numériques de l'étude.

Dans ce chapitre, je présenterai quelques notions théoriques sous-jacentes aux méthodes numériques et expérimentales utilisées dans l'article. L'objectif de cette section est de faciliter l'interprétation et d'indiquer les limites de notre étude.

2.1. Mécanique moléculaire

Les molécules biologiques sont des objets quantiques et, pour être théoriquement rigoureux, devraient être étudiées via les lois de la mécanique quantique, c'est-à-dire, en solutionnant l'équation de Schrodinger [**1**]. Or, les processus biologiques d'intérêt, comme le repliement des protéines, l'interaction avec des ligands, etc., sont impossibles à étudier selon les lois de la mécanique quantique, car: (1) ils se produisent sur des échelles de temps couvrant plusieurs ordres de magnitude, de la femto-seconde pour la vibration des liens, à la seconde pour le repliement des protéines, et (2) ils peuvent impliquer des centaines/milliers d'atomes/molécules simultanément. De plus, pour pouvoir étudier le système à l'équilibre et pouvoir comparer avec l'expérience, il faut pouvoir échantillonner de façon exhaustive les états importants avec les bonnes proportions, données par les facteurs de Boltzmann [**28**]. Pour la majorité des systèmes d'intérêt, il s'agit d'un processus lent et coûteux computationnellement. Ainsi, afin d'étudier ces systèmes, il faut avoir recours à des approximations.

La *mécanique moléculaire* offre une simplification comparativement aux lois de la mécanique quantique ce qui permet l'étude de processus biologiques plus vastes sur des échelles de temps plus longues. Dans cette approche, les atomes sont décrites par des sphères rigides localisées à la position du noyau et reliées entre elles par des ressorts. Les électrons sont considérés implicitement. Les interactions entre les atomes sont décrits par un champ de force (ou potentiel), un ensemble d'équations décrivant l'énergie du système (V) en fonction des coordonnées atomiques (**r**).

2.1.1. Champ de force

Les champs de force (ou potentiel) sont un ensemble de fonctions mathématiques simples permettant d'estimer les interactions (l'énergie) du système en fonction des coordonnées atomiques (**r**). Les interactions sont habituellement divisées en deux types: les interactions liées, qui font intervenir un (ou plusieurs) lien(s) atomique(s) et les interactions non-liées. De façon générale, le potentiel peut s'écrire:

$$\begin{aligned} V(\mathbf{r}_{ij}) &= V_{liée} + V_{non-liée} \\ V_{liée} &= V_{lien}(\mathbf{r}_{ij}) + V_{angle}(\theta_{ijk}) + V_{dièdre}(\phi_{ijkl}) \\ V_{non-liée} &= V_{LJ}(\mathbf{r}_{ij}) + V_{Coulomb}(\mathbf{r}_{ij}) \end{aligned} \tag{2.1.1}$$

Les termes de chacune des équations seront explicités dans ce qui suit.

2.1.1.1. Interactions liées. Les potentiels de chacun des termes des interactions liées sont donnés par les équations suivantes:

$$V_{lien}(r_{ij}) = \frac{1}{2} k^b_{ij} \left(r_{ij} - r^0_{ij} \right)^2 \tag{2.1.2}$$

$$V_{angle}(\theta_{ijk}) = \frac{1}{2} k^\theta_{ijk} \left(\theta_{ijk} - \theta^0_{ijk} \right)^2 \tag{2.1.3}$$

$$V_{diedre}(\phi_{ijkl}) = \sum_{n=1}^{N} V_n (1 + \cos\left(n \cdot \phi_{ijkl} - \phi^0_n \right)) \tag{2.1.4}$$

Visuellement, les interactions liées et le potentiel associé sont présentés à la Figure 2.1.

La première interaction liée est le **lien atomique** qui, tel que mentionné précédemment, est modélisé comme un ressort. Le potentiel associé aux liens atomiques est donné par l'équation 2.1.2. Il s'agit d'un simple potentiel harmonique centré autour de la longueur d'équilibre du lien r^0 avec une constante rappel k^b.

La seconde interaction liée met en scène deux liens atomiques. Il s'agit de l'**angle de valence**, c'est-à-dire, de l'angle formé par deux liens atomiques à partir d'un atome commun. Le potentiel associé aux angles de valence est donné par l'équation 2.1.3. Comme pour le lien atomique, il s'agit d'un simple potentiel harmonique avec une constante de rappel k^θ centré autour de la valeur d'angle d'équilibre θ^0.

Finalement, la troisième interaction liée met en scène trois liens atomiques. Il s'agit de l'**angle dièdre**, c'est-à-dire, de l'angle entre les plans formés par deux liens atomiques à partir d'un troisième lien, commun aux deux plans. Le potentiel associé aux angles dièdres est donné par l'équation 2.1.4. Il s'agit d'une série de Fourier tronquée au terme N avec une amplitude V_n et un déphasage ϕ_n^0.

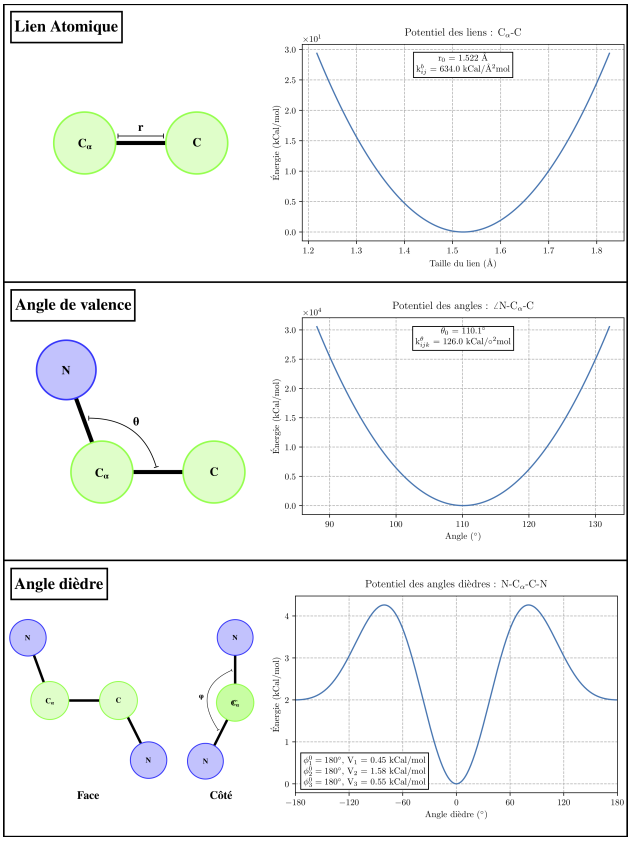


Fig. 2.1. Interactions liées et le potentiel associé. De haut en bas, on retrouve les liens atomiques, les angles de valence et les angles dièdres. À gauche, on retrouve une représentation visuelle de l'interaction et à droite le potentiel associé à cette interaction. Tous les paramètres sont tirés de AMBER99SB*-ILDN [**29**].

2.1.1.2. Interactions non-liées. Les interactions non-liées sont les interactions de répulsion/dispersion et les interactions électrostatiques dont les équations associées sont présentées ci-dessous.

$$V_{LJ}(r_{ij}) = \epsilon \left[\left(\frac{r_0}{r_{ij}} \right)^{12} - 2 \left(\frac{r_0}{r_{ij}} \right)^{6} \right] \tag{2.1.5}$$

$$V_{Coulomb}(r_{ij}) = \frac{q_i q_j}{4\pi\epsilon_0\epsilon_r r_{ij}} \tag{2.1.6}$$

Visuellement, les interactions non-liées et le potentiel associé sont présentés à la Figure 2.2.

Les interactions de répulsion dispersion (dispersion de London) sont décrites par un potentiel de Lennard-Jones donné à l'équation 2.1.5 où r_0 est la position du minimum et ϵ est la profondeur du minimum. Le terme en r^{-6} est dérivé de la mécanique quantique et est associé à l'attraction entre atomes neutres due aux fluctuations de la densité de charge [**1**,**28**]. Le terme en r^{-12} est une grossière approximation de la répulsion des couches électroniques (principe d'exclusion de Pauli) entre deux atomes à très courtes distances [**1**, **28**].

Finalement, les interactions électrostatiques entre les charges, qui peuvent être partielles, associées aux différents atomes sont décrites par le potentiel de Coulomb, donné à l'équation 2.1.6. Le paramètre q_i est la charge associée à l'atome i tandis qu'ϵ_r et ϵ_0 sont respectivement la constante diélectrique relative et la constante diélectrique du vide respectivement. Il est à noter que dans la majorité des potentiels classiques, les charges associées à chacun des atomes sont fixes. En d'autres mots, les effets de polarisation ne sont pas considérés. Cette simplification permet d'accélérer l'échantillonnage, un des facteurs limitants pour la comparaison avec l'expérience.

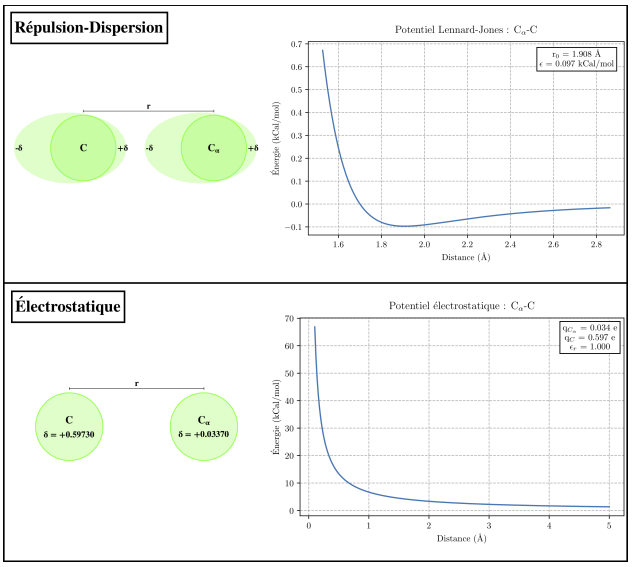


Fig. 2.2. Interactions non-liées et le potentiel associé. De haut en bas, on retrouve les interactions répulsives/attractives et les interactions électrostatiques. À gauche, on retrouve une représentation visuelle de l'interaction et à droite le potentiel associé à cette interaction. Tous les paramètres sont tirés de AMBER99SB*-ILDN [**29**].

2.1.1.3. Protéine et Ligand. La forme du potentiel, c'est-à-dire, les expressions mathématiques données précédemment, est sensiblement la même pour la majorité des potentiels tout-atome courant; AMBER [**30**], OPLS [**31**] et CHARMM [**32**,**33**]. Les différences pour ces potentiels viennent surtout des distinctions de philosophie dans l'optimisation des paramètres. Dans notre étude sur SARS-CoV-2, nous avons utilisé le potentiel AMBERff14SB [**30**] pour

décrire les protéines et le potentiel AMBER généralisé ("generalized AMBER forcefield", GAFF) [**34**], l'extension de AMBER pour tous les types de molécules, pour le ligand.

AMBERff14SB [**30**] propose une nouvelle paramétrisation des angles dièdres de la chaîne principale et des principaux angles dièdres des chaînes latérales afin de combler certaines problématiques observées pour AMBERff99SB [**35**]; problèmes au niveau des rotamères des chaînes latérales et de la balance des structures secondaires [**30**]. Pour les angles dièdres de la chaîne principale, AMBERff14SB propose une modification empirique du potentiel afin de compenser le manque de données de mécanique quantique pour la transition $\beta \rightarrow ppII$ dans la paramétrisation de AMBERff99SB. Pour ce faire, les auteurs ont réalisé des simulations sur la tri-alanine (petit peptide composé de trois alanines) avec AMBERff99SB à 300K avec le modèle d'eau TIP3P. De ces simulations, les populations des paires pour les angles dièdres ϕ/ψ (par tranche de 5°) furent extraites. Avec les équations de Karplus, ils ont converti cette grille en valeurs de couplage scalaire (J-coupling) qui furent ensuite comparées aux mesures de ^{3}J($H_N H_\alpha$) de résonnance magnétique nucléaire [**30**]. Pour ce qui est des angles dièdres principaux des chaînes latérales (χ), les paramètres du potentiel furent optimisés à l'aide de comparaisons avec des calculs de mécanique quantique (RHF/6-31G*). À nouveau, des simulations avec AMBERff99SB furent réalisées sur des dipeptides pour tous les résidus sauf ALA, GLY et PRO. Les dipeptides sont initialement en conformation α ($\phi = -60°$,$\psi = -45°$) ou β($\phi = -130°$,$\psi = 135°$). À l'aide d'un algorithme génétique, les amplitudes (V_n) et les déphasages (ϕ_n^0) furent optimisés afin de minimiser les différences d'énergie entre les calculs de mécanique quantique et les simulations de mécanique moléculaire. AMBERff14SB fut testé sur quelques petites protéines/peptides et a permis d'obtenir une meilleure distribution des rotamères et une stabilisation plus optimale des hélices α, tout en conservant la stabilité au niveau des feuillets-β.

Pour ce qui est du ligand, nous avons utilisé le potentiel AMBER généralisé [**34**], GAFF, qui fut développé afin d'étendre les capacités de AMBER à l'étude des molécules autres que les protéines ou l'ADN/ARN. Tout particulièrement, le développement de GAFF s'est fait avec pour objectif l'étude et le développement de médicaments. Le potentiel GAFF introduit 35 nouveaux types d'atomes dans AMBER; cinq pour le carbone, huit pour l'azote, trois pour l'oxygène, cinq pour le soufre, huit pour le phosphore, six pour l'hydrogène et quatre pour les halogènes (F, Cl, Br, I). Ces nouveaux types d'atomes sont associés à des types d'atomes, des hybridations, des aromaticités et/ou des environnements locaux qui ne sont pas retrouvés au niveau des protéines/acides nucléiques. La forme utilisée pour le potentiel est la même que dans AMBER, et est décrite par les équations 2.1.2 à 2.1.6. La paramétrisation du potentiel pour ces nouveaux types d'atomes s'est effectuée en quatre étapes principales. Premièrement, les paramètres de van der Waals (équation 2.1.5), sont directement tirés de AMBERff99 [**36**] et ne sont pas optimisés d'avantage. Deuxièmement, les charges partielles,

de l'équation 2.1.6, sont dérivées à partir du potentiel électrostatique avec contraintes ("restrained electrostatic potential" ou RESP) [**37**,**38**]. Dans cette méthode, la charge de chacun des types d'atome est optimisée afin de reproduire le potentiel électrostatique calculé à partir de calculs de mécanique quantique [**37**]. Des contraintes sont ajoutées au processus d'optimisation afin de fixer les charges des atomes lourds à une valeur désirée tout en optimisant les charges des atomes d'hydrogène [**37**]. Troisièmement, les paramètres des liens atomiques et des angles de valence, équations 2.1.2 et 2.1.3, sont déterminés avec une relation empirique à partir des valeurs à l'équilibre. Ces dernières sont tirées de simulations avec AMBERff99, de calculs de mécanique quantique ou via les structures expérimentales. Finalement, la dernière étape d'optimisation du potentiel est l'ajustement des paramètres des angles dièdres de l'équation 2.1.4. Pour ce faire, les amplitudes (V_n) de l'équation 2.1.4 furent optimisées à l'aide d'une recherche exhaustive [**39**] pour reproduire leurs profils énergétiques générés à l'aide de calculs de mécanique quantique. La paramétrisation de GAFF fut testée en mesurant les déviations structurelles sur 74 petites molécules après une minimisation de l'énergie. GAFF mena à un RMSD moyen de 0.26 Å comparativement aux structures expérimentales.

2.1.1.4. Solvant. En plus des protéines et des ligands, une des molécules les plus cruciales pour les processus biologiques, et donc pour nos modélisations, est l'eau. En effet, la majorité des processus moléculaires de la biologie se fait en milieu aqueux. L'eau joue un rôle crucial dans l'effet hydrophobe ("hydrophobic collapse"), phénomène fondamental pour le repliement des protéines, dans l'écrantage des charges, etc. [**28**]. Les orbitales électroniques de la molécule d'eau mènent à une forme tétraédrique où deux orbitales sont occupées par les atomes d'hydrogène et deux autres sont occupées par des doublets non-liants [**1**, **28**]. Cette séparation des charges entre les hydrogènes, chargés positivement(accepteur), et les deux doublets non-liants, chargés négativement (donneur), mène à l'apparition d'un dipôle. Ce dipôle permet la formation de ponts-H par les molécules d'eau, une interaction cruciale pour de nombreux phénomènes biologiques.

Au niveau numérique, un des modèles d'eau les plus populaires est TIP3P, développé au début des années 1980 par le groupe de recherche de William L. Jorgensen [**40**]. Il s'agit d'un modèle à géométrie fixe composé de trois sites qui sont positionnés au niveau de l'atome d'oxygène et des deux atomes d'hydrogène. La géométrie du modèle est présentée à la Figure 2.3. TIP3P est caractérisé par les paramètres suivants: l'angle entre les atomes H-O-H est de 104.52°, la longueur du lien atomique O-H est de 0.9572 Å, la charge de l'oxygène est de -0.834e et la charge des hydrogènes est de 0.417e. Pour ce qui est du potentiel, TIP3P est composé d'un terme attractif/répulsif, identique à l'équation 2.1.5, et d'un terme électrostatique, identique à l'équation 2.1.6. La paramétrisation de TIP3P fut réalisée pour tenter de reproduire divers facteurs thermodynamiques (chaleur spécifique à

pression constante, densité, énergie de vaporisation) et structurels (fonction de distribution radiale) à 25° et 1 atm lors de simulation de Monte-Carlo [**40**, **41**].

De multiples autres modèles numériques pour l'eau ont aussi été développés avec le temps avec chacune des méthodologies de paramétrisation et des géométries distinctives. Par exemple, TIP4P [**40**] est un modèle à quatre sites et TIP5P [**42**] est un modèle à cinq sites. Le choix du modèle d'eau approprié varie selon la méthodologie utilisée et le système d'intérêt. Dans notre étude sur les molécules thérapeutiques contre la COVID-19, nous avons arrêté notre choix sur TIP3P, car il s'agit du modèle d'eau utilisé pour la paramétrisation du potentiel AMBERff14 [**30**] que nous avons utilisé pour la description des protéines de notre système.

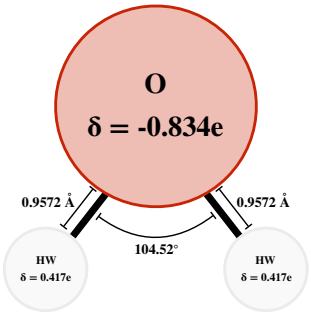


Fig. 2.3. Géométrie du modèle d'eau TIP3P. L'atome d'oxygène et les deux atomes d'hydrogène sont présentés respectivement en rouge et en gris. δ indique la charge partielle de chacun des atomes.

2.2. Dynamique moléculaire

Il est toujours important de rappeler que les propriétés macroscopiques d'un système, qu'elles soient d'équilibre ou dynamiques, sont données par la moyenne sur un ensemble de structures générées avec la bonne distribution statistique (poids de Boltzmann) [**28**, **43**]. La méthode de dynamique moléculaire est appropriée pour l'étude tant des propriétés d'équilibre que des propriétés dynamiques d'un système [**43**].

Le principe d'ergodicité stipule que les moyennes calculées sur un très grand nombre de systèmes sont identiques aux moyennes calculées sur un seul système après une très longue évolution temporelle.

La dynamique moléculaire fait évoluer un système dans le temps en solutionnant les équations du mouvement classique; les lois de Newton. Cette méthode est essentiellement décrite via les deux équations suivantes:

$$\frac{\mathrm{d}^2\mathbf{r}}{\mathrm{d}t^2} = \mathbf{M}^{-1}\mathbf{F} \tag{2.2.1}$$

$$\mathbf{F}_i = -\vec{\nabla}V(r_i) \tag{2.2.2}$$

La première équation, équation 2.2.1, est la seconde loi de Newton et décrit la relation entre les forces ($\mathbf{F}$) appliquées sur les atomes, leur masse ($\mathbf{M}$) et leur accélération ($\ddot{\mathbf{r}}$). La seconde équation, quant à elle, permet de dériver les forces ($\mathbf{F}$) à partir du potentiel V. Plus spécifiquement, les forces sont données par l'opposé du gradient du potentiel selon l'équation 2.2.2. En combinant ces deux équations, on peut faire évoluer le système dans le temps à l'aide d'un procédé en trois étapes:

(1) À partir des positions, $\mathbf{r}$, des atomes du système, on évalue le gradient du potentiel afin d'obtenir les forces, $\mathbf{F}$, agissant sur chacun d'eux.
(2) En utilisant la deuxième loi de Newton, on calcule la vitesse de chacun des atomes.
(3) Avec les vitesses, on calcule les nouvelles positions de chacun des atomes.

2.2.1. Intégration

Numériquement, l'intégration de la deuxième loi de Newton, équation 2.2.1, se fait à partir de l'algorithme saute-mouton (ou "Leap-Frog") dans lequel les positions au temps $t + \Delta t$ sont déterminées à partir des vitesses au temps $t + \frac{1}{2}\Delta t$. Les vitesses aux demi-pas de temps, $\mathbf{v}(t + \frac{1}{2}\Delta t)$ et les nouvelles positions, $\mathbf{r}(t + \Delta t)$, sont données par les équations suivantes:

$$\begin{aligned} \mathbf{v}(t + \frac{1}{2}\Delta t) &= \mathbf{v}(t - \frac{1}{2}\Delta t) + \frac{\Delta t}{m}\mathbf{F}(t) \\ \mathbf{r}(t + \Delta t) &= \mathbf{r}(t) + \mathbf{v}(t + \frac{1}{2}\Delta t)\Delta t \end{aligned} \tag{2.2.3}$$

2.2.2. Thermostat et Barostat

Les mesures expérimentales effectuées sur les systèmes moléculaires biologiques se font à des conditions bien précises de température, de pression, etc. Ces conditions expérimentales sont particulièrement importantes puisqu'elles définissent les ensembles statistiques qui sont le fondement de la mécanique statistique. Par exemple, un système dont le volume (V), et le nombre de particules (N) sont constants et qui est en contact avec un réservoir thermique, de telle sorte que la température (T), est constante, représente l'ensemble canonique ou NVT. Un système dont le nombre de particules (N), est fixe en contact avec un réservoir thermique, de telle sorte que la température (T) est constante, et un "réservoir d'espace", de telle sorte que la pression (P) est constante, représente l'ensemble isotherme-isobare ou NPT.

Pour pouvoir étudier ces différents ensembles statistiques à l'aide de méthodes numériques, il faut pouvoir reproduire l'effet du réservoir de chaleur et/ou du "réservoir d'espace" à l'intérieur de nos simulations. Pour ce faire, on doit modifier les équations du mouvement de la dynamique moléculaire afin d'inclure l'effet du thermostat et du barostat.

Dans notre étude sur les molécules thérapeutiques contre la COVID-19, nos simulations furent réalisées dans l'ensemble isotherme-isobare (NPT).

La température du système peut être déterminée à partir du théorème d'équipartition et de l'énergie cinétique du système, K, selon l'équation:

$$K = \frac{1}{2}\sum_{i}^{N} m_i v_i^2 = \frac{1}{2} N_{dl} k_B T$$

où N est le nombre d'atomes, m est la masse, v est la vitesse, T est la température et k_B est la constante de Boltzmann. Finalement, N_{dl} est le nombre de degrés de liberté et vaut $N_{dl} = 3N - N_c - N_{com}$ où N_c est le nombre de contraintes et N_{com} est le nombre de degrés de liberté associés au centre de masse et vaut trois.

Pour garder la température constante, nous avons utilisé le thermostat de Nosé-Hoover [**44**, **45**]. Cette méthode introduit un réservoir thermique à l'Hamiltonien du système ainsi qu'un terme de friction aux équations du mouvement. Ce terme de friction est proportionnel à la vitesse de chacune des particules et à un paramètre de friction, ξ. Ce paramètre de friction est une quantité dynamique associée à sa propre quantité de mouvement p_ξ et sa propre équation du mouvement. Avec ce réservoir thermique, la seconde loi de Newton, équation 2.2.1, devient:

$$\frac{\mathrm{d}^2 \mathbf{r}_i}{\mathrm{d}t^2} = \frac{\mathbf{F}_i}{m_i} - \frac{p_\xi}{Q}\frac{\mathrm{d}\mathbf{r}_i}{\mathrm{d}t}$$

avec Q une constante appelée le paramètre de masse du réservoir. L'équation du mouvement associée au réservoir thermique dépend de la différence entre la température actuelle, T, et la température de référence, T_0, selon l'équation suivante:

$$\frac{\mathrm{d}p_\xi}{\mathrm{d}t} = (T - T_0)$$

Le thermostat de Nosé-Hoover contrôle la température [**44**, **45**] à l'aide d'une relaxation oscillatoire [**43**]. Par contre, un des problèmes principaux avec le thermostat de Nosé-Hoover est qu'il peut mener à un comportement non-ergodique, c'est-à-dire que l'échantillonnage ne sera pas complet, même après un temps infiniment long. Pour prévenir ce phénomène, on utilise des chaînes de Nosé-Hoover [**46**]. Avec cette approche, le thermostat de Nosé-Hoover est contrôlé par un second thermostat de Nosé-Hoover qui est lui-même contrôlé par un troisième thermostat de Nosé-Hoover et ainsi de suite, de façon récursive. Dans la limite où le nombre de thermostats de la chaîne de Nosé-Hoover, N, tend vers l'infini ($N \to \infty$), alors on retrouve le comportement ergodique désiré. Par contre, cela complexifie passablement les

équations du mouvement qui deviennent donc:

$$\begin{aligned}
\frac{\mathrm{d}^2\vec{r_i}}{\mathrm{d}t^2} &= \frac{\vec{F_i}}{m_i} - \frac{p_{\xi 1}}{Q_1}\frac{\mathrm{d}\vec{r_i}}{\mathrm{d}t} \\
\frac{\mathrm{d}p_{\xi 1}}{\mathrm{d}t} &= (T - T_0) - p_{\xi 1}\frac{p_{\xi 2}}{Q_2} \\
\frac{\mathrm{d}p_{\xi i=1,\cdots,N}}{\mathrm{d}t} &= \left(\frac{p_{\xi i-1}^2}{Q_{i-1}} - k_B T\right) - p_{\xi i}\frac{p_{\xi i+1}}{Q_{i+1}} \\
\frac{\mathrm{d}p_{\xi N}}{\mathrm{d}t} &= \left(\frac{p_{\xi N-1}^2}{Q_{N-1}} - k_B T\right)
\end{aligned}$$

La pression, quant à elle, peut se calculer à partir du théorème du viriel et mène à l'équation:

$$\begin{aligned}
P &= \frac{2}{V}(K - \Xi) \\
\Xi &= \sum_i^N \mathbf{F}_i \cdot \mathbf{r}_i
\end{aligned}$$

où Ξ est le viriel, $\mathbf{F}$ est la somme des forces agissant sur la particule i, V est le volume et K est l'énergie cinétique.

Pour ce qui est de la pression, nous avons utilisé le barostat de Parrinello-Rahman [**47**]. Cet algorithme donne, en théorie, l'ensemble NPT exact. Pour ce faire, les dimensions de la boite, $\mathbf{b}$ sont contrôlées à l'aide de leur propre équation du mouvement:

$$\frac{\mathrm{d}\mathbf{b}^2}{\mathrm{d}t^2} = V\mathbf{W}^{-1}\mathbf{b}'^{-1}(\mathbf{P} - \mathbf{P_0}) \tag{2.2.4}$$

où V est le volume de la boîte, $\mathbf{W}$ est une matrice qui permet de contrôler l'amplitude du couplage, $\mathbf{P}$ est la pression actuelle et $\mathbf{P_0}$ est la pression de référence. Ainsi les équations du mouvement en présence du barostat deviennent:

$$\frac{\mathrm{d}^2\mathbf{r}_i}{\mathrm{d}t^2} = \frac{\mathbf{F_i}}{m_i} - \mathbf{M}\frac{\mathrm{d}\mathbf{r}_i}{\mathrm{d}t} \tag{2.2.5}$$

$$\mathbf{M} = \mathbf{b}^{-1}\left[\mathbf{b}\frac{\mathrm{d}\mathbf{b}'}{\mathrm{d}t} + \frac{\mathrm{d}\mathbf{b}}{t}\mathbf{b}'\right]\mathbf{b}'^{-1} \tag{2.2.6}$$

En combinant le thermostat et le barostat dans nos simulations, on obtient alors l'ensemble statistique NPT associé au système, ce qui correspond à l'ensemble statistique des méthodes expérimentales.

2.2.3. Contraintes

Un des principaux problèmes des simulations de dynamique moléculaire est relié au temps d'échantillonnage. En effet, les phénomènes biochimiques d'intérêt se produisent sur des échelles temporelles couvrant une quinzaine d'ordres de grandeur: de quelques dizaines de

femtosecondes pour la vibration des liens atomiques jusqu'à la micro-seconde/seconde pour le repliement des protéines [**1**]. Pour pouvoir intégrer correctement les équations du mouvement, via l'équation 2.2.3, il faut choisir un pas de temps suffisamment petit afin de capturer l'échelle de temps la plus courte, soit la période d'oscillation des liens atomiques. Ainsi, une liaison atomique avec un atome d'hydrogène a une période d'oscillation $\sim$ 10 fs. En utilisant une dizaine d'intégrations sur cette période, on obtient un pas de temps maximal de $\sim$ 1 fs, ce qui nécessite $\sim 10^{11}$ étapes d'intégration pour atteindre la micro-seconde. Afin d'accélérer la simulation, la longueur des liens atomiques est généralement contrainte à leur valeur d'équilibre. En contraignant les liens atomiques, la prochaine période d'oscillation à considérer est celle des angles atomiques à $\sim$ 20 fs [**1**, **43**]. Avec le même raisonnement, on peut donc augmenter le pas de temps à $\sim$ 2 fs et accélérer d'un facteur deux notre simulation. Divers algorithmes ont été développés pour contraindre les liens atomiques.

Dans notre étude, LINCS [**48**] ("LINear Constraint Solver"), fut utilisé pour restreindre tous les liens à leur longueur d'équilibre au niveau de la protéine et du ligand. LINCS est basé sur la méthode des multiplicateurs de Lagrange. Au système, on ajoute K contraintes, g_i, selon l'équation:

$$g_i(\mathbf{r}) = |\mathbf{r}_{i1} - \mathbf{r}_{i2}| - d_i = 0, i = 1, \cdots, K$$

où $r_{i/j}$ est la position de deux atomes connectés par un lien atomique et d est la longueur d'équilibre dudit lien. Avec ces contraintes, la seconde loi de Newton, donnée à l'équation 2.2.1, se ré-écrit comme:

$$-\mathbf{M}\frac{\mathrm{d}^2\mathbf{r}}{\mathrm{d}t^2} = \frac{\partial}{\partial \mathbf{r}}\left(\mathbf{V} - \lambda \cdot \mathbf{g}\right)$$

où $\mathbf{M}$ est une matrice diagonale $3N \times 3N$ contenant les masses des particules et λ sont les multiplicateurs de Lagrange.

Tel que mentionné précédemment, on peut résoudre cette équation du mouvement en utilisant la méthode d'intégration numérique du saute-mouton, décrite par l'équation 2.2.3. On obtient alors les nouvelles positions en présence des contraintes, sur les liens atomiques, $\mathbf{r}_{n+1}$, en fonction des positions sans contraintes $\mathbf{r}^{unc}_{n+1}$:

$$\mathbf{r}_{n+1} = \mathbf{r}^{unc}_{n+1} - \mathbf{M}^{-1}\mathbf{B}_n\left(\mathbf{B}_n\mathbf{M}^{-1}\mathbf{B}_n^{\mathbf{T}}\right)^{-1}\left(\mathbf{B}_n\mathbf{r}^{unc}_{n+1} - \mathbf{d}\right) \tag{2.2.7}$$

où $\mathbf{B}$ est une matrice $K \times 3N$ contenant les directions des contraintes: $B_{hi} = \frac{\partial g_h}{\partial r_i}$.

À cette étape, c'est la projection du nouveau lien p_i sur la direction de l'ancien lien qui est de la longueur d'équilibre. On doit maintenant corriger l'effet de la rotation pour fixer la longueur du lien elle-même à sa valeur d'équilibre. Pour ce faire, p_i est fixé selon:

$$p_i = \sqrt{2d_i^2 - l_i^2}$$

où l_i est la longueur du lien après la première étape de projection. En considérant tout ce qui fut mentionné précédemment, les positions finales r^*_{n+1} corrigées de telle sorte que les liens sont à leur longueur d'équilibre sont données par:

$$\mathbf{r}^*_{n+1} = (\mathbf{I} - \mathbf{T}_n\mathbf{B}_n)r_{n+1} + \mathbf{T}_n\mathbf{p}$$

où $\mathbf{I}$ est la matrice identité, $\mathbf{T} = \mathbf{M}^{-1}\mathbf{B}^T\left(\mathbf{B}\mathbf{M}^{-1}\mathbf{B}^T\right)^{-1}$ est une matrice $3N \times K$ qui transforme les coordonnées contraintes dans les coordonnées cartésiennes.

En résumé, l'algorithme contient trois étapes. La première est de calculer les nouvelles positions sans l'application de contraintes r^{unc}_{n+1}. La seconde consiste à fixer la projection du nouveau lien sur l'ancien lien à zéro. Finalement, la troisième applique la correction pour l'effet de rotation.

Pour ce qui est des molécules d'eau, qui forment généralement près de 80% du système simulé, les liens sont contraints avec l'algorithme SETTLE [**49**] qui fut spécialement développé pour ces molécules. Les molécules d'eau doivent rester rigides durant la simulation. En effet, le potentiel des molécules d'eau fut paramétrisé lors de simulations de Monte-Carlo où elles étaient rigides [**40**]. Pour ce faire, SETTLE représente les molécules d'eau comme un triangle où les sommets A, B et C correspondent respectivement à l'atome d'oxygène et aux deux atomes d'hydrogène et les trois arêtes correspondent aux deux liens atomiques H-O et à la distance H-H. SETTLE détermine analytiquement les trois angles ϕ, ψ et θ permettant de retrouver la géométrie idéale à partir de la géométrie obtenue après une intégration des équations du mouvement.

2.3. Minimisation de l'énergie

Les algorithmes de minimisation permettent le réarrangement local des atomes du système afin d'atteindre un minimum énergétique, c'est-à-dire, une configuration où les forces sont nulles. Dans cette section, je discuterai des deux principales méthodes de minimisation de l'énergie utilisées dans notre article; l'algorithme du col ("steepest descent") et l'algorithme du gradient conjugué ("conjugate gradient") [**43**]. On retrouve les mêmes deux éléments clés dans ces méthodes [**1**]; (1) la détermination de la direction de recherche (d) et (2) une recherche linéaire afin de déterminer l'amplitude du déplacement (h). Elles prennent la forme de l'équation suivante:

$$r_{n+1} = r_n + d_n h_n$$

Pour l'algorithme du col, la direction de recherche est simplement déterminée par la direction des forces, correspondant au gradient du potentiel selon l'équation 2.2.2, avec:

$$d_n = \frac{\mathbf{F}_n}{\max\left(|\mathbf{F}_n|\right)}$$

Pour ce qui est de la méthode du gradient conjugué, la direction de recherche est déterminée à partir de l'expression récursive suivante:

$$\mathbf{d}_{k+1} = -\mathbf{F}_{k+1} + \beta_{k+1}\mathbf{d}_k$$

de telle sorte que les nouvelles directions de recherche soient orthogonales aux précédentes. Cela permet d'éviter de chercher de nombreuses fois dans la même direction. Plusieurs expressions existent pour β_{k+1}. Dans la méthode de Polak-Ribière, β est déterminée par l'expression suivante:

$$\beta_n^{PR} = \frac{\mathbf{F}_n^T\left(\mathbf{F}_n - \mathbf{F}_{n-1}\right)}{\mathbf{F}_{n-1}^T\mathbf{F}_{n-1}}$$

Une fois la direction de recherche fixée, il faut maintenant déterminer l'amplitude du déplacement, h, à l'aide d'une recherche linéaire. La valeur initiale de h est déterminée arbitrairement, mais elle est ensuite mise à jour régulièrement en fonction de l'énergie des nouvelles positions $U(\mathbf{r}_{n+1})$ par rapport à l'énergie des anciennes positions $U(\mathbf{r}_n)$. Si $U(\mathbf{r}_{n+1}) < U(\mathbf{r}_n)$, les nouvelles positions sont acceptées et l'amplitude du déplacement est augmentée $h_{n+1} = 1.2h_n$. Si $U(\mathbf{r}_{n+1}) \geq U(\mathbf{r}_n)$, les nouvelles positions sont refusées et l'amplitude du déplacement est diminuée $h_n = 0.2h_n$.

Ce processus est répété jusqu'à l'obtention d'un critère seuil sur l'amplitude des forces.

2.4. Interactions protéines/ligands

Certaines petites molécules ont un potentiel thérapeutique et peuvent servir aux traitements de diverses maladies. Pour ce faire, elles doivent pouvoir se lier à leurs cibles et mener à une réponse physiologique sécuritaire et mesurable [**50**]. L'identification de nouveaux médicaments est un processus lent, coûteux et complexe allant de l'identification initiale d'une potentielle molécule thérapeutique jusqu'aux essais cliniques [**50**]. L'identification de molécules thérapeutiques se base sur des procédés expérimentaux qui sont souvent lents et coûteux [**50, 51**]. Pour ces raisons, les méthodes numériques présentent un complément intéressant aux méthodes expérimentales pour étudier les liaisons récepteurs/ligands.

Deux facteurs sont particulièrement importants à considérer pour le développement de molécules thérapeutiques; (1) les sites de liaison récepteurs/ligands et (2) les affinités de liaison récepteurs/ligands. Différentes méthodes numériques furent développées pour étudier ces problématiques. Chacune d'entre elles se positionne différemment au niveau du plan rapidité-vs-complexité de la Figure 0.1 et présente différents avantages et inconvénients.

Dans le coin supérieur droit, du plan rapidité-vs-complexité, rapide et simple, de la Figure 0.1, on retrouve des méthodes comme l'amarrage moléculaire ("Molecular Docking"). L'amarrage moléculaire permet d'estimer rapidement les sites et les affinités de liaison récepteur/ligand ce qui en fait une des méthodes numériques les plus populaires dans le

développement thérapeutique. À l'autre extrémité, coin inférieur gauche, du plan rapidité/complexité, lent et complexe, (Figure 0.1), on retrouve les méthodes de perturbations alchimiques ("Alchemical Perturbation Methods") qui permettent, en principe, une estimation très précise de l'affinité de liaison récepteur/ligand, mais qui requièrent un échantillonnage exhaustif des conformations récepteur/ligand en plus de nombreux intermédiaires, souvent non-physiques [**51**, **52**]. Entre les deux, on retrouve les méthodes point-finaux ("end-point"), comme la méthode combinée de mécanique moléculaire, poisson-boltzmann et surface accessible au solvant (MM/PBSA). Ces méthodes sont plus précises que le simple amarrage moléculaire et sont plus rapides que les calculs de perturbations alchimiques [**51**, **52**].

Dans les sections suivantes, je discuterai avec plus de détails les méthodes d'amarrage moléculaire et de MM-PBSA qui furent utilisées dans notre étude sur les molécules thérapeutiques contre la COVID-19.

2.4.1. Amarrage moléculaire

Comme mentionné précédemment, l'amarrage moléculaire est une méthode simple et rapide permettant de déterminer les sites et les affinités de liaison récepteurs/ligands [**53**,**54**]. L'amarrage moléculaire est généralement composé de deux éléments clés. Le premier élément est une fonction de score qui permet de discriminer les bonnes liaisons des mauvaises et dont le minimum est associé à la meilleure liaison possible. Le second élément est un algorithme d'échantillonnage qui permet d'explorer l'image de la fonction de score afin d'en identifier le minimum. De nombreuses méthodes d'amarrage moléculaire furent développées avec chacune leurs spécificités au niveau de la fonction de score et de l'algorithme d'échantillonnage [**54**]. Dans cette section, je ne vais m'attarder qu'à une seule d'entre elles, appelée AutoDock VINA [**55**], qui fut utilisée dans notre étude des molécules thérapeutiques contre la COVID-19.

Intéressons-nous premièrement à la fonction de score d'Autodock VINA. Il s'agit un potentiel empirique/"knowlegde-base" permettant d'estimer l'affinité de liaison récepteur/ligand. Pour chacune des paires de types d'atome, ij, le potentiel est décrit par les

équations suivantes:

$$\text{gauss1}(d_{ij}) = w_{\text{gauss1}} \cdot \exp\left[-\left(\frac{d_{ij}}{0.5\text{Å}}\right)^2\right]$$

$$\text{gauss2}(d_{ij}) = w_{\text{gauss2}} \cdot \exp\left[-\left(\frac{d_{ij} - 3\text{Å}}{2\text{Å}}\right)^2\right]$$

$$\text{répul}(d_{ij}) = \begin{cases} w_{\text{répul}} \cdot d_{ij}^2 & \text{, if } d_{ij} < 0 \\ 0 & \text{, if } d_{ij} \geq 0 \end{cases}$$

$$\text{hydro}(d_{ij}) = \begin{cases} w_{\text{hydro}} & \text{, if } d_{ij} < 0.5\text{Å} \\ w_{\text{hydro}} \cdot \left(-d_{ij} + 1.5\text{Å}\right) & \text{, if } 0.5\text{Å} \leq d_{ij} \leq 1.5\text{Å} \\ 0 & \text{, if } d_{ij} > 1.5\text{Å} \end{cases}$$

$$\text{HB}(d_{ij}) = \begin{cases} w_{HB} & \text{, if } d_{ij} < -0.7\text{Å} \\ w_{HB} \cdot \left(-\frac{10}{7} d_{ij}\right) & \text{, if } -0.7\text{Å} \leq d_{ij} \leq 0\text{Å} \\ 0 & \text{, if } d_{ij} > 0\text{Å} \end{cases}$$

Les trois premiers termes ("gauss1", "gauss2" et "répul") décrivent les interactions attractives/répulsives et ils sont utilisés pour toutes les paires de types d'atome. Le terme "hydro" est un terme additionnel entre les paires de types d'atome hydrophobes. Le terme "HB" quant à lui décrit la formation de ponts hydrogènes et il n'est utilisé qu'entre les paires de types d'atome où cette possibilité d'interaction existe. La distance entre deux atomes, ij est définie comme $d_{ij} = r_{ij} - R_i - R_j$ avec R_x le rayon de van der Waals associé au type d'atome x. Finalement, les paramètres w_X sont différents poids associés à chacun des termes. Ces poids ont été optimisés à partir de mesures expérimentales d'affinités et des préférences conformationnelles de différents complexes récepteur/ligand [**55**]. La valeur de ces poids pour chacune des interactions est donnée au Tableau 2.1.

Potentiel	Poids (w)
gauss1	-0.0356
gauss2	-0.00516
répulsion	0.840
hydro	-0.0351
HB	-0.587
N_{rot}	0.0585

Tableau 2.1. Poids d'AutoDock VINA. La colonne de gauche présente le terme du potentiel et la colonne de droite présente le poids associé à celui-ci.

Passons maintenant à l'algorithme d'échantillonnage d'Autodock VINA. L'objectif de cet algorithme est d'explorer différents positionnements du ligand au niveau du récepteur et de

déterminer la valeur de la fonction de score associée à ceux-ci. Pour ce faire, Autodock VINA utilise un algorithme itératif nommé "Iterated Local Search global optimizer". À chacune des étapes de cet algorithme, un changement conformationnel est réalisé, suivi d'une minimisation locale. Plus spécifiquement, l'algorithme d'échantillonnage d'Autodock VINA est basé sur un système de coordonnées internes décrit dans l'article d'Abagyan *et coll.* [**56**]. Brièvement, autour d'un axe aléatoire passant par le centre de masse du ligand, une rotation suivie d'une translation est réalisée sur trois atomes de celui-ci. Suivant ce mouvement, le ligand complet est ensuite reconstruit en fonction des coordonnées internes. À cette nouvelle position, la fonction de score est minimisée à l'aide de la méthode de Broyden-Fletcher-Goldfarb-Shanno (BFGS) pour permettre une réorganisation locale des atomes des chaînes latérales du récepteur autour du ligand. Finalement, chacune des étapes est acceptée ou refusée selon un critère de Metropolis.

Autodock VINA retourne les meilleures structures (meilleures valeurs de la fonction de score explorées) du complexe récepteur/ligand. Les structures du complexe sont ensuite classées selon l'estimation de l'énergie libre de liaison. Autodock VINA dérive l'énergie libre de liaison récepteur/ligand à partir de la partie intermoléculaire de la fonction de score et le degré de flexibilité du ligand selon l'équation:

$$s_i = \frac{c_{inter}}{1 + w_{N_{rot}} N_{rot}}$$

où N_{rot} est le nombre de liens "flexibles" du ligand.

Tel que mentionné précédemment, l'amarrage moléculaire permet rapidement d'obtenir une estimation des sites de liaison du ligand sur un récepteur et ainsi permettre l'estimation au niveau de nombreux ligands en un temps relativement court. Par contre, cette vitesse d'exécution est obtenue à partir de simplifications au niveau de la méthode. Tout particulièrement, la fonction de score d'Autodock VINA reste assez imprécise. Elle permet généralement de discriminer si une molécule est un ligand ou un non-ligand, mais ne peut discriminer entre différents ligands dont les affinités de liaison sont comparables [**51**, **52**]. Pour obtenir une estimation plus précise, il faut se tourner vers des méthodes plus complexes et donc plus lentes.

2.4.2. MM/PBSA

MM/PBSA est une méthode d'estimation de l'énergie libre offrant un compromis entre l'amarrage moléculaire et les calculs de perturbations alchimiques; elle est plus précise que les fonctions de score du premier et elle requiert moins de ressources computationnelles que les seconds [**51**, **52**]. Cette méthode combine trois éléments principaux, la mécanique moléculaire, la résolution des équations de Poisson-Boltzmann et l'estimation de la surface accessible au solvant, d'où le nom MM/PBSA pour *M*olecular *M*echanics *P*oisson-*B*oltzmann

*S*urface *A*rea. La méthode MM/PBSA est basée sur un échantillonnage des états finaux (le complexe en solution et possiblement le récepteur seul en solution et le ligand seul en solution), contrairement aux méthodes de perturbations alchimiques qui demandent en plus un échantillonnage important de nombreux états intermédiaires.

L'énergie libre de liaison, ΔG_{bind}, entre un récepteur et un ligand est donnée par l'équation suivante

$$\Delta G_{bind} = \langle G_{RL} \rangle - \langle G_R \rangle - \langle G_L \rangle \tag{2.4.1}$$

où R indique le récepteur, L indique le ligand et $\langle\rangle$ dénote une moyenne sur un ensemble d'états.

Dans la méthode de MM/PBSA, l'énergie libre de chacun des états (RL, R, L) est estimée via l'équation:

$$G_x = \langle E_{mm} \rangle - TS + \langle G_{solv} \rangle \tag{2.4.2}$$

Cette équation comporte trois composantes; (1) L'évaluation de l'énergie interne (E_{mm}), (2) l'évaluation de l'énergie libre de solvatation (G_{solv}) et (3) l'évaluation de l'entropie (S). La forme de chacun de ces termes sera explicitée dans ce qui suit.

L'**énergie interne** du système s'écrit:

$$E_{mm} = E_{\text{lié}} + E_{\text{élec}} + E_{\text{vdw}} \tag{2.4.3}$$

où $E_{\text{lié}}$ est le potentiel des interactions liées (lien, angle, angle dièdre), $E_{\text{élec}}$ est l'énergie électrostatique et E_{vdw} est l'énergie des interactions attractives/répulsives. Tous ces termes sont calculés directement à partir du champ de force de mécanique moléculaire (MM) avec laquelle la simulation fut réalisée.

L'**énergie libre de solvatation** (G_{solv}) est divisée en deux parties, une polaire et une apolaire via l'équation:

$$G_{solv} = G_{polaire} + G_{apolaire} \tag{2.4.4}$$

La **partie polaire** de l'énergie libre de solvatation est estimée à l'aide d'un solvant implicite déterminé en solutionnant les équations de Poisson-Boltzmann (PB). Le modèle de Poisson-Boltzmann est dérivé avec l'hypothèse d'un champ moyen ("Potential of Mean Force"), et s'intéresse à la distribution des charges positives et des charges négatives présentes dans le solvant (diélectrique) autour d'une charge spécifique q_0.

Le potentiel électrostatique en tout point $\phi(\mathbf{r})$ d'une certaine distribution de charges $\rho(\mathbf{r})$ est donné par l'équation de Poisson:

$$\nabla^2 \phi(\mathbf{r}) = -\frac{\rho(\mathbf{r})}{\epsilon \epsilon_0} \tag{2.4.5}$$

La complexité du problème revient à déterminer la distribution des charges $\rho(\mathbf{r})$ autour de q_0. Dans le cas où l'on assume que les charges sont en solution, de telle sorte qu'elles peuvent se réarranger librement autour de q_0, alors la distribution des charges dépend uniquement de r (symétrie sphérique). La distribution moyenne des charges $\langle r(r)\rangle$ autour de q_0 est donnée par la fonction de distribution radiale $\mathrm{rdf}(r)$ de chacun des types d'ions présents. Si on considère deux types d'ions, un de charge $+e$ et un de charge $-e$, respectivement à concentration égale en solution, $n = N_-/V = N_+/V$, alors on obtient:

$$\langle \rho_0(r)\rangle = ne \cdot [\mathrm{rdf}_+(r) - \mathrm{rdf}_-(r)] \tag{2.4.6}$$

En utilisant la relation entre la fonction de distribution radiale et le PMF, décrite à la section 5.1.2, on peut alors réécrire l'équation précédente par:

$$\langle \rho_0(r)\rangle = ne \cdot \left[e^{-\beta \mathrm{PMF}_+(r)} - e^{-\beta \mathrm{PMF}_-(r)}\right]$$

En utilisant cette distribution de charges dans l'équation de Poisson, équation 2.4.5, alors on obtient:

$$\nabla^2\psi_0(r) = -\frac{ne}{\epsilon\epsilon_0} \cdot \left[e^{-\beta \mathrm{PMF}_+(r)} - e^{-\beta \mathrm{PMF}_-(r)}\right] \tag{2.4.7}$$

Si l'équation 2.4.7 est exacte, avec un certain nombre de suppositions, elle est par contre inutilisable puisque la forme des PMFs est indéterminée. Pour obtenir l'équation de Poisson-Boltzmann, on fait l'hypothèse que les PMFs dépendent uniquement du potentiel électrostatique ψ_0 et d'aucun autre type d'interaction (comme les interactions attractives/répulsives) [**28**], de telle sorte qu'il peut s'écrire sous la forme:

$$\mathrm{PMF}_\pm(r) \simeq \pm e\psi_0(r)$$

On obtient alors l'équation de Poisson-Boltzmann:

$$\nabla^2\psi_0(r) = -\frac{ne}{\epsilon\epsilon_0} \cdot \left[e^{-e\beta\psi_0(r)} - e^{e\beta\psi_0(r)}\right] \tag{2.4.8}$$

$$= \frac{2ne}{\epsilon\epsilon_0} \sinh\left(e\beta\psi_0(r)\right) \tag{2.4.9}$$

Dans le cas où $\psi_0(r) << k_BT$ alors $\sinh x \approx x$ et l'équation de Poisson-Boltzmann devient:

$$\nabla^2\psi_0(r) \simeq \kappa^2\psi_0(r)$$

Il s'agit de l'équation de Debye-Hückel où κ est la longueur de Debye et est donnée par l'expression: $\frac{2ne^2\beta}{\epsilon\epsilon_0}$.

La **partie apolaire** de l'énergie libre de solvatation est approximée proportionnelle à la surface accessible au solvant (SASA):

$$G_{apolaire} = \gamma A + b \tag{2.4.10}$$

où A est la SASA, γ est un coefficient associé à la tension de surface et b est un paramètre ajustable.

Finalement, la dernière partie importante de l'équation 2.4.2 est l'**entropie** (S). Une panoplie de techniques peut être utilisée pour estimer la contribution de l'entropie à l'énergie libre de liaison, dont une des plus populaires est probablement l'analyse des modes normaux ("normal mode analysis") [**52**]. Par contre, l'estimation de l'entropie est généralement le facteur limitant de la vitesse de calcul de l'équation 2.4.2 en plus de présenter de fortes fluctuations ce qui rend l'estimation de l'énergie libre de liaison peu précise [**52**, **57**]. Pour ces raisons, l'entropie est généralement négligée.

En résumé, la méthode de MM/PBSA estime l'énergie libre de liaison entre un ligand et son récepteur via l'équation 2.4.2 comportant quatre composantes; (1) l'énergie interne estimée à partir du potentiel de mécanique moléculaire, (2) la partie polaire de l'énergie libre de solvatation estimée en solutionnant l'équation de Poisson-Boltzmann, (3) la partie apolaire de l'énergie libre de solvatation estimée à l'aide de la surface accessible au solvant et (4) l'entropie, qui est généralement négligée.

Avec MM/PBSA, on peut maintenant évaluer les énergies libres du complexe récepteur/ligand, du récepteur et du ligand de l'équation 2.4.1. Le seul élément manquant est l'ensemble sur lequel les moyennes ($\langle\rangle$) de cette équation doivent être évaluées. Rigoureusement, ces moyennes ($\langle\rangle$) doivent être calculées de trois simulations (ensembles) distinctes: une simulation du complexe en solution pour G_{RL}, une simulation du récepteur seul en solution pour G_R et une simulation du ligand seul en solution pour G_L. Il s'agit de la méthode 3T-MM/PBSA et l'équation 2.4.1 devient

$$\Delta G_{bind} = \langle G_{RL} \rangle_{RL} - \langle G_R \rangle_R - \langle G_L \rangle_L$$

Par contre, en pratique, on utilise plutôt les résultats d'une seule simulation sur le complexe (RL) en solution (1T-MM/PBSA) et l'équation 2.4.1 devient:

$$\Delta G_{bind} = \langle G_{RL} - G_R - G_L \rangle_{RL}$$

De cette seule simulation, on extrait les atomes désirés soit les atomes du complexe, les atomes du récepteur et les atomes du ligand pour estimer respectivement G_{RL}, G_R et G_L. Cette simplification est valide si on assume que l'échantillonnage du récepteur et du ligand en complexe est similaire au récepteur et ligand seul en solution [**52**]. Comme une seule simulation du complexe est requise, 1T-MM/PBSA est plus rapide que 3T-MM/PBSA. De plus, et ce malgré le fait qu'elle néglige les changements conformationnels de la protéine et du ligand lors de l'association, la méthode 1T-MM/PBSA, est généralement plus précise que la méthode 3T-MM/PBSA, entre autres, parce que les contributions $E_{\text{lié}}$ de l'énergie interne (équation 2.4.3) s'annulent [**51**].

Malgré tout, la méthode MM-PBSA présente un nombre important de limitations. Entre autres, l'estimation de l'énergie interne E_{MM} contient toutes les approximations entrant dans le champ de force utilisé. Tout particulièrement, la polarisation n'est généralement pas considérée et $E_{\text{élec}}$ est sous-estimée [**51**]. De plus, il y a incohérence au niveau du solvant entre les simulations et les calculs de MM/PBSA; les premières sont faites avec un solvant explicite tandis que les seconds sont faits avec un solvant implicite [**51**]. L'utilisation d'un solvant implicite est requise pour les calculs de MM/PBSA afin d'éviter les larges fluctuations de l'énergie mesurées avec un solvant explicite [**52**].

Pour terminer sur une note plus concrète, dans notre étude, nous avons utilisé l'outil g_MMPBSA [**57**] pour calculer l'énergie libre de liaison à partir des trajectoires obtenues avec GROMACS [**58**]. Cet outil fait appel à APBS [**59**] pour solutionner l'équation de Poisson-Boltzmann à l'aide d'une méthode itérative.

2.4.3. Méthodes expérimentales

La partie numérique de notre étude sur des molécules thérapeutiques contre la COVID-19 fut complémentée grâce à deux méthodes expérimentales: des essais immuno-enzymatiques de liaison de ligands (ELISA) et des expériences de résonance plasmon surface (SPR). Dans cette section, je décrirai l'essentiel des bases théoriques de chacune des méthodes qui permettront de comprendre l'analyse présentée dans notre étude.

2.4.3.1. Essais de liaison de ligand. Les essais de liaison de ligand ("Ligand Binding Assay") sont des méthodes expérimentales qui visent la détection et la quantification de la liaison entre un ligand et son récepteur. Pour ce faire, la méthode utilise la mesure de l'absorbance de l'échantillon. L'absorbance est la capacité d'un milieu à absorber la lumière. En effet, lorsque la lumière passe à travers un milieu, l'intensité de celle-ci à sa sortie se retrouve diminuée à cause des différents phénomènes physiques se passant dans celui-ci; absorbance, réflexion, diffraction, etc. Mathématiquement, l'absorbance est définie par le ratio entre l'intensité de la lumière entrante (I_0) et sortante (I) d'un milieu selon l'équation:

$$A = \log\left(\frac{I_0}{I}\right)$$

La loi de Beer-Lambert permet de relier les mesures d'absorbance à la concentration (c) d'un échantillon selon:

$$A = \epsilon \cdot c \cdot l \tag{2.4.11}$$

où ϵ est l'absorptivité molaire et l est le chemin optique.

Les détails du protocole ELISA utilisé dans notre étude sont présentés à la Figure 2.4 et ci-dessous:

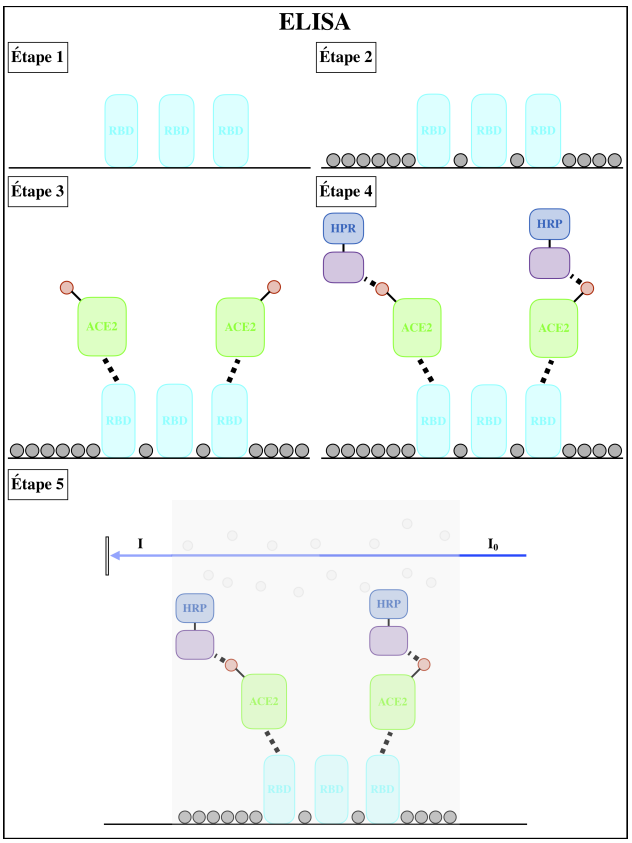


Fig. 2.4. Protocole d'ELISA utilisé dans l'étude COVID. Les molécules de RBD, les bloqueurs, ACE2, la biotine, la streptavidine, la peroxydase de raifort et les substrats chromogènes sont présentés respectivement en turquoise, noir, vert, rouge, mauve, bleu et gris. L'absorbance est calculée à 450 nm grâce au ratio de l'intensité lumineuse initiale (I_0) et l'intensité lumineuse finale (I).

(1) Le fond des microplaques ELISA est recouvert du RBD de la protéine *Spike* de SARS-CoV2.

(2) Après un nettoyage, on ajoute une solution contenant un agent bloquant qui permet de couvrir la surface non-recouverte par le RBD du fond des microplaques et ainsi éviter que les autres éléments ne se lient au fond de celles-ci.

(3) On ajoute ensuite les molécules de ACE2 préalablement liées de façon covalente avec une molécule de biotine. Les molécules de ACE2 vont se lier aux molécules de RBD fixées au fond des microplaques.

(4) Après un nouveau nettoyage, on ajoute une solution contenant des molécules de streptavidine liées à de la peroxydase de raifort (HRP). La steptavidine se lie avec une haute affinité et une haute spécificité avec la biotine liée à ACE2.

(5) Après un nouveau nettoyage, on ajoute une solution contenant des substrats chromogènes. La peroxydase de raifort convertit ces molécules chromogènes en molécules colorées.

(6) À l'aide d'un spectrophotomètre, on mesure l'absorbance de l'échantillon à 450 nm. Tel que mentionné précédemment, les mesures d'absorbance sont liées à la concentration de l'échantillon via l'équation 2.4.11

En d'autres mots, si les molécules de ACE2/biotine se lient aux molécules de RBD fixées au fond des microplaques, alors on observera un changement de couleur de notre échantillon qui se traduira par un changement au niveau de son absorbance. Pour tester l'impact de nos molécules thérapeutiques, on ajoute entre l'étape deux et trois une solution contenant nos ligands. Si ceux-ci préviennent la liaison entre ACE2 et RBD, alors on observera une diminution de la présence de molécules colorées qui se traduira par une diminution de l'absorbance.

2.4.3.2. Biosenseur par résonance des plasmons de surface. Les biosenseurs par résonance de plasmon de surface (SPR) sont des méthodes très populaires permettant d'étudier la dynamique de liaison récepteur/ligand en temps réel [**60**]. Ces méthodes sont basées sur le phénomène de résonance des plasmons de surface; un phénomène lumière/matière permettant de mesurer les changements d'indice de réfraction à proximité d'une mince surface métallique. Plus spécifiquement, la résonance des plasmons de surface se produit lorsque de la lumière est réfléchie par une mince surface métallique. Alors, une partie de la lumière incidente interagit avec les électrons délocalisés de la couche métallique (plasmons), et l'intensité lumineuse réfléchie s'en trouve diminuée [**60**]. Sans entrer dans les détails, l'orientation de la lumière incidente est cruciale pour l'observation de ce phénomène [**60**] et l'orientation de résonnance dépend fortement des conditions à proximité de cette interface.

Ce phénomène permet de développer des biosenseurs SPR particulièrement adaptés pour étudier les interactions récepteurs/ligands. Un schéma de biosenseur SPR typique est présenté à la Figure 2.5. À proximité de la couche métallique, la liaison récepteurs/ligands mène à un changement de l'indice de réfraction, qui cause un changement de l'angle de résonnance avec les plasmons. Ainsi, en mesurant l'intensité de la lumière réfléchie par la couche métallique en fonction de l'angle d'incidence de la lumière, on peut détecter en temps-réel les changements de l'angle de résonnance qui sont associés aux interactions récepteurs/ligands.

Les biosenseurs SPR permettent d'étudier la dynamique sous-tendant les interactions récepteurs/ligands. La dynamique d'association récepteurs/ligands est associée à l'équilibre suivant [**28**, **60**]:

$$R + L \underset{k_d}{\overset{k_a}{\rightleftharpoons}} RL$$

où R indique le récepteur, L indique le ligand, k_a est la constante d'association et k_d la constante de dissociation. Avec cet équilibre, on peut écrire la concentration du complexe,

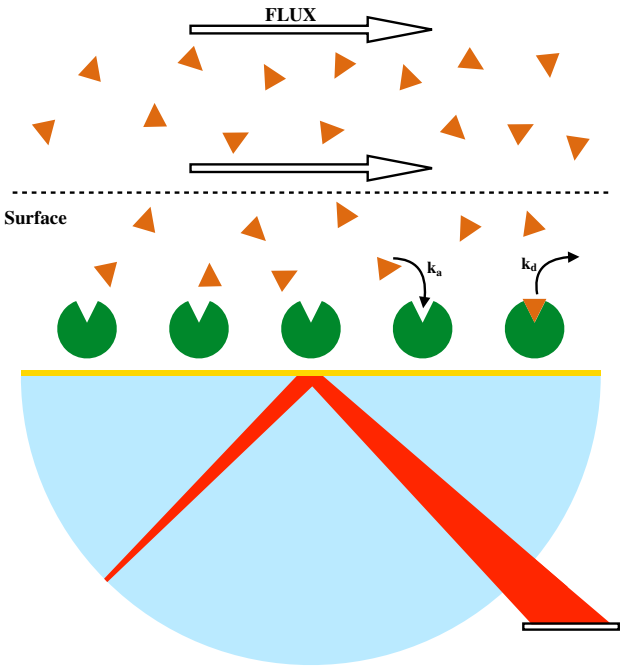


Fig. 2.5. Schéma d'un montage SPR typique. La surface métallique, cruciale pour la résonance des plasmons, est présentée en couleur or et les récepteurs liés à celle-ci sont présentés en vert. Le ligand est présenté en orange.

$[RL]$, en fonction du temps avec l'équation différentielle suivante:

$$\frac{\mathrm{d}[RL]}{\mathrm{d}t} = k_a[R][L] - k_d[RL] \tag{2.4.12}$$

À proximité de la surface métallique, la quantité de récepteurs libres (R_l) et donc disponibles pour les interactions avec le ligand est limitée. Sa concentration est donnée par l'équation suivante:

$$[R_l] = [R_{max}L] - [RL] \tag{2.4.13}$$

où $[R_{max}L]$ est la concentration maximale possible du complexe (à saturation).

On peut alors combiner l'équation 2.4.12 et l'équation 2.4.13 et obtenir l'équation suivante pour la dynamique récepteur/ligand à proximité de la surface métallique:

$$\frac{\mathrm{d}[RL]}{\mathrm{d}t} = k_a[L]\left([R_{max}L] - [RL]\right) - k_d[RL]$$

Finalement, le biosenseur SPR ne mesure pas directement la concentration du complexe, mais bien la réponse au niveau de l'angle de résonance causée par les changements conformationnels à la surface métallique. En assumant que la réponse, P, est proportionnelle à la concentration du complexe $[RL]$, alors l'équation précédente devient:

$$\frac{\mathrm{d}P}{\mathrm{d}t} = k_a[L]P_{max} - P\left(k_a[L] + k_d\right) \tag{2.4.14}$$

où P_{max} est la réponse maximale (à saturation). Il s'agit de l'équation gouvernant les mesures par biosenseur SPR. Sa résolution est essentielle pour comprendre les mesures réalisées par ces biosenseurs.

Une expérience classique avec biosenseur SPR se sépare habituellement en trois étapes; la phase d'association, la phase d'équilibre et la phase de dissociation.

Dans la phase d'association, on commence l'expérience en injectant en continu une solution contenant les ligands d'intérêts qui pourront interagir avec les récepteurs situés à proximité de la surface métallique. On doit donc résoudre l'équation 2.4.14, une équation différentielle linéaire de premier ordre. Pour ce faire, on utilise la méthode classique en posant $P = uv$ et en utilisant la règle de dérivation d'un produit $\mathrm{d}R/\mathrm{d}t = u\mathrm{d}v/\mathrm{d}t + \mathrm{d}u/\mathrm{d}tv$.

On obtient alors que la réponse en fonction du temps est donnée par l'équation suivante:

$$P = P_0 \left(1 - \exp\left(-(k_d + k_a[L])t\right)\right) \tag{2.4.15}$$

où $P_0 = \frac{k_a[L]P_{max}}{k_a[L]+k_d}$ est la réponse lorsque t=0.

En attendant un temps suffisant, le système atteint l'équilibre et le nombre de complexes demeure constant (autant d'association que de dissociation). En reprenant l'équation 2.4.14, on peut alors déterminer la réponse à l'équilibre (P_{eq}) qui est constant:

$$P_{eq} = \frac{k_a[L]P_{max}}{k_a[L] + k_d} \tag{2.4.16}$$

Finalement, la dernière étape du protocole est la dissociation. Dans celle-ci, on cesse l'injection de la solution contenant le ligand pour une solution tampon, permettant d'éliminer les ligands dissociés. On peut donc résoudre l'équation 2.4.14 avec $[L] = 0$ et on obtient:

$$P = P_0 \exp\left(-k_d t\right) \tag{2.4.17}$$

En résumé, la réponse SPR associée aux interactions récepteurs/ligands est décrite par les équations 2.4.15, 2.4.16 et 2.4.17. Les phases d'association et de dissociation sont décrites par des exponentielles tandis que la phase d'équilibre est décrite par une constante. Un profil de réponse classique d'un biosenseur SPR pour étudier les interactions récepteurs/ligands est présenté à la Figure 2.6. En faisant varier la concentration du ligand ($[L]$), on peut alors obtenir une série de courbes de réponse et ajuster les valeurs des paramètres des équations 2.4.15, 2.4.16 et 2.4.17 afin de reproduire le mieux possible les courbes expérimentales. Ainsi, on peut obtenir les valeurs d'intérêt k_a et k_d (et ultimement les constantes d'équilibre $K_D = k_d/k_a$ et $K_A = k_a/k_d$) qui permettent de caractériser le degré d'affinité entre le récepteur et le ligand.

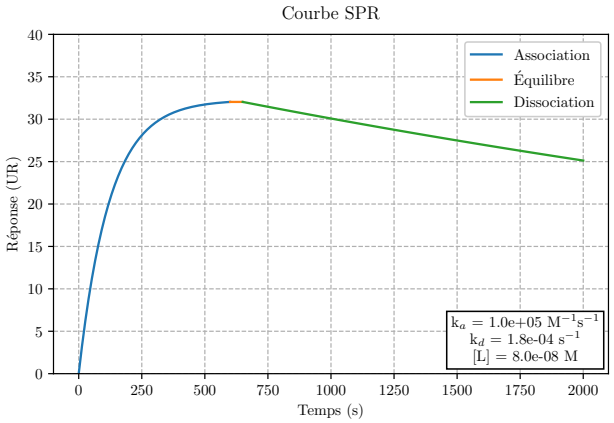


Fig. 2.6. Profil classique de réponse de biosenseur SPR pour l'étude des interactions récepteurs/ligands. Réponse SPR (en unité de réponse) en fonction du temps. Les étapes d'association, d'équilibre et de dissociation sont présentées respectivement en bleu, orange et vert. Les paramètres utilisés sont $k_a = 1 \times 10^5$ $M^{-1}s^{-1}$, $k_d = 1.8 \times 10^{-4}$ s^{-1} ($K_D = 1.8 \times 10^{-9}$ M) et $[L] = 80$ nM. Ces paramètres furent choisis afin de produire une courbe comparable à celle entre corilagin/RBD présentée à la Figure 3.5.

Chapitre 3

Corilagin and 1,3,6-Tri-O-galloy-β-D-glucose: potential inhibitors of SARS-CoV-2 variants

par

Vincent Binette[1], Sébastien Côte[1], Mohamed Haddad[2], Phuong Trang Nguyen[3], Sébastien Bélanger[4], Steve Bourgault[5], Charles Ramassamy[6], Roger Gaudreault[7] et Normand Mousseau[7]

([1]) Département de physique, Université de Montréal, Case postale 6128, succursale Centre-ville, Montréal, QC, H3C 3J7 Canada.

([2]) Centre Armand-Frappier Santé Biotechnologie, 531 boulevard des Prairies, Laval, QC, H7V 1B7 Canada.

([3]) Département de Chimie, Université du Québec à Montréal, 2101 Rue Jeanne-Mance, Montreal, QC, H2X 2J6 Canada.

([4]) Department of Physics, McGill University, 3600 University Street, Montreal, QC, H3A 2T8 Canada

([5]) Département de Chimie, Université du Québec à Montréal, 2101 Rue Jeanne-Mance, Montreal, QC, H2X 2J6 Canada.

([6]) Centre Armand-Frappier Santé Biotechnologie, 531 boulevard des Prairies, Laval, QC, H7V 1B7 Canada.

([7]) Département de physique, Université de Montréal, Case postale 6128, succursale Centre-ville, Montréal, QC, H3C 3J7 Canada.

Cet article a été soumis dans Physical Chemistry Chemical Physics.

Mes contributions et le rôle des coauteurs:

Au niveau de notre étude des molécules thérapeutiques contre la COVID, mes principales contributions se retrouvent le long de trois axes principaux:

(1) **Conception de la méthode**: Au niveau de la méthodologie, ma principale contribution a été de concevoir et de réaliser le protocole numérique, incluant: les protocoles d'amarrage moléculaire (choix des régions d'intérêt, des paramètres, etc.), les protocoles de dynamique moléculaire (choix du potentiel et des paramètres de simulation) et le protocole de calcul d'énergie libre par MM/PBSA.

(2) **Réalisation des simulations**: À ce niveau, ma principale contribution a été de réaliser les amarrages molécules et de réaliser les dynamiques moléculaires au niveau des complexes protéines/ligands. Tout particulièrement, j'ai réalisé l'entièreté des simulations en lien avec les mutants.

(3) **Analyse, visualisation et publication**: J'ai contribué à l'analyse des résultats, notamment en développant les figures de l'article et en comparant avec la littérature. Finalement, j'ai contribué à toutes les étapes d'écriture de l'article et des corrections en tant que premier auteur.

Aussi au niveau numérique, *Sébastien Côté* a travaillé à la conception des protocoles de simulations, au développement des outils d'analyse et de visualisation, tout particulièrement au niveau des protéines seules ou en complexe, sans ligand, en plus de participer à l'écriture de l'article. *Sébatien Bélanger* a réalisé les calculs et simulations préliminaires servant à la mise en place du protocol final. *Roger Gaudreault* a participé à la conception de l'article, tout particulièrement au niveau du choix des ligands, en plus de contribuer à son écriture. Tout le côté numérique (conception, analyse), ainsi que l'écriture de l'article, s'est fait sous la supervision du professeur *Normand Mousseau*.

En plus des résultats numériques, l'article présente aussi des résultats expérimentaux. *Mohamed Haddad* et le professeur *Charles Ramassamy* ont réalisé toutes les expériences ELISA tandis que *Phuong Trang Nguyen* et professeur *Steve Bourgault* ont réalisées les expériences de SPR.

Abstract. The COVID-19 disease caused by the virus SARS-CoV-2, first detected in December 2019, is still emerging through virus mutations. Although almost under control in some countries due to effective vaccines that are mitigating the worldwide pandemic, the urgency to develop additional vaccines and therapeutic treatments is imperative. In this work, the natural polyphenols corilagin and 1,3,6-tri-O-galloy-β-D-glucose (TGG) are investigated to determine the structural basis of inhibitor interactions as potential candidates to inhibit SARS-CoV-2 viral entry into target cells. First, the therapeutic potential of the ligands are assessed on the ACE2/wild-type RBD. We first use molecular docking followed by molecular dynamics, to take into account the conformational flexibility that plays a significant role in ligand binding and that cannot be captured using only docking, and then analyze more precisely the affinity of these ligands using MMPBSA binding free energy. We show that both ligands bind to the ACE2/wild-type RBD interface with good affinities which might prevent the ACE2/RBD association. Second, we confirm the potency of these ligands to block the ACE2/RBD association using a combination of surface plasmon resonance and biochemical inhibition assays. These experiments confirm that TGG and, to a lesser extent, corilagin, inhibit the binding of RBD to ACE2. Both experiments and simulations show that the ligands interact preferentially with RBD, while weak binding is observed with ACE2, hence, avoiding potential physiological side-effects induced by the inhibition of ACE2. In addition to the wild-type RBD, we also study numerically three RBD mutations (E484K, N501Y and E484K/N501Y) found in the main SARS-CoV-2 variants of concerns. We find that corilagin could be as effective for RBD/E484K but less effective for the RBD/N501Y and RBD/E484K-N501Y mutants, while TGG strongly binds at relevant locations to all three mutants, demonstrating the significant interest of these molecules as potential inhibitors for variants of SARS-CoV-2.
Keywords: SARS-CoV-2 mutants, Molecular dockings, Molecular dynamics, Free-energy computation

3.1. Introduction

The COVID-19 disease, first detected in late December 2019 in Wuhan China, has quickly spread worldwide leading to 145 million reported cases and 3.1 million deaths as of April 2021 [**61**]. This disease is caused by the virus SARS-CoV-2 of the coronavirus family, which is characterized by a lipid envelope sealing a genome made of a single positive RNA strand. To replicate its genome, SARS-CoV-2 has to penetrate and hijack the translation center of a host cell.

To do so, SARS-CoV-2 uses its *Spike* protein to bind to the angiotensin-converting enzyme 2 (ACE2) [**18**, **22**], a receptor found at the cell surface in a wide variety of human organs, such as the heart, the liver, the kidneys and alveoli [**20**] and that plays a decisive regulating function in the renin-angiotensin system (RAS) [**19**]. More specifically, the crystal structure of the *Spike* protein and ACE2 interface [**21**] shows that the interactions with ACE2 are mediated by the receptor binding motif (RBM) of the receptor binding domain (RBD) of the *Spike* protein.

The SARS-CoV-2 virus is constantly changing due to evolutionary pressure and mutations in its crucial *Spike*-protein were observed all over the world. The B.1.1.7 variant, initially observed in the United Kingdom, includes a mutation at position 501 of the RBD where the asparagine is replaced by a tyrosine (N501Y) [**62**]. The B.1.351 variant, identified first in South Africa, counts multiple mutations on the RBD including K417N, E484K and N501Y [**63**]. The B.1.1.28 variant, originating in Brazil, and its descendent, the P.1 mutant, contains multiple mutations on the RBD, including K417T, E484K and N501Y [**64**]. Among these mutations, the E484K and N501Y mutations, present in the above variants, could be critical as they are located on the RBD of the *Spike* protein and they have been shown experimentally to confer enhanced affinity for ACE2 [**65**].

Many strategies to prevent virus-induced infection aim at using small molecules to mitigate one (or many) of the steps of the SARS-CoV-2 mechanism of action [**23**]. On the *Spike* protein, two main regions could potentially be targeted; (1) The RBM of the RBD that directly interacts with ACE2 [**21**] and, although not located directly at the interface, (2) the furin cleavage sites of the *Spike* protein have been shown to significantly affect the binding affinity between the *Spike* protein and ACE2 [**22**, **66**, **67**]. In this work, we only focused our attention on the former.

One of the promising approaches considered for reducing the transmission of SARS-CoV-2 is the use of polyphenols because they are natural compounds found in plants and their therapeutic potential is already well documented for different diseases, e.g., neurodegenerative [**68**–**70**], cardiovascular [**71**], antihypertensive [**72**], cancers [**71**, **73**], HIV [**74**, **75**], and antiviral [**76**,**77**] including antiviral drug candidates for SARS-CoV-1 and COVID-19 [**78**,**79**]. Since the beginning of the actual pandemic, the potential of polyphenols against SARS-CoV-2 has been widely investigated [**17**, **80**]. For instance, molecular docking of amentoflavone, a natural compound found in *Ginkgo Biloba*, on SARS-CoV-2 *Spike* protein showed a high binding affinity [**81**]. Other results from molecular docking and MD simulation on SARS-CoV-2 *Spike* protein identified fisetin, kaempferol and quercetin, all natural compounds found in many fruits and vegetables, as having a high binding affinity and a network of interactions that could disrupt the interaction with ACE2 [**82**].

Here, we focus on two naturally occurring polyphenols that are promising therapeutic compounds against SARS-CoV-2; corilagin ($C_{27}H_{22}O_{18}$) and TGG ($C_{27}H_{24}O_{18}$). Both molecules share a very similar structures; corilagin phenolic rings (R3-R6) are joined compared to TGG, making the former rigid and the latter flexible [**83**, **84**]. Both molecules have very low toxicity even at high dosages [**85**] as well as promising therapeutic properties [**86**]. For example, corilagin was described as having anti-hypertensive [**87**], anti-inflammatory and antioxidant [**88**] properties. On the other hand, the less studied TGG is closely related to the tetra-TGG molecule, a promising therapeutic compound against SARS-CoV-1 [**78**].

In this work, we probe in more details the crucial interactions between corilagin/TGG and the *Spike* protein/ACE2 interface. First, we use MD simulations on the ligand-protein complexes to take into account the conformational flexibility that plays a significant role in ligand binding and that cannot be captured using only docking [**89**, **90**]. Second, we analyze more precisely the affinity of these ligands using MMPBSA binding free energy [**91**]. These numerical predictions and methodology are validated using experimental tools; Surface Plasmon Resonance (SPR) as well as Binding Inhibitor Assay (ELISA). Finally, we assess numerically the impact of emerging SARS-CoV-2 mutations (E484K, N501Y and E484K/N501Y) of the variants of concerns on the binding affinity of corilagin and TGG to RBD.

3.2. Materials and Methods

We investigate the mechanisms and binding affinity of corilagin and TGG with ACE2 and RBD using a combination of simulations (molecular docking, molecular dynamics and MMPBSA free energy calculations) and experiments (surface plasmon resonance and binding inhibitor assay).

3.2.1. MD simulations

As a first step, we perform 500-ns MD simulation on the ACE2/RBD complex as well as on ACE2 and RBD alone to evaluate their stability and fluctuations. Each system was prepared as follow: (1) the system undergoes an energy minimization step in vacuum using sequentially the steepest descent (SD) and conjugate gradient (CG) algorithms; (2) it is then solvated with explicit water molecules (TIP3P) inserted to fill the dodecahedron box; (3) and ions are added until neutrality; (4) the solvent configurational energy is minimized using sequentially the SD and CG algorithms with all non-hydrogen atoms of the protein kept in place using harmonic restraints; (5) the whole system is equilibrated in the NVT ensemble at 300 K over 10 ns, while maintaining harmonic restraints on non-hydrogen atoms; (6) this is followed by a 10-ns NPT equilibration also with harmonic constraints on non-hydrogen atoms; and (7) a full molecular dynamics (MD) simulation in the NPT ensemble, without any restraint.

All simulations are run with GROMACS v2019.3 [**58**]. The all-atom AMBER14sb forcefield [**30**] is used for the parameters of the protein. The temperature is kept at 300 K using the Nosé-Hoover thermostat [**44**, **45**] with a coupling constant of 0.1 ps. This temperature is the same as the one used for AMBER14sb's parametrization and testing, and it is in line with the temperature in the experiments of our study. Counter ions (Na^+ and Cl^-) were added to obtain neutrality. The pressure is fixed at 1 atm using the Parrinello-Rahman barostat [**47**] with a coupling constant of 2 ps. We apply a cutoff of 1 nm for both the van der Waals and electrostatic interactions. Long-range electrostatic interactions are computed

using Particle Mesh-Ewald [**92**, **93**]. Bond lengths are constrained using LINCS [**48**] and water geometry are constrained using SETTLE [**49**].

The ACE2/RBD complex, ACE2 and RBD simulations are analyzed on the 250 to 500 ns interval (see next section). The ACE2/RBD complex simulation is used to quantify the contacts, H-bonds and salt-bridges between ACE2 and RBD in order to characterize the ability of the ligands to block those interactions. The ACE2 and RBD simulations are used to determine an ensemble of configurations representative of their flexibility in order to take it into account while performing the docking of the ligands. Their main configurations are identified using Daura's clustering algorithm [**94**] with a cutoff of 0.15 nm on the backbone atoms. Clusters containing at least 5% of the total population, four clusters for the RBD and three clusters for ACE2, are considered for docking (more details in section 3.2.2).

We also perform 100-ns MD simulations for three RBD mutants (E484K, N501Y and E484K with N501Y) using the protocol described above. These simulations were started from the center of the biggest cluster of the RBD simulation. The three RBD mutants (E484K, N501Y and E484K with N501Y) are generated using PyMOL [**95**].

Identification of the interactions between ACE2 and RBD in the complex. The interactions between ACE2 and RBD in terms of contacts, H-bonds and salt-bridges are determined from a 500-ns MD simulation on the ACE2/RBD complex, starting from the experimental structure determined using X-rays crystallography (PDB:6M0J) [**21**] (SFig. B.1A). The interface of the ACE2/RBD complex remains globally similar to the crystal structure, particularly on the 250-500 ns convergence interval, as shown by the backbone-RMSD of the interface (0.22 ± 0.02 nm) and the probability of the secondary structure motifs for both ACE2 and RBD (SFig. B.1B-C). In terms of secondary structure, the propensity of α-helices, β-sheets, turns and coils for ACE2 and RBD are essentially the same as in the crystal structure, except for a drop from 11% in the crystal to 3 ± 2% in the RBD α-helix propensity over the simulation.

During the simulation, most ACE2-RBD contacts are between the A1A2 segment (residues 19-83, helix-helix) or the HS segment (residues 322-362, helix-sheet-sheet) of ACE2 and the RBM segment (residues 438-506, mainly disordered with a small helix and small sheets) of RBD, as shown in SFig. B.2A. Overall, most of these ACE2-RBD contacts are present in the crystal structure: experimental contacts are observed 72 ± 6% of the time during the simulation, with this percentage significantly rising up to 89 ± 4% when using a slightly less stringent distance threshold of 0.6 nm for the contact definition during the simulation (instead of 0.4 nm as in the experiment). More precisely for RBD, there are only four residues that interact with ACE2 in the crystal structure, but that interact with ACE2 less than 60% of the time during the simulation: Lys-417, Gly-446, Gly-447 and Glu-484. On the other hand, three more residues of RBD interact with ACE2 during the simulation:

Phe-490, Pro-491 and Leu-492. In terms of H-bonds between ACE2 and RBD, all those observed in the crystal structure (Asp-30, Gln-42, Tyr-83 and Lys-353 on ACE2's side and Gly-446, Asn-487 and Gly-502 on RBD's side) are also present to varying degrees during the simulation (SFig. B.2B). Moreover, other relevant H-bonds are observed during the simulation because it takes into account the flexibility of the complex coming from it being in a solvated environment at 300 K and 1 atm. In terms of salt-bridges, the crystal D30-K417 salt-bridge is the most populated in our simulation and two new salt-bridges (E223-K458 and E37-R403) are also observed (SFig. B.2C).

A more detailed comparison of the H-bonds present in the crystal structure is also presented in Table B.1. After addition of the hydrogen atoms in the crystal structure and minimization, we found that five out of the thirteen H-bonds identified in Table 1 of *Lan et al.* [**21**] satisfy our distance and angle criteria for H-bonds identification: Lys-417(RBD)/Asp30(ACE2), Asn-487(RBD)/Gln-24(ACE2), Asn-487(RBD)/Tyr-83(ACE2), Tyr-489(RBD)/Tyr-83(ACE2) and Tyr-505(RBD)/Glu-37(ACE2). In our MD simulation, most of these H-bonds are unstable and only Lys-417(RBD)/Asp-30(ACE2) and TYR505(RBD)-GLU37(ACE2) are formed with over 25% occurrence rate. In addition to these H-bonds, the minimization of the crystal structure leads to the formation of 10 new H-bonds. In our MD simulation, only Tyr-449(RBD)/Asp-38(ACE2), Gln-493(RBD)/GLU-35(ACE2) and Gln-493(RBD)/Lys-31(ACE2) are stable with an occurrence rate of 40.64%, 36.95% and 34.82% respectively (Table B.1).

Assessing the structural flexibility of the starting ACE2 and RBD structures. The representative configurations of ACE2 in solution are identified from the 500-ns MD simulation on ACE2 alone, starting from its structure in the crystal complex (PDB:6M0J) (SFig. B.3A). We establish that the simulation is converged after the first 250 ns by looking at the backbone RMSD and the secondary structure as a function of time (SFig. B.3B-C). ACE2 keeps a structure similar to when it is in the complex as shown by the backbone RMSD on the whole (0.29 ± 0.02 nm) and on the segments A1A2 and HS at the interface with RBD (0.26 ± 0.02 nm). In terms of per residue secondary structure, slightly longer α-helices are observed in the A1A2 segment during the simulation, while slightly longer β-sheets are observed in the HS segment compared to the crystal structure in complex form (SFig. B.3D).

Similarly, the representative configurations of RBD in solution are identified from the 500-ns MD simulation on RBD alone, starting from its structure in the crystal complex (PDB:6M0J) [**21**] (SFig. B.4A), with convergence also achieved after 250 ns (SFig. B.4B-C). While RBD deviates more from the complex structure than ACE2, it stays relatively near from it as shown by the backbone RMSD: 0.37 ± 0.03 nm on the whole RBD and 0.40 ± 0.05 nm on the RBM segment at the interface with ACE2. In terms of per residue secondary structure, the helix and sheets fluctuate with other motifs, while some sheets are

slightly longer during the simulation (SFig. B.4D). In particular, residues 443-447, 474-489 and 500-505 of the RBM, which are interacting with ACE2 in the complex, show the highest degree of fluctuations.

3.2.2. Molecular docking

The docking of corilagin and TGG are performed using AutoDock VINA v1.1.2 [**55**]. The protein flexibility is taken into account by performing the docking on the center of all clusters representing at least 5% of the sampled population. The ligand flexibility is considered by VINA's methodology. VINA's estimation of the "correctness" of the poses is done based on a simplified physics-based potential with empirically determined weights [**55**]. Every atomic pair is affected by a steric term, and, depending on the pair type, a hydrophobic term and a hydrogen bond term [**55**]. However, it is important to note that docking and scoring techniques are often simplified for efficiency and the predicted binding affinities only weakly correlates with experimental predictions [**96**]. The region of interest during the docking involves residues found at the ACE2/RBD interface in the complex. On ACE2, the region of interest involves two long α-helices between residues 19 and 83 (referred to as A1A2) and small α-helix followed by a small β-sheet between residues 322 and 362 (referred to as HS). The docking is targeted on these regions using a box with x, y, z dimensions of 22.640, 52.072 and 14.633 Å, respectively. On the RBD, the region of interest is composed of residues 438 to 506 (referred to as the RBM) and docking is focused on this region using a 27.625 $\times$ 43.358 $\times$ 26.026 Å box. Both molecular systems can be visualized in Figure B.6. VINA's exhaustiveness parameter is set to 100. The conversion between PDB and PDBQT format is done using Open Babel v3.1.0 [**97**].

3.2.3. Protein-ligand simulations

We use the best prediction (highest binding affinity) generated by AutoDock VINA as the starting point for additional MD simulations for each combination of the five proteins – ACE2, RBD, RBD(E484K), RBD(N501Y) and RBD(E484K-N501Y) – and two ligands – corilagin and TGG. The same simulation protocol described in section 3.2.1 is used to launch a 100-ns simulation for each of the 10 systems.

The ligand parameters were determined using the generalized AMBER forcefield (GAFF) [**34**] with their partial charges determined using the RESP protocol [**37**, **38**] with ANTECHAMBER [**34**, **98**]. The electrostatic potential of each ligand has been computed using HF6-31G*//HF6-31G* with Gaussian16 [**99**]. The initial conformations of corilagin and TGG used for those computations were taken from previously published work [**83**] where they were determined using PM3 semi-empirical Molecular Orbital Theory. The

initial structures of the ligands are shown in Figure B.5. All files are converted GROMACS compatible format with the help of ACPYPE [**100**].

3.2.4. Analysis

The analysis of the MD simulations is done using a combination of GROMACS tools [**58**] and in-house scripts. Secondary structures (SS) are determined using DSSP [**101**]. Hydrogen bonds (H-bonds) are defined using a 0.35 nm donor-acceptor cutoff and a 30° hydrogen-donor-acceptor angle cutoff. Contacts are defined with a 0.40 nm cutoff, the same cutoff used in the analysis of the experimental structure [**21**]. Salt-bridges are defined using a 0.40 nm distance cutoff between the oppositely charged groups [**102**]. Molecular visualization is done using PyMOL [**95**] and ligand/protein interaction visualization using LigPlot+ [**103**, **104**]. Daura's algorithm is used for clusterization [**94**].

3.2.5. Binding free-energy

The MMPBSA method [**91**] is used to estimate the protein/ligand binding free-energy (ΔG_{bind}), defined by

$$\Delta G_{bind} = \langle G_{RL} - G_R - G_L \rangle_{RL}$$

where G_{RL}, G_R and G_L are the free-energy of the receptor/ligand complex, of the receptor alone and of the ligand alone, respectively. We use a single trajectory MMPBSA computation: the conformation of the complex (RL), receptor (R) and ligand (L) are all taken from a unique MD trajectory. The bracket pair $\langle\rangle$ represent an ensemble average over all receptor/ligand conformations. More specifically, the free-energy G is estimated according to

$$\Delta G = U + G_{\text{solvation}} - TS,$$

where U is the internal energy, computed using the AMBER14sb forcefield field, $G_{solvation}$ is the solvation free-energy and is usually decomposed into a polar part, computed by solving the Poisson-Boltzmann equation and a non-polar part that depends on the solvent accessible surface area (SASA), T is the temperature and S is the entropy [**51**]. The MMPBSA method offers a relatively quick and easy way to estimate the binding free-energy. It does, however, makes a few crude approximation: the solvation is considered implicitly and thus possibly neglects crucial water molecules at the binding site. Moreover, the entropic part of the equation is often neglected (as in this study) [**51**, **52**]. In spite of these limitations, the MMPBSA method has proven to be useful for refining the results of docking predictions [**51**].

MMPBSA computations are done with g_mmpbsa utility [**57**], which uses APBS [**59**] for computing the polar part of the solvation free-energy. The dielectric constants of the solute and solvent are set to 2 and 80 respectively. The surface tension (γ) is set to 0.0226778 kJ/(mol Å^2) and the temperature at 300 K. The results are computed from the

convergence interval of the ligand-protein MD simulations using 40 ps snapshots. A 500-steps bootstrap analysis is used to compute the average and standard deviation of the free energy.

3.2.6. Products

Corilagin (β-1-O-Galloyl-3,6-(R)-hexahydroxydiphenoyl-D-Glucose), with molecular formula $C_{27}H_{22}O_{18}$ and molecular weight of 634.45 g mol^{-1}, was obtained from Cayman Chemical (USA). The powder material, C.A.S. 23094-69-1, is natural in origin, with purity >98%. TGG (1,3,6-tri-O-galloy-β-D-glucose) with molecular formula $C_{27}H_{24}O_{18}$ and molecular weight of 636.46 g mol^{-1}, was obtained from MuseChem (USA). The powder material, C.A.S. 18483-17-5, is natural in origin, with purity 98.23%. Host Cell Receptor Binding Domain (RBD) (RayBiotech, cat number: 230-30162) was expressed at Arg319-Phe541 region in human embryonic kidney (HEK293) cells with a His-tag at C-terminal. The protein was supplied as a 0.2 μm filtered solution in PBS (pH 7.4) with purity > 95%. Recombinant Human ACE2 Protein was purchased from Bioss Inc. (Cat number: BS-46110P). Recombinant Human ACE2 Biotinylated Protein was purchased from (R&D Systems).

3.2.7. Surface Plasmon Resonance (SPR)

SPR analyses are performed using a Biacore T200 instrument (GE Healthcare). S1-RBD and ACE2 recombinant proteins are respectively immobilized on a carboxymethylated dextran CM5 sensor chip (GE Healthcare) using an amine-coupling strategy. Briefly, the sensor chip surface is activated with a 1:1 mixture of N-hydroxysuccinimide (NHS) and 3-(N,N-dimethylamino)-propyl-N-ethylcarbondiimide (EDC). Recombinant protein solutions (20 μg/ml) are injected at a flow rate of 10 μl/min using HBS-N running buffer (10 mM HEPES, 150 mM NaCl, pH 7.4) to reach a level of immobilization of 200 RU. Surfaces (protein and reference) are blocked by the injection of an ethanolamine-HCl solution. Binding kinetics of TGG and corilagin over the immobilized recombinant proteins sensor chip are evaluated in HBS-N buffer with increasing polyphenol concentrations (1 to 100 nM) at a flow rate of 20 μl/min. Association time is set at 180 sec and dissociation time is extended up to 1,200 seconds. The sensor chip surface is regenerated by injecting 15 μl of a 10 mM glycine solution, pH 3. For ACE2/RBD interactions, the binding partner is injected over the counterpart-functionalized surface with concentrations from 1 to 100 nM and surface is regenerated with 15 μl of a 50 mM NaOH solution. For inhibition assay, 50 nM RBD recombinant protein is pre-incubated for 30 min at room temperature (RT) with increasing polyphenol concentrations and the mixtures are subsequently injected over an ACE2 functionalized CM5 surface. Binding sensograms are obtained by subtracting the reference flow cell (without protein). Experiments are performed at least in duplicate and data analysis is

performed using the BIA evaluation software package (GE Healthcare) and fit to a one-site (1:1 molecular ratio) Langmuir binding model.

3.2.8. SARS CoV-2 RBD Spike Protein and Human ACE2 Binding Inhibitor Assay

The capacity of TGG and corilagin to inhibit the binding between the RBD Spike protein and the human ACE2 was assessed at different concentrations from 0.1 to 10 µM, by ELISA. For this, ELISA plates were coated with 0.5 μg/ml of RBD Spike protein and kept overnight at 4°C. Plates were then rinsed three times with the washing buffer (0.05% Tween 20 in phosphate buffered saline (PBS)) and then blocked with the blocking buffer (1% bovine serum albumin (BSA) in PBS) by incubating for 1 hour at 37°C. After three washing, one hundred microliters of biotinylated human ACE2 protein, diluted at 0.5 μg/ml in the blocking buffer, were added to each well and incubated at 37°C for 1 h. After washing with the same washing buffer, diluted peroxidase-conjugated streptavidin was added to each well and incubated at 37°C for 30 min. Following three washes, chromogenic substrate solution was added to each well and incubated at 37°C for 30 min followed by 50 μL of the stop solution (2N H_2SO_4). The absorbance was then read at 450 nm. To note, a concentration response curve for the human ACE2 protein (0.015 to 2 μg/ml) was established to confirm a concentration-dependent increase of the absorbance at 450 nm (Figure B.7A). For the competition assay, different concentrations of TGG and corilagin were incubated with immobilized RBD Spike protein for 1 hour at 37°C before the addition of the human ACE2 protein.

3.2.9. TGG, corilagin and human ACE2 binding assay

To study the possible binding of TGG or corilagin and their mixture to the human ACE2, the ELISA ACE2 detection kit (R&D Systems) was used with some modifications. Plates were coated with 0.5 μg/ml of human ACE2 antibody which can bind to the extracellular region of the ACE2 protein (AA 18-740) at room temperature during overnight. After washing, wells were blocked with the blocking buffer. For the competition assay, biotinylated human ACE2 was mixed with various amounts of TGG or corilagin or their mixture for 1 hour at 37°C. After incubation, the mixture of human ACE2 and polyphenols was added to the coated wells and incubated for 1 hour at 37°C. After washing, diluted peroxidase-conjugated streptavidin was added to each well and incubated at 37°C for 30 min. Chromogenic substrate was added to each well after washing and incubated at 37°C for 30 min. The absorbance was then read at 450 nm in a fluorescent microplate reader. To note, a concentration response curve for the human ACE2 protein (0.015 to 2 μg/ml) was established to confirm a concentration-dependent increase in absorbance at 450 nm (Figure B.7B).

3.2.10. Statistical analysis for binding assays

Data were analyzed using the GraphPad Prism program. For the inhibitory effects of TGG, corilagin and their mixture on the SARS CoV-2 Spike protein RBD and human ACE2 interactions, statistical analyses were performed using One-way ANOVA analysis followed by the Dunnett's t-test. A p value less than 0.05 was considered statistically significant.

3.3. Results

3.3.1. The impact of Corilagin and TGG on the ACE/RBD Wild-Type interactions

3.3.1.1. Molecular Docking. In order to probe the possible interaction sites between the ligands (corilagin/TGG) and ACE2/wild-type RBD (WT-RBD), we first carry out molecular docking using AutoDock VINA [**55**, **103**]. To consider the protein flexibility, we use representative structures extracted from the MD simulations run independently on ACE2 and RBD as described in Sect. 3.2.1. The docking simulations result in a wide variety of predicted conformations characterized by varying binding affinities and positions at the interface. Conformations sampled on RBD (WT and mutants) are shown as two-dimensional occurrence map as a function of VINA's binding energy and fraction of contacts with interface residues on Figure B.8 for corilagin and Figure B.9 for TGG.

Docking on the RBD-WT for corilagin leads to docked positions with VINA docking energy ranging from -5.8 to -8.1 kcal/mol and a fraction of contacts with interface residues going from 0.20 to 0.45. For the RBD-WT and TGG, the VINA docking energy spectrum is narrow and lower, from -7.0 to -8.8 kcal/mol, with a broader fraction of contacts from 0.20 to 0.55 among the identified docked conformations. Interestingly, many docked TGG conformations are characterized by a low binding energy and a high fraction of contacts with interface residues.

3.3.1.2. Molecular Dynamics. Molecular docking by itself takes into account only limited protein and ligand conformational dynamics. Yet, molecular flexibility is critical for a reliable and predictable characterization [**89**, **90**]. Thus, we perform a 100-ns MD simulation on the best predicted ACE2/RBD-WT with corilagin/TGG complexes given by Autodock VINA in order to allow for local rearrangements both on the protein and ligand sides.

ACE2. ACE2's structure at the interface remains very stable when in contact with either corilagin or TGG (Figure B.10). After docking, the A1A2 and HS backbone RMSD against the experimental structure is 0.25 ± 0.01 and 0.25 ± 0.01 nm for corilagin and TGG respectively as compared to 0.26 ± 0.02 nm without ligand. Moreover, both the A1A2 and HS secondary structures remain unaffected by the presence of ligand: the α-helix content is

$53 \pm 3\%$ (corilagin) and $53 \pm 1\%$ (TGG) in the presence of the ligands as compared to $55 \pm 1\%$ without the ligands, and the β-sheet content remains at $15 \pm 2\%$ (corilagin) and $15 \pm 2\%$ (TGG), while it is $16 \pm 1\%$ without ligand. The initial docked positions of the corilagin, located in between the HS segment and the middle of the A1 helix, and TGG, located on the flexible loops of the HS segments, are also very stable and show little deformation during the MD.

RBD. The structure of the WT RBM segment is also only weakly affected by the ligands (Figures 3.1 and 3.2). Indeed, the average backbone RMSD measured with respect to the experimental structure computed is 0.37 ± 0.02 nm, 0.44 ± 0.05 nm for corilagin and TGG respectively as compared to 0.40 ± 0.05 nm for the system without the ligands. The secondary structure is largely unaffected by the presence of the ligand. With a bound ligand, its α-helix content is $5 \pm 2\%$ (corilagin) and $4 \pm 3\%$ (TGG), compared to $6 \pm 3\%$ without it; the β-sheet content is $22 \pm 6\%$ with corilagin, and $16 \pm 4\%$ with TGG, as compared to $16 \pm 4\%$ without any ligand. The solvent accessible surface area of the WT RBM is of 49 ± 2 nm^2. This relative stability on the protein side is reflected on the binding conformations of corilagin and TGG found with VINA, that both remain very stable over the MD simulation (Figures 3.1 and 3.2).

3.3.1.3. Interactions and Binding Energies. Molecular docking predictions use a simplified binding affinity score. In order to refine the estimation of the binding energies, we turn to the MMPBSA technique. We also characterize the interactions network between the proteins (ACE2/RBD-WT) and the ligands (Corilagin/TGG) and their evolution over the MD simulations.

ACE2. The binding affinities of both ligands with ACE2 during the MD simulations are compared using the VINA score as well as the MMPBSA free energy as explained in Section 3.2 (Table 3.2). The average binding affinity of corilagin/ACE2 is -6.1 ± 0.5 kcal/mol (VINA score) and -0.1 ± 0.2 kcal/mol (MMPBSA). In spite of this negligible binding affinity, corilagin remains associated with ACE2 during the entire MD simulation, demonstrating at least the presence of a metastable state, once binding occurs. The binding affinity of TGG/ACE2 is more favorable with -6.0 ± 0.4 kcal/mol (VINA) and -14.4 ± 0.2 kcal/mol (MMPBSA). The LigPlot interaction maps between corilagin/TGG and ACE2 for the center of the biggest cluster (total population of 84% for corilagin and 98% for TGG) are shown on Figure B.11. Corilagin is stabilized by nine H-bonds with multiple residues of the A1A2 segment (Asp30 twice, His34 and Glu37 twice) as well as with residues Arg393 (twice), Gln388 and Phe390. Four nonpolar contacts are formed with residues of the A1A2 segment (Asn33), the HS segment (Lys353 and Gly354) and the rest of the ACE2 (Pro389). For its part, TGG is forming four H-bonds with residues of the HS segment (Met323, Gln325 and Asp350 twice) and it is stabilized by a large number of nonpolar contacts, mainly with the

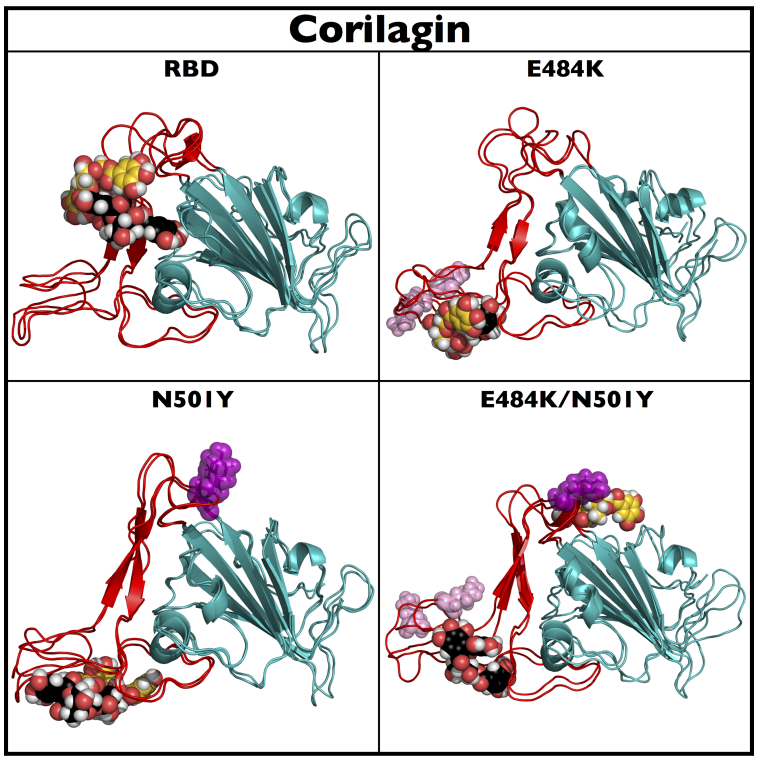


Fig. 3.1. The docked position of corilagin on RBD and the three mutants. The RBM segment and the rest of the RBD are shown respectively in red and teal. Residues 484 and 501, both the location of tested mutation, are shown in pink and purple respectively. The ligand in black and gold is respectively the conformation after docking and the center of the biggest cluster sampled during the converged part of the MD simulation respectively.

HS segment (Asn322, Thr324, Gly326, Gly352, Gly354, Asp355 and Phe356) and the rest of ACE2 (Pro321, Met383, Ala386 and Arg393).

RBD. The binding affinities of both ligands with RBD during the MD simulations are compared using the VINA score as well as the MMPBSA free energy as explained in Section 3.2 and shown in Table 3.2. The corilagin/RBD binding affinity is -5.0 ± 0.5 kcal/mol (VINA) and -7.2 ± 0.1 kcal/mol (MMPBSA) and that for TGG/RBD is similar in terms of VINA (-6.0 ± 0.4 kcal/mol), but more favorable in terms of MMPBSA (-12.8 ± 0.4 kcal/mol). The LigPlot interaction maps between the corilagin/TGG and RBD for the center of the biggest cluster (total population of 96% for corilagin and 52% for TGG) are shown in Figures 3.3 and 3.4, respectively. Corilagin is stabilized by the formation of four H-bonds with Tyr449, Gln493, Ser494 and Gly496 as well as four nonpolar contacts with Tyr495, Gln498, Gly504 and Tyr505. On the other hand, TGG is stabilized by the formation of three H-bonds with Arg454, Glu471 and Pro491 and six nonpolar contacts with Leu455, Phe456, Arg457, Lys458, Thr470 and Leu492.

3.3.1.4. Corilagin/TGG ability to disrupt the ACE2/wild-type RBD interactions. Using a combination of molecular docking and molecular dynamics simulations, we find that both

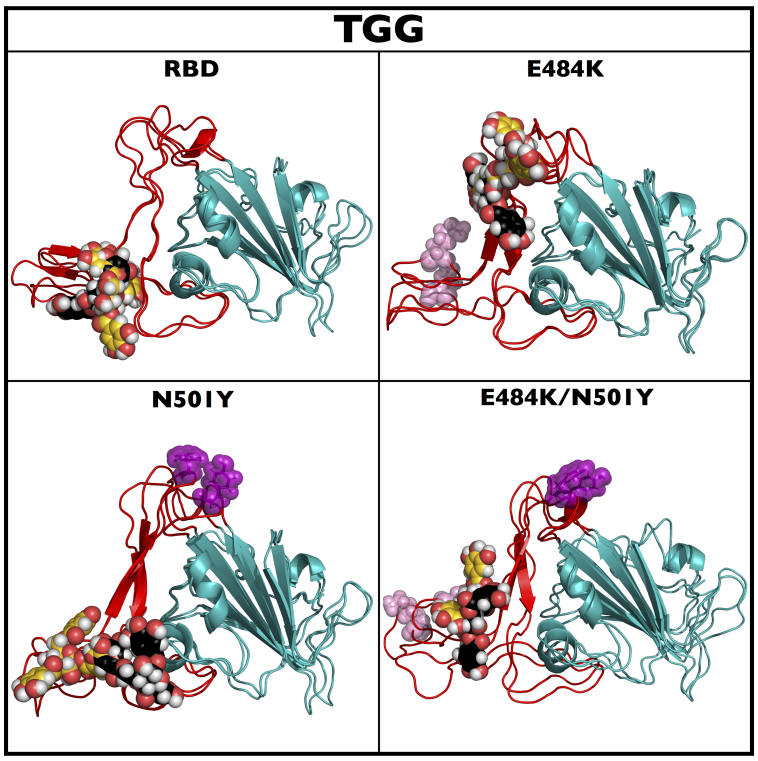


Fig. 3.2. The docked position of TGG on RBD and the three mutants. The RBM segment and the rest of the RBD are shown respectively in red and teal. Residues 484 and 501, both the location of tested mutation, are shown in pink and purple respectively. The ligand in black and gold is respectively the conformation after docking and the center of the biggest cluster sampled during the converged part of the MD simulation respectively.

ligands, corilagin and TGG, are able to interact with ACE2 and the RBD-WT. We analyze below whether those interactions are compatible with a disruption of the ACE2/RBD-WT association.

ACE2. For ACE2/corilagin, although the binding energy is low (-0.1 ± 0.2 kcal/mol according to MMPBSA), the corilagin's localization on ACE2 is compatible with the disruption of many interface residues of the A1A2 segment (Asp30, His34, Glu37) and of the HS segment (Lys353, Gly354, Arg393).

For ACE2/TGG, TGG interacts only with the HS segment (Lys353, Gly354, Asp355, Arg393) and not with the A1A2 segment (SFig. B.10) suggesting that it is predominantly interacting with a small portion of the ACE2's residues involved at the interface with RBD-WT. In this position, TGG might not be able to disrupt significantly the association of ACE2 with the RBD-WT.

RBD-WT. Our MMPBSA calculation shows that corilagin binds much more strongly to WT RBD than to ACE2. In its preferred binding site, corilagin interacts with five residues at the RBD/ACE2 interface – Tyr449, Gln493, Gly496, Gln498 and Tyr505 – and forms a H-bond with two of these – Tyr449 and Gly496 (Figure 3.1). Since these five residues

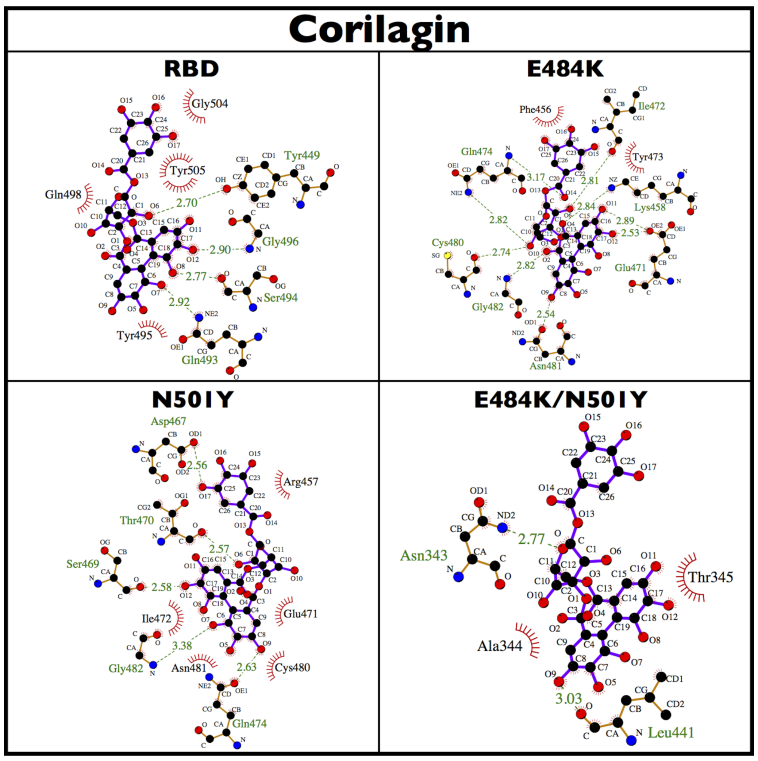


Fig. 3.3. Corilagin interaction maps. The interaction maps of corilagin with wildtype RBD (top left), RBD/E484K (top right), RBD/N501Y (bottom left) and RBD/E484K-N501Y (bottom right) are shown for the center of the biggest cluster computed on the convergence interval using the protein backbone atoms and ligand non-hydrogen atoms. The nonpolar contacts, defined by a distance smaller than 0.40 nm, between the ligand and the protein are shown as red arcs. H-bonds and their donor/acceptor distance are shown in green. All figures were generated using LigPlot [**103**, **104**].

form H-bonds with ACE2 in the complex, the presence of corilagin could interfere with the formation of these H-bonds and impair the complexation.

Our MMPBSA calculation also shows that the association between WT RBD and TGG is strong. TGG interacts with many residues involved at the ACE2-RBD interface in the complex such as nonpolar residues Leu455, Phe456, Phe490, Pro491 (with which it forms a H-bond) and Leu492 as well as polar residues: Lys458 and Gln493 (Table 3.1). Notably, these two polar residues are involved in a salt-bridge (Lys458) and a H-bond (Gln493) with ACE2.

In summary, both corilagin and TGG could have the potential to disrupt the interaction between ACE2 and the WT RBD. However, our numerical results shows that the disruption would be more important (better binding energy and better network of interactions) on the side of the WT RBD than on the side of ACE2, leaving the crucial physiological functions of ACE2 untouched.

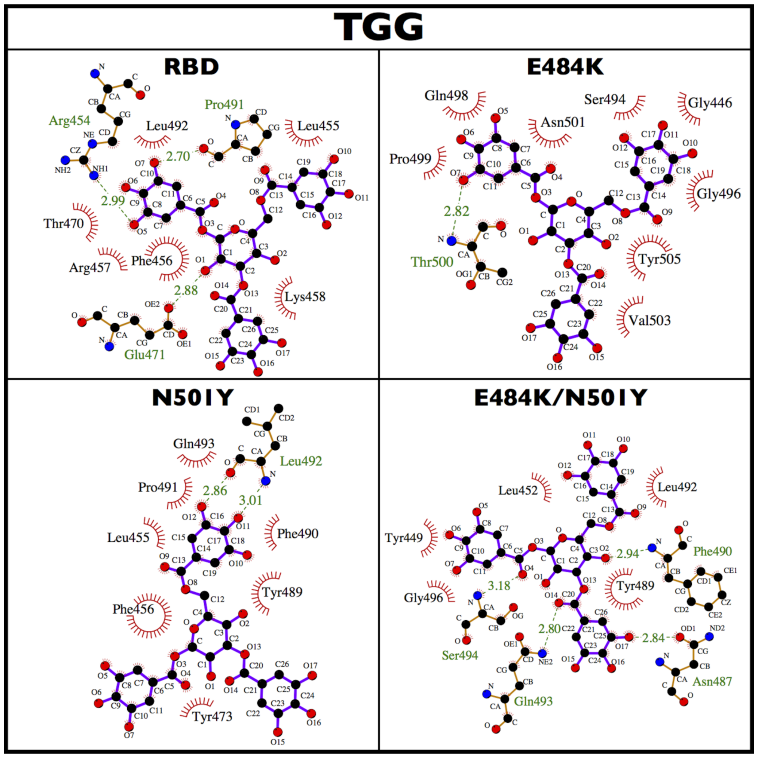


Fig. 3.4. TGG interaction maps. The interaction maps of TGG with wildtype RBD (top left), RBD/E484K (top right), RBD/N501Y (bottom left) and RBD/E484K-N501Y (bottom right) for the center of the biggest cluster computed on the convergence interval using the protein backbone atoms and ligand non-hydrogen atoms. The nonpolar contacts, defined by a distance smaller than 0.40 nm, between the ligand and the protein are shown as red arcs. H-bonds and their donor/acceptor distance are shown in green. All figures were generated using LigPlot [**103**, **104**].

3.3.1.5. Surface Plasmon Resonance. In order to validate our numerical observations, we turn to experiments for confirmation. First, we perform SPR measurements to determine the binding kinetics of TGG and corilagin to ACE2 and RBD.

The recombinant proteins ACE2 and RBD are respectively immobilized on carboxymethylated dextran sensor chips. The results (sensograms) show that both polyphenols bind avidly to the immobilized RBD (Figure 3.5A). Fitting the sensograms to a one-site (1:1 molecular ratio) binding model leads to dissociation constant (K_D) in the low nanomolar range, i.e., 1.8 nM for corilagin/RBD and 1.3 nM for TGG/RBD. In sharp contrast, no significant binding of neither corilagin, nor TGG, to ACE2 is observed by SPR over the range of 1 to 80 nM concentrations (Figure 3.5B). This observation indicates that the binding of polyphenols to RBD has clear specificity.

Next, we validate the interactions between ACE2 and RBD by means of two different experimental configurations; (i) binding of ACE2 to immobilized RBD and (ii) binding of RBD to immobilized ACE2. As expected, we observe strong interactions between these two

Tableau 3.1. ACE2-RBD contacts blocked by the ligands. RBD(RBM) residues blocked by corilagin (Cor.) and TGG. The RBM residues showed are those that are specifically involved in a contact pair with ACE2 that is formed with a probability of at least a 60% during the ACE2-RBD complex MD simulation. Nonpolar, polar, positively charged and negatively charged residues are shown respectively in gray, green, red and blue. The formation of a contact with the ligand is shown in gray. The presence of a H-bond or a salt-bridge is indicated by *HB* and *SB*, respectively. The star (*) indicates that a H-bond is present in the experimental structure. The dagger (†) beside SB for Lys458 indicates that this residue was added to the table even if its contact probability with ACE2 is less than 60% (45%) because it forms a salt-bridge with E23 of ACE2.

RBM contacts blocked by the ligands									
In contact with A1A2/HS in the complex		WT		E484K		N501Y		E484-N501Y	
		Cor.	TGG	Cor.	TGG	Cor.	TGG	Cor.	TGG
Tyr-449	HB	x HB							x
Tyr-453	HB								
Leu-455			x				x		
Phe-456			x	x			x		
Lys-458	SB†/HB		x	x			x		
Tyr-473	HB			x			x		
Ala-475									
Gly-476	HB								
Phe-486									
Asn-487	HB*								
Tyr-489	HB						x		x
Phe-490	HB		x				x		x HB
Pro-491			x HB				x		
Leu-492			x				x HB		x
Gln-493	HB	x	x				x		x HB
Gly-496	HB	x HB			x				x
Gln-498	HB	x			x				
Pro-499					x				
Thr-500	HB				x HB				
Asn-501	HB				x HB				
Gly-502	HB*								
Tyr-505	HB	x			x				

proteins (Figure 3.5C) with K_D of 41 nM (ACE2 to immobilized RBD) and 63 nM (RBD to immobilized ACE2), in agreement with recent studies [**21**, **105**, **106**].

Finally, the capacity of TGG and corilagin to inhibit the RBD/ACE2 interaction is evaluated by pre-incubating 50 nM RBD for 30 mins in presence, or absence, of increasing concentrations of polyphenols before injecting the mixtures onto an ACE2-functionalized sensor chip. Strikingly, 12.5 to 50 nM of TGG, or corilagin, fully inhibit the binding of RBD to surface-immobilized ACE2 (Figure 3.5D).

3.3.1.6. Binding assays. To complement the results obtained by MD and SPR, we also investigate the ability of TGG and corilagin to inhibit the interaction between the SARS-CoV-2 RBD protein and the human ACE2 using binding assays. We find that the incubation of SARS-CoV-2 RBD Spike protein with TGG or corilagin results in a significant reduction of the interaction between RBD and ACE2, e.g., from 45% up to 75% for concentrations from 0.1 to 10 μM (Figure 3.6A-B). Moreover, the mixture of TGG with corilagin, from 0.1

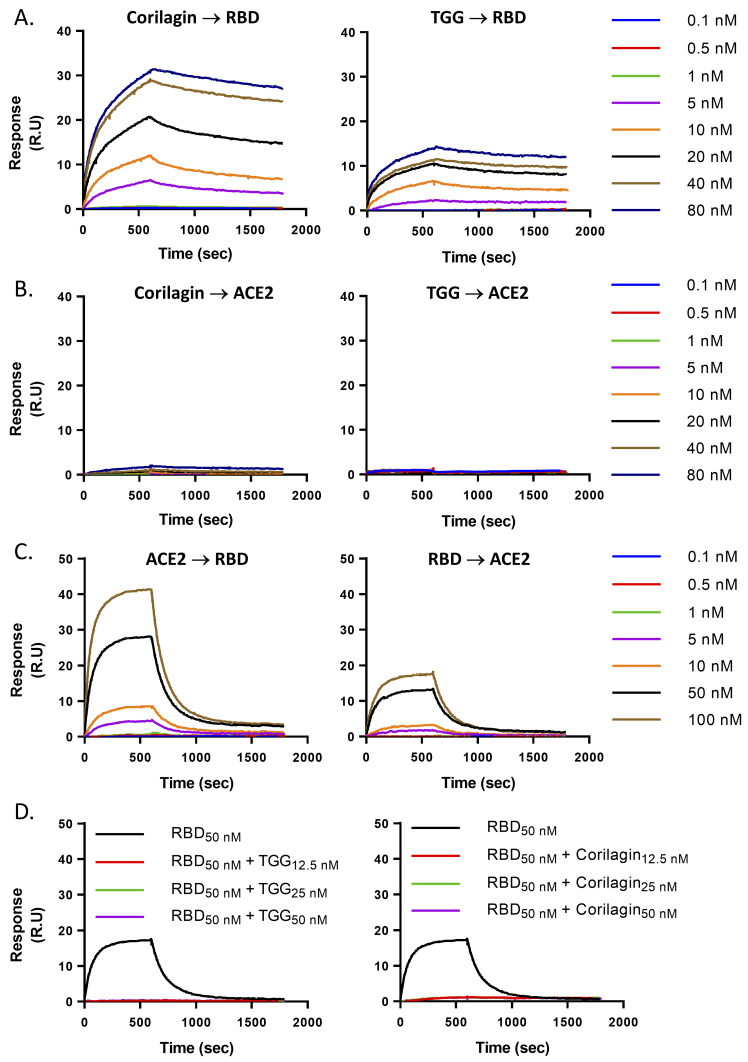


Fig. 3.5. Characterization of molecular interactions by surface plasmon resonance. A, B) Binding kinetics of corilagin and TGG on immobilized (**A**) RBD and (**B**) ACE2. The recombinant proteins RBD and ACE2 are respectively immobilized on a CM5 sensor chip and increasing concentrations of polyphenols are injected to evaluate binding kinetics. **C**) Kinetics of ACE2 binding to immobilized RBD (left panel) and kinetics of RBD binding to immobilized ACE2 (right panel). **D**) Pre-incubation of RBD (50 nM) for 30 minutes with increasing concentrations of corilagin or TGG inhibit the binding of RBD to immobilized ACE2.

to 5 μM, inhibits 50% of the interaction and does not potentiate the inhibitory effect of each compound (Figure 3.6C).

Next, we evaluate whether the inhibition of the RBD-ACE2 interaction by TGG and corilagin is associated with a preferential binding of these polyphenols to RBD in comparison to ACE2. For this, the ACE2 antibody is immobilized and the binding of ACE2 protein in absence or in presence of polyphenols is evaluated by ELISA. Strikingly, TGG and corilagin, used alone or in combination, do not reduce avidly the recognition of the ACE2 protein by

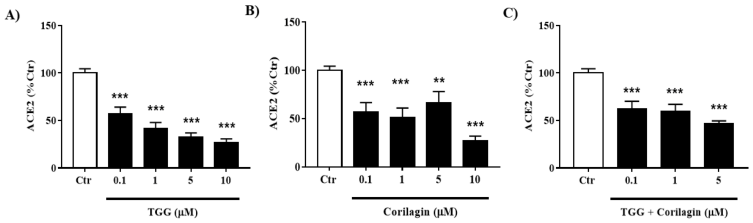


Fig. 3.6. Inhibitory effects of TGG, corilagin and their mixture on the interaction between SARS CoV-2 Spike protein and human ACE2. TGG (**A**) and corilagin (**B**) are tested at different concentrations (0.1, 1, 5 and 10 μM) and their mixture (**C**) (0,1, 1, 5 μM) to evaluate their ability to inhibit the binding of immobilized Spike protein (0.5 μg/ml) to human biotin labeled ACE2 (0.5 μg/ml), by using the ELISA assay. The absorbance values at 450 nm of human ACE2 (0.5 μg/ml) are set to 100%. Results are expressed as mean ± standard error of the mean (SEM) of two (combined effect) or three independent assays. Statistical analysis is performed using the One-way ANOVA followed by the Dunnett's post hoc test with $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$ compared to human ACE2 (0.5 μg/ml).

the anti-ACE2 (Figure 3.7A-C). These results suggest that the inhibition of the binding of the WT RBD to ACE2 is mainly mediated by the binding of these polyphenols to the RBD protein, and less to the ACE2 protein.

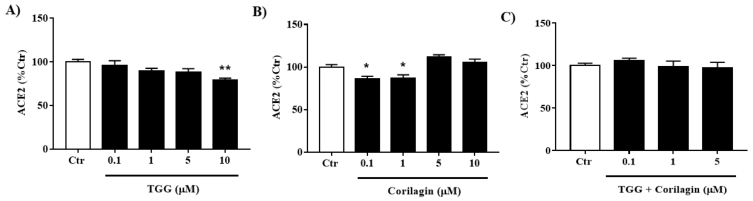


Fig. 3.7. Inhibitory effects of TGG, corilagin and their mixture on the interaction between human ACE2 and ACE2 antibody (18-740 AA). TGG (**A**) and corilagin (**B**) are tested at different concentrations (0.1, 1, 5 and 10 μM) and their mixture (C) (0,1, 1, 5 μM) to study their ability to inhibit the binding of immobilized ACE2 antibody (0.5 μg/ml) to human biotin labeled ACE2 (0.5 μg/ml), by using the ELISA assay. The absorbance values at 450 nm of human ACE2 (0.5 μg/ml) were set to 100%. Results are expressed as mean ± standard error of the mean (SEM) of two (combined effect) or three independent assays. Statistical analysis was performed using the One-way ANOVA followed by the Dunnett's post hoc test with $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$ compared to human ACE2 (0.5 μg/ml).

3.3.2. The impact of Corilagin and TGG on mutant RBDs

Our results from both numerical techniques (molecular docking and molecular dynamics) and experimental techniques (SPR and ELISA assay), show that both corilagin and TGG could interfere with the binding of the WT RBD to ACE2, by primarily interacting on the side of the RBD. However, due to evolutionary pressure, the SARS-CoV-2 virus is in constant mutations and led to the rise of three main variants of concerns with many critical mutations

on the RBD; E484K (B.1.351, B.1.1.28) and N501Y (B.1.1.7, B.1.351, B.1.1.28). In order to probe the therapeutic potential of our two molecules, we also test their binding with three mutant RBDs, with mutation E484K, N501Y and E484K/N501Y, using the described numerical protocol we used on the WT RBD and that was validated using SPR and ELISA assay.

The impact of the mutations on RBD's stability. First, the impact of the mutants E484K, N501Y and E484K/N501Y on RBD are analyzed using 100-ns MD simulations starting from the center of the biggest cluster sampled during the MD on the WT. These systems converge quickly after about 40 ns as shown by the backbone RMSD and the secondary structure as a function of time (SFig. B.12). The similarity of the RBM segment between the three mutants RBDs and the WT is characterized using three parameters: the backbone-RMSD compared to experimental structure, the secondary structure (SS) and the solvent accessible surface area (SASA).

In terms of RMSD, the RBM segment of the three mutant sequences closely resembles the WT experimental structure with 0.34 ± 0.03 nm (E484K), 0.41 ± 0.04 nm (N501Y) and 0.32 ± 0.05 nm (E484K/N501Y). The secondary structure is also similar in terms of the α-helical content with $3 \pm 3\%$ (E484K), $3 \pm 3\%$ (N501Y) and $4 \pm 3\%$ (E484K/N501Y) as well as in terms of the β-sheet content with $20 \pm 3\%$ (E484K), $24 \pm 6\%$ (N501Y) and $24 \pm 6\%$ (E484K/N501Y). In terms of SASA, the RBM of the three mutant systems is equally exposed to the solvent: 47 ± 1 nm^2 (E484K), 47 ± 1 nm^2 (N501Y) and 49 ± 1 nm^2 (E484K/N501Y). Overall, these results indicate that the structure of the RBM segment of RBD is not much affected by the mutations and closely resembles our results on the WT RBD (presented in section 3.3.1.2) as well as with the crystal structure of RBD in complex with ACE2.

Beyond the overall structural stability, we also assess the potential impact of the mutations on RBD interactions with ACE2, since this association is most critical for SARS-CoV-2's cell recognition. To do so, we focus our attention on the residues of RBD that are in contact with ACE2 when they form a complex. We compute the difference between the average SASA of these residues for all simulations (WT, E484K, N501Y and E484K/N501Y) and the experimental structure of RBD alone (Figure 3.8). In agreement with what we observe in terms of RMSD and SS, these mutations have little impact on region at the interface, as most residues have similar accessibility. The only notable difference is observed for E484K for which the accessibility of Thr500 increases and the accessibility of Asn501 decreases significantly. For the two RBDs with the N501Y mutations, we do observe a slight increase in the accessibility for residue 501 compared to the WT; this is likely due to the size increase between TYR and ASN as we do not observe any significant structural change.

Overall, the RBD residues involved at the interface with ACE2 stay accessible even in the presence of these mutations, hence, could still potentially interact with ACE2.

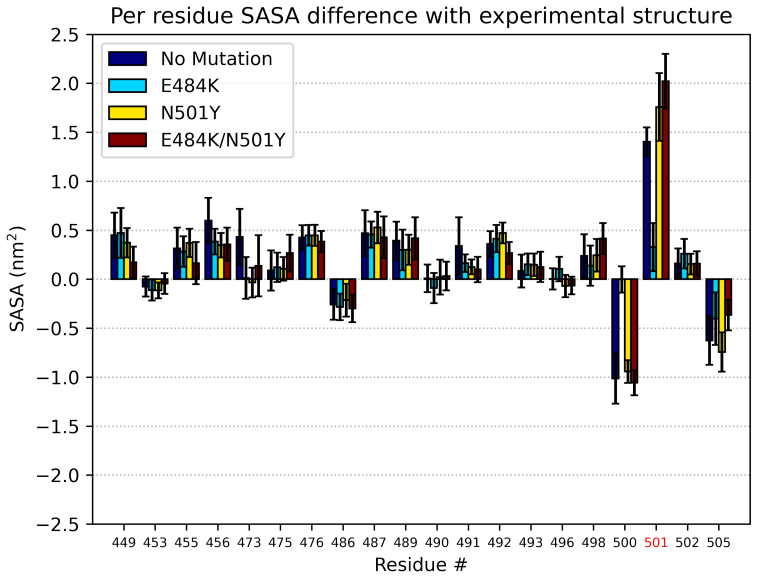


Fig. 3.8. Mutations effect on the solvent accessibility of RBD alone. RBD's per residue solvent accessible surface area (SASA) difference between the MD and the experimental structure. Only the residues of RBD interacting with ACE2 (contact probability greater than 60% during the MD simulation) in the complex structure are shown. The red residue number indicates the position of a mutation. The SASA of wildtype (blue), E484K (teal), N501Y (yellow) and E484K/N501Y (red) are compared. The error bars correspond to the standard deviation over the 250-500 ns interval.

3.3.3. Corilagin and TGG interactions with ACE2 and RBD

3.3.3.1. Molecular Docking. Using the same molecular docking protocol used for the WT, the predicted interaction sites for corilagin and TGG with the three mutants were generated. Although, there are very few structural differences between the WT and the mutants, as shown in section 3.3.2, we clearly see changes in the generated ensembles of docked conformations.

For the RBD/E484K and RBD/E484K-N501Y mutant, the two-dimensional occurrence map is very similar to the one of the WT for corilagin (Figure B.8) and TGG (Figure B.9; a very similar range of binding energies and fraction of contacts with interface residues are sampled, although, for both mutation, the distribution is slightly shifted toward higher binding energies. In contrast, with the N501Y mutation alone, the generated corilagin conformations have similar binding energy (from −5.4 to −6.5 kcal/mol), but are characterized by a wide variety of interface contact fraction (from 0.05 to 0.60). The same is observed for the docking results obtained using TGG (Figure B.9).

3.3.3.2. Molecular Dynamics. As with the WT, we follow the molecular docking with 100 ns of molecular dynamics simulation on the best predicted RBD/ligand complex for the three mutant RBD and the two ligands. Convergence of all systems is reached at 75 ns (except for RBD/E484K-N501 with corilagin, for which, the convergence is reached at 90 ns) as monitored by the RMSD on the proteins' backbone atoms and the ligands' heavy atoms (SFig. B.13).

RBD/E484K. The RMSD with respect to the experimental structure computed on the backbone atoms remains small: 0.33 ± 0.02 nm and 0.31 ± 0.07 nm for corilagin and TGG, respectively, compared with 0.34 ± 0.03 nm without the ligands. The secondary structure remains also stable in terms of α-helix content ($1 \pm 2\%$, $0 \pm 0\%$ and $3 \pm 3\%$ for corilagin, TGG and without the ligands, respectively) and β-sheet content ($19 \pm 2\%$, 17 ± 4 and $20 \pm 3\%$, respectively).

On the ligand side, the position of corilagin remains very stable, while that of TGG moves slightly away from the β-sheet of RBM in order to interact with the flexible loop between GLN498 and GLY504 (Figures 3.1 and 3.2).

RBD/N501Y. The backbone RMSD measured against the experimental structure also remains small with 0.36 ± 0.02 nm and 0.39 ± 0.03 nm for corilagin and TGG, respectively, compared with 0.41 ± 0.04 nm in the absence of a ligand. The secondary structure also stays very stable upon ligand addition: the α-helix content is $4 \pm 2\%$, $0 \pm 0\%$ and $3 \pm 3\%$ and the β-sheet content being $25 \pm 5\%$, $23 \pm 4\%$ and $24 \pm 6\%$ for respectively corilagin, TGG and without the ligands.

Corilagin stays near its initial docking position during the MD, while a small movement is observed for TGG (Figures 3.1 and 3.2). From its initial position, between the loop associated with residues 405-424 and the 455-461 loop, TGG moves slightly to interact with the 455-458 loop and the 489-493 loop. The latter portion of the loop contains residues that interact directly with ACE2 in the complex.

RBD/E484K-N501Y. The backbone RMSD measured against the experimental structure remains also small for this system: 0.32 ± 0.05 nm, 0.33 ± 0.05 nm and 0.32 ± 0.05 nm for corilagin, TGG and without the ligand, respectively. The secondary structure stays similar to WT: the α-helix content is $4 \pm 3\%$, $5 \pm 2\%$ and $4 \pm 3\%$ and the β-sheet content is $24 \pm 6\%$, $23 \pm 5\%$ and $24 \pm 6\%$ for corilagin, TGG and without the ligands, respectively.

After around 80 ns of MD, starting from the VINA docked position, corilagin disassociates from the RBD and reassociates with it at around 90 ns, but at a completely new localization, in the region between the RBM and the rest of the RBD, near the mutation N501Y (Figure 3.1). In this new position, the corilagin doesn't interact with RBM's residues involved in the association with ACE2. On the other hand, TGG is really stable at its initial docked position and stays near it for the entire MD calculation (Figure 3.2).

3.3.3.3. Interactions and Binding Energies. As for the WT RBD, the structure of the mutated RBD – E484K, N501Y and E484K/N501Y – are not affected significantly by corilagin or TGG over the 100-ns MD simulations and the VINA docking positions are generally fairly stable. Yet, small local rearrangements, which are made possible by the dynamical trajectories, contribute to a better exploration of their interactions, modifying the binding affinity and the ligand/RBD contact network as described in the following paragraphs.

RBD/E484K. The ligands' binding affinities with RBD/E484K during the MD simulations starting from an initial VINA docking are shown in Table 3.2. The computed corilagin/RBD binding affinity is -6.3 ± 0.4 kcal/mol (VINA) and -7.0 ± 0.2 kcal/mol (MMPBSA), while it is weaker for TGG/RBD when measured with VINA (-4.5 ± 0.4 kcal/mol), but stronger with MMPBSA (-10.9 ± 0.4 kcal/mol). The LigPlot interaction maps between corilagin/TGG and RBD/E484K for the center of the biggest cluster (total population of 92% for corilagin and 34% for TGG) are shown in Figures 3.3 and 3.4, respectively. Corilagin is stabilized by nine H-bonds involving seven RBM's residues (Lys458, Glu471 twice, Ile472, Gln474 twice, Cys480, Asn481, Gly482) and only two nonpolar contacts (Phe456 and Tyr473). For its part, TGG is only making one H-bond with Thr500, but has eight nonpolar contacts with Gly446, Ser494, Gly496, Gln498, Pro499, Asn501, Val503 and Tyr505.

RBD/N501Y. The binding affinities with RBD/N501Y are shown in Table 3.2. For corilagin/RBD, it is found to be -6.4 ± 0.5 kcal/mol with VINA and -11.3 ± 0.1 kcal/mol with MMPBSA, while the binding affinity for TGG/RBD is weaker with VINA (-3.8 ± 0.5 kcal/mol), but stronger with MMPBSA (-12.1 ± 0.3 kcal/mol). The LigPlot interaction maps between corilagin/TGG and RBD/N501Y for the center of the biggest cluster (total population of 99% for corilagin and 73% for TGG) are shown in Figures 3.3 and 3.4, respectively. Corilagin is forming five H-bonds with RBD (Asp467, Ser469, Thr470, Gln474 and Gly482) and is involved in five nonpolar contacts (Arg457, Gly471, Ile472, Cys480 and Asn481). TGG is again mainly stabilized by nonpolar contacts (Leu455, Phe456, Tyr473, Tyr489, Phe490, Pro491 and Gln493) and two H-bonds with Leu494 (mainchain).

RBD/E484K-N501Y. The binding affinities with RBD/E484K+N501Y are presented in Table 3.2. For corilagin/RBD, two binding sites are presented: the first site (from 20-70 ns), located at the interface, before the dissociation and the second site (90-100 ns), outside the interface, after the reassociation. For the first site, the binding affinities are -4.3 ± 0.7 kcal/mol (VINA) or -15.0 ± 0.1 kcal/mol (MMPBSA). In spite this high binding affinity computed with MMPBSA, corilagin still dissociates of the interface after around 80 ns. On the new binding sites, after reassociation, the binding energy is -4.1 ± 1.0 kcal/mol (VINA) or -1.4 ± 0.3 kcal/mol (MMPBSA). For TGG/RBD, we computed a binding affinity of: -4.9 ± 0.4 kcal/mol with VINA and -15.1 ± 0.1 kcal/mol with MMPBSA. Although the computed MMPBSA binding energy is similar to the one computed for corilagin on the first site, TGG stays strongly associated to the RBD for the entire simulated 100 ns. The

LigPlot interaction map between corilagin/TGG and RBD/E484K-N501Y for the center of the biggest cluster (total population of 40% and 93% for corilagin and TGG, respectively) is shown in Figures 3.3 and 3.4, respectively. For corilagin, the interactions with the RBD are minimal, with the formation of only two H-bonds with Asn343 and Leu441 and two nonpolar contacts with Ala344 and Thr345. On the other hand, TGG is stabilized by the formation of four H-bonds with Asn487, Phe490, Gln493 and Ser494 and five nonpolar contacts with Tyr449, Leu452, Tyr489, Leu492 and Gly496.

Tableau 3.2. The binding affinity between corilagin/TGG and ACE2/RBD. The second column shows the VINA binding affinity for the best pose found during docking. The third column shows the average VINA binding affinity computed over the interval of convergence (90-100 ns for RBD/E484K-N501Y with Corilagin and 75-100 ns for the rest) of the ligand-protein MD simulations. The fourth column shows the MMPBSA binding free energy computed over the same interval using the g_mmpbsa tools [**57**]. The average and standard deviation are computed using a 500-steps of bootstrap analysis and 40 ps snapshots on the interval of convergence (90-100 ns for RBD/E484K-N501Y with corilagin and 75-100 ns for the rest). For RBD/E484K-N501Y with corilagin, we compare the results for the two binding sites. Site 1 is the corilagin's localization before its disassociation (from 20-70 ns). Site 2 is the corilagin's localization after its reassociation (from 90-100 ns)

Binding affinity (kcal/mol)			
	Docking VINA	MD VINA	MD MMPBSA
ACE2			
Corilagin	-7.3	-6.1 ± 0.5	-0.1 ± 0.2
TGG	-7.7	-6.0 ± 0.4	-14.4 ± 0.2
Wildtype RBD			
Corilagin	-8.1	-5.0 ± 0.5	-7.2 ± 0.1
TGG	-8.8	-6.0 ± 0.4	-12.8 ± 0.4
RBD/E484K			
Corilagin	-7.2	-6.3 ± 0.4	-7.0 ± 0.2
TGG	-7.8	-4.5 ± 0.4	-10.9 ± 0.4
RBD/N501Y			
Corilagin	-6.5	-6.4 ± 0.5	-11.3 ± 0.1
TGG	-7.1	-3.8 ± 0.5	-12.1 ± 0.3
RBD/E484K-N501Y			
Corilagin, Site 1	-7.5	-4.3 ± 0.7	-15.0 ± 0.1
Corilagin, Site 2	-7.5	-4.1 ± 1.0	-1.4 ± 0.3
TGG	-7.9	-4.9 ± 0.4	-15.1 ± 0.1

3.3.3.4. Corilagin ability to disrupt the ACE2/mutated RBD interactions. For **RBD/E484K**, the MMPBSA binding free energy of corilagin is almost identical to that of the WT. Yet, Table 3.1 shows that corilagin only interacts with three different interfacial residues in this case: Phe456; Lys458, which forms a salt-bridge with ACE2;

and Tyr473, which forms a H-bond with ACE2. Thus, this association is most likely less efficient to prevent the ACE2/RBD complex formation than for the WT.

For **RBD/N501Y**, the MMPBSA binding free energy of corilagin is significantly more negative than for the WT and E484K. However, in this binding site, corilagin is not involved in interactions with crucial residues of the interface (Table 3.1). It is rather located on the other side of the loop involved in the formation of the interface. It interacts with residues Glu471, Ile472 and Gln474, instead of the residues of the other side of the loop such as Tyr473, Ala475 or Gly476. Therefore, despite the increased binding affinity, corilagin would potentially be less effective when the N501Y mutation is present.

For **E484K/N501Y**, corilagin dissociates from its initial position at around 80 ns, then associates again at around 90 ns, but at a completely new location outside the RBM (Figure 3.1). This new docked position is characterized by a very small MMPBSA binding free energy. Moreover, Table 3.1 shows that, at this site, corilagin is not interacting with any RBD's residues found at the interface.

Overall, we find that corilagin binds to WT RBD and E484K with a high binding affinity in a position that could directly disrupt association of RBD with ACE2. However, when adding the N501Y mutation (alone or with E484K), corilagin binds to the RBD in a region outside of the interface and would likely have no direct effect on the RBD-ACE2 complex formation, reducing its interest as an inhibitor.

3.3.3.5. TGG ability to disrupt the ACE2/mutated RBD interactions. For **RBD/E484K**, the MMPBSA binding free energy is favorable, but TGG's location is different than in the WT as it interacts mainly with residues at the C-terminal end of the RBM (Table 3.1). In this position, TGG interacts with many residues forming H-bonds with ACE2 such as Gly496, Gln498, Thr500, Asn501 and Tyr505, and it even forms two H-bonds with Thr500 and Asn501.

For **RBD/N501Y**, TGG binds with an affinity similar to WT. Table 3.1 shows that, contrary to what is observed for corilagin, TGG takes position directly at the interface and makes contact with many nonpolar (Leu455, Phe456, Phe490, Pro491, Leu492 with H-bond formation) and polar (Lys458, Tyr473, Tyr489 and Gln493) residues.

For **E484K/N501Y**, TGG shows the highest binding affinity of all systems studied here with -15.1 ± 0.1 kcal/mol. At this position, TGG makes many nonpolar (Phe490 with H-bond, Leu492 and Gly496) and polar (Tyr449, Tyr489 and Gln493 with H-bond) contacts with residues found at the interface between ACE2 and RBD.

In summary, TGG binds to all RBD sequences studied here (without and with mutations) with high MMPBSA binding free energy. Moreover, for all these sequences, TGG binds to the protein at locations that could hinder the formation of ACE2/RBD complex and, potentially,

reduce the ability of SARS-Cov-2's *Spike* protein to bind to the ACE2 protein on human cells.

3.4. Discussion

The current article is constructed in two parts. In the first part, we study the therapeutic potential of two small molecules, corilagin and TGG, to disrupt the association between ACE2 and the wild-type (WT) RBD; a crucial step of the infection by the virus. To do so, we used a combination of numerical tools (molecular dockings, molecular dynamics and MMPBSA free energy calculations) and experimental tools (SPR and ELISA assay). In the second part, we study, using the same array of numerical tools, the impact of the main mutations (E484K and N501Y) of the variants of concerns (B.1.1.7, B.1.351, B.1.1.28).

3.4.1. Best ligand targeting ACE2 and the wild-type RBD

The first part of the article focus on the interactions between ACE2 and the WT RBD. Numerically, the ability of corilagin and TGG to potentially impair the association of ACE2 and WT RBD, is summarized in Figure B.14 for the A1A2-RBM segments and Figure B.15 for the HS-RBM segments; all segments containing residues involved in the ACE2/RBD association. We show that both ligands could interfere with the interaction between ACE2 and the WT RBD, more likely on the RBD's side as both ligands localization and binding energies are better than on the ACE2's side.

In order to validate our numerical conclusions and our numerical methodology, we tested the impact of corilagin and TGG on the association between ACE2 and the WT RBD using SPR and ELISA binding assay.

Our SPR experiments identify a strong association between ACE2 and the RBD, characterized by a dissociation constant of 41 nM (ACE2 with immobilized RBD) and 63 nM (RBD with immobilized ACE2) (Figure 3.5C). These values are compatible with recent experimental results from other research groups [**21**, **105**, **106**]. Both ligands are binding to RBD with dissociation constants in the low nanomolar range: 1.8 nM and 1.3 nM for corilagin and TGG, respectively (Figure 3.5A). Moreover, SPR shows that the incubation of corilagin or TGG with RBD fully inhibits its binding to immobilized ACE2 (Figure 3.5D). Consequently, our ligands have much more affinity compared to quercetin/ACE2 (K_D of 4.830 μM) and quercetin/RBD (K_D of 2.210 μM) [**106**].

In addition to SPR experiments, we also test the binding of these ligands with RBD and ACE2 using biochemical inhibition assays. The results show that the addition of TGG or corilagin reduces from 45% (0.1 μM) and up to 75% (10 μM) the binding between ACE2 and RBD (Figure 3.6A-B). Moreover, a mixture of both ligands reduces the binding by 50%, independently of the ligand concentrations (between 0.1 μM to 5 μM) (Figure 3.6C); i.e., no

synergy is observed when pre-mixing both ligands. More specifically, supporting the SPR results, we find that the ACE2/RBD binding inhibition by TGG and corilagin is mediated by their interactions with the RBD as there is no significant reduction in the ACE2/anti-ACE2 binding when adding TGG, corilagin or both (Figure 3.7A-C).

Interestingly, TGG and corilagin inhibit the interaction between the human ACE2 receptor and Spike protein RBD at 0.1 μM, which is quite low as compared to the antiviral activities of tetra-TGG against SARS CoV (EC50 4.5 μM) [**78**] and the inhibition of the binding of SARS CoV-2 spike protein RBD to ACE2 based on AlphaLISA assay (IC50 of 5.5 μM) [**107**]. The concentration to inhibit the ACE2-RBD interaction found here is also lower than the EC50 of the polyphenol resveratrol (4.48 μM) on SARS-CoV-2 replication in Vero cell culture [**108**] or the estimated concentration of quercetin i.e., greater than 25 μM [**109**].

Natural polyphenolic compounds were reported to be sources of antiviral candidates against SARS-CoV-2, e.g. in terms of coronaviral entry inhibitors, protease inhibitors and coronavirus replication inhibitors [**110**]. However, it is important to note that some polyphenolic compounds such as magnolol and rosmarinic acid could increase the activity or expression of ACE2, and therefore aggravate SARS-CoV-2 infection [**111**].

SARS CoV-2 induces death and injury of virus-infected cells and tissues which could be caused by high levels of inflammatory cytokines release as IL-1β, TNF-α and IL-6 [**112**,**113**]. Interestingly, in addition to the inhibition of the binding ACE2-RBD, TGG and corilagin possess a wide range of biologic properties including anti-inflammatory, antioxidant and low toxicity. Corilagin could efficiently reduce inflammation with the reduction of the release of pro-inflammatory cytokines TNF-α, IL-1β and IL-6 through blocking the NF-κB pathway [**114**–**116**].

Therefore, both SPR and biochemical inhibition essays show that corilagin and TGG bind to the RBD domain in a way that disrupt the ACE2/RBD interaction, in agreement with our modeling results. These findings suggest that corilagin and TGG can be useful as multi target treatment against the WT SARS CoV-2 infection.

3.4.2. The impact of the mutation of the RBD and therapeutic potential of the ligands

With the numerical protocol validated by our experimental results from SPR and ELISA assays, we also test the therapeutic potential of both molecules against three mutations (RBD/E484K, RBD/N501Y and RBD/E484K-N501Y) found in the main variants of concerns.

3.4.2.1. The impacts of RBD mutations on its structural ensemble. Recent results from experiments using pseudoparticles showed that the *Spike*-protein of the B.1.1.7, B.1.351 and B.1.1.248 mutants present no difference in terms of stability and cell entry kinetics compared

to the SARS-CoV-2 WT [**25**]. Although mutant RBDs alone in solution have not been studied in the literature, they have been heavily studied in complex with ACE2. A recent Cryo-EM structure of the RBD with the N501Y mutation in complex with ACE2 showed that there was no significant changes in terms of secondary, quaternary and binding site structures compared to the RBD-WT/ACE2 complex [**117**]. The ACE2/RBD mutant complex was also studied using numerical techniques. MD simulations and principal component analysis were realized by Nelson *et al.* [**118**] on the ACE2 complex with RBD with N501Y mutation, E484K mutation and K417N+E484K+N501Y mutations. The authors showed that the complex with the E484K mutation adopts conformations that are mostly similar to the WT, while the complex with the N501Y mutation adopts conformations that are very different than the WT. The triple mutant conformational space more closely resembles the one from the WT or the E484K mutant than the N501Y mutant. Finally, Dehury *et al.* [**119**] studied the impact of multiple alanine point mutations on the ACE2/RBD complex using MD simulations. They found that the complex was stable during their simulation with no noteworthy changes in terms of secondary and quaternary structure for all mutants tested. The highest backbone RMSD they measured was 0.33 ± 0.09 nm for the N501A mutant system compared to 0.25 ± 0.03 nm for the WT. All these results taken together shows that the mutations have little impact on the stability, the cell's entry kinetics and the structure in the complex of RBD.

These results are compatible with what we observed for the RBD in solution; the mutations have very little impact on RBD's structure in terms of backbone RMSD, secondary structure and on the solvent accessibility of crucial residues involved in interactions with ACE2 (Figure 3.8). This last result suggests that the nature of the interface with ACE2 is probably similar for the WT and three mutant systems we tested. However, only extensive free-energy calculations, that are beyond the object of this work, could provide information on the impact of these mutations on the binding affinity between the two proteins. Here, these simulations are used to identify conformational ensembles representative of the mutants in order to evaluate their impact on the binding with corilagin and TGG.

3.4.2.2. Therapeutic potential of the ligands. Since the start of the pandemic, a multitude of therapeutic techniques were developed to fight against the virus; from vaccine [**120**], to monoclonal antibody approved for emergency use by the FDA [**121**, **122**]. Although, the usage of small molecules as drugs that could be used against SARS-CoV-2 was heavily studied [**17**, **80**], no drugs were able to be designed yet for widespread and efficient usage. However, the recent spread of multiple SARS-CoV-2 variants from the United Kingdom, South Africa and Brazil, imposes a reassessment of the efficacy of currently used treatment, as well as additional effort in drugs development. Recent results for the Novavax vaccine showed that its efficacy drop significantly for the variant from South-Africa (between 50% and 60%) compared to SARS-CoV-2 WT (89%) [**120**]. Additionally, a number of experimental [**25**,

123, **124**] and numerical [**125**] results shows that SARS-CoV-2 mutants with the E484K mutation, like the one from Brazil and South-Africa, were partially, if not fully, resistant to the antibodies approved for emergency use by the FDA. Computed binding affinities between these antibodies and the RBD with the E484K mutation were heavily reduced compared to the WT [**125**].

In our study, we find that the E484K mutation does not impact the therapeutic potential, contrary to what is found for the vaccine and antibodies, of either corilagin or TGG; both bind the RBD in a location that could prevent crucial interactions with ACE2 with affinities similar to the WT. On the other hand, we find that the N501Y mutation, present in the variants from the United-Kingdom, South-Africa and Brazil, highly impairs the therapeutic potential of corilagin. When only the N501Y mutation is present, corilagin binds to the RBD with a high affinity but doesn't interact with residues crucial for interaction with ACE2. The therapeutic potential of corilagin is even lower when both the N501Y and E484K mutations are present as it is unstable on the RBM. On the other hand, the N501Y mutation, alone or in pair with the E484K mutation, doesn't impact the therapeutic potential of TGG; TGG binds to the RBD at a relevant location to disrupt its interaction with ACE2 and with a high binding affinity.

Conclusion

In this study, the combination of numerical and experimental data shows that two natural polyphenols, corilagin and 1,3,6-tri-O-galloy-β-D-glucose (TGG) could play a protective role in reducing the potency of WT SARS-CoV-2 by disrupting the S-protein-RBD/ACE2 receptor interface stability or the ability of the RBD of the S-protein to recognize the ACE2 receptor.

Combining molecular modelling, including molecular dynamics and protein-ligand docking, with SPR and ELISA assays, we demonstrate that the observed inhibition of the binding of Spike RBD to human ACE2 is caused mainly by the binding of these polyphenols to the RBD protein, with dissociation constants in the low nanomolar range: 1.8 nM and 1.3 nM for corilagin and TGG, respectively. Such preference would have the potential to limit physiological side-effects induced by the inhibition of ACE2.

In addition, we use the same numerical protocol to study the impact of RBD mutated sequences associated with three dominant variants — the B.1.1.7 variant, the B.1.351 variant, and the B.1.1.28, first identified in the United Kingdom, in South Africa and in Brazil respectively —, focusing on mutations affecting the interface : RBD/E484K, RBD/N501Y and RBD/E484K-N501Y.

Analysing the impact of the potential inhibitors by identifying docking sites using AutoDock VINA and further assessing the role of flexibility by running MD on the most stable

ligand-protein configurations, using MMPBSA to compute the binding free-energy, we show that both molecules have the potential to bind more strongly to mutants RBD than ACE2, similarly to what is observed for WT RBD. Both also bind well to the RBD/E484K mutant compared to the WT, albeit with a significantly increased binding free energy for TGG compared to corilagin (-12.8 ± 0.4 kcal/mol vs -7.2 ± 0.1 kcal/mol for WT and -10.9 ± 0.4 vs -7.0 ± 0.2 kcal/mol for RBD/E484K), values that compare well with other potential inhibitors [**126**]. For the structures with the N501Y mutant (RBD/N501Y and RBD/E484K-N501Y), corilagin's binding localization is outside of the RBM's region that is interacting with ACE2, so that the recognition of ACE2 by SARS-CoV-2's *Spike* protein could still take place. On the other hand, TGG is as effective on these mutants than it is on the WT and RBD/E484K mutant.

This work strongly supports the need for further experimental assessments to evaluate the ligands' selectivity towards the virus vs. other binding sites as well as to establish their *in vivo* behavior.

Author Contributions

Vincent Binette: Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing; **Sébastien Côté**: Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing; **Mohamed Haddad**: Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft; **Phuong Trang Nguyen**: Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft; **Sébastien Bélanger**: Investigation, Methodology; **Steve Bourgault**: Conceptualization, Funding acquisition, Investigation, Project administration, Supervision, Visualization, Writing – original draft; **Charles Ramassamy**: Conceptualization, Funding acquisition, Investigation, Project administration, Supervision, Visualization, Writing – original draft; **Roger Gaudreault**: Conceptualization, Funding acquisition, Project administration, Supervision, Writing – original draft, Writing – review & editing; **Normand Mousseau**: Conceptualization, Funding acquisition, Investigation, Project administration, Supervision, Visualization, Writing – original draft, Writing – review & editing;

Acknowledgments

The authors acknowledge support from the Natural Science and Engineering Council of Canada (NSERC) and the Fonds de recherche du Québec - Nature et technologies (FRQ-NT #275304). They are grateful to Calcul Québec and Compute Canada for generous allocation of computer time. The authors would also like to thank La Fondation Famille Lemaire, the R. Howard Webster Foundation and Fruitomed Inc for their valuable support.

3.5. Addendum: On the study's limitations

In this article, we present a combination of experimental and computational tools to study the therapeutic potential of two polyphenol molecules: corilagin and TGG. The results presented here, although promising, are still very limited. The main limitation is that both the experimental and numerical techniques are not done under realistic physiological conditions. Indeed, only a segment of the SARS-CoV-2 *spike*-protein is used (the RBD) and the entire membrane environment around ACE2 is neglected. We do not know how Corilagin and TGG might interact with many of these physiologically relevant biomolecules, absent from our experiments. Our study should be thought of as the first of many steps into the therapeutic potential of these molecules.

Chapitre 4

Repliement des protéines: une introduction

Les protéines sont des nanomachines jouant une panoplie de rôles fondamentaux pour tout organisme biologique. De la transcription et traduction de l'ADN, à la catalyse des réactions chimiques en passant par le transport membranaire, les fonctions remplies par les protéines sont aussi vastes et distinctes qu'elles sont essentielles pour la survie de la cellule. De plus, les protéines sont aussi impliquées dans de nombreuses maladies. Comme exemple, on peut nommer la protéine *Spike* du virus SARS-CoV-2 qui est fondamentale pour la propagation du virus SARS-CoV-2 à l'origine de la pandémie mondiale de COVID-19 [**2**,**3**] ou la protéine amyloïde-β associée à la maladie d'Alzheimer. Si elles peuvent être associées à l'apparition de maladies, les protéines peuvent aussi servir de traitement. En effet, de petits peptides aux propriétés thérapeutiques pourraient être cruciaux pour contrer la résistance aux antibiotiques [**127**–**129**]. Ainsi, une connaissance préalable de la structure des protéines peut se révéler cruciale pour le développement thérapeutique.

4.1. Structures des protéines

La vision classique de la biologie moléculaire stipule que c'est la structure des protéines qui est à l'origine de leurs fonctions. Comme leurs fonctions sont particulièrement importantes, l'étude de leurs structures est donc primordiale.

Les protéines sont composées de l'union de petites sous-unités similaires portant le nom d'acide aminé. Naturellement, on retrouve 20 types d'acides aminés différents. La composition de chacun d'entre eux est divisée en deux parties; la chaîne principale et la chaîne latérale. La chaîne principale, composée des atomes lourds N, Cα, C et O, est commune à tous les acides aminés. Les différents acides aminés sont connectés entre eux via la chaîne principale à l'aide de la formation d'un lien peptidique, montré à la Figure 4.1. L'arrangement local de la chaîne principale est caractérisé par trois angles dièdres qui décrivent l'orientation relative de quatre atomes (ijkl) liés par des liens atomiques. Le premier de ces angles dièdres, ω, est défini par les atomes $C_\alpha^{(i)}$-$C^{(i)}$-$N^{(i+1)}$-$C_\alpha^{(i+1)}$. À cause de la double

liaison entre l'atome d'oxygène et l'atome de carbone à l'origine de fortes contraintes stériques, cet angle est très peu flexible, prenant habituellement la valeur de 180° associée à la configuration *trans* [**1**]. Dans cette configuration, les quatre atomes formant ω sont dans le même plan, appelé plan peptidique. Les deux autres angles dièdres, ϕ, défini par les atomes $C^{(i-1)}$-$N^{(i)}$-$C_\alpha^{(i)}$-$C^{(i)}$, et ψ, défini par les atomes $N^{(i)}$-$C_\alpha^{(i)}$-$C^{(i)}$-$N^{(i+1)}$, sont beaucoup plus flexibles que ω et peuvent adopter un grand nombre de valeurs, présentées dans le célèbre diagramme de Ramachandran à la Figure 4.2.

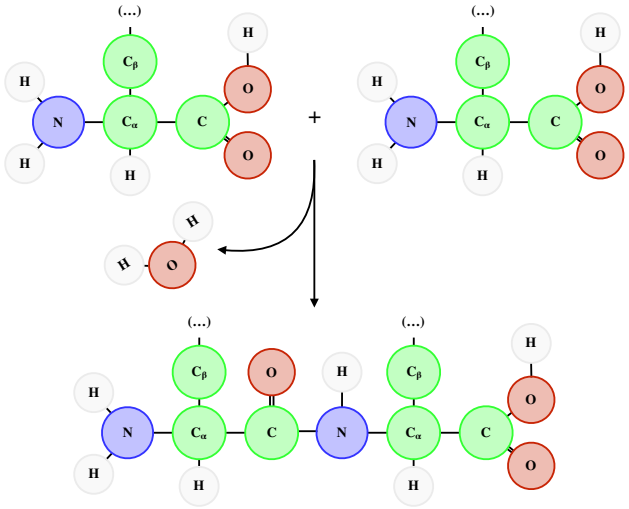


Fig. 4.1. Structure de la chaîne principale et lien peptidique. L'azote, les carbones (C, C_α), l'oxygène et l'hydrogène sont présentés respectivement en bleu, vert, rouge et gris. De haut en bas, on voit la formation du lien peptidique via la libération d'une molécule d'eau.

La chaîne latérale est quant à elle formée d'un ou plusieurs atomes connectés à la chaîne principale via le C_α. La chaîne latérale est unique pour chacun des acides aminés et ce sont les propriétés de la chaîne latérale qui donnent les caractéristiques distinctives à chacun des acides aminés. La Figure 4.3 présente les différentes chaînes latérales.

Plus globalement, la structure des protéines est divisée en plusieurs niveaux d'organisation [**1**], désignés par la structure *primaire*, *secondaire*, *tertiaire* et *quaternaire*.

La **structure primaire** correspond à la séquence linéaire des acides aminés composant une protéine. Selon l'hypothèse thermodynamique, postulée par Christian B. Anfinsen *et coll.* au début des années 1960 [**130**], ce sont les propriétés physico-chimiques des acides aminés qui encodent la structure globale de la protéine, ou, en d'autres mots, la connaissance de la structure primaire serait suffisante pour déterminer la structure globale. L'hypothèse thermodynamique est cruciale dans les méthodes de prédiction et nous y reviendrons donc par la suite.

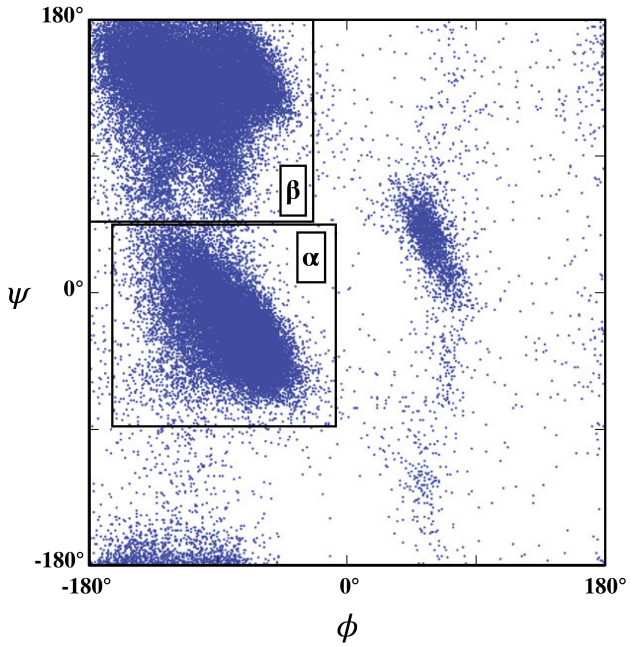


Fig. 4.2. Diagramme de Ramachandran. La figure est adaptée du livre de Tamar Schlick [**1**]. Les zones appelées α et β correspondent respectivement aux paires d'angles dièdres ϕ/ψ associées aux structures secondaires d'hélice-α et de feuillet β respectivement.

La **structure secondaire** correspond au repliement local des acides aminés qui mène à l'obtention de motifs communs à de nombreuses protéines. On dénote deux classes principales: les hélices (α et dans une moindre mesure les 310 et les π) et les feuillets β (parallèles ou anti-parallèles). Celles-ci sont présentées à la Figure 4.4. Ces deux types de structure secondaire possèdent des angles de torsion ϕ/ψ, présentés à la Figure 4.2, et un réseau de ponts-hydrogène, présentés à la Figure 4.4, caractéristiques.

La **structure tertiaire** correspond quant à elle au repliement global, ou tridimensionnel, de la protéine. Ce repliement mène à la formation de motifs de grandes échelles entre les différents éléments de la structure secondaire: protéine α, β, α/β ou $\alpha+\beta$ [**1**].

Finalement, de multiples protéines peuvent s'associer les unes avec les autres pour former des complexes, aussi appelés **structure quaternaire**.

Les structures tertiaire et quaternaire sont associées à la fonction de la protéine. Des exemples classiques sont l'hémoglobine qui est un assemblage de quatre sous-unités, quatre protéines, permettant le transport d'oxygène ou les canaux potassiques qui sont des tétramères assemblés autour d'un pore permettant aux ions de franchir la membrane cellulaire [**1**].

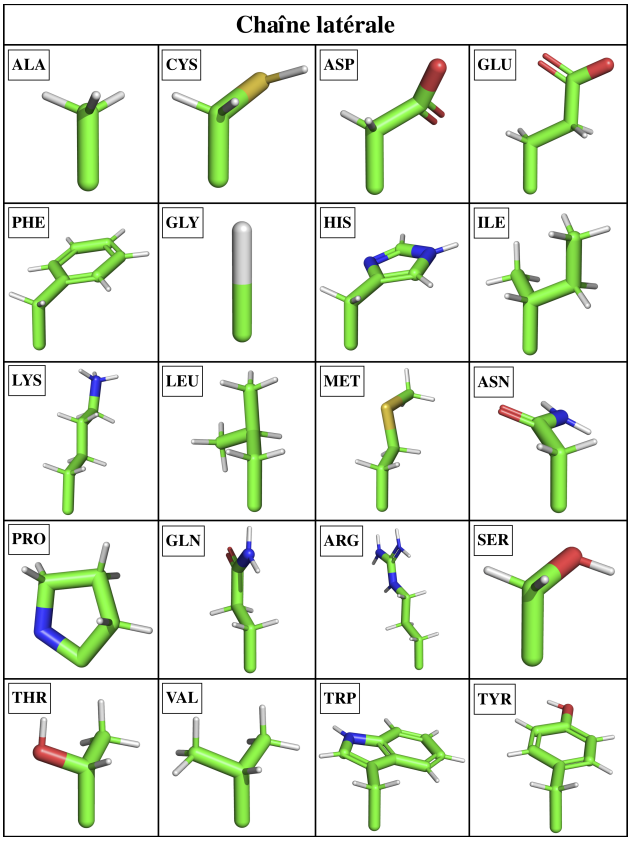


Fig. 4.3. Structure des chaînes latérales. La chaîne latérale (et le carbone-α) de chacun des acides aminés. Les atomes de carbones, d'oxygène, d'azote et d'hydrogène sont présentés respectivement en vert, rouge, bleu et blanc. Les images furent réalisées avec PYMOL [**95**].

4.2. Repliement des protéines

Le repliement des protéines décrit le processus par lequel la structure tridimensionnelle (structure tertiaire ou quaternaire) est obtenue à partir de la séquence d'acides aminés (structure primaire). L'étude de ce phénomène est d'une grande importance puisque la structure d'une protéine est associée à sa fonction.

Le repliement des protéines est gouverné par les lois de la physique. L'hypothèse thermodynamique stipule que la structure native d'une protéine, c'est-à-dire la structure tridimensionnelle qu'elle adopte en condition physiologique, correspond à un minimum global de l'énergie libre [**131**, **132**]. Une fois cela mentionné, la question du repliement des protéines demeure extraordinairement complexe comme le met en évidence le fameux paradoxe de Levinthal [**133**] développé par Cyrus Levinthal à la fin des années 1960. Cette expérience de la pensée vise à mettre en évidence la quantité incroyable de conformations accessibles aux protéines. Prenons une protéine de 100 acides aminés, soit 99 liens peptidiques (99 angles dièdres ϕ et 99 angles dièdres ψ). Supposons maintenant que chacune des paires ϕ/ψ puisse prendre trois valeurs distinctes. Alors, le nombre total de conformations différentes que peut

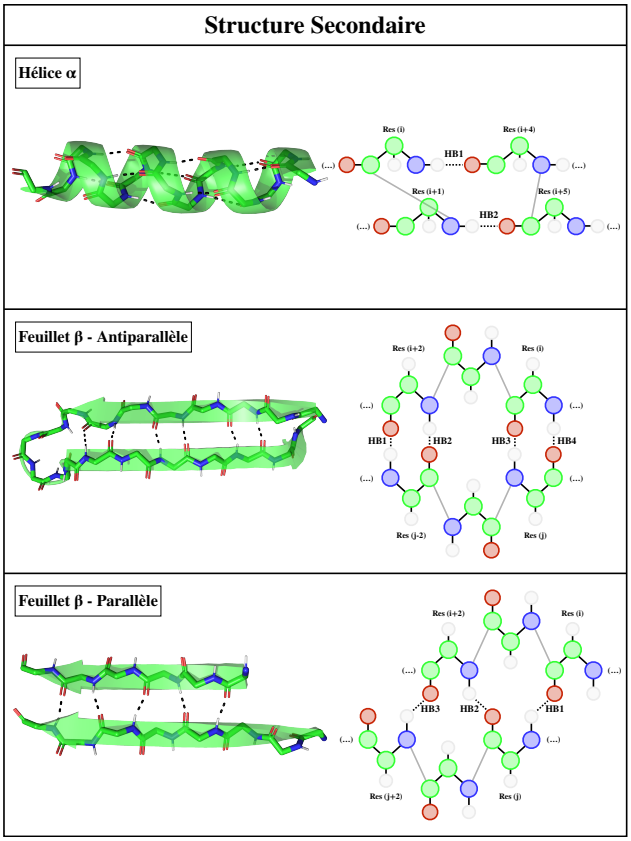


Fig. 4.4. Schéma des principales structures secondaires. De haut en bas, on retrouve les hélices α, les feuillets β antiparallèles et les feuillets β parallèles. À gauche, on retrouve un schéma de la structure, tiré de la protéine 1PGB, et réalisé avec PYMOL [**95**]. À droite, on retrouve les ponts-H, représentés par des lignes pointillées, caractéristiques de chacune de ces structures. Seuls les atomes de la chaîne principale sont présentés.

prendre la protéine est de $3^{99} \approx 1 \times 10^{47}$. Si chacune des conformations peut-être évaluée rapidement, de l'ordre de la nanoseconde, alors trouver le minimum d'énergie libre via l'exploration exhaustive de chacune des conformations accessibles prend un temps bien supérieur à l'âge de l'univers ($\approx 1 \times 10^{17}$ s). Or, le repliement des protéines se fait généralement sur des échelles de temps de l'ordre de la milliseconde à la seconde [**1**]. Très clairement, le paysage d'énergie libre sous-jacent le repliement des protéines doit prendre une forme particulière pour éviter un tel paradoxe.

Revenons à nouveau à l'hypothèse thermodynamique afin de discuter de certains points importants dans la compréhension du paysage d'énergie libre associé au repliement. (1) La structure native est *unique*, c'est-à-dire qu'aucune autre structure n'est comparable en termes d'énergie libre. (2) La structure native est *stable*, c'est-à-dire qu'elle n'est pas affectée par des changements modestes de l'environnement, ou, pour le mettre en termes énergétique, que les barrières autour du minimum sont élevées et abruptes. Finalement, (3) la structure native est *accessible*, c'est-à-dire que le chemin au niveau du paysage d'énergie libre pour

arriver à la structure native de la structure désordonnée est relativement lisse. Toutes ces considérations sont englobées dans un modèle décrivant le paysage d'énergie libre du repliement sous la forme d'un entonnoir [**131**, **132**], comme présenté à la Figure 4.5. Ce paysage d'énergie libre est caractérisé par un très grand nombre de structures désordonnées sans les interactions caractéristiques de la structure native (haute entropie, haute énergie) et un beaucoup plus petit nombre de structures natives formant un grand nombre d'interactions favorables, mais étant peu flexibles (basse entropie, basse énergie). La forme du potentiel est telle que, majoritairement, les étapes du repliement apportent une diminution de l'énergie en descendant le long de l'entonnoir [**131**, **132**].

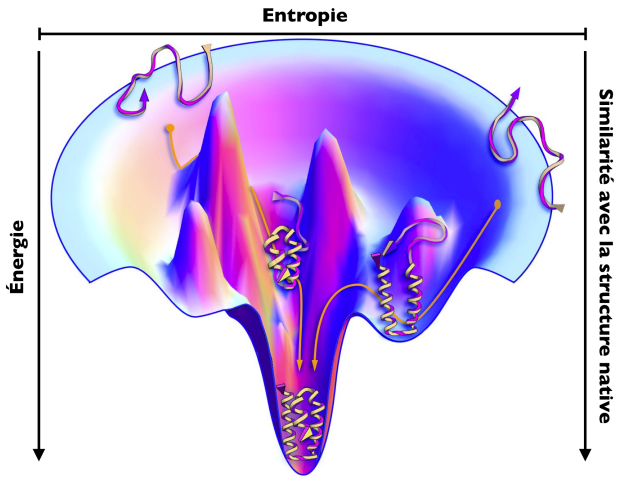


Fig. 4.5. Paysage d'énergie libre en forme d'entonnoir. L'image est adaptée de l'article de Ken A Dill *et coll.* [**131**]. Paysage énergétique caractéristique du repliement des protéines. Les niveaux d'entropie et d'énergie sont présentés respectivement horizontalement et verticalement.

4.3. Similitude structurelle

La question de comment quantifier les similarités entre deux structures peut sembler de prime abord simple, mais est en réalité particulièrement complexe. Dans cette section, je décrirai quelques méthodes pour la quantification des similarités structurelles.

4.3.1. RMSD

Une des méthodes les plus connues pour quantifier les similarités structurelles est la racine de l'erreur quadratique moyenne ou RMSD pour "root mean square deviation". Le RMSD est défini comme étant la somme des carrés des distances entre les mêmes atomes d'une structure modèle (M) à une structure de référence (T). Mathématiquement, le RMSD est

défini selon l'équation:

$$\mathrm{RMSD}(M,T) = \frac{1}{N}\sum_{i}^{N} |\mathbf{R}\mathbf{M}_i - \mathbf{T}_i|^2 \tag{4.3.1}$$

où N est le nombre d'atomes à comparer entre les deux structures et $\mathbf{R}$ est une matrice orthonormale 3×3 qui fait la rotation des coordonnées des atomes $\mathbf{M}_i$ sur $\mathbf{T}_i$. La mesure du RMSD requiert préalablement la superposition entre la structure modèle (M) et la structure de référence (T) de telle sorte que le RMSD soit minimisé. La première étape de cette superposition est le positionnement des centroïdes de M et T à l'origine du système de coordonnées. Par la suite, on doit déterminer la matrice R qui minimise le RMSD, soit l'équation 4.3.1. Pour ce faire, on utilise l'algorithme de Kabsch [**134**, **135**] qui utilise la méthode des multiplicateurs de Lagrange pour déterminer la matrice $\mathbf{R}$ optimale. Ainsi, la matrice de rotation $\mathbf{R}$ est donnée par:

$$\mathbf{R} = \mathbf{W}\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & d \end{bmatrix}\mathbf{V}^\top$$

où d est le signe du déterminant de la matrice de covariance $\mathbf{C} = \mathbf{M}\mathbf{T}^\top$ et $\mathbf{W}$ et $\mathbf{V}$ sont des matrices orthonormales déterminées à partir de la décomposition en valeurs singulières de $\mathbf{C}$.

Le RMSD a une borne inférieure de 0, si M et T sont identiques et aucune borne supérieure. Si le RMSD est une méthode classique pour quantifier les similarités entre deux structures, il possède un nombre de propriétés problématiques. Premièrement, le RMSD est informatif uniquement s'il est relativement petit et distribué de façon uniforme au niveau de la structure [**136**, **137**]. En effet, le RMSD est particulièrement sensible aux grandes déviations locales, même si elles sont localisées dans des régions plus flexibles comme les extrémités N/C-terminales. Ainsi une amplitude "moyenne" du RMSD peut difficilement être associée à une bonne ou mauvaise similarité structurelle [**136**, **137**]. Deuxièmement, le RMSD ne favorise pas nécesssairement l'obtention de modèles avec la bonne stéréochimie ou les bonnes interactions atomiques, comme les ponts-H [**138**]. Troisièmement, le RMSD requiert une étape de superposition, ce qui ralentit son calcul.

Pour contrer certains problèmes du RMSD, divers autres scores de quantification des similarités structurelles furent développés. Deux de ceux-ci, le BC-score et le CAD-score, sont présentés dans ce qui suit.

4.3.2. BC-score

Le BC-score [**137**] est basé sur la formule de Binet-Cauchy. Plus spécifiquement, le BC-score est basé sur la comparaison entre les volumes des tétraèdres formés par les trios

d'atomes C_α et son centroïde. Un de ces tétraèdres est présenté à la Figure 4.6. Le volume d'un tel tétraèdre est donné par le déterminant d'une matrice 3×3, X, contenant les coordonnées atomiques de chacun des C_α selon: $V = \frac{1}{6} \det X$. Plus spécifiquement, le BC-score est défini avec:

$$BC(M,T) = \frac{\det\left(M^\top T\right)}{\sqrt{\det\left(M^\top M\right) \det\left(T^\top T\right)}}$$

où M et T sont des matrices $N \times 3$ contenant respectivement les coordonnées des atomes du C_α de la structure modèle et de la structure de référence.

Le BC-score est borné entre -1, correspondant à la situation où M est l'image miroir de T et 1, correspondant à la situation où M est identique T (à une déformation près). Un des avantages du BC-score est qu'il est indépendant de la rotation. Ainsi, contrairement au RMSD, il n'est pas nécessaire de calculer la matrice de rotation R permettant la superposition des structures; le calcul du BC-score sur une large base de données (cas test) est plus de 50 fois plus rapide que le calcul du RMSD [**137**]. Finalement, contrairement au RMSD, le BC-score est indépendant de la taille des protéines étudiées.

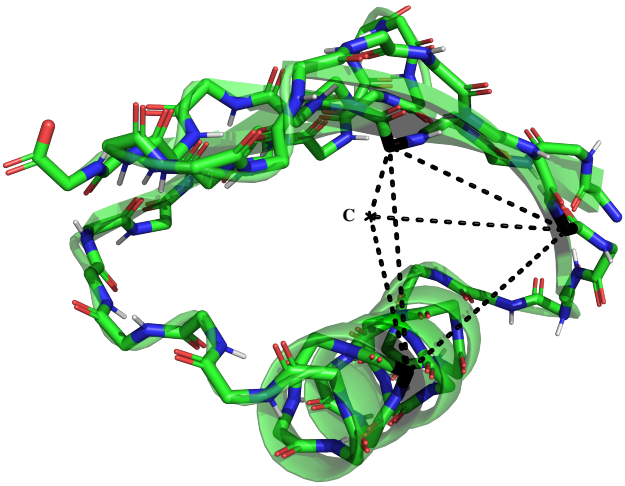


Fig. 4.6. Exemple d'un tétraèdre à la base du calcul du BCscore. Le tétraèdre est formé par trois carbones-α et le centre géométrique de tous les carbones-α. Cette figure présente, pour la protéine 1PGB, un tétraèdre formé par le centre géométrique des carbones α (C) et les carbones α des résidus LEU5, THR18 et TYR32.

4.3.3. CAD-score

Le repliement des protéines se fait sous l'action des interactions atomiques ou, en d'autres mots, des différents contacts au niveau des résidus de la protéine. Cette idée est derrière le développement du CAD-score, pour **C**ontact **A**rea **D**ifference-based score [**136**]. Avec ce score, la similarité structurelle est déterminée en fonction de la quantité de contacts présents

dans la structure de référence qui est reproduite par la structure modèle. Ainsi, on définit le CAD comme étant:

$$\mathrm{CAD}_{(i,j)} = \left| T_{(i,j)} - M_{(i,j)} \right|$$

où (i,j) indique toutes les paires de résidus dont la surface de contacts dans la structure de référence, $T_{(i,j)}$, est non-nulle. Pour imposer la symétrie entre la sur-prédiction et la sous-prédiction des contacts, on utilise plutôt la mesure bornée:

$$\mathrm{CAD}^{borne}_{(i,j)} = \min\left(\mathrm{CAD}_{(i,j)}, T_{(i,j)}\right)$$

En utilisant cette valeur, le CAD-score est alors défini comme étant la somme de tous les contacts de la structure de référence $T_{(i,j)}$ qui sont reproduits par la structure modèle $M_{(i,j)}$ normalisée par la quantité totale de contacts dans la structure de référence. Mathématiquement, il est défini comme:

$$\text{CAD-score} = 1 - \frac{\sum_{(i,j)\in G} \mathrm{CAD}^{borne}_{(i,j)}}{\sum_{(i,j)\in G} T_{(i,j)}}$$

où G correspond à toutes les paires de résidus en contact dans la structure de référence (T). Ainsi, le CAD-score est borné entre 0, dans le cas où aucun contact de la structure de référence n'est reproduit dans la structure modèle et 1, dans le cas où tous les contacts de la structure de référence sont reproduits identiquement par la structure modèle. Les contacts entre les résidus sont déterminés via la construction d'un diagramme de Voronoi [**136**, **139**, **140**]. Dans cette méthode, une cellule de Voronoi est générée autour de chacun des atomes lourds du système. Pour un atome particulier, cette cellule est définie par tous les points qui sont plus près de cet atome que de n'importe quel autre. Deux atomes sont considérés en contact si leurs cellules de Voronoi partagent un certain nombre de points et la surface de contacts entre les deux atomes est définie par la surface commune des cellules de Voronoi. Comme ce calcul se fait au niveau atomique, il faut uniquement combiner les atomes de chacun des résidus pour obtenir la surface de contacts inter-résidus requis pour le calcul du CAD-score.

Une étude comparative entre le CAD-score et d'autres scores populaires a montré que le CAD-score possède plusieurs propriétés uniques et intéressantes [**138**]. Premièrement, le CAD-score favorise les structures adoptant la bonne stéréochimie (liens atomiques, angles atomiques, angles dièdres, collisions entre atomes) selon MolProbity [**141**], comparativement aux autres scores comme le RMSD. Deuxièmement, le CAD-score promeut une meilleure reproduction des ponts hydrogène. Troisièmement, le CAD-score est peu sensible aux déviations dans les régions flexibles comme les extrémités N/C-terminales.

4.4. Prédiction structurelle

Dans les sections précédentes, les bases de la structure des protéines, et la façon dont la structure est liée à la fonction, et la théorie générale associée au repliement des protéines furent présentées. Dans celle-ci, je discuterai des méthodes numériques de prédictions *de novo* de la structure tridimensionnelle des protéines.

Un des objectifs principaux de la bio-informatique est la prédiction de la structure tridimensionnelle des protéines à partir uniquement de sa séquence en acides aminés aussi appelée prédiction *de novo*. Les progrès associés à cet objectif sont évalués lors de la rencontre bisannuelle CASP [**142**].

Les algorithmes de prédictions structurelles sont généralement basés sur deux éléments clés: (1) une fonction de score qui permet de discriminer les "bonnes" prédictions des "mauvaises" et (2) une méthode d'échantillonnage permettant d'explorer l'espace défini par la fonction de score.

Avec les années, de multiples méthodes furent développées pour la prédiction des protéines et des peptides, utilisant chacune une approche distincte au niveau de la fonction de score et/ou de la méthode d'échantillonnage. Certaines méthodes sont basées sur des observations/paramètres dérivés des lois de la physique. Par exemple, la méthode PEPstr [**143**] (et son extension aux acides aminés non-standards, PEPstrMOD [**144**]) fait la prédiction de la structure des peptides grâce à l'observation de la prévalence des coudes β chez ceux-ci. Cette observation est ensuite convertie en une série de contraintes qui sont ajoutées lors de simulations de dynamique moléculaire. La suite-AWSEM [**145**] utilise comme fonction de scores le potentiel gros-grains AWSEM [**146**] en plus de contraintes tirées de données d'homologie et de co-évolution lors de simulations de recuit simulé. Finalement, une méthode comme PEP-FOLD [**147**, **148**], qui sera décrite en détail à la section 5.3, utilise le principe d'alphabet structurel combiné à un potentiel gros-grain, sOPEP, pour la simulation des petits peptides.

Dans les dernières années, diverses méthodes utilisent les avancées de l'apprentissage machine, et plus spécifiquement l'apprentissage profond, afin de prendre avantage de l'énorme quantité de données de co-évolution, de données structurelles etc. [**149**–**151**]. La méthode RaptorX [**152**–**154**] utilise un réseau de neurones, ResNet, pour, à partir des données d'alignement, dériver les distributions des distances et orientations interatomiques. Ce potentiel est ensuite minimisé à l'aide d'un algorithme du gradient. De façon très similaire, la méthode AlphaFold2 [**12**], développée par Google, a obtenu des résultats d'une qualité inégalée lors de la rencontre CASP 2020. Pour une séquence donnée, un réseau de neurones est utilisé pour dériver les distances interatomiques et les angles dièdres. Ce potentiel est ensuite minimisé à l'aide d'un algorithme du gradient. Finalement, la méthode APPTest [**155**], spécialisée pour les petits peptides, utilise un réseau de neurones afin de dériver des contraintes au niveau

des distances interatomiques et des angles dièdres. Le potentiel est ensuite minimisé à l'aide de simulations de recuit simulé.

Chapitre 5

Méthode pour l'amélioration de PEP-FOLD

La prédiction structurelle des protéines et des petits peptides est un des principaux objectifs de la bio-informatique. Dans ce chapitre, nous nous intéresserons à une de ces méthodes, PEP-FOLD, dont la version améliorée fait l'objet de l'article du chapitre 6.1. PEP-FOLD est une méthode simplifiée pour la prédiction structurelle des petits peptides. PEP-FOLD est composé de deux éléments cruciaux: un alphabet structurel et un potentiel gros-grain. Ce potentiel gros-grain s'appelle sOPEP et fait partie de la famille des potentiels OPEP ("Optimized Potential for Efficient peptide-structure Prediction"), des potentiels gros-grain dont le développement fut initié à la fin des années 90 par Philippe Derreumaux [**156**].

Le présent chapitre sera divisé en trois sections et présentera les bases théoriques nécessaires pour comprendre la méthode PEP-FOLD, tout particulièrement en lien avec les améliorations réalisées dans l'article présenté au chapitre suivant. Dans un premier temps, je ferai une brève introduction aux potentiels gros-grain et présenterai deux notions théoriques fondamentales: la fonction de distribution radiale et le potentiel de champ moyen. Dans un second temps, je ferai l'historique des potentiels gros-grain OPEP, dont le développement s'est fait en parallèle avec la méthode PEP-FOLD. Plus spécifiquement, le potentiel sOPEP, de la famille OPEP, joue un rôle fondamental à l'intérieur de PEP-FOLD. Les divers éléments-clés de OPEP, pertinents pour sOPEP, seront présentés. Finalement, la dernière section portera directement sur PEP-FOLD, notamment par la présentation détaillée de la méthode, et mettra l'accent sur les améliorations apportées et présentées au chapitre suivant.

5.1. Quelques notions théoriques

5.1.1. Potentiel gros-grain

Les systèmes biologiques d'intérêt couvrent de vastes dimensions, tant au niveau spatial que temporel. Malgré les avancées au niveau technologique, l'étude de ces systèmes requiert parfois l'utilisation de méthodes numériques plus simples et plus rapides que les méthodes

numériques classiques. En effet, nous avons présenté au chapitre précédent l'étude de deux protéines (RBD et ACE2) et de petites molécules thérapeutiques à l'aide de méthodes numériques classiques. Pour un tel système, en incluant l'ajout du solvant pour décrire les conditions physiologiques, la modélisation inclut déjà plus de 180 000 atomes. Or, ce système est déjà une représentation assez simplifiée de la réalité; le RBD est un fragment de la protéine réelle (protéine *Spike*), ACE2 est une protéine membranaire, etc. Ainsi, pour pouvoir étudier des systèmes plus grands sur des échelles de temps plus longues, il faut développer des méthodes simplifiées.

Une des stratégies de simplification est le développement de potentiel gros-grain. Ces potentiels regroupent divers atomes du système en un seul "atome effectif" et permettent de focaliser sur les caractéristiques d'intérêt et d'intégrer les autres. Dans son article [**157**], William G. Noid présente trois philosophies différentes pour le développement de potentiels gros-grain. Les modèles gros-grain "bottom-up" sont basés sur des résultats empiriques obtenus à partir de modèles plus complexes, par exemple, les résultats de simulations avec un potentiel tout-atome. Contrairement aux modèles "bottom-up", les modèles gros-grain "top-down" sont plutôt développés afin de reproduire les phénomènes d'intérêt observés expérimentalement. Finalement, les potentiels "knowledge-based" sont dérivés des statistiques observées au niveau des structures obtenues expérimentalement, c'est-à-dire avec les données de la *Protein Data Bank*. Cette dernière philosophie contraste avec les deux premières, plutôt basées sur des lois physiques: tirées de la paramétrisation des méthodes complexes de référence pour "bottom-up" ou directement des méthodes expérimentales pour "top-down" [**157**].

Les notions de fonction de distribution radiale et de potentiel de champ moyen sont cruciales pour le développement des potentiels "bottom-up" et "knowledge-based". Ainsi, elles seront présentées dans les deux sections suivantes.

5.1.2. Fonction de distribution radiale

La fonction de distribution radiale, $\mathrm{RDF}_{AB}(r)$ décrit la probabilité de retrouver une particule de type B à une distance r d'une particule de type A relativement à une distribution uniforme de B [**28**, **43**]. La fonction de distribution radiale se calcule à partir de l'équation suivante:

$$\begin{aligned}\mathrm{RDF}_{AB}(r) &= \frac{\langle \rho_B(r) \rangle}{\langle \rho_B \rangle_{local}} \\ &= \frac{1}{\langle \rho_B \rangle_{local}} \frac{1}{N_A} \sum_{j \in A}^{N_A} \sum_{j \in B}^{N_B} \frac{\delta\left(r_{ij} - r\right)}{4\pi r^2}\end{aligned}$$

où $\langle \rho_B(r) \rangle$ est la densité de particule de type B à une distance r et $\langle \rho_B \rangle_{local}$ est la densité de particule de type B sur tout l'espace considéré. Pour calculer la fonction de distribution radiale à partir des résultats de simulation, on peut diviser l'espace en une série de coquilles

dont le rayon est compris entre r et dr et calculer l'histogramme associé, normalisé par une distribution uniforme, qui elle est proportionnelle à l'aire de ladite coquille considérée $4\pi r^2$.

5.1.3. Potentiel de champ moyen

Les systèmes hautement multidimensionnels, comme les systèmes biologiques, sont particulièrement complexes à étudier et à visualiser. Le potentiel de champ moyen ("potential of mean force" (PMF)) est une méthode permettant d'étudier/visualiser ces systèmes complexes. Plus spécifiquement, il s'agit des distributions projetées sur un nombre restreint de coordonnées d'intérêt. Le PMF est aussi appelé profil d'énergie libre, puisque les facteurs de Boltzmann associés au PMF donnent les probabilités d'observation le long des coordonnées d'intérêt [**28**]. Mathématiquement parlant, le PMF est défini par les facteurs de Boltzmann comme suit:

$$\rho(x,y,\cdots) = \exp -\beta \cdot \mathrm{PMF}(x,y,\cdots) \tag{5.1.1}$$

où $\beta = \frac{1}{k_B T}$ Le PMF peut se calculer aisément à partir d'une simulation numérique (avec un bon échantillonnage) en prenant le logarithme de l'histogramme déterminé le long de variables d'intérêt, comme la RDF.

Les PMFs sont particulièrement importants dans le développement des potentiels *knowledge-based*. Dans ce type de potentiel, les fonctions énergétiques et les paramètres sont dérivés des PMFs obtenus à partir de la RDF calculée sur une banque de protéine obtenue expérimentalement comme la *Protein Data Bank*.

Il est par contre important de mentionner que la RDF, et par la même occasion le PMF dérivé de celle-ci, représente rarement le véritable potentiel sous-jacent à la génération de la RDF. Par exemple, la RDF obtenue pour un liquide dont les interactions sont décrites par un simple potentiel de Lennard-Jones est caractérisée par de multiples maxima, et ce, même si le potentiel sous-jacent est caractérisé par un seul minimum en r_0 [**28**].

5.2. Historique des potentiels OPEP

La famille de potentiel OPEP est un des éléments-clés de la méthode PEP-FOLD pour la prédiction structurelle des petits peptides. Les potentiels OPEP sont des potentiels gros-grain hybrides. En effet, ils sont en partie "knowledge-based", c'est-à-dire basés sur les statistiques de la *Protein Data Bank*, et en partie "physic-based", c'est-à-dire dérivés d'interactions physiques.

Depuis sa conception, la famille de potentiel OPEP a été appliquée avec succès pour l'étude de phénomènes biophysiques variés [**147**, **148**, **158**–**164**]. Une des applications principales de OPEP est l'étude des protéines amyloïdes, tout particulièrement de la protéine amyloïde-β (Aβ) impliquée dans la maladie d'Alzheimer. Avec OPEP, des études sur Aβ

permirent de mettre en évidence les conformations distinctes entre Aβ_{40} et Aβ_{42} tant au niveau du monomère [**159**] que du dimère [**160**] et d'étudier la taille du noyau critique [**164**] d'Aβ avec l'étude de petits fragments d'Aβ allant jusqu'au décamère. De plus, OPEP fut utilisé dans un protocole de simulation hiérarchique, alliant simulations gros-grain avec OPEP, amarrage moléculaire et simulations tout-atome, pour l'étude de molécules thérapeutiques agissant sur Aβ [**161**]. OPEP fut aussi utilisé avec succès pour l'étude de systèmes ADN/ARN [**158**], l'amarrage peptide/protéine [**163**] et l'étude des effets hydrodynamiques [**162**]. Finalement, une des applications les plus pertinentes pour cette thèse est l'utilisation d'une variante de OPEP, sOPEP, pour la prédiction *de novo* de la structure des peptides à l'aide de la méthode PEP-FOLD [**147**, **148**].

5.2.1. OPEPv1

La première version du potentiel gros-grain OPEP a été développée par Philippe Derreumaux [**156**] pour la prédiction structurelle de petites protéines à l'aide de simulation de Monte-Carlo. La représentation simplifiée de OPEPv1 décrit chacun des acides aminés par un centre d'interaction, pour ALA et PRO, ou deux centres d'interaction, un pour la chaîne principale et un pour la chaîne latérale, pour tous les autres acides aminés. Pour ce qui est du potentiel d'OPEPv1, l'énergie totale est déterminée à partir de sept termes distincts; (1) les interactions liées (E_L) incluant les liens atomiques, les angles de valences et les angles dièdres impropres, (2) les interactions d'attraction/répulsion, incluant un terme entre les chaînes latérales ($E_{SC,SC}$) et la chaîne principale (E_{C_α,C_α}), (3) les ponts hydrogènes, qui sont décrits de façon explicite (E_{HB1}), (4) la coopérativité entre des ponts hydrogènes des structures secondaires d'hélice α et de feuillet β (E_{HB2}), (5) la propension pour chacun des résidus à adopter une structure secondaire α (E_P^α) ou β (E_P^β), (6) une pénalité pour les résidus précédents la proline E_{X-P} et (7) les ponts di-sulfure (E_{ss}). L'énergie totale prend donc la forme de l'équation suivante:

$$\begin{aligned} E =& w_H E_{HB1} + w_{HH} E_{HB2} + w_L E_L + w_{SC} E_{SC,SC} + w_A E_{C_\alpha,C_\alpha} + \\ & \sum_{20} w_P^\alpha E_P^\alpha + \sum_{20} w_P^\beta E_P^\beta + w_P E_{X-P} + w_{ss} E_{ss} \end{aligned}$$

Dans cette équation, les paramètres w, 47 au total, sont des poids permettant d'ajuster la balance relative entre chacun des termes énergétiques. Ces poids furent optimisés afin de maximiser les différences d'énergie entre les conformations natives et non-natives pour quatre petits peptides; un α, un β, un $\alpha\beta$ et un $\beta\beta\alpha$. Le potentiel fut ensuite testé sur un ensemble de 20 petites protéines à l'aide de simulation de Monte-Carlo. OPEPv1 a permis de discriminer entre les structures natives et non-natives générées pour ces protéines tests en plus de générer des ensembles conformationnels compatibles avec la structure expérimentale.

5.2.2. OPEPv3

La troisième version de OPEP, OPEPv3 [**165**], apporte une panoplie de nouveautés, tant au niveau de la représentation gros-grain que de la forme du potentiel, tout en conservant certaines idées originales à OPEPv1 [**156**].

Un premier changement significatif entre OPEPv1 et OPEPv3 est au niveau de la représentation gros-grain. Avec OPEPv3, la forme moderne de la représentation gros-grain de OPEP fut développée. La chaîne principale est représentée en tout-atome avec un centre d'interaction pour respectivement les atomes N, H, Cα, C et O. La chaîne latérale est quant à elle représentée par un seul centre d'interaction, sauf pour la proline dont la chaîne latérale est en tout-atome. La représentation gros-grain de OPEPv3 est présentée à la Figure 5.1. Le positionnement de chacun des pseudo-atomes correspondant à la chaîne latérale fut dérivé à partir du centre de masse des atomes lourds calculé sur l'ensemble des rotamères. Pour ce qui est des rayons de chacun des pseudo-atomes, ils furent dérivés à partir des distributions des distances entre paires de chaînes latérales en contact extraites d'un ensemble de structures tirées de la *Protein Data Bank*. Plus spécifiquement, les valeurs des rayons sont optimisées pour minimiser un facteur de moindre carré avec la moyenne de ces distributions pour chacune des paires [**166**].

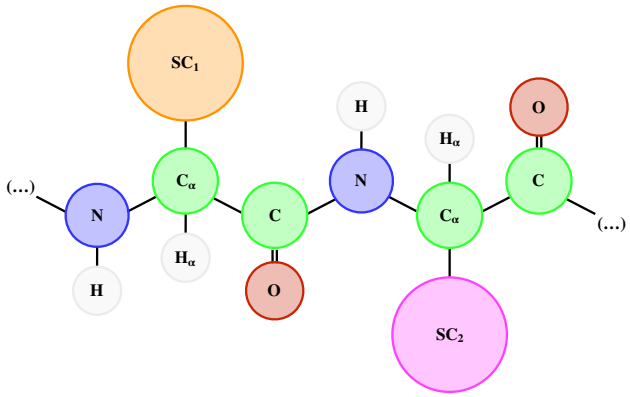


Fig. 5.1. Représentation gros-grain des potentiels OPEP. La chaîne principale est représentée en tout-atome. Les atomes d'azote, d'oxygène, d'hydrogène et de carbone sont présentés respectivement en bleu, en rouge, en gris et en vert. La chaîne latérale est représentée par un seul centre d'interaction, présenté ici en orange et en magenta.

Au niveau de la fonction d'énergie, OPEPv3 conserve une grande partie des particularités d'OPEPv1; des ponts hydrogènes explicites, un terme de coopérativité pour les ponts hydrogènes impliqués dans les hélices α et les feuillets β et une propension de structures secondaires par résidus. Globalement, le potentiel OPEPv3 est composé des interactions liées et non-liées.

Les interactions liées comprennent un terme associé aux liens atomiques, aux angles de valence (ϕ et ψ), aux angles dièdres et aux angles dièdres impropres et décrites par l'équation:

$$
\begin{aligned}
E_{\text{lié}} &= E_{\text{lien}} + E_{\text{angle}} + E_{\text{imp}} + E_{\phi} + E_{\psi} \\
E_{\text{lien}}(r_{ij}) &= w_{\text{lien}} \sum_{\text{lien}} k^{b}_{ij} \left(r_{ij} - r^{0}_{ij} \right)^2 \\
E_{\text{angle}}(\theta_{ijk}) &= w_{\text{angle}} \sum_{\text{angle}} k^{\theta}_{ijk} \left(\theta_{ijk} - \theta^{0}_{ijk} \right)^2 \\
E_{\text{imp}}(\omega_{ijkl}) &= w_{\omega} \sum_{imp} k^{\omega}_{ijkl} \left(\omega_{ijkl} - \omega^{0}_{ijkl} \right)^2 \\
E_{\phi}(\phi_{ijkl}) &= w_{\phi} \sum_{\phi} k^{\phi}_{ijkl} \left(\phi_{ijkl} - \phi^{0}_{ijkl} \right)^2 \\
E_{\psi}(\psi_{ijkl}) &= w_{\psi} \sum_{\psi} k^{\psi}_{ijkl} \left(\psi_{ijkl} - \psi^{0}_{ijkl} \right)^2
\end{aligned}
$$

Les énergies associées aux liens (E_{lien}), aux angles (E_{angle}) et aux angles dièdres impropres (E_{imp}) sont simplement décrites par un potentiel harmonique autour de la valeur d'équilibre $r^0_{ij}/\theta^0_{ijk}/\omega^0_{ijkl}$ avec une constante de rappel k. Pour ce qui est des angles dièdres ϕ/ψ, le potentiel présenté ci-dessus est en réalité un potentiel à fond plat où les valeurs de ϕ^0/ψ^0 sont données par:

$$
\phi^0/\psi^0 = \begin{cases} \phi/\psi & \text{si } \phi_{\min}/\psi_{\min} \leq \phi/\psi \leq \phi_{\max}/\psi_{\max} \\ \min\left(\phi/\psi - \phi_{\min}/\psi_{\min}, \phi/\psi - \phi_{\max}/\psi_{\max}\right) & \text{sinon} \end{cases}
$$

où $\phi_{\min}/\psi_{\min}$ et $\phi_{\max}/\psi_{\max}$ sont les bornes délimitant la taille du fond plat.

Pour ce qui est des interactions non-liées, elles se divisent en deux classes; les interactions de van der Waals et les ponts hydrogène.

Dans OPEPv3, les interactions de van der Waals sont décrites par un potentiel attractif/répulsif ou un potentiel uniquement répulsif, selon le signe du paramètre ϵ_{ij}. Plus spécifiquement, il est donné par l'équation:

$$
\begin{aligned}
E_{VdW} =& \epsilon_{ij} \left(\left(\frac{r^0_{ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{r^0_{ij}}{r_{ij}} \right)^{6} \right) H(\epsilon_{ij}) \\
& - \epsilon_{ij} \left(\frac{r^0_{ij}}{r_{ij}} \right)^{6} H(-\epsilon_{ij})
\end{aligned}
$$

où r^0_{ij} est la distance d'équilibre donnée par la moyenne des rayons de van der Waals de l'atome i et j et ϵ_{ij} est l'énergie minimale et est tirée d'une matrice d'interaction déterminée à partir des statistiques obtenues sur les structures déterminées expérimentalement (*knowledge-based*) [**167**]. $H(x)$ est la fonction de Heaviside et vaut 1 lorsque $x \geq 1$ et 0 sinon et détermine la forme attractive/répulsive ou simplement répulsive du potentiel. La forme

attractive/répulsive est utilisée pour les paires d'acides aminés apolaires et paires d'acides aminés de charges opposées.

Comme pour sOPEPv1 (voir le point (3) de la section précédente), OPEPv3 considère les ponts hydrogène explicitement selon l'équation suivante:

$$
\begin{aligned}
E_{HB}\left(r_{ij},\alpha_{ij}\right) = \quad & \epsilon_{\alpha}^{HB} \sum_{ij,j=i+4} \mu(r_{ij}) \cdot \nu(\alpha_{ij}) + \epsilon_{\beta}^{HB} \sum_{ij,j>4} \mu(r_{ij}) \cdot \nu(\alpha_{ij}) \\
\mu(r_{ij}) = \quad & \epsilon_{ij} \cdot \left[5 \left(\frac{\sigma}{r_{ij}} \right)^{12} - 6 \left(\frac{\sigma}{r_{ij}} \right)^{10} \right] \\
\nu(\alpha_{ij}) = \quad & \begin{cases} \cos^2(\alpha_{ij}) & \text{if } \alpha_{ij} > 90^{\circ} \\ 0 & \text{otherwise} \end{cases}
\end{aligned}
$$

où σ est la valeur d'équilibre de la liaison hydrogène et est fixée à 1.8 Å. Les ponts hydrogène caractéristiques des hélices α et des feuillets β sont considérés explicitement.

Finalement, comme dans OPEPv1 (voir les points (4) et (5) de la section précédente), OPEPv3 considère un terme de coopérativité lorsque l'on retrouve deux ponts hydrogènes caractéristiques des hélices α ou des feuillets β au niveau de la structure secondaire. La coopérativité est donnée par l'équation:

$$
\begin{aligned}
E_{coop}(r_{ij},r_{kl}) &= \epsilon_{\alpha}^{coop} \sum C(r_{ij},r_{kl}) \times \Delta(ijkl) + \epsilon_{\beta}^{coop} \sum C(r_{ij},r_{kl}) \times \Delta'(ijkl) \\
C(r_{ij},r_{kl}) &= \exp\left(-0.5\left(r_{ij}-\sigma\right)^2\right) \cdot \exp\left(-0.5\left(r_{kl}-\sigma\right)^2\right) \\
\Delta(ijkl) &= \begin{cases} 1 & \text{if } (k,l) = (i+1,j+1) \\ & \text{and } (j,l) = (i+4,k+4) \\ 0 & \text{otherwise} \end{cases} \\
\Delta'(ijkl) &= \begin{cases} 1 & \text{if } (k,l) = (i+2,j-2) \\ & \text{or } (k,l) = (i+2,j+2) \\ 0 & \text{otherwise} \end{cases}
\end{aligned}
$$

Similairement à OPEPv1, la balance relative entre chacun des termes du potentiel est déterminée à partir de l'optimisation des poids w associés à chacun d'entre eux. Ces paramètres furent optimisés afin de maximiser la reconnaissance de la structure native parmi un ensemble de diverses structures. Ces diverses structures contre lesquelles l'énergie de la native est optimisée sont appelées leurres.

Un jeu de 12 protéines-test fut développé afin de former l'ensemble de paramétrisation. Ces protéines sont 1abz(α, 40 aa), 1dv0(α, 47 aa), 1e0m (β, 37 aa), 1orc (α/β, 71 aa), 1pgb (α/β, 56 aa), 2gb1f (un fragment β de 2gb1 entre les résidus 41-56), 1qhk(α/β, 47 aa), 1shg(β,

62 aa), 1ss1 (α, 62 aa), 1vii (α, 36 aa), 2ci2 (α/β, 83 aa) et 2cro-fisa (α, 71aa) [**148**]. Les trois critères suivants ont motivé le choix de ces protéines: (1) elles sont stables en solution sans présence de ponts disulfures et composées uniquement des acides aminés standards, (2) elles sont composées d'une bonne balance entre les hélices α et les feuillets β et finalement (3) elles sont suffisamment petites pour permettre un bon échantillonnage de leur espace conformationel.

Pour chacune de ces protéines-test, un ensemble de leurres fut généré à l'aide de diverses méthodes d'échantillonnage; dynamique moléculaire, reconstruction greedy [**168**], méthode ART et enfilage ("threading") sur diverses topologies. En moyenne, 550 leurres furent générés pour chacune des protéines-test (entre 430 et 928). Ces leurres furent classifiés en fonction de leur degré de similarité avec la structure expérimentale tel que déterminé avec le TM-score [**169**]. Trois classes de leurres ont ainsi été définies: les leurres natifs (N) correspondant à la structure expérimentale, les leurres comme-natifs(L) moyennement similaires à la structure expérimentale (TM ≥ 0.5) et les leurres non-natifs (M) aucunement similaires à la structure expérimentale.

Finalement, cette classification permet de définir un ensemble d'inégalités sur lequel les paramètres de OPEPv3 furent optimisés. Cet ensemble d'inégalité est défini par:

$$E(N_i) < E(L_j), \text{pour i,j}$$
$$E(N_i) < E(M_k), \text{pour i,k}$$
$$E(L_j) < E(M_k), \text{pour j,k}$$

où i,j,k est respectivement le nombre de leurres dans les classes N, L et M respectivement.

À l'aide d'un algorithme génétique, les poids de OPEPv3 furent optimisés en maximisant la quantité d'inégalités résolues. Les résultats de l'optimisation furent testés sur un ensemble de protéines-test indépendant composé de 16 protéines avec en moyenne 1307 leurres chacun (entre 213 et 11519).

5.2.3. OPEPv4

Dans OPEPv4 [**170**], la forme du potentiel non-lié fut modifiée à partir d'une forme développée pour l'ADN [**158**]. Ce potentiel a été développé de telle sorte que: (1) à courte distance, on retrouve une loi de puissance répulsive pour décrire la répulsion entre les couches électroniques, (2) à large distance on retrouve une décroissance exponentielle pour offrir plus de flexibilité considérant la description gros-grain et (3) que la largeur du puits puisse être

contrôlée par la position du minimum (r^0_{ij}). Le potentiel ainsi utilisé est le suivant:

$$E_{VdW} = E_{AR4} - \epsilon_{i,j}\left(\frac{r^0_{ij}}{r_{ij}}\right)^8 H(-\epsilon_{i,j})$$

$$E_{AR4} = \epsilon_{i,j}\left(\left(\frac{G(r^0_{ij})}{r_{ij}}\right)e^{-2r_{ij}} + 0.6563701\tanh\left[2(r_{ij} - r^0_{ij} - 0.5) - 1\right]\right)H(\epsilon_{i,j})$$

$$G(r^0_{ij}) = -0.7\exp\left[\frac{2r^0_{ij} - 1}{5.0}\right]\left(r^0_{ij} - 0.5\right)$$

où $G(r^0_{ij})$ est l'expression contrôlant la largeur du potentiel en fonction de r^0_{ij} [**171**]. Tout comme les précédentes versions de OPEP, le potentiel pour les chaînes latérales dépend de leur type. Certaines sont attractives/répulsives ($\epsilon_{i,j}$ positif), tandis que d'autres sont simplement répulsives ($\epsilon_{i,j}$ négatif), contrôlées par la fonction de Heaviside ($H(x)$). On note que dans cette formulation, la partie répulsive du potentiel est plus douce (exposant 8 au lieu de 12).

En plus de ces modifications au niveau de la forme du potentiel, un problème au niveau de la stabilisation des hélices α fut aussi corrigé. Pour ce faire, des termes spécifiques aux contacts dans les hélices, entre les résidus (i,i+3) et (i,i+4), furent ajoutés pour les 11 paires d'interactions suivantes: Lys-Glu, la Lys-Asp et la Glu-Arg pour les interactions (i,i+3) et la Lys-Glu, la Lys-Asp, la Glu-Arg, l'Asp-Arg, la Lys-Gln, la Lys-Leu, l'Ala-Arg, l'Ala-Gln, l'Ala-Glu, la Leu-Glu, et l'Ile-Lys pour les interactions (i,i+4). Ces interactions sont décrites par le potentiel attractif/répulsif et sont affectées de paramètres distincts. En effet, les $\epsilon_{i,j}$ de ces interactions furent optimisés à l'aide d'un protocole similaire à celui d'OPEPv3 tout en gardant tous les autres paramètres fixes.

OPEPv4 permit de bien discriminer la structure native minimisée d'un ensemble de leurres pour les protéines 1PGB, 2CI2, 1SHG et 1ABZ, ensemble développé pour l'optimisation d'OPEPv3 [**165**]. Par la suite, OPEPv4 [**170**] fut testé en simulation de dynamique moléculaire à 300K. Après 20-30 ns, les structures de 17 protéines aux structures secondaires variées sont en moyenne à 3.1 Å de RMSD, calculé sur les C_α du coeur rigide, c'est-à-dire les C_α dont le positionnement varie peu entre les différents modèles expérimentaux. Les simulations furent prolongées jusqu'à 100 ns pour cinq d'entre elles sans changement significatif des résultats. Des simulations d'échange de répliques en température permit d'identifier le bon paysage d'énergie libre pour quatre petites protéines (trpzip1, trpzip2 et les fragments p16-31 et H1 de la protéine Prion) [**170**]. Finalement, la transition d'hélice α à agrégation amyloïde de la petite protéine ccβ fut reproduite [**170**].

5.2.4. OPEPv5

La version 5 de OPEP [**172**] propose une re-paramétrisation complète des paramètres entre chaînes latérales chargées: Arg-Asp, Lys-Asp, Lys-Asp et Lys-Glu. Pour ce faire, les

développeurs d'OPEPv5 ont utilisé la méthode d'inversion de Boltzmann itérative ("Iterative Boltzmann Inversion", IBI). Cette méthode permet d'obtenir le potentiel qui peut reproduire les RDFs (voir chapitre 5.1.2) obtenues à partir de simulations tout-atome. Plus spécifiquement, la méthode IBI permet d'obtenir les RDFs de façon itérative selon le protocole suivant. (1) À partir d'un potentiel initial, généralement déterminé à partir de l'équation de Boltzmann et de la RDF, $V_{\mathrm{PMF}}(r) = -k_B T \ln \mathrm{RDF}(r)$ (discuté à la section 5.1.3), une simulation du système à l'étude est réalisée. (2) La nouvelle RDF est calculée de cette simulation. (3) Le potentiel est corrigé à partir des différences entre la RDF de référence et celle obtenue via l'équation: $V^{n+1}_{\mathrm{PMF}}(r) = V^{n}_{\mathrm{PMF}}(r) - \alpha k_B T \ln\left(\frac{\mathrm{RDF}^n(r)}{\mathrm{RDF}_{ref}(r)}\right)$, où α est un paramètre contrôlant l'amplitude de la correction du potentiel.

Pour OPEPv5 en particulier, la méthode IBI a été utilisée afin de reproduire les fonctions de distribution radiale des paires de chaînes latérales chargées. Les RDFs ont été obtenues à partir de simulations de chacune des paires avec le champ de force tout-atome OPLS [**173**] avec solvant décrit explicitement avec TIP3P [**40**] à 303 K et 1 atm. Des groupements NH2 et COOH sont placés aux extrémités N-terminale et C-terminale respectivement. Pour obtenir le nouveau potentiel d'OPEPv5, ces simulations furent répétées entre 100 et 200 itérations lors de l'application de la méthode IBI.

Les potentiels furent dérivés sur les paires de résidus seules en solution [**172**], indépendamment des autres paramètres du potentiel, selon le protocole décrit au paragraphe précédent. Comme mentionné à la section 5.2.2, la relation entre les différents termes du potentiel OPEP est gérée à partir d'un poids w associé à chacun des termes. Pour les nouveaux paramètres des paires d'acides aminés chargés, un paramètre f est introduit afin d'échelonner ces interactions avec les autres de OPEP, restées inchangées. Pour déterminer la meilleure valeur de f, plusieurs tests furent réalisés. Dans un premier temps, la stabilité de six petites protéines, entre 37 et 85 acides aminés, fut testée avec des simulations de dynamique moléculaire pour différentes valeurs de f. Les meilleurs résultats, en termes de RMSD sur les C_α comparativement à la structure de référence, sont obtenus avec $f = 0.9$ avec une moyenne de 3.77 Å. Les résultats se détériorent avec un RMSD sur les C_α de 4.00 Å et 5.88 Å pour respectivement $f = 1.0$ et $f = 1.3$. Dans un second temps, la stabilité d'une petite hélice α de 13 résidus, correspondant au N-terminal de la ribonucléase-C, fut simulée avec de l'échange de répliques en température. Pour ce système, les résultats ne sont pas améliorés comparativement à OPEPv4. Pour un troisième temps, une simulation d'échange de répliques en température sur une petite épingle β, correspondant à un fragment de 16 résidus de la protéine G, fut réalisée. Contrairement à l'hélice α, un paramètre $f = 1$ mena à une meilleure reproduction de la température de fusion, une meilleure reproduction des ponts-H et une meilleure reproduction de la structure secondaire qu'avec OPEPv4. Finalement, une simulation d'échange de répliques en température fut réalisée sur le petit ccβ. Ce petit peptide de 17 résidus forme une hélice α à basse température et forme des fibres amyloïdes à

plus haute température. En fixant $f = 1.1$ ou $f = 1.3$, la stabilité de l'hélice α de ccβ est similaire avec OPEPv5 et OPEPv4, mais la transition vers l'agrégation amyloïde (feuillet β) est améliorée avec OPEPv5. En bref, en fixant adéquatement le paramètre d'échelonnage, f, OPEPv5 se compare avantageusement à OPEPv4 sur les tests décrits précédemment.

5.2.5. OPEPv6

OPEPv6 [**174**] modifie légèrement la forme du potentiel d'OPEPv5 afin de rendre OPEP compatible avec un nouveau protocole de modélisation du pH. Ce nouveau protocole utilise un algorithme de titration appelé "Fast Proton Titration Scheme" afin de mettre à jour de la protonation des chaînes latérales Glu, Asp, Tyr, Cys, Lys, His et Arg durant une simulation de dynamique moléculaire. Dans OPEPv5, uniquement les interactions entre Asp^-/Glu^- et Arg^+/Lys^+ furent modifiées. Or, pour ce qui est des charges inverses, il manque les interactions entre Asp^-/Glu^- et His^+. Comme la fonction de distribution radiale obtenue à l'aide de simulations tout-atome entre Asp^-/Glu^- et His^+ est similaire à celle entre Glu^- et Lys^+, le même potentiel est utilisé, sans optimisation supplémentaire. Dans OPEPv5, les interactions entre paires de résidus de même charge, par exemple Asp^- avec Glu^-, sont décrites à l'aide du potentiel attractif/répulsif décrit précédemment et ne dépend donc pas de leur état de protonation. Afin de considérer explicitement ces effets, un nouveau potentiel est introduit pour ces paires d'interactions. Il prend la forme d'une interaction de Coulomb écrantée donnée par l'équation:

$$U_{\text{répul}} = 332\frac{z_i z_j e^{-\kappa_c r_{ij}}}{\epsilon_s r_{ij}}$$

où z_i est la charge du résidu i, κ_c est la constante d'écrantage de Debye-Hückel et ϵ_s est la permittivité diélectrique du solvant. Ainsi, en combinant toutes ces modifications, le potentiel entre les différentes chaînes latérales s'écrit alors comme:

$$V_{SC,SC}(i,j) = \begin{cases} V^{IP}, \text{Si i/j ont des charges opposées} \\ U^{\text{répul}}, \text{Si i/j ont les mêmes charges} \\ V^{LJ}, \text{sinon} \end{cases}$$

où V^{IP} est le potentiel développé pour les paires chargées dans OPEPv5, $U^{\text{répul}}$ est le potentiel de Coulomb écranté développé dans OPEPv6 et finalement V^{LJ} est le potentiel attractif/répulsif de OPEPv4.

5.3. PEP-FOLD

Les petits peptides sont particulièrement intéressants puisque certains peuvent avoir des propriétés thérapeutiques cruciales pour la lutte contre certaines maladies et pour contrer la résistance aux antibiotiques [**127**–**129**]. Or les petits peptides présentent un défi particulier

par rapport aux protéines plus larges. En effet, la structure adoptée peut grandement varier pour la même séquence, qu'il s'agisse d'un peptide ou d'un fragment de protéines plus grande [**175**]. PEP-FOLD [**147**, **148**, **176**] est une méthode simplifiée, et disponible gratuitement en ligne [**177**], pour faire la prédiction *de novo* de la structure des peptides et des petites protéines; c'est-à-dire la prédiction structurelle uniquement à partir de la séquence en acide aminé. Dans cette section, je décrirai dans un premier temps les divers éléments de la méthode PEP-FOLD, dont le processus global est présenté à la Figure 5.4. Finalement, je discuterai de possibles pistes d'amélioration, dont certaines sont l'objet de l'article présenté au chapitre suivant.

5.3.1. Acides aminés à alphabet structurel

À la base de PEP-FOLD, on retrouve un alphabet structurel ("structural alphabet") dérivé d'un modèle de Markov caché ("Hidden Markov Model" (HMM)) [**178**]. Dans cet alphabet structurel, chacune des lettres correspond à un fragment de quatre acides aminés et décrite par quatre paramètres, tel que présenté à la Figure 5.2: les trois distances inter-atomiques entre les Cα non-liés, $d_1 = d\{C_{\alpha 1} - C_{\alpha 3}\}$, $d_2 = d\{C_{\alpha 1} - C_{\alpha 4}\}$ et $d_3 = d\{C_{\alpha 2} - C_{\alpha 4}\}$, et d_4 la projection de $C_{\alpha 4}$ sur le vecteur normal du plan formé par les trois premiers. Les lettres adjacentes dans la séquence sont partiellement superposées, c'est-à-dire que les atomes $C_{\alpha 2}$, $C_{\alpha 3}$ et $C_{\alpha 4}$ de la lettre à la position (i) correspondent aux atomes $C_{\alpha 1}$, $C_{\alpha 2}$ et $C_{\alpha 3}$ de la lettre à la position (i+1).

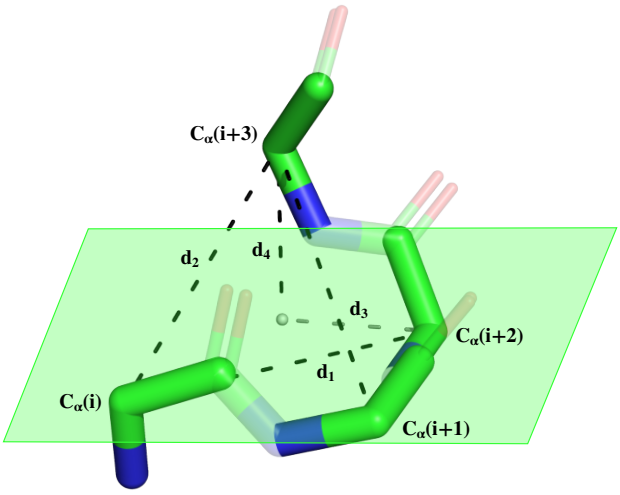


Fig. 5.2. Représentation des quatre distances définissant une lettre de l'alphabet structurel. Pour la lettre A de l'alphabet structurel, les quatre distances d_1, d_2, d_3 et d_4 sont présentées. Le plan dans lequel se situe les trois premiers carbones-α est représenté par le parallélogramme vert.

Les paramètres, d_1, d_2, d_3, d_4, de chacun des fragments ainsi que les taux de transition entre chacun d'entre eux furent appris simultanément lors de l'apprentissage du modèle de

Markov caché à l'aide de l'algorithme d'espérance-maximisation ("expectation-maximization algorithm") sur deux banques représentant chacune 56000 lettres dérivées d'un ensemble de 250 protéines. Un HMM et un SA furent dérivés en variant le nombre de lettres dans l'alphabet entre 12 et 33. La description maximale est atteinte pour un SA contenant 27 lettres [**178**], identifiées par les lettres majuscules A-Z et le a minuscule et présentées à la Figure 5.3.

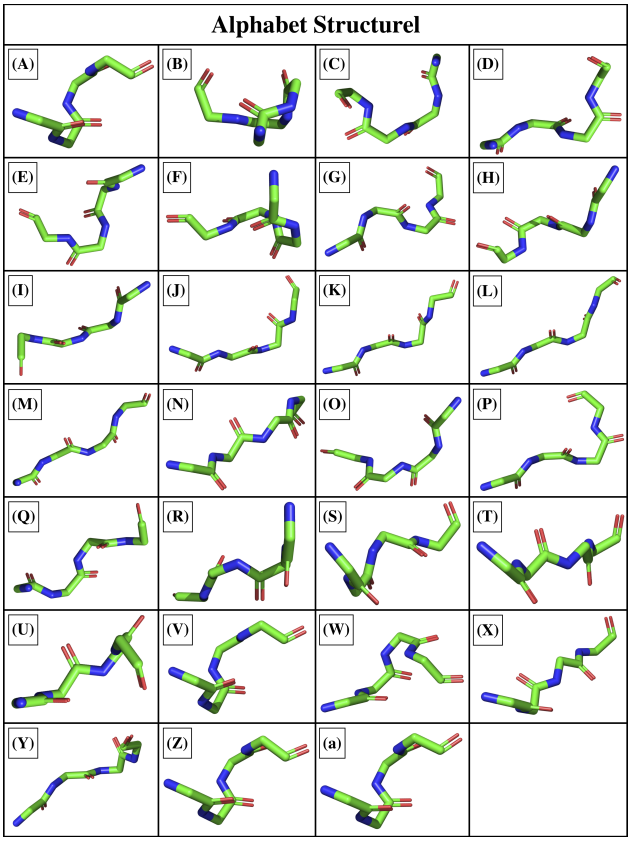


Fig. 5.3. Représentation des lettres de l'alphabet structurel de PEP-FOLD. Un fragment représentatif (fragment 0) est présenté pour chacune des lettres de l'alphabet structurel (a,A-Z) utilisé par PEP-FOLD.

La présence de chacune de ces lettres dans les différentes structures secondaires fut analysée à l'aide de STRIDE [**179**]. Les lettres A, a, V et W sont associées aux structures d'hélice-α, les lettres L, N, M, T et X sont associées aux structures de feuillet-β et la lettre B est associée aux structures d'hélice-310 [**178**]. Ainsi l'alphabet structurel peut être conçu comme une généralisation du concept de structure secondaire. Afin d'augmenter la variabilité structurelle, chacune des lettres de l'alphabet structurel est associée à une série de fragments représentatifs, déterminés à partir d'un algorithme de "clusterization" avec un RMSD seuil de 0.5 Å. Ainsi, à l'intérieur de PEP-FOLD, la représentation tridimensionnelle des protéines est donc encodée dans l'espace unidimensionnelle des lettres de l'alphabet structure.

La prédiction structurelle *de novo* requiert la prédiction de la structure tridimensionnelle uniquement à partir de sa séquence en acide aminé. La première étape de PEP-FOLD est donc de traduire la séquence d'acides aminés dans l'espace de l'alphabet structurel.

Pour ce faire, PEP-FOLD utilise une machine à vecteurs de support ("support vector machine", SVM) [**147**, **148**]. Pour déterminer les probabilités de chacune des lettres de l'alphabet structurel associées à un fragment, le SVM prend comme donnée d'entrée une matrice 20×8. Les colonnes correspondent aux quatre acides aminés du fragment en plus des deux acides aminés les précédents et les suivants. Les lignes correspondent à la matrice de score spécifique à la position ("position-specific score matrix") dérivée par PSI-BLAST [**180**] sur la banque de séquence d'UniRef30 [**151**]. Pour une protéine de taille L, le SVM retourne une matrice $L-3 \times 27$ correspondant aux probabilités à chacune des positions d'être représentée par les 27 lettres de l'alphabet structurel [**147**, **148**]. Cette étape est présentée au panneau (2) de la Figure 5.4.

Finalement, la dernière étape, une nouveauté de PEP-FOLD3 [**147**], est l'utilisation des prédictions du SVM et d'un algorithme "Forward-Backtrack" afin de convertir la séquence d'acides aminés en une série de séquences composées des lettres de l'alphabet structurel. Cette étape correspond au panneau (3) de la Figure 5.4.

5.3.2. Reconstruction tridimensionnelle

Une fois que l'ensemble de séquences dans l'ensemble des lettres de l'alphabet structurel est généré à partir la séquence d'acides aminés, il faut maintenant les assembler afin d'obtenir la structure tridimensionnelle associée. Pour ce faire, PEP-FOLD utilise une version stochastique de l'algorithme "greedy" [**168**]. Pour ce faire, la chaîne polypeptidique est allongée d'un acide aminé à la fois, via la superposition des fragments de l'alphabet structurel. Plus spécifiquement, les trois premiers Cα du fragment à ajouter sont superposés sur les trois derniers Cα du fragment précédent, via une minimisation du RMSD. Cette opération est répétée de façon itérative jusqu'à atteindre les extrémités. La reconstruction est présentée au panneau (4) de la Figure 5.4.

Par contre, commme mentionné à la section précédente, chacune des lettres de l'alphabet structurel, i, est associée à un certain nombre de fragments représentatifs, l_i. Ainsi, pour une protéine composée de L acides aminés, décrite par L-3 lettres, le nombre de possibilités de reconstruction tridimensionnelle est donnée par $\Pi_i^{L-3} l_i$, un nombre qu'il n'est pas raisonnable de considérer de façon exhaustive, même pour de petites protéines.

Avec l'algorithme "greedy", la reconstruction s'effectue en conservant uniquement un certain nombre de structures à chacune des étapes. En effet, une chaîne de taille i est allongée en considérant exhaustivement tous les fragments de la lettre associée pour donner une chaîne de taille $i+1$. Si le nombre de structures excède un nombre seuil, H, alors uniquement les

H "meilleures" structures seront conservées pour l'ajout du prochain résidu, de telle sorte, qu'à chacune des étapes de la reconstruction, le nombre de structures considérées ne dépasse jamais $H \times l_i$. Dans PEP-FOLD, les H "meilleures" structures sont les structures de plus basse énergie, déterminée selon le potentiel sOPEP [**148**], décrit à la prochaine section. Un des problèmes d'une telle procédure est que l'objectif est d'obtenir la meilleure structure globale, alors que la sélection est effectuée au fur et à mesure de la reconstruction. En d'autres mots, un sous-segment de haute énergie qui serait éliminé par l'algorithme "greedy" pourrait ultimement mener au minimum global lorsque la reconstruction est complétée.

Afin de contrer ces problématiques, un nouvel algorithme "greedy" fut développé en considérant un élément stochastique dans la sélection des structures à conserver [**168**]. Dans ce nouvel algorithme, un certain nombre, B, de structures sont sélectionnées selon le critère de plus basse énergie et un certain nombre, R, de structures sont sélectionnées aléatoirement, de telle sorte que le nombre total de structures conservées, H, est donné par $H = B + R$.

Ce nouvel algorithme se compare avantageusement comparativement à l'algorithme "greedy" ordinaire lors de la reconstruction de 16 protéines guidée par un potentiel de Go. Le nouvel algorithme "greedy" mena à des prédictions entre 1.5 et 4.8 Å de RMSD tandis que l'algorithme "greedy" classique mena à des structures désordonnées [**168**]. Cette étape du protocole PEP-FOLD est présentée au panneau (5) de la Figure 5.4.

5.3.3. sOPEP

Tel que mentionné dans la section précédente, la reconstruction des modèles en trois dimensions se fait via un assemblage "greedy" des différents fragments associés aux lettres de l'alphabet structurel. Durant la reconstruction, une partie des structures est conservée ou rejetée selon un critère énergétique. Dans l'article original [**168**], un potentiel de Go avait été utilisé. Dans PEP-FOLD, l'énergie des assemblages préliminaires est déterminée à partir du potentiel sOPEP [**148**] ("simplified OPEP"). sOPEP est un potentiel gros-grain de la famille OPEP et partage donc de nombreuses caractéristiques avec les différentes versions de OPEP qui ont été présentées en détail à la section 5.2.

La représentation gros-grain de sOPEP est similaire à celle de OPEP, décrite à la section 5.2.2, à une différence près; la chaîne latérale de la proline n'est plus considérée en tout-atome, mais bien comme un seul centre d'interaction. Pour ce qui est du potentiel, tous les termes restent inchangés, sauf les interactions entre les chaînes latérales. En effet, l'assemblage des fragments dans l'espace discret de PEP-FOLD est sensiblement différent de la dynamique moléculaire pour laquelle OPEP fut développé et mène à un nombre important de collisions stériques. Pour éviter ce problème, le potentiel entre chaînes latérales fut

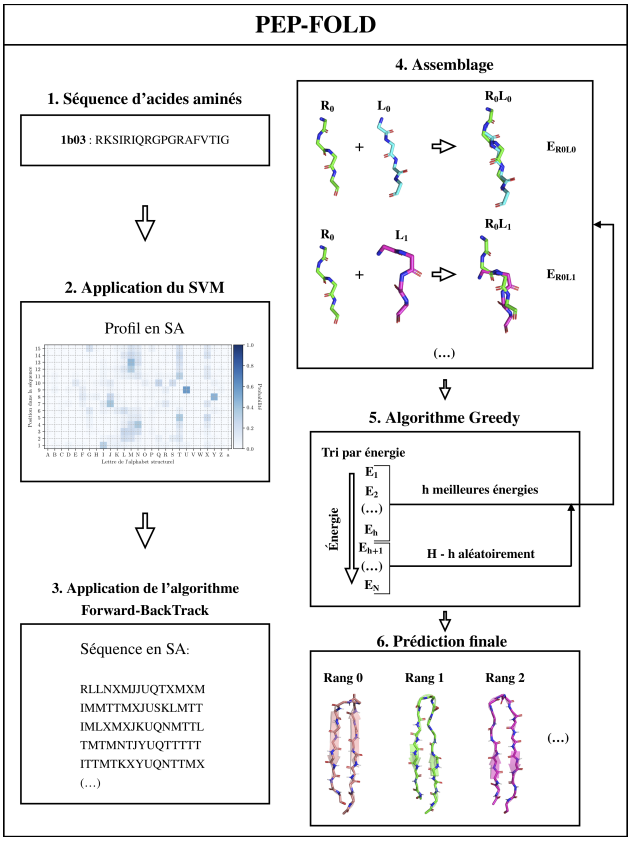


Fig. 5.4. PEP-FOLD. Présentation du protocole PEP-FOLD. À partir de la séquence d'acides aminés (1), un SVM est utilisé afin de générer les probabilités que chacune des positions soit décrite par chacune des lettres de l'alphabet structure (2). Par la suite, un algorithme de Foward-Backtrack est appliqué afin de générer un nombre désiré de séquences dans l'espace des lettres de l'alphabet structurel (3). Par la suite, ces séquences sont assemblées (4) via la superposition des fragments et un algorithme greedy (5). Une fois la reconstruction terminée, on obtient les prédictions tridimensionnelles (6). L'exemple présenté ici est pour la protéine 1b03.

modifié afin de le rendre moins contraignant. Le nouveau potentiel est décrit par l'équation:

$$E_{SC,SC}(r_{ij}) = \begin{cases} -\epsilon_{ij} C(r_{ij})^6, & \text{si } \epsilon_{ij} < 0 \\ \epsilon_{ij} \left(C(r_{ij})^{12} - 2C(r_{ij})^6 \right), & \text{sinon} \end{cases} \tag{5.3.1}$$

Dans le cas où $\epsilon_{ij} > 0$, alors, $C\left(r_{ij}\right)$ prend la forme suivante:

$$C\left(r_{ij}\right) = \frac{r_{ij}^0 - p_{ij}}{r_{ij}^0 - p_{ij}}$$

$$p_{ij} = \frac{r_{ij}^0 - \sqrt[6]{2} R_{ij}^0}{1 - \sqrt[6]{2}}$$

où R_{ij}^0 est la valeur du zéro du potentiel. Ainsi, pour contrôler le caractère répulsif du potentiel à courte distance, le potentiel sOPEP contrôle la position de l'asymptote à l'aide du paramètre p, lui-même fonction de r^0 et R^0.

Tandis que pour $\epsilon_{ij} < 0$, $C\left(r_{ij}\right)$ prend la forme suivante:

$$C\left(r_{ij}\right) = \frac{2R_{ij}^0 - r_{ij}^0}{r_{ij}}$$

Les valeurs de R_{ij}^0 furent déterminées avec une approche similaire à celle utilisée pour dériver les valeurs de r_{ij}^0 pour OPEPv3 (voir section 5.2.2). La distribution des distances entre les chaînes latérales fut dérivée à partir d'une banque de structures extraite de la *Protein Data Bank* avec moins de 30% de similarité de séquence. Pour le potentiel attractif/répulsif ($\epsilon_{ij} > 0$) et répulsif ($\epsilon_{ij} > 0$), les valeurs de R_{ij}^0 correspondent respectivement au quantile 0.1 et 0.2 de la distribution des distances.

Finalement, pour considérer les modifications apportées à la forme du potentiel, les poids associés aux interactions SC-SC furent ré-optimisés à l'aide du même protocole que pour OPEPv3 [**165**], décrit à la section 5.2.2. Le jeu de protéines et les leurres formant l'ensemble de paramétrisation du protocole d'optimisation sont identiques à celui de OPEPv3, sauf l'exclusion de la protéine betanova.

5.4. Objectifs des travaux

La méthode PEP-FOLD est composée de trois éléments-clés: (1) un alphabet structurel et un ensemble de fragments permettant de décrire de façon discrète la structure des protéines, (2) un SVM permettant de convertir les séquences d'acides aminés dans l'espace des lettres de l'alphabet structurel et (3) un potentiel gros-grain permettant de discriminer les prédictions natives des prédictions non-natives, appelé sOPEP.

Nous avons décidé de retravailler deux de ceux-ci, soit les fragments de l'alphabet structurel et la formulation du potentiel gros-grain sOPEP. Un nouveau protocole a été utilisé afin de développer une nouvelle librairie de fragments. Premièrement, le processus de superposition a été revisité afin d'augmenter les similarités avec la superposition de l'algorithme "greedy". Deuxièmement, un nouveau processus de clusterisation est utilisé pour réunir les fragments similaires entre eux. Pour ce qui est du potentiel sOPEP, nous avons révisé la formulation du potentiel pour les interactions non-liées, afin de corriger certaines propriétés étranges de celui-ci. Premièrement, la version attractive/répulsive permet de contrôler la position de l'asymptote du potentiel via les paramètres r^0 et R^0. Or certaines combinaisons de ces paramètres peuvent mener l'asymptote du potentiel assez loin dans les distances négatives, ce qui diminue grandement le caractère répulsif à très courte distance. Deuxièmement, la version répulsive du potentiel est uniquement affectée de l'exposant 6, ce qui le rend passablement moins répulsif qu'un potentiel de Lennard-Jones classique où la répulsion

est de l'ordre 12. Afin de corriger ces problèmes, mais de garder la flexibilité requise pour l'assemblage discret et la description gros-grain, nous avons utilisé la version généralisée du potentiel de Lennard-Jones, soit le potentiel de Mie [**181**]. De plus, nous avons développé un protocole d'optimisation des paramètres de sOPEP spécifique à PEP-FOLD, contrairement à son optimisation initiale plutôt axée sur la dynamique moléculaire.

Chapitre 6

A generalized attraction-repulsion potential and revisited fragment library improves PEP-FOLD peptide structure prediction

par

Vincent Binette[1], Normand Mousseau[1] et Pierre Tuffery[2]

([1]) Départment de physique, Université de Montréal, Case postale 6128, succursale Centre-ville, Montréal, QC, H3C 3J7 Canada.

([2]) Université de Paris, INSERM U1133, CNRS UMR 8251, F-75205 Paris, France.

Cet article a été soumis dans Journal of Chemical Theory and Computation.

Mes contributions et le rôle des coauteurs:

Au niveau des améliorations de la méthode PEP-FOLD, mes principales contributions se retrouvent le long de deux axes principaux:

(1) **Développement et conception de la méthode**: Ma première contribution a été de déterminer les modifications à apporter au potentiel sOPEP. Ces modifications incluent la nouvelle forme des interactions non-liées et le choix de négliger les propensions de structure secondaire des acides aminés dans le calcul de la coopérativité

en plus de déterminer le choix des paramètres à optimiser. Ma seconde contribution se situe au niveau de l'optimisation du potentiel. Tout particulièrement, de déterminer les protéines ciblées, de développer l'ensemble de paramétrisation et de validation et de choisir et de développer le protocole d'optimisation.

(2) **Analyse, visualisation et publication**: J'ai contribué à l'analyse des résultats, notamment en développant les figures de l'article et en réalisant les comparaisons avec les méthodes d'apprentissage machine. Finalement, j'ai contribué à toutes les étapes d'écriture de l'article et des corrections en tant que premier auteur.

Tout ce travail s'est fait sous la supervision du professeur *Normand Mousseau*, en plus de ses contributions au niveau de la conception, de l'analyse des résultats et de l'écriture de l'article. Finalement le professeur *Pierre Tuffery* a contribué à la conception et la réalisation du projet, tout particulièrement en ce qui a trait au développement de la librairie révisée de fragments, en plus de l'analyse et de l'écriture de l'article.

Abstract. Fast and accurate structure prediction is essential to the study of peptide function, molecular targets and interactions and has been the subject of considerable efforts in the past decade. In this work, we present improvements to the popular, simplified PEP-FOLD technique for small peptides structure prediction. PEP-FOLD originality is three fold : it uses (i) a predetermined structural alphabet, (ii) a sequential algorithm to reconstruct the tri-dimensional structures of these peptides in a discrete space using a fragment library, and (iii) it assesses the energy of these structures using a coarse-grained representation in which all the backbone atoms but the α-hydrogen are present, and the side chain corresponds to a unique bead. In former versions of PEP-FOLD, a van der Waals formulation was used for non bonded interactions, each side chain being associated with a fixed radius. Here, we explore the relevance of using instead a generalized formulation in which not only the optimal distance of interaction and the energy at this distance are parameters, but also the distance at which the potential is zero. This allows each side chain to be associated with a different radius and potential energy shape, depending on its interaction partner, and in principle, to make more effective the coarse-grained representation. In addition, PEP-FOLD's new version is associated with an updated library of fragments. We show that these modifications lead to important improvements for many of the problematic targets identified with former PEP-FOLD version while maintaining already correct predictions. The improvement is both in terms of model ranking and model accuracy. We also compare PEP-FOLD enhanced version to state-of-the-art techniques for both peptide and structure predictions; APPTest, RaptorX and AlphaFold2. We find that the new predictions are superior, in particular with respect to the prediction of small β-targets, to those of APPTest and RaptorX and bring, with its original approach, additional understanding on folded structures, even when less precise than AlphaFold2. With their strong physical influence, the revised structural library and coarse-grained potential offer, however, means for a deeper understanding of the nature of folding and open a solid basis for studying flexibility and other dynamical properties not accessible to IA structure prediction approaches.
Keywords: peptide's structure prediction, coarse-grain potential

6.1. Introduction

Proteins are macromolecules involved in a wide variety of crucial biological processes. Their functions are determined by their tri-dimensional structure as well as their dynamics and thermodynamics properties. Thus the characterization of protein folding is of great interest in molecular biology, particularly how the tri-dimensional structure of protein is encoded in its amino acid sequence [**131**]. Since the end of the Human Genome Project, the development of next-generation sequencing lead to a drastic decrease in cost and drastic increase in the number and diversity of determined genomes [**182**] and leads to a gap between the number of known sequences and known structures. The development of fast and accurate protein structure predictions techniques is required to not only study the characteristics of the protein themselves but also their interactions with partners such as peptides and small molecules. Indeed, protein structure predictions also play a key role in the design of new

therapeutic molecules as the discovery and development of 210 new molecules approved by the US Food and Drug Administration between 2010 and 2016 were facilitated by structural information available in the *Protein Data Bank* [**183**].

Progress in the numerical predictions of protein's tri-dimensional structure is monitored by the Critical Assessment of Techniques for Protein Structure Prediction (CASP) meetings [**142**]. In recent years, the utilization of multiple sequence alignments (MSA) with protein sequences derived from genomic sequencing taken from huge data sets combined with very successful machine learning techniques, such as RaptorX [**152**–**154**], RosettaFold [**184**] or the now state-of-the-art AlphaFold2 [**12**] has led to tremendous improvement of the predicted results nearing experimental accuracy for some targets.

However, the CASP meetings are mainly focused on fairly large proteins of a couple hundreds (to a few thousands) amino acids. For example, only three targets tested in CASP14 have less than 70 amino acids. However, many small peptides, of less than a few dozens amino acids, have interesting properties. For instance, antimicrobial peptides of such size [**127**,**129**] could be crucial in the mitigation of antibiotic resistance, which according to some experts, could lead to 10 millions yearly deaths by 2050 [**128**]. Newly emerging interfering peptides also belong to these sizes [**185**].

Small peptides present a unique challenge compared to large proteins [**175**] and multiple computational approaches utilizing a wide variety of techniques have been developed to target specifically the peptide secondary and tertiary structure prediction. For example, PSSP-MVIRT is a successful deep-learning method for the prediction of peptide's secondary structure [**186**]. PEPstr [**143**] (and its extension to non-standard amino acid, PEPstr-Mod [**144**]) utilizes the observation on the prevalence of β-turn secondary structure to add constraints on molecular dynamics simulation to predict peptide's tertiary structure. In the parallel microgenetic algorithm (PMGA) [**187**] techniques, peptides' structure predictions are done by utilizing a genetic algorithm with backbone dihedral angle correlations for sampling a density functional theory derived fitness function. Finally, the recently developped APPTest [**155**] was developped by combining distances/angles constrains derived by a neural network with simulated annealing, resulting in great structural predictions for small peptides.

In this study, we present improvements to PEP-FOLD [**147**,**148**,**176**] a quick and highly simplified approach for small peptides structure prediction. The PEP-FOLD approach is freely available as a webserver [**188**] and as been used in a variety of applications, such as the very recent research of a SARS-CoV-2 treatment [**189**–**191**]. The PEP-FOLD approach is based on three main features: (1) the concept of structural alphabet (SA), (2) discrete fragment assembly and (3) a coarse-grained energy function. We present here improvements to two of these key features.

First, the fragment library was reworked to better sample the conformational variability associated with the letters of the structural alphabet. Second, we revisit the coarse-grained energy function. The coarse-grained energy function used in PEP-FOLD is based on the **O**ptimized **P**otential for **E**fficient Structure **P**rediction (OPEP). Compared to other forcefields such as that of CABS-fold [**192**], major differences come from the coarse grained representation and from the treatment of hydrogen bonds. The OPEP representation includes all atoms from the backbone except the α hydrogen, and represents side chains using only one bead. This detailed backbone representation makes possible an explicit account for hydogen bonds, necessary to support the OPEP specific treatment for cooperativity in hydrogen bonds, that favors secondary structure formation during folding. Over the years, the OPEP forcefield has been successfully applied to a wide variety of biophysical application [**162**], including the self-assembly of amyloid protein [**4**], associated with many neurodegenerative diseases like Alzheimer, the study of DNA/RNA systems [**158**], the peptide/protein docking [**163**] and many more. More specifically, PEP-FOLD's predictions are guided by a **s**implified version of the OPEP forcefield named sOPEP, that ignores most of the bonded energy terms due to the PEP-FOLD' specific assembly procedure that does not occur in a continuous space, but in a discretized space using a limited number of fragments representative of the structural alphabet [**148**]. This rigid assembly process challenges the relevance of the non-bonded energy terms that are based on a van der Waals formulation. In OPEP, each particle is associated with one radius. This fixed radius can reveal problematic optimally parametrize interactions that can occur under contradictory circumstances. For instance a large radius could be relevant for a large side-chain interacting with another large side-chains but irrelevant for interactions with small side-chains or the beads describing the backbone, leading to high energy values for inter-bead distances observable in structures. Several ways to overcome this kind of limitation have been proposed in the literature such as the use of soft-core potentials [**193**], or variations in the exponent values of the van der Waals terms, as proposed by Mie a very long time ago [**181**] or more recently in the context of long-range corrections for dispersion interactions in inhomogeneous simulations [**194**]. However, none of these solutions addresses satisfactorily the requirement to have simultaneous control over the optimal distance $r0$, the energy at this distance, and the distance at which the energy is 0. In a previous study, we had proposed a formulation making it possible for disulfide bonds [**195**]. Here we generalize it to any exponent combination.

In former studies, the optimization of sOPEP was done on large ensembles of decoys generated with a wide variety of sampling techniques; molecular dynamics, threading, greedy assembly etc. These sampling algorithm have different search spaces compared to PEP-FOLD, which could have an impact on the effectiveness of sOPEP. Second, the classification score used for the optimization was the TMscore [**169**], a score based mainly on geometric factors (mean distance between corresponding Cα atoms) while sOPEP energy terms mainly

involves inter-atomic interactions (contacts, explicit hydrogen-bonds etc.). Finally, only a small portion of the parameters were optimized while most of them were derived from experimental structures, with little consideration for interactions inter-dependence.

In this study, we present a reworking of the non-bonded interactions of sOPEP as well as the re-optimization of all its energy components. We analyze how the combination of the newly improved fragments library and the newly optimized sOPEP potential impacts PEP-FOLD predictions' quality, and we compare it to state-of-the-art approaches for both peptide and structure predictions.

6.2. Materials and Methods

6.2.1. PEP-FOLD

PEP-FOLD relies on a Hidden Markov Model (HMM) derived structural alphabet (SA) [**178**]. It consists of 27 letters that correspond to fragments of four residues overlapping by three residues. Thus, the 3D-conformation of a peptide of length L can be described by L-3 SA fragments.

More specifically, PEP-FOLD's prediction of the 3D-structure from the amino acid sequence is performed according to a three steps protocol:

(1) **SA profile prediction**: PEP-FOLD first converts the amino acid sequence into a sequence of letters taken from the structural alphabet (SA). This is achieved by using a support vector machine (SVM) that takes as input a matrix of eight series of 20 values. The series correspond to the four amino acids in the fragments, extended by two on each side. The 20 values correspond to the position-specific scoring matrix as determined by PSI-BLAST [**180**]. The result is a SA profile that gives the probability of each segment of the protein to be described by each of the 27 letters of the SA. Following the SVM prediction, the Forward-Backtrack algorithm (FBT) or a Taboo Sampling algorithm is then used to generate a specified number of sub-optimal sequences in the SA letters space from the SA profile [**147**].

(2) **Tri-dimensional reconstruction**: Starting from the generated sequences in the SA letters space, PEP-FOLD then generates the 3D-conformations associated with each sequence. This is done via a rigid assembly of the fragments associated with each SA letter sequentially. The polypeptide chain is built by adding amino acid by amino acid, starting from an initial fragment of 4 amino-acids. At each step, all possible conformations are generated by superimposing the fragments associated with the SA letter at the current position to the last three amino-acids of the conformations generated at the previous step. A modified greedy algorithm [**168**] is used to filter the generated conformations at each step of reconstruction. A portion of the structures

are kept based on their predicted energy according to the sOPEP forcefield [**148**] with the rest selected at random among the remaining structures.

(3) **Monte-Carlo**: As a final step, the resulting unique conformation associated with each of the generated sequences in the SA letters space is then refined using a Monte-Carlo procedure; at each Monte-Carlo step, a fragment is randomly replaced by another and the modification is accepted based on a Metropolis criteria.

For a more thorough description of the PEP-FOLD's protocol, we refer the reader to *Lamiable et al.*'s article [**147**].

6.2.2. Library of fragments

The first part of PEP-FOLD, which is revisited here, is the fragment library. The structure of a non redundant collection of proteins was decoded as a series of strings of SA letters using the Viterbi or the forward-backward algorithms (see Camproux et al. [**178**]). Fragments of 4 amino acids associated with each of the 27 letters were collected, and for each letter, the clustering the fragments associated with them allowed to identify representative fragments of the letter, in a number depending on the conformational variability of the letter.

Two main changes are made with respect to the initial design of the library of fragments. The first one concerns the approach used to superimpose the fragments in order to generate a distance matrix between each of them. While, originally, superimposition was performed using the backbone of the four amino-acids of the fragments, we opt here to superimpose only the three first amino acids to measure the RMSd between the fragments. This modification delivers a scheme that is expected to be more consistent with the HMM concept, as it allows a better measurement of the diversity of the position of the fourth amino acid.

The second is a change in the clustering itself. Instead of using dynamics clustering, we now use the Ward algorithm, as implemented in the *hclust* module of R, using the squared dissimilarity values (ward.D2). The resulting tree is then used to identify clusters separated by some arbitrary cut-off value. A similar value was used for all SA letters. In order to keep the calculation tractable, only a limited number of fragments was randomly drawn from the complete sets. A number of 5000 was found sufficient to ensure a satisfactory reproducibility. Cluster centroids used for the fragment assembly correspond to the fragment the closest to all other members of the cluster. Finally, outliers clusters, i.e, those whose effectiveness are less than 2.5% of the number of fragments, are discarded. This threshold, which would allow up to 40 equally distributed clusters, is much lower than the expected frequency of well populated clusters which number are, in practice, of the order of 15 to 20. From now on, the original and the updated libraries will be referred to as Lib1 and Lib2 respectively.

6.2.3. sOPEP

One of PEP-FOLD's particularities is the use of a physics-based/knowledge-based coarse-grained potential, sOPEP, to discriminate between structures. This description plays a crucial role in guiding the 3D-reconstruction as well as in the Monte-Carlo refinement step.

The coarse-grained representation used in sOPEP is based on the OPEP [**148**, **156**] forcefield representation. The OPEP forcefield is a coarse-grained model where each amino acid is represented by a total of six pseudo-atoms as shown in Figure 6.1: the backbone is represented by five pseudo-atoms for atoms N, H, C_α, C and O and a single pseudo-atom is used to represent the SC. The position for the side-chain (i) is fixed based on the C(i-1), N(i) and C_α(i) positions and using predetermined centroid values [**165**].

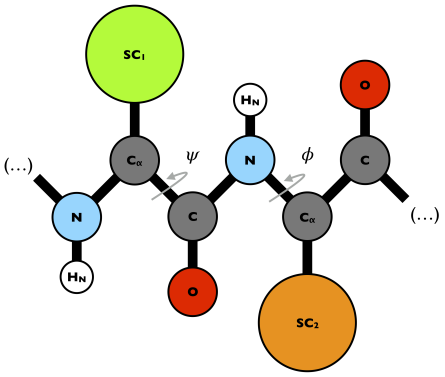


Fig. 6.1. Coarse-grained representation in sOPEP. The backbone is represented in a all-atoms format with all backbone atoms but HA — N, H_N, C_α, C and O — present while side-chains are represented by a single interaction center.

The sOPEP potential is a variation on OPEP targeted at the specific structural alphabet approach of PEP-FOLD ; it is composed of three main energy terms: bonded interactions (dihedral angles), non-bonded interactions (repulsion/dispersion effects) and explicit hydrogen bonds (with secondary structure cooperativity). In the following, the original formulation/parametrization is referred to as sOPEPv1, while the formulation/parametrization introduced here is called sOPEPv2. The complete formulation of sOPEPv2, as well as its key differences with sOPEPv1, are presented below.

6.2.3.1. Bonded Potential. The only bonded interaction considered in sOPEPv1 is the dihedral angle ϕ (between atoms C(i-1), N(i), Cα(i) and C(i)). In addition to dihedral angle ϕ, sOPEPv2 also accounts for the dihedral angle ψ (between atoms N(i), Cα(i), C(i) and N(i+1)). These two dihedral angles are of crucial importance in the description of protein conformations as demonstrated by the well-known Ramachandran Plot. Since, in PEP-FOLD, the geometry is mainly imposed by the superimposition of the discrete SA letters, the impact of this addition is minimal, but is added here for completeness.

As PEP-FOLD constrains the backbone by a rigid association of the fragments of the SA letters, sOPEPv2 uses a simple flat-bottomed quadratic potential to described the energy associated with dihedral angles ϕ described by:

$$E_{rama}(\phi_i) = \epsilon_\phi \left(\phi_i - \phi_{0\_sc\_i}\right)^2$$

where $\phi_{0\_sc\_i} = \phi$ within the interval $[\phi_{low\_sc\_i}, \phi_{high\_sc\_i}]$ and $\phi_{0\_sc\_i} = \min(\phi - \phi_{low\_sc\_i}, \phi - \phi_{high\_sc\_i})$ outside of the interval $\phi_{low\_sc\_i}$ and $\phi_{high\_sc\_i}$ are specific to each amino acid type.

sOPEPv2 uses the same equations for describing the dihedral ψ angle with adapted parameters.

6.2.3.2. Non-Bonded Potential. The potential associated with repulsion/dispersion effects was slightly reworked in sOPEPv2. For side-chain–side-chain interactions, sOPEPv1 adopted a dual formulation using either a repulsive term or a repulsive/attractive term based on the identify of the atoms pair [**148**]. In sOPEPv2, the repulsive/attractive term is adopted for all side-chains pairs, and for all non-bonded particle interactions, excluding HN which is only considered for hydrogen bonds.

The repulsion/dispersion effects are described using the following potential given:

$$E_{vdw\_ij} = \epsilon_{ij} \times \left[\frac{m}{n-m}\left(\frac{r_{ij}^0}{r_{ij}}\right)^n - \frac{n}{n-m}\left(\frac{r_{ij}^0}{r_{ij}}\right)^m\right] \tag{6.2.1}$$

where ϵ_{ij} is the potential depth and r_{ij}^0 is the position of the potential minimum function of atomic types for i and j. The combination of exponents, n and m, gives the relationship between the position of the potential minimum (r^0) and the position where it is zero ($gR0$):

$$gR0 = \left(\frac{m}{n}\right)^{\frac{1}{n-m}} r_0 \tag{6.2.2}$$

It is thus possible to have control over the well depth, its position and the position where the potential is zero, but the slope at $gR0$ cannot be adjusted independently. This formulation makes it possible, to some extent, to limit the impact of the representation of the side-chains using only one bead. sOPEPv1 parameters include ϵ_{ij} and $gR0_{ij}$ specific to each pseudo-atom type pair and potential minimum defined by the sum of individual pseudo-atom type radius : $r_{ij}^0 = r_i^0 + r_j^0$. sOPEPv2 retains sOPEPv1 description for ϵ_{ij} and $gR0_{ij}$ (using a n_{ij}/m_{ij} combination) and optimizes r_{ij}^0 for each heavy atom type pair specifically. Moreover, as described above, all pseudo-atom pair interactions include the attractive and repulsive terms. To make it compatible with sOPEPv1, the initial value for ϵ is set at 0.05 kcal/mol, similarly to side-chain/backbone and backbone/backbone interactions.

6.2.3.3. Explicit Hydrogen Bond. Hydrogen bonds are considered explicitly in the OPEP family of potentials. sOPEPv2 keeps the same formulation as sOPEPv1: a hydrogen bond

between residue (i) and residue (j) is characterized by the hydrogen/acceptor distance r_{ij} and the donor/hydrogen/acceptor angle α_{ij}. The hydrogen bond potential is defined as follows:

$$E_{HB}\left(r_{ij},\alpha_{ij}\right) = \epsilon_{\alpha}^{HB} \sum_{ij,j=i+4} \mu(r_{ij}) \cdot \nu(\alpha_{ij}) + \epsilon_{\beta}^{HB} \sum_{ij,j>4} \mu(r_{ij}) \cdot \nu(\alpha_{ij}) \tag{6.2.3}$$

$$\mu(r_{ij}) = \epsilon_{ij} \cdot \left[5\left(\frac{\sigma}{r_{ij}}\right)^{12} - 6\left(\frac{\sigma}{r_{ij}}\right)^{10}\right] \tag{6.2.4}$$

$$\nu(\alpha_{ij}) = \begin{cases} \cos(\alpha_{ij}) & \text{if } \alpha_{ij} > 90^{\circ} \\ 0 & \text{otherwise} \end{cases} \tag{6.2.5}$$

where σ is the position of the potential minimum and ϵ is the potential depth. We distinguish between α-helix-like hydrogen bonds defined by O(i)-H(i+4) and other hydrogen bonds. Hydrogen bonds between a pair of residues separated by less than four amino acids are not considered.

sOPEP also includes a cooperativity term between hydrogen bonds motifs present in secondary structure. In sOPEPv1, the cooperativity formulation involves a per-residue cooperativity propensity associated with α-helix and β-sheet [**148**, **165**]. sOPEPv2 integrates the cooperativity formulation of sOPEPv1 but does not include a residues-specific cooperativity propensity.

The cooperativity, which involves pairs of hydrogen bonds (between residues (i) and (j) and residues (k) and (l)), is used to stabilize secondary structure motifs and distinguishes between α-helix cooperativity and β-sheet cooperativity. The cooperativity energy is given by the following:

$$E_{coop}(r_{ij},r_{kl}) = \epsilon_{\alpha}^{coop} \sum C(r_{ij},r_{kl}) \times \Delta(ijkl) + \epsilon_{\beta}^{coop} \sum C(r_{ij},r_{kl}) \times \Delta'(ijkl)$$

$$C(r_{ij},r_{kl}) = \exp\left(-0.5\left(r_{ij}-\sigma\right)^2\right) \cdot \exp\left(-0.5\left(r_{kl}-\sigma\right)^2\right)$$

$$\Delta(ijkl) = \begin{cases} 1 & \text{if } (k,l) = (i+1,j+1) \\ & \text{and } (j,l) = (i+4,k+4) \\ 0 & \text{otherwise} \end{cases}$$

$$\Delta'(ijkl) = \begin{cases} 1 & \text{if } (k,l) = (i+2,j-2) \\ & \text{or } (k,l) = (i+2,j+2) \\ 0 & \text{otherwise} \end{cases}$$

6.2.4. Optimization Protocol

The optimization of the sOPEPv2 parameters follows the basic optimization scheme developed for earlier versions of OPEP [**148**] and sOPEP [**148**]. The optimization process is designed to allow sOPEPv2 to discriminate between conformations almost identical to the experimental structure (native conformations), conformations resembling the experimental structure (near-native conformations) and the rest of the possible conformations (non-native conformations) without imposing additional biases associated with intermediary approximations such as all-atom forcefields.

More specifically, the decoys' classification defines a remarkably simple set of inequalities :

$$\begin{aligned} E(N_i) &< E(L_j), \text{for all i,j} \\ E(N_i) &< E(M_k), \text{for all i,k} \\ E(L_j) &< E(M_k), \text{for all j,k} \end{aligned} \tag{6.2.6}$$

where $E(X)$ is the sOPEP energy of a decoy X_h, being the $h = i,j,k$ element of the $X = N,L,M$, where N, L and M correspond to Native, Near-Native and Non-Native class of decoys, respectively (see below). The optimization scheme uses these inequalities to classify an ensemble of decoys, on which the parameters are optimized.

The following sections will describe how: (1) decoys are classified; (2) protein targets are selected for the parametrization/validation ensemble; (3) decoys are generated for each protein targets; and finally (4) the optimization score and protocol are defined.

6.2.4.1. Decoys Classification. In order to define the set of inequalities given in Equation 6.2.6, it is necessary to adopt a criterion for decoys classification as there no unique way to set up the classes. The optimization of sOPEPv1 [**148**, **165**] used a decoy classification based on the TMscore [**169**]. Here, we select, rather, the CAD-score, a score based on the similarity of interatomic contacts [**136**, **138**], to classify decoys into the non-native, near-native and native class. This score presents features that make it particularly suitable for optimizing sOPEPv2 : (i) it is based on interatomic contacts, and it was shown (ii) to be more accurate both in terms of the HB network similarity and (iii) to give a more realistic stereochemical features according to MOLPROBITY [**141**] as compared to other highly used score such as the TMscore (sOPEPv1), the GDT-TS and the RMSD [**138**]. These features are well aligned with the sOPEP forcefield, which is based on inter-atomic interactions, explicit hydrogen bonds and hydrogen bond cooperativity in secondary structure.

The CAD-score [**136**, **138**] is defined as follow:

$$CAD = 1 - \left(\frac{\sum_{(i,j)} CAD^{bounded}_{(i,j)}}{\sum_{(i,j)} T_{(i,j)}} \right)$$
$$CAD^{bounded}_{(i,j)} = \min(CAD_{(i,j)}, T_{(i,j)})$$
$$CAD_{(i,j)} = |T_{(i,j)} - M_{(i,j)}|$$

where $T_{(i,j)}$ and $M_{(i,j)}$ are the contact area between residue (i) and (j) for the target structure and model structure respectively. The contact area is estimated using a Voronoi diagram of the heavy atoms described by hard-sphere with radius corresponding to their van der Waals radius. In order to compute the CAD-score in the coarse-grained representation of OPEP, we define new atom types corresponding to the OPEP side-chains with the radius taken from sOPEPv1. In the following, we refer to this score as CAD-CG.

Our classification is based on two main elements. We first consider the empirical distribution of the all-atom CAD score (CAD-AA) (and the cumulative distribution) presented by Olechnovic *et al.* [**138**]. The overwhelming majority of the score is distributed between values of 0.3 and 0.7. More than 80% of the structures have a CAD-AA below 0.60 while more than 90% of the structures have a CAD-AA below 0.65. The second factor is based on the highest CAD-CG predictions generated by sOPEP1/Lib1. After visual inspection, we determine, in agreement with Olechnovic's observations, that a CAD-CG above 0.60 is associated with largely correct secondary structure predictions while a CAD-CG of above 0.65 is associated with accurate secondary and tertiary structure. Thus, the Native, Near-Native and Non-Native classes are characterized by CAD-score ranges of [0.65,1.00], [0.60,0.65[and [0.00,0.60[respectively.

6.2.4.2. Selection of protein targets. The parametrization ensemble for optimizing sOPEPv1 contained 12 proteins or protein fragments; 1abz(α, 40 aa), 1dv0(α, 47 aa), 1e0m (β, 37 aa), 1orc (α/β, 71 aa), 1pgb (α/β, 56 aa), 2gb1f (a β fragment of 2gb1 spanning residue 41-56), 1qhk(α/β, 47 aa), 1shg(β, 62 aa), 1ss1 (α, 62 aa), 1vii (α, 36 aa), 2ci2 (α/β, 83 aa), and 2cro-fisa (α, 71aa) [**148**].

In order to improve the original sOPEP parametrization ensemble, we probe the *Protein Data Bank* for protein targets with the following characteristics: sequences that (1) have 70 amino acids or less, (2) are monomers, (3) contain only standard amino acids, (4) have a structure determined in a PH between 5.5 and 8.5, (5) are not membrane proteins, (6) are not making interactions with ions or ligands and (7) show no more than 30% sequence similarity with others in the set. An additional 6 targets with more than 30% sequence similarity were added to the validation ensembles when they were considered in previous PEP-FOLD's publication [**147**, **148**, **176**] (see SM for listing).

This leads to 135 protein targets, that we further divide into a parametrization and a validation ensemble. For each protein target, we generate the reference structure for the CAD-CG score computation in the following manner: we extract the first model from the *Protein Data Bank*, we minimize it using the all-atom forcefield AMBER99sb*-ILDN [**196**] using the GROMACS software [**58**] and then we convert it to sOPEP coarse-grain representation.

To try and minimize potential problems in the SVM part of PEP-FOLD and really focus on the potential optimization, we further classify the targets based on whether or not sOPEP1/Lib1 is able (i) to generate native predictions for the target, irrespective of their energy, and (2) to correctly assign a low energy to the native prediction with respect to near and non-native structures. The classification protocol is presented in Figure 6.2. For each target, we first generate 500 PEP-FOLD predictions with Lib1/sOPEPv1 and assign the resulting structure to one of the three classes using the CAD-CG (see section Decoys Classification).

If the lowest energy prediction is in the same class (Native/Near-Native) as the best generated prediction, the sequence target is placed in the *Generated/Correctly Classified* (G/CC) category because PEP-FOLD predictions (Lib1/sOPEPv1) is already able to provide reliable folding for this target. If, on the contrary, none of the generated structure is classified as native or near-native, the target sequence is placed in the *Not Generated* (NG) category: sOPEP1/Lib1's fails to produce a satisfactory folding. Finally, if predicted native or near-native structures are generated but do not correspond to the lowest energy prediction, the target sequence is placed in the *Generated/Incorrectly Classified* (G/IC) category, meaning that, for this target, sOPEP1/Lib1 is able to generate a good structural prediction, but its energy is high with respect to non-native and near-native structures.

Out of the 135 protein targets, 48, 64 and 23 sequences are placed in the G/CC, G/IC and NG categories, respectively. The full list of targets is presented in the SM.

For the optimization, we focus our attention on the targets from the G/IC ensemble since, for these targets, the potential is the primary hurdle to improvement of the predictions. From the 64 targets of the G/IC ensemble, we select 25 targets with special care given to contact and structural diversity in order to build the parametrization set. These selected targets are: 1b03(β, 18 aa), 1bhi(α/β, 38 aa), 1cpz(α/β, 68 aa), 1e0n(α/β, 27 aa), 1fex(α, 59 aa), 1g2h(α, 61 aa), 1go5(α, 69 aa), 1i6c(β, 39 aa), 1jjs(α, 50 aa), 1spw(β, 39 aa), 1uxd(α, 65 aa), 1wcn(α, 70 aa), 1yiu(α/β, 37 aa), 1z4h(α/β, 66 aa), 1zv6(α, 68 aa), 1zxg(α, 59 aa), 2b7e(α, 59 aa), 2bby(α/β, 69 aa), 2dt6(α, 64 aa), 2fmr(α/β, 65 aa), 2l92(β, 50 aa), 2l93(α/β, 55 aa), 2lma(α, 22 aa), 2mwf(β, 32 aa) and 2ysb(β, 49 aa). More specifically, the optimization ensemble is composed of 11 α-protein, six β-protein and seven α/β-protein. To show the diversity of included contacts, the side-chain contact frequency of the targets in the parametrization ensemble is presented in Figure A.1 . Only the MET-MET contact

is absent from the selected experimental structures. We also note few contacts with CYS, because targets forming disulfide bond are excluded.

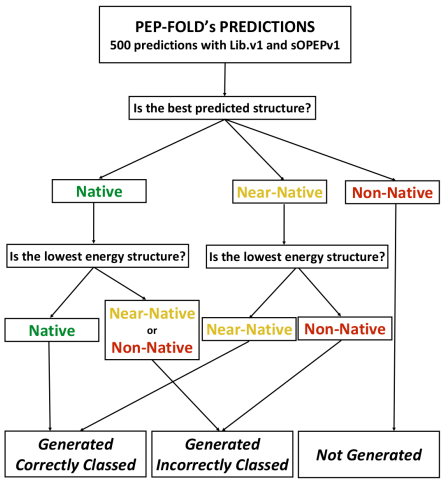


Fig. 6.2. Target classification based on PEP-FOLD predictions. Each protein target is classified in one of three ensemble: *Generated/Correctly Classified* (G/CC), *Generated/Incorrectly Classified* (G/IC) and *Not Generated* (NG). Targets for which the best predicted structure is in the non-native class are placed in ensemble NG (no predicted structure in the Native or Near-Native class). Targets for which the lowest energy structure is in a worst class than the best predicted structure are placed in ensemble (G/IC). Finally, targets for which the lowest energy structure is in the same class as the best predicted structure are placed in ensemble G/CC.

6.2.4.3. Decoys generation. We then generate decoys on which to optimize the parameters for each protein targets identified previously.

In sOPEPv1 optimization, decoys were generated using multiple techniques: (1) molecular Dynamics simulation, (2) threading, (3) greedy assembly and (4) simulated annealing [**165**]. Between 430 and 928 were generated for each target for an average of 550 decoys per target.

In the present work, all decoys are generated directly with the PEP-FOLD's protocol. We use 500 sub-optimal sequences in the SA letters space generated using the FBT algorithm. For the greedy algorithm, we use a heap size of 300, of which 100 are selected based the sOPEP energy while 200 are randomly selected. Finally, the structures are refined using 30 000 Monte-Carlo steps.

6.2.4.4. Parameters optimization. Using the inequalities of equation 6.2.6 based on our decoys classification, we define the optimization score as follow:

$$\text{Score} = -1.0 \cdot \frac{N_{tot}}{T} \sum_{t}^{T} \frac{1}{N_t} \sum_{i}^{C} \sum_{ii}^{D_t^i} \sum_{j>i}^{C} \sum_{jj}^{D_t^j} H\left(E_{jj} - E_{ii}\right)$$

where N_{tot} is the total number of inequalities, T is the total number of targets, N_t is the number of inequalities associated with target t, C is the total number of decoys' class (Native, Near-Native and Non-native) included in the evaluation and D_t^i is the number of decoys for target t in class i. The sum over i and j is done over all decoys' class from Native to Non-Native. $H\left(E_{jj} - E_{ii}\right)$ is the Heaviside function and equals 1 if the energy of decoy ii is smaller than the energy of decoy jj: $E_{jj} - E_{ii} > 0$. To prevent that improvement in the resolution of inequalities be dominated by a single target with more decoys, the score is normalized by the total number of inequalities for each target N_t.

The optimization of sOPEPv2 parameters is done using Particle SWARM Optimization (PSO) [**197**] as implemented in the *pyswarm* python package. This optimization technique works by moving a set of particles (here, the parameters' value), each representing a candidate solution, iteratively in the search space according to the following velocity and position equations:

$$\overrightarrow{V}_i(t+1) = \omega \overrightarrow{V}_i(t) + \phi_p r_p \left(\overrightarrow{p}_i - \overrightarrow{X}_i\right) + \phi_g r_g \left(\overrightarrow{g}_i - \overrightarrow{X}_i\right)$$
$$\overrightarrow{X}_i(t+1) = \overrightarrow{X}_i(t) + \overrightarrow{V}_i(t+1)$$

where ω, ϕ_p, ϕ_g represent the inertia, the "cognitive" coefficient and the "social" coefficient, respectively. $\overrightarrow{p}$ is the best position visited by each particle individually and $\overrightarrow{g}$ is the global best position visited by the swarm. r_p and r_g are real random number between 0 and 1. In this work, we use the default parameters: $\omega = 0.5$, $\phi_p = 0.5$ and $\phi_g = 0.5$ with initial velocities set randomly according to a uniform distribution.

Only a small fraction of the parameters were directly optimized for sOPEPv1 : the parameters $r0$ and $gR0$ for the repulsion/dispersion interactions were determined directly from the distances distributions computed on the *Protein Data Bank* and were not optimized further; only the ϵ parameters of the repulsion/dispersion interactions were optimized [**148**].

For sOPEPv2, all parameters are re-optimized. The pair potential involves 300 pairs of heavy atom type (210 side-chain/side-chain, 80 side-chain/main-chain and 10 main-chain/main-chain), each each associated with 4 parameters — ϵ, r0, and the n and m (giving the value of gR0), for a total of 1200 parameters. For the HB and cooperativity interactions, we have a total of five parameters for the $\epsilon_{\alpha/\beta}^{HB}$, $\epsilon_{\alpha/\beta}^{coop}$ and σ. Finally, for the ϕ/ψ potential, we have the two ϵ and the 40 lower/higher limits $\phi_{low/high\_sc\_i}/\psi_{low/high\_sc\_i}$ (one per amino acid). In preliminary tests to maximize the speed and efficacy of the optimization, increasing the number of particles from 100 to 250 improves the best score by $\sim 25\%$. Further increasing the number of particles from 250 to 500 and to 1000 leads to more modest improvements of respectively, $\sim 5\%$ and $\sim 8\%$. Therefore, we select to use 500 particles for a maximum of 75 iterations or until the score is stable for 10 iterations, whichever comes first. A final

optimization step is also tested with 750 particles, but no improvements to the final score is noted.

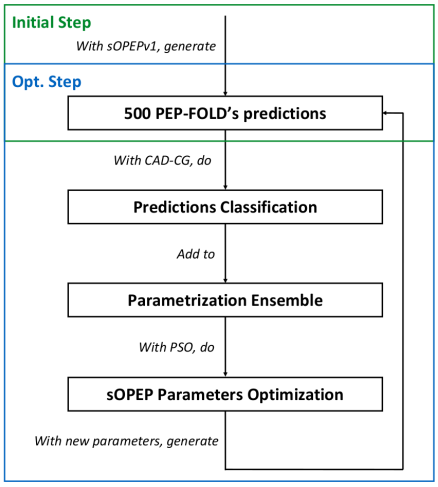


Fig. 6.3. Overview of the optimization protocol. In the initial step of the optimization (top green frame), the parametrization ensemble composed of 500 PEP-FOLD predictions for each targets. The sOPEP parameters are then optimized utilizing an iterative procedure (bottom blue frame), in which the parametrization ensemble is improved by adding newly generated PEP-FOLD predictions.

6.2.4.5. Iterated optimization procedure. To take onto account the close relationship between the conformational search and the forcefield, we use an iterative optimization procedure that is based specifically on PEP-FOLD generated decoys, as described Figure 6.3.

The optimization process for sOPEPv2 is presented in Figure 6.3. Each iteration involves the following three steps:

(1) All 1287 (1200 non-bonded, five for hydrogen bonds and cooperativity and 82 for dihedral angles) sOPEP parameters are optimized. Ten independent optimizations (randomly generated SWARM's positions and velocities) are launched. Only the optimized parameters leading to the best score are used for the next steps.

(2) With the optimized parameters, new PEP-FOLD predictions are generated, as described in section Decoys generation on the protein sequences of the parametrization ensemble.

(3) These newly generated decoys, that reflect the biases of the optimized potential, are added to the optimization ensemble. This approach allows the fitting procedure to include regions of the search space that could be available using the new parameters, mainly new wrong predictions with good energies.

In the full optimization cycle for sOPEP2, this whole procedure is repeated five times leading to stable results. After the update of the library of fragments, from v1 to v2, we further optimize the bonded parameters, taking into consideration the difference in the local

superposition of the new fragment. To reinforce this improvement, we use a more stringent score that requires a threshold of 0.6 for BC-WDC [**137**] in addition to the CAD-CG for the definition of native and near-native decoys. The BC-WDC is a non-local score based on the volume defined by the tetrahedron formed by trio of Cα and the geometric center of the protein. This added constraint helps with the identification of correct domain orientation, for which a local score such as the CAD score is less sensitive to [**138**].

6.3. Comparison with state-of-the-art techniques

In order to probe the quality of PEP-FOLD's predictions on small peptides, we compare our results with three state-of-the-art machine learning techniques: the APPTest server [**155**], the RaptorX-server [**152**–**154**] and AlphaFold2 [**12**].

APPTest uses constraints on distances and dihedral angles determined with a neural network in simulated annealing simulation for the prediction of small peptides' structure [**155**]. It was recently tested against other software for the structural predictions of small peptides and showed a high rate of success.

RaptorX uses an ultra-deep residual neural network (ResNet) on multiple sequence alignment to predict the inter-atomic distances and orientations probability distribution. Then a gradient-based minimization is used to build a 3-D model from the potential derived by ResNet. The RaptorX-server had excellent results in CASP12 and CASP13 [**152**–**154**].

AlphaFold2 [**12**] works by feeding a deep neural network with multiple sequence alignment features obtained from the UniRef90 [**151**], BFD [**150**], and MGnify [**149**] databases. One particularity of AlphaFold2 is that is uses a novel attention-based deep learning architecture. This first step results in a sequence specific probability distribution for inter-atomic distances and dihedral angles. The derived potential is then minimized via a gradient descent algorithm. AlphaFold2 showed tremendous results on the targets of CASP14 [**12**].

In order to compare the results with PEP-FOLD, the all-atoms predictions of APPTest, the RaptorX-server and AlphaFold2 are converted into the sOPEP coarse-grained representation (the main-chain stays unchanged) before comparison.

6.4. Results

6.4.1. Updated library of fragments

In the original library (Lib1), a total of 182 4-residue fragments had been identified for the 27 structural alphabet (SA) letters. Seven of these fragments were associated with SA letters corresponding to α-helix (A,a,V,W), while 17 were associated with SA letters corresponding to β-sheet (L,N,M,T,X).

Using the new strategy, and testing with decreasing the clustering cut-off from 2.0 Å^2 down to 1.5 Å^2, the number of separate clusters increases from 161 to 210. We select 1.9 Å^2 as a reasonable compromise between the number of clusters and the effectiveness of structure reconstruction.

This updated library, dubbed Lib2, contains 166 fragments, seven of which are associated with SA letters corresponding to a α-helix conformation and 28 to β-sheets. This increase in the number of fragments associated with β-strands is a direct consequence of the change of strategy in fragment superimposition prior to clustering.

6.4.2. Optimization of the sOPEPv2 parameters

To separate the impact of upgrading the fragment library from that of revising the force field, we first perform a full optimization cycle with five optimization steps, on decoys generated using the original fragment library. Step to step improvements of fraction of unsolved inequalities are presented in Figure 6.4. Before optimization, with 500 decoys per target, 64.5% of the total number of inequalities are solved with the un-optimized second version of the potential, compared to 67.5 % for the original potential. The optimization leads to an improvement of 24.6 % in the number of unsolved inequalities. After five optimization steps, with 2500 decoys per target, associated with a 2.4% improvement in the number of unsolved inequalities, the re-optimized potential is able to solve 75.5 % of the total number of inequalities.

To take into consideration the difference in the local superposition of the new fragments associated with Lib2, we only re-optimize the bonded potential (Phi/Psi parameters) while keeping the non-bonded parameters fixed to their previously optimized values. Additionally, we use a slightly more stringent classification score as described in Section Iterated optimization procedure. The optimization is performed over 500 newly generated PEP-FOLD predictions using Lib2 and sOPEPv2 for each target in the parametrization set. This new optimization step leads to a 2.8 % improvements in the number of unsolved inequalities. Since only a small improvement in the number of unsolved inequalities is observed, in addition to the fact that only the torsion angles parameters are optimized (82 parameters out of 1200 for the non-bonded parameters), we consider the optimization converged after this single step.

The optimization has noticeable impact on the non-bonded energy terms. The exact values of the parameters are provided in the Supplementary Information Table A.1. The updated potential for a few interactions is presented in Figures A.6 and A.4 . The optimization affects the $r0$ values only slightly and its average variation during the optimization optimization is of only -0.01 $\pm$ 0.79. Few large deviations occur for side-chain-side-chain interactions. The largest decrease occurs for the ASP-TRP pair (difference 1.91 Å) and the largest increase

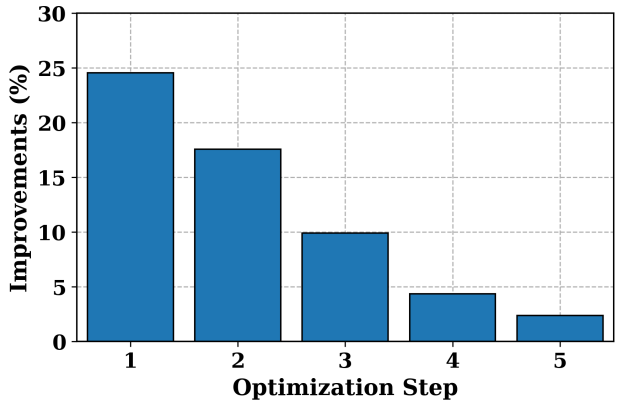


Fig. 6.4. Improvements in the number of unsolved inequalities during the optimization protocol. For each optimization step (x-axis), the improvements are presented as the additional fraction in the number of unsolved inequalities over the previous optimization step.

for HIS-GLN (1.82 Å) and MET-GLN (1.80 Å). For the ϵ parameters, we observe an average deviation of -0.09 ± 0.19 Å, with a largest decrease CYS-CYS (-0.9 kcal/mol) and largest increase for PHE-MET (0.66 kcal/mol). After optimization, ϵ for ASN-LEU and THR-LEU are at the maximum allowed value, indicating that these interactions stay mainly repulsive even with the attractive/repulsive formulation. Larger deviations are observed for $gR0$, that tends to decrease. On average the difference is -0.20 ± 0.67 Å, with minimal and maximal deviations of -2.07 and 1.61 Å for ASN-THR and GLN-SER, respectively. In sOPEPv1, the repulsive strength at shorts distances is controlled by moving the asymptotic divergence around zero instead of directly changing the exponents which are still 12-6 [**148**]. With this in mind, the most striking difference is observed for the exponents n and m. Their variation during the optimization is on average, of -4.96 ± 3.95 and -3.13 ± 2.36, respectively. For close to 160 interactions involving the side chain pseudo-atom, n tends to be close to 4 (compared to the initial value of 12), while m ranges from only 0.6 to close to only 4. Although such exponents are less repulsive than the original 12-6, the fact that the asymptote stays at zero for a MIE potential can still lead to sharper repulsion at shorter distances as can be seen for PHE-PHE, or PHE-TYR in the supplementary Figure A.6 . Strikingly, it is mostly side-chain side-chain interactions that are modified, whereas side-chain backbone interactions or backbone-backbone interactions tend to be less impacted.

Overall, we observe that, for the interactions that were already attractive/repulsive in sOPEPv1, two thirds (62/93) are more permissive at low distances with a smaller value of gR0 in sOPEPv2. These interactions are mainly between apolar/apolar residues (such as ILE-ILE, ILE-LEU, LEU-VAL etc.), between pairs of aromatic residues (PHE-PHE, PHE-TRP, PHE-TYR, TRP-TRP and TRP-TYR) and between some pairs of oppositively charged residues (ASP-ARG, GLU-LYS). The updated potential for a few interactions is presented

in Figure A.4 . For their part, attractive/repulsive interactions that are less permissive in sOPEPv2 are mainly between polar/polar residues (ASN-GLN, ASN-ASN, GLN-TYR etc.) or between polar/apolar residues (MET-GLN, HIS-PRO etc.). For the interactions that with the repulsive formulation in sOPEPv1, modified to an attrative/repulsive formulation in sOPEPv2, we observe that sOPEPv2 is less permissive at shorter distances; energies at 2 Å and 2.5 Å are higher for sOPEPv1 for only 30(/117) and 31(/117) interaction pairs, respectively. These interactions involves mainly the small polar residue SER, with polar and charged residues (ASP, ASN, GLN, ARG, TYR) and the small polar residue THR, as well as the positively charged residues LYS, ARG and, depending on the pH, HIS (HIS-LYS, LYS-ARG, ARG-ARG, LYS-ARG, HIS-ARG).

6.4.3. Impact on structure Prediction

In order to simplify the following discussion, we focus on the results associated with sOPEP1/Lib1 (Lib1/sOPEPv1) and sOPEP2/Lib2 (Lib2/sOPEPv2). Results for the other studied combinations (Lib1/sOPEPv2 and Lib2/sOPEPv1) are all presented in Tables A.2 , A.3 and A.4 .

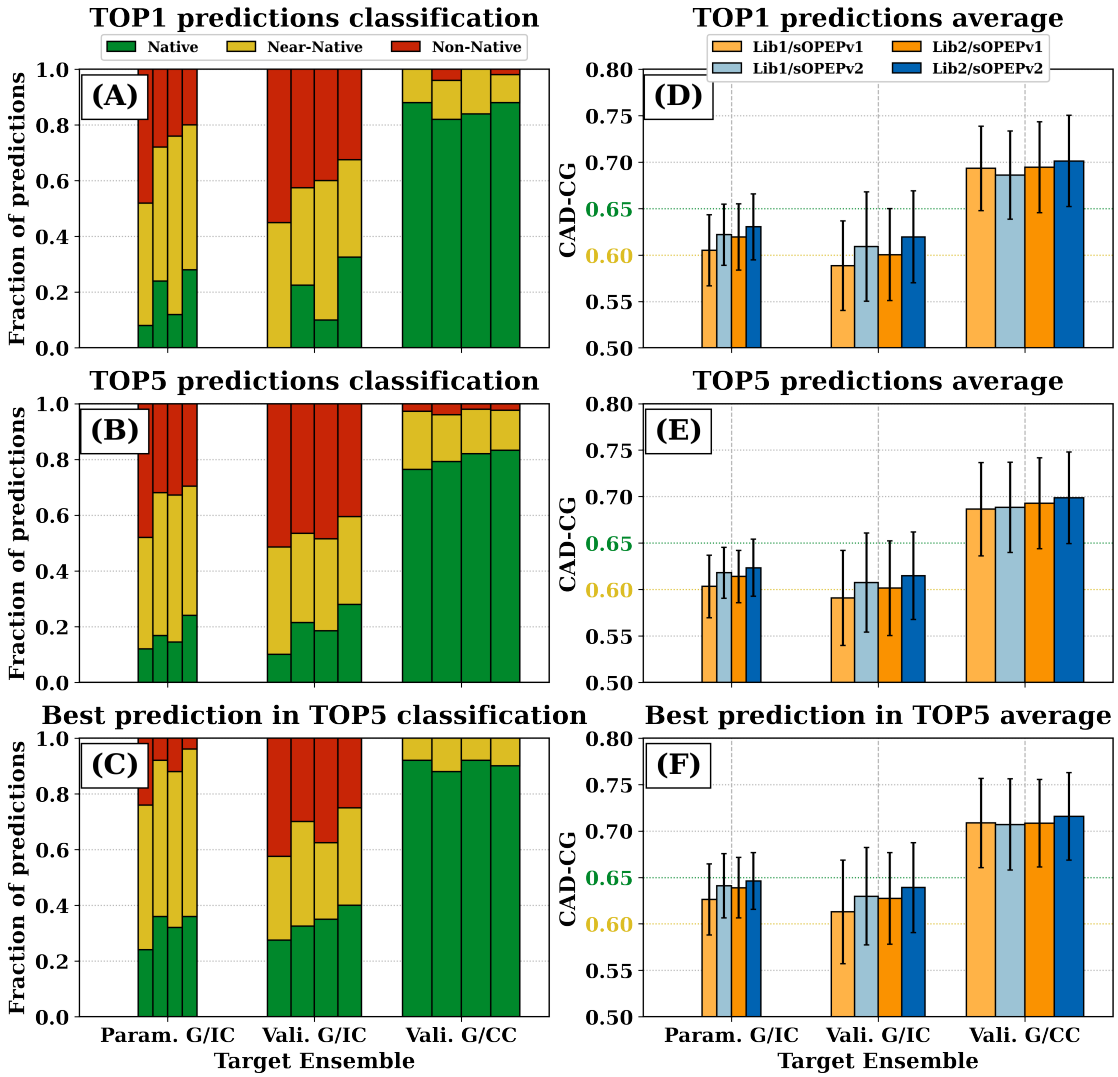


Fig. 6.5. Classification of PEP-FOLD's predictions and average CAD-CG of PEP-FOLD's predictions. x-axis: name of the proteins' sets. *Param. G/IC*, *Vali. G/IC* and *Vali. G/CC* refer to the parametrization, the validation G/IC and the validation G/CC set containing 25, 39 and 48 proteins, respectively. Bar width are proportional to the number of proteins of the set. **Left** side: classification of PEP-FOLD's predictions. The native, near-native and non-native classification are shown in green, yellow and red, respectively. For each set, the four columns represent from left to right, the original library/original potential, the original library/re-optimized potential, the new library/original potential and the new library/re-optimized potential respectively. Panels (A), (B) and (C): fraction of proteins per class considering the lowest energy model only (TOP1), the five lowest energy models (TOP5), the best CAD-CG in the TOP5, respectively. **Right** side: average CAD-CG of 3D predictions. For each set, the four columns represent from left to right, the original library/original potential (light orange), the original library/re-optimized potential (light blue), the new library/original potential (orange) and new library/re-optimized potential (blue) respectively. The CAD-CG associated with the near-native and native classification are shown respectively in yellow and green (y-axis).

The parametrization ensemble. We first consider the impact of sOPEP re-optimization on the 25 protein target sequences retained in the parametrization ensemble (see Table A.2 for details), using either Lib1 or Lib2. Results for this ensemble are presented as the left bar of each panel in Figure 6.5. Overall, one notes an improvement in model quality moving from sOPEP1 to sOPEP2, and an improvement in model quality moving from Lib1 to Lib2. This was expected due to the optimization protocol.

Considering the *best rank only* (TOP1) (Figure 6.5, Panel A), out of 25 targets, the optimization increases the number of targets having native and near-native conformations from 2 and 11 up to 6 and 12, respectively, using Lib1. Using Lib2, this increases up to 7 and 13. The combined impact of sOPEPv2/Lib2 leads to a decrease by more than a factor two in the number of targets with non-native predictions, from 12 (48%) to 5 (20%) targets, The average CAD-CG (Figure 6.5, Panel D), increases from 0.605 ± 0.038 - slightly above the near-native threshold (0.6), up to 0.630 ± 0.035, well into the near-native interval ([0.6,0.65[).

Considering the *best in TOP5* prediction, presented in Panel (C) of Figure 6.5, the same trends are observed. For sOPEP1/Lib1, six targets have a native prediction in the TOP5, and six have only non-native predictions in the TOP5. For sOPEP2/Lib2, nine targets have a native prediction in the TOP5 and only one target has only non-native predictions (1jjs - see Table A.11 and the discussion). The average CAD-CG of the best in 5 predictions - Panel (F) - slightly increases from 0.626 ± 0.038 up to 0.646 ± 0.030.

Finally, Panels (B) and (E) present an analysis over the *five lowest energy* (TOP5) predictions. Using sOPEP1/Lib1, only 12% of the predictions in the TOP5 are native while 48% are non-native. With sOPEP2/Lib2, the fraction of native predictions in the five lowest-energy structures for the parametrization targets increases to 24% while the number of non-native predictions decreases to 30%. The average CAD-CG values increase from 0.603 ± 0.034, slightly above the near-native threshold (0.6) up to 0.623 ± 0.031, corresponding to the near-native definition.

In summary, for the parametrization ensemble, not only the optimization makes it possible to generate better models among the TOP5, but these are of better quality on average. sOPEP2 outperforms sOPEP1, Lib2 outperforms Lib1, and there is an added value in combining sOPEP2 and Lib2.

The validation (G/IC) ensemble. As shown Figure 6.5, the transferability of the improvements observed for the parametrization ensemble to the 40 protein targets of the validation (G/IC) ensemble is obvious (Details regarding this ensemble are presented in Tables A.3 , A.8 and A.12).

Considering only the lowest energy model - Panels (A) and (D), using sOPEP1/Lib1, zero(/40) of the lowest-energy structures correspond to the native state of a target, with 18(/40) classified as near-native and 22(/40) as non-native. With sOPEP2/Lib2, 13(/40)

of the generated lowest energy structures correspond to a native state, 14(/40) are near-native and only 13(/40) are non-native. This is a clear improvement. The average CAD-CG is 0.589 ± 0.048 for sOPEP1/Lib1, a score corresponding to the non-native classification. Using sOPEP2/Lib2, it increases up to 0.620 ± 0.049, well above the near-native threshold of 0.6. As before, sOPEP2/Lib2 gives the best results among the various combinations of force-field / library of fragments.

Considering the *best in TOP5* prediction - Panels (C) and (F), and using sOPEP1/Lib1, 11(/40) sequences have at least one native structure among the prediction with the lowest five energy and 17 have only non-native among those. With sOPEP2/Lib2, 16(/40) sequences have a predicted native structure among the TOP5 and only 10 have only non-native structures among the TOP5. The associated CAD-CG averages are of 0.613 ± 0.056 - corresponding to the near-native class, and of 0.639 ± 0.048, closer to our native threshold of 0.65. The only target with a native prediction in the TOP5 with sOPEP1/Lib1 but only non-native prediction with sOPEP2/Lib2 is 5y22 (see the discussion).

Finally, considering the *five lowest energy* (TOP5) predictions - Panels (B) and (E), 10% of the predictions in the TOP5 are in the native class while 52% of the predictions in the TOP5 are non-native using sOPEP1/Lib1, whereas using sOPEP2/Lib2 almost triples the fraction of native predictions in the TOP5, to 28%, while the number of non-native predictions in the TOP5 decreases to 41%.

In summary, a clear improvement is observed for the targets that were incorrectly ranked using sOPEP1/Lib1.

The validation (G/CC) ensemble. We now look at the 50 protein targets of the validation (G/CC) ensemble that were correctly generated and ranked using sOPEP1/Lib1. Results for this target ensemble are presented in Table A.4 , A.9 and A.13 .

Overall, the results correspond to the expectation of a preserved performance. This is observed in terms of the *lowest energy* prediction - Panels (A) and (D) of Figure 6.5, for which the lowest energy prediction is native for 44(/50) protein targets and non-native for zero(/50) protein targets using sOPEP1/Lib1, while it native for 44(/48) targets and non-native for one structure (1rzs), but with a CAD-CG of 0.597, i.e., very close to near-native with a CAD score, using sOPEP2/Lib2. The average CAG-CG is 0.686 ± 0.050 with sOPEP1/Lib1 with a very slight improvement, at 0.699 ± 0.049, for sOPEP2/Lib2.

Considering the *best in TOP5* prediction, the results are very similar for all potential/library pairs with, for example, 0.709 ± 0.048 for sOPEP1/Lib1 and 0.716 ± 0.047 for sOPEP2/Lib2. The same is observed considering the *five lowest energy* (TOP5) predictions with 76% of the predictions in the TOP5 native and only 3% non-native for sOPEP1/Lib1 and 83% and 2% of the predictions are respectively native and non-native for sOPEP2/Lib2.

Overall, improving predictions of targets correctly predicted by sOPEP1/Lib1 does not lead to a deterioration for those correctly predicted: sOPEP2/Lib2 leads to similar or slightly better predictions than sOPEP1/Lib1 for almost all protein targets tested.
The NG ensemble. Using our classification procedure of the targets, described in section Selection of protein targets, we identified a series of 23 proteins for which sOPEP1/Lib1 is *unable* to generate a native(or near-native) prediction (ensemble NG), irrespective of its energetic classification. These proteins are mainly longer sequences (19 out of 23 are between 50 and 70 amino acids, i.e. longer than the original PEP-FOLD maximal size of 50) dominated by β-sheet secondary structures (10, 9 and 4 out of 23 are respectively β-protein, α/β-protein and α-protein).

Modifications at the level of the library of fragments and the potential have no impact of PEP-FOLD ability to generate native predictions for these targets; as for sOPEP1/Lib1, sOPEP2/Lib2 only generates non-native predictions. In order to better understand where the limitations lie and whether they are related to the discrimination by sOPEPv2, we compute the energy of the experimental structure, following a minimization. The energy of the experimental structure is then ranked compared to the 3D predictions, as shown in Table 6.1. For 14 sequences out of 23, the energy of the native structure is lower than that of the best prediction, with the native structure of an additional sequence being positioned within the top 5 predictions. A more thorough analysis of the significance of these results is provided in section PEP-FOLD's limitations .

6.4.3.1. Improvements. To better understand the underlying effects of revised PEP-FOLD, we focus on the protein targets, among all ensembles, that see their lowest energy prediction change classification when going from sOPEP1/Lib1 (Lib1/sOPEPv1) to sOPEP2/Lib2 (Lib2/sOPEPv2). All predictions using sOPEPv2/Lib2 are available in the SM. A total of 32 targets are shifted to a better class, as shown in Table 6.2: 14 go from non-native to near-native, 12 go from near-native to native and 6 move directly from non-native to native. Only six target sequences move down in classification with sOPEP2/Lib2, with respect to the sOPEP1/Lib1 as shown in Table 6.3: three move from native to near-native and three from near-native to non-native. When analyzing the best structure in the five lowest energy (TOP5), we do however note the presence of at least a native prediction for 2m8j and at least a near-native prediction for the five others (1g2h, 2l4j, 1qpm, 1rzs and 2wqg). The classification of the best prediction in the TOP5 deteriorates for only two out the these six targets (1qpm and 2wqg).

In terms of secondary structure, out of a total of 60 α-targets, 13 are improved (1zv6, 2dt6, 2lma, 1ify, 1q1v, 1zrj, 1zwv, 2bn6, 2coo, 2jof, 2k2a, 2msu and 2cp9) and three are deteriorated (1g2h, 1rzs and 2wqg). Out of the 31 β-targets, 12 of them move up in classification (1b03, 1i6c, 1spw, 2l92, 2mwf, 2ysb, 2jtm, 2k57, 2ysh, 2zaj, 1wr3 and 1wr4) and

NG Target	LowE Energy (kCal/mol)	Native Energy (kCal/mol)	Native Rank
1gyf (α/β, 62)	-138	-184	0
1nd9 (α/β, 49)	-132	-98	501
1ne3 (β, 68)	-137	-149	0
1qxf (β, 66)	-170	-183	0
1vpu (α, 45)	-111	-40	501
1y2y (β, 68)	-149	-38	501
2cw1 (α/β, 65)	-180	-196	0
2do3 (β, 69)	-163	-213	0
2dy8 (α/β, 69)	-158	-173	0
2eqi (β, 69)	-127	-184	0
2gdl (α, 31)	-58	-21	501
2jrr (β, 67)	-147	-169	0
2jtv (α/β, 65)	-150	-182	0
2kaf (α/β, 67)	-176	-218	0
2l8d (β, 66)	-140	-208	0
2lhc (α, 56)	-161	-111	231.5
2lss (α/β, 70)	-162	-251	0
2m2l (α/β, 67)	-147	-138	4.5
2m4y (β, 56)	-128	-70	493.5
2m7o (α/β, 70)	-191	-205	0
2mck (α, 69)	-149	-127	202.5
2mdu (β, 29)	-76	-70	18.5
2xk0 (β, 69)	-158	-174	0

Tableau 6.1. Energy ranking for targets in the *NG* ensemble. LowE and Native Energy: energy of the lowest energy model generated using sOPEP2/Lib2, and experimental structure using sOPEPv2, respectively. Native Rank: ranking of the experimental structure compared to models generated using sOPEP2/Lib2. PEP-FOLD's predictions are ordered from 1 to 500 in order of increasing energy; rank 0 means that the experimental structure has a lower energy than all predictions while a rank of 501 means the experimental structure has a higher energy than all predictions.

two of them (2l4j and 2m8j) down. Finally, out of the 21 α/β-targets, seven of them see improved predictions (2fmr, 1k8b, 1pgb, 2a63, 2kt2, 2l4m and 2v0e) with one of them (1qpm) deteriorating.

Overall, sOPEP2/Lib2 generates improved predictions across targets, irrespective of length: out of the 62 targets below 50 amino acids, 16 targets are improved (1b03, 1i6c, 1spw, 2l92, 2lma 2mwf, 2ysb, 1ify, 1zrj, 2bn6, 2jof, 2msu, 2ysh, 2zaj, 1wr3 and 1wr4) and only two targets move down in classification (2l4j and 2m8j). Similar results are obtained for longer sequences with 15 targets (1zv6, 2d6, 2fmr, 1k8b, 1pgb, 1q1v, 1zwv, 2a63, 2coo, 2jtm, 2k2a, 2k57, 2kt2, 2l4m, 2v0e and 2cp9) out of the 50 between 50 and 70 amino acids moving to a higher classification and predictions for four targets (1g2h, 1qpm, 1rzs and 2wqg) deteriorating.

Improved Proteins				
	sOPEPv1 - Lib1		sOPEPv2 - Lib2	
Target	LowE	Best in TOP5	LowE	Best in TOP5
1b03 (β, 18)	0.567 (0.280)	0.592 (0.496)	0.621 (0.553)	0.621 (0.553)
1i6c (β, 39)	0.589 (0.683)	0.645 (0.776)	0.608 (0.741)	0.608 (0.741)
1spw (β, 39)	0.590 (0.232)	0.652 (0.902)	0.683 (0.841)	0.693 (0.927)
1zv6 (α, 68)	0.545 (-0.159)	0.563 (-0.289)	0.625 (0.516)	0.629 (0.267)
2dt6 (α, 64)	0.630 (0.231)	0.636 (-0.146)	0.650 (0.604)	0.665 (0.718)
2fmr (α/β, 65)	0.562 (0.392)	0.571 (0.043)	0.680 (0.939)	0.680 (0.939)
2l92 (β, 50)	0.579 (0.016)	0.610 (0.845)	0.631 (0.314)	0.647 (0.502)
2lma (α, 22)	0.555 (0.101)	0.560 (0.031)	0.642 (0.051)	0.659 (0.047)
2mwf (β, 32)	0.575 (0.803)	0.625 (0.893)	0.671 (0.881)	0.675 (0.860)
2ysb ((β, 49)	0.642 (0.799)	0.659 (0.861)	0.658 (0.823)	0.697 (0.882)
1ify (α, 49)	0.638 (-0.363)	0.689 (0.912)	0.688 (0.850)	0.706 (0.764)
1k8b (α/β, 52)	0.499 (0.485)	0.526 (-0.023)	0.607 (0.660)	0.607 (0.660)
1pgb (α/β, 56)	0.551 (0.452)	0.580 (0.762)	0.664 (0.898)	0.664 (0.898)
1q1v (α, 70)	0.639 (0.613)	0.678 (0.873)	0.697 (0.904)	0.697 (0.904)
1zrj (α, 50)	0.639 (0.782)	0.682 (0.812)	0.687 (0.843)	0.695 (0.851)
1zwv (α, 58)	0.594 (0.335)	0.627 (0.286)	0.668 (0.877)	0.668 (0.877)
2a63 (α/β, 66)	0.643 (-0.540)	0.648 (0.651)	0.692 (0.945)	0.739 (0.955)
2bn6 (α, 33)	0.585 (0.027)	0.599 (0.007)	0.685 (0.865)	0.701 (0.878)
2coo (α, 70)	0.646 (0.539)	0.654 (0.191)	0.670 (0.761)	0.706 (0.935)
2jof (α, 20)	0.639 (0.294)	0.715 (0.492)	0.680 (0.156)	0.680 (0.156)
2jtm (β, 60)	0.540 (-0.008)	0.589 (-0.188)	0.602 (0.715)	0.669 (0.849)
2k2a (α, 70)	0.632 (-0.043)	0.634 (0.151)	0.685 (0.924)	0.704 (0.924)
2k57 (β, 61)	0.579 (0.304)	0.634 (0.899)	0.631 (0.204)	0.631 (0.204)
2kt2 (α/β, 69)	0.585 (0.764)	0.611 (0.488)	0.607 (0.789)	0.633 (0.624)
2l4m (α/β, 69)	0.546 (0.314)	0.553 (-0.432)	0.604 (0.195)	0.604 (0.195)
2msu (α, 20)	0.572 (0.128)	0.584 (0.170)	0.607 (0.036)	0.624 (-0.004)
2v0e (α/β, 55)	0.570 (0.070)	0.572 (0.244)	0.604 (0.849)	0.612 (0.849)
2ysh (β, 40)	0.646 (0.839)	0.646 (0.839)	0.654 (0.819)	0.665 (0.882)
2zaj (β, 49)	0.635 (0.784)	0.656 (0.824)	0.676 (0.826)	0.695 (0.807)
1wr3 (β, 36)	0.645 (0.837)	0.645 (0.837)	0.664 (0.808)	0.671 (0.882)
1wr4 (β, 36)	0.631 (0.859)	0.657 (0.884)	0.671 (0.879)	0.676 (0.837)
2cp9 (α, 64)	0.636 (-0.361)	0.685 (0.787)	0.689 (0.719)	0.693 (0.761)

Tableau 6.2. Proteins for which the classification of the lowest energy prediction (TOP1) is improved. The notations are identical to that of table 6.1. Columns two and three present the results for the lowest energy prediction and the best prediction in the TOP5 for sOPEPv1/Lib1, while columns four and five present the same results for sOPEPv2/Lib2. Each column present the quality assessment in terms of CAD-CG and, in parenthesis, BC-WDC. Color coding: CAD-CG scores corresponding to the native, near-native and non-native classification are shown respectively in green, yellow and red.

The lowest energy prediction for both sOPEP1/Lib1 and sOPEP2/Lib2 for the 6 sequences that move from non-native to native class are presented in Figure 6.6. Improved predictions can be subtle, introducing a turn or perfecting the alignment, but they can also be fundamental, correcting badly predicted secondary structure as shown by these examples.

Deteriorated Proteins				
	sOPEPv1 - Lib1		sOPEPv2 - Lib2	
Target	LowE	Best in TOP5	LowE	Best in TOP5
1g2h (α, 61)	0.645 (0.447)	0.645 (0.447)	0.579 (0.140)	0.628 (0.428)
2l4j (β, 46)	0.607 (0.787)	0.618 (0.913)	0.593 (0.747)	0.635 (0.863)
1qpm (α/β, 69)	0.685 (0.925)	0.685 (0.925)	0.630 (0.300)	0.646 (0.301)
1rzs (α, 61)	0.631 (0.476)	0.638 (0.665)	0.597 (0.227)	0.625 (0.355)
2m8j (β, 48)	0.652 (0.776)	0.652 (0.776)	0.634 (0.794)	0.666
2wqg (α, 51)	0.703 (0.940)	0.705 (0.925)	0.624 (0.839)	0.646 (0.856)

Tableau 6.3. Targets for which the classification of the lowest energy prediction (TOP1) is deteriorated. The notations are identical to that of table 6.2.

For 1pgb, the sOPEP1/Lib1 prediction only identified two out of the four β-strand and the alignment of the α-helix and β-sheet is off. This prediction has a CAD-CG of 0.551 and a BC-WDC of 0.452. The prediction with sOPEP2/Lib2 correctly identify the four β-strands and their alignment is fairly well reproduced, although a small deviation in the alignment of the α-helix remains. The new prediction has a CAD-CG of 0.664 and and BC-WDC of 0.898. For 1spw, the sOPEP1/Lib1 prediction incorrectly predicts a small helix around residues 34 to 36 and while both small β-strands are present, their alignment is incorrect. For this prediction the CAD-CG is 0.590 and the BC-WDC is 0.232. sOPEP2/Lib2 correctly reproduce the secondary structure elements and their alignment leading to a CAD-CG of 0.683 and a BC-WDC of 0.841.

While almost all secondary structure motifs for 1zwv are correctly predicted with sOPEP1/Lib1, their alignment is completely wrong, leading to a CAD-CG of 0.594 and a BC-WDC of 0.335. It is correctly predicted with sOPEP2/Lib2, leading to a CAD-CG of 0.668 and a BC-WDC of 0.877.

sOPEP1/Lib1 overstabilized α-helical structures for 2bn6, predicting a single straight helix as its best structure leading to low CAD-CG (0.585) and BC-WDC (0.027) scores. sOPEP2/Lib2 correctly generates a turn, breaking the two α-helix, with best predicted structure showing CAD-CG of 0.685 and a BC-WDC of 0.865.

Finally, for 2fmr, sOPEP1/Lib1 has mostly the correct α-helical content, but the β-sheet content is seriously underestimated leading to an overall alignment that is off, as shown by the low CAD-CG (0.562) and BC-WDC (0.392). sOPEP2/Lib2 correctly identify 100% of the α-helical residues and 75% of the β-sheet residues, as well as the overall alignment with a CAD-CG of 0.680 and a BC-WD of 0.939.

The predictions' quality assessment is, in addition to the CAD-CG score, computed with the BCscore [**137**] defined by the residues of the well-defined core (BC-WDC). The results are presented in Figure A.2 for the lowest energy prediction (Panel A), the five lowest energy prediction (Panel B) and the best prediction in the five lowest energy (Panel C). The

improvements observed in terms of CAD-CG are also compatible with the observed trend in terms of BC-WDC. For the lowest energy prediction, the BC-WDC goes from 0.347 ± 0.448 to 0.502 ± 0.380, from 0.330 ± 0.346 to 0.502 ± 0.380 and from 0.624 ± 0.400 to 0.707 ± 0.350 for the parametrization G/IC, the validation G/IC and the validation G/CC ensemble respectively. For the five lowest energy predictions, the BC-WDC goes from 0.327 ± 0.290 to 0.445 ± 0.345, from 0.327 ± 0.290 to 0.445 ± 0.345 and from 0.626 ± 0.361 to 0.721 ± 0.275 for the parametrization G/IC, the validation G/IC and the validation G/CC ensemble respectively. Finally, the BC-WDC of the best prediction in the five lowest energy goes from 0.601 ± 0.274 to 0.652 ± 0.317, from 0.601 ± 0.274 to 0.652 ± 0.317 and from 0.810 ± 0.283 to 0.827 ± 0.193. Theses observations are compatible with those obtained for the CAD-CG.

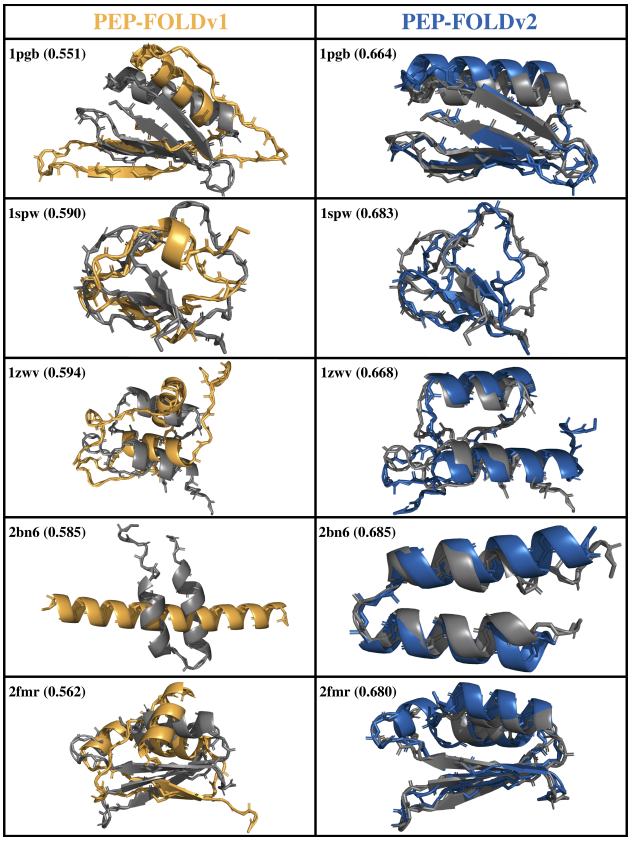


Fig. 6.6. Lowest energy predictions for proteins going from non-native to native predictions. The left and right columns show the results for the original library with the original potential, in orange, and the new library with the new potential, in blue, respectively. The experimental structure is shown in gray. The structures are aligned on C_α of residues of the well-defined core as presented on the *Protein Data Bank*. Pictures were generated using Pymol [**95**] and secondary structure elements were determined using STRIDE [**179**].

6.4.4. Comparison

We now compare the predictions from sOPEP2/Lib2 to three state-of-the-art machine learning technique: APPTest server [**155**], the RaptorX-server [**152**–**154**] and AlphaFold2 [**12**].

APPTest is limited to sequences of 40 amino acids or less. For the very small peptides, 25 amino acids or less, the results with this tool are very good with an average CAD-CG of 0.698 ± 0.084. In the tested targets, the score is mainly dragged down by 1b03, for which the β-sheet secondary structure is not correctly reproduced (CAD-CG of 0.507), and 1s4j, which is predicted as almost fully extended (CAD-CG of 0.532). For larger targets between 26 and 40 amino acids, the prediction quality is decreased compared to the smaller targets with an average CAD-CG of 0.634 ± 0.075. As shown on Panel (A) of Figure 6.7, this is mainly due to the β-sheet targets (CAD-GG score: 0.612 ± 0.077) as α-helical target are very well predicted (CAD-GG score: 0.722 ± 0.046). This is caused by a slight shift in the hydrogen bonds network between the β-strands, leading to incorrect prediction of side-chain/side-chain interactions, captured by the CAD score. Overall, panel (A) and (B) of Figure 6.7 show that APPTest does better on smaller, α-helical targets, while PEP-FOLD does better for larger, β-sheet targets as it predicts the correct hydrogen bonds network between the strands. Only two sequences from the NG ensemble are below the APPTest threshold of 40 amino acids: 2gdl and 2mdu. Similarly to PEP-FOLD using sOPEP2/Lib2, APPTest has trouble with these two targets with a CAD-CG of 0.534 and 0.592, respectively. When considering the best prediction in the TOP5, APPTest results are improved for six targets, notably for 2luf (from near-native) and 2mwf (from near-native) with a native prediction (Table A.13).

For the RaptorX-server a clear distinction of strengths and weaknesses, both in term of secondary structure and length, emerges with respect to sOPEP2/Lib2. Panel (A) of Figure 6.7 shows that the predictions for α-targets of the RaptorX-server are slightly better than those obtained using sOPEP2/Lib2 (average CAD-CG of 0.688 ± 0.082 vs 0.676 ± 0.059), while sOPEP2/Lib2's predictions are more reliable for β-targets (average CAD-CG of 0.647 ± 0.055 for sOPEP2/Lib2 vs 0.559 ± 0.082 for RaptorX-server). In terms of target length, Panel (B) of Figure 6.7 shows that, for smaller targets, between 26 to 50 amino acids, sOPEP2/Lib2 gives better predictions than RaptorX (average CAD-CG 0.680 ± 0.063 vs 0.612 ± 0.108) while RaptorX predictions are better for longer targets, between 50 and 70 amino acids (average CAD-CG of 0.635 ± 0.049 vs 0.665 ± 0.094). For the NG ensemble, RaptorX-server predictions are also below its average over the other sequences : for 9 of the 23 predicted targets in the NG ensemble are classified as native, with one near-native and 13 non-native. Raptor-X results are improved for only two targets, when we focus on the best prediction in the TOP5, notably for 2ysi (from non-native) with a native prediction (Table A.13).

For its part, AlphaFold2's predictions are excellent for all secondary structure type, with an average CAD-CG of above 0.75 for α, β and α/β targets (Figure 6.7(A)) and for all protein lengths, with an average CAD-CG of above 0.70 for targets below 26 amino acids and above 0.75 for targets between 25 and 50 amino acids and between 50 and 70 amino acids (Figure 6.7(B)). For the smallest targets tested, the average of AlphaFold2 is only dragged down by 1b03, for which the hydrogen bond network of the β-sheet is shifted, leading to incorrect side-chain/side-chain interactions and a CAD-CG in the non-native realm (0.583) and 1s4j, which is predicted as a small α-helix as opposed to the native structure with two turns and no secondary structure elements. Finally, AlphaFold2 is the only method tested here that is able to correctly predict the overwhelming majority of the targets in the NG ensembles. Only four are incorrectly predicted by AlphaFold2, 1nd9, 1vpu, 1y2y and 2gdl with a CAD-CG of respectively 0.558, 0.534, 0.494 and 0.595. For AlphaFold2, considering the best prediction inside the TOP5 does change the results with a single target, 2kya, that goes from the non-native to the near-native class (Table A.12).

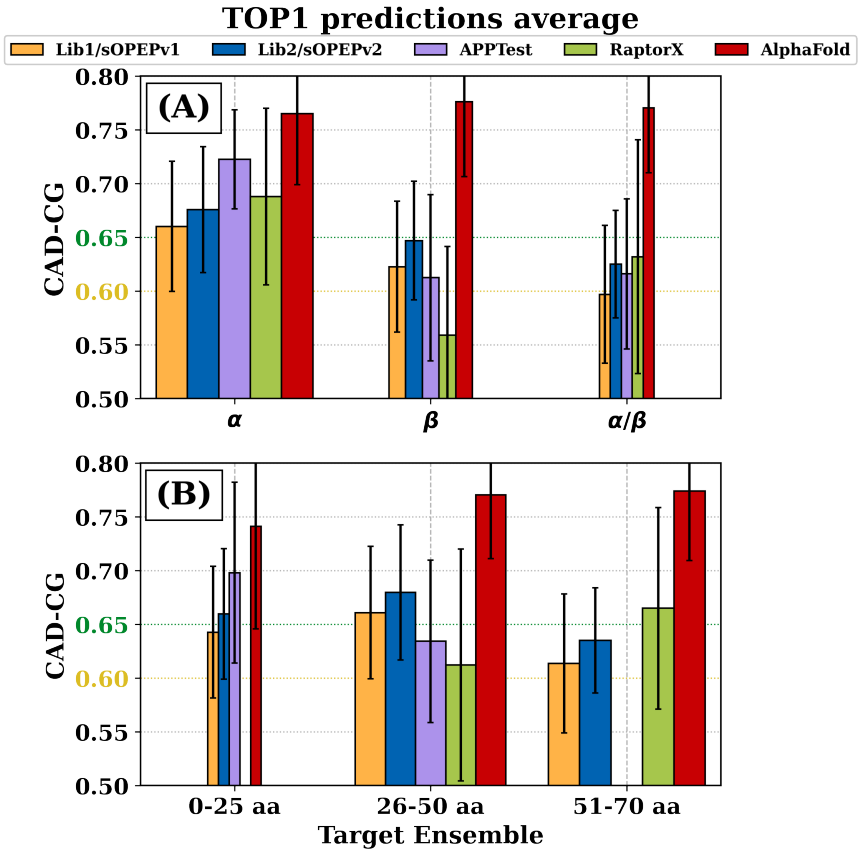


Fig. 6.7. Average CAD-CG score for the TOP1 prediction using five prediction approaches. Panel (A): results when targets are classified by structural class, with respectively 60, 32 and 21 proteins in the α, β and α/β categories. Panel (B): results when targets are classified by length, with respectively 17, 48 and 50 targets with less than 26 amino acids, between 26 and 50 amino acids and between 51 and 70 amino acids. Of note: PEP-FOLD is usually limited to up to only 50 amino-acids. The RaptorX-server minimum accepted length is 26 amino acids. The APPTest does not consider sequences with more than 40 amino acids. The CAD-CG associated with the near-native and native classification are shown respectively in yellow and green (y-axis). Protein targets from the *NG* ensemble are excluded for this figure.

6.5. Discussion

PEP-FOLD [**147**, **148**, **176**] is a quick, simplified and successful approach to peptide's structure prediction that is freely available as a web-server [**177**]. It is specialized in the structure prediction of small peptides up to 50 amino-acids. In this study we question PEP-FOLD applicability from 50 amino acids to 70, but still keep the focus on relatively short sequences as these present a unique challenge compared to the prediction of larger protein [**175**].

In this study, we present improvements to two of the core aspects of PEP-FOLD: an updated library of fragments (Lib2) and a re-optimized version of sOPEP (sOPEPv2). sOPEPv2 introduces an new formulation for non-bonded interactions, and it is parametrized by using a self-consistent iterative process, using a philosophy similar to that developed for optimizing OPEP [**165**] and sOPEPv1 [**148**]. The parametrization is designed to maximize the discrimination on an ensemble of decoys classified using simple criteria on the CAD score, with no further information on the distributions of inter-atomic distances. Despite real, the improvement associated with the update of the library of fragments alone appears limited compared to that brought by generalized formulation used to describe non-bonded interactions, for sizes up to 50 amino acids, as can be seen Table A.5 . Furthermore, despite its simplicity (discrete assembly, coarse-grain potential etc.), the updated version of PEP-FOLD presented here shows improvements compared to state-of-the-art machine learning approaches such as APPTest and RaptorX.

6.5.1. Dependence on target size and secondary structure

sOPEP2/Lib2, with updated fragments library and re-optimized potential, improves the accuracy of predicted structures for targets of 50 amino acids or less compared to sOPEP1/Lib1. The average CAD-CG goes from 0.656 to 0.675, placing most of these proteins in the native class, as we define it, while the average BC-WDC goes up from 0.519 to 0.596 (shown on Figure A.3 and in Table A.5). In spite of the introduction of longer peptides, results for smaller peptides do not deteriorate but improve both in terms of CAD-CG and BC-WDC.

As expected with a focus on longer sequences, sOPEP2/Lib2 delivers a pronounced improvement for peptides between 50 to 70 amino acids of the results in terms of CAD-CG, from 0.614 to 0.635, placing most of these proteins in the near-native class, with the average BC-WDC moving from 0.373 to 0.608. This improvement, that does not, for such sizes, bring PEP-FOLD to the level of performance of an approach such as AlphaFold, strongly suggests, however, that the generalized formulation proposed here is an effective direction. For such large sizes, other aspects of PEP-FOLD can limit the effective generation of accurate models, namely the sequential assembly process in a discrete space.

Considering the secondary structure class of the targets (see Figure A.7 and Table A.6), α-proteins tend to be more often correctly predicted by sOPEP2/Lib2, as compared to β-proteins and α/β-proteins. This trend was also observed with sOPEP1/Lib1, although we nevertheless note important improvement in terms of tertiary structure in terms of CAD-CG and BC-WDC A.7 , and it is likely first related to PEP-FOLD's assembly process rather than to the force field. Indeed, because α-helices are local in structure, correctly predicted structural alphabet letters associated with α-helices can therefore be immediately identified

as favorable, during the amino-acid by amino-acid model generation process, whereas this is not possible for β-strands. More specifically, the lowest-energy structures predicted by sOPEP1/Lib1 and sOPEP2/Lib2 reproduce 93% and 96% of the experimental α-structures, respectively for α-targets (shown on Figure A.8).

By definition, β-targets contain no experimental α-helix. This feature is perfectly reproduced with PEP-FOLD : the folds contain no α-helix, hence a 100% success rate for predictions. For α/β-targets, the predicted amount of correct experimental α-structure is also very high, with 94% for both sOPEP1/Lib1 and sOPEP2/Lib2, respectively.

β-sheets are reproduced perfectly as slight deviations in the hydrogen bonds network between strands can lead to new interactions between the side-chains, a mismatch that can be measured using the CAD-CG score [**138**]. With sOPEP2/Lib2, we can clearly see improvements in the prediction of β-sheets. For the β-proteins, the average CAD-CG goes from 0.623 to 0.647 (shown on Figure A.7) and the average reproduced β-sheet from the experimental structure increases from 0.817% to 0.871 % (shown on Figure A.8). For the α/β-proteins, the change is even more noticeable with the CAD-CG going from 0.597 to 0.625 and the average of reproduced β-sheet content from the experimental structure going from 45% to 78% for sOPEP1/Lib1 and sOPEP2/Lib2 respectively. Beyond the re-optimized potential, the use of the new library of fragments also contributes to this improvement as the number of fragments associated with β-sheet letters of the SA goes from 17 to 28. This leads to more residues adopting the correct β-sheet secondary structure, as shown in Figure A.8, with both sOPEPv2 and, to a lesser extent, sOPEPv1 when using Lib2.

6.5.2. PEP-FOLD's limitations

In spite of the overall prediction improvements realized with the revision of the sOPEP potential, we identify a few proteins for which PEP-FOLD is unable to make a correct predictions within the five lowest energy.

One of these proteins, 1jjs (α, 50 amino-acids), is in the parametrization ensemble. For this target, sOPEPv2/Lib2 over-stabilizes two of the three α-helices (from residue 5-13 and 19-31 and compared to 2-14 and 25-31) and predicts a fourth helix between residues 45-49. Additionally, the alignment of the first helix is off. A near-native structure (CAD-CG of 0.600 and BC-WDC of 0.566), is however present just outside the TOP5, at rank 8, as shown in Table 6.4.

In the validation G/IC ensemble, PEP-FOLD is unable to identify a near-native or native prediction among the five lowest energy structure for 10 out of the 40 targets; five of which are β-protein (1ed7, 1k91, 2m6o, 2mdj and 2mi6), two are α/β-protein (1f0z, 1n87), two are α-protein (2kya, 5y22) and one has no secondary structure elements(1s4j). To identify the source of this difficulty, we compute the energy of the experimental structure with

Target	Native Rank	First non Non-Native Prediction CAD-CG (BC-WDC)	Rank
1s4j (Coil, 13)	501	0.610 (-0.525)	122
5y22 (α, 22)	236.5	0.724 (0.977)	61
2kya (α, 34)	2.5	0.611 (0.017)	119
1k91 (β, 37)	0	0.601 (-0.059)	10
1ed7 (β, 45)	0	0.602 (0.769)	102
2m6o (β, 48)	4.5	0.621 (0.877)	17
1jjs (α, 50)	158.5	0.600 (0.566)	8
1n87 (α/β, 56)	3.5	0.602 (0.871)	10
2mdj (β, 56)	217.5	0.613 (-0.445)	122
2mi6 (β, 62)	0	0.612 (0.693)	13
1f0z (α/β, 66)	0	0.600 (0.867)	77

Tableau 6.4. Ranking of incorrectly predicted targets. For each target incorrectly predicted within the five lowest energy, the ranking of the experimental structure and the quality assessment in terms of CAD-CG(BC-WDC) and ranking of the first non Non-native prediction are presented respectively for column 2 to 4. PEP-FOLD's predictions are ordered from 1 to 500 in order of increasing energy; rank 0 means that the experimental structure has a lower energy than all predictions while a rank of 501 means the experimental structure has a higher energy than all predictions.

the re-optimized potential after relaxation and compare its ranking with sOPEP2/Lib2's predictions. For four of these 10 sequences (1ed7, 1f0z, 1k91 and 2mi6), the experimental structure ranks before the best prediction and for two others (1n87 and 2m6o), the experimental structure ranks in the five lowest energy predictions, as presented in Table 6.4. For 1k91 and 1n87, a near-native prediction is present just outside the TOP5 at rank 10 for both (see Table 6.4).

We now have a look at the remaining sequences. For 1s4j, sOPEPv2/Lib2 predicts a small β-hairpin, similarly to sOPEPv1/Lib1 instead of the two turns and no secondary structure elements of the native structure. For 2kya, the SVM predicts the position of the α-helix around residues 24 to 33 while the experimental α-helix is around residues 12 to 28 and it also predicts non-existing β-strand around residues 11 to 20. Finally, for 5y22, sOPEPv2/Lib2 predicts that the second half of the α-helix, between residue 3-15 in the experimental structure, instead forms a small β-hairpin. With sOPEPv1/Lib1, the second half of the experimental α-helix is instead mainly disordered, except for the fourth prediction which correctly predicts the correct α-helix (see Table A.12). For sOPEPv2/Lib2, the correctly predicted structure is not present in the five lowest energy predictions. 5y22 is the only case for which the results in terms of the best prediction in the TOP5 is deteriorated by using sOPEPv2/Lib2 compared to sOPEPv1/Lib1.

Similarly to what we observed for 2kya, we find some limitation for the SVM on the targets from the Not-Generated (NG) ensemble as we identify multiple incorrect secondary

structure predictions made by the SVM. For example, in 2gdl (α, 31 aa) the SVM predicts the localization of the α-helix around residues 21 to 29, instead of around residues 5 to 18 in the experimental structure. For 1vpu, (α, 45 aa), the experimental helix around residues 23 to 28 is shifted in the SVM predictions to around residues 26-35, in addition to helix between residues 39-43 not being identified by the SVM. Finally, for 2lhc, (α, 56 aa), two of the three experimental α-helix, between residues 9 to 14 and residues 39 to 51, are identified as β-sheet by the SVM.

Together, these results show that the updated library and potentials are able to identify correctly the native structure of most the problematic sequences. This confirms that the simplified representation adopted here, both in terms of structure, including the coarse-graining of the side chain, and interactions, manages to capture the essential features responsible for folding.

The results also show that, for most sequences, the SVM approach to structure prediction excels at generating the relevant structures, both secondary and tertiary, that can then be classified using the energy model. With the current structural alphabet, however, this approach can fail for a relatively small subset of sequences, particularly, for sequences where the tertiary structure is essential to enforce the secondary structure. While a more detailed analysis of these cases could allow us to better understand the delicate balance between these two levels of organization for some sequences, the SVM remains a powerful tool for exploring the structures of peptidic sequences.

6.6. Conclusion

Small peptides can play an important role in the development of novel therapeutic approaches [**127**, **129**] and represent a unique challenge compared to larger proteins. Indeed, the same amino acid sequence can adopt very different structures whether it is a peptide or a fragment of a larger protein [**142**, **175**]. In this work, we present improvements to the popular, freely available online [**177**], PEP-FOLD method for small peptides structure predictions and extend its application from sequences of up 50 amino acids to 70.

These improvements focus of two aspects of PEP-FOLD. First, using a new superimposition and clusterization scheme, we update PEP-FOLD's library of fragments associated with each letter of the structural alphabet (SA). This leads to an overall decrease in the total number of fragments, from 182 to 166 but with a larger number of fragments associated with β-sheet letters (from 17 to 28). Second, the parameters of the sOPEP force field, used in PEP-FOLD for prediction's classification during (and after) greedy assembly of the fragments, are re-optimized using an iterative self-consistent process. sOPEP2/Lib2 leads to improved predicted structures for targets found problematic with sOPEP1/Lib1, both in terms of the lowest energy and five lowest energy prediction, while maintaining the

quality for targets already correctly predicted by sOPEP1/Lib1. While PEP-FOLD is the only approach of this study not going to the all-atom level and using a discrete space search, sOPEP2/Lib2's predictions compare well with other state-of-the-art protein/peptide structure prediction techniques — the recently developed APPTest [**155**], RaptorX [**152**–**154**] — but is behind the recently proposed AlphaFold2 [**12**].

Therefore, with its overall high reliability for shorter sequences (50 amino acids and less), the original approach retained by PEP-FOLD, including the use of a structural alphabet, of a sequential growth algorithm and of a rich coarse-grained potential optimized using a very general classification scheme, this improved version of PEP-FOLD offers a solid prediction tool that can provide physical insights onto the folding process. The analysis presented with this revised parametrization underlines, in particular, the importance of better understanding the link between between tertiary and secondary structure, particularly for these smaller fragments, but also the strength of the local approach retained here. As updated, sOPEP2/Lib2 remains, therefore, an important tool for structure prediction of short sequences. In addition, the quality of the structure prediction provides a strong support for the simplified sOPEP2 potential, developed here, that could serve as a solid basis for dynamical studies, unreachable by purely IA folding techniques.

6.7. Acknowledgement

Vincent Binette is grateful to the Fonds de recherche du Québec – Nature et technologie through a post-graduate fellowship and to MITACS for a travel grant. This program is partially support through a Discovery Grant of the the Natural Sciences and Engineering Research Council of Canada to Normand Mousseau. This research was made possible through generous computer allocations from Calcul Québec (www.calculquebec.ca) and Compute Canada (www.computecanada.ca).

6.8. Supplementary material

Detailed comparison for all tested targets for all tested prediction techniques.

6.9. Addendum

To test the statistical significance of the improvements to PEP-FOLD, the p-values between Lib1-sOPEPv1 and Lib2-sOPEPv2, associated with Figure 6.5 are reported in Table 6.5. For the Param. G/IC and Vali. G/IC, the improvements are statistically significant with small p-values (<0.05). On the other hand, for already correctly predicted targets of the Vali. G/CC, the differences are not statistically significant with high p-values

(>0.05). Taken together, these results show that with Lib2 and sOPEPv2, the new PEP-FOLD improves significantly the predictions for problematic targets for Lib1-sOPEPv1 while preserving the quality of the correctly predicted targets with Lib1-sOPEPv1.

Finally, the analysis of the statistical significance was also realized for the comparison between PEP-FOLD (Lib2-sOPEPv2) and other popular prediction techniques (APPTest, RaptorX and AlphaFold2), shown in Figure 6.7. The results are presented in Table 6.6. Between PEP-FOLD (Lib2-sOPEPv2) and APPTest, we see that the differences are statistically significant for the β-targets and the targets of medium length (26-50aa) with small p-values (<0.05). On the other hand, the differences are not statistically significant for the α-targets, the α/β-targets and the small targets (0-25aa). Between PEP-FOLD (Lib2-sOPEPv2) and Raptor-X, we observe significant differences for β-targets, the medium length targets (26-50aa) and the high length targets (51-70aa). For the first two classes, PEP-FOLD is giving better results while for the third classes, RaptorX gives better results. Finally, we observe that the improvements obtained by using AlphaFold2 are statistically significant for all target classes.

	p-value		
	Param. G/IC	Vali. G/IC	Vali. G/CC
TOP1	0.02	0.01	0.41
TOP5	0.04	0.03	0.22
Best in TOP5	0.05	0.03	0.46

Tableau 6.5. Statistical significance of PEP-FOLD's improvements. The statistical significance was computed on the CAD-CG score of the TOP1, TOP5 and best in TOP5 predictions (row 1 to 3), as presented in Figure 6.5. Results that are significantly better for PEP-FOLD with Lib2 and sOPEPv2 are presented in green.

	p-value		
	APPTest	RaptorX	AlphaFold2
α	0.06	0.51	0.00
β	0.02	0.00	0.00
α/β	0.61	0.80	0.00
0-25 aa	0.10	-	0.01
26-50 aa	0.04	0.00	0.00
51-70 aa	-	0.05	0.00

Tableau 6.6. Statistical significance of the comparison between PEP-FOLD and other popular prediction techniques. The statistical significance was computed on the CAD-CG score of the TOP1 predictions, as presented in Figure 6.7. The top and bottom parts of the table present the results in term of secondary structure and sequence length respectively. Results that are significantly better for PEP-FOLD are presented in green.

Chapitre 7

Préambule au développement de aaOPEP

Tel que discuté aux chapitres précédents (chapitre 5), les systèmes biologiques d'intérêt peuvent être particulièrement complexes à étudier à l'aide de méthodes traditionnelles. Les potentiels gros-grain sont des alternatives aux potentiels tout-atome classiques et permettent d'étudier des systèmes plus gros sur des échelles de temps plus longues en apportant des simplifications importantes au niveau de la représentation du système et de ses interactions. Les potentiels OPEP, présentés en détail à la section 5.2, sont des potentiels gros-grain qui furent utilisés avec succès dans de nombreuses applications [**147**, **148**, **158**–**164**]. Notamment, les améliorations apportées à sOPEP ont permis d'améliorer de façon considérable les prédictions de la méthode PEP-FOLD [**198**]. Ainsi, ces potentiels sont des outils intéressants pour étudier des systèmes offrant des défis particuliers aux méthodes traditionnelles, tant expérimentales que numériques. Les systèmes moléculaires à l'origine de la maladie d'Alzheimer sont un de ces défis particuliers. De ce fait, leurs études requièrent le développement de méthodes adaptées et simplifiées. Pour s'attaquer à ce problème, nous développons le potentiel simplifié aaOPEP dont le développement se fera spécifiquement pour l'étude de la maladie d'Alzheimer. aaOPEP portera la philosophie des potentiels gros-grain OPEP en régime tout-atome.

Le présent chapitre présentera les notions théoriques et méthodologiques pertinentes pour le développement de aaOPEP. Il est divisé en deux sections. Dans la première section, je présenterai les processus moléculaires associés à l'apparition de la maladie d'Alzheimer et j'expliciterai les raisons qui font de ces systèmes un défi particulier pour les méthodes numériques traditionnelles. Dans la seconde, je présenterai la méthode d'échange de répliques, une méthode d'échantillonnage avancée que nous utilisons dans le développement de aaOPEP

7.1. Maladie d'Alzheimer

Aux États-Unis, la maladie d'Alzheimer est la sixième cause de décès la plus importante et la cinquième chez les 65 ans et plus, une proportion qui va en augmentant avec le vieillissement de la population [**199**]. Parmi les symptômes les plus importants de la maladie, on retrouve les pertes de mémoire, les changements d'humeur et de personnalité, les difficultés de cognition et la confusion [**199**]. L'apparition des symptômes est associée à la détérioration et à la mort des neurones. Encore aujourd'hui, aucun traitement ne permet de prévenir ou de ralentir cette destruction des neurones à l'origine de la maladie [**199**–**202**].

7.1.1. Amyloïde-β

Au niveau moléculaire, la maladie d'Alzheimer est caractérisée par l'accumulation de fibres amyloïdes, composées du petit peptide amyloïde-β (Aβ), à l'extérieur des neurones, perturbant le processus de la synapse, et l'accumulation de protéines tau à l'intérieur des neurones, empêchant le transport de nutriments et le fonctionnement normal des neurones [**199**].

Selon l'hypothèse de la cascade amyloïde, c'est l'accumulation de fibres amyloïdes qui serait à l'origine de l'apparition de la maladie [**200**]. Par contre, plus récemment, l'hypothèse de la cascade amyloïde est modifiée et l'origine de la toxicité serait plutôt associée aux petits oligomères d'Aβ [**4**, **200**].

Au niveau physiologique, Aβ est issu du clivage d'une grande protéine membranaire, appelée protéine précurseur de l'amyloïde ("amyloid precursor protein", APP), par les enzymes β-sécrétases et γ-sécrétases [**201**, **203**]. Cette dernière joue des rôles importants dans le développement du cerveau, la régulation des synapses, le transport du fer, etc. [**201**, **203**]. Quant à lui, Aβ, en conditions normales, joue un rôle important au niveau du système immunitaire [**203**] ou est crucial pour certaines capacités cérébrales comme la mémoire [**201**]. Les monomères d'Aβ s'assemblent les uns avec les autres pour former des oligomères de tailles variées, du dimère au dodécamère, ou des structures plus grandes comme les fameuses fibres amyloïdes caractéristiques de la maladie [**203**].

7.1.2. Protéine intrinsèquement désordonnée

Comme la toxicité serait principalement associée aux petits oligomères d'Aβ [**4**, **200**], l'étude de ceux-ci est donc cruciale pour le développement thérapeutique. Or, l'étude d'Aβ représente un défi particulièrement important. En effet, la vision classique de la biologique moléculaire stipule que c'est la structure des protéines qui est à l'origine de leurs fonctions. Or, des nuances furent apportées à cette théorie suivant la découverte de protéines sans structure tridimensionnelle stable, mais jouant tout de même de nombreux rôles importants chez la cellule. Ces protéines, appelées protéines intrinsèquement désordonnées (PID), pourraient représenter jusqu'à 30% du protéome humain [**204**] et représentent un défi particulier pour

les méthodes expérimentales classiques; Aβ fait partie de la classe des PIDs. En effet, les méthodes expérimentales classiques offrent généralement une vision d'ensemble, c'est-à-dire que les résultats correspondent à une moyenne sur les conformations présentes. Or, les PIDs sont caractérisées par un ensemble conformationnel vaste et hétérogène mal caractérisé par la moyenne [**4**, **204**]. À cause de son caractère désordonné et de sa propensition à l'agrégation, l'étude des petits oligomères d'Aβ à l'aide de méthodes expérimentales est particulièrement complexe [**8**, **202**] et la majorité des données expérimentales sont de faible résolution.

En plus de représenter un défi important au niveau expérimental, les PIDs représentent aussi un défi particulier pour les méthodes numériques. En effet, de nombreuses simulations au niveau des PIDs ont mis en lumière de multiples lacunes dans la paramétrisation des champs de force classiques [**204**]. De plus, le phénomène d'agrégation et de fibrillation d'Aβ est fondamentalement multi-échelle, tant au niveau temporel que spatial [**203**]; du monomère d'Aβ contenant quelques centaines d'atomes jusqu'à la fibre amyloïde pouvant se composer de millier de monomères. Pour pouvoir étudier des systèmes plus grands, sur des échelles de temps plus longues, il faut donc développer des méthodes simplifiées.

7.2. Échange de répliques Hamiltonien

Bien que fortement utilisée, la dynamique moléculaire, telle que discutée à la section 2.2, a des limitations. En plus des très longues échelles de temps requises pour étudier la majorité des phénomènes biologiques d'intérêt, la probabilité d'observer un système dans un certain état est donnée par les facteurs de Boltzmann $P(r) \propto \exp -\frac{U(r)}{k_B T}$. Ainsi, les systèmes pour lesquels les barrières énergétiques sont beaucoup plus importantes que $k_B T$ ont tendance à rester "piégés" dans des états métastables. Diverses méthodes dites méthodes avancées d'échantillonnage furent développées afin de s'attaquer à cette limitation. La méthode d'échange de répliques Hamiltonien [**205**, **206**] est une de ces méthodes.

Cette méthode est basée sur le concept de répliques, soit différentes copies du système qui sont simulées en parallèle. Avec l'échange de répliques en température traditionnel, chacune des répliques est simulée à une température différente. Une réplique "froide" est simulée à la température désirée afin d'obtenir les caractéristiques réelles du système. Une réplique "chaude" est simulée à une température plus élevée pour accélérer l'échantillonnage (la probabilité d'observer un état d'énergie $U(r)$ augmente avec la température selon les facteurs de Boltzmann). Finalement, une série de répliques "tièdes", simulées à des températures intermédiaires, permettent de faire la transition entre la réplique "froide" et "chaude". Un des problèmes avec l'échange de répliques en température est que ce paramètre est intensif, c'est-à-dire qu'il caractérise le système dans son entièreté. Ainsi, augmenter la température du système affecte non seulement la(les) molécule(s) d'intérêt, mais aussi les molécules d'eau, composant, dans de nombreuses situations, la majorité du système simulé. Ainsi le

nombre de répliques nécessaires pour faire la transition entre la réplique "froide" et la réplique "chaude" croit rapidement avec la taille du système [**205**, **206**].

L'idée derrière l'échange de répliques Hamiltonien est similaire à l'échange de répliques en température où une réplique "froide", une réplique "chaude" et une série de répliques "tièdes" sont simulées en parallèle. Par contre, au lieu de températures différentes, les répliques sont simulées avec un potentiel différent. Les différences de potentiel permettent de diminuer l'amplitude des barrières énergétiques et d'accélérer l'échantillonnage. Comme le potentiel est une caractéristique extensive, on peut donc cibler les molécules d'intérêt dont le potentiel sera affecté et le nombre de répliques nécessaire s'en retrouve diminué. Dans le protocole REST2 [**206**], le potentiel protéine/protéine (E_{pp}), le potentiel protéine/solvant, (E_{pw}) et le potentiel solvant/solvant (E_{ww}) de chacune des répliques est échelonné à l'aide du facteur $\frac{\beta_m}{\beta_0}$ selon l'équation suivante:

$$U_m^{REST2}(X) = \frac{\beta_m}{\beta_0} U_{pp}(X) + \sqrt{\frac{\beta_m}{\beta_0}} U_{pw}(X) + U_{ww}(X)$$

Plus spécifiquement, pour obtenir un tel échelonnage, les charges, les profondeurs de puits (ϵ) du potentiel de Lennard-Jones et les amplitudes du potentiel des angles dièdres sont échelonnés d'un facteur $\sqrt{\frac{\beta_m}{\beta_0}}$, $\frac{\beta_m}{\beta_0}$ et $\sqrt{\frac{\beta_m}{\beta_0}}$ respectivement [**205**].

La transition entre la réplique "froide" et "chaude" se fait via l'échange des coordonnées et des vitesses entre chacune des répliques à un interval de temps déterminé via une procédure de Monte-Carlo. Afin de préserver à chacun des potentiels (ou températures) les bonnes statistiques, ces échanges sont acceptés ou refusés avec une probabilité α donné via l'équation de "detailed balance" suivante:

$$\alpha = \min\left(1, \exp\left(\frac{-V_i(r_j) + V_i(r_i)}{k_B T_i}\right) + \left(\frac{-V_j(r_i) + V_j(r_j)}{k_B T_j}\right)\right)$$

où $V_i(r_j)$ est l'énergie du système de la réplique i calculée avec les coordonnées de la réplique j et T_i est la température de la réplique i. Comme on peut le constater, les échanges dépendent du recoupement en énergie entre les différentes répliques. Comme la valeur des fluctuations d'énergie autour de l'équilibre dépend de $N^{-\frac{1}{2}}$ où N est le nombre d'atomes, on peut donc voir comment la méthode de l'échange de répliques Hamiltonien requiert un nombre moins important de répliques intermédiaires que son homologue en température.

Chapitre 8

Développements préliminaires de aaOPEP

La famille de potentiels gros-grain OPEP fut utilisée avec succès dans de nombreuses études numériques du processus d'agrégation d'Aβ [**159**, **160**, **164**]. Il y a malgré tout place à amélioration, car même les potentiels tout-atome de pointe présentent certaines lacunes dans l'étude des PIDs [**202**]. La première piste d'amélioration concerne la paramétrisation des potentiels OPEP. En effet, ils sont caractérisés par une série de poids qui sont optimisés afin de discriminer d'un ensemble de leurres la structure native de protéines possédant des structures tridimensionnelles bien définies [**165**]. La seconde piste d'amélioration est plus fondamentale et concerne la description gros-grain d'OPEP qui peut mener à certains problèmes d'empaquetage au niveau des boucles [**207**] et des protéines globulaires [**208**]. Ces deux limitations sont à l'origine du développement du potentiel aaOPEP, qui portera les idées d'OPEP en régime tout-atome et permettra d'étudier les processus d'agrégation d'Aβ et le potentiel thérapeutique de petites molécules.

Bien que le projet ne soit pas complété, différents éléments et outils d'analyse furent mis en place pour assurer sa bonne initiation. Ils seront présentés dans ce chapitre. Dans un premier temps, je discuterai de la conception des divers éléments de aaOPEP (types d'atome, termes du potentiel etc.). Dans un second temps, je présenterai la stratégie d'optimisation du potentiel afin de reproduire les données obtenues au niveau de simulations tout-atome pour Aβ. Finalement, je présenterai quelques outils qui furent implémentés dans le simulateur OPEP et qui permettront de simuler de façon exhaustive avec aaOPEP afin d'en confirmer la validité. Finalement, je présenterai la feuille de route pour la suite du projet.

8.1. Conception du potentiel

La description atomique de aaOPEP est tout-atome et basée sur les potentiels AMBER [**29**]. Les types de chacun des atomes d'aaOPEP sont tirés d'AMBER [**29**], mais, pour conserver l'idée d'OPEP d'interactions spécifiques entre chaînes latérales, les types d'atomes d'AMBER sont augmentés d'un identifiant spécifique à la chaîne latérale. Ainsi, des 19

(neuf carbones, cinq azotes, trois oxygènes et deux soufres) types d'atomes d'AMBER que l'on retrouve au niveau des chaînes latérales, on obtient 52 types d'atomes pour aaOPEP.

Dans aaOPEP, les interactions attractives/répulsives, les interactions électrostatiques et les effets de l'environnement (solvant, ions, etc.) sont réunis en une seule interaction effective. Cette interaction effective est décrite par un potentiel de MIE prenant la forme suivante:

$$V(r) = \epsilon \left[\left(\frac{m}{n-m} \right) \left(\frac{r_0}{r} \right)^n - \left(\frac{n}{n-m} \right) \left(\frac{r_0}{r} \right)^m \right]$$

où ϵ, r_0, n, m sont respectivement la profondeur du puit, la position du minimum et les exposants. Il s'agit de la formulation générale du potentiel de Lennard-Jones classique caractérisé par $n = 12$ et $m = 6$.

Pour chaque type d'atome i, on définit quatre paramètres associés à notre interaction effective: r_i^0, ϵ_i, n_i et m_i. Les interactions entre paires sont décrites par les règles de combinaison suivantes:

$$\begin{aligned} \epsilon_{ij} &= \sqrt{\epsilon_i \cdot \epsilon_j} \\ r_{ij}^0 &= \frac{1}{2} \left(r_i^0 + r_j^0 \right) \\ n_{ij} &= \sqrt{n_i \cdot n_j} \\ m_{ij} &= \sqrt{m_i \cdot m_j} \end{aligned}$$

Comme pour les potentiels OPEP, les ponts-hydrogène de la chaîne principale sont décrits explicitement et la formation de structure secondaire d'hélice-α ou de feuillet-β est stabilisée via une interaction de coopérativité [**165**]. Dans un premier temps, les paramètres des interactions liées (liens atomiques, angles de valence et angles dièdres) sont tirés directement d'AMBERff99SB*-ILDN [**29**].

8.2. Paramétrisation initiale

La paramétrisation de aaOPEP se fera avec la philosophie "bottom-up" [**157**], c'est-à-dire qu'elle se fera en reproduisant les résultats de simulations obtenues à l'aide d'un potentiel tout-atome. Le processus initial de paramétrisation est présenté dans ce qui suit.

Pour obtenir une première approximation adéquate des paramètres du potentiel, nous avons premièrement simulé les 210 différentes paires d'acides aminés en solution à l'aide de simulation de dynamique moléculaire de 300 ns. De ces simulations, nous avons extrait les fonctions de distribution radiale (voir la section 5.1.2) associées à chacun des types d'atomes de aaOPEP. Le champ de force AMBER99sb*-ILDN a été utilisé pour les protéines [**29**, **196**] et TIP3P pour les molécules d'eau. [**40**]. Lorsque nécessaire, le système est neutralisé par l'ajout d'ions (Na^+ ou Cl^-). La température est maintenue à 300K via l'utilisation du thermostat Nosé-Hoover [**44**, **45**] avec une constante de couplage de 0.1ps. La pression est

maintenue à 1 atm via le barostat de Parrinello-Rahman [**47**] et une constante de couplage de 2 ps. Un cutoff de 1nm est utilisé autant pour les interactions de van der Waals que pour les interactions électrostatiques. L'algorithme LINCS [**48**] est utilisé pour contraindre les liens avec hydrogène à leur longueur d'équilibre tandis que l'algorithme SETTLE [**49**] est utilisé pour préserver la géométrie des molécules d'eau.

Les potentiels de champ moyen ont été générés via l'équation 5.1.1 à partir de ces fonctions de distribution radiale. Les paramètres du potentiel de MIE de chacun des types d'atomes (ϵ, r^0, n, m) furent ensuite déterminés à l'aide d'une optimisation par essaim particulaire [**209**] ("Particle SWARM Optimization") afin de minimiser les carrés des distances entre la fonction optimisée et les potentiels de champ moyen obtenus des simulations. À titre d'exemple, les résultats pour les paires d'atomes CT(LYS)-CT(MET), CT(GLU)-CT(ILE) et CA(PHE)-N(GLN) sont présentés respectivement à la ligne (A), (B) et (C) de la Figure 8.1. Comme les différents termes du potentiel sont interdépendants, cette paramétrisation initiale sera peaufinée lors d'optimisations plus poussées directement sur notre système d'intérêt. Le protocole sera décrit dans les sections suivantes.

8.3. Simulation d'Amyloïde-β

Comme nous voulons appliquer aaOPEP plus spécifiquement au phénomène d'agrégation d'Aβ, une PID pour laquelle même les potentiels tout-atome de pointe peinent à étudier [**204**], la paramétrisation des paramètres doit être effectuée directement au niveau du système d'intérêt. Afin de pouvoir calculer les quantités d'intérêt, nous avons réalisé des simulations d'échange de répliques Hamiltonien pour le monomère et le dimère d'Aβ. Additionnellement, comme nous voulons aussi étudier le potentiel thérapeutique de petites molécules de la famille des polyphénols [**70**], des simulations d'échange de répliques Hamiltonien pour le monomère et le dimère d'Aβ en présence de la Corilagine et du TGG ont été réalisées. Cette partie du projet étant toujours en phase embryonnaire, les résultats ne seront pas présentés dans cette section.

8.3.1. Protocole des simulations

Configuration initiale. Nous sommes repartis des simulations [**159**, **160**] gros-grain réalisées avec OPEP3.2 [**165**] pour le monomère [**159**] et le dimère [**160**] d'Aβ_{40}. La structure initiale pour nos simulations correspond au centre du plus gros cluster obtenu à l'aide de l'algorithme décrit dans l'article de Daura *et coll.* [**94**]. Cette structure fut ensuite convertie en tout-atome à l'aide du programme SCWRL4 [**210**].

Protocole de simulation. Dans ce travail toutes les simulations furent réalisées avec GROMACS [**58**]. La protéine est décrite à l'aide du champ de force AMBERff99SB*-ILDNP [**29**, **211**] et l'eau est décrite à l'aide du modèle TIP3P [**40**]. Pour le monomère

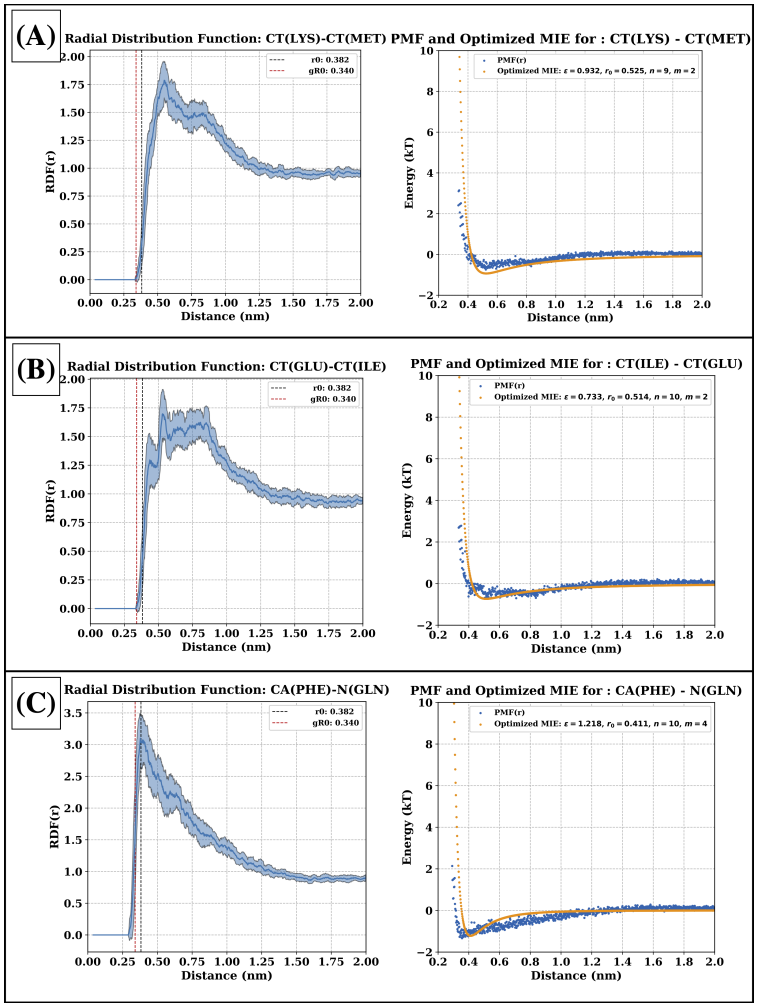


Fig. 8.1. Exemples de potentiel de MIE de aaOPEP. Les fonctions de distributions radiales obtenues via 300 ns de dynamique moléculaire sont présentées à gauche (en bleu). Les lignes verticales noire et rouge correspondent respectivement à la valeur du zéro et à la valeur du minimum du potentiel pour cette paire d'atomes dans AMBERff99SB*-ILDN. Le PMF, obtenu de la courbe de gauche, et le potentiel de MIE optimisé sont présentés à droite respectivement en bleu et en orange.

et le dimère, la préparation des systèmes a ensuite été réalisée de façon identique en suivant un protocole en cinq étapes: (1) Le système est placé dans une boîte dodécaédrique ($\alpha = 60°$, $\beta = 90°$, $\gamma = 60°$) en laissant 1.25 nm entre la protéine et le côté de la boîte. (2) L'énergie du système dans le vide est minimisée en appliquant successivement l'algorithme du col et l'algorithme du gradient conjugué. (3) Le solvant est ajouté, 3569 et 5543 molécules d'eau pour le monomère et le dimère respectivement, ainsi que les ions (Na^+), afin d'atteindre la neutralité. (4) L'énergie du système est ensuite minimisée de nouveau en appliquant successivement l'algorithme du col et l'algorithme du gradient conjugué. Des contraintes sont appliquées au niveau des positions de la protéine. (5) Le système est équilibré à l'aide d'une nanoseconde de simulation en NVT puis NPT, toujours en gardant les contraintes au niveau

des positions de la protéine. La configuration résultante est présentée au panneau (A) de la Figure 8.2 pour le monomère et de la Figure 8.3 pour le dimère. La température est gardée constante à 298 K l'aide du thermostat de Bussi-Donadio-Parrinello [**212**]. Les liens sont contraints avec LINCS [**48**] et SETTLE [**49**] pour la protéine et le solvant respectivement, permettant un pas d'intégration de 2 fs. Une distance seuil de 1 nm est utilisée pour les interactions attractives/répulsives et pour les interactions électrostatiques. Le calcul des interactions électrostatiques longue portée est réalisé à l'aide de la méthode de "Particle Mesh Ewald" (PME) [**92**, **93**].

Les simulations d'échange de répliques Hamiltonien sont réalisées à l'aide du protocole REST2 [**206**] grâce au programme GROMACS [**58**] en combinaison avec PLUMED [**205**, **213**]. Pour le monomère et le dimère, respectivement 24 répliques et 40 répliques furent générées à l'aide d'un facteur d'échelonnage allant de 1.0 à 0.3 suivant une distribution géométrique. Les échanges entre chacune des répliques sont tentés à chaque 4 ps (2000 pas de temps). Tant pour le monomère que le dimère, ce protocole résulte en des taux d'échange entre 20% et 40%. Les index des répliques au facteur d'échelonnage 1.0 (réplique 0) sont présentées au panneau (B) de la Figure 8.2 et de la Figure 8.3 pour le monomère et le dimère respectivement.

Convergence et analyse. L'estimation de la convergence se fait exclusivement au niveau de la réplique non-échelonnée (réplique 0), étant donné qu'il s'agit de la réplique associée à la version physique du potentiel. Pour déterminer la convergence, la stabilité en fonction du temps de deux paramètres, le rayon de giration et la surface accessible au solvant, est caractérisée. Dans un premier temps, cette caractérisation s'est faite visuellement, en traçant les paramètres en fonction du temps. Dans un second temps, et pour confirmer le choix d'intervalle de convergence déterminée par analyse visuelle, les distributions des rayons de giration et de la surface accessible au solvant sur l'intervalle sont comparées quantitativement en calculant l'intersection entre les distributions. Pour deux histogrammes, H_1 et H_2 avec un nombre identique de "bins", N, l'intersection (HI) est donnée par l'équation suivante:

$$HI = \sum_{i=1} N \min\left(H_1(i), H_2(i)\right)$$

Tout comme l'estimation de la convergence, l'analyse de clusterisation est réalisée uniquement au niveau de la réplique non-échelonnée. La clusterisation est réalisée sur l'intervalle de convergence à l'aide de l'algorithme décrit dans le papier de Daura *et coll.* [**94**] et un cutoff de 0.2 nm.

8.3.2. Monomère

Les inspections visuelles en fonction du temps des deux paramètres choisis, le rayon de giration et la surface accessible au solvant, sont présentées respectivement au panneau (C)

de la Figure 8.2 et au panneau (E) de la Figure 8.2. Le panneau (C) montre que le rayon de giration est assez stable au niveau des 200 dernières nanosecondes de notre simulation. La même chose est observée au panneau (E) pour la surface accessible au solvant. Ces deux mesures indiquent que la convergence est atteinte à partir de 600 ns, laissant 200 ns pour l'analyse (600-800 ns). Pour confirmer l'analyse visuelle de façon plus quantitative, l'intervalle de convergence est séparé en deux sous-parties indépendantes de 100 ns chacune sur lesquelles les distributions des rayons de giration et de la surface accessible au solvant sont comparées. Le panneau (D) de la Figure 8.2, présente les résultats pour le rayon de giration. L'intersection entre les distributions est élevée avec une valeur de 0.728. La même chose est observée au niveau de la surface accessible au solvant au panneau (F) de la Figure 8.2. L'intersection des distributions est élevée à 0.832. Ainsi, pour nos deux mesures, tant l'analyse visuelle que l'analyse quantitative montrent que la convergence est atteinte à partir de 600 ns pour le monomère.

8.3.3. Dimère

Une analyse similaire à celle réalisée pour le monomère est présentée à la Figure 8.3 pour le dimère. Les panneaux (C) et (D) montrent les mesures du rayon de giration, tandis que les panneaux (E) et (F) présentent les mesures au niveau de la surface accessible au solvant. Visuellement, on remarque, pour les deux chaînes seules ou combinées, que le rayon de giration et la surface accessible au solvant sont stables au niveau des 200 dernières nanosecondes. L'analyse visuelle est confirmée par une mesure quantitative de la similarité des distributions (panneaux (D) et (F)). Les intersections des distributions pour les deux dernières tranches de 100 ns sont de 0.758 pour le rayon de giration et 0.841 pour la surface accessible au solvant. Ces deux mesures combinées indiquent que la convergence est atteinte à partir de 350 ns pour le dimère.

8.3.4. Clusterisation

Pour aaOPEP, nous avons choisi comme première cible pour l'optimisation de reproduire les résultats obtenus via les simulations tout-atome décrites précédemment. Ainsi, aaOPEP a comme objectif de discriminer entre les "bonnes" et les "mauvaises" strutures d'Aβ. Par "bonnes" et "mauvaises", on entend les structures observées souvent et moins souvent au niveau des simulations tout-atome. Pour identifier ces structures, une clusterisation des résultats est réalisée. 674 clusters furent identifiés pour le monomère et 357 pour le dimère. Les populations de chacun des clusters sont présentées à la Figure 8.4.

Pour la suite du projet, nous utiliserons l'idée derrière le protocole d'optimisation de OPEP [**165**] et présenté exhaustivement à la section 5.2.2. Les conformations obtenues de notre analyse de cluster ont été classifiées selon la taille des clusters. Les centres des clusters

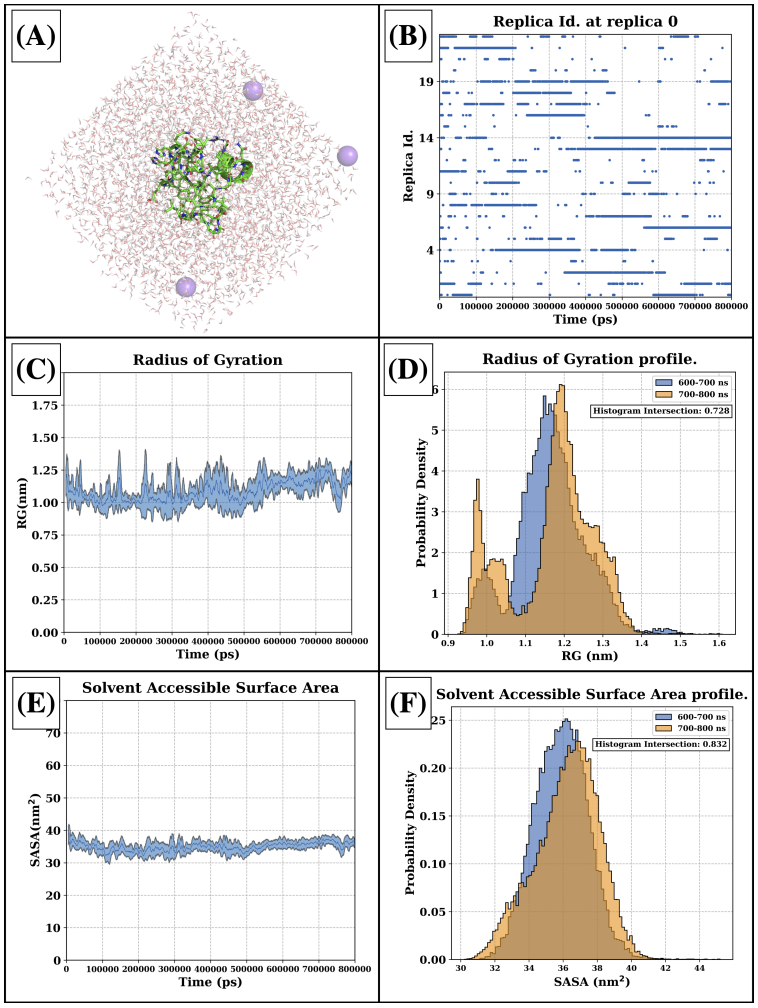


Fig. 8.2. Convergence du monomère d'Aβ_{40}. Le panneau (A) présente le configuration initiale. Le panneau (B) présente l'index de la réplique présente à la réplique sans échelonnage de l'énergie potentielle (réplique 0) en fonction du temps. Les panneaux (C) et (D) présentent respectivement le rayon de giration en fonction du temps et la distribution des rayons de giration dans les deux dernières tranches de 100 ns de la simulation (600-700 ns en bleu et 700-800 ns en orange). Finalement, les panneaux (E) et (F) présentent respectivement la surface accessible au solvant en fonction du temps et la distribution des surfaces accessibles au solvant pour les deux dernières tranches de 100 ns de la simulation (600-700 ns en bleu et 700-800 ns en orange).

représentant plus de 5% , entre 1% et 5% et moins de 1% des structures totales correspondent respectivement aux leurres natifs, comme-natifs et non-natifs. De la simulation du monomère, nous obtenons trois leurres natifs (24% de la population totale), 16 leurres comme-natifs (33% de la population totale) et 655 leurres non-natifs. De la simulation du dimère, nous obtenons cinq leurres natifs (54% de la population totale), 11 leurres comme-natifs (28% de la population totale) et 341 leurres non-natifs. Les structures des leurres natifs du monomère et du dimère sont présentées à la Figure 8.4.

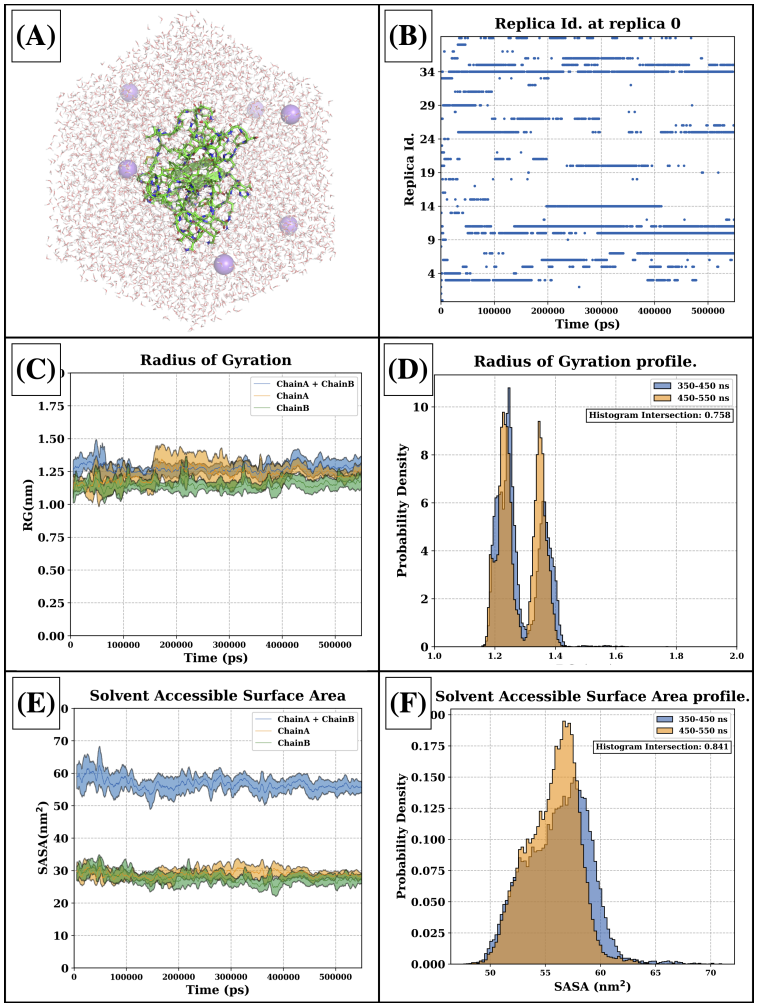


Fig. 8.3. Convergence du dimère d'Aβ_{40}. Le panneau (A) présente la configuration initiale. Le panneau (B) présente l'index de la réplique sans échelonnage de l'énergie potentielle (réplique 0) en fonction du temps. Le panneau (C) présente le rayon de giration en fonction du temps; la chaîne A, la chaîne B et les deux chaînes sont présentées respectivement en orange, vert et bleu, Le panneau (D) présente la distribution des rayons de giration des deux chaînes dans les deux dernières tranches de 100 ns de la simulation (350-450 ns en bleu et 450-550 ns en orange). Le panneau (E) présente respectivement la surface accessible au solvant en fonction du temps; la chaîne A, la chaîne B et les deux chaînes sont présentées respectivement en orange, vert et bleu. Finalement, le panneau (F) présente la distribution des surfaces accessibles au solvant des deux chaînes pour les deux dernières tranches de 100 ns de la simulation (350-450 ns en bleu et 450-550 ns en orange).

8.4. Outil: simulateur OPEP

Finalement, le dernier élément manquant est l'outil qui permettra de simuler les systèmes à l'aide d'aaOPEP. Cet outil est le simulateur OPEP. Il s'agit d'un code maison dont le développement a été initialisé par Sébastien Côté. Le simulateur OPEP est composé d'une

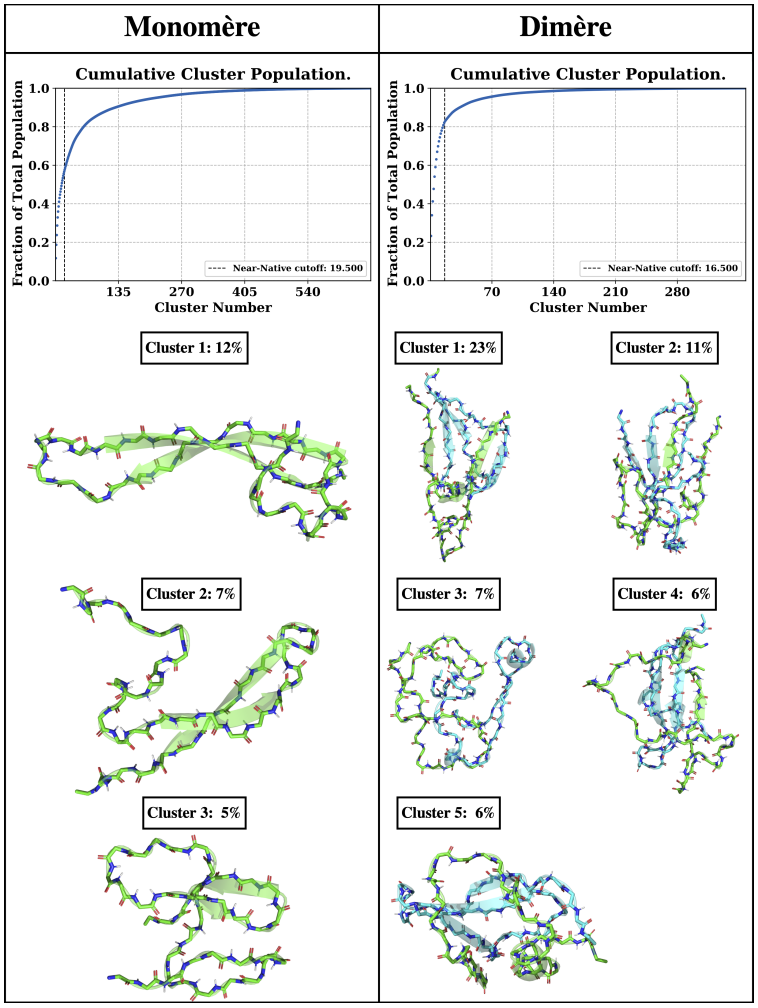


Fig. 8.4. Clusterisation du monomère et dimère d'Aβ. Les résultats au niveau du monomère et du dimère sont présentés respectivement à gauche et à droite. La population cumulative en fonction des clusters est présentée dans le haut des colonnes. En dessous, on retrouve les centres des clusters que nous avons classifiés comme natif. La clusterisation a été réalisée à l'aide de l'algorithme de l'article de Daura *et coll.* [**94**].

série d'utilitaires, permettant de générer les fichiers de coordonnées et les fichiers de topologie adaptés pour les champs de force OPEP, et d'un code principal permettant de réaliser des simulations de dynamique moléculaire avec OPEP. Une présentation plus détaillée du simulateur OPEP est présentée dans la thèse de Sébastien Côté [**171**].

Pour s'attaquer à l'étude du processus d'agrégation d'Aβ et à la paramétrisation de aaOPEP, deux ajouts d'importance furent implémentés dans le simulateur OPEP. Le premier est une parallélisation multi-coeurs du calcul des forces avec openMP. À l'intérieur du simulateur OPEP, la boîte de simulation est sous-divisée en une série de boîtes plus petites. En choisissant adéquatement la taille de ces sous-boîtes en fonction des valeurs seuils pour le

calcul des interactions attractives/répulsives, on peut alors simplifier le calcul des interactions longues portées. En effet, les voisins des atomes de chacune des boîtes se retrouvent nécessairement dans (i) la même boîte ou (ii) une des 26 sous-boîtes voisines. On peut donc calculer de façon indépendante les énergies et les forces des sous-boîtes et communiquer l'information uniquement à la fin du calcul. Cette décomposition doit par contre être améliorée. Comme OPEP ne modélise pas explicitement le solvant, la décomposition actuelle mène à des débalancements importants entre les calculs effectués par chacun des coeurs. Malgré ce problème, la nouvelle parallélisation permet de faciliter l'étude de systèmes plus grands sur de plus longues échelles de temps.

Le second est l'implémentation de l'algorithme d'échange de répliques en température [**214**]. L'algorithme est similaire à ce qui est décrit à la section 7.2, mais au lieu de simuler les différentes versions du système à des potentiels différents, des versions du système des températures différentes sont simulées. L'implémentation de cet algorithme a été réalisée à l'aide de MPI: chacune des versions du système est simulée de manière indépendante et, à certains moments prédéterminés, les positions et les vitesses sont échangées à l'aide de MPI. Comme le solvant est implicite dans OPEP, le nombre de répliques nécessaires à la simulation reste limité et cet algorithme permettra de tester efficacement la paramétrisation de aaOPEP au niveau d'Aβ.

8.5. Perspective

Tous les éléments sont maintenant en place afin de réaliser l'optimisation de aaOPEP:

- La conception du potentiel, incluant la définition des types d'atomes et des formes des fonctions énergétiques, est complétée.
- Les modifications nécessaires du générateur de topologie et du code de dynamique moléculaire du simulateur OPEP ont été apportées afin de rendre l'utilisation de aaOPEP possible. De plus, deux implémentations cruciales pour l'utilisation du simulateur sont complétées; une parallélisation au niveau du calcul des forces et une méthode pour effectuer des simulations d'échange de répliques en température.
- Des simulations pour le monomère et le dimère d'Aβ ont été obtenues à l'aide d'un champ de force tout-atome de pointe. Celles-ci serviront de référence dans le développement de la méthode simplifiée aaOPEP.
- Les outils pour l'optimisation du potentiel sont complétés et prêts à être utilisés.

Dans un premier temps, l'optimisation se limitera aux interactions non-liées, incluant les ponts hydrogène et la coopérativité caractéristique des potentiels OPEP. Dans un second temps, les paramètres liés, notamment des cruciaux angles de torsion de la chaîne principale, ϕ et ψ, et des chaînes latérales, χ, devront être optimisés. Pour ce faire, nous adapterons le protocole développé pour l'optimisation de sOPEP et décrit en détail au chapitre 6.1. Au

lieu d'utiliser PEP-FOLD pour générer les leurres, ceux-ci seront extraits de simulations de dynamique moléculaire réalisées avec le simulateur OPEP.

Finalement, en plus de l'étude du processus d'agrégation d'Aβ, nous désirons aussi étudier le potentiel thérapeutique de petites molécules de la famille des polyphénols, notamment la Corilagine et TGG. Ainsi, la dernière étape est donc d'adapter aaOPEP pour le traitement de ces petites molécules. Bien que cela ne soit pas présenté dans ce document, les simulations de référence du monomère et du dimère d'Aβ avec Corilagine et TGG ainsi que certaines analyses préliminaires sont déjà complétées. Ultimement, aaOPEP offrira une méthode hybride combinant les caractéristiques clés ayant fait le succès des potentiels gros-grain OPEP, mais en régime tout-atome.

Conclusion

Les protéines sont des nanomachines jouant une panoplie de rôles fondamentaux pour la survie de tout organisme. De plus, elles sont aussi impliquées dans le développement de maladies, comme la protéine *Spike* du virus SARS-CoV-2 associé à la COVID-19 [**2**, **3**], ou la protéine Aβ associée à l'apparition de la maladie d'Alzheimer [**4**, **200**]. D'un autre côté, les protéines peuvent aussi jouer des rôles d'agent thérapeutique qui pourraient être cruciaux pour freiner la résistance aux antibiotiques [**127**–**129**]. La structure des protéines est étroitement reliée à leurs fonctions [**1**]. Elles adoptent leurs structures lors d'un processus de repliement complexe déterminé par les lois de la physique. L'étude des caractéristiques structurelles, mais aussi dynamiques des protéines, est essentielle afin de comprendre les processus moléculaires du repliement des protéines, des interactions avec leurs partenaires, etc.

La bio-modélisation numérique est un domaine multi-disciplinaire qui, grâce à l'amélioration constante des ordinateurs et des algorithmes, est maintenant un partenaire essentiel aux méthodes expérimentales pour l'étude des protéines. Un des avantages de la bio-modélisation numérique est sa capacité à étudier les phénomènes au niveau atomique et en temps réel. De plus, elle peut prendre avantage des géantes bases de données développées grâce à des années de travail d'expérimentateurs en cristallographie aux rayons-X, en résonnance magnétique nucléaire, en cryo-microscopie électronique, etc. En effet, la *Protein Data Bank* contient plus de 150 000 structures disponibles en ligne et gratuitement [**8**]. Qu'elles soient "knowledge-based" ou basées sur les lois de la physique, simples et rapides ou précises et lentes, les méthodes développées par la bio-modélisation numérique permettent d'étudier les multiples facettes des processus moléculaires des protéines.

8.6. Développement thérapeutique contre la COVID-19

Les propriétés macroscopiques d'un système sont données par la mécanique statistique et correspondent à des moyennes sur un ensemble d'états générés avec la bonne statistique, donnée par les poids de Boltzmann [**28**, **43**]. En bio-modélisation numérique, une des méthodes les plus populaires pour générer cet ensemble d'états est la dynamique moléculaire.

La première partie de cette thèse visait l'étude du potentiel thérapeutique de petites molécules contre les variants du virus SARS-CoV-2 à l'origine de la COVID-19. Pour ce faire, nous avons utilisé un protocole numérique unissant divers outils numériques puissants: amarrages moléculaires, simulations de dynamique moléculaire à l'aide de potentiels tout-atome de pointe et calculs d'énergie libre de MM/PBSA.

De plus, nos méthodes numériques furent complémentées de méthodes expérimentales d'essais immuno-enzymatiques et des expériences de résonance plasmon de surface. Cette étude montre clairement la synergie des méthodes numériques et expérimentales.

Au niveau du "wild-type" de SARS-CoV-2, nos simulations numériques ont montré que le TGG et, dans une moindre mesure, la Corilagine, avaient une capacité à se lier à l'interface avec ACE2 d'une manière qui pourrait prévenir l'association. Ces résultats ont été confirmés par nos mesures expérimentales, tant au niveau des essais immuno-enzymatiques que des expériences de résonance plasmon de surface. Dans les deux cas, TGG et Corilagine permettent de prévenir l'association RBD/ACE2, en se liant surtout au niveau du RBD. Cette dernière propriété est particulièrement intéressante puisque ces molécules préserveraient les fonctions essentielles de ACE2. À l'aide de notre protocole numérique, dont les conclusions sont confirmées au niveau du "wild-type" par deux méthodes expérimentales, nous avons par la suite étudié trois mutations clés du RBD: E484K, N501Y et E484K/N501. Dans les trois cas, le potentiel thérapeutique du TGG est conservé malgré ces mutations, quant à la Corilagine, nous avons observé qu'elle est efficace contre le mutant E484K, mais perdrait son potentiel thérapeutique pour les mutants N501Y et E484K/N501.

8.7. Prédictions structurelles par PEP-FOLD

Dans la première partie de cette thèse, nous avons utilisé des méthodes numériques traditionnelles; simulations de dynamique moléculaire à l'aide d'un potentiel tout-atome avec solvant implicite. Malgré tous les avantages de ces méthodes traditionnelles, les systèmes biologiques d'intérêt couvrent une vaste gamme d'échelles spatiales et temporelles et peuvent être difficiles à étudier avec celles-ci. Nous avons notamment observé ces limitations lors de notre étude sur les molécules thérapeutiques contre la COVID-19. Ainsi, pour étudier les systèmes plus grands sur des échelles de temps plus longues, il faut développer des méthodes simplifiées. Les potentiels gros-grain font partie de ces méthodes de simplifications.

Ainsi, un des éléments importants de la deuxième partie de cette thèse était le développement et l'amélioration de méthodes simplifiées, notamment d'un potentiel gros-grain, dans le cadre d'une méthode de prédiction structurelle. La prédiction structurelle des protéines est particulièrement importante puisque leur structure est associée à leur fonction. Selon l'hypothèse thermodynamique [**130**], les propriétés physico-chimiques des acides aminés sont à l'origine de la structure tridimensionnelle des protéines. Ainsi, il est possible de prédire la

structure des protéines, uniquement à partir de leur séquence en acides aminés; ce que l'on appelle prédiction *de novo*.

La seconde partie de cette thèse portait sur les améliorations de la méthode de prédiction *de novo* PEP-FOLD pour les peptides et les petites protéines. PEP-FOLD est une méthode simplifiée pour la prédiction structurelle dont un des éléments-clés est l'utilisation d'un potentiel gros-grain dérivé de la famille OPEP, sOPEP. Dans ce travail, deux des trois éléments clés de PEP-FOLD furent revisités; (1) La librairie de fragments de l'alphabet structurel et (2) le potentiel gros-grain sOPEP.

En combinant la version révisée de la librairie et du potentiel, nous observons des améliorations importantes de la qualité des prédictions avec PEP-FOLD, tant en termes de CAD-score qu'en termes de BC-score. Ces améliorations s'observent autant au niveau de la prédiction de plus basse énergie (TOP1), que des cinq prédictions de plus basses énergies (TOP5) que de la meilleure prédiction du TOP5. La version améliorée de PEP-FOLD fut comparée de façon exhaustive à diverses méthodes de pointe utilisant les avancées récentes de l'intelligence artificielle. Malgré son modèle discret et gros-grain, les résultats de PEP-FOLD sont comparables à ceux de Raptor-X [**152**] et d'APPTest [**155**]. En particulier, PEP-FOLD semble mieux prédire les petites cibles β grâce à une bonne reproduction des ponts hydrogène. Comme Raptor-X et APPTest, les prédictions de PEP-FOLD restent inférieures à celles d'AlphaFold2. Par contre, en utilisant un champ de force dont la formulation est basée sur des interactions physiques, notre méthode permet d'étudier plus facilement la flexibilité et la dynamique; sOPEP est effet dérivé des potentiels OPEP qui furent utilisés avec succès en dynamique moléculaire [**147**, **148**, **158**–**164**]. Finalement, notre étude a permis d'identifier un certain nombre de pistes d'amélioration pour une future version de PEP-FOLD.

8.8. Méthode simplifiée pour l'étude de l'Alzheimer

Finalement, la dernière partie de cette thèse continue l'exploration du développement de potentiels simplifiés commencé à la section précédente. Au lieu de la prédiction *de novo* de la structure des protéines, ce potentiel simplifié sera adapté au processus de l'agrégation et de la fibrillation d'Aβ, un phénomène associé à l'apparition de la maladie d'Alzheimer et particulièrement complexe à étudier à l'aide de méthodes numériques (et expérimentales) traditionnelles. En effet, les monomères Aβ, composés de quelques centaines d'atomes seulement, s'associent les uns avec les autres pour former des oligomères de tailles diverses et cette association mène ultimement à la formation de fibres amyloïdes composées de plusieurs milliers de monomères.

Pour étudier ce phénomène, la dernière section de cette thèse présente la mise en place et quelques résultats préliminaires dans le développement de aaOPEP, un nouveau potentiel

pour l'étude d'Aβ. aaOPEP s'inspire à la fois des potentiels tout-atome comme AMBER, notamment via la définition des types d'atomes, et des potentiels gros-grains de la famille OPEP, via la formulation du potentiel (l'inclusion explicite des ponts hydrogène, la stabilisation des structures secondaires via un terme de coopérativité entre ponts hydrogène, le traitement implicite du solvant, etc.) et le processus de paramétrisation (discrimination sur une banque de leurres). Bien que le projet soit en phase préliminaire, une grande partie du travail de mise en place est complétée. Premièrement, des simulations d'échange de répliques Hamiltonien ont été réalisées sur le monomère et le dimère d'Aβ afin de servir de référence pour le processus d'optimisation. Des simulations de ces systèmes en présence de petites molécules thérapeutiques ont aussi été réalisées. Deuxièmement, le code du simulateur OPEP fut amélioré afin d'y ajouter deux éléments essentiels pour la simulation avec aaOPEP: une parallélisation du calcul des forces et des énergies et une méthode d'échange de répliques. Finalement, nos simulations de référence ont servi à développer l'ensemble des leurres requis pour la paramétrisation.

Références bibliographiques

[1] Tamar Schlick. *Molecular modeling and simulation: an interdisciplinary guide*, volume 2. Springer, 2010.

[2] Marco Cascella, Michael Rajnik, Abdul Aleem, Scott Dulebohn, and Raffaela Di Napoli. Features, evaluation, and treatment of coronavirus (covid-19). *StatPearls*, 2021.

[3] Philip V'kovski, Annika Kratzel, Silvio Steiner, Hanspeter Stalder, and Volker Thiel. Coronavirus biology and replication: implications for sars-cov-2. *Nature Reviews Microbiology*, pages 1–16, 2020.

[4] Jessica Nasica-Labouze, Phuong H Nguyen, Fabio Sterpone, Olivia Berthoumieu, Nicolae-Viorel Buchete, Sebastien Cote, Alfonso De Simone, Andrew J Doig, Peter Faller, Angel Garcia, et al. Amyloid β protein and alzheimer's disease: When computer simulations complement experimental studies. *Chemical reviews*, 115(9):3518–3563, 2015.

[5] Tamar Schlick, Stephanie Portillo-Ledesma, Christopher G Myers, Lauren Beljak, Justin Chen, Sami Dakhel, Daniel Darling, Sayak Ghosh, Joseph Hall, Mikaeel Jan, et al. Biomolecular modeling and simulation: a prospering multidisciplinary field. *Annual Review of Biophysics*, 50:267–301, 2021.

[6] Tamar Schlick and Stephanie Portillo-Ledesma. Biomolecular modeling thrives in the age of technology. *Nature Computational Science*, 1(5):321–331, 2021.

[7] David E Shaw, JP Grossman, Joseph A Bank, Brannon Batson, J Adam Butts, Jack C Chao, Martin M Deneroff, Ron O Dror, Amos Even, Christopher H Fenton, et al. Anton 2: raising the bar for performance and programmability in a special-purpose molecular dynamics supercomputer. In *SC'14: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 41–53. IEEE, 2014.

[8] Emiliano Brini, Carlos Simmerling, and Ken Dill. Protein storytelling through physics. *Science*, 370(6520), 2020.

[9] Marco De Vivo, Matteo Masetti, Giovanni Bottegoni, and Andrea Cavalli. Role of molecular dynamics and related methods in drug discovery. *Journal of medicinal chemistry*, 59(9):4035–4061, 2016.

[10] Jiri Sponer, Giovanni Bussi, Miroslav Krepl, Pavel Banáš, Sandro Bottaro, Richard A Cunha, Alejandro Gil-Ley, Giovanni Pinamonti, Simón Poblete, Petr Jurečka, et al. Rna structural dynamics as captured by molecular simulations: a comprehensive overview. *Chemical reviews*, 118(8):4177–4338, 2018.

[11] wwPDB consortium. Protein data bank: the single global archive for 3d macromolecular structure data. *Nucleic acids research*, 47(D1):D520–D528, 2019.

[12] Andrew W Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander WR Nelson, Alex Bridgland, et al. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, 2020.

[13] World health organization weekly epidemiological update. https://www.who.int/publications/m/item/weekly-epidemiological-update-on-covid-19---21-september-2021. Accessed: 2021-09-22.

[14] Ben Hu, Hua Guo, Peng Zhou, and Zheng-Li Shi. Characteristics of sars-cov-2 and covid-19. *Nature Reviews Microbiology*, 19(3):141–154, 2021.

[15] Erin K Stokes, Laura D Zambrano, Kayla N Anderson, Ellyn P Marder, Kala M Raz, Suad El Burai Felix, Yunfeng Tie, and Kathleen E Fullerton. Coronavirus disease 2019 case surveillance—united states, january 22–may 30, 2020. *Morbidity and Mortality Weekly Report*, 69(24):759, 2020.

[16] Sven Ullrich and Christoph Nitsche. The sars-cov-2 main protease as drug target. *Bioorganic & Medicinal Chemistry Letters*, page 127377, 2020.

[17] Ananda da Silva Antonio, Larissa Silveira Moreira Wiedemann, and Valdir Florêncio Veiga-Junior. Natural products' role against covid-19. *RSC Advances*, 10(39):23379–23393, 2020.

[18] Peng Zhou, Xing-Lou Yang, Xian-Guang Wang, Ben Hu, Lei Zhang, Wei Zhang, Hao-Rui Si, Yan Zhu, Bei Li, Chao-Lin Huang, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *nature*, 579(7798):270–273, 2020.

[19] Wentao Ni, Xiuwen Yang, Deqing Yang, Jing Bao, Ran Li, Yongjiu Xiao, Chang Hou, Haibin Wang, Jie Liu, Donghong Yang, et al. Role of angiotensin-converting enzyme 2 (ace2) in covid-19. *Critical Care*, 24(1):1–10, 2020.

[20] Satarudra Prakash Singh, Manisha Pritam, Brijesh Pandey, and Thakur Prasad Yadav. Microstructure, pathophysiology, and potential therapeutics of covid-19: A comprehensive review. *Journal of Medical Virology*, 93(1):275–299, 2021.

[21] Jun Lan, Jiwan Ge, Jinfang Yu, Sisi Shan, Huan Zhou, Shilong Fan, Qi Zhang, Xuanling Shi, Qisheng Wang, Linqi Zhang, et al. Structure of the sars-cov-2 spike receptor-binding domain bound to the ace2 receptor. *Nature*, 581(7807):215–220, 2020.

[22] Markus Hoffmann, Hannah Kleine-Weber, Simon Schroeder, Nadine Krüger, Tanja Herrler, Sandra Erichsen, Tobias S Schiergens, Georg Herrler, Nai-Huei Wu, Andreas Nitsche, et al. Sars-cov-2 cell entry depends on ace2 and tmprss2 and is blocked by a clinically proven protease inhibitor. *cell*, 181(2):271–280, 2020.

[23] Haibo Zhang, Josef M Penninger, Yimin Li, Nanshan Zhong, and Arthur S Slutsky. Angiotensin-converting enzyme 2 (ace2) as a sars-cov-2 receptor: molecular mechanisms and potential therapeutic target. *Intensive care medicine*, 46(4):586–590, 2020.

[24] World health organization coronavirus (covid-19) dashboard. https://covid19.who.int/. Accessed: 2021-09-21.

[25] Markus Hoffmann, Prerna Arora, Rüdiger Groß, Alina Seidel, Bojan F. Hörnich, Alexander S. Hahn, Nadine Krüger, Luise Graichen, Heike Hofmann-Winkler, Amy Kempf, Martin S. Winkler, Sebastian Schulz, Hans-Martin Jäck, Bernd Jahrsdörfer, Hubert Schrezenmeier, Martin Müller, Alexander Kleger, Jan Münch, and Stefan Pöhlmann. Sars-cov-2 variants b.1.351 and p.1 escape from neutralizing antibodies. *Cell*, 184(9):2384–2393.e12, 2021.

[26] Markus Hoffmann, Heike Hofmann-Winkler, Nadine Krüger, Amy Kempf, Inga Nehlmeier, Luise Graichen, Prerna Arora, Anzhalika Sidarovich, Anna-Sophie Moldenhauer, Martin S Winkler, et al. Sars-cov-2 variant b. 1.617 is resistant to bamlanivimab and evades antibodies induced by infection and vaccination. *Cell Reports*, 36(3):109415, 2021.

[27] World health organization tracking sars-cov-2 variants. https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/. Accessed: 2021-09-22.

[28] Daniel M Zuckerman. *Statistical physics of biomolecules: an introduction*. CRC Press, 2010.

[29] Kresten Lindorff-Larsen, Stefano Piana, Kim Palmo, Paul Maragakis, John L Klepeis, Ron O Dror, and David E Shaw. Improved side-chain torsion potentials for the amber ff99sb protein force field. *Proteins: Structure, Function, and Bioinformatics*, 78(8):1950–1958, 2010.

[30] James A Maier, Carmenza Martinez, Koushik Kasavajhala, Lauren Wickstrom, Kevin E Hauser, and Carlos Simmerling. ff14sb: improving the accuracy of protein side chain and backbone parameters from ff99sb. *Journal of chemical theory and computation*, 11(8):3696–3713, 2015.

[31] Michael J Robertson, Julian Tirado-Rives, and William L Jorgensen. Improved peptide and protein torsional energetics with the opls-aa force field. *Journal of chemical theory and computation*, 11(7):3499–3509, 2015.

[32] Robert B Best, Xiao Zhu, Jihyun Shim, Pedro EM Lopes, Jeetain Mittal, Michael Feig, and Alexander D MacKerell Jr. Optimization of the additive charmm all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain $\chi 1$ and $\chi 2$ dihedral angles. *Journal of chemical theory and computation*, 8(9):3257–3273, 2012.

[33] Jing Huang, Sarah Rauscher, Grzegorz Nawrocki, Ting Ran, Michael Feig, Bert L De Groot, Helmut Grubmüller, and Alexander D MacKerell. Charmm36m: an improved force field for folded and intrinsically disordered proteins. *Nature methods*, 14(1):71–73, 2017.

[34] Junmei Wang, Romain M Wolf, James W Caldwell, Peter A Kollman, and David A Case. Development and testing of a general amber force field. *Journal of computational chemistry*, 25(9):1157–1174, 2004.

[35] Viktor Hornak, Robert Abel, Asim Okur, Bentley Strockbine, Adrian Roitberg, and Carlos Simmerling. Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins: Structure, Function, and Bioinformatics*, 65(3):712–725, 2006.

[36] Junmei Wang, Piotr Cieplak, and Peter A Kollman. How well does a restrained electrostatic potential (resp) model perform in calculating conformational energies of organic and biological molecules? *Journal of computational chemistry*, 21(12):1049–1074, 2000.

[37] Christopher I Bayly, Piotr Cieplak, Wendy Cornell, and Peter A Kollman. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the resp model. *The Journal of Physical Chemistry*, 97(40):10269–10280, 1993.

[38] Wendy D Cornell, Piotr Cieplak, Christopher I Bayly, and Peter A Kollman. Application of resp charges to calculate conformational energies, hydrogen bond energies, and free energies of solvation. *Journal of the American Chemical Society*, 115(21):9620–9631, 2002.

[39] Junmei Wang and Peter A Kollman. Automatic parameterization of force field by systematic search and genetic algorithms. *Journal of Computational Chemistry*, 22(12):1219–1228, 2001.

[40] William L Jorgensen, Jayaraman Chandrasekhar, Jeffry D Madura, Roger W Impey, and Michael L Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of chemical physics*, 79(2):926–935, 1983.

[41] William L Jorgensen and Julian Tirado-Rives. Potential energy functions for atomic-level simulations of water and organic and biomolecular systems. *Proceedings of the National Academy of Sciences*, 102(19):6665–6670, 2005.

[42] Michael W Mahoney and William L Jorgensen. A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. *The Journal of chemical physics*, 112(20):8910–8922, 2000.

[43] MJ Abraham, D Van Der Spoel, E Lindahl, and B Hess. The gromacs development team, gromacs user manual version 2021, 2021.

[44] Shuichi Nosé. A unified formulation of the constant temperature molecular dynamics methods. *The Journal of chemical physics*, 81(1):511–519, 1984.

[45] William G Hoover. Canonical dynamics: Equilibrium phase-space distributions. *Physical review A*, 31(3):1695, 1985.

[46] Glenn J Martyna, Michael L Klein, and Mark Tuckerman. Nosé–hoover chains: The canonical ensemble via continuous dynamics. *The Journal of chemical physics*, 97(4):2635–2643, 1992.

[47] Michele Parrinello and Aneesur Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied physics*, 52(12):7182–7190, 1981.

[48] Berk Hess, Henk Bekker, Herman JC Berendsen, and Johannes GEM Fraaije. Lincs: a linear constraint solver for molecular simulations. *Journal of computational chemistry*, 18(12):1463–1472, 1997.

[49] Shuichi Miyamoto and Peter A Kollman. Settle: An analytical version of the shake and rattle algorithm for rigid water models. *Journal of computational chemistry*, 13(8):952–962, 1992.

[50] James P Hughes, Stephen Rees, S Barrett Kalindjian, and Karen L Philpott. Principles of early drug discovery. *British journal of pharmacology*, 162(6):1239–1249, 2011.

[51] Samuel Genheden and Ulf Ryde. The mm/pbsa and mm/gbsa methods to estimate ligand-binding affinities. *Expert opinion on drug discovery*, 10(5):449–461, 2015.

[52] Ercheng Wang, Huiyong Sun, Junmei Wang, Zhe Wang, Hui Liu, John ZH Zhang, and Tingjun Hou. End-point binding free energy calculation with mm/pbsa and mm/gbsa: strategies and applications in drug design. *Chemical reviews*, 119(16):9478–9508, 2019.

[53] Leonardo G Ferreira, Ricardo N Dos Santos, Glaucius Oliva, and Adriano D Andricopulo. Molecular docking and structure-based drug design strategies. *Molecules*, 20(7):13384–13421, 2015.

[54] Nataraj S Pagadala, Khajamohiddin Syed, and Jack Tuszynski. Software for molecular docking: a review. *Biophysical reviews*, 9(2):91–102, 2017.

[55] Oleg Trott and Arthur J Olson. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, 31(2):455–461, 2010.

[56] Ruben Abagyan, Maxim Totrov, and Dmitry Kuznetsov. Icm—a new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation. *Journal of computational chemistry*, 15(5):488–506, 1994.

[57] Rashmi Kumari, Rajendra Kumar, Open Source Drug Discovery Consortium, and Andrew Lynn. g_mmpbsa - a gromacs tool for high-throughput mm-pbsa calculations. *Journal of chemical information and modeling*, 54(7):1951–1962, 2014.

[58] Mark James Abraham, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy C Smith, Berk Hess, and Erik Lindahl. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1:19–25, 2015.

[59] Nathan A Baker, David Sept, Simpson Joseph, Michael J Holst, and J Andrew McCammon. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proceedings of the National Academy of Sciences*, 98(18):10037–10041, 2001.

[60] Richard BM Schasfoort. *Handbook of surface plasmon resonance*. Royal Society of Chemistry, 2017.

[61] World health organization weekly epidemiological update. `https://www.who.int/publications/m/item/weekly-epidemiological-update---10-march-2021`. Accessed: 2021-03-13.

[62] Andrew Rambaut, Nick Loman, Oliver Pybus, Wendy Barclay, Jeff Barrett, Alesandro Carabelli, Tom Connor, Tom Peacock, David L Robertson, and Erik Volz. Preliminary genomic characterisation of an emergent sars-cov-2 lineage in the uk defined by a novel set of spike mutations. `https://virological.`

org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by
563.

[63] Houriiyah Tegally, Eduan Wilkinson, Marta Giovanetti, Arash Iranzadeh, Vagner Fonseca, Jennifer Giandhari, Deelan Doolabh, Sureshnee Pillay, Emmanuel James San, Nokukhanya Msomi, et al. Emergence of a sars-cov-2 variant of concern with mutations in spike glycoprotein. *Nature*, pages 1–8, 2021.

[64] Paola Cristina Resende, João Felipe Bezerra, Romero Henrique Teixeira de Vasconcelos, Ighor Arantes, Luciana Appolinario, Ana Carolina Mendonça, Anna Carolina Paixao, Ana Carolina Duarte Rodrigues, Thauane Silva, Alice Sampaio Rocha, et al. Spike e484k mutation in the first sars-cov-2 reinfection case confirmed in brazil, 2020. https://virological.org/t/spike-e484k-mutation-in-the-first-sars-cov-2-reinfection-case-confirmed-in-brazil-2020/584.

[65] Tyler N Starr, Allison J Greaney, Sarah K Hilton, Daniel Ellis, Katharine HD Crawford, Adam S Dingens, Mary Jane Navarro, John E Bowen, M Alejandra Tortorici, Alexandra C Walls, et al. Deep mutational scanning of sars-cov-2 receptor binding domain reveals constraints on folding and ace2 binding. *Cell*, 182(5):1295–1310, 2020.

[66] Baofu Qiao and Monica Olvera de la Cruz. Enhanced binding of sars-cov-2 spike protein to receptor by distal polybasic cleavage sites. *ACS nano*, 14(8):10616–10623, 2020.

[67] Bryan A Johnson, Xuping Xie, Adam L Bailey, Birte Kalveram, Kumari G Lokugamage, Antonio Muruato, Jing Zou, Xianwen Zhang, Terry Juelich, Jennifer K Smith, et al. Loss of furin cleavage site attenuates sars-cov-2 pathogenesis. *Nature*, 591(7849):293–299, 2021.

[68] Charles Ramassamy. Emerging role of polyphenolic compounds in the treatment of neurodegenerative diseases: a review of their intracellular targets. *European journal of pharmacology*, 545(1):51–64, 2006.

[69] Roger Gaudreault and Normand Mousseau. Mitigating alzheimer's disease with natural polyphenols: a review. *Current Alzheimer Research*, 16(6):529–543, 2019.

[70] Roger Gaudreault, Vincent Hervé, Theo GM van de Ven, Normand Mousseau, and Charles Ramassamy. Polyphenol-peptide interactions in mitigation of alzheimer's disease: Role of biosurface-induced aggregation. *Journal of Alzheimer's Disease*, 81(1):1–23, 2021.

[71] Alessandra Durazzo, Massimo Lucarini, Eliana B Souto, Carla Cicala, Elisabetta Caiazzo, Angelo A Izzo, Ettore Novellino, and Antonello Santini. Polyphenols: A concise overview on the chemistry, occurrence, and human health. *Phytotherapy Research*, 33(9):2221–2243, 2019.

[72] Ta-chen Lin, Feng-lin Hsu, and Juei-Tang Cheng. Antihypertensive activity of corilagin and chebulinic acid, tannins from lumnitzera, racemosa. *Journal of Natural Products*, 56(4):629–632, 1993.

[73] Jiajia Xu, Gongye Zhang, Yinping Tong, Jiahui Yuan, Yuanyue Li, and Gang Song. Corilagin induces apoptosis, autophagy and ros generation in gastric cancer cells in vitro. *International journal of molecular medicine*, 43(2):967–979, 2019.

[74] F Notka, GR Meier, and R Wagner. Inhibition of wild-type human immunodeficiency virus and reverse transcriptase inhibitor-resistant variants by phyllanthus amarus. *Antiviral research*, 58(2):175–186, 2003.

[75] Frank Notka, Georg Meier, and Ralf Wagner. Concerted inhibitory activities of phyllanthus amarus on hiv replication in vitro and ex vivo. *Antiviral Research*, 64(2):93–102, 2004.

[76] Liang-Tzung Lin, Wen-Chan Hsu, and Chun-Ching Lin. Antiviral natural products and herbal medicines. *Journal of traditional and complementary medicine*, 4(1):24–35, 2014.

[77] Sang-Gu Yeo, Jae Hyoung Song, Eun-Hye Hong, Bo-Ra Lee, Yong Soo Kwon, Sun-Young Chang, Seung Hyun Kim, Sang won Lee, Jae-Hak Park, and Hyun-Jeong Ko. Antiviral effects of phyllanthus urinaria containing corilagin against human enterovirus 71 and coxsackievirus a16 in vitro. *Archives of pharmacal research*, 38(2):193–202, 2015.

[78] Ling Yi, Zhengquan Li, Kehu Yuan, Xiuxia Qu, Jian Chen, Guangwen Wang, Hong Zhang, Hongpeng Luo, Lili Zhu, Pengfei Jiang, et al. Small molecules blocking the entry of severe acute respiratory syndrome coronavirus into host cells. *Journal of virology*, 78(20):11334–11339, 2004.

[79] Lanying Du, Yuxian He, Yusen Zhou, Shuwen Liu, Bo-Jian Zheng, and Shibo Jiang. The spike protein of sars-cov—a target for vaccine and therapeutic development. *Nature Reviews Microbiology*, 7(3):226–236, 2009.

[80] Ines L Paraiso, Johana S Revel, and Jan F Stevens. Potential use of polyphenols in the battle against covid-19. *Current Opinion in Food Science*, 32:149–155, 2020.

[81] Kateryna Miroshnychenko and Anna V. Shestopalova. Combined use of amentoflavone and ledipasvir could interfere with binding of spike glycoprotein of sars-cov-2 to ace2: The results of molecular docking study. *10.26434/chemrxiv.12377870.v1*, 5 2020.

[82] Preeti Pandey, Jitendra Subhash Rane, Aroni Chatterjee, Abhijeet Kumar, Rajni Khan, Amresh Prakash, and Shashikant Ray. Targeting sars-cov-2 spike protein of covid-19 with naturally occurring phytochemicals: an in silico study for drug development. *Journal of Biomolecular Structure and Dynamics*, pages 1–11, 2020.

[83] Roger Gaudreault, Theo GM van de Ven, and Michael A Whitehead. Molecular modeling of poly (ethylene oxide) model cofactors; 1, 3, 6-tri-o-galloyl-β-d-glucose and corilagin. *Molecular modeling annual*, 8(3):73–80, 2002.

[84] R Gaudreault, TGM van de Ven, and MA Whitehead. Theoretical studies of the interactions between complex molecules in the gas phase. *Journal of Physical Chemistry A*, pages 3692–3702, 2006.

[85] B Uma Reddy, Ranajoy Mullick, Anuj Kumar, Geetika Sharma, Paromita Bag, Chaitrali Laha Roy, Govindarajan Sudha, Himani Tandon, Pratik Dave, Ashutosh Shukla, et al. A natural small molecule inhibitor corilagin blocks hcv replication and modulates oxidative stress to reduce liver damage. *Antiviral research*, 150:47–59, 2018.

[86] Xuan Li, Yuan Deng, Zhizhong Zheng, Wen Huang, Lianghua Chen, Qingxuan Tong, and Yanlin Ming. Corilagin, a promising medicinal herbal agent. *Biomedicine & Pharmacotherapy*, 99:43–50, 2018.

[87] Juei-Tang Cheng, Ta-Chen Lin, and Feng-Lin Hsu. Antihypertensive effect of corilagin in the rat. *Canadian journal of physiology and pharmacology*, 73(10):1425–1429, 1995.

[88] Feng Jin, Du Cheng, Jun-Yan Tao, Shu-Ling Zhang, Ran Pang, Yuan-Jin Guo, Pian Ye, Ji-Hua Dong, and Lei Zhao. Anti-inflammatory and anti-oxidative effects of corilagin in a rat model of acute cholestasis. *BMC gastroenterology*, 13(1):1–10, 2013.

[89] Veronica Salmaso and Stefano Moro. Bridging molecular docking to molecular dynamics in exploring ligand-protein recognition process: An overview. *Frontiers in pharmacology*, 9:923, 2018.

[90] Jennifer Loschwitz, Anna Jäckering, Monika Keutmann, Maryam Olagunju, Raphael J Eberle, Monika Aparecida Coronado, Olujide O Olubiyi, and Birgit Strodel. Novel inhibitors of the main protease enzyme of sars-cov-2 identified via molecular dynamics simulation-guided in vitro assay. *Bioorganic Chemistry*, 111:104862, 2021.

[91] Peter A. Kollman, Irina Massova, Carolina Reyes, Bernd Kuhn, Shuanghong Huo, Lillian Chong, Matthew Lee, Taisung Lee, Yong Duan, Wei Wang, Oreola Donini, Piotr Cieplak, Jaysharee Srinivasan, David A. Case, and Thomas E. Cheatham. Calculating structures and free energies of complex

molecules: Combining molecular mechanics and continuum models. *Accounts of Chemical Research*, 33(12):889–897, 2000. PMID: 11123888.

[92] Tom Darden, Darrin York, and Lee Pedersen. Particle mesh ewald: An n log (n) method for ewald sums in large systems. *The Journal of chemical physics*, 98(12):10089–10092, 1993.

[93] Ulrich Essmann, Lalith Perera, Max L Berkowitz, Tom Darden, Hsing Lee, and Lee G Pedersen. A smooth particle mesh ewald method. *The Journal of chemical physics*, 103(19):8577–8593, 1995.

[94] Xavier Daura, Karl Gademann, Bernhard Jaun, Dieter Seebach, Wilfred F Van Gunsteren, and Alan E Mark. Peptide folding: when simulation meets experiment. *Angewandte Chemie International Edition*, 38(1-2):236–240, 1999.

[95] LLC Schrödinger and Warren DeLano. Pymol.

[96] Zhe Wang, Huiyong Sun, Xiaojun Yao, Dan Li, Lei Xu, Youyong Li, Sheng Tian, and Tingjun Hou. Comprehensive evaluation of ten docking programs on a diverse set of protein–ligand complexes: the prediction accuracy of sampling power and scoring power. *Physical Chemistry Chemical Physics*, 18(18):12964–12975, 2016.

[97] Noel M O'Boyle, Michael Banck, Craig A James, Chris Morley, Tim Vandermeersch, and Geoffrey R Hutchison. Open babel: An open chemical toolbox. *Journal of cheminformatics*, 3(1):33, 2011.

[98] Junmei Wang, Wei Wang, Peter A Kollman, and David A Case. Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of molecular graphics and modelling*, 25(2):247–260, 2006.

[99] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox. Gaussian~16 Revision C.01, 2016. Gaussian Inc. Wallingford CT.

[100] Alan W Sousa Da Silva and Wim F Vranken. Acpype-antechamber python parser interface. *BMC research notes*, 5(1):367, 2012.

[101] Wolfgang Kabsch and Christian Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers: Original Research on Biomolecules*, 22(12):2577–2637, 1983.

[102] David J Barlow and JM Thornton. Ion-pairs in proteins. *Journal of molecular biology*, 168(4):867–885, 1983.

[103] Andrew C Wallace, Roman A Laskowski, and Janet M Thornton. Ligplot: a program to generate schematic diagrams of protein-ligand interactions. *Protein engineering, design and selection*, 8(2):127–134, 1995.

[104] Roman A Laskowski and Mark B Swindells. Ligplot+: multiple ligand–protein interaction diagrams for drug discovery, 2011.

[105] Daniel Wrapp, Nianshuang Wang, Kizzmekia S Corbett, Jory A Goldsmith, Ching-Lin Hsieh, Olubukola Abiona, Barney S Graham, and Jason S McLellan. Cryo-em structure of the 2019-ncov spike in the prefusion conformation. *Science*, 367(6483):1260–1263, 2020.

[106] Boyu Pan, Senbiao Fang, Ju Zhang, Ya Pan, Han Liu, Yun Wang, Min Li, and Liren Liu. Chinese herbal compounds against sars-cov-2: Puerarin and quercetin impair the binding of viral s-protein to ace2 receptor. *Computational and structural biotechnology journal*, 18:3518–3527, 2020.

[107] Quinlin M Hanson, Kelli M Wilson, Min Shen, Zina Itkin, Richard T Eastman, Paul Shinn, and Matthew D Hall. Targeting ace2–rbd interaction as a platform for covid-19 therapeutics: Development and drug-repurposing screen of an alphalisa proximity assay. *ACS Pharmacology & Translational Science*, 3(6):1352–1360, 2020.

[108] Minghui Yang, Jinli Wei, Ting Huang, Luping Lei, Chenguang Shen, Jinzhi Lai, Min Yang, Lei Liu, Yang Yang, Guoshi Liu, et al. Resveratrol inhibits the replication of severe acute respiratory syndrome coronavirus 2 (sars-cov-2) in cultured vero cells. *Phytotherapy Research*, 35(3):1127–1129, 2020.

[109] Gary Williamson and Asimina Kerimi. Testing of natural products in clinical trials targeting the sars-cov-2 (covid-19) viral spike protein-angiotensin converting enzyme-2 (ace2) interaction. *Biochemical pharmacology*, page 114123, 2020.

[110] Yasanandana Supunsiri Wijayasinghe, Pravin Bhansali, Ronald E Viola, Mohammad A Kamal, and Nitesh Kumar Poddar. Natural products: A rich source of antiviral drug lead candidates for the management of covid-19. *Current pharmaceutical design*, 26:1–25, 2020.

[111] Arquimedes Gasparotto, Sara Emília Lima Tolouei, Francislaine Aparecida dos Reis Lívero, Francielli Gasparotto, Thaise Boeing, and Priscila de Souza. Natural agents modulating ace-2: A review of compounds with potential against sars-cov-2 infections. *Current Pharmaceutical Design*, 27(13):1588–1596, 2021.

[112] Matthew Zirui Tay, Chek Meng Poh, Laurent Rénia, Paul A MacAry, and Lisa FP Ng. The trinity of covid-19: immunity, inflammation and intervention. *Nature Reviews Immunology*, 20(6):363–374, 2020.

[113] Chaolin Huang, Yeming Wang, Xingwang Li, Lili Ren, Jianping Zhao, Yi Hu, Li Zhang, Guohui Fan, Jiuyang Xu, Xiaoying Gu, et al. Clinical features of patients infected with 2019 novel coronavirus in wuhan, china. *The lancet*, 395(10223):497–506, 2020.

[114] Lei Zhao, Shu-Ling Zhang, Jun-Yan Tao, Ran Pang, Feng Jin, Yuan-Jin Guo, Ji-Hua Dong, Pian Ye, Hong-Yang Zhao, and Guo-Hua Zheng. Preliminary exploration on anti-inflammatory mechanism of corilagin (beta-1-o-galloyl-3, 6-(r)-hexahydroxydiphenoyl-d-glucose) in vitro. *International immunopharmacology*, 8(7):1059–1064, 2008.

[115] Xiao-Rong Dong, Ming Luo, Li Fan, Tao Zhang, Li Liu, Ji-Hua Dong, and Gang Wu. Corilagin inhibits the double strand break-triggered nf-κb pathway in irradiated microglial cells. *International journal of molecular medicine*, 25(4):531–536, 2010.

[116] Kumju Youn, Seonah Lee, Woo-Sik Jeong, Chi-Tang Ho, and Mira Jun. Protective role of corilagin on ab25–35-induced neurotoxicity: Suppression of nf-kb signaling pathway. *JOURNAL OF MEDICINAL FOOD*, 19(10):901–911, 2016.

[117] Xing Zhu, Dhiraj Mannar, Shanti S Srivastava, Alison M Berezuk, Jean-Philippe Demers, James W Saville, Karoline Leopold, Wei Li, Dimiter S Dimitrov, Katharine S Tuttle, et al. Cryo-electron microscopy structures of the n501y sars-cov-2 spike protein in complex with ace2 and 2 potent neutralizing antibodies. *PLoS biology*, 19(4):e3001237, 2021.

[118] Gard Nelson, Oleksandr Buzko, Patricia R Spilman, Kayvan Niazi, Shahrooz Rabizadeh, and Patrick R Soon-Shiong. Molecular dynamic simulation reveals e484k mutation enhances spike rbd-ace2 affinity

and the combination of e484k, k417n and n501y mutations (501y. v2 variant) induces conformational change greater than n501y mutant alone, potentially resulting in an escape mutant. *BioRxiv*, 2021.

[119] Budheswar Dehury, Vishakha Raina, Namrata Misra, and Mrutyunjay Suar. Effect of mutation on structure, function and dynamics of receptor binding domain of human sars-cov-2 with host cell receptor ace2: a molecular dynamics simulations study. *Journal of Biomolecular Structure and Dynamics*, pages 1–15, 2020.

[120] Meredith Wadman and Jon Cohen. Novavax vaccine delivers 89% efficacy against covid-19 in uk—but is less potent in south africa. *Science*, 2021.

[121] Alina Baum, Dharani Ajithdoss, Richard Copin, Anbo Zhou, Kathryn Lanza, Nicole Negron, Min Ni, Yi Wei, Kusha Mohammadi, Bret Musser, et al. Regn-cov2 antibodies prevent and treat sars-cov-2 infection in rhesus macaques and hamsters. *Science*, 370(6520):1110–1115, 2020.

[122] Peter Chen, Ajay Nirula, Barry Heller, Robert L Gottlieb, Joseph Boscia, Jason Morris, Gregory Huhn, Jose Cardona, Bharat Mocherla, Valentina Stosor, et al. Sars-cov-2 neutralizing antibody ly-cov555 in outpatients with covid-19. *New England Journal of Medicine*, 384(3):229–237, 2021.

[123] Pengfei Wang, Manoj S Nair, Lihong Liu, Sho Iketani, Yang Luo, Yicheng Guo, Maple Wang, Jian Yu, Baoshan Zhang, Peter D Kwong, et al. Antibody resistance of sars-cov-2 variants b. 1.351 and b. 1.1. 7. *Nature*, pages 1–6, 2021.

[124] Marek Widera, Alexander Wilhelm, Sebastian Hoehl, Christiane Pallas, Niko Kohmer, Timo Wolf, Holger F Rabenau, Victor M Corman, Christian Drosten, Maria JGT Vehreschild, et al. Bamlanivimab does not neutralize two sars-cov-2 variants carrying e484k in vitro. *medRxiv*, 2021.

[125] Leyun Wu, Cheng Peng, Zhijian Xu, and Weiliang Zhu. Predicting the potential effect of e484k mutation on the binding of 28 antibodies to the spike protein of sars-cov-2 by molecular dynamics simulation and free energy calculation. *10.26434/chemrxiv.13897091.v1*, 2021.

[126] Samuel K Kwofie, Emmanuel Broni, Seth O Asiedu, Gabriel B Kwarko, Bismark Dankwa, Kweku S Enninful, Elvis K Tiburu, and Michael D Wilson. Cheminformatics-based identification of potential novel anti-sars-cov-2 natural compounds of african origin. *Molecules*, 26(2):406, 2021.

[127] Charles H Chen and Timothy K Lu. Development and challenges of antimicrobial peptides for therapeutic applications. *Antibiotics*, 9(1):24, 2020.

[128] Cassandra Willyard. The drug-resistant bacteria that pose the greatest health threats. *Nature News*, 543(7643):15, 2017.

[129] Min-Duk Seo, Hyung-Sik Won, Ji-Hun Kim, Tsogbadrakh Mishig-Ochir, and Bong-Jin Lee. Antimicrobial peptides for therapeutic applications: a review. *Molecules*, 17(10):12276–12286, 2012.

[130] Christian B Anfinsen, Edgar Haber, Michael Sela, and FH White Jr. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proceedings of the National Academy of Sciences of the United States of America*, 47(9):1309, 1961.

[131] Ken A Dill and Justin L MacCallum. The protein-folding problem, 50 years on. *science*, 338(6110):1042–1046, 2012.

[132] José Nelson Onuchic and Peter G Wolynes. Theory of protein folding. *Current opinion in structural biology*, 14(1):70–75, 2004.

[133] Cyrus Levinthal. How to fold graciously. *Mossbauer spectroscopy in biological systems*, 67:22–24, 1969.

[134] Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, 32(5):922–923, 1976.

[135] Wolfgang Kabsch. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, 34(5):827–828, 1978.

[136] Kliment Olechnovič, Eleonora Kulberkytė, and Česlovas Venclovas. Cad-score: a new contact area difference-based function for evaluation of protein structural models. *Proteins: Structure, Function, and Bioinformatics*, 81(1):149–162, 2013.

[137] Frédéric Guyon and Pierre Tuffery. Fast protein fragment similarity scoring using a binet–cauchy kernel. *Bioinformatics*, 30(6):784–791, 2014.

[138] Kliment Olechnovic, Bohdan Monastyrskyy, Andriy Kryshtafovych, Ceslovas Venclovas, and Alfonso Valencia. Comparative analysis of methods for evaluation of protein models against native structures. *Bioinformatics*, 1:8, 2018.

[139] Brendan J McConkey, Vladimir Sobolev, and Marvin Edelman. Quantification of protein surfaces, volumes and atom–atom contacts using a constrained voronoi procedure. *Bioinformatics*, 18(10):1365–1373, 2002.

[140] Deok-Soo Kim, Youngsong Cho, and Donguk Kim. Euclidean voronoi diagram of 3d balls and its computation via tracing edges. *Computer-Aided Design*, 37(13):1412–1424, 2005.

[141] Vincent B Chen, W Bryan Arendall, Jeffrey J Headd, Daniel A Keedy, Robert M Immormino, Gary J Kapral, Laura W Murray, Jane S Richardson, and David C Richardson. Molprobity: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D: Biological Crystallography*, 66(1):12–21, 2010.

[142] Andriy Kryshtafovych, Torsten Schwede, Maya Topf, Krzysztof Fidelis, and John Moult. Critical assessment of methods of protein structure prediction (casp)—round xiii. *Proteins: Structure, Function, and Bioinformatics*, 87(12):1011–1020, 2019.

[143] Harpreet Kaur, Aarti Garg, and Gajendra Pal Singh Raghava. Pepstr: a de novo method for tertiary structure prediction of small bioactive peptides. *Protein and peptide letters*, 14(7):626–631, 2007.

[144] Sandeep Singh, Harinder Singh, Abhishek Tuknait, Kumardeep Chaudhary, Balvinder Singh, S Kumaran, and Gajendra PS Raghava. Pepstrmod: structure prediction of peptides containing natural, non-natural and modified residues. *Biology direct*, 10(1):1–19, 2015.

[145] Shikai Jin, Vinicius G Contessoto, Mingchen Chen, Nicholas P Schafer, Wei Lu, Xun Chen, Carlos Bueno, Arya Hajitaheri, Brian J Sirovetz, Aram Davtyan, et al. Awsem-suite: a protein structure prediction server based on template-guided, coevolutionary-enhanced optimized folding landscapes. *Nucleic acids research*, 48(W1):W25–W30, 2020.

[146] Aram Davtyan, Nicholas P Schafer, Weihua Zheng, Cecilia Clementi, Peter G Wolynes, and Garegin A Papoian. Awsem-md: protein structure prediction using coarse-grained physical potentials and bioinformatically based local structure biasing. *The Journal of Physical Chemistry B*, 116(29):8494–8503, 2012.

[147] Alexis Lamiable, Pierre Thévenet, Julien Rey, Marek Vavrusa, Philippe Derreumaux, and Pierre Tuffery. Pep-fold3: faster de novo structure prediction for linear peptides in solution and in complex. *Nucleic acids research*, 44(W1):W449–W454, 2016.

[148] Julien Maupetit, Philippe Derreumaux, and Pierre Tuffery. A fast method for large-scale de novo peptide and miniprotein structure prediction. *Journal of computational chemistry*, 31(4):726–738, 2010.

[149] Alex L Mitchell, Alexandre Almeida, Martin Beracochea, Miguel Boland, Josephine Burgin, Guy Cochrane, Michael R Crusoe, Varsha Kale, Simon C Potter, Lorna J Richardson, et al. Mgnify: the microbiome analysis resource in 2020. *Nucleic acids research*, 48(D1):D570–D578, 2020.

[150] Martin Steinegger, Milot Mirdita, and Johannes Söding. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nature methods*, 16(7):603–606, 2019.

[151] UniProt Consortium. Uniprot: a worldwide hub of protein knowledge. *Nucleic acids research*, 47(D1):D506–D515, 2019.

[152] Jinbo Xu, Matthew Mcpartlon, and Jin Li. Improved protein structure prediction by deep learning irrespective of co-evolution information. *Nature Machine Intelligence*, pages 1–9, 2021.

[153] Jinbo Xu. Distance-based protein folding powered by deep learning. *Proceedings of the National Academy of Sciences*, 116(34):16856–16865, 2019.

[154] Sheng Wang, Siqi Sun, Zhen Li, Renyu Zhang, and Jinbo Xu. Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS computational biology*, 13(1):e1005324, 2017.

[155] Patrick Brendan Timmons and Chandralal M Hewage. Apptest is a novel protocol for the automatic prediction of peptide tertiary structures. *Briefings in Bioinformatics*, 22:bbab308, 2021.

[156] Philippe Derreumaux. From polypeptide sequences to structures using monte carlo simulations and an optimized potential. *The Journal of chemical physics*, 111(5):2301–2310, 1999.

[157] William George Noid. Perspective: Coarse-grained models for biomolecular systems. *The Journal of chemical physics*, 139(9):09B201_1, 2013.

[158] Samuela Pasquali and Philippe Derreumaux. Hire-rna: a high resolution coarse-grained energy model for rna. *The journal of physical chemistry B*, 114(37):11957–11966, 2010.

[159] Sébastien Côté, Philippe Derreumaux, and Normand Mousseau. Distinct morphologies for amyloid beta protein monomer: Aβ1–40, aβ1–42, and aβ1–40 (d23n). *Journal of chemical theory and computation*, 7(8):2584–2592, 2011.

[160] Sébastien Côté, Rozita Laghaei, Philippe Derreumaux, and Normand Mousseau. Distinct dimerization for various alloforms of the amyloid-beta protein: Aβ1–40, aβ1–42, and aβ1–40 (d23n). *The journal of physical chemistry B*, 116(13):4043–4055, 2012.

[161] Yassmine Chebaro, Ping Jiang, Tong Zang, Yuguang Mu, Phuong H Nguyen, Normand Mousseau, and Philippe Derreumaux. Structures of aβ17–42 trimers in isolation and with five small-molecule drugs using a hierarchical computational procedure. *The Journal of Physical Chemistry B*, 116(29):8412–8422, 2012.

[162] Fabio Sterpone, Simone Melchionna, Pierre Tuffery, Samuela Pasquali, Normand Mousseau, Tristan Cragnolini, Yassmine Chebaro, Jean-Francois St-Pierre, Maria Kalimeri, Alessandro Barducci, et al. The opep protein model: from single molecules, amyloid formation, crowding and hydrodynamics to dna/rna systems. *Chemical Society reviews*, 43(13):4871–4893, 2014.

[163] Philipp Kynast, Philippe Derreumaux, and Birgit Strodel. Evaluation of the coarse-grained opep force field for protein-protein docking. *BMC biophysics*, 9(1):1–17, 2016.

[164] Thanh Thuy Tran, Phuong H Nguyen, and Philippe Derreumaux. Lattice model for amyloid peptides: Opep force field parametrization and applications to the nucleus size of alzheimer's peptides. *The Journal of chemical physics*, 144(20):205103, 2016.

[165] Julien Maupetit, P Tuffery, and Philippe Derreumaux. A coarse-grained protein force field for folding and structure prediction. *Proteins: Structure, Function, and Bioinformatics*, 69(2):394–408, 2007.

[166] Julien Maupetit. *Génération ab initio de modèles protéiques à partir de représentations discrètes des protéines et de critères d ' énergie simplifiés.* PhD thesis, Sorbonne Paris Cité, 2007.

[167] Marcos R Betancourt and D Thirumalai. Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. *Protein science*, 8(2):361–369, 1999.

[168] Pierre Tuffery, Frédéric Guyon, and Philippe Derreumaux. Improved greedy algorithm for protein structure reconstruction. *Journal of computational chemistry*, 26(5):506–513, 2005.

[169] Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710, 2004.

[170] Yassmine Chebaro, Samuela Pasquali, and Philippe Derreumaux. The coarse-grained opep force field for non-amyloid and amyloid proteins. *The Journal of Physical Chemistry B*, 116(30):8741–8752, 2012.

[171] Sébastien Côté. *Développements et applications de méthodes computationnelles pour l'étude de l'agrégation des protéines amyloïdes*. PhD thesis, Université de Montréal, 2016.

[172] Fabio Sterpone, Phuong H Nguyen, Maria Kalimeri, and Philippe Derreumaux. Importance of the ion-pair interactions in the opep coarse-grained force field: parametrization and validation. *Journal of chemical theory and computation*, 9(10):4574–4584, 2013.

[173] William L Jorgensen and Julian Tirado-Rives. The opls [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *Journal of the American Chemical Society*, 110(6):1657–1666, 1988.

[174] Fernando Luís Barroso da Silva, Fabio Sterpone, and Philippe Derreumaux. Opep6: A new constant-ph molecular dynamics simulation scheme with opep coarse-grained force field. *Journal of chemical theory and computation*, 15(6):3875–3888, 2019.

[175] Harinder Singh, Sandeep Singh, and Gajendra Pal Singh Raghava. Peptide secondary structure prediction using evolutionary information. *bioRxiv*, page 558791, 2019.

[176] Yimin Shen, Julien Maupetit, Philippe Derreumaux, and Pierre Tuffery. Improved pep-fold approach for peptide and miniprotein structure prediction. *Journal of chemical theory and computation*, 10(10):4745–4758, 2014.

[177] Bertrand Néron, Hervé Ménager, Corinne Maufrais, Nicolas Joly, Julien Maupetit, Sébastien Letort, Sébastien Carrere, Pierre Tuffery, and Catherine Letondal. Mobyle: a new full web bioinformatics framework. *Bioinformatics*, 25(22):3005–3011, 2009.

[178] Anne-Cloude Camproux, Romain Gautier, and Pierre Tuffery. A hidden markov model derived structural alphabet for proteins. *Journal of molecular biology*, 339(3):591–605, 2004.

[179] Dmitrij Frishman and Patrick Argos. Knowledge-based protein secondary structure assignment. *Proteins: Structure, Function, and Bioinformatics*, 23(4):566–579, 1995.

[180] Stephen F Altschul, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402, 1997.

[181] Gustav Mie. Zur kinetischen theorie der einatomigen körper. *Annalen der Physik*, 316(8):657–697, 1903.

[182] Sara Goodwin, John D McPherson, and W Richard McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6):333, 2016.

[183] John D Westbrook and Stephen K Burley. How structural biologists and the protein data bank contributed to recent fda new drug approvals. *Structure*, 27(2):211–217, 2019.

[184] Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N Kinch, R Dustin Schaeffer, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 2021.

[185] Heriberto Bruzzoni-Giovanelli, Valerie Alezra, Nicolas Wolff, Chang-Zhi Dong, Pierre Tuffery, and Angelita Rebollo. Interfering peptides targeting protein–protein interactions: the next generation of drugs? *Drug Discovery Today*, 23(2):272–285, 2018.

[186] Xiao Cao, Wenjia He, Zitan Chen, Yifan Li, Kexin Wang, Hongbo Zhang, Lesong Wei, Lizhen Cui, Ran Su, and Leyi Wei. Pssp-mvirt: peptide secondary structure prediction based on a multi-view deep learning architecture. *Briefings in Bioinformatics*, 2021.

[187] Xiao Ru and Zijing Lin. Genetic algorithm embedded with a search space dimension reduction scheme for efficient peptide structure predictions. *The Journal of Physical Chemistry B*, 125(15):3824–3829, 2021.

[188] Julien Maupetit, Philippe Derreumaux, and Pierre Tuffery. Pep-fold: an online resource for de novo peptide structure prediction. *Nucleic acids research*, 37(suppl_2):W498–W503, 2009.

[189] Irene Maffucci and Alessandro Contini. In silico drug repurposing for sars-cov-2 main proteinase and spike proteins. *Journal of proteome research*, 19(11):4637–4648, 2020.

[190] Abhishek Singh, Mukesh Thakur, Lalit Kumar Sharma, and Kailash Chandra. Designing a multi-epitope peptide based vaccine against sars-cov-2. *Scientific reports*, 10(1):1–12, 2020.

[191] Muhammad Tahir ul Qamar, Abdur Rehman, Kishver Tusleem, Usman Ali Ashfaq, Muhammad Qasim, Xitong Zhu, Israr Fatima, Farah Shahid, and Ling-Ling Chen. Designing of a next generation multiepitope based vaccine (mev) against sars-cov-2: Immunoinformatics and in silico approaches. *PloS one*, 15(12):e0244176, 2020.

[192] Maciej Blaszczyk, Michal Jamroz, Sebastian Kmiecik, and Andrzej Kolinski. Cabs-fold: server for the de novo and consensus-based prediction of protein structure. *Nucleic acids research*, 41(W1):W406–W411, 2013.

[193] Michael Levitt. Protein folding by restrained energy minimization and molecular dynamics. *Journal of molecular biology*, 170(3):723–764, 1983.

[194] Jane cek Jiří, Olivier Said-Aizpuru, and Patrice Paricaud. Long range corrections for inhomogeneous simulations of mie n–m potential. *Journal of chemical theory and computation*, 13(9):4482–4491, 2017.

[195] Pierre Thévenet, Yimin Shen, Julien Maupetit, Frederic Guyon, Philippe Derreumaux, and Pierre Tuffery. Pep-fold: an updated de novo structure prediction server for both linear and disulfide bonded cyclic peptides. *Nucleic acids research*, 40(W1):W288–W293, 2012.

[196] Stefano Piana, Kresten Lindorff-Larsen, and David E Shaw. How robust are protein folding simulations with respect to force field parameterization? *Biophysical journal*, 100(9):L47–L49, 2011.

[197] Russell Eberhart and James Kennedy. Particle swarm optimization. In *Proceedings of the IEEE international conference on neural networks*, volume 4, pages 1942–1948, 1995.

[198] Vincent Binette, Normand Mousseau, and Pierre Tuffery. A generalized attraction–repulsion potential and revisited fragment library improves pep-fold peptide structure prediction. *Journal of Chemical Theory and Computation*, 2022.

[199] Association Alzheimer's. 2021 alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 17(3):327–406, 2021.

[200] Nunilo Cremades and Christopher M Dobson. The contribution of biophysical and structural studies of protein self-assembly to the design of therapeutic strategies for amyloid diseases. *Neurobiology of disease*, 109:178–190, 2018.

[201] Pu Chun Ke, Marc-Antonie Sani, Feng Ding, Aleksandr Kakinen, Ibrahim Javed, Frances Separovic, Thomas P Davis, and Raffaele Mezzenga. Implications of peptide assemblies in amyloid diseases. *Chemical Society Reviews*, 46(21):6492–6531, 2017.

[202] Andrew J Doig and Philippe Derreumaux. Inhibition of protein aggregation and amyloid formation by small molecules. *Current opinion in structural biology*, 30:50–56, 2015.

[203] Guo-fang Chen, Ting-hai Xu, Yan Yan, Yu-ren Zhou, Yi Jiang, Karsten Melcher, and H Eric Xu. Amyloid beta: structure, biology and structure-based therapeutic development. *Acta Pharmacologica Sinica*, 38(9):1205–1235, 2017.

[204] Robert B Best. Computational and theoretical advances in studies of intrinsically disordered proteins. *Current opinion in structural biology*, 42:147–154, 2017.

[205] Giovanni Bussi. Hamiltonian replica exchange in gromacs: a flexible implementation. *Molecular Physics*, 112(3-4):379–384, 2014.

[206] Lingle Wang, Richard A Friesner, and BJ Berne. Replica exchange with solute scaling: a more efficient version of replica exchange with solute tempering (rest2). *The Journal of Physical Chemistry B*, 115(30):9431–9438, 2011.

[207] Jean-François St-Pierre and Normand Mousseau. Large loop conformation sampling using the activation relaxation technique, art-nouveau method. *Proteins: Structure, Function, and Bioinformatics*, 80(7):1883–1894, 2012.

[208] L Dupuis and Normand Mousseau. Understanding the ef-hand closing pathway using non-biased interatomic potentials. *The Journal of chemical physics*, 136(3):035101, 2012.

[209] Riccardo Poli. Analysis of the publications on the applications of particle swarm optimisation. *Journal of Artificial Evolution and Applications*, 2008, 2008.

[210] Georgii G Krivov, Maxim V Shapovalov, and Roland L Dunbrack Jr. Improved prediction of protein side-chain conformations with scwrl4. *Proteins: Structure, Function, and Bioinformatics*, 77(4):778–795, 2009.

[211] Abil E Aliev, Martin Kulke, Harmeet S Khaneja, Vijay Chudasama, Tom D Sheppard, and Rachel M Lanigan. Motional timescale predictions by molecular dynamics simulations: case study using proline and hydroxyproline sidechain dynamics. *Proteins: Structure, Function, and Bioinformatics*, 82(2):195–215, 2014.

[212] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *The Journal of chemical physics*, 126(1):014101, 2007.

[213] Gareth A Tribello, Massimiliano Bonomi, Davide Branduardi, Carlo Camilloni, and Giovanni Bussi. Plumed 2: New feathers for an old bird. *Computer Physics Communications*, 185(2):604–613, 2014.

[214] Yuji Sugita and Yuko Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chemical physics letters*, 314(1-2):141–151, 1999.

Annexe A

Supporting Figures: A generalized attraction-repulsion potential and revisited fragment library improves PEP-FOLD peptide structure prediction

A.1. Tested Targets

The 25 targets placed in the parametrization set (from the G/IC ensemble) are the following: 1b03 (β, 18), 1bhi (α/β, 38), 1cpz (α/β, 68), 1e0n (β, 27), 1fex (α, 59), 1g2h (α, 61), 1go5 (α, 69), 1i6c (β, 39), 1jjs (α, 50), 1spw (β, 39), 1uxd (α, 65), 1wcn (α, 70), 1yiu (β, 37), 1z4h (α/β, 66), 1zv6 (α, 68), 1zxg (α, 59), 2b7e (α, 59), 2bby (α/β, 69), 2dt6 (α, 64), 2fmr (α/β, 65), 2l92 (β, 50), 2l93 (α/β, 55), 2lma (α, 22), 2mwf (β, 32) and 2ysb (β, 49).

The rest of the targets from the G/IC ensemble were used to build a first validation set. It is composed of the following 39 targets: 1e0m (β, 37), 1ed7 (β, 45), 1f0z (α/β, 66), 1gyz (α, 62), 1ify (α, 49), 1k8b (α/β, 52), 1k91 (β, 37), 1n87 (α/β, 56), 1pgb (α/β, 56), 1q1v (α, 70), 1rij (α, 23), 1s4j (Coil, 13), 1u97 (α, 69), 1zrj (α, 50), 1zwv (α, 52), 2a63 (α/β, 66), 2bn6 (α, 33), 2bzt (α, 66), 2coo (α, 70), 2jof (α, 20), 2jtm (β, 60), 2k2a (α, 70), 2k57 (β, 55), 2kac (α/β, 64), 2kt2(α/β 69), 2kya(α, 34), 2l4j (β, 46), 2l4m (α/β, 69), 2lrx (α/β, 45), 2m6o (β, 48), 2mdj (β, 56), 2mi6 (β, 62), 2msu (α, 20), 2oru (β, 20), 2v0e(α/β, 55), 2ysh(β, 40), 2zaj(β, 49), 4c26 (α/β, 66), 5y22 (α, 22), 5ykl (α, 19)

Targets which were already correctly predicted by PEP-FOLDv1 were placed in the G/CC ensemble. The following 48 targets were identified: 1bwx(α, 39), 1by0(α, 27), 1cok(α, 68), 1du6(α, 64), 1dv0(α, 47), 1e0l(β, 37), 1f4i(α, 45), 1jrj(β, 39), 1k1v(α, 41), 1p9c(α, 45), 1pgy(α, 47), 1pv0(α, 46), 1qpm(α/β, 69), 1r4g(α, 53), 1rq6(α, 62), 1rzs(α, 61), 1uao(β, 10), 1wji(α, 63), 1wr3(β, 36), 1wr4(β, 36), 1wr7(β, 41), 1yyb(α, 27), 2cp9(α, 64

), 2dmv(β, 43)„ 2e5t(α, 46), 2ekk(α, 47), 2evq (β) (12), 2fce(α, 70), 2j8p(α, 49), 2jnh(α, 46), 2k76(α/β, 30), 2k9d(α, 44), 2ki0(α/β, 36), 2kz9(α, 69), 2l0g(α, 32), 2l54(α/β, 63), 2luf(α, 20), 2m8j(β, 43), 2n16(α, 20), 2p81(α, 44), 2wqg(α, 51), 2wxc(α, 47), 2ysc(β, 39), 2ysf(β, 40), 2ysg(β, 40), 2ysi(β, 40), 5t7q(α, 21), 6g4v(α, 24), 6r2x(α, 25)

Finally, targets for which no native/near-native predictions were generated by PEP-FOLDv1 were placed in the NG ensemble. The following 23 proteins were identified: 1gyf (α/β, 62), 1nd9 (α/β, 49), 1ne3 (β, 68), 1qxf (β, 66), 1vpu (α, 45), 1y2y (β, 68), 2cw1 (α/β, 65), 2do3 (β, 69), 2dy8 (α/β, 69), 2eqi (β, 69), 2gdl (α, 31), 2jrr (β, 67), 2jtv (α/β, 65), 2kaf (α/β, 67), 2l8d (β, 66), 2lhc (α, 56), 2lss (α/β, 70), 2m2l (α/β, 67), 2m4y (β, 56), 2m7o (α/β, 70), 2mck (α, 69), 2mdu (β, 29), 2xk0 (β, 69).

In order to compare with previously published version of PEP-FOLD [**147**, **148**, **176**], the following six targets with more than 30% of sequence identity were added: 1bwx (with 1fvy), 1e0l (with 2ysi), 1jrj (with 1rij), 1pgb (with 2lhc), 2l4j (with 1wmv) and 1f4i (with 1dv0).

Pair	r0	ϵ	gR0	n	m
ALA ALA	4.38	-0.24	3.89	18.00	2.91
ALA CYS	4.62	-0.25	3.42	3.99	2.70
ALA ASP	4.51	-0.07	2.93	5.19	0.77
ALA GLU	4.50	-0.04	2.60	4.06	0.60
ALA PHE	4.98	-0.30	3.35	3.99	1.46
ALA GLY	4.34	-0.04	2.65	4.00	0.84
ALA HIS	4.58	-0.04	3.29	3.99	2.23
ALA ILE	5.44	-0.56	3.50	3.99	1.13
ALA LYS	5.12	-0.03	2.93	3.99	0.59
ALA LEU	4.83	-0.23	3.74	4.23	3.59
ALA MET	5.77	-0.17	4.56	5.06	3.52
ALA ASN	4.73	-0.03	3.60	3.99	3.39
ALA PRO	5.37	-0.05	4.67	15.90	2.38
ALA GLN	5.96	-0.04	4.86	5.28	4.49
ALA ARG	7.16	-0.05	4.10	3.99	0.59
ALA SER	5.44	-0.06	4.40	6.06	3.56
ALA THR	4.41	-0.03	2.76	3.99	0.96
ALA VAL	4.04	-0.37	3.46	10.00	3.85
ALA TRP	6.41	-0.41	3.66	3.99	0.59
ALA TYR	5.79	-0.18	3.31	3.99	0.59
CYS CYS	4.52	-2.24	3.24	6.68	1.00
CYS ASP	5.05	-0.05	3.44	3.99	1.57
CYS GLU	4.08	-0.04	2.89	3.99	1.99
CYS PHE	5.46	-0.40	4.04	5.06	2.05
CYS GLY	4.26	-0.07	2.97	6.19	0.92
CYS HIS	5.85	-0.15	3.83	3.99	1.25
CYS ILE	5.81	-0.69	4.21	4.12	2.27
CYS LYS	5.13	-0.06	3.78	3.99	2.63
CYS LEU	6.60	-0.46	4.86	7.28	1.09
CYS MET	6.67	-0.32	4.78	4.06	2.15
CYS ASN	6.32	-0.02	4.81	8.21	1.23
CYS PRO	5.27	-0.19	4.70	11.05	6.88
CYS GLN	5.27	-0.03	3.07	4.13	0.62
CYS ARG	6.46	-0.02	3.69	3.99	0.59
CYS SER	4.60	-0.05	3.49	3.99	3.27
CYS THR	4.55	-0.07	2.60	3.99	0.59
CYS VAL	3.90	-0.61	3.04	4.55	3.46
CYS TRP	6.35	-0.64	3.63	3.99	0.59
CYS TYR	6.08	-0.18	3.74	3.99	0.88
ASP ASP	5.72	-0.04	4.35	8.13	1.22
ASP GLU	5.07	-0.06	2.91	4.02	0.60
ASP PHE	4.99	-0.05	3.58	3.99	2.22
ASP GLY	3.94	-0.05	2.93	3.99	2.85
ASP HIS	6.78	-0.27	4.33	4.98	0.74
ASP ILE	6.16	-0.04	4.52	3.99	2.54
ASP LYS	5.93	-0.42	4.32	3.99	2.45
ASP LEU	4.51	-0.02	3.38	7.74	1.16
ASP MET	5.25	-0.06	3.00	3.99	0.59
ASP ASN	5.22	-0.10	3.78	3.99	2.34
ASP PRO	6.83	-0.06	5.58	11.03	1.65
ASP GLN	6.34	-0.05	4.43	3.99	1.85
ASP ARG	5.02	-0.86	3.63	3.99	2.31
ASP SER	5.16	-0.05	3.85	3.99	2.89
ASP THR	4.44	-0.04	3.21	5.34	1.56
ASP VAL	3.92	-0.05	2.99	3.99	3.38
ASP TRP	4.46	-0.04	3.06	4.42	1.44
ASP TYR	5.79	-0.06	4.34	7.71	1.15
GLU GLU	4.15	-0.04	2.98	3.99	2.22
GLU PHE	5.97	-0.05	3.60	3.99	0.78
GLU GLY	4.26	-0.04	3.02	3.99	2.06
GLU HIS	6.65	-0.09	4.50	3.99	1.52
GLU ILE	6.79	-0.05	4.18	4.59	0.68
GLU LYS	4.48	-0.42	3.81	12.06	2.57
GLU LEU	7.66	-0.04	5.36	3.99	1.87
GLU MET	6.12	-0.03	3.51	3.99	0.60
GLU ASN	6.65	-0.01	4.82	3.99	2.37
GLU PRO	6.60	-0.03	5.16	9.06	1.35
GLU GLN	6.98	-0.04	5.27	3.99	3.14
GLU ARG	5.89	-0.41	4.22	5.19	1.53
GLU SER	5.48	-0.03	3.14	3.99	0.59
GLU THR	5.32	-0.06	4.02	3.99	3.17
GLU VAL	6.81	-0.03	4.52	3.99	1.35
GLU TRP	8.12	-0.17	4.65	3.99	0.59
GLU TYR	6.58	-0.17	4.78	3.99	2.40
PHE PHE	5.31	-0.83	3.07	3.99	0.63
PHE GLY	5.35	-0.04	3.61	5.00	1.06
PHE HIS	4.94	-0.13	3.32	4.56	1.19
PHE ILE	6.77	-0.68	4.70	6.11	0.91
PHE LYS	6.55	-0.04	3.82	3.99	0.66
PHE LEU	5.98	-0.99	4.18	3.99	1.84
PHE MET	5.61	-0.22	3.21	3.99	0.59
PHE ASN	6.54	-0.02	4.32	3.99	1.32
PHE PRO	5.89	-0.21	4.62	8.57	1.56
PHE GLN	5.46	-0.04	3.58	3.99	1.26
PHE ARG	5.27	-0.04	3.01	3.99	0.59
PHE SER	6.36	-0.04	3.64	3.99	0.59
PHE THR	5.17	-0.05	3.80	4.11	2.54
PHE VAL	4.61	-0.72	2.88	4.74	0.71
PHE TRP	6.44	-0.19	3.68	3.99	0.59
PHE TYR	5.80	-0.44	3.33	4.02	0.60
GLY GLY	3.31	-0.27	2.45	3.99	2.73
GLY HIS	5.03	-0.07	3.38	5.61	0.84
GLY ILE	4.28	-0.02	3.25	3.99	3.35
GLY LYS	4.21	-0.05	2.83	4.33	1.31
GLY LEU	4.03	-0.05	3.01	4.08	2.82
GLY MET	5.51	-0.05	3.80	3.99	1.72
GLY ASN	4.97	-0.07	3.35	3.99	1.48
GLY PRO	5.00	-0.00	4.34	11.12	4.06
GLY GLN	5.45	-0.04	3.26	3.99	0.75
GLY ARG	4.91	-0.02	2.81	3.99	0.59
GLY SER	4.73	-0.01	3.48	4.05	2.61

GLY THR	3.35	-0.03	2.73	5.31	4.51
GLY VAL	4.72	-0.05	3.60	3.99	3.39
GLY TRP	5.01	-0.34	2.88	4.02	0.60
GLY TYR	5.23	-0.03	4.07	7.44	1.80
HIS HIS	7.47	-0.19	5.05	3.99	1.51
HIS ILE	6.44	-0.04	4.43	3.99	1.68
HIS LYS	6.62	-0.03	5.04	6.96	1.62
HIS LEU	6.63	-0.06	4.16	3.99	0.97
HIS MET	5.95	-0.16	3.40	3.99	0.59
HIS ASN	7.16	-0.01	4.89	4.03	1.58
HIS PRO	5.56	-0.03	4.59	8.16	3.15
HIS GLN	7.91	-0.02	4.53	3.99	0.60
HIS ARG	7.78	-0.02	4.45	3.99	0.59
HIS SER	6.05	-0.04	4.30	3.99	2.07
HIS THR	5.01	-0.05	3.82	4.03	3.36
HIS VAL	5.68	-0.04	4.01	3.99	1.99
HIS TRP	7.08	-0.80	4.05	3.99	0.59
HIS TYR	6.97	-0.15	4.21	3.99	0.79
ILE ILE	4.60	-0.85	3.73	6.19	3.65
ILE LYS	5.98	-0.05	3.49	3.99	0.66
ILE LEU	5.13	-1.38	4.38	13.73	2.15
ILE MET	5.06	-0.83	4.15	6.90	3.62
ILE ASN	5.62	-0.05	4.69	5.99	5.09
ILE PRO	4.91	-0.05	3.74	3.99	3.39
ILE GLN	5.56	-0.01	4.50	6.72	3.12
ILE ARG	6.06	-0.06	3.47	4.00	0.60
ILE SER	6.04	-0.01	4.08	4.21	1.38
ILE THR	4.42	-0.07	3.37	4.34	3.08
ILE VAL	5.55	-0.58	4.49	10.57	1.58
ILE TRP	5.91	-0.61	3.38	3.99	0.59
ILE TYR	6.91	-0.39	3.95	3.99	0.59
LYS LYS	6.76	-0.04	4.63	5.92	0.88
LYS LEU	7.33	-0.04	4.73	5.08	0.76
LYS MET	6.72	-0.05	4.93	4.61	2.16
LYS ASN	5.29	-0.10	3.55	4.12	1.37
LYS PRO	5.39	-0.05	3.49	3.99	1.16
LYS GLN	5.03	-0.06	3.33	4.01	1.33
LYS ARG	8.10	-0.04	6.18	8.09	1.29
LYS SER	4.56	-0.05	3.47	3.99	3.36
LYS THR	6.76	-0.06	4.26	4.84	0.72
LYS VAL	5.61	-0.04	4.10	3.99	2.49
LYS TRP	7.36	-0.17	4.21	3.99	0.59
LYS TYR	6.65	-0.48	3.80	3.99	0.59
LEU LEU	5.91	-0.90	4.55	8.54	1.28
LEU MET	4.14	-0.71	3.16	3.99	3.39
LEU ASN	5.46	-0.08	3.12	3.99	0.59
LEU PRO	7.06	-0.05	6.07	7.13	6.06
LEU GLN	6.25	-0.04	5.17	7.86	3.38
LEU ARG	6.71	-0.03	4.65	3.99	1.75
LEU SER	4.21	-0.04	3.21	4.01	3.41
LEU THR	6.34	-0.08	4.55	3.99	2.19
LEU VAL	5.47	-1.38	4.38	4.84	4.12
LEU TRP	6.81	-0.86	4.27	4.78	0.71
LEU TYR	7.14	-0.27	4.09	3.99	0.59
MET MET	5.16	-0.37	3.50	3.99	1.54
MET ASN	5.82	-0.04	3.65	3.99	0.97
MET PRO	5.89	-0.14	4.21	3.99	2.13
MET GLN	7.81	-0.00	5.28	3.99	1.51
MET ARG	6.76	-0.06	3.88	4.02	0.60
MET SER	5.60	-0.05	4.33	4.22	3.56
MET THR	6.13	-0.05	3.92	3.99	1.08
MET VAL	6.25	-0.51	4.75	3.99	3.31
MET TRP	7.35	-0.44	4.20	3.99	0.59
MET TYR	6.37	-0.34	3.95	4.68	0.70
ASN ASN	6.24	-0.05	4.90	9.28	1.39
ASN PRO	4.93	-0.06	3.41	3.99	1.74
ASN GLN	7.28	-0.04	5.39	3.99	2.73
ASN ARG	6.77	-0.04	4.29	3.99	1.03
ASN SER	6.28	-0.06	5.01	4.76	4.05
ASN THR	3.83	-0.03	2.19	3.99	0.59
ASN VAL	5.95	-0.03	4.48	3.99	3.10
ASN TRP	6.35	-0.07	4.24	3.99	1.40
ASN TYR	6.81	-0.04	4.89	6.71	1.00
PRO PRO	4.96	-0.09	3.13	3.99	1.01
PRO GLN	4.50	-0.06	2.65	3.99	0.69
PRO ARG	6.96	-0.01	4.62	3.99	1.35
PRO SER	4.98	-0.08	3.63	4.40	2.19
PRO THR	6.01	-0.03	4.46	4.25	2.58
PRO VAL	6.04	-0.06	4.40	3.99	2.44
PRO TRP	4.74	-0.19	2.72	4.01	0.60
PRO TYR	5.64	-0.25	3.23	3.99	0.60
GLN GLN	6.69	-0.04	4.94	4.34	2.44
GLN ARG	6.01	-0.07	3.44	3.99	0.59
GLN SER	6.94	-0.04	5.74	11.77	1.76
GLN THR	5.10	-0.04	3.89	3.99	3.39
GLN VAL	5.64	-0.03	4.20	5.50	1.89
GLN TRP	7.10	-0.05	4.07	3.99	0.60
GLN TYR	6.88	-0.05	5.21	6.58	1.68
ARG ARG	7.44	-0.04	5.52	7.46	1.11
ARG SER	5.22	-0.08	3.80	3.99	2.46
ARG THR	6.83	-0.08	3.90	3.99	0.59
ARG VAL	6.74	-0.06	3.86	3.99	0.59
ARG TRP	7.64	-0.44	4.37	3.99	0.59
ARG TYR	7.25	-0.36	4.22	4.12	0.61
SER SER	4.62	-0.08	3.47	7.80	1.17
SER THR	5.10	-0.04	4.38	7.91	5.36
SER VAL	3.64	-0.04	2.76	3.99	3.26
SER TRP	5.20	-0.04	3.27	3.99	0.99
SER TYR	6.62	-0.08	4.57	3.99	1.72
THR THR	3.74	-0.06	2.81	3.99	3.05
THR VAL	5.84	-0.03	4.64	6.66	2.64
THR TRP	5.12	-0.03	2.93	3.99	0.59

THR TYR	6.21	-0.03	4.60	3.99	2.76
VAL VAL	4.80	-0.97	4.11	14.38	2.15
VAL TRP	7.52	-0.24	4.49	3.99	0.75
VAL TYR	6.86	-0.29	4.45	3.99	1.18
TRP TRP	5.95	-0.62	3.42	3.99	0.61
TRP TYR	5.41	-0.23	3.09	3.99	0.59
TYR TYR	6.56	-0.21	3.76	3.99	0.59
ALA N	3.96	-0.00	3.70	17.83	11.68
ALA CA	4.46	-0.00	3.59	6.08	3.37
ALA C	4.42	-0.00	3.77	8.27	4.65
ALA O	3.82	-0.00	3.48	15.99	6.68
CYS N	4.04	-0.00	3.72	18.00	7.29
CYS CA	5.86	-0.00	4.81	8.95	2.51
CYS C	4.22	-0.00	3.75	16.06	3.68
CYS O	3.78	-0.00	3.40	12.10	7.15
ASP N	4.22	-0.00	3.46	7.49	3.19
ASP CA	6.28	-0.01	5.39	10.21	3.89
ASP C	5.69	-0.00	5.09	12.58	6.04
ASP O	3.81	-0.00	3.36	11.61	5.04
GLU N	4.18	-0.00	3.66	9.21	6.12
GLU CA	6.95	-0.00	6.00	8.62	5.21
GLU C	5.02	-0.00	4.31	9.66	4.16
GLU O	5.36	-0.00	4.63	9.95	4.44
PHE N	5.20	-0.00	4.55	11.89	4.26
PHE CA	6.89	-0.00	5.99	11.09	4.31
PHE C	4.44	-0.00	3.84	13.05	3.11
PHE O	4.50	-0.00	4.08	16.33	5.86
GLY N	4.00	-0.00	3.35	7.03	4.39
GLY CA	4.76	-0.00	3.91	8.12	2.91
GLY C	3.72	-0.00	3.26	13.39	3.74
GLY O	3.35	-0.00	3.09	18.00	8.51
HIS N	4.71	-0.00	4.37	16.17	11.40
HIS CA	5.76	-0.00	4.95	12.52	2.92
HIS C	4.42	-0.00	4.07	14.62	9.84
HIS O	3.38	-0.00	2.77	7.57	3.21
ILE N	3.28	-0.00	2.97	13.50	7.17
ILE CA	5.07	-0.00	4.66	15.36	8.89
ILE C	4.68	-0.00	4.15	12.51	5.23
ILE O	5.62	-0.00	4.91	9.14	5.99
LYS N	6.01	-0.00	4.76	7.97	1.95
LYS CA	6.51	-0.00	5.73	10.62	5.60
LYS C	4.33	-0.00	4.00	15.78	9.84
LYS O	5.36	-0.00	4.79	11.63	6.49
LEU N	4.47	-0.00	3.62	6.22	3.50
LEU CA	5.07	-0.00	4.32	7.25	5.38
LEU C	5.20	-0.00	4.64	11.91	6.18
LEU O	3.86	-0.00	3.33	10.53	3.97
MET N	3.63	-0.00	3.27	15.78	5.41
MET CA	5.84	-0.00	5.31	14.58	7.33
MET C	5.37	-0.00	4.50	9.21	3.12
MET O	4.09	-0.00	3.71	13.85	7.33
ASN N	4.46	-0.00	3.96	11.50	5.82
ASN CA	5.08	-0.00	4.61	15.69	6.33
ASN C	4.86	-0.00	4.46	15.47	8.15
ASN O	3.90	-0.00	3.57	13.68	9.29
PRO N	3.25	-0.00	2.75	12.48	2.28
PRO CA	5.51	-0.00	5.01	16.31	6.06
PRO C	5.02	-0.00	4.49	12.18	6.43
PRO O	3.07	-0.00	2.59	12.96	1.94
GLN N	5.20	-0.00	4.54	11.57	4.34
GLN CA	4.12	-0.00	3.64	12.39	5.06
GLN C	3.52	-0.00	3.24	14.91	9.58
GLN O	5.38	-0.00	4.83	12.35	6.78
ARG N	5.42	-0.00	4.61	10.50	3.26
ARG CA	6.68	-0.00	5.45	7.42	3.07
ARG C	6.77	-0.00	5.79	7.55	5.37
ARG O	4.91	-0.00	4.23	7.99	5.61
SER N	4.66	-0.00	3.81	7.86	2.86
SER CA	3.53	-0.01	3.17	13.20	6.00
SER C	4.12	-0.00	3.57	11.07	4.03
SER O	4.96	-0.00	4.19	10.65	2.82
THR N	4.67	-0.00	4.38	16.78	14.26
THR CA	4.45	-0.00	3.73	8.87	3.36
THR C	5.13	-0.00	4.68	18.00	5.95
THR O	4.23	-0.00	3.06	6.46	1.18
VAL N	3.50	-0.00	3.04	10.01	4.79
VAL CA	5.58	-0.00	4.87	10.81	4.75
VAL C	5.88	-0.00	5.05	7.98	5.39
VAL O	3.84	-0.00	3.37	10.86	5.18
TRP N	4.77	-0.00	4.00	7.90	3.94
TRP CA	6.67	-0.00	5.81	9.01	5.72
TRP C	5.53	-0.00	4.75	12.18	2.98
TRP O	4.97	-0.00	4.51	15.31	6.68
TYR N	5.71	-0.00	5.06	13.33	4.70
TYR CA	5.06	-0.00	4.51	12.64	5.66
TYR C	6.80	-0.00	5.94	10.30	5.08
TYR O	5.53	-0.00	4.36	9.43	1.41
N N	3.40	-0.00	3.03	12.29	5.83
N CA	4.62	-0.00	4.31	16.00	12.2
N C	3.10	-0.00	2.52	6.98	3.19
N O	2.68	-0.00	2.31	13.21	2.85
CA CA	4.88	-0.00	4.23	9.72	4.91
CA C	3.88	-0.00	3.45	11.89	5.80
CA O	3.74	-0.00	2.82	5.24	2.25
C C	4.74	-0.00	4.21	12.20	5.46
C O	3.08	-0.00	2.82	13.37	9.94
O O	2.82	-0.00	2.49	11.70	4.94

Tableau A.1. sOPEP2 non bonded potential parameters.

TOP1				
Library - Potential	Native (/25)	Near-Native (/25)	Non-Native (/25)	Average CAD-CG (BC-WDC)
Lib1 - sOPEPv1	2	11	12	0.605 ± 0.038 (0.347 ± 0.448)
Lib1 - sOPEPv2	6	12	7	0.622 ± 0.033 (0.443 ± 0.380)
Lib2 - sOPEPv1	3	16	6	0.619 ± 0.036 (0.558 ± 0.314)
Lib2 - sOPEPv2	7	13	5	0.630 ± 0.035 (0.549 ± 0.325)
TOP5				
Library - Potential	Native (/125)	Near-Native (/125)	Non-Native (/125)	Average CAD-CG (BC-WDC)
Lib1 - sOPEPv1	15	50	60	0.603 ± 0.034 (0.334 ± 0.356)
Lib1 - sOPEPv2	21	64	40	0.618 ± 0.027 (0.385 ± 0.325)
Lib2 - sOPEPv1	18	66	41	0.614 ± 0.028 (0.408 ± 0.348)
Lib2 - sOPEPv2	30	58	37	0.623 ± 0.031 (0.482 ± 0.290)
Best in TOP5				
Library - Potential	Native (/25)	Near-Native (/25)	Non-Native (/25)	Average CAD-CG (BC-WDC)
Lib1 - sOPEPv1	6	13	6	0.626 ± 0.038 (0.599 ± 0.322)
Lib1 - sOPEPv2	9	14	2	0.641 ± 0.034 (0.613 ± 0.332)
Lib2 - sOPEPv1	8	14	3	0.639 ± 0.033 (0.620 ± 0.279)
Lib2 - sOPEPv2	9	15	1	0.646 ± 0.030 (0.669 ± 0.253)

Tableau A.2. Comparison of PEP-FOLD's predictions for the parametrization (G/IC) ensemble. The parametrization (G/IC) ensemble contains 25 protein targets. The table is split into three parts. From top to bottom, we show the results for the lowest energy prediction, the TOP5 lowest energy predictions and the best prediction inside the TOP5. The first column describes the library/potential pairing, the next three columns are the number of decoys generated in the Native, Near-Native and Non-Native class respectively and the last column shows the average CAD-CG score and in parenthesis, the BC score computed on the WDC defined on the *Protein Data Bank*.

TOP1				
Library - Potential	Native (/40)	Near-Native (/40)	Non-Native (/40)	Average CAD-CG (BC-WDC)
Lib1 - sOPEPv1	0	18	22	0.589 ± 0.048 (0.330 ± 0.346)
Lib1 - sOPEPv2	9	14	17	0.609 ± 0.059 (0.441 ± 0.395)
Lib2 - sOPEPv1	4	20	16	0.600 ± 0.050 (0.373 ± 0.379)
Lib2 - sOPEPv2	13	14	13	0.620 ± 0.049 (0.502 ± 0.380)
TOP5				
Library - Potential	Native (/200)	Near-Native (/200)	Non-Native (/200)	Average CAD-CG (BC-WDC)
Lib1 - sOPEPv1	20	77	103	0.591 ± 0.051 (0.327 ± 0.290)
Lib1 - sOPEPv2	43	64	93	0.607 ± 0.053 (0.435 ± 0.351)
Lib2 - sOPEPv1	37	66	97	0.601 ± 0.051 (0.367 ± 0.313)
Lib2 - sOPEPv2	56	63	81	0.615 ± 0.047 (0.445 ± 0.345)
Best in TOP5				
Library - Potential	Native (/40)	Near-Native (/40)	Non-Native (/40)	Average CAD-CG (BC-WDC)
Lib1 - sOPEPv1	11	12	17	0.613 ± 0.056 (0.601 ± 0.274)
Lib1 - sOPEPv2	13	15	12	0.630 ± 0.052 (0.647 ± 0.340)
Lib2 - sOPEPv1	14	11	15	0.627 ± 0.049 (0.590 ± 0.325)
Lib2 - sOPEPv2	16	14	10	0.639 ± 0.048 (0.652 ± 0.317)

Tableau A.3. Comparison of PEP-FOLD's predictions for the validation G/IC ensemble. The validation G/IC ensemble contains 40 protein targets. The table is split into three parts. From top to bottom, we show the results for the lowest energy prediction, the TOP5 lowest energy predictions and the best prediction inside the TOP5. The first column describe the library/potential pairing, the next three columns are the number of decoys generated in the Native, Near-Native and Non-Native class respectively and the last column shows the average CAD-CG score and in parenthesis, the BC score computed on the WDC defined on the *Protein Data Bank*.

TOP1				
Library - Potential	Native (/50)	Near-Native (/50)	Non-Native (/50)	Average CAD-CG (BC-WDC)
Lib1 - sOPEPv1	44	6	0	0.693 ± 0.045 (0.624 ± 0.400)
Lib1 - sOPEPv2	41	7	2	0.686 ± 0.048 (0.628 ± 0.426)
Lib2 - sOPEPv1	42	8	0	0.694 ± 0.049 (0.692 ± 0.327)
Lib2 - sOPEPv2	44	5	1	0.701 ± 0.049 (0.707 ± 0.350)
TOP5				
Library - Potential	Native (/250)	Near-Native (/250)	Non-Native (/250)	Average CAD-CG (BC-WDC)
Lib1 - sOPEPv1	191	52	7	0.686 ± 0.050 (0.626 ± 0.361)
Lib1 - sOPEPv2	198	42	10	0.688 ± 0.049 (0.670 ± 0.317)
Lib2 - sOPEPv1	205	40	5	0.693 ± 0.049 (0.677 ± 0.319)
Lib2 - sOPEPv2	208	36	6	0.699 ± 0.049 (0.721 ± 0.275)
Best in TOP5				
Library - Potential	Native (/48)	Near-Native (/48)	Non-Native (/48)	Average CAD-CG (BC-WDC)
Lib1 - sOPEPv1	46	4	0	0.709 ± 0.048 (0.810 ± 0.283)
Lib1 - sOPEPv2	44	6	0	0.707 ± 0.049 (0.789 ± 0.298)
Lib2 - sOPEPv1	46	4	0	0.708 ± 0.047 (0.828 ± 0.242)
Lib2 - sOPEPv2	45	5	0	0.716 ± 0.047 (0.827 ± 0.193)

Tableau A.4. Comparison of PEP-FOLD's predictions for the validation G/CC ensemble. The validation G/CC ensemble contains 50 protein targets. The table is split into three parts. From top to bottom, we show the results for the lowest energy prediction, the TOP5 lowest energy predictions and the best prediction inside the TOP5. The first column describe the library/potential pairing, the next three columns are the number of decoys generated in the Native, Near-Native and Non-Native class respectively and the last column shows the average CAD-CG score and in parenthesis, the BC score computed on the WDC defined on the *Protein Data Bank*.

0-50 amino acids				
Library - Potential	Native (/65)	Near-Native (/65)	Non-Native (/65)	Average CAD-CG (BC-WDC)
Lib1 - sOPEPv1	37	13	15	0.656 ± 0.062 (0.519 ± 0.418)
Lib1 - sOPEPv2	39	15	11	0.658 ± 0.065 (0.560 ± 0.426)
Lib2 - sOPEPv1	37	19	9	0.662 ± 0.066 (0.601 ± 0.391)
Lib2 - sOPEPv2	47	9	9	0.675 ± 0.063 (0.596 ± 0.406)
51-70 amino acids				
Library - Potential	Native (/50)	Near-Native (/50)	Non-Native (/50)	Average CAD-CG (BC-WDC)
Lib1 - sOPEPv1	9	22	19	0.614 ± 0.065 (0.373 ± 0.402)
Lib1 - sOPEPv2	17	18	15	0.629 ± 0.052 (0.475 ± 0.397)
Lib2 - sOPEPv1	12	25	13	0.624 ± 0.054 (0.488 ± 0.332)
Lib2 - sOPEPv2	17	23	10	0.635 ± 0.049 (0.608 ± 0.312)

Tableau A.5. Comparison of PEP-FOLD's lowest energy predictions by peptide's length. The table is split into two parts. From top to bottom, we show the results for peptides of length 0 to 50 amino acids and of length 51 to 70 amino acids respectively. The first column describe the library/potential pairing, the next three columns are the number of decoys generated in the Native, Near-Native and Non-Native class respectively and the last column shows the average CAD-CG score and in parenthesis, the BC score computed on the WDC defined on the *Protein Data Bank*.

Alpha				
Library - Potential	Native (/60)	Near-Native (/60)	Non-Native (/60)	Average CAD-CG (BC-WDC)
Lib1 - sOPEPv1	28	21	9	0.660 ± 0.061 (0.407 ± 0.452)
Lib1 - sOPEPv2	37	16	5	0.669 ± 0.056 (0.461 ± 0.458)
Lib2 - sOPEPv1	31	18	9	0.665 ± 0.063 (0.497 ± 0.427)
Lib2 - sOPEPv2	37	7	6	0.676 ± 0.059 (0.585 ± 0.403)
Beta				
Library - Potential	Native (/32)	Near-Native (/32)	Non-Native (/32)	Average CAD-CG (BC-WDC)
Lib1 - sOPEPv1	12	8	12	0.623 ± 0.061 (0.635 ± 0.305)
Lib1 - sOPEPv2	13	10	9	0.628 ± 0.056 (0.671 ± 0.307)
Lib2 - sOPEPv1	13	12	7	0.633 ± 0.058 (0.714 ± 0.235)
Lib2 - sOPEPv2	18	8	6	0.647 ± 0.055 (0.667 ± 0.285)
Alpha/Beta				
Library - Potential	Native (/21)	Near-Native (/21)	Non-Native (/21)	Average CAD-CG (BC-WDC)
Lib1 - sOPEPv1	4	5	12	0.597 ± 0.064 (0.362 ± 0.391)
Lib1 - sOPEPv2	4	6	11	0.607 ± 0.052 (0.490 ± 0.379)
Lib2 - sOPEPv1	3	13	5	0.612 ± 0.050 (0.496 ± 0.280)
Lib2 - sOPEPv2	6	9	6	0.625 ± 0.050 (0.576 ± 0.328)

Tableau A.6. Comparison of PEP-FOLD's lowest energy predictions by peptide's structural class. The table is split into three parts. From top to bottom, we show the results for α-targets, β-targets and α/β-targets respectively. The first column describe the library/potential pairing, the next three columns are the number of decoys generated in the Native, Near-Native and Non-Native class respectively and the last column shows the average CAD-CG score and in parenthesis, the BC score computed on the WDC defined on the *Protein Data Bank*.

Parametrization G/IC					
	sOPEPv1 - Lib1	sOPEPv2 - Lib2	APPTest	RaptorX	AlphaFold2
1b03 (β) (18)	0.567 (0.280)	0.621 (0.553)	0.507 (-0.513)	-	0.606 (0.579)
2lma (α) (22)	0.555 (0.101)	0.642 (0.051)	0.772 (0.946)	-	0.706 (0.158)
1e0n (β) (27)	0.691 (0.918)	0.666 (0.625)	0.549 (0.443)	0.508 (-0.313)	0.781 (0.980)
2mwf (β) (32)	0.575 (0.803)	0.671 (0.881)	0.609 (0.930)	0.500 (0.610)	0.740 (0.944)
1yiu (β) (37)	0.687 (0.946)	0.701 (0.873)	0.631 (0.913)	0.482 (-0.640)	0.847 (0.986)
1bhi (α/β) (38)	0.557 (-0.362)	0.552 (0.383)	0.547 (0.833)	0.536 (-0.281)	0.741 (0.956)
1i6c (β) (39)	0.589 (0.683)	0.608 (0.741)	0.535 (0.859)	0.464 (-0.350)	0.676 (0.815)
1spw (β) (39)	0.590 (0.232)	0.683 (0.841)	0.619 (0.908)	0.533 (0.769)	0.780 (0.981)
2ysb (β) (49)	0.642 (0.799)	0.658 (0.823)	-	0.644 (0.884)	0.824 (0.995)
1jjs (α) (50)	0.574 (-0.158)	0.572 (0.028)	-	0.580 (-0.028)	0.573 (0.009)
2l92 (β) (50)	0.579 (0.016)	0.631 (0.314)	-	0.622 (0.845)	0.824 (0.977)
2l93 (α/β) (55)	0.607 (0.586)	0.611 (0.743)	-	0.622 (0.866)	0.772 (0.946)
1fex (α) (59)	0.622 (0.623)	0.609 (0.340)	-	0.702 (0.944)	0.770 (0.979)
1zxg (α) (59)	0.584 (0.758)	0.599 (0.836)	-	0.656 (0.947)	0.688 (0.944)
2b7e (α) (59)	0.619 (-0.604)	0.631 (-0.318)	-	0.721 (0.972)	0.816 (0.989)
1g2h (α) (61)	0.646 (0.447)	0.580 (0.140)	-	0.515 (0.562)	0.661 (0.538)
2dt6 (α) (64)	0.630 (0.231)	0.650 (0.604)	-	0.778 (0.969)	0.859 (0.991)
1uxd (α) (65)	0.620 (-0.458)	0.635 (0.817)	-	0.599 (0.890)	0.775 (0.967)
2fmr (α/β) (65)	0.562 (0.392)	0.680 (0.939)	-	0.746 (0.962)	0.777 (0.988)
1z4h (α/β) (66)	0.577 (-0.038)	0.584 (0.260)	-	0.429 (-0.598)	0.696 (0.923)
1zv6 (α) (68)	0.545 (-0.159)	0.625 (0.516)	-	0.675 (0.848)	0.758 (0.976)
1cpz (α/β) (68)	0.622 (0.908)	0.647 (0.908)	-	0.478 (0.862)	0.823 (0.986)
2bby (α/β) (69)	0.643 (0.647)	0.628 (0.764)	-	0.749 (0.948)	0.769 (0.952)
1go5 (α) (69)	0.634 (0.311)	0.638 (0.372)	-	0.710 (0.920)	0.729 (0.925)
1wcn (α) (70)	0.607 (0.766)	0.631 (0.702)	-	0.769 (0.981)	0.855 (0.997)

Tableau A.7. Comparison for targets in the parametrization G/IC set - TOP1. Each column show the CAD-CG (and BC-WDC in parenthesis). For each target, we recall its secondary structure class and size in amino-acids. The next two columns show the results of the lowest energy prediction (TOP1) PEP-FOLDv1 and PEP-FOLDv2 respectively. The next three column show the results using APPTest, RaptorX and AlphaFold2 respectively. Color coding: CAD-CG scores corresponding to the native, near-native and non-native classification are shown respectively in green, yellow and red.

Validation G/IC					
	sOPEPv1 - Lib1	sOPEPv2 - Lib2	APPTest	RaptorX	AlphaFold2
1s4j (Coil) (13)	0.553 (-0.061)	0.523 (-0.270)	0.532 (0.010)	-	0.551 (-0.013)
5ykl (α) (19)	0.649 (-0.003)	0.651 (0.039)	0.691 (0.662)	-	0.661 (0.112)
2jof (α) (20)	0.639 (0.294)	0.680 (0.156)	0.703 (0.709)	-	0.879 (0.957)
2msu (α) (20)	0.572 (0.128)	0.607 (0.036)	0.663 (0.886)	-	0.703 (0.917)
2oru (β) (20)	0.604 (0.509)	0.605 (0.198)	0.657 (0.896)	-	0.704 (0.683)
5y22 (α) (22)	0.578 (0.030)	0.590 (-0.050)	0.770 (0.985)	-	0.806 (0.988)
1rij (α) (23)	0.637 (0.007)	0.649 (0.295)	0.743 (0.780)	-	0.775 (0.943)
2bn6 (α) (33)	0.585 (0.027)	0.685 (0.865)	0.698 (0.890)	0.691 (0.878)	0.734 (0.900)
2kya (α) (34)	0.591 (0.006)	0.565 (0.025)	0.694 (0.897)	0.588 (-0.318)	0.568 (-0.271)
1e0m (β) (37)	0.614 (0.779)	0.635 (0.839)	0.443 (0.900)	0.443 (-0.579)	0.781 (0.976)
1k91 (β) (37)	0.570 (0.282)	0.541 (0.096)	0.538 (-0.205)	0.551 (-0.319)	0.808 (0.829)
2ysh (β) (40)	0.646 (0.839)	0.654 (0.819)	0.537 (0.783)	0.506 (-0.623)	0.722 (0.958)
1ed7 (β) (45)	0.491 (0.351)	0.527 (0.085)	-	0.485 (0.596)	0.729 (0.965)
2lrx (α/β) (45)	0.598 (0.553)	0.599 (-0.164)	-	0.586 (0.566)	0.680 (0.732)
2l4j (β) (46)	0.607 (0.787)	0.594 (0.747)	-	0.545 (0.784)	0.704 (0.981)
2m6o (β) (48)	0.580 (0.859)	0.588 (0.937)	-	0.481 (0.506)	0.682 (0.970)
1ify (α) (49)	0.638 (-0.363)	0.688 (0.850)	-	0.735 (0.981)	0.818 (0.986)
2zaj (β) (49)	0.635 (0.784)	0.676 (0.826)	-	0.601 (-0.642)	0.791 (0.989)
1zrj (α) (50)	0.639 (0.782)	0.687 (0.843)	-	0.696 (0.956)	0.863 (0.997)
1k8b (α/β) (52)	0.499 (0.485)	0.607 (0.660)	-	0.636 (0.757)	0.641 (0.777)
2v0e (α/β) (55)	0.570 (0.070)	0.604 (0.849)	-	0.737 (0.968)	0.780 (0.975)
1n87 (α/β) (56)	0.507 (0.146)	0.568 (0.195)	-	0.550 (0.764)	0.760 (0.979)
1pgb (α/β) (56)	0.551 (0.452)	0.664 (0.898)	-	0.753 (0.961)	0.907 (0.997)
2mdj (β) (56)	0.556 (0.444)	0.575 (0.266)	-	0.618 (0.312)	0.611 (0.724)
1zwv (α) (58)	0.594 (0.335)	0.668 (0.877)	-	0.627 (0.824)	0.711 (0.841)
2jtm (β) (60)	0.540 (-0.008)	0.602 (0.715)	-	0.751 (0.926)	0.789 (0.954)
2k57 (β) (61)	0.579 (0.304)	0.631 (0.204)	-	0.754 (0.980)	0.835 (0.994)
1gyz (α) (62)	0.604 (0.783)	0.622 (0.765)	-	0.680 (0.960)	0.742 (0.974)
2mi6 (β) (62)	0.476 (-0.026)	0.535 (0.008)	-	0.583 (0.852)	0.828 (0.983)
2kac (α/β) (64)	0.634 (0.698)	0.606 (0.640)	-	0.557 (0.443)	0.853 (0.988)
1f0z (α/β) (66)	0.489 (0.528)	0.534 (0.050)	-	0.498 (0.640)	0.772 (0.985)
2bzt (α) (66)	0.611 (0.052)	0.628 (0.713)	-	0.788 (0.986)	0.837 (0.990)
2a63 (α/β) (66)	0.643 (-0.540)	0.692 (0.945)	-	0.564 (0.489)	0.798 (0.991)
4c26 (α/β) (66)	0.568 (0.117)	0.595 (0.849)	-	0.602 (0.946)	0.720 (0.975)
1u97 (α) (69)	0.639 (0.566)	0.626 (0.689)	-	0.691 (0.969)	0.713 (0.974)
2kt2 (α/β) (69)	0.585 (0.764)	0.607 (0.789)	-	0.504 (0.875)	0.764 (0.975)
2l4m (α/β) (69)	0.546 (0.314)	0.604 (0.195)	-	0.758 (0.988)	0.855 (0.996)
1q1v (α) (70)	0.639 (0.613)	0.697 (0.904)	-	0.718 (0.950)	0.768 (0.970)
2coo (α) (70)	0.646 (0.539)	0.670 (0.761)	-	0.731 (0.970)	0.826 (0.996)
2k2a (α) (70)	0.632 (-0.043)	0.685 (0.924)	-	0.622 (0.869)	0.756 (0.941)

Tableau A.8. Comparison for targets in the validation G/IC set - TOP1. Notations and color coding are similar to that of Table A.7

Validation G/CC					
	sOPEPv1 - Lib1	sOPEPv2 - Lib2	APPTest	RaptorX	AlphaFold2
1uao (β) (10)	0.697 (0.770)	0.722 (0.594)	0.694 (0.158)	-	0.878 (0.973)
2e4e (Coil) (10)	0.715 (0.767)	0.705 (0.851)	0.703 (0.854)	-	0.814 (0.959)
2evq (β) (12)	0.719 (0.849)	0.738 (0.876)	0.858 (0.980)	-	0.911 (0.997)
2luf (α) (20)	0.654 (0.917)	0.643 (0.918)	0.625 (0.952)	-	0.666 (0.892)
2n16 (α) (20)	0.633 (0.679)	0.656 (0.116)	0.688 (0.922)	-	0.700 (0.782)
5t7q (α) (21)	0.693 (0.968)	0.668 (0.980)	0.746 (0.954)	-	0.711 (0.934)
6g4v (α) (24)	0.732 (0.472)	0.756 (0.612)	0.744 (0.307)	-	0.782 (0.529)
6r2x (α) (25)	0.727 (0.961)	0.758 (0.958)	0.770 (0.967)	-	0.779 (0.983)
1by0 (α) (27)	0.764 (0.856)	0.783 (0.828)	0.777 (0.845)	0.766 (0.844)	0.774 (0.810)
1yyb (α) (27)	0.816 (0.860)	0.804 (0.801)	0.810 (0.843)	0.814 (0.815)	0.812 (0.832)
2k76 (α/β) (30)	0.744 (-0.080)	0.752 (0.315)	0.712 (0.306)	0.788 (0.908)	0.837 (0.905)
2l0g (α) (32)	0.732 (0.022)	0.756 (0.682)	0.709 (0.605)	0.781 (0.898)	0.822 (0.971)
1wr3 (β) (36)	0.645 (0.837)	0.664 (0.808)	0.666 (0.938)	0.465 (-0.479)	0.753 (0.979)
1wr4 (β) (36)	0.631 (0.859)	0.671 (0.879)	0.587 (0.891)	0.486 (-0.648)	0.799 (0.976)
2ki0 (α/β) (36)	0.685 (0.811)	0.672 (0.669)	0.589 (0.324)	0.704 (0.785)	0.721 (0.784)
1e0l (β) (37)	0.652 (0.311)	0.681 (0.802)	0.629 (0.929)	0.520 (0.668)	0.805 (0.990)
1jrj (β) (39)	0.742 (0.970)	0.730 (0.935)	0.724 (0.970)	0.756 (0.878)	0.861 (0.996)
1bwx (α) (39)	0.685 (0.716)	0.697 (0.710)	0.676 (0.852)	0.658 (0.407)	0.696 (0.669)
2ysc (β) (39)	0.666 (0.852)	0.671 (0.644)	0.548 (0.677)	0.735 (0.880)	0.821 (0.982)
2ysf (β) (40)	0.677 (0.947)	0.672 (0.895)	0.571 (0.779)	0.576 (0.551)	0.803 (0.991)
2ysg (β) (40)	0.716 (0.904)	0.730 (0.868)	0.607 (-0.560)	0.568 (-0.442)	0.797 (0.977)
2ysi (β) (40)	0.688 (0.930)	0.695 (0.931)	0.655 (0.855)	0.583 (-0.448)	0.807 (0.988)
1k1v (α) (41)	0.683 (0.252)	0.785 (0.913)	-	0.785 (0.953)	0.820 (0.982)
1wr7 (β) (41)	0.654 (0.780)	0.693 (0.927)	-	0.549 (-0.552)	0.792 (0.985)
2dmv (β) (43)	0.703 (0.943)	0.726 (0.942)	-	0.553 (-0.402)	0.807 (0.983)
2m8j (β) (43)	0.653 (0.776)	0.634 (0.794)	-	0.503 (-0.364)	0.801 (0.983)
2p81 (α) (44)	0.706 (0.978)	0.724 (0.778)	-	0.461 (-0.605)	0.715 (0.981)
2k9d (α) (44)	0.725 (0.949)	0.708 (0.933)	-	0.650 (0.828)	0.765 (0.963)
1f4i (α) (45)	0.652 (0.727)	0.679 (-0.410)	-	0.761 (0.981)	0.815 (0.989)
1p9c (α) (45)	0.680 (0.572)	0.674 (0.613)	-	0.660 (0.721)	0.715 (0.745)
1pv0 (α) (46)	0.718 (-0.106)	0.762 (0.953)	-	0.794 (0.986)	0.845 (0.992)
2e5t (α) (46)	0.780 (0.227)	0.777 (0.923)	-	0.585 (0.528)	0.780 (0.933)
2jnh (α) (46)	0.659 (0.725)	0.680 (-0.242)	-	0.672 (0.649)	0.795 (0.979)
1dv0 (α) (47)	0.653 (-0.413)	0.661 (0.814)	-	0.713 (0.973)	0.804 (0.991)
1pgy (α) (47)	0.677 (0.837)	0.689 (0.822)	-	0.718 (0.936)	0.726 (0.933)
2ekk (α) (47)	0.693 (0.696)	0.744 (0.929)	-	0.583 (0.881)	0.859 (0.996)
2wxc (α) (47)	0.667 (0.769)	0.742 (0.966)	-	0.582 (0.903)	0.780 (0.989)
2j8p (α) (49)	0.689 (-0.670)	0.681 (-0.688)	-	0.697 (-0.693)	0.706 (-0.662)
2wqg (α) (51)	0.703 (0.940)	0.625 (0.839)	-	0.652 (0.944)	0.761 (0.992)
1r4g (α) (53)	0.771 (0.978)	0.684 (0.762)	-	0.719 (-0.665)	0.807 (0.988)
1rzs (α) (61)	0.632 (0.476)	0.597 (0.227)	-	0.502 (0.614)	0.845 (0.989)
1rq6 (α) (62)	0.640 (0.272)	0.613 (0.732)	-	0.718 (0.953)	0.745 (0.959)
1wji (α) (63)	0.678 (0.671)	0.707 (0.917)	-	0.790 (0.990)	0.817 (0.995)
2l54 (α/β) (63)	0.660 (0.234)	0.679 (0.905)	-	0.732 (0.953)	0.744 (0.956)
2cp9 (α) (64)	0.636 (-0.361)	0.689 (0.719)	-	0.722 (0.974)	0.836 (0.997)
1du6 (α) (64)	0.689 (0.871)	0.696 (0.951)	-	0.699 (0.926)	0.756 (0.957)
1cok (α) (68)	0.656 (0.926)	0.667 (0.821)	-	0.728 (0.964)	0.746 (0.967)
1qpm (α/β) (69)	0.686 (0.925)	0.631 (0.300)	-	0.740 (0.927)	0.747 (0.975)
2kz9 (α) (69)	0.837 (0.758)	0.821 (0.722)	-	0.823 (0.636)	0.835 (0.644)
2fce (α) (70)	0.670 (0.294)	0.666 (0.805)	-	0.611 (0.805)	0.795 (0.975)

Tableau A.9. Comparison for targets in the validation G/CC set - TOP1. Notations and color coding are similar to that of Table A.7

Validation IG					
	sOPEPv1 - Lib1	sOPEPv2 - Lib2	APPTest	RaptorX	AlphaFold2
2mdu (β) (29)	0.425 (-0.052)	0.491 (0.182)	0.592 (0.869)	0.542 (0.722)	0.726 (0.922)
2gdl (α) (31)	0.447 (-0.005)	0.473 (0.007)	0.534 (0.138)	0.451 (0.065)	0.594 (0.091)
1vpu (α) (45)	0.509 (0.384)	0.539 (0.560)	-	0.544 (-0.396)	0.527 (-0.579)
1nd9 (α/β) (49)	0.534 (0.663)	0.544 (0.686)	-	0.553 (0.677)	0.562 (0.802)
2lhc (α) (56)	0.474 (-0.223)	0.477 (-0.420)	-	0.458 (-0.464)	0.727 (0.964)
2m4y (β) (56)	0.503 (0.112)	0.528 (0.615)	-	0.561 (0.904)	0.668 (0.859)
1gyf (α/β) (62)	0.489 (0.248)	0.515 (-0.089)	-	0.707 (0.898)	0.778 (0.985)
2cw1 (α/β) (65)	0.521 (-0.375)	0.548 (-0.092)	-	0.677 (0.897)	0.750 (0.964)
2jtv (α/β) (65)	0.485 (0.357)	0.545 (-0.122)	-	0.754 (0.987)	0.803 (0.992)
1qxf (β) (66)	0.526 (0.501)	0.498 (0.356)	-	0.707 (0.967)	0.774 (0.983)
2l8d (β) (66)	0.572 (0.512)	0.505 (-0.056)	-	0.734 (0.983)	0.790 (0.993)
2jrr (β) (67)	0.456 (-0.103)	0.502 (-0.053)	-	0.581 (0.820)	0.755 (0.991)
2kaf (α/β) (67)	0.524 (0.127)	0.475 (0.137)	-	0.516 (0.367)	0.851 (0.993)
2m2l (α/β) (67)	0.527 (-0.233)	0.574 (-0.414)	-	0.601 (0.904)	0.743 (0.988)
1ne3 (β) (68)	0.465 (-0.048)	0.493 (0.174)	-	0.695 (0.970)	0.736 (0.977)
1y2y (β) (68)	0.466 (-0.427)	0.493 (0.143)	-	0.484 (-0.201)	0.494 (-0.013)
2do3 (β) (69)	0.557 (0.566)	0.500 (-0.365)	-	0.742 (0.983)	0.823 (0.998)
2dy8 (α/β) (69)	0.536 (-0.127)	0.570 (0.528)	-	0.521 (0.861)	0.761 (0.993)
2eqi (β) (69)	0.455 (0.298)	0.470 (-0.121)	-	0.496 (0.780)	0.825 (0.995)
2mck (α) (69)	0.529 (0.351)	0.524 (0.396)	-	0.477 (0.516)	0.839 (0.992)
2xk0 (β) (69)	0.533 (0.398)	0.547 (0.351)	-	0.739 (0.989)	0.784 (0.992)
2lss (α/β) (70)	0.471 (-0.121)	0.512 (0.548)	-	0.572 (0.861)	0.739 (0.974)
2m7o (α/β) (70)	0.553 (0.498)	0.585 (-0.099)	-	0.693 (0.889)	0.750 (0.940)

Tableau A.10. Comparison for targets in the validation NG set - TOP1. Notations and color coding are similar to that of Table A.7

Parametrization G/IC					
	sOPEPv1 - Lib1	sOPEPv2 - Lib2	APPTest	RaptorX	AlphaFold2
1b03 (β) (18)	0.592 (0.496)	0.622 (0.553)	0.591 (-0.727)	-	0.606 (0.579)
2lma (α) (22)	0.604 (0.031)	0.659 (0.047)	0.791 (0.969)	-	0.706 (0.158)
1e0n (β) (27)	0.707 (0.897)	0.699 (0.694)	0.593 (0.694)	0.522 (-0.086)	0.792 (0.975)
2mwf (β) (32)	0.664 (0.893)	0.676 (0.860)	0.671 (0.943)	0.505 (0.211)	0.743 (0.950)
1yiu (β) (37)	0.687 (0.946)	0.701 (0.873)	0.631 (0.913)	0.482 (-0.640)	0.847 (0.986)
1bhi (α/β) (38)	0.558 (-0.091)	0.634 (0.734)	0.607 (0.830)	0.536 (-0.281)	0.758 (0.971)
1i6c (β) (39)	0.645 (0.776)	0.609 (0.741)	0.561 (0.697)	0.507 (-0.628)	0.688 (0.816)
1spw (β) (39)	0.653 (0.902)	0.694 (0.927)	0.645 (0.926)	0.533 (0.769)	0.789 (0.984)
2ysb (β) (49)	0.689 (0.861)	0.698 (0.882)	-	0.644 (0.884)	0.834 (0.997)
1jjs (α) (50)	0.574 (-0.183)	0.588 (-0.129)	-	0.580 (-0.028)	0.577 (0.009)
2l92 (β) (50)	0.634 (0.845)	0.647 (0.502)	-	0.622 (0.845)	0.824 (0.977)
2l93 (α/β) (55)	0.610 (0.430)	0.619 (0.714)	-	0.622 (0.866)	0.779 (0.945)
1fex (α) (59)	0.622 (0.623)	0.635 (0.860)	-	0.702 (0.944)	0.777 (0.979)
1zxg (α) (59)	0.613 (0.882)	0.638 (0.857)	-	0.656 (0.947)	0.695 (0.946)
2b7e (α) (59)	0.621 (-0.298)	0.631 (-0.318)	-	0.721 (0.972)	0.817 (0.990)
1g2h (α) (61)	0.646 (0.447)	0.628 (0.423)	-	0.515 (0.562)	0.664 (0.569)
2dt6 (α) (64)	0.637 (-0.146)	0.666 (0.718)	-	0.790 (0.972)	0.859 (0.991)
1uxd (α) (65)	0.662 (0.905)	0.656 (0.852)	-	0.599 (0.890)	0.776 (0.975)
2fmr (α/β) (65)	0.572 (0.043)	0.680 (0.939)	-	0.753 (0.966)	0.781 (0.988)
1z4h (α/β) (66)	0.589 (-0.183)	0.602 (0.212)	-	0.473 (0.631)	0.704 (0.930)
1cpz (α/β) (68)	0.622 (0.908)	0.647 (0.908)	-	0.478 (0.862)	0.823 (0.986)
1zv6 (α) (68)	0.564 (-0.289)	0.630 (0.267)	-	0.675 (0.848)	0.768 (0.978)
1go5 (α) (69)	0.634 (0.253)	0.638 (0.372)	-	0.722 (0.921)	0.730 (0.937)
2bby (α/β) (69)	0.643 (0.647)	0.628 (0.764)	-	0.749 (0.948)	0.778 (0.956)
1wcn (α) (70)	0.619 (0.748)	0.631 (0.702)	-	0.769 (0.981)	0.860 (0.997)

Tableau A.11. Comparison for targets in the parametrization G/IC set - Best in TOP5 (five lowest energy models). Notations and color coding are similar to that of Table A.7

Validation G/IC					
	sOPEPv1 - Lib1	sOPEPv2 - Lib2	APPTest	RaptorX	AlphaFold2
1s4j (Coil) (13)	0.553 (-0.061)	0.548 (-0.133)	0.546 (0.002)	-	0.551 (-0.013)
5ykl (α) (19)	0.659 (0.044)	0.670 (0.035)	0.697 (0.714)	-	0.678 (0.187)
2jof (α) (20)	0.715 (0.492)	0.681 (0.156)	0.736 (0.819)	-	0.879 (0.957)
2msu (α) (20)	0.585 (0.170)	0.624 (-0.004)	0.685 (0.854)	-	0.708 (0.913)
2oru (β) (20)	0.604 (0.509)	0.635 (0.318)	0.756 (0.968)	-	0.732 (0.771)
5y22 (α) (22)	0.734 (0.958)	0.590 (-0.050)	0.787 (0.988)	-	0.819 (0.987)
1rij (α) (23)	0.674 (0.593)	0.687 (0.533)	0.743 (0.780)	-	0.787 (0.951)
2bn6 (α) (33)	0.600 (0.007)	0.701 (0.878)	0.698 (0.890)	0.691 (0.878)	0.736 (0.905)
2kya (α) (34)	0.591 (0.006)	0.576 (0.151)	0.757 (0.932)	0.588 (-0.224)	0.601 (-0.202)
1e0m (β) (37)	0.643 (0.891)	0.674 (0.857)	0.443 (0.900)	0.472 (0.330)	0.787 (0.978)
1k91 (β) (37)	0.570 (0.282)	0.590 (0.105)	0.597 (0.388)	0.582 (-0.089)	0.809 (0.824)
2ysh (β) (40)	0.647 (0.839)	0.665 (0.882)	0.609 (0.905)	0.529 (-0.698)	0.729 (0.958)
1ed7 (β) (45)	0.493 (0.348)	0.581 (0.729)	-	0.516 (0.641)	0.743 (0.967)
2lrx (α/β) (45)	0.613 (0.468)	0.604 (0.060)	-	0.589 (0.245)	0.682 (0.739)
2l4j (β) (46)	0.619 (0.913)	0.636 (0.863)	-	0.588 (0.818)	0.708 (0.978)
2m6o (β) (48)	0.580 (0.859)	0.593 (0.904)	-	0.512 (-0.496)	0.694 (0.965)
1ify (α) (49)	0.690 (0.912)	0.707 (0.764)	-	0.749 (0.977)	0.821 (0.990)
2zaj (β) (49)	0.657 (0.824)	0.696 (0.807)	-	0.612 (0.803)	0.800 (0.989)
1zrj (α) (50)	0.682 (0.812)	0.695 (0.851)	-	0.708 (0.967)	0.863 (0.997)
1k8b (α/β) (52)	0.526 (-0.023)	0.608 (0.660)	-	0.636 (0.757)	0.644 (0.745)
2v0e (α/β) (55)	0.572 (0.244)	0.612 (0.849)	-	0.737 (0.968)	0.786 (0.977)
1n87 (α/β) (56)	0.529 (-0.249)	0.568 (0.195)	-	0.550 (0.764)	0.765 (0.980)
1pgb (α/β) (56)	0.580 (0.762)	0.664 (0.898)	-	0.763 (0.978)	0.909 (0.998)
2mdj (β) (56)	0.570 (0.349)	0.586 (-0.497)	-	0.624 (-0.049)	0.621 (-0.518)
1zwv (α) (58)	0.627 (0.286)	0.668 (0.877)	-	0.627 (0.824)	0.719 (0.845)
2jtm (β) (60)	0.589 (-0.188)	0.669 (0.849)	-	0.751 (0.926)	0.805 (0.957)
2k57 (β) (61)	0.634 (0.899)	0.632 (0.204)	-	0.788 (0.978)	0.835 (0.994)
1gyz (α) (62)	0.663 (0.933)	0.643 (0.797)	-	0.684 (0.959)	0.744 (0.977)
2mi6 (β) (62)	0.520 (0.337)	0.551 (0.349)	-	0.583 (0.852)	0.832 (0.985)
2kac (α/β) (64)	0.655 (0.888)	0.639 (0.787)	-	0.574 (0.314)	0.858 (0.990)
1f0z (α/β) (66)	0.510 (0.408)	0.588 (0.791)	-	0.523 (0.761)	0.772 (0.985)
2bzt (α) (66)	0.623 (0.710)	0.642 (0.745)	-	0.788 (0.986)	0.847 (0.992)
2a63 (α/β) (66)	0.648 (0.651)	0.739 (0.955)	-	0.566 (0.522)	0.804 (0.991)
4c26 (α/β) (66)	0.585 (0.825)	0.604 (0.843)	-	0.602 (0.946)	0.727 (0.975)
1u97 (α) (69)	0.639 (0.566)	0.650 (0.553)	-	0.712 (0.972)	0.721 (0.975)
2kt2 (α/β) (69)	0.612 (0.488)	0.633 (0.624)	-	0.522 (0.788)	0.778 (0.981)
2l4m (α/β) (69)	0.554 (-0.432)	0.604 (0.195)	-	0.767 (0.990)	0.855 (0.996)
1q1v (α) (70)	0.678 (0.873)	0.697 (0.904)	-	0.757 (0.951)	0.793 (0.983)
2coo (α) (70)	0.654 (0.191)	0.706 (0.935)	-	0.733 (0.970)	0.996 (0.996)
2k2a (α) (70)	0.634 (0.151)	0.705 (0.924)	-	0.622 (0.869)	0.756 (0.941)

Tableau A.12. Comparison for targets in the validation G/IC set - Best in TOP5.
Notations and color coding are similar to that of Table A.7

Validation G/CC					
	sOPEPv1 - Lib1	sOPEPv2 - Lib2	APPTest	RaptorX	AlphaFold2
1uao (β) (10)	0.733 (0.902)	0.722 (0.594)	0.711 (0.119)	-	0.879 (0.972)
2e4e (Coil) (10)	0.745 (0.968)	0.722 (0.794)	0.717 (0.943)	-	0.842 (0.963)
2evq (β) (12)	0.750 (0.869)	0.738 (0.876)	0.867 (0.986)	-	0.911 (0.997)
2luf (α) (20)	0.654 (0.917)	0.643 (0.918)	0.667 (0.961)	-	0.702 (0.901)
2n16 (α) (20)	0.633 (0.679)	0.656 (0.116)	0.731 (0.932)	-	0.700 (0.782)
5t7q (α) (21)	0.693 (0.968)	0.668 (0.980)	0.758 (0.941)	-	0.717 (0.949)
6g4v (α) (24)	0.766 (0.571)	0.756 (0.612)	0.769 (0.378)	-	0.782 (0.529)
6r2x (α) (25)	0.733 (0.982)	0.758 (0.958)	0.772 (0.960)	-	0.782 (0.983)
1by0 (α) (27)	0.773 (0.845)	0.795 (0.834)	0.787 (0.855)	0.766 (0.844)	0.782 (0.828)
1yyb (α) (27)	0.816 (0.860)	0.824 (0.818)	0.826 (0.830)	0.814 (0.815)	0.828 (0.830)
2k76 (α/β) (30)	0.760 (0.859)	0.779 (0.619)	0.737 (-0.093)	0.788 (0.908)	0.847 (0.915)
2l0g (α) (32)	0.762 (0.788)	0.774 (0.652)	0.709 (0.605)	0.787 (0.874)	0.826 (0.969)
1wr3 (β) (36)	0.645 (0.837)	0.671 (0.882)	0.666 (0.938)	0.518 (0.527)	0.762 (0.981)
1wr4 (β) (36)	0.657 (0.884)	0.676 (0.837)	0.605 (0.885)	0.486 (-0.648)	0.808 (0.972)
2ki0 (α/β) (36)	0.685 (0.811)	0.687 (0.543)	0.591 (0.405)	0.704 (0.785)	0.733 (0.822)
1e0l (β) (37)	0.653 (0.832)	0.687 (0.829)	0.629 (0.929)	0.568 (0.890)	0.805 (0.990)
1bwx (α) (39)	0.688 (0.729)	0.699 (0.745)	0.676 (0.852)	0.671 (0.450)	0.703 (0.668)
1jrj (β) (39)	0.742 (0.970)	0.746 (0.920)	0.732 (0.977)	0.757 (0.986)	0.866 (0.994)
2ysc (β) (39)	0.666 (0.852)	0.695 (0.804)	0.585 (0.609)	0.746 (0.896)	0.836 (0.982)
2ysf (β) (40)	0.677 (0.947)	0.710 (0.879)	0.628 (0.794)	0.579 (-0.590)	0.820 (0.991)
2ysg (β) (40)	0.726 (0.904)	0.730 (0.868)	0.607 (-0.560)	0.568 (-0.442)	0.813 (0.978)
2ysi (β) (40)	0.713 (0.956)	0.722 (0.815)	0.655 (0.855)	0.699 (0.889)	0.810 (0.984)
1k1v (α) (41)	0.736 (0.764)	0.785 (0.913)	-	0.785 (0.953)	0.820 (0.982)
1wr7 (β) (41)	0.667 (0.852)	0.693 (0.927)	-	0.550 (0.056)	0.802 (0.988)
2dmv (β) (43)	0.703 (0.943)	0.726 (0.942)	-	0.597 (-0.604)	0.819 (0.985)
2m8j (β) (43)	0.653 (0.776)	0.666 (0.799)	-	0.528 (0.679)	0.813 (0.984)
2k9d (α) (44)	0.728 (0.951)	0.724 (0.942)	-	0.650 (0.828)	0.766 (0.961)
2p81 (α) (44)	0.746 (0.964)	0.725 (0.922)	-	0.461 (-0.605)	0.745 (0.991)
1f4i (α) (45)	0.706 (0.951)	0.705 (0.851)	-	0.761 (0.981)	0.824 (0.992)
1p9c (α) (45)	0.680 (0.572)	0.702 (0.738)	-	0.672 (0.724)	0.715 (0.745)
1pv0 (α) (46)	0.718 (-0.106)	0.783 (0.985)	-	0.799 (0.981)	0.847 (0.992)
2e5t (α) (46)	0.801 (0.947)	0.804 (0.818)	-	0.585 (0.528)	0.823 (0.983)
2jnh (α) (46)	0.669 (0.769)	0.690 (-0.380)	-	0.688 (0.669)	0.800 (0.981)
1dv0 (α) (47)	0.653 (-0.413)	0.692 (0.958)	-	0.720 (0.974)	0.804 (0.991)
1pgy (α) (47)	0.724 (0.865)	0.715 (0.925)	-	0.718 (0.936)	0.736 (0.933)
2ekk (α) (47)	0.772 (0.987)	0.744 (0.929)	-	0.583 (0.881)	0.859 (0.996)
2wxc (α) (47)	0.703 (0.859)	0.742 (0.966)	-	0.582 (0.903)	0.798 (0.992)
2j8p (α) (49)	0.704 (-0.678)	0.712 (0.835)	-	0.698 (-0.691)	0.706 (-0.662)
2wqg (α) (51)	0.705 (0.962)	0.646 (0.856)	-	0.706 (0.966)	0.764 (0.992)
1r4g (α) (53)	0.771 (0.978)	0.739 (0.937)	-	0.774 (0.944)	0.809 (0.987)
1rzs (α) (61)	0.639 (0.665)	0.626 (0.355)	-	0.533 (0.457)	0.857 (0.989)
1rq6 (α) (62)	0.640 (0.272)	0.639 (0.627)	-	0.718 (0.953)	0.760 (0.971)
1wji (α) (63)	0.717 (0.976)	0.707 (0.917)	-	0.791 (0.991)	0.819 (0.995)
2l54 (α/β) (63)	0.660 (0.234)	0.679 (0.905)	-	0.744 (0.950)	0.750 (0.956)
1du6 (α) (64)	0.713 (0.924)	0.708 (0.896)	-	0.699 (0.926)	0.761 (0.964)
2cp9 (α) (64)	0.685 (0.787)	0.693 (0.761)	-	0.734 (0.983)	0.843 (0.997)
1cok (α) (68)	0.656 (0.926)	0.692 (0.939)	-	0.738 (0.966)	0.748 (0.969)
1qpm (α/β) (69)	0.686 (0.925)	0.647 (0.301)	-	0.743 (0.934)	0.761 (0.977)
2kz9 (α) (69)	0.837 (0.758)	0.833 (0.503)	-	0.823 (0.636)	0.846 (0.677)
2fce (α) (70)	0.684 (0.843)	0.706 (0.799)	-	0.611 (0.805)	0.819 (0.989)

Tableau A.13. Comparison for targets in the validation G/CC set - Best ni TOP5. Notations and color coding are similar to that of Table A.7

Validation IG					
	sOPEPv1 - Lib1	sOPEPv2 - Lib2	APPTest	RaptorX	AlphaFold2
2mdu (β) (29)	0.466 (-0.565)	0.512 (0.164)	0.592 (0.869)	0.563 (0.768)	0.747 (0.926)
2gdl (α) (31)	0.474 (-0.143)	0.482 (-0.173)	0.555 (-0.023)	0.470 (0.035)	0.594 (0.091)
1vpu (α) (45)	0.523 (0.553)	0.546 (-0.283)	-	0.558 (0.357)	0.542 (0.345)
1nd9 (α/β) (49)	0.539 (0.655)	0.558 (0.695)	-	0.553 (0.677)	0.566 (0.810)
2lhc (α) (56)	0.547 (-0.481)	0.516 (-0.081)	-	0.515 (-0.496)	0.735 (0.968)
2m4y (β) (56)	0.516 (0.160)	0.548 (0.664)	-	0.561 (0.904)	0.672 (0.872)
1gyf (α/β) (62)	0.517 (0.292)	0.563 (0.587)	-	0.707 (0.898)	0.792 (0.984)
2cw1 (α/β) (65)	0.542 (-0.187)	0.549 (-0.066)	-	0.694 (0.898)	0.751 (0.962)
2jtv (α/β) (65)	0.510 (0.136)	0.545 (-0.122)	-	0.754 (0.987)	0.803 (0.992)
1qxf (β) (66)	0.526 (0.501)	0.556 (0.050)	-	0.722 (0.971)	0.774 (0.983)
2l8d (β) (66)	0.572 (0.512)	0.531 (-0.160)	-	0.735 (0.985)	0.790 (0.993)
2jrr (β) (67)	0.488 (-0.113)	0.502 (-0.053)	-	0.620 (0.780)	0.755 (0.991)
2kaf (α/β) (67)	0.524 (0.127)	0.492 (-0.025)	-	0.516 (0.367)	0.851 (0.993)
2m2l (α/β) (67)	0.550 (0.445)	0.582 (0.739)	-	0.601 (0.904)	0.743 (0.988)
1ne3 (β) (68)	0.511 (0.149)	0.526 (0.012)	-	0.721 (0.977)	0.736 (0.977)
1y2y (β) (68)	0.479 (-0.316)	0.493 (0.143)	-	0.490 (-0.215)	0.494 (-0.037)
2do3 (β) (69)	0.557 (0.566)	0.521 (0.394)	-	0.747 (0.982)	0.831 (0.998)
2dy8 (α/β) (69)	0.549 (0.047)	0.592 (0.372)	-	0.521 (0.861)	0.765 (0.988)
2eqi (β) (69)	0.462 (0.462)	0.538 (0.508)	-	0.505 (0.786)	0.831 (0.996)
2mck (α) (69)	0.553 (0.082)	0.546 (-0.021)	-	0.480 (-0.348)	0.839 (0.992)
2xk0 (β) (69)	0.533 (0.398)	0.550 (0.796)	-	0.739 (0.989)	0.788 (0.990)
2lss (α/β) (70)	0.471 (-0.121)	0.512 (0.548)	-	0.572 (0.861)	0.739 (0.974)
2m7o (α/β) (70)	0.553 (0.498)	0.585 (-0.099)	-	0.697 (0.886)	0.750 (0.940)

Tableau A.14. Comparison for targets in the validation NG set - Best in TOP5. Notations and color coding are similar to that of Table A.7

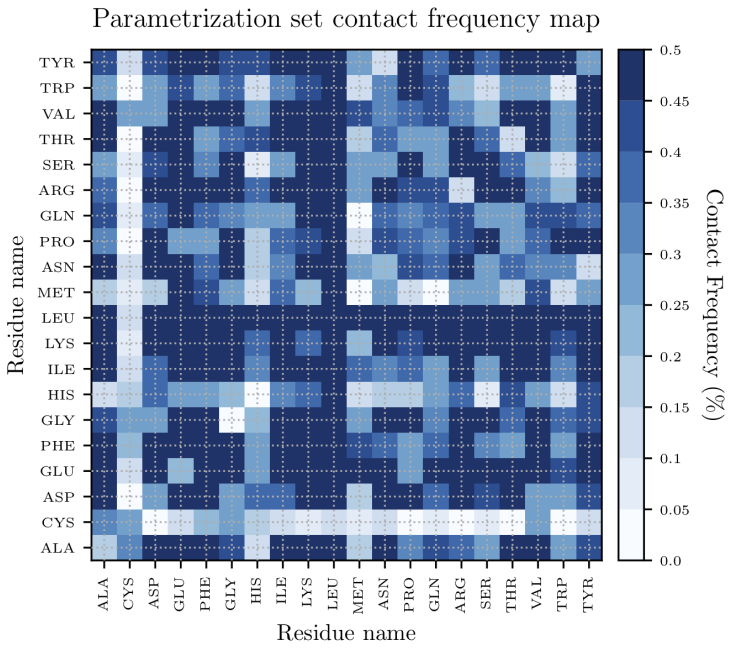


Fig. A.1. Parametrization set all-atom contact map. A contact between two side-chains is considered if the distance between at least one heavy atom pairs follows $r_{ij} < R_i + R_j + d$, where R_x is the van der Waals radius of atom type X and d is a cutoff here fixed at 1.5Å. Neighbouring residues is the amino acid sequence were not considered.

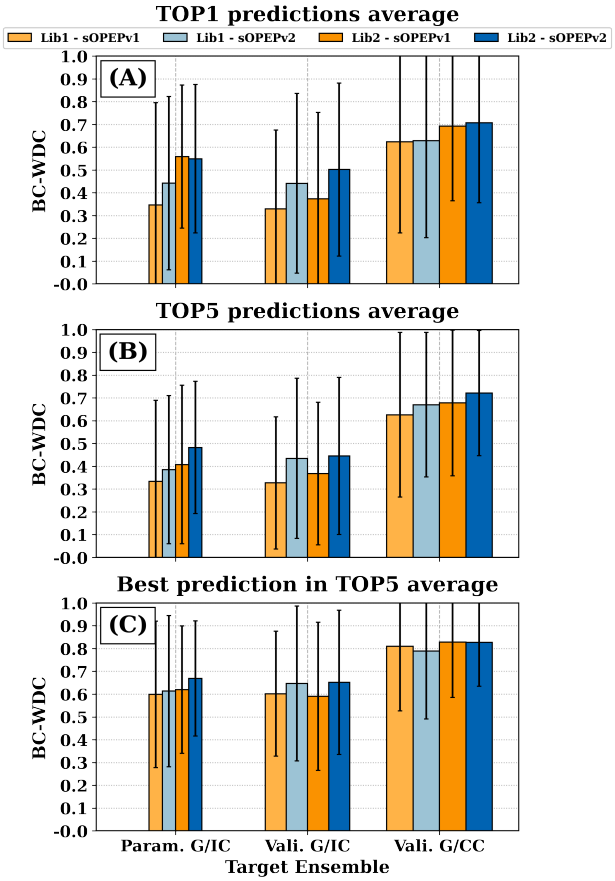


Fig. A.2. PEP-FOLD' models average BC-WDC. The x-axis is the name of our proteins' sets. In this case, *Param.*, *To Improve* and *Good* refers to the parametrization, the validation "To Improve" and the validation "Good" set containing respectively 25, 40 and 50 proteins. The bar width is proportional to the number of proteins in each set. For each set, the four columns represent from left to right, the original library/original potential (light orange), the original library/re-optimized potential (light blue), the new library/original potential (orange) and new library/re-optimized potential (blue) respectively. Panel (A) shows the average BC-WDC [**137**] for the lowest energy (TOP1) predictions. Panel (B) shows the the average BC-WDC for the five lowest energy (TOP5) predictions. Panel (C) shows the average BC-WDC for the prediction with the best CAD-CG in the TOP5. The WDC is taken from the *Protein Data Bank* validation report.

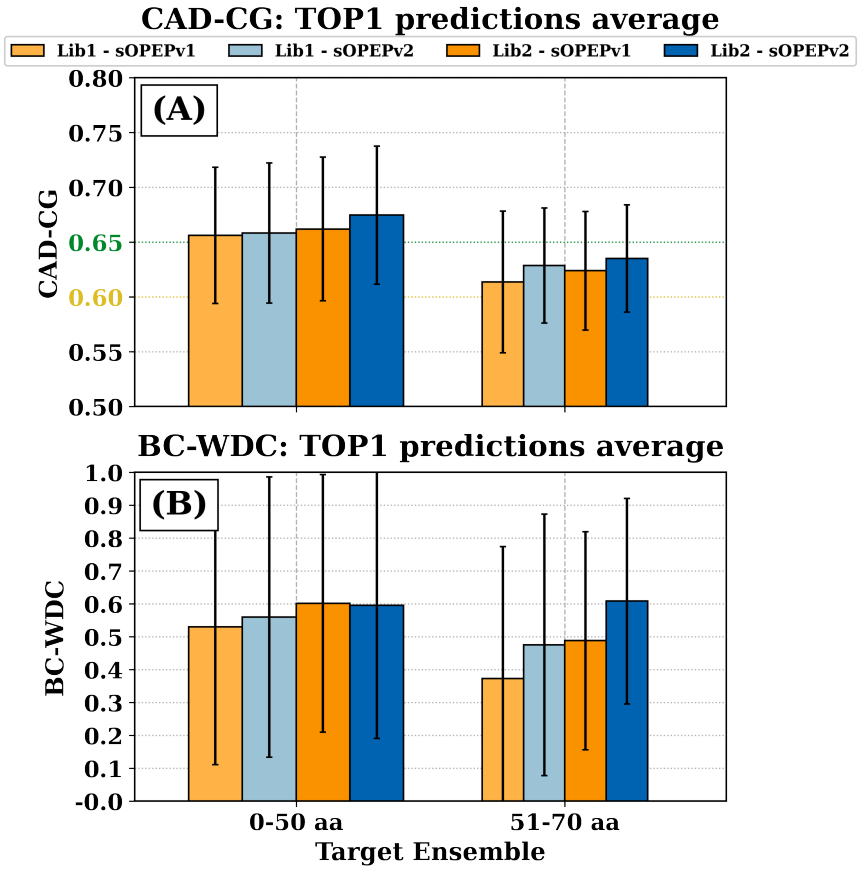


Fig. A.3. PEP-FOLD' models average scores by target size The x-axis is the name of our proteins' sets. In this case, *0-50aa* and *50-70aa* refers to the proteins of length between 0 and 50 amino acids (inclusively) and to the proteins of length between 50 (exclusively) and 70 amino acids set containing respectively 65 and 50 proteins. The bar width is proportional to the number of proteins in each set. For each set, the four columns represent from left to right, the original library/original potential (light orange), the original library/re-optimized potential (light blue), the new library/original potential (orange) and new library/re-optimized potential (blue) respectively. Panel (A) shows the average CAD-CG for the lowest energy (TOP1) predictions. Panel (B) shows the the average BC-WDC for the lowest energy (TOP1) predictions. The WDC is taken from the *Protein Data Bank* validation report.

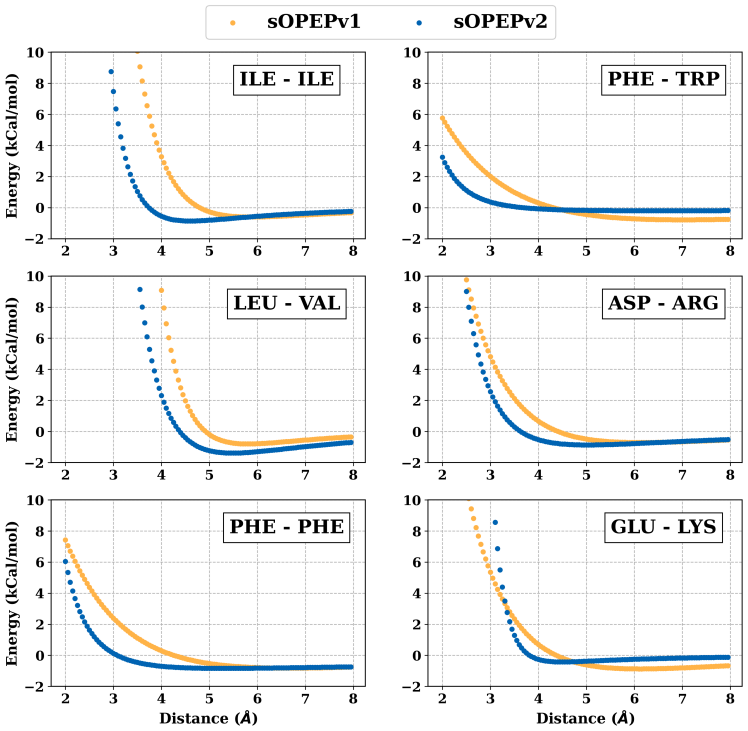


Fig. A.4. Updated attractive/repulsive potential of sOPEPv1 compared to sOPEPv2. The updated sOPEPv2 potential (in blue) compared to the sOPEPv1 potential (in orange). For these interactions, sOPEPv2 is more permissive at short distances than sOPEPv1.

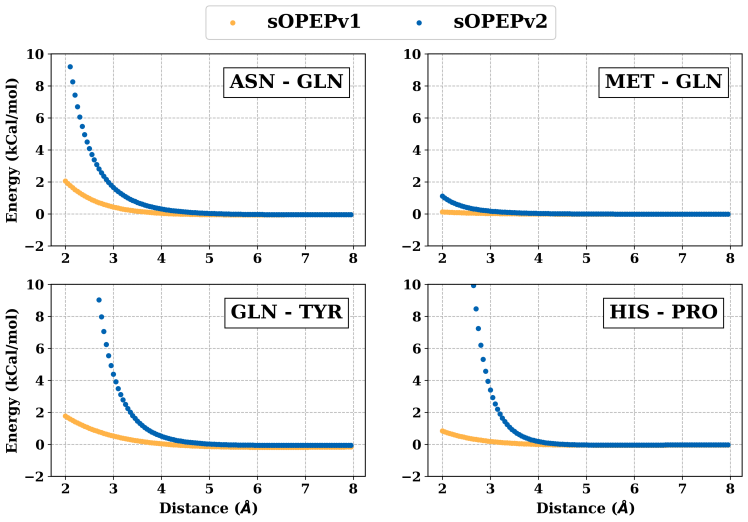


Fig. A.5. Updated attractive/repulsive potential of sOPEPv1 compared to sOPEPv2. The updated sOPEPv2 potential (in blue) compared to the sOPEPv1 potential (in orange). For these interactions, sOPEPv2 is less permissive at short distances than sOPEPv1.

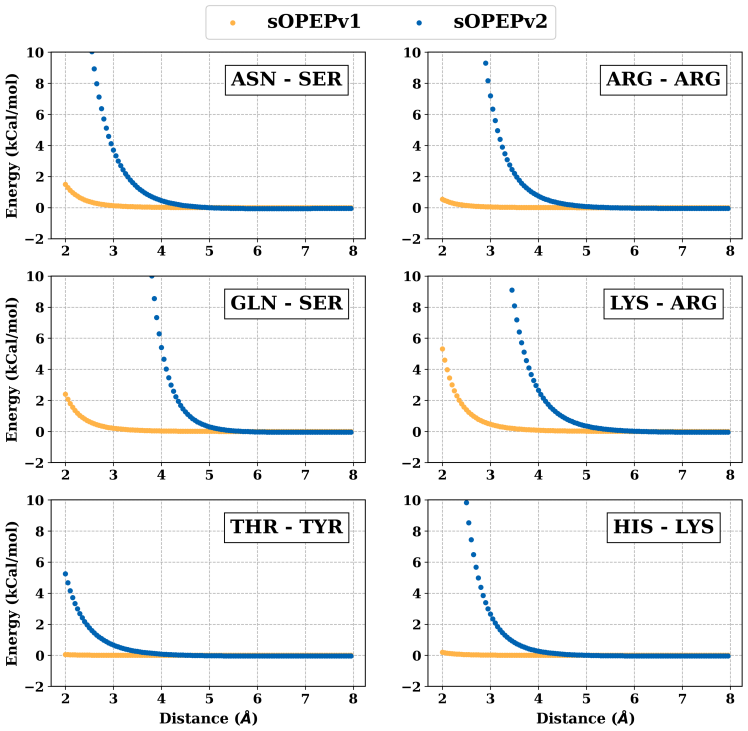


Fig. A.6. Updated repulsive potential of sOPEPv1 compared to sOPEPv2. For these interactions, sOPEPv2 is less permissive at short distances than sOPEPv1.

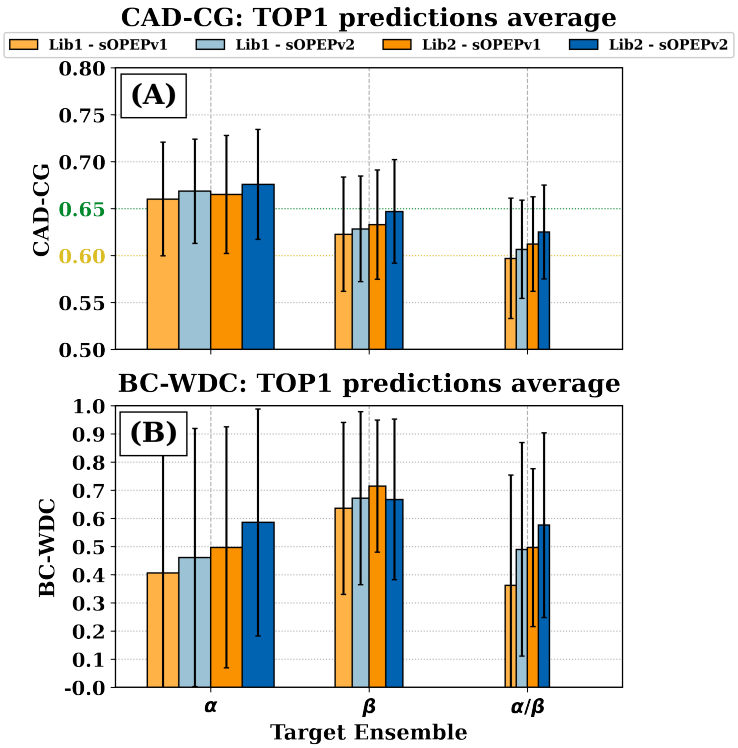


Fig. A.7. PEP-FOLD' models average scores by secondary structure class The x-axis is the name of our proteins' sets with α, β, α/β referring to the proteins of α, β and α/β secondary structure set containing respectively 52, 31 and 21 proteins. The bar width is proportional to the number of proteins in each set. For each set, the four columns represent from left to right, the original library/original potential (light orange), the original library/re-optimized potential (light blue), the new library/original potential (orange) and new library/re-optimized potential (blue) respectively. Panel (A) shows the average CAD-CG for the lowest energy (TOP1) predictions. Panel (B) shows the the average BC-WDC for the lowest energy (TOP1) predictions. The WDC is taken from the *Protein Data Bank* validation report.

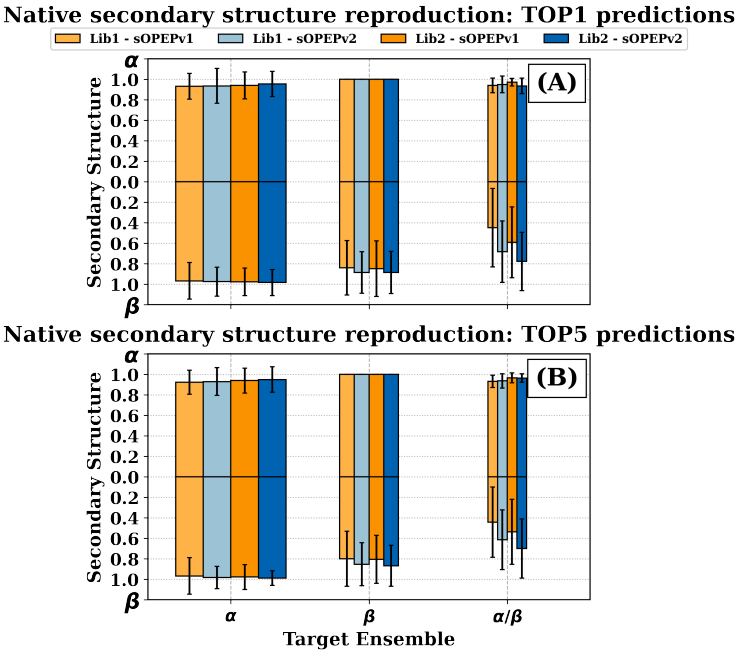
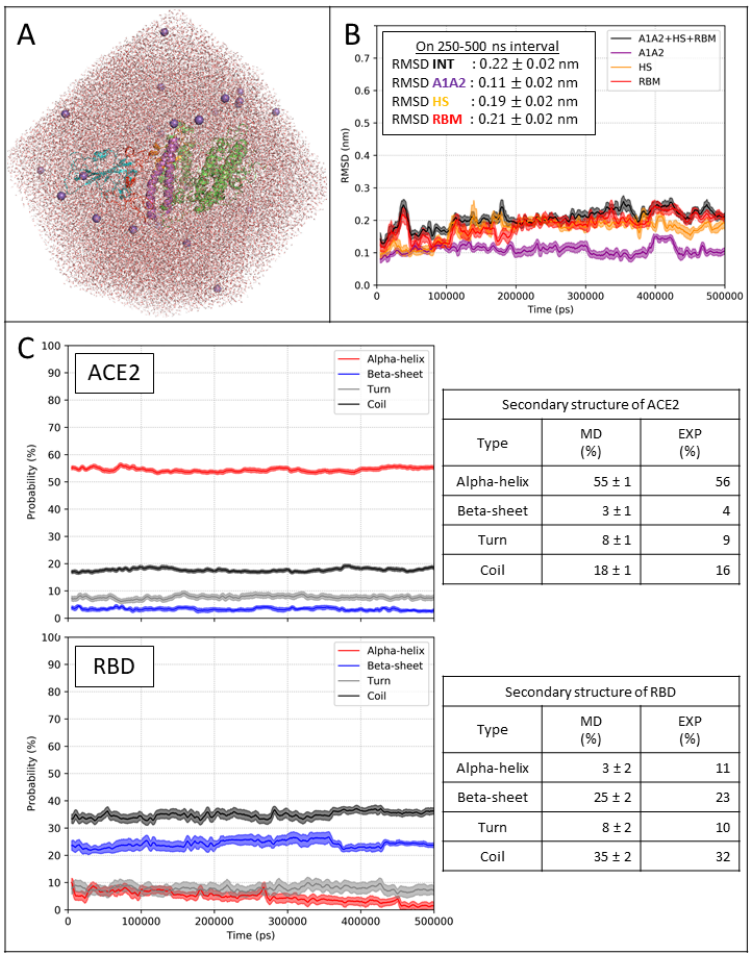


Fig. A.8. Native secondary structure reproduction by secondary structure class. The x-axis is the name of our proteins' sets with α, β, α/β referring to the proteins of α, β and α/β secondary structure set containing respectively 60, 32 and 21 proteins. The bar width is proportional to the number of proteins in each set. For each set, the four columns represent from left to right, the original library/original potential (light orange), the original library/re-optimized potential (light blue), the new library/original potential (orange) and new library/re-optimized potential (blue) respectively. Bars above and below the zero line shows α-helix and β-sheet native secondary structure reproduction respectively. Panel (A) shows the average native secondary structure reproduction for the lowest energy (TOP1) predictions. Panel (B) shows the the average native secondary structure reproduction for the five lowest energy (TOP5) predictions. Secondary structure assignments were done using STRIDE [**179**].

Annexe B

Supporting Figures: Corilagin and 1,3,6-Tri-O-galloy-β-D-glucose: potential inhibitors of SARS-CoV-2 variants



Secondary structure of ACE2		
Type	MD (%)	EXP (%)
Alpha-helix	55 ± 1	56
Beta-sheet	3 ± 1	4
Turn	8 ± 1	9
Coil	18 ± 1	16

Secondary structure of RBD		
Type	MD (%)	EXP (%)
Alpha-helix	3 ± 2	11
Beta-sheet	25 ± 2	23
Turn	8 ± 2	10
Coil	35 ± 2	32

Fig. B.1. Convergence of the ACE2-RBD complex MD simulation. (A) The simulated system with ACE2 in magenta (A1A2, residues 19-83), orange (HS, residues 322-362) and green (else) as well as RBD in red (RBM, residues 438-506) and teal (else). **(B)** Root mean square deviation (RMSD) on the backbone atoms (N, Cα, C and O) from the experimental crystal structure of the complex (PDB:6M0J) as a function of time for the regions at the interface (A1A2 and HS for ACE2 and RBM for RBD). The average and the standard deviation of the RMSD on the converged interval (250-500 ns) are shown in the inset. **(C)** The DSSP secondary structure as a function of time for ACE2 and RBD. The average and the standard deviation of the secondary structure on the converged interval are compared to the values computed on the experimental structure (EXP) on the right side. Only the α-helix, β-sheet, turn and coil content are shown. The difference with 100% is associated rest of the DSSP secondary structure classes (β-bridge, bend, 310-helix and π-helix). **(B-C)** The figures depict the running average using a 5-ns time window. The $\pm 1\sigma$ interval is shown by the semitransparent region around the curve.

Tableau B.1. Experimental and numerical comparison of H-bonds between ACE2-RBD. The first and second columns indicate the residues (and atoms) of respectively RBD and ACE2, involved in the formation of a H-bond, while the third column indicates the donor/acceptor length. The first three columns were computed on the minimized crystal structure [**21**]. Bold donor/acceptor pairs were also identified in the experimental paper [**21**]. The last two columns show the occurrence and the donor/acceptor length computed on the convergence interval (250-500 ns) of our MD of the ACE2/RBD complex.

Hydrogen Bonds				
RBD's atom	ACE2's atom	EXP length (Å)	MD occurrence (%)	MD length (Å)
K417(NZ)	D30(OD1)	2.7	17.16	2.89 ± 0.19
K417(NZ)	**D30(OD2)**	2.7	25.74	2.84 ± 0.16
Y449(OH)	D38(OD2)	2.6	40.64	2.73 ± 0.18
E484(OE2)	K31(NZ)	2.8	0.01	2.86 ± 0.07
N487(ND2)	**Q24(OE1)**	2.8	3.06	2.92 ± 0.16
N487(OD1)	**Y83(OH)**	2.7	5.06	2.85 ± 0.18
Y489(OH)	**Y83(OH)**	2.8	12.95	3.33 ± 0.99
Q493(OE1)	K31(NZ)	2.9	34.82	2.80 ± 0.11
Q493(NE2)	E35(OE1)	2.7	36.95	2.89 ± 0.16
Q498(NE2)	D38(OD2)	2.9	9.57	2.95 ± 0.21
Q498(NE2)	Q42(NE2)	3.0	0.61	3.19 ± 0.17
Q498(NE2)	Q42(OE1)	3.1	0.30	2.98 ± 0.14
T500(O)	N330(ND2)	3.0	0.00	-
T500(OG1)	D355(OD2)	2.7	7.34	2.84 ± 0.22
Y505(OH)	**E37(OE2)**	2.6	27.59	2.76 ± 0.19

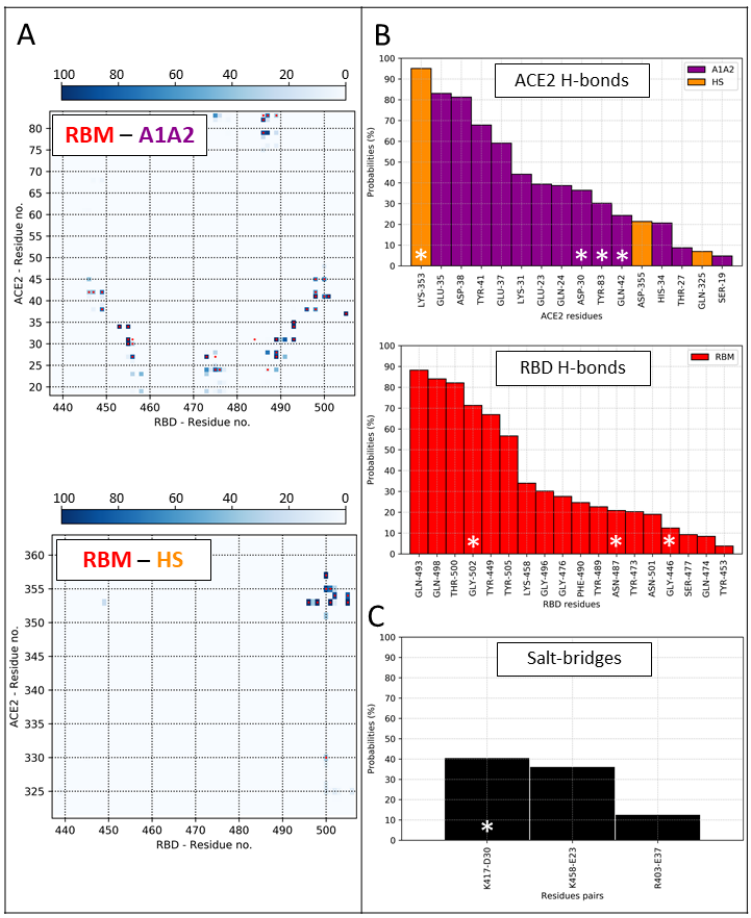


Fig. B.2. Contacts between ACE2 and RBD during the MD simulation. (A) Probability contact maps between A1A2 (residues 19-83 of ACE2) and RBM (residues 438-506 of RBD) as well as between HS (residues 322-362 of ACE2) and RBM (residues 438-506 of RBD). A contact is considered between two residues if the distance between any pair of atoms is smaller than 0.40 nm. The presence of a contact in the experimental crystal complex (PDB:6M0J) is shown by a red dot. **(B)** H-bonds probability of ACE2 residues with RBD (top) and of RBD residues with ACE2 (bottom). A H-bond is considered when the donor-acceptor distance is less than 0.35 nm and when the hydrogen-donor-acceptor angle is less than 35 degrees. The involvement of each residue in a H-bond in the experimental crystal structure is indicated by the white star. **(C)** Salt-bridges probability of ACE2 with RBD. A salt-bridge is considered when the distance between two oppositely charged groups is less than 0.40 nm. The presence of the salt-bridges in the crystal structure is shown by the white star. **(A-B-C)** All probabilities are computed on the converged interval (250-500 ns).

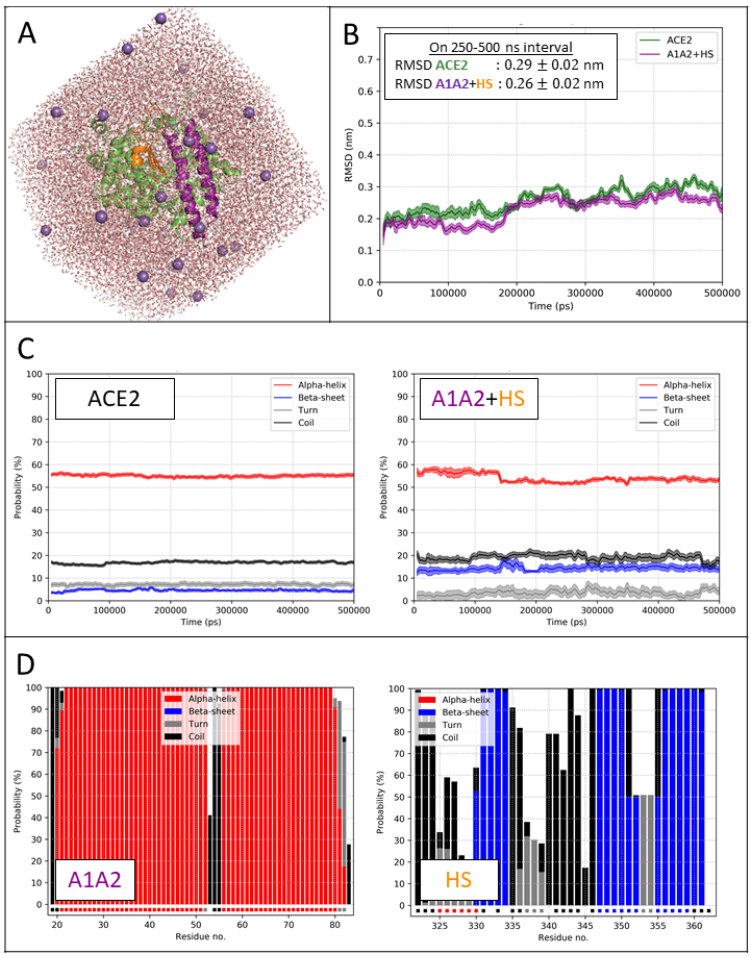


Fig. B.3. Convergence of the ACE2 MD simulation. (A) The simulated system with ACE2 in magenta (A1A2, residues 19-83), orange (HS, residues 322-362) and green (else). **(B)** Root mean square deviation (RMSD) on the backbone atoms (N, Cα, C and O) from the structure of ACE2 in the experimental crystal of the complex (PDB:6M0J) as a function of time for the whole protein (ACE2) and the regions at the interface (A1A2 and HS). The average and the standard deviation of the RMSD on the converged interval (250-500 ns) are shown in the inset. **(C)** The DSSP secondary structure as a function of time for ACE2 (left) and A1A2+HS (right). **(B-C)** The figures depict the running average using a 5-ns time window. The $\pm 1\sigma$ interval is shown by the semitransparent region around the curve. **(D)** The DSSP per residue secondary structure for A1A2 (left) and HS (right) on the converged interval (250-500 ns). The experimental secondary structure for each residue is illustrated by the square below the 0 mark. White means other secondary structure.

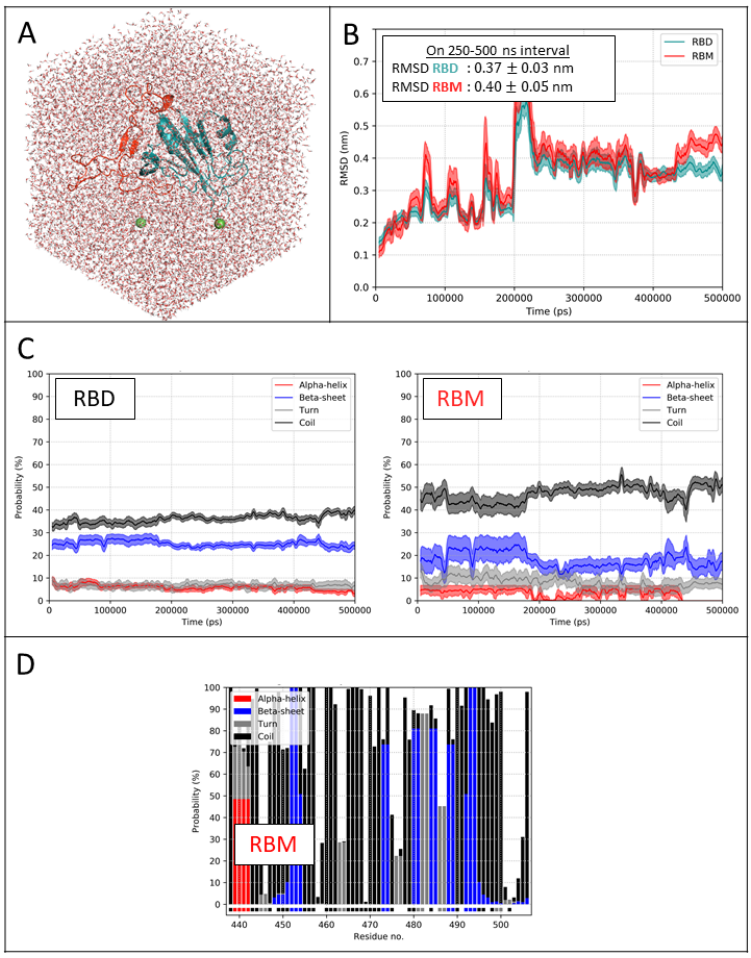


Fig. B.4. Convergence of the RBD MD simulation. (A) The simulated system with RBD in red (RBM, residues 438-506) and teal (else). **(B)** Root mean square deviation (RMSD) on the backbone atoms (N, Cα, C and O) from the structure of RBD in the experimental crystal of the complex (PDB:6M0J) as a function of time for the whole protein (RBD) and the region at the interface (RBM). The average and the standard deviation of the RMSD on the converged interval (250-500 ns) are shown in the inset. **(C)** The DSSP secondary structure as a function of time for RBD (left) and RBM (right). **(B-C)** The figures depict the running average using a 5-ns time window. The $\pm 1\sigma$ interval is shown by the semitransparent region around the curve. **(D)** The DSSP per residue secondary structure for RBM on the converged interval (250-500 ns). The experimental secondary structure for each residue is illustrated by the square below the 0 mark. White means other secondary structure.

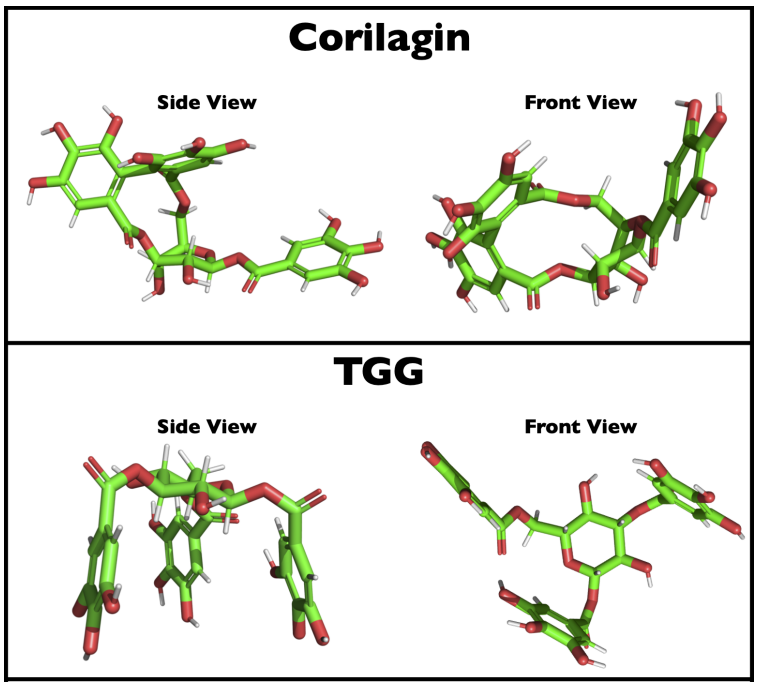


Fig. B.5. The initial structure of each ligand. (TOP) Corilagin (BOTTOM) TGG. The figures were generated using PyMOL [**95**].

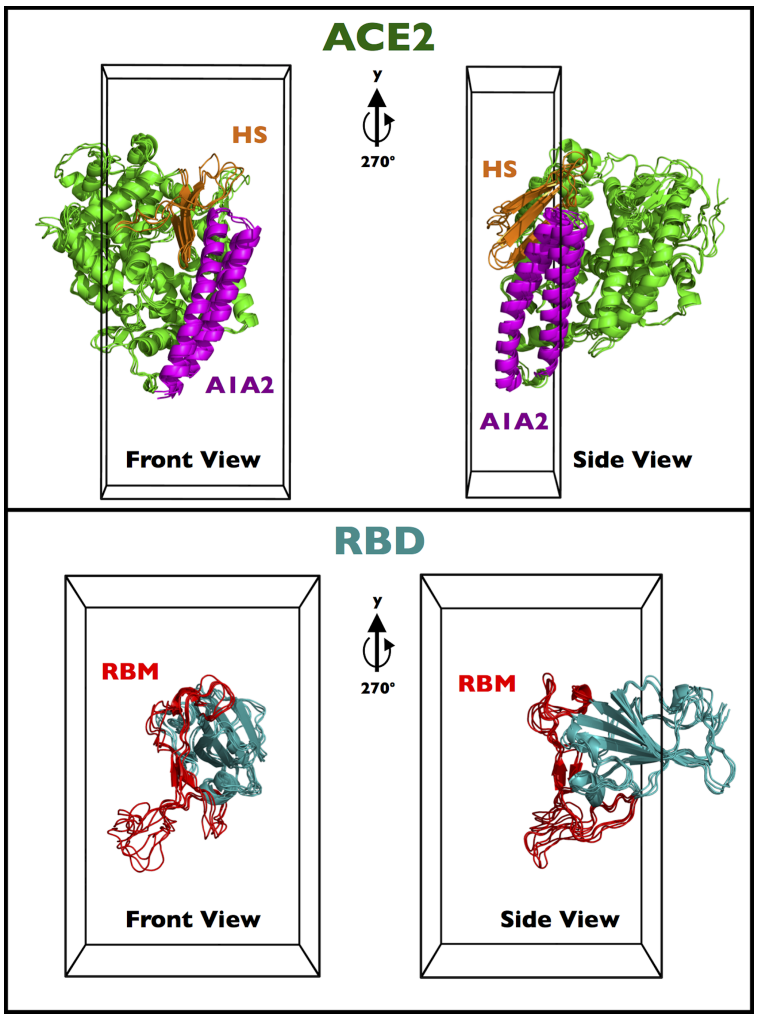


Fig. B.6. Configuration ensembles and box used for the molecular docking. (Top) Box used for the docking on ACE2. The box is centered around the point (6.5427, 8.3338, 6.5367) nm with a size of 2.2640 nm, 5.2072 nm, 1.4633 nm respectively in x, y and z. The A1A2 region (residues 19-83) and the HS region (residues 322-362) are shown respectively in magenta and orange. **(Bottom)** Box used for docking on the RBD. The box is centered around the point (6.13835, 6.8906, 1.1736) nm with a size of 2.7625 nm, 4.3358 nm, 2.6026 nm respectively in x, y and z. The RBM segment (residues 438-506) is shown in red.

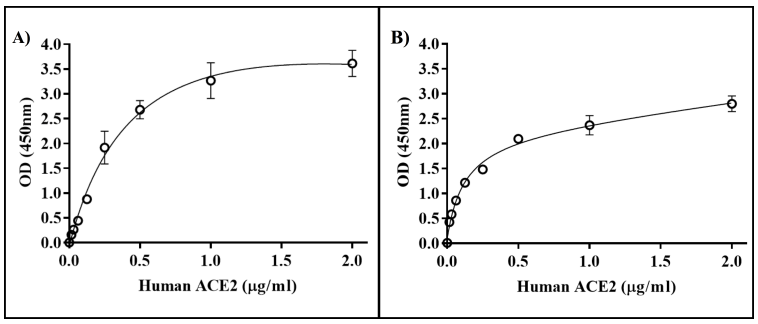


Fig. B.7. A) Human ACE2 protein binding to immobilized SARS-CoV-2 RBD Spike protein (0.5 μg/ml) using an increasing dose of human ACE2 protein (0,015 to 2 μg/ml). **B) Human ACE2 protein binding to immobilized ACE2 antibody** (0.5 μg/ml) using an increasing dose of human ACE2 protein (0,015 to 2 μg/ml). Results are expressed as mean ± standard error.

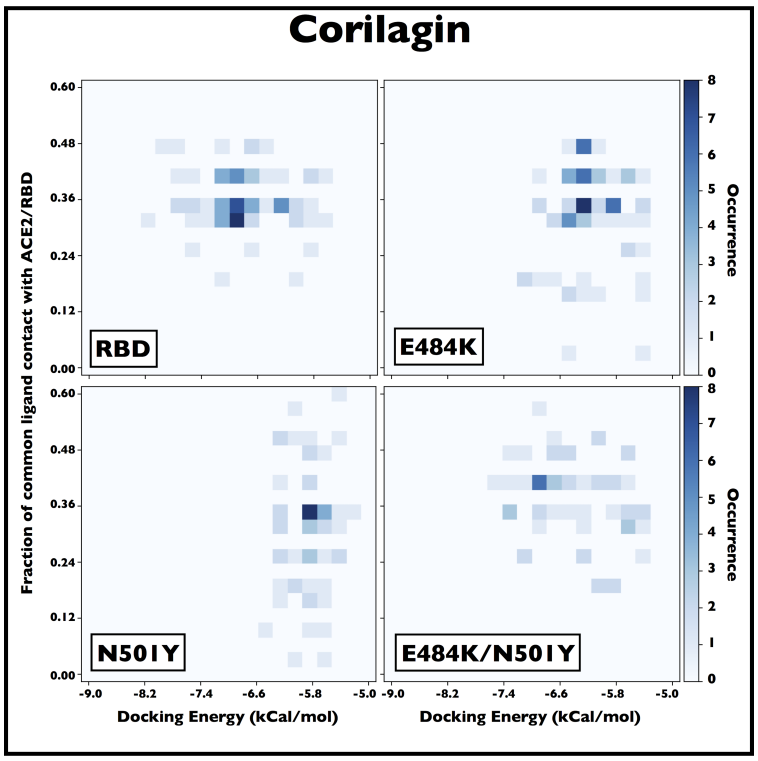


Fig. B.8. Corilagin's ability of the generated docking conformations to block the ACE2-RBD complex formation. The docking conformations generated by AutoDock VINA [**55**] as a function of its binding energy (x-axis) and fraction of ACE2/RBD contacts such conformation is able to black (y-axis). The occurrence of such conformation is shown as the z-axis.

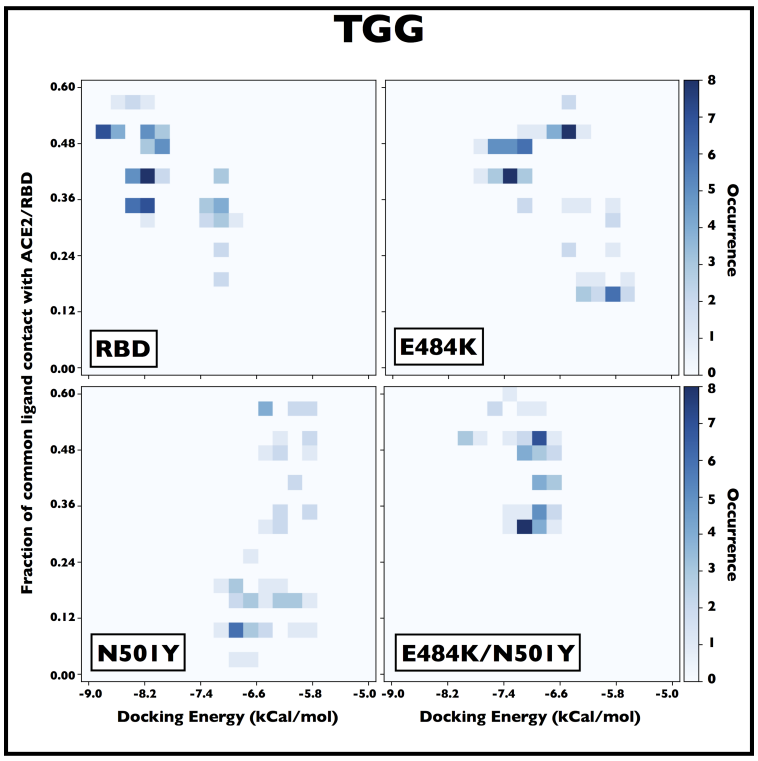


Fig. B.9. TGG's ability of the generated docking conformations to block the ACE2-RBD complex formation. The docking conformations generated by AutoDock VINA [**55**] as a function of its binding energy (x-axis) and fraction of ACE2/RBD contacts such conformation is able to black (y-axis). The occurrence of such conformation is shown as the z-axis.

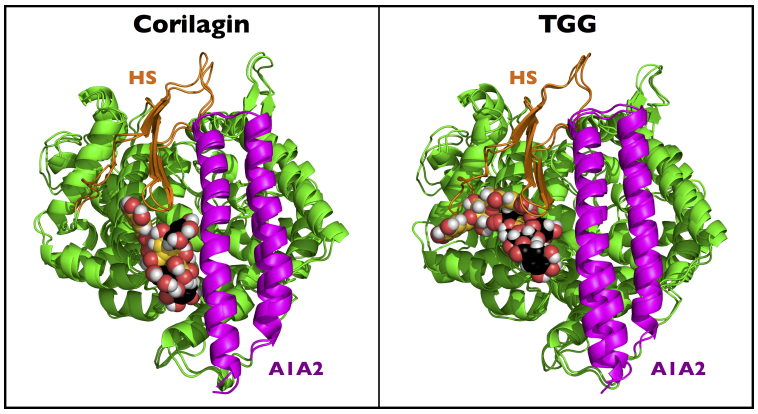


Fig. B.10. Ligands poses on ACE2. The docked position of corilagin (right) and TGG (left) on the ACE2 protein. The A1A2 segment, the HS segment and the rest of the RBD is shown in purple, orange and green respectively. The ligand in black and gold is respectively the conformation after docking and the center of the biggest cluster sampled during MD simulation respectively.

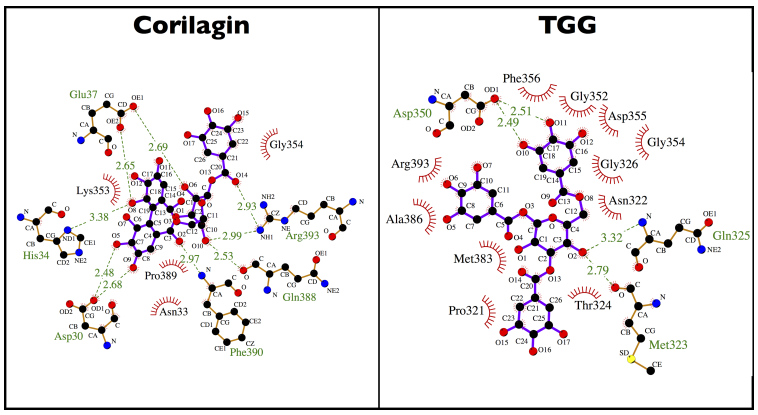


Fig. B.11. Ligands interaction map with ACE2. The interaction maps of corilagin (right) and TGG (left) with ACE2 are shown for the center of the biggest cluster computed on the convergence interval using the protein backbone atoms and ligand non-hydrogen atoms. The nonpolar contacts, defined by a distance smaller than 0.40 nm, between the ligand and the protein are shown as red arcs. H-bonds and their donor/acceptor distance are shown in green. All figures were generated using LIGPLOT [**103**, **104**].

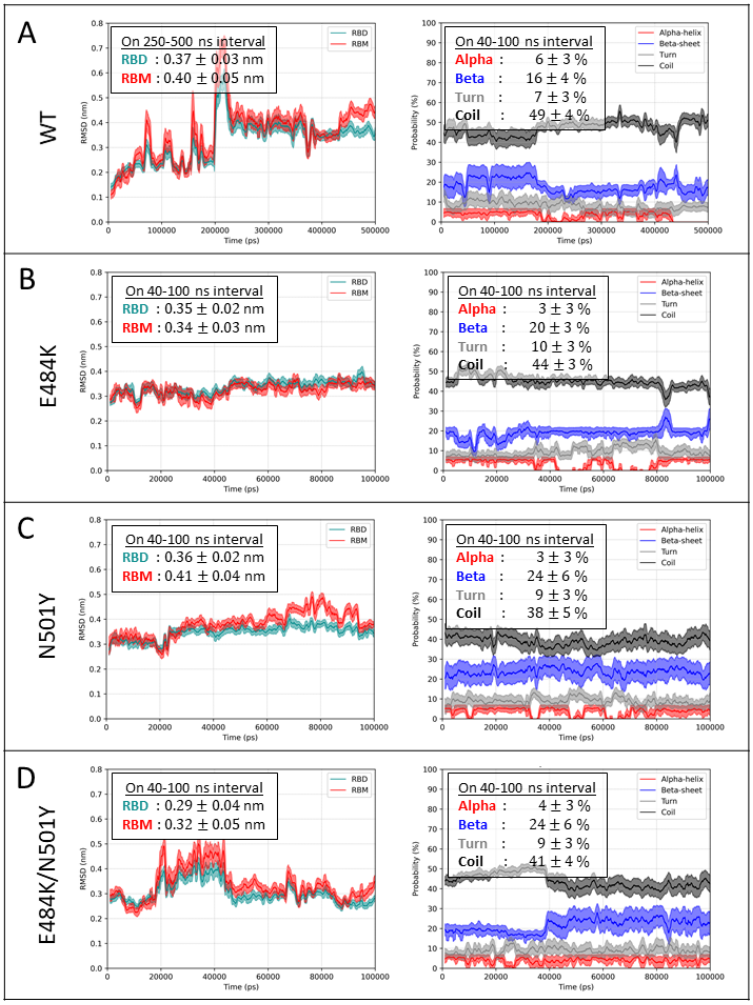


Fig. B.12. Convergence of the RBD mutants MD simulations. Backbone RMSD on the N, Cα, C and O atoms from the wild-type experimental structure and DSSP secondary structure as a function of time for (A) wildtype RBD, (B) E484K, (C) N501Y and (D) E484K/N501Y mutations.

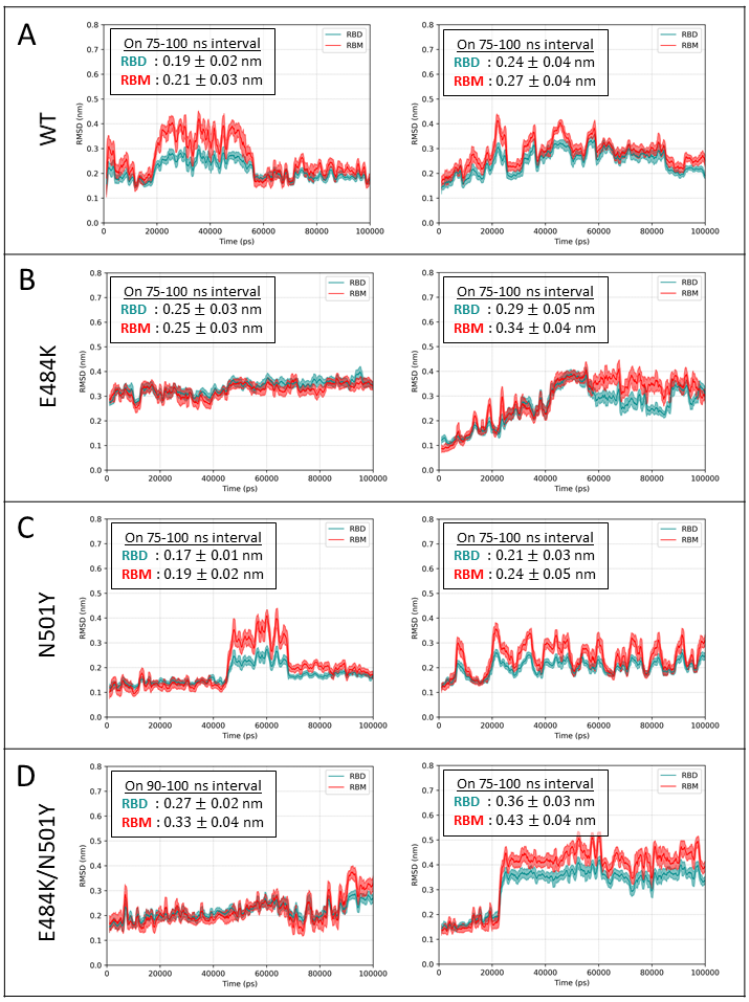


Fig. B.13. Convergence of the ligands–RBD MD simulations. Convergence is assessed by monitoring the RMSD on the backbone atoms of RBD (N, Cα, C and O) and the non-hydrogen atoms of the ligands from the initial structure as a function of time for (A) wildtype RBD, (B) E484K, (C) N501Y and (D) E484K/N501Y mutations. To the left, corilagin–RBD. To the right, TGG–RBD.

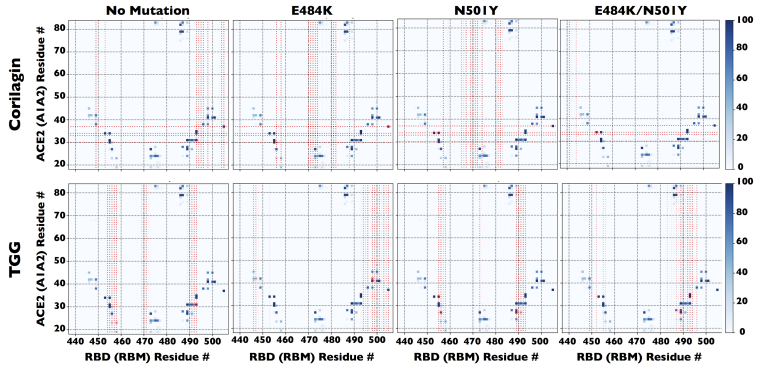


Fig. B.14. Ligand-protein contact network on A1A2 and RBM. The contact probability map between the A1A2 segment of ACE2 (vertical axis) and the RBM segment of RBD (horizontal axis) is shown in blue. These probabilities were computed on the converged interval of the MD simulation on the ACE2-RBD complex. The red disks indicate the residues blocked by corilagin (top row) and TGG (bottom row) during our ligand-protein MD simulations on ACE2, RBD, RBD/E484K, RBD/N501Y and RBD/E484K-N501Y. The size of the circles is proportional to the interaction probability with the residues.

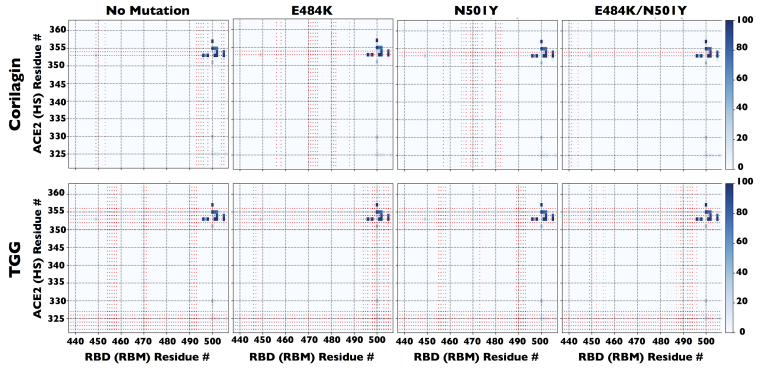


Fig. B.15. Ligand-protein contact network on HS and RBM. The contact probability map between the HS segment of ACE2 (vertical axis) and the RBM segment of RBD (horizontal axis) is shown in blue. These probabilities were computed on the converged interval of the MD simulation on the ACE2-RBD complex. The red disks indicate the residues blocked by corilagin (top row) and TGG (bottom row) during our ligand-protein MD simulations on ACE2, RBD, RBD/E484K, RBD/N501Y and RBD/E484K-N501Y.