Computational methods for protein recognition : application to O-GlcNAcylation prediction and SARS-CoV-2 interactions

Théo Mauri

Université de Lille - Faculté des Sciences et Technologies Ecole Doctorale Biologie-Santé (ED 446)

THÈSE DE DOCTORAT D'UNIVERSITÉ

Méthodes informatiques pour la reconnaissance des protéines : application à la prédiction de la *O*-GlcNAcylation et aux interactions du SARS-CoV-2

Computational methods for protein recognition: application to *O*-GlcNAcylation prediction and SARS-CoV-2 interactions

Présentée publiquement le 16 Décembre 2022 par

Théo MAURI

En vue de l'obtention du titre de

Docteur de l'Université de Lille, Spécialité : Biochimie et biologie moléculaire

Composition du Jury		
M. Raphaël GUEROIS	Directeur de recherche - Institut de Biologie Intégrative de la Cellule (I2BC) - CEA, Université Paris-Saclay, CNRS	Rapporteur
Mme Sophie SACQUIN-MORA	Directrice de recherche - Laboratoire de Biochimie Théorique (UPR9080) - Institut de Biologie Physico-Chimique	Rapporteuse
M. Christophe BIOT	Professeur des universités - Unité de Glycobiologie Structurale et Fonctionnelle - Université de Lille, CNRS	Examinateur **Président du Jury**
Mme Stéphanie OLIVIER-VAN STICHELEN	Assistant professor - Department of Biochemistry - Medical College of Wisconsin, Milwaukee, USA	Examinatrice
Mme Caroline SMET-NOCCA	Maîtresse de conférences - Inserm, CHU Lille, Institut Pasteur de Lille, U1167 - RID-AGE - Risk Factors and Molecular Determinants of Aging-Related Diseases - Université de Lille	Examinatrice
M. Tony LEFEBVRE	Professeur des universités - Unité de Glycobiologie Structurale et Fonctionnelle - Université de Lille, CNRS	Membre invité
M. Marc LENSINK	Directeur de recherche - Unité de Glycobiologie Structurale et Fonctionnelle - Université de Lille, CNRS	Directeur de thèse

Remerciements

Je tiens à remercier l'Unité de Glycobiologie Structurale et Fonctionnelle (UGSF), dirigée par le Pr. Christophe D'HULST puis par le Dr. Yann GUERARDEL, de m'avoir accueilli durant toutes ces années. Et je remercie à nouveau le Pr Yann GUERARDEL de m'avoir permis de participer aux Commissions Mixtes de la Facultés des Sciences et Techniques pour le département de Biologie. Cela m'a permis de mieux comprendre les enjeux pour la création de postes au sein de l'Université.

Je souhaite remercier l'ensemble des membres de mon jury de thèse sans qui elle ne pourrait avoir lieu. Je remercie notamment le Dr. Sophie SACQUIN-MORA et le Dr. Raphaël GUEROIS qui me font l'honneur d'être la rapporteuse et le rapporteur de ce manuscrit de thèse. Je remercie également les Dr. Caroline SMET-NOCCA et Dr. Stéphanie OLIVIER - VAN STICHELEN et le Pr. Christophe BIOT d'avoir accepté d'examiner ma thèse. Je remercie également le Pr. Tony LEFEBVRE d'avoir accepté de participer à cette défense en tant que membre invité. Cela compte beaucoup pour moi.

I sincerely thank my PhD supervisor Dr. Marc Ferdinand LENSINK. Thank you for your PhD subject proposal and thank you for your confidence, this sharing of knowledge and these evenings of board games too few because of the COVID-19. I enjoyed evolving by your sides. I discovered the protein-protein interaction world thanks to you. I am grateful to you for letting me actively participate in CAPRI during this health crisis to help understand this infection. My English level improved a lot thanks to you as well as my autonomy and my reflection.

I truly thank all the Computational Molecular System Biology members, gone or still present. Guillaume, the best colleague I will ever have and one the best people I know. Your rigor, your scientific reflection and your daily joy are for me a model to follow. Thank you for your two years supervision during my apprenticeship. I have grown a lot by your side. Julie, for helping me to remind that biology is not always with a computer and for our various talks. Clarisse, even if you are very far now (yes a stair is a lot), in my mind you're still from our team. Thank you for everything and I wish you all the best for your future and I hope

you'll find a permanent position as soon as possible. Jérôme, thank you for the great discussions and debates whether at the lab or around a pint of beer (or more). I also learnt a lot from you. Ralf, thank you for your kindness. I'm sad that I was not able to see you more because of this health crisis. Physics is still complicated for me but I have a better understanding thanks to you. Shubham, thank you for our nerd talks and meme sharing. I wish you all the best for the second half of your PhD. I would also like to thank the two interns that I had the chance to supervise for two months. It was a great but a bit exhausting experience. I wish you both the best for the following.

Je souhaite remercier sincèrement l'équipe "*O*-GlcNAcylation, signalisation cellulaire et cycle cellulaire" pour nos échanges de qualités. Je vous remercie tous de m'avoir écouté sans rechigner lorsque je parlais trop info et particulièrement le Pr. Ikram EL YAZIDI-BELKOURA et le Dr. Anne-Sophie EDOUART-VERCOUTTER qui ont su être patientes avec moi. Je vous admire tous beaucoup et je sais que la *O*-GlcNAcylation n'aura un jour plus de secret pour le monde et que ce sera en grande partie grâce à vous. Alors que j'écris ces phrases, j'ai une pensée pour les doctorantes (Maïté, Moyira, Ninon et Sadia) que j'ai pu côtoyer au cours de mes cinq années et qui sont désormais docteures. C'était un plaisir de partager avec vous et je souhaite à la relève (Awatef, Jodie, Ferdinand et Dimitri) plein de courage pour votre thèse mais je sais que vous êtes bien entourés. Je remercie également Marlène, Quentin et Stephan qui m'ont toujours accueilli avec le sourire. Je remercie une nouvelle fois le Pr. Tony LEFEBVRE, pour avoir partager avec moi ses connaissances sur la *O*-GlcNAcylation et d'avoir cru en moi pour la prédiction de sites *O*-GlcNAcylés. Désolé que ça n'ait pas marché mais cette *O*-GlcNAcylation nous rendra tous fous! Je suis déçu de ne pas avoir eu assez de temps à consacrer sur l'interaction beta-caténine/OGT. Merci pour ta bonne humeur et ton aide précieuse que tu as toujours su m'apporter peu importe les contraintes.

Je remercie l'ensemble des personnes faisant ou ayant fait partie de l'UGSF. Merci pour cette ambiance familiale qui permet des échanges scientifiques de qualité. Je remercie particulièrement le Dr. Corentin SPRIET qui est devenu un ami. Je te remercie de m'avoir donné goût à la vulgarisation scientifique et de m'avoir fait confiance pour Xperium et les différentes Fêtes de la Science. Que de bons moments en mémoire. Nos bières de dernière

minute me manqueront. Je te souhaite plein de bonheur à toi et les deux femmes de ta vie. Essayes de ne pas trop penser à moi quand tu joueras à “Où est Charlie”. En parlant de Fête de la Sciences je remercie également Coralie, Angelina, Mélanie, Xavier, Jean-François, Vincent et Yannick, pour ces animations et discussion de qualité. Je remercie à nouveau le Pr. Christophe BIOT, pour ta confiance, nos discussions et échanges musicaux de qualité. Si tu viens sur Rouen, n'hésites pas à te manifester.

Je remercie également le Dr. Laurence MENU-BOUAOUICHE et le Pr. Muriel BARDOR pour m’avoir accueilli au sein de votre laboratoire Glyco-MEV lors de mes premiers stages. Je vous remercie sincèrement pour la confiance que vous m’avez portée pour monter l'alternance avec l’UGSF qui a changé ma vie. Merci pour vos retours et votre savoir être scientifique qui m‘ont beaucoup appris. Je remercie vraiment Laurence pour avoir cru en moi et qui restera à jamais celle qui m’a permis de découvrir la bioinformatique structurale et la glycobiologie.

Je remercie le Dr. Jessica ANDREANI, Pr. Pascal TOUZET et le Dr. Alexandre BONVIN, membres de mon CSI. Je vous remercie pour vos précieux conseils apportés lors de ces moments d’échanges qui m’ont permis de bien évoluer tout au long de ma thèse. Je suis sincèrement déçu de ne pas avoir pu vous rencontrer physiquement malgré différents évènements tels que EMBO Courses qui s’est déroulé en visio et JOBIM 2022.

Je remercie également les adhérents et membres du bureau de l’association des docteurs et doctorants de Biologie Santé de Lille BioAddoct que j’ai eu la chance de présider pendant deux ans. Merci pour ces moments de partage.

Je me dois aussi de remercier ceux sans qui cette thèse n’aurait été possible: mes amis. Mes amis d’enfance Raphy, Roro et Xaviminou, qui continuent toujours de m’écouter et me supporter après plus de 27 ans maintenant. Merci à Xav et Raph d’être toujours présents malgré votre vie portugaise. Xaviminou, merci pour tous ces memes et ces partages qui m’ont bien soutenu. Reste comme tu es car tu es génial. Mon Roro, je te remercie pour tout ce que tu es, pour tes précieux conseils d’homme expérimenté en termes de thèse mais aussi pour ces games accompagnées ou non de ta douce. Reste toi aussi comme tu es, tu es

formidable. En parlant d'elle, je remercie sincèrement Milou pour sa bonne humeur et ses encouragements. Merci à vous deux pour les sessions Apétarot en ligne ou Mario Kart. Vous m'avez beaucoup apporté. Je vous souhaite plein de bonheur pour le futur et notamment pour votre nouvelle aventure qu'est la maison et plein d'autres j'espère.

Je remercie également mes amis de longues dates qui m'ont permis de changer de monde un soir toutes les deux semaines: Clément H, Clément M, Louis, Maël le sale breton, Maki et Philou le Filou ou devrais-je dire Sir Kamtag, Klehm, Dragar, Tazouk, Père Spasfon et Tabal, mes joueurs de JdR. Merci pour ces fou-rires et ce non avancement qui me permettent de préparer un scenar qui dure finalement une thèse! Merci particulièrement à Maki, Maël et Philippe pour notre conversation qui me soutient beaucoup.

Merci aussi à tous ceux qui viennent nous voir dans notre lointaine région qu'est le Nord. Je pense notamment à Mélanie et Romain le frère, Simon et Marina, Églantine, Benramine, Julien, Tom et Justine. Merci aussi à Laura et Alex pour ces petits week-ends dans le "Sud". Merci à tous pour ces bons moments.

Enfin je remercie à nouveau Guillaume, pour tous nos moments, nos chansons, nos cris, nos balades à vélo pour éviter d'en faire et l'odeur du pamplemousse. Merci sincèrement pour ces 5 ans de vrai partage de vie. Tu es et resteras le maître. Signé le dernier des Mauri (se prononce MaOri).

Enfin je remercie ma famille et je commence par ma belle-famille: Anne et André, merci pour tout, vos encouragements, votre intérêt pour ma thèse et votre soutien. Je souhaite ensuite remercier ma petite famille. Tout d'abord mes parents, ceux qui m'ont toujours supportés, de par leur amour et leurs sourires, financièrement, pour avoir payé mes études qui ont été longues et je vous en serai éternellement reconnaissant. Merci pour tout, grâce à vous j'ai pu m'épanouir dans la vie personnelle et professionnelle. Moman, peut-être qu'un jour tu arriveras à prononcer *O*-GlcNAcylation; en tout cas, moi, je crois en toi. Papou je sais que tu es fier de moi. Je vous aime. Je remercie aussi chaleureusement mon petit frère. Thibthib tu es une des personnes à laquelle je tiens le plus au monde. Tu es une personne superbe. Crois en toi comme tu crois en moi. Je t'aime.

Pour finir, j'aimerai remercier la personne qui représente à la fois ma meilleure amie mais aussi ma famille, toi, Julia, chouchoute, mon amoureuse. Même si tu as dû subir des

discussions en anglais de sujets qui ne te parlaient pas du tout, tu m'as toujours soutenu. Tu as changé de vie pour moi, pour cette thèse et je t'en serai éternellement reconnaissant. Merci pour ton soutien, ta motivation et ta compréhension lors des longues soirées que je passais à travailler. Tu es mon phare qui me guide dans la vie et je n'en serais pas là aujourd'hui sans toi. Je n'arrive pas à exprimer à quel point je te suis reconnaissant. Merci pour les rires, les pleurs, les voyages et cette belle aventure qu'est notre amour. J'ai hâte de découvrir ce qui nous attend car ça sera avec toi. Merci pour cette belle aventure qui s'annonce encore plus merveilleuse. Merci de porter en toi notre futur qui saura nous apporter un autre bonheur. Je t'aime toi et notre petit pois.

Preface

This PhD thesis has been carried out under the supervision of Dr. Marc Ferdinand Lensink within the Unité de Glycobiologie Structurale et Fonctionnelle, Faculté des Sciences et Technologies, Lille University, Villeneuve d'Ascq, in the Computational Molecular System Biology team. It was financed by the Ministère de l'Education Nationale, de l'Enseignement Supérieur et de la Recherche (MENESR) and was the subject of publications and oral and poster communications presented below:

Publications:

Mauri T, Menu-Bouaouiche L, Bardor M, Lefebvre T, Lensink MF and Brysbaert G. (2021). *O*-GlcNAcylation Prediction: An Unattained Objective. *Advances and Application in Bioinformatics and Chemistry.*

Lensink MF, Brysbaert G, **Mauri T** *et al.* (2021). Prediction of protein assemblies, the next frontier: The CASP14-CAPRI experiment. *Proteins.*

Previous publications:

Brysbaert G, **Mauri T**, and Lensink MF. (2018). Comparing protein structures with RINspector in Cytoscape. *F1000Research*.

Brysbaert G, **Mauri T**, de Ruyck J and Lensink MF. (2019). Identification of Key Residues in Proteins Through Centrality Analysis and Flexibility Prediction with RINspector. *Current Protocol in Bioinformatics.*

Oral presentation:

Mauri T, Brysbaert G, Bates AP, Wodak JS and Lensink MF. "Analysis of SARS-CoV-2 and human protein interactions: a CAPRI-COVID Round". JOBIM 2022 (Journées Ouvertes en Biologie, Informatique et Mathématiques), Rennes, France.

Mauri T, Menu-Bouaouiche L, Bardor M, Lefebvre T, Lensink MF and Brysbaert G. "*O*-GlcNAcylation Prediction: An Unattained Objective". Journée André Verbert 2021, Lille, France

Poster presentations:

Brysbaert G, **Mauri T**, and Lensink MF. “Comparing protein structures with RINspector in Cytoscape”. EMBO2021 Course: Integrative modelling of biomolecular interactions. Izmir, Turkey (Zoom): Flash talk + Poster

Mauri T, Menu-Bouaouiche L, Bardor M, Lefebvre T, Lensink MF and Brysbaert G. “O-GlcNAcylation Prediction: An Unattained Objective”. JOBIM2019, Nantes, France

Brysbaert G, **Mauri T**, and Lensink MF. “Comparing protein structures with RINspector in Cytoscape”. JOBIM2018, Marseille, France.

Curriculum Vitae

THÉO MAURI

Doctorant en Bioinformatique - UMR 8576 UGSF

@ mauritheopro@gmail.com +33 (0)6 65 57 81 27 6 allée des moissons, 59510 Forest-sur-Marque, France
in theo-mauri

EXPÉRIENCE

Doctorant

Unité de Glycobiologie Structurale et Fonctionnelle, CNRS - Université de Lille, Équipe Computational Molecular Systems Biology

Octobre 2019 – Septembre 2022 Villeneuve d'Ascq, FR

Référent: **Dr Lensink Marc F.**

@ marc.lensink@univ-lille.fr

+33 (0)3 62 53 17 28

Développement de méthodes d'analyse d'interactions protéine-protéine
Développement d'un outil de prédiction de sites de O-GlcNAcylation

- Machine learning
  - Random Forest
  - Support Vector Machine
  - Gradient Boosting Tree
- Développement
  - Gestion de version
  - Documentation
- Analyses de séquences protéiques

Analyses de complexes entre protéines du SARS-CoV-2 et de l'humain

- Clustering
  - Hierarchical Clustering
- Analyses tri-dimensionnelles
  - Réseau d'interaction de résidues
  - Interactions protéine-protéine
  - Identification d'interfaces
- Développement d'une méthode pour trouver un consensus sans template
- Création d'une page web

Analyses de l'interaction entre la β-caténine et la O-GlcNAc Transferase

- Modélisation:
  - Recherche de template
  - Identification de résidus centraux
  - Identification d'interfaces
  - Docking protéine - protéine
  - Docking protéine - peptide
- Minimisation d'énergie
- Dynamique moléculaire

Projets annexes:

- Encadrement d'étudiants: 2 stagiaires niveau M1 - 2 mois / Gestion de projet M2 - Responsable bioinformatique du projet
- Étude *in silico* de l'interaction entre des sous parties de ribosomes de différentes lignées *S. nutens* empêchant leur hybridation
- Étude *in silico* de la dimerisation de la FAS et du rôle de la *O*-GlcNAcylation

Apprenti - Master 2

UGSF, Équipe Computational Molecular Systems Biology, Dr Marc F. Lensink

Septembre 2017 – Août 2019 Villeneuve d'Ascq, FR

Référent: **M. Brysbaert Guillaume**

@ guillaume.brysbaert@univ-lille.fr

+33 (0)3 62 53 17 32

Développement d'une méthode de prédiction de sites de *O*-GlcNAcylation

- Séquences protéiques:
  - Compositions en acides aminés
  - Tests statistiques de proportions
- Structures secondaires:
  - Prédictions de structures secondaires
  - Prédiction de la flexibilité
- Structures tertiaires:
  - Modélisation de protéines
  - Calculs de modes normaux
  - Calculs d'accessibilité

DIPLOMES

Master BIMS (BioInformatique, Modélisation et Statistique)

Université de Rouen, Normandie

2019 – Master 2: Apprentissage de 2 ans

Licence de Biochimie, Biologie Moléculaire, Cellulaire et Physiologie

Université de Rouen, Normandie

2016

COMPÉTENCES

Développement:

Python | R | Perl | Bash/Unix | Git | JavaScript | HTML & CSS | PHP | MySQL | LaTeX | VSCodium | Rstudio

Logiciels/Outils:

Random Forest | Support Vector Machine | Gradient Boosting Tree | PyMOL | Cytoscape | AlphaFold | I-TASSER | Gitlab | Naccess | ElNemo

Domaine:

Apprentissage supervisé | Bases de données | Modélisation | Docking | Interaction protéine-protéine | Biologie Structurale

Professionnelles:

Autonomie | Médiation Scientifique | Encadrement | Communication | Rédaction | Veille Bibliographique

ANGLAIS

Linguaskill 179/180+

Listening
Reading
Speaking

EXPÉRIENCE

Stagiaire - Master 1

Laboratoire Glyco-MEV, Université de Rouen, Normandie

Mars 2017 – Juillet 2017 — Mont Saint Aignan, FR

Référent: **Dr. Menu-Bouaouiche Laurence**

@ laurence.menu-bouaouiche@univ-rouen.fr

+33 (0)3 62 53 17 32

Comparaison de méthodes de prédiction de *O*-Glycosylation

- Création d'un jeu de données
- Analyses des différentes méthodes de classification par apprentissage

Animateur commercial

Projef

Juin 2013 – Avril 2017 — Normandie, FR

Vente de produits dans des magasins de bricolage les samedis

Facteur

La Poste

Juin 2012 – Aout 2012 — Grand-Couronne, FR

Organisation et distribution du courrier en voiture, vélo et à pied

PUBLICATIONS

- **Mauri T**, Menu-Bouaouiche L, Bardor M, Lefebvre T, Lensink M, Brysbaert G O-GlcNAcylation Prediction: An Unattained Objective. *Adv Appl Bioinform Chem.* 2021 Jun 8.
- Lensink MF, Brysbaert G, **Mauri T** *et al.* Prediction of protein assemblies, the next frontier: The CASP14-CAPRI experiment. *Proteins.* 2021 Dec;89(12):1800-1823
- Brysbaert G, **Mauri T**, Lensink MF. Comparing protein structures with RINspector automation in Cytoscape. *F1000Res.* 2018 May 9 ;7:563.
- Brysbaert G, **Mauri T**, de Ruyck J, Lensink MF. Identification of Key Residues in Proteins Through Centrality Analysis and Flexibility Prediction with RINspector. *Curr Protoc Bioinformatics.* 2019 Mar;65(1):e66

COMMUNAUTÉ SCIENTIFIQUE

- **Création et animation d'un workshop:** *Workshop - Rentrée de l'EDBSL*
- **Xperium Saison 4 - Contrat 64h (2020-2022):**
  - Présentation du monde de la recherche publique à un public hétérogène
  - Sujet: *Quand la chimie éclaire le vivant*
  - Création d'un stand
  - Vulgarisation scientifique adaptée à différents niveaux (3ème - industriels)
- **Participation dans la communauté CAPRI (Critical Assessment of PRediction of Interactions)**
- **Création et animation de séminaires**

CONFÉRENCES

- *JOBIM 2022 (Journées Ouvertes en Biologie, Informatique et Mathématiques)* - Rennes : Présentation orale
- *Journée André Verbert 2021* - Colloque des doctorants
- *EMBO2021 Course: Integrative modelling of biomolecular interactions.* Izmir (Zoom): Flash talk + Poster
- *JOBIM 2019* - Nantes: Poster
- *JOBIM 2018* - Marseille: Poster

FORMATIONS

Cytoscape for the visualisation and analysis of biological networks

2 jours

Python - Module initiation

3 jours

Python - Module perfectionnement

3 jours

GTC NVIDIA

Génomique, Drug discovery et NLP

4 jours

Pratique éthique du métier de chercheur

1 jour

Révéler le potentiel entrepreneurial des doctorants

2 jours

Formation aux métiers du conseil

2.5 jours

VIE ASSOCIATIVE

BioAddoct
Association des docteurs et doctorants en Biologie Santé

Président

Janv 2020 – Janv 2022

ASSEMBR
ASSociation Étudiante du Master de Bioinformatique de Rouen

Président

Juil 2017 – Juil 2019

LOISIRS

Musique

Guitare
Piano
Banjo

Sport

Handball
Running
Randonnée

Jeu de Rôle

Maître du jeu

MOBILITÉ

Permis B + Voiture

Abstract

Interactions between proteins are one of the foundations of the development of life and their identification and understanding are still major elements of fundamental and applied research. In this context, the focus is on post-translational modifications of proteins that can alter their efficiency and lifetime. In addition, specific interactions between proteins can now be studied at the atomic level thanks to the development of experimental methods for solving the structures of protein complexes. However, these methods still do not always provide the expected results and their cost, whether financial or in terms of time, may prevent the understanding of certain phenomena, particularly during the emergence of a health crisis such as COVID-19. This is why, in parallel, computational methods such as molecular docking or molecular dynamics have been developed. This thesis is situated in these two contexts: firstly, the prediction of *O*-GlcNAcylation sites, a post-translational modification, catalyzed by a single enzyme called OGT, which has been extensively studied and implicated in different diseases such as cancer, Alzheimer's disease and type 2 diabetes. Secondly, in the context of COVID-19, interactions between human and viral proteins were highlighted through a world-wide study, in which the CAPRI protein docking experiment proposed several of these interactions to expert modelers of protein complexes in order to better understand the mechanisms of COVID-19.

The prediction of *O*-GlcNAcylation sites is not a new research field, as some tools for this type of prediction already exist. We have created a new data set, in order to compare and differentiate these. As the different algorithms consistently showed too many false positives, we developed an improvement based on a larger dataset but also on structural characteristics. However, the results still show too much heterogeneity to allow a safe prediction. Additional results support the theory that chaperone proteins are required for the enzyme to recognise its substrate. In order to better understand the mechanisms of this modification, the interaction between beta-catenin and OGT was specifically studied. This interaction has been shown to be involved in colorectal cancer and is therefore of particular interest.

To establish the veracity of the proposed models for the interactions between the human and SARS-CoV-2 proteins, a method based on the consensus of all the models produced was developed. Initial test results showed this method to be effective. We therefore tested its predictive capacity on a new and larger dataset provided by CAPRI. Once again, the developed method showed good results. It was then compared with pre-existing scoring algorithms on a similar benchmark and demonstrated improved results. The method also showed that the interaction models between viral and human proteins are not as reliable as desired.

Keywords: protein-protein interactions, O-GlcNAcylation, prediction, modeling, COVID-19, CAPRI

Résumé

Les interactions entre les protéines sont l'une des bases du développement de la vie. Leur identification et compréhension sont toujours des éléments majeurs de la recherche fondamentale et appliquée. Dans cette optique, on s'intéresse aux modifications post-traductionnelles des protéines qui ont la capacité d'altérer leur efficacité et leur durée de vie. Les interactions spécifiques entre protéines sont désormais étudiées au niveau atomique grâce au développement des méthodes expérimentales pour résoudre des structures de complexes protéiques. Cependant, ces méthodes ne permettent toujours pas d'obtenir les résultats escomptés et leur coût, que ce soit financier ou en termes de temps, peut empêcher la compréhension de certains phénomènes, notamment lors d'émergence de crise sanitaire comme le COVID-19. C'est pourquoi, en parallèle, des méthodes informatiques telles que l'amarrage moléculaire ou la dynamique moléculaire ont été développées. Cette thèse se situe dans ces deux contextes: dans un premier temps, la prédiction de sites de *O*-GlcNAcylation, une modification post-traductionnelle, catalysée par une seule enzyme appelée OGT, très étudiée qui est impliquée dans différentes maladies telles que le cancer, la maladie d'Alzheimer et le diabète de type 2. Dans un second temps, et ceci dans le contexte du COVID-19, des interactions entre les protéines humaines et virales ont été mises en avant mais avec la montée rapide de cas d'infection et les méthodes expérimentales étant trop longues, une expérimentation mondiale appelée CAPRI a proposé plusieurs des ces interactions aux modélisateurs du monde entier.

La prédiction de sites de *O*-GlcNAcylation n'est pas une recherche récente car des outils proposent déjà cette possibilité. Afin de les comparer, une base de données a été créée pour les différencier. Comme les différents logiciels montraient un trop grand nombre de faux positifs, une amélioration basée sur cette plus grande base de données mais aussi sur des caractéristiques structurelles a été proposée. Malgré cela, les résultats montrent une trop grande hétérogénéité pour permettre une prédiction sûre. Des résultats supplémentaires appuient la théorie du besoin de protéines auxiliaires pour permettre à l'enzyme la reconnaissance de son substrat. Afin de mieux comprendre les mécanismes de cette modification, l'interaction entre la beta-caténine et l'OGT a été étudiée

spécifiquement. En effet, cette interaction a été montrée comme étant impliquée dans le cancer colorectal et révèle donc un intérêt particulier.

Pour établir la véracité des modèles proposés pour les interactions entre les protéines du SARS-CoV-2 et de l'humain, une méthode basée sur le consensus de tous les modèles produits a été développée. Au vu des premiers résultats, cette méthode semblait performante. C'est pourquoi sa capacité de prédiction a été testée sur une nouvelle grande base de données, fournie par CAPRI. Une fois encore, la méthode développée a montré de bons résultats. Elle a ensuite été comparée aux logiciels de scoring actuels et montre ici de meilleurs résultats. Hélas, cette méthode montre que les modèles d'interaction entre les protéines virales et humaines ne sont pas aussi fiables que souhaités.

Mots clés: intéractions protéine-protéine, O-GlcNAcylation, prédiction, modélisation, COVID-19, CAPRI

Abbreviations

A

A: Alanine
Å: Ångström

Acc: Accuracy

AF: AlphaFold

AFM: AlphaFold-Multimer

AO: Adjacency Overlap

AP-MS: Affinity-Purification–Mass Spectrometry

APC: Adenomatous Polyposis Coli

API: Application Programming Interface

ASA: Accessible Solvent Area Curve

AUC: Area Under

B

β-TrCp: β-transducin repeat containing protein

C

C: Cysteine
CADD: Computer-Assisted Drug Discovery
CAPRI: Critical Assessment of PRedicted Interactions
CASP: Critical Assessment of protein Structure Prediction
CAZy: Carbohydrate-Active enZymes
CCDS: Consensus CDS
CDI: Catalytic domain I
CDII: Catalytic domain II
CDS: CoDing Sequence
CK1α: Casein Kinase 1 α
COVID-19: CoronaVirus Disease 2019
CRC: ColoRectal Cancer
CTNB1: Catenin beta-1
Cryo-EM: Cryogenic Electron Microscopy

D

D: Aspartic acid
D-box: Destruction box
DFT: Discrete Fourier Transformation
DNA: DesoxyriboNucleic Acid

E

E: Glutamic acid
EXOSC8: Exosome component 8

F

F: Phenylalanine
FDR: False Detection Rate
FN: False Negative
fP: False Positive
FFT: Fast Fourier Transformation
Fru: Fructose
FTIR: Fourier-Transform InfraRed spectroscopy

G

G: Glycine
GBM: Gradient Boosting Machine
GBT: Gradient Boosting Tree
GDP: Guanosine DiPhosphate
GlcNAc: *N*-AcetylGlucosamine
Glu: Glutamic acid
GNN: Graph NEural Network
GSK3β: Glycogen Synthase Kinase 3β
GT: Glycosyl Transferase
GTP: Guanosine TriPhosphate

H

H: Histidine
HBP: Hexosamine Biosynthesis Pathway
HCF-1: Host Cell Factor-1

HK: HexoKinase

HMOX1: Heme oxygenase 1

I

I: Isoleucine
I2H: *In silico* 2 Hybrid
Int-D: Intermediate Domain

K

K: Lysine

L

L: Leucine

M

M: Methionine
MCl: Markov Clustering
MD: Molecular Dynamics
MERS: Middle Eastern Respiratory Syndromes
Met: Methionine
MS: Mass Spectrometry
MSA: Multiple Sequence Alignment

N

N: Asparagine

NMR: Nuclear Magnetic Resonance
NN: Neural Network
Nsp: Non-structural protein
NUTF2: NUclear Transport Factor 2

O

OGA: *O*-GlcNAcase
***O*-GlcNAcylation**: *O*-linked β-*N*-acetylglucosamine
OGT: O-linked N-acetylglucosamine transferase
mOGT: mitochondrial OGT
ncOGT: nucleic and cytosolic OGT
sOGT: small OGT
ORF3a: ORF3a protein

P

P: Proline
P: Phosphate
PCA: Principal Component Analysis
PCAs: Protein-fragment Complementation Assays
PDB: Protein Data Bank
PDBe: european Protein Data Bank
PGM: PhosphoGlucoMutase
PhD: Doctor of Philosophy
PMIDS: PubMed IDs
PTM: Post-Translational Modification
PPI: Protein-Protein Interaction
PPV: Positive Predictive Value
Pr: Precision

Q

Q: Glutamine

R

R: Arginine
RF: Random Forest
RINs: Residue Interaction Networks
RhoA: Transforming protein RhoA
RMSD: Root Mean Square Deviation
iRMS: interface RMSD
lRMS:ligand RMSD
sRMS: side-chain RMSD
RNA: RiboNucleic Acid
ROC: Receiver Operating Characteristic

S

S: Serine
SARS-CoV-2: Severe Acute Respiratory Syndrome CoronaVirus 2
SAS: Surface Solvent Accessibility
SAXS: Small-Angle X-ray Scattering
SBDD: Structure-Based Drug Design
SBVS: Structure-Based Virtual Screening
SCM: Side-Chain center of Mass
Ser: Serine
Sn: Sensitivity
Sp: Specificity
SVM: Support Vector Machine

T

T: Threonine
TAP: Tandem Affinity Purification
TCF: T-Cell Factor
TN: True Negative
TNR: True NEgative Rate
TP: True Positive
TPR: TetratricoPeptide Repeats
TPR: True Positive Rate

U

UDP: Uridine DiPhosphate
USD: United States Dollar

V

V: Valine

W

W: Tryptophan
WHO: World Health Organization
Wnt: Wingless integration

Y

Y: Tyrosine
Y2H: Yeast 2 Hybrid

Table of Figures

List of tables

I. General introduction

A. Proteins

Proteins are biomolecules which participate in a variety of functions in life. They are composed of a chain of amino acids bonded by peptidic linkages at the output of ribosomes. The ensemble of all proteins is called the proteome, and it corresponds to transcription of the information contained inside the genome. The residue chains are considered a protein if the size is above 20-30 residue long, under this value, they are called peptides. Proteins are defined by their sequences, which are themselves defined by the coding sequences of the DNA. DNA sequences can be divided in codons, which are sets of 3 nucleotides that together encode one residue. As there are 4 codons, 64 combinations are possible. In nature, 20 residues can be found meaning that multiple codons can code for the same amino acid.

Proteins are considered to have up to 4 different levels of structure: the first one is the amino acid sequence, the secondary structure is compound of different possibilities: helices, sheet, loop or coiled part determined by the backbone torsion angles in the amino acid residues, the tertiary structure also called three-dimensional structure is the native form of the structure in the organism and the last one which is not possible for every protein is called the functional assembly and is the interaction between at least two three dimensional proteins (see Figure 1). The function of a protein is related to its structure.

Secondary structures will play a role in the final form of the protein and so in its function. The main helices found in protein structures are alpha helices; the form of helice is due to the protein backbone angles and the name of the helice depends on the number of residues found in one turn (3.6/3.7 amino acid per turn for an α-helix). These helices can have two opposite directions, the most common is the clockwise turn and the less frequent is the anticlockwise. This structure is stabilized by the hydrogen bonds between the carbonyl group and the amide group of the amino acid but also by van der Waals forces[1]. The other most common secondary structure is beta-sheet (β-sheet) which corresponds to the alignment of adjacent amino acid chains stabilized by hydrogen bonds. β-sheets can be divided into two categories: parallel and antiparallel depending on the orientation of the

polypeptide chains. The antiparallel β-sheet is found more often than the parallel one and the reason may be that the stability is higher thanks to the straight hydrogen bonds. The most common number of chains is around 6 strands and the distance between two chains is generally around 7 Ångström (Å) and the length of one plate is between 6 and 15 residues. The turn between the chains is called beta turn and is composed of 4 residues including proline, the only residue with a particular backbone. It allows a rotation thanks to the linkage between the carbonyl group and the amide group of the first and fourth amino acids respectively. This turn is part of what is called a loop [2]. Loop structures are parts of protein joining the other secondary structures (helices and beta-strands) with a length of 2 to 6 residues. These loops are responsible for the change of the chain direction and give the three-dimensional shape of the protein. The last category is the coil structure and it is not considered as a true secondary structure but more like a conformation. It is part of a protein with no secondary structure such as alpha-helix or beta-sheet but coil structure is participating in the stable conformation. Some parts of the coil region have a disordered region called random coil structure and are known to have a major role in ligand recognition or binding [3].

At the output of the ribosomes, some proteins can start to structure themselves and then be functional. Others will be modified by Post Translational Modifications (PTMs) in different cell compartments/organel such as golgi apparatus and the endoplasmic reticulum but also can be modified during their life inside a cytosol. These modifications will have an impact on the structure and therefore the activity of the protein right after the transcription or during the folding but also later in the protein life according to the cell cycle. Proteins are the living molecules inside the cell compared to the other molecules which are inert. They can interact with other molecules as protein, peptide, DNA, RNA, lipids or carbohydrates. With this capacity of interacting, proteins are involved in many biological processes and can have activating, inhibitory or enzymatic activities, amongst others as signaling and transport.

As their structure is important for their functions, knowing it would help understanding their mechanisms. To that, different experiments have been developed through the years, starting with X-ray crystallization in the beginning of the second half of the twentieth century, by Max Perutz and Sir John Cowdery Kendrew [4]. Since then

experimental methods have been improved and new methods have been developed (NMR, FTIR) and we are now able to perform cryo-electron microscopy on large macromolecular assemblies [5,6]. In parallel, as protein structures are not always easy to obtain, alternative methods as *in silico* experiments have emerged to predict structural information of proteins such as secondary, tertiary and quaternary structure but also surface solvent accessibility (SAS) and flexibility. Thanks to technological progress, these methods have also improved a lot lately and allow better prediction.

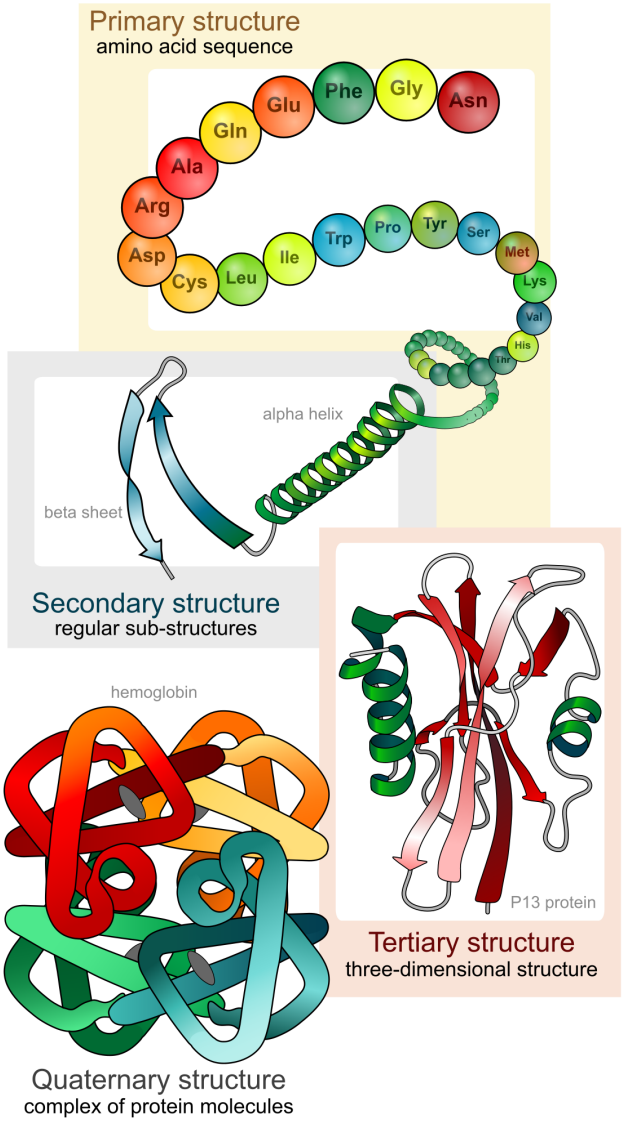


Figure 1. **Representation of the four levels of protein's structure**

(source: Wikipedia)

To answer the need for prediction of secondary structure, many software have been proposed, mostly based on protein sequences and now with the larger amount of data with machine learning. It is the case of the PSI-PRED and SPIDER3 algorithm which can predict

secondary structure, the dihedral angles and even the Accessible Solvent Area (ASA). These software help predict the tertiary structure of proteins and are often found in pipelines for three-dimensional structure prediction. With the improvement of secondary structure prediction, came best performance for 3D prediction, using template or *ab initio* algorithms. These algorithms used information from already resolved structures but also some co-evolution signals.

To evaluate the performance of the protein structure prediction an experiment was created in 1994 called CASP for Critical Assessment of protein Structure Prediction. Nowadays, CASP experiments are at a number of 15 rounds and show a recent major breakthrough in the protein 3D structure prediction. At the beginning of such prediction, the improvement of algorithms made a big jump, then with the improvement of experiments and the growing number of structures and so of data, machine learning brought a next generation of predictors. The top 10 servers according to the last CASP experiment from 2020 are the following: BAKER-ROSETTASERVER, *QUARK*, *ZHANG-SERVER* (also known as I-TASSER), RAPTORX, FEIG-S, TFOLD, T-FOLD-IDT, *ZHANG-CETHREADER*, *ZHANG-TBM* and TFOLD-CAT [7]. All of these servers are using deep learning and most of them are from the same lab (*Zhang* and Tfold), some of their algorithms are explained in Table 1. AlphaFold2 was not, at this moment, available as a server.

Name (type of algorithm)	Summarized protocol
QUARK (Deep learning)	Initially D-QUARK, it has been upgrade with 4 features:(i) a new MSA collection tool, DeepMSA2; (ii) a contact-based domain boundary prediction algorithm, FUpred; (iii) a residual convolutional neural network-based method, DeepPotential; (iv) optimized spatial restraint energy potentials to guide the structure assembly simulations
ZHANG-SERVER (Deep learning)	Initially I-Tasser, it has been upgrade with the same 4 features as QUARK:(i) a new MSA collection tool, DeepMSA2; (ii) a contact-based domain boundary prediction algorithm, FUpred; (iii) a residual convolutional neural network-based method, DeepPotential; (iv) optimized spatial restraint energy potentials to guide the structure assembly simulations

RAPTORX (Deep learning) [8]	First distance and contact prediction with ResNet, then MSA with 3 iterations of HHblits and 1 of Jackhmmer. Then if it template-based modeling, the RaptroX-TBM is used with DeepThreader otherwise it is RaptorX-DeepModeller which uses contact prediction, alignment and coevolutionary information through Deep Learning

Table 1. **Summarized protocols of 3 deep learning server predictors**

These results showed the improvement of deep learning algorithms for protein structure prediction. Indeed, with the Google DeepMind software called AlphaFold2 using deep learning on multiple alignment matrices with coevolution signals, structure prediction has been made easily available with good results [9,10]. AlphaFold publication was released in 2020 and showed very good results during the CASP13 (2018) edition and better results during CASP14 (2020) with their second version of AlphaFold (AlphaFold2) compared to other predictors as it can be seen on Figure 2.

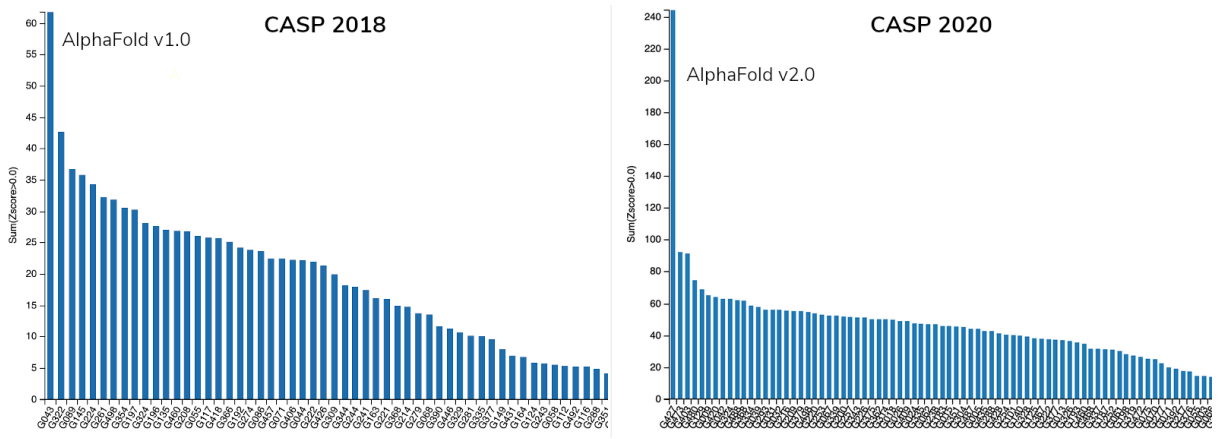


Figure 2. **Rankings for the 2018 and 2020 CASP competitions**

The main information here is that AlphaFold outperformed the other predictor groups. (From Foldit)

B. Protein-protein interaction

Interactions between proteins are known to be involved in many biological processes, from life development regarding the cell cycle with signal transduction to control and promote cell division. But these interactions can also be found in host-pathogen relations and through various diseases such as cancer [11,12]. Protein function can be regulated

by other proteins but also modified through post-translational modifications such as phosphorylation or glycosylation. Their function can also be regulated by peptides or small ligands and the need to highlight them is important to understand the different pathways but also to find interaction inhibitors or modulators in case of associated disease pathologies.

Even if the interactions between proteins can be predicted or found by experimental analyses, most of the time the results will be binary indicating if a protein A will interact with protein B but without giving any details. As no detail is given, the process on how the protein interacts is still unknown and there is a need to find where this interaction occurs while scanning the accessible surface area for example (ASA). If there is a need to understand the mechanism or if the interaction is a problem and finding an inhibitor is needed like in drug design/development then having a more detailed interaction even at an atomic level can be necessary. To this, scientists have developed a lot of experimental techniques such as X-ray crystallography, tandem affinity purification, affinity chromatography, coimmunoprecipitation, protein arrays, protein fragment complementation, phage display and NMR spectroscopy [13] .

As many experimental techniques exist to highlight or find specific Protein-Protein Interaction (PPI), they are still time and money consuming. A relatively cheap alternative is to predict these interactions by computational means [14]. Indeed, even if they are predictions, some results are not possible to obtain experimentally yet, and these theoretical results can help to be close to reality. Predictive results can be discussed with a confidence score, representing the knowledge on which it is based and easily interpretable by humans, or algorithms alike. These *in silico* methods are to be seen complementary to experimental ones. If the latter are more accurate according to their conditions, the computational algorithms win time, orient research, find new approaches and other.

An extensive list of methods to detect PPI is listed in Table 2 1.

Approach	**Technique**	**Summary**

In vitro	Tandem affinity purification-mass spectroscopy (TAP-MS)	TAP-MS is based on the double tagging of the protein of interest on its chromosomal locus, followed by a two-step purification process and mass spectroscopic analysis
	Affinity chromatography	Affinity chromatography is highly responsive, can even detect weakest interactions in proteins, and also tests all the sample proteins equally for interaction
	Coimmunoprecipitation	Coimmunoprecipitation confirms interactions using a whole cell extract where proteins are present in their native form in a complex mixture of cellular components
	Protein microarrays (H)	Microarray-based analysis allows the simultaneous analysis of thousands of parameters within a single experiment
	Protein-fragment complementation	Protein-fragment complementation assays (PCAs) can be used to detect PPI between proteins of any molecular weight and expressed at their endogenous levels
	Phage display (H)	Phage-display approach originated in the incorporation of the protein and genetic components into a single phage particle
	X-ray crystallography	X-ray crystallography enables visualization of protein structures at the atomic level and enhances the understanding of protein interaction and function
	NMR spectroscopy	NMR spectroscopy can even detect weak protein-protein interactions
In vivo	Yeast 2 hybrid (Y2H) (H)	Yeast two-hybrid is typically carried out by screening a protein of interest against a random library of potential protein partners
	Synthetic lethality	Synthetic lethality is based on functional interactions rather than physical interaction
In silico	Ortholog-based sequence approach	Ortholog-based sequence approach based on the homologous nature of the query protein in the annotated protein databases using pairwise local sequence algorithm
	Domain-pairs-based	Domain-pairs-based approach predicts

sequence approach	protein interactions based on domain-domain interactions
Structure-based approaches	Structure-based approaches predict protein-protein interaction if two proteins have a similar structure (primary, secondary, or tertiary)
Gene neighborhood	If the gene neighborhood is conserved across multiple genomes, then there is a potential possibility of the functional linkage among the proteins encoded by the related genes
Gene fusion	Gene fusion, which is often called a Rosetta stone method, is based on the concept that some of the single-domain containing proteins in one organism can fuse to form a multidomain protein in other organisms
In silico 2 hybrid (I2H)	The I2H method is based on the assumption that interacting proteins should undergo coevolution in order to keep the protein function reliable
Phylogenetic tree	The phylogenetic tree method predicts the protein-protein interaction based on the evolution history of the protein
Phylogenetic profile	The phylogenetic profile predicts the interaction between two proteins if they share the same phylogenetic profile
Gene expression	The gene expression predicts interaction based on the idea that proteins from the genes belonging to the common expression-profiling clusters are more likely to interact with each other than proteins from the genes belonging to different clusters
Protein-protein modeling	The protein-protein modeling predicts atomic interaction between two proteins based on template or with a molecular docking or hybrid method

Table 2 **Summary of PPI detection methods**

(Adapted version from Rao *et al.* (2014))

For the rest of this manuscript, the methods used and developed are limited to protein-protein and peptide-protein docking, and PTM prediction based on sequences.

1. Protein-protein interaction for Post Translational Modifications

Post Translational Modifications (PTMs) are modifications which occur on proteins after their translation at the output of the ribosome. These interactions are known to be important for the cell as they may regulate protein activity and folding [15]. PTMs increase the complexity of the proteome by expanding the proteins' functionalities. Nowadays, we estimate a diversity of modification superior to 300, localized on 15 proteinogenic residue side chains or protein backbone [16]. These modifications can occur in the cytosol, in the nucleus, but mainly occur in the Endoplasmic Reticulum and the Golgi organelle. The PTM machinery is a crosstalk of enzymes which control the modifications, thereby defining their cellular function [15]. They are made possible thanks to enzymes which modify protein through their catalytic activities. But some PTMs are enzyme free: glycation, carbamylation, carbonylation and the spontaneous isopeptide bond formation [17–20]. In general, enzymes can be classified in three different categories: "Writers", "Erasers" and "Readers". The first category corresponds to enzymes which transfer modifying groups on the side chain of residues like kinases or glycosyltransferases which catalyze the addition of a phosphate or sugar group, respectively. This addition can be done onto a previously attached group, leading thusly to linearly extended or even branched polymers. The second category, "eraser" enzymes like phosphatases are proteins which catalyze the removal of these modifications on the substrate. The final group, "reader" enzymes, are often proteins that transduce a PTM-dependent function with high affinity to specific PTMs. The PTM crosstalk is summarized in Figure 3 from Leutert *et al.*, 2021 [15].

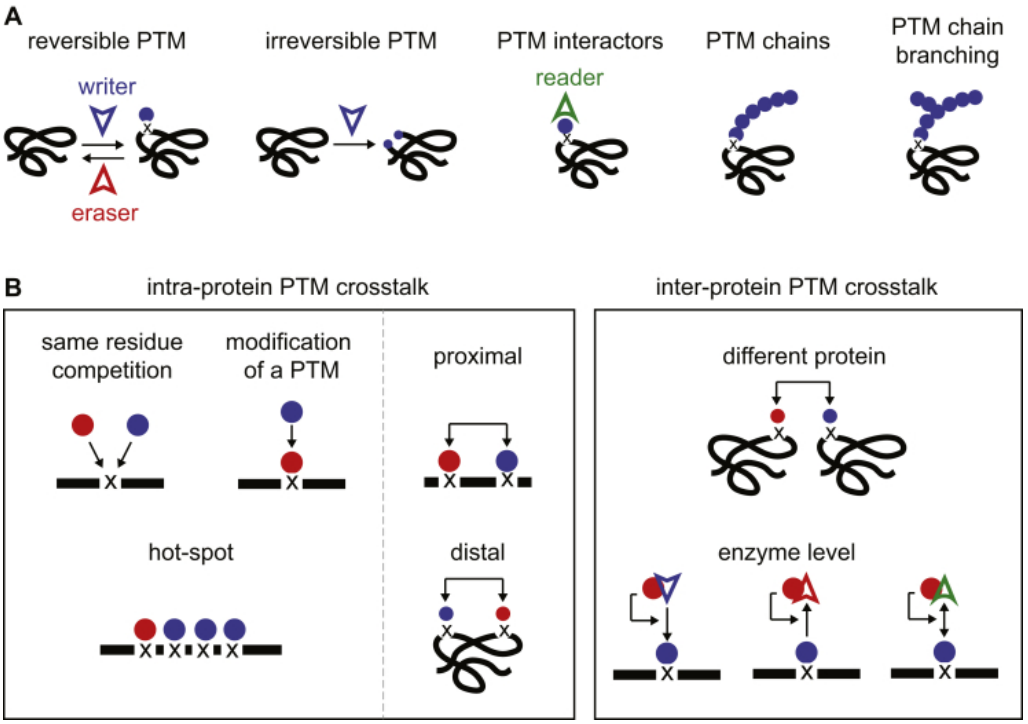


Figure 3**. The Toolbox for PTM crosstalk**

A, the PTM machinery includes proteins that write, erase, and read the PTM. PTMs come in reversible and non-reversible forms, as monomers, polymers, and branched polymers.
B, different modes of PTM crosstalk are separated based on intra- or inter-protein crosstalk. Proteins are illustrated in black, different PTMs as blue and red circles, the modification site is depicted as "x."
(from Leutert et al., 2021)[15]

These different modifications induce the need of the enzyme to interact with at least their substrate.

2. Phosphorylation: kinase interaction

Phosphorylation is one of the most studied PTMs. Involved in many cellular activities like cellular signaling, apoptosis, cell growth and differentiation[21], its deregulation is often found in diseases like cancers. Indeed, the tyrosine kinase family includes the greatest number of oncoproteins. This modification is catalyzed by two kinds of enzymes; kinases and phosphorylases, which respectively add or remove a phosphate moiety on the hydroxyl group of serines, threonines and tyrosines [22]. This modification leads to a modification of the substrate conformation or localization and even activates or deactivates it [23,24]. The number of kinases and phosphatases is very high and has been estimated to correspond to 2 to 5% of the total human genome. This large amount of enzymes can be explained by the high number of proteins that can be phosphorylated. In fact, thousands of sites can be modified

by phosphorylation, and it is estimated that they represent up to 30% of the proteome [23]. The multiple kinases can be grouped into families and subfamilies, according to which substrate is recognized. The specific interaction between a kinase and its substrate depends on specific patterns. Various *in silico* algorithms attempt to predict the phosphorylation site based on these different kinase families using a plethora of different algorithms[25]. This modification is (often) in competition with another PTM called *O*-GlcNAcylation (*O*-linked β-*N*-acetylglucosaminylation) [24,26].

3. *O*-GlcNAcylation: one enzyme to modify them all

O-GlcNAcylation consists of the addition of a *N*-acetylglucosamine (GlcNAc) moiety onto hydroxyl group of serines or threonines of nucleus and cytosolic proteins [27]. The nucleotide sugar is derived from the glucose through the Hexosamine Biosynthesis Pathway (HBP) shown on Figure 4 [28]. This pathway is nutrient dependent and as such, the activity of this modification can be over-regulated by (mal)-nutrition and lead to various diseases as Alzheimer' disease, diabetes or cancers [29]. The addition of the GlcNAc is in competition with phosphorylation as they target the same residue[26]. But, unlike phosphorylation, *O*-GlcNAcylation is catalyzed by only two enzymes: the OGT (*O*-GlcNAc Transferase, which attaches an *O*-GlcNAc group) and OGA (*O*-GlcNAcase, which removes an *O*-GlcNAc group). The first enzyme is composed of two main domains: a catalytic domain and a recognition domain called TPR Domain (TetratricoPeptide Repeats). The catalytic domain first recognizes the nucleotide sugar and attached it in its pocket, then once the substrate is available, the GlcNAc moiety will be added by the action of the histidine 498 [30]. The second domain is a supra helix domain involved in substrate recognition. The number of TPR will depend on the isoforms. Indeed, this enzyme exists in three different isoforms: ncOGT for nucleus and cytosolic OGT, sOGT stands for small OGT and the mOGT for mitochondrial OGT. All exhibit the same catalytic domain, but differ in the number of TPRs: 13.5 for ncOGT (nucleus and cytosolic OGT), 2.5 for sOGT (small OGT) and 9.5 for mOGT (mitochondrial OGT) [31]. The enzyme responsible for the removal of the nucleotide sugar is the OGA which recognizes the GlcNAc and removes it through hydrolysis [32]. Today, we hypothesize that thousands of proteins are *O*-GlcNAcylated and this modification has been shown to be involved in various diseases such as cancer, diabetes and Alzeihmer's disease [33].

The need to have a better understanding of *O*-GlcNAcylation even when effectuated by only a single enzyme is still relevant. This is largely due to the fact that only few of the possible *O*-GlcNAcylation sites are in fact *O*-GlcNAcylated. As the number of proteins estimated to be *O*-GlcNAcylated is huge, the exact sites are mostly unknown. The experimental methods to highlight the sites are time and money consuming so the need to develop a *O*-GlcNAcylation prediction software has risen. Many labs have tried and claim to have a good sensitivity while having based their algorithms on protein sequences [34–36]. But the mechanism behind the substrate is still unknown even if an asparagine ladder has been shown to be important in the OGT activity [37]. There are hypotheses that *O*-GlcNAcylation needs scaffold proteins to be able to recognize target substrates.

Also, OGT has been identified as a protease, in particular in the case of Human HCF-1 where a region called HCF-1_{PRO} repeat has been shown to be a binding site for the OGT [38]. This particular region contains 26 amino acids with a conserved 20 amino acid core sequence shared among vertebrate species [39]. The proteolyse of the HCF-1 protein leads to two subunits (N- and C-), closely attached, involved in the regulation of distinct phases of the cell-division cycle [38].

OGT is able to interact with a very high amount of proteins, with different roles and even with different catalytic activities. But the recognition of the difference in such a number of different substrates is still an unresolved process.

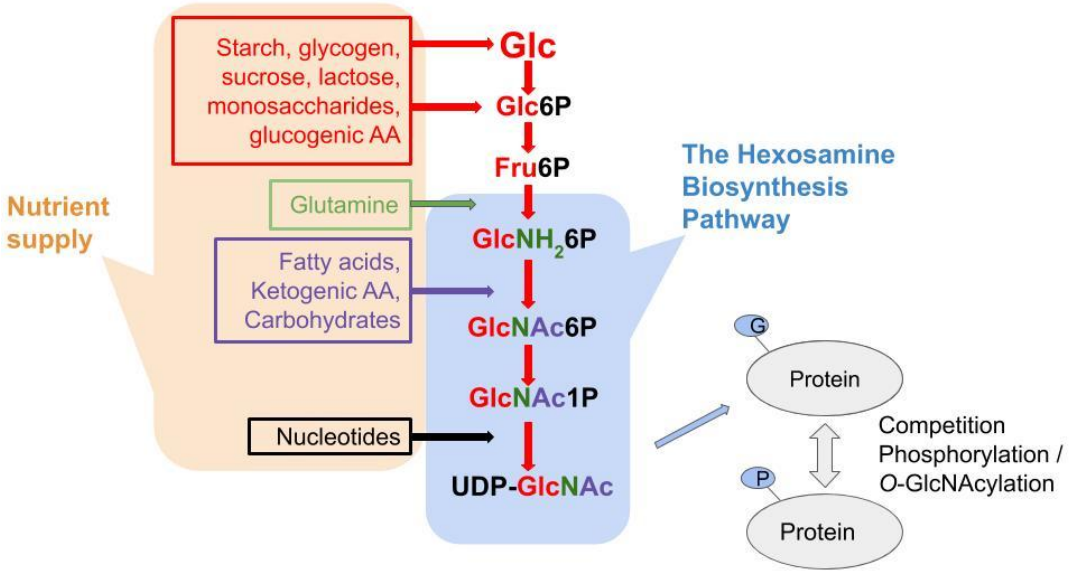


Figure 4. **Schematic representation of the HBP and *O*-GlcNAcylation/phosphorylation competition**

C. Prediction of protein-protein interactions

1. Docking and PPI modeling

a) Protein-protein docking and modeling

In silico predictions of protein-protein interactions became more and more important during the last years. Indeed, with the improvement of computational technologies in parallel with the high number of information retrieved with the new experiments leading to a better understanding of the important information implicated in protein-protein interactions. This information, including interaction surface prediction, contact prediction but also biochemical interaction and co-evolution of residues, led to a significant increase in the amount of data that could be used by predictors and their algorithms. Recently, the interest of big companies such as Google to enter the world of protein-protein interaction prediction created a breakthrough showing the power of machine learning or precisely deep learning when such amount of data is available. Unfortunately, this major step still needs to be improved because protein-protein interactions are not only complex structures but also need additional information such as modifications like glycosylation or interface prediction.

The protein-protein interaction simulations start with finding the putative interface and then dock the two molecules together. Starting in the early 1970s, the prediction of protein-protein interactions has been developed with bioinformatics by predicting interaction based on sequence similarity. This was followed by solvent accessibility calculations to identify the potential interaction surface with PDB structures studied in 1999 by Lo Conte *et al.* [40,41]. In parallel, an alternative approach had been developed, relying on atomic packing analyses. This method is based on the complementary surfaces forming compact interfaces with few cavities and close-packed atoms [40]. This theory used Voronoi calculations, proposed by Richards (1974) and Finney (1975) and applied by Janin and Chothia (1976) [42–44]. Voronoi calculations are the transformation of atoms to polyhedra to calculate covalent or noncovalent bonds between proteins. But considering the interface with Voronoi calculations reduces the accuracy because the atoms must be completely

surrounded and only one third of the residues contributing to the interface have zero accessibility to the surface.

The predicting and the simulating of protein-protein interaction by docking of proteins instead of small ligands was first considered by Wodak and Janin in 1978 [45]. A rigid body search of six degrees of freedom (five rotational and one translational) was used to bring the two molecules in direct contact. As it was very expensive in terms of the number of calculations needed, an approximation was made in 1980 where residues were modeled as balls.

Today, protein-protein docking remains one of the central and challenging problems in computational structural biology and many labs are working on developing new methods to predict such interactions [46]. With respect to the early days' calculations, immense progress can be observed.

In this paragraph, the most commonly used methods for protein structure prediction that are also being used for prediction of protein complexes will be briefly described. One of the main principles on which structure prediction is based is a minimization of the free energy of the system. For example, energy embedding has been introduced by Crippen in early 1980's [37]; this algorithm basically puts a conformation with a very low energy inside a high dimensional shape. This shape, which is a three-dimensional space, is then reduced. After the constraints are enlarged on the structure while keeping the energy minima[48]. A molecular conformation can be represented in Cartesian coordinates, or using a reduced space such as dihedral angles, but it can also be described by a squared matrix with *n* rows and columns containing the distance between all *n* atoms of a molecule. Protein force fields can use this matrix and calculate interaction energetics such as electrostatics, hydrogen bonds and bond stretching, to name but a few [48]. Usually the protein structure prediction energy is found using Monte Carlo technique which is not that recent because it was exposed for molecular docking in 1985 by T. Noguti and N. Go [49]. Indeed, the Monte Carlo theory shows that the protein structure will have random conformations like random walk and compare the structure energy between two conformations and select the one with the lowest energy and then proceed to go into increasingly lower-energy structures (see

Figure 5). But this kind of method can lead to a local minima which is not the lowest energy score conformation. Other algorithms have been created to minimize this energy like molecular dynamics (MD). This method is a computational simulation of the possible physical movements of atoms or particles given, for a fixed period, possibilities for atoms to interact with each other, becoming a simulation of a dynamic evolution. During this simulation, the energy is augmented thanks to a kinetic energy leading to more conformational spaces and many minimums. Most of the time, the forces between the atoms and their potential energies are calculated using interatomic potentials or molecular mechanics force fields. Molecular docking is a powerful tool to study the interaction between the receptor and the ligand at the molecular point of view. The different results are analyzed by different scoring functions. Software of molecular docking can be divided into two categories: flexible ligand search or flexible protein docking. First one usually uses three algorithms : systematic, stochastic and simulation methods. The second one uses Monte Carlo and molecular dynamics methods. Different software apply these different methods. According to Gurung *et al.* (2021), the most used tools are AutoDock[50], AutoDock Vina[51], GOLD[52], CDOCKER[53], FlexX[54], Surflex[55], GLIDE[56], DOCK6[57] and SwissDock[58].

To estimate the position of the atoms in interaction between ligand and receptor, molecular dynamics simulations use a suitable force field. This force field allows us to determine the overall energy of the complex system[59]. A non-exhaustive list of different MD simulation software exist and have been summarized in the Table 3 [60].

Software	Key Features	Simulation system
GROMACS[61]	GROMACS (Groningen MAchine for chemical simulation) is an efficient and versatile MD program with source code that is suited for the simulation of biological (macro) molecules in aqueous and membrane environments. The program can be run on single processors or parallel computer systems and is compatible with various force fields such as GROMOS, OPLS, AMBER, and ENCAD force fields.	Proteins, lipids, carbohydrate, nucleic acids
AMBER[62]	Amber is an extensively used biomolecular simulation program with an assembly of codes that are designed to work together. It is a collection of codes that are designed to work together and principally divided into three major step-system preparation (antechamber, LEaP programs), simulation (sander), and trajectory analysis (ptraj analysis program).	Proteins, nucleic acids, carbohydrates

CHARMM[63]	CHARMM (chemistry at HARvard molecular mechanics) is a widely used molecular simulation program that is primarily designed to study biological molecules such as proteins, peptides, lipids, nucleic acids, carbohydrates, and small molecule ligands. The calculations are based on different energy functions (quantum mechanical-molecular mechanical force fields, all-atom classical potential energy functions) and models such as explicit solvent, implicit solvent, and membrane models.	Proteins, lipids, carbohydrates, nucleic acids
NAMD	NAMD is a high-performance biomolecular simulation program that employs the prioritized message-driven execution capabilities of the charm+ +/converse parallel runtime system compatible with parallel supercomputers and workstation cluster	Proteins, lipids, carbohydrates, nucleic acids
Desmond (https://www.schrodinger.com/products/desmond)	Desmond is a powerful molecular dynamic simulation program designed by D. E. Shaw with considerable speed, accuracy, and scalability. It supports explicit solvent simulations with periodic boundary conditions and can be used to model explicit membrane systems under various conditions.	Proteins, lipids
Tinker[64]	Tinker is a molecular modeling and dynamic package written primarily in a standard Fortran 95 with OpenMP extensions. It supports a wide variety of classical molecular simulations particularly biomolecular calculations and offers various force fields including the modern polarizable atomic multipole-based AMOEBA model.	Proteins, nucleic acids
LAMMPS (https://lammps.sandia.gov)	LAMMPS (large-scale atomic/molecular massively parallel simulator) is a classical molecular dynamic code for materials modeling. It has potentials for soft matter (biomolecules, polymers), solid-state materials (metals, semiconductors), and coarse-grained or mesoscopic systems.	Proteins, lipids, carbohydrates, nucleic acids
DL_POLY[65]	DL_POLY is a general purpose molecular dynamic simulation package, which allows the study of liquids of large complexity. The code is developed using the replicated data (RD) parallelization strategy.	Membranes, proteins

Table 3. **A summary of commonly used molecular dynamic (MD) simulation software**

(from Gurung *et al.*, 2021)

As some proteins are really big, the interface surface can be very wide and so the contact with the solvent is larger which creates long range electrostatic interactions. Thus a method to simplify space search has been created called Fast Fourier Transform (FFT). This technique is based on the calculation of the Discrete Fourier transform (DFT) which consists of the transformation of a variable of its own dimension like space into a frequency [66]. The FFT will convert the DFT matrix into a product of sparse (matrices composed with a lot of zero) to increase the speed of calculus [67].

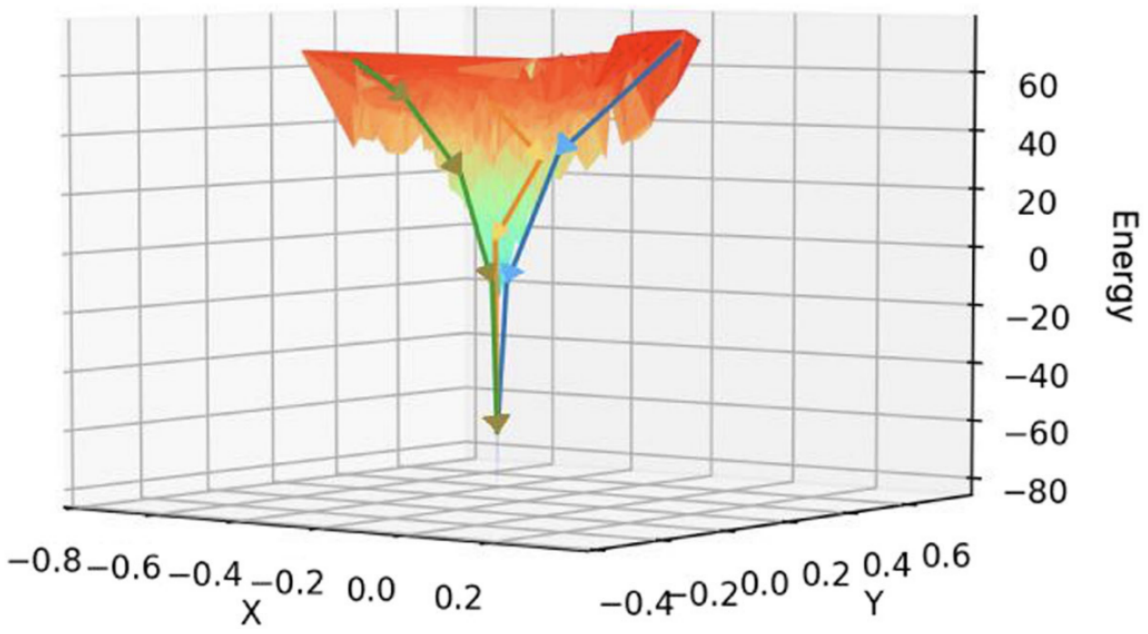


Figure 5. **Conformation searching process using the Monte Carlo technique**

The protein structure prediction approaches usually employ the Monte Carlo technique to search the conformation with the lowest energy. An execution of conformation search will generate a path of conformations, e.g., the lines in blue and yellow.

(from Wang *et al.* 2019 [68])

Nowadays, two main families of docking algorithms have been set up: template-based docking or *ab initio* docking (also call free docking) which uses the techniques described above (see Figure 6).

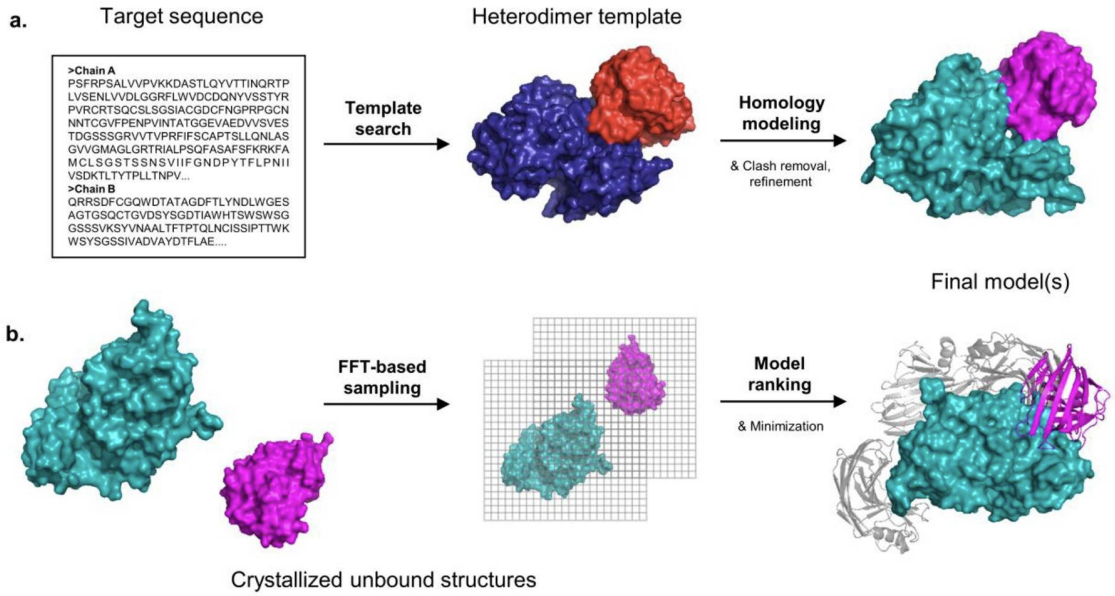


Figure 6. **General comparison of template-based and free docking methods for an example heterodimer target**

(a) A template-based method begins with the target sequences, using a template search to identify an existing heterodimer template from the PDB (Protein Data Bank). Homology modeling is used to map the target sequence onto the template structure. The model is then refined to its final form. (b) In free docking the two components must be individually crystallized. Millions of protein conformations are evaluated, often through the use of an FFT-based algorithm. Final models of the heterodimer are ranked and minimized.
(From Porter *et al.* 2019 [69])

The first one uses experimental structure knowledge available to orient the molecular docking if homologous complexes exist in public databases. This method is more and more powerful as the number of complex structures is increased each day in particular with the expansion of Cryo-EM [69]. Free docking only needs structure of the participants (if they don't exist they can be modeled but it will be a prediction based on prediction this reduces the overall reliability of the prediction).

To perform *ab initio* docking there are plenty of different methods and algorithms based on many features (see Table 4). Also along with all these methods, the docking software can be divided into two main classes: rigid and flexible-body docking. Rigid-body docking is the most frequent docking technique. Indeed, as torsion angles, bond angles and length are rigid, the calculations are less time and resources consuming. Flexible-body docking as their name suggests consists in adding flexibility inside docking. But this flexibility increases the number of degrees of freedom which results in a raise of processing but also the amount of false positive results [70]. In a publication of Desta *et al.* (2020), rigid-body docking (CLUSPRO) is shown to have more positive results (5/10) than flexible-body docking but the latter's results get a better quality (found as top 1 prediction model) [71].

Software	Main features	Protocols
MDockPP [72]	- FFT - ITScorePP - Flexibility - Physics-based potentials	Reduced-model FFT-based protein docking optimized by ITScorePP
ATTRACT [73]	- Coarse-grained representation - Normal modes	Docking using pseudo-atoms attractivity/repulsivity then minimization on normal

	- Molecular Dynamics	modes
HawkDock [74]	- ATTRACT - Clustering - MM/GBSA[75]	Docking with ATTRACT followed by scoring with ATTRACT score and HawkRank and then clustering and re-ranking. Possibility to get key residues
SwarmDock [76,77]	- Energy based - Normal modes - Swarm optimization	Docking using local docking and particle swarm optimization using normal modes
HDOCK [78]	- FFT - Shape-based pairwise scoring function	The server use meany third party software but can be only proceed to dock using FFT and home made scoring function
HADDOCK(-CG) [79,80]	- Energy minimization - Refinement - (Martini CG)	Random orientation followed by energy minimization-driven rigid body docking then semi flexible and flexible refinement. (MARTINI coarse-grained force field as option)
LZerD [81]	- Local 3D Zernike descriptors (3DZDs) - Surface features	Proteins are first transformed using 3DZDs into soft surface representation and then docked and complexes are clustered
CLUSPRO [82]	- Energy - Clustering - FFT	FFT sampling followed by clustering and CHARMM minimization according to different energies
pyDockWEB (and pyDock) [83,84]	- FFT - electrostatics - desolvation energy	Docking is performed using ASA-based solvation, Coulombic electrostatics and van der Waals energy
FRODOCK [85]	- Energy optimization - Rigid-body	First receptor is converted into grid maps, the a

	- Rotational search - Clustering	docking 6D search is performed (3D rotations + 3D translations) and then clustered
InterEvDock2 [86]	- Evolutionary and biological information	Docking using third party software as FRODOCK, then clustered using FRODOCK clustering and then scored using InterEvDock and SOAP-PP

Table 4. **Non exhaustive list of docking software with summarized features and protocols**

To be able to compare those prediction software/methods an experiment modeled on CASP called CAPRI (Critical Assessment of Predicted Interactions) has been set up. This experiment takes place in Rounds and will be explained in more detail later in this manuscript. In the last Round, twenty groups of modelers, including six web servers, participated [87]. In general, web servers are shown to have lower quality models than human groups with the exception of MDOCKPP and LZERD which obtained results on par with human groups [72,88]. According to this CAPRI Round, the best methods are the one developed on Baker, Venclovas and Seok predictor groups (see Figure 7).

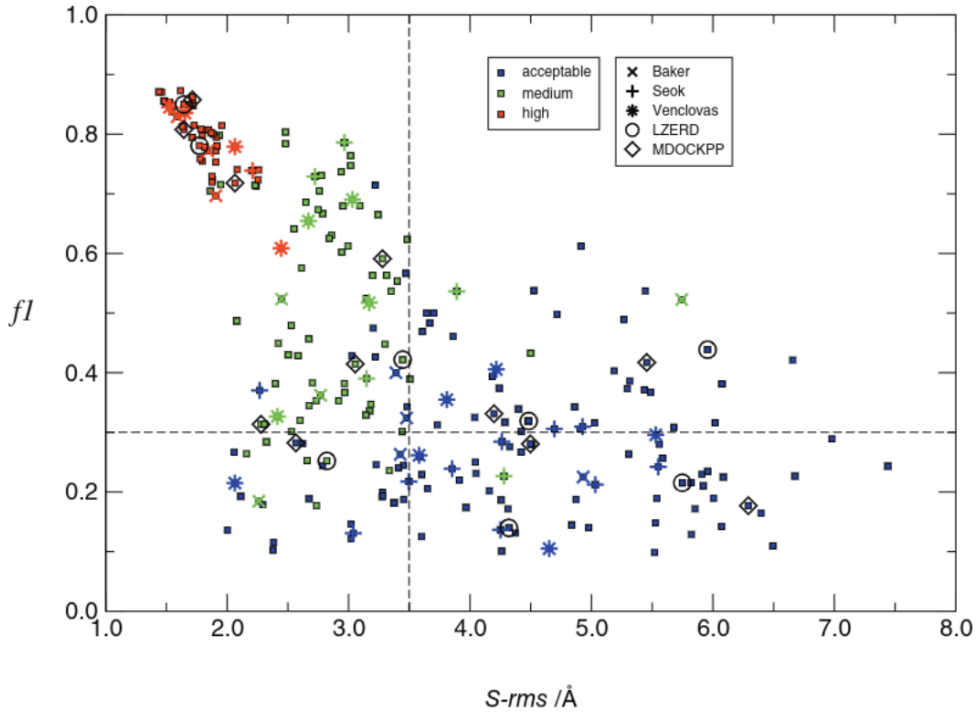


Figure 7. ***f1* as a function of *S-rms***.

Each point in the figure represents the best model of a predictor group for each of the 23 interfaces. Individual points are color-coded following the CAPRI model quality as follows: yellow: incorrect; blue: acceptable; green: medium; red: high. The results for the best predictors (Baker, Seok, and Venclovas) and servers (LZERD, MDOCKPP) are highlighted. *f1* and *S-rms* are respectively a function of the recall, and precision in modeling the residue-residue contact at the binding interface and the root mean square deviation of sidechain atoms of residues at the binding interface. The upper left quadrant features the best models, with *S-rms* values below 3.5 Å and *f1* values above 0.3, corresponding to mostly medium and high-quality models
(from Lensink *et al.* 2021[87])

Another method to be able to assess the quality of interaction prediction structure is the DockQ score which is a combination score. It combines scores also used by CAPRI called f_{nat}, i-rms and s-rms into a score between 0 and 1 [89].

Baker's research to predict structure is based on amino acid co-evolution and multiple sequence alignment of homologous proteins used by machine learning. Their software, called RoseTTAFold, is also able to predict *de novo* design proteins from sequence only [90]. This method combined with AlphaFold showed better results than RoseTTAFold or AlphaFold alone to predict protein complexes. But as machine learning needs experimental results to train, the development of cryoelectron microscopy to obtain large assemblies at high resolution will allow better and better results.

Venclovas' lab has been using a workflow depending on the context knowledge for the CASP14-CAPRI round, varying according to the presence of templates for the interaction, partial templates or no templates at all [91]. When a template is found, comparative modeling is performed; otherwise, a template is searched by profile-profile identification. If none of this method works, DALI is performed to make queries for PDB searches [92]. Once a multimeric template is found by one of these methods, the complex is generated by a multichain modeling [93]. The template-based docking is performed when the template is available, or if a protein has similar annotations. In that case, the chains are modeled on the template and then relaxed to remove steric clashes using the same methods as in the case of free docking [91]. For the free docking, two different docking methods have been used depending is the complex is hetero or homomeric [94,95]. All the models produced are then ranked by a proper solution called VoroMQA implemented in the VoroMQA web server [96–98]. Then, the top 100-500 models are relaxed [99–101].

Seok's group predicts protein interaction structure thanks to GALAXY pipelines and then adds some human insight regarding the binding site identification, template and model selection using literature and public database information (see Figure 8)[102,103]. The automated parts use GalaxyPPDock, an in-house method developed for *ab initio* protein-protein docking using space annealing algorithm[104]. This server is a rigid-body docking algorithm based on FFT (Fast Fourier Transform) which will produce 50 models then analyzed by GalaxyPPDock.

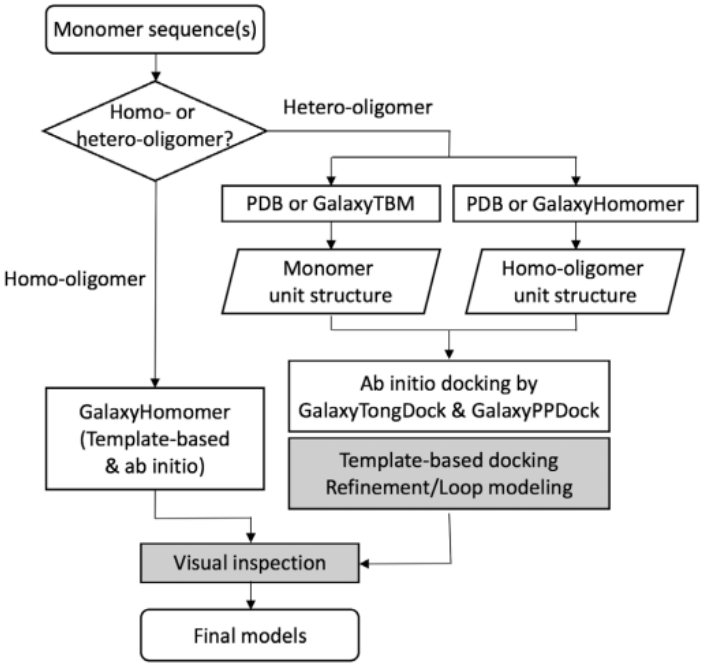


Figure 8. **Overall procedure for protein-oligomer structure prediction**

Human prediction involved additional procedures colored in gray (From Park *et al.* 2020)

Most docking or protein-protein modeling software create a large number of models and this amount has to be reduced to only keep the more likely ones. Scoring or filtering functions have been developed for this purpose. These functions are often specific to the software on which it is applied to but some are standalone software with the only goal to score a model whatever its provenance. Scoring methods can be based on different categories such as physics-based potential, interface shape, knowledge based statistical potentials, evolution of interface residues, machine learning and deep learning using interface structure [105].

These features are mostly not used alone but in combination with others. The main scoring software are listed in the Table 5 below:

Scoring Software	Features category	Principles
AccuRefiner [106]	- Physics-based potentials - Interface shape - Deep learning	RMSD of atom positions combined with deep learning
Degiacomi (2019) [107] **AutoEncoder**	- Physics-based potentials - Deep learning - Molecular Dynamics	Neural network with structures obtain after Molecular Dynamics to characterize conformational space
MaSIF (Molecular Surface Interaction Fingerprinting) [108]	- Geometric features - Chemical features - Polar coordinates - Deep learning	Every feature is combined into a map to learn a soft grid and get specific layers
iScore [109]	- Physics-based potentials - graph kernel - Evolutionary conservation	Combining evolutionary, topological and energetic information for scoring docked conformations
Kingsley *et al.* (2016) [110]	- Potential of mean force calculations - RMSD - steered Molecular Dynamics (sMD)	Using sMD to calculate difference between the highest and lowest force and then umbrella sampling
Lu *et al.* (2003) [111]	- Statistical interfacial pair potentials - Multithreading	The pairwise amino acid preference to interact across a protein-protein interface is analyzed and pair potentials

		constructed
ITScore-PP and ITScore-PP(SCM) [112]	- Interatomic pair potentials - Iterative method - Side-chain center of mass (SCM)	Improving the interatomic pair potentials by iteration, until the pair potentials can distinguish true binding modes Using the (SCM) to represent a residue
PROCOS [113]	- Machine learning (SVM)	Calculating a probability-like measure to be native for a given complex with a model based on features of false and native complexes
Nadaradjane *et al.* (2018) [114] **InterEvDock** [115]	- Co-evolution with Multiple Sequence Alignment (MSA)	Exploiting coevolution constraints in protein-protein docking methods
3D Zernike descriptors (LZerD) [81,116]	- Shape complementarity - Kd-tree nearest neighbor algorithm	Using 3D Zernike descriptor to capture shape complementarity with a molecular mechanics-based
GNN-Dove [105]	- Graph Neural Network (GNN) - Intermolecular Interactions	Interface area is translated into a network with gate-augmented attention mechanism
HADDOCK2.2 [117]	- Physics-based potentials - Molecular shape	Linear combination of various energies and buried surface area

Table 5. **Non exhaustive list of scoring algorithms with summarized features and principles**

Another main protein-protein complex also appeared in the last few years which is AlphaFold v2 (AF), as described in the protein part of the introduction this method is breakthrough for protein modeling. But the last version that allows AlphaFold v2 to model complexes called AlphaFold-Multimer (AFM) is available but there is no official publication. Nevertheless, the paper is available on BioRXiv as their code so it is possible to use

AlphaFold-Multimer[118]. Since the first release, improvements have been made to avoid clashes between atoms which was the main problem. In their BiorXiv preprint, AFM showed already good results on a dataset (*Recent-PDB-Multimers*) before removing the clash as Table 6 shows. Recent-PDB-Multimers is a homology-reduced set of 4,433 recent protein complexes from PDB which have been then reduced to 2,603 to remove complexes with at least 40% of template similarity [118].

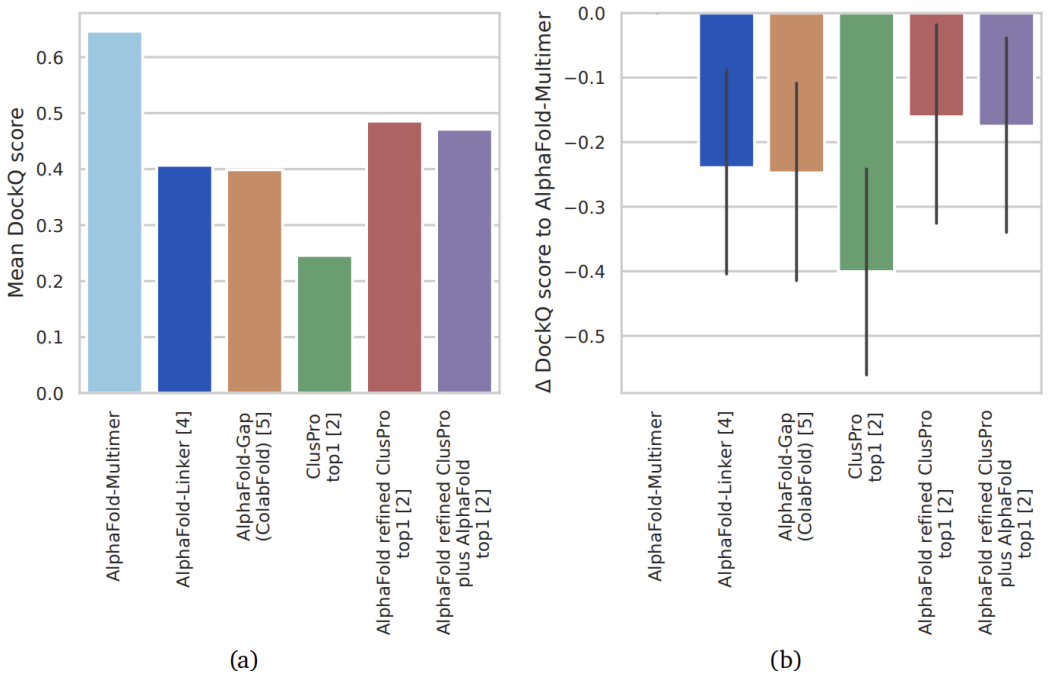
They also compared AFM to other docking algorithm as ClusPro but also version created from the first version of AF known as AlphaFold-Linker, ColabFold, AlphaFold refined Cluspro and AlphaFold refined Cluspro plus AlphaFold which are also software tested and showed in BioRXiv [119]. The results are shown in Figure 9.

Type of interfaces	Mean DockQ score	Incorrect (%)	Acceptable (%)	Medium (%)	High (%)
Homomeric	0.523	30.70	9.83	25.10	34.30
Heteromeric	0.479	32.70	11.90	33.10	22.30

Table 6. **Performance on the *Recent-PDB-Multimers* dataset, evaluated on homology-reduced chain pairs , with low training set similarity broken down into DockQ categories**

Incorrect: 0 ≤ DockQ < 0.23 ; Acceptable: 0.23 ≤ DockQ < 0.49; Medium: 0.49 ≤ DockQ < 0.80; High: 0.80 ≤ DockQ (adapted from Evans *et al.* 2021 preprint)

As for AlphaFold v2, AFM adapted their methods so they can be applied on protein complexes. As complexes multiply the time and resources to calculate the models, the AF system has been trained on cropped segments of proteins which are contiguous blocks of residues up to 384 amino acids. Also, another bunch of details have been changed to fit the new problematic: changing the loss taken into account for permutation symmetry among identical chains, the multiple sequence alignment are paired into alignment to reveal inter-chain genetic information for example.



(a) (b)

Figure 9. **The performance of AlphaFold-Multimer against several published baselines is shown on a dataset, consisting of 17 heterodimer targets with low training set homology**

AlphaFoldLinker is AlphaFold with a 21 residue linker of repeated Glycine-Glycine-Serine residues, similar to previous AlphaFold modifications. AlphaFold-Gap (ColabFold [120]), version from 2021-08-16, is a published system that runs AlphaFold with a gap between residue indices between chains, uses MMSeqs2 for genetics, includes MSA pairing and does not include templates. ClusPro, AlphaFold refined ClusPro, and AlphaFold refined ClusPro plus AlphaFold are all systems and results based on combining the docking algorithm ClusPro with AlphaFold, results are as reported in [119]. Error bars represent a 95% confidence interval around the mean. (from Evans *et al.* 2021, preprint)

AFM has been such a breakthrough that a lot of AF v2 based software have been developed such as ColabFold, OmegaFold, Uni-Fold and even a python package which uses AlphaFold-Multimer called AphaPullDown [121–123]. This shows the major evolution of the *de novo* protein and protein complex structure prediction thanks to this brand new method not even truly published yet.

b) Protein-peptide docking

Protein-peptide docking is also a very challenging domain whether it is to understand the mechanics really important in many cellular processes, as peptides mediate 40% of protein-protein interactions or in drug design. Indeed, peptides are promising drug candidates but it requires a good characterization of the interaction between them and the

target molecules [124]. But the difficulty of experimentally resolving protein-peptide structures results in a low amount of such structures in Protein Data Bank (PDB) [125] compared to the number of protein-protein complexes available. So there is a need for effective and efficient computational algorithms, methods development to complement experimental techniques[126]. The *in silico* methods to produce protein-peptide models are first to find the peptide binding site then the docking. There is a complementary step according to the research purpose which is the design of inhibitory peptides. Finding peptide binding can be done through analyzing the solvent surface area but also regarding amino acid residue probes [126,127].

The first method is used by the PeptiMap protocol. This protocol is based on experimental observation, regarding NMR experiments but also crystals, that small ligands of cariant length and polarity tend to bind on protein surface regions where other bigger ligands interact [128,129]. In their protocol, PeptideMap uses Fast Fourier Transform (FFT)-based method to map the surface of solvent to identify binding sites [130–132]. But the FTMap method is modified to fit for peptide-protein binding sites[127].

The second method is the one used by the ACCLUSTER web server. It is based on the hypothesis that peptide binding sites are composed of residues that can form good chemical interaction [126].

Another method is doing semi-rigid docking. This is the case of PepSite-Finder which considers the protein receptor as rigid and the peptide is represented by a number of different conformations[133]. These conformations are retrieved in the PepDB database with multiple protein-peptide complexes obtained at resolution of less than 2 Å [134]. The size of the peptides are between 5 and 15 amino acids long.

Globally, we can divided the protein-peptide docking into 3 main categories: template-based docking, local docking and global docking which can been summarized as shown in Figure 10 from Ciemny *et al.* 2018 [124]. Templated-based docking for protein-peptide is the same protocol as for protein-protein docking except that there is a lower amount of data. This method uses already resolved complexes as scaffolds to build the

models. Template-based docking provides good results if the inputs are close to the templates. If no template exists between the protein and a peptide it is possible to use the known interaction interface of the protein (from a protein-protein complex for example). Local docking uses the information provided by the user to perform the docking of the peptide in a particular interface of the receptor. This information can be obtained by experimental data or by prediction software as explained earlier. As the interaction area is provided, docking is able to be performed as flexible-body docking. The final method is called global docking and is the method of choice when no information about the complex is available. The global docking will usually perform rigid-body docking for protein and peptide. Some software generates many peptide conformations from the peptide sequence provided. A list of protein-peptide docking software has been retrieved in the Table 7.

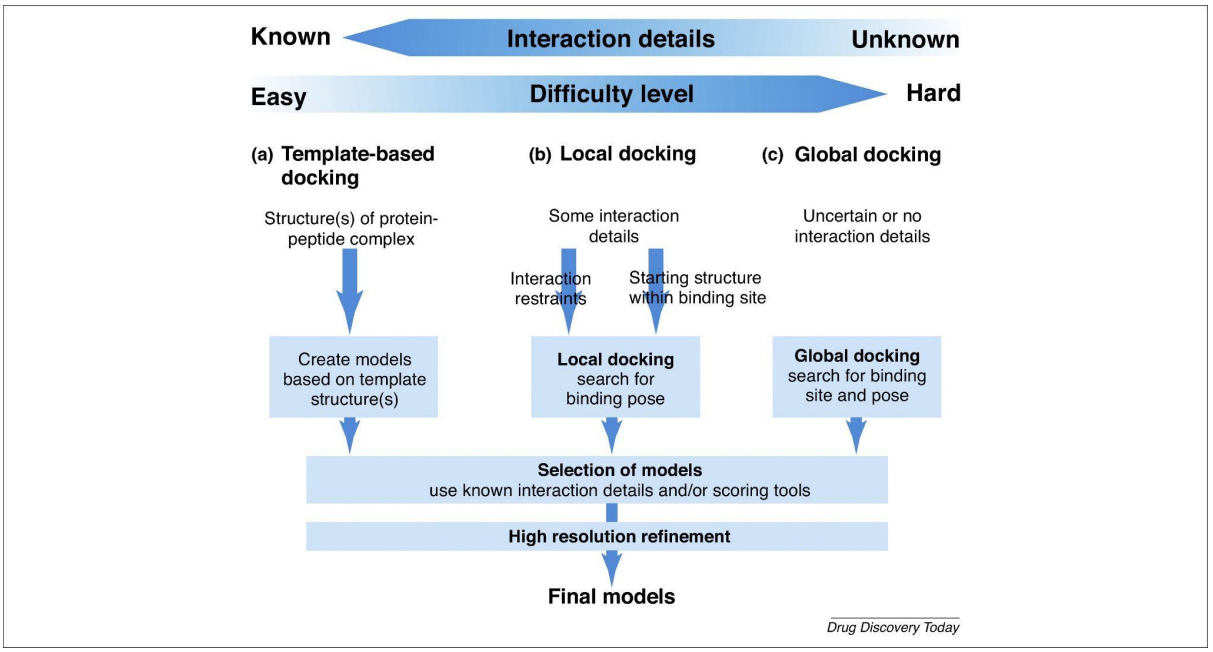


Figure 10. **Typical pipelines for protein–peptide molecular docking**

Docking methods can be divided into three categories according to the amount of required input data: (a) template-based methods that utilize knowledge about the structure of similar complexes (templates); (b) local docking methods that require some knowledge about the binding site; and (c) global docking methods that assume no knowledge about the peptide beyond its sequence. From Ciemny *et al.* (2018) [124]

Nowaday, there is at least 3 main challenges regarding protein-peptide docking [124]:

- Flexibility: be able to model significant conformational changes whether it is for the receptor or the peptide itself.

- Scoring: The scoring functions developed for protein-protein docking are not as efficient for protein-peptide filtering and the need to have dedicated scoring functions is lacking.
- Integrative modeling: As peptide can be very flexible, adding some NMR data for instance could help to identify the possible conformations or even native contacts. cryo-EM and SAXS could also help to find the shape of the complex and guide the protein-peptide docking [135].

Method Server	**Description**
GalaxyPepDock [136] http://galaxy.seoklab.org/pepdock and a standalone version	Template-based docking procedure: (i) search for templates based on structure and interaction similarity; (ii) model building by energy-based optimization; (iii) energy-based scoring; and (iv) refinement of final structures
PepComposer http://biocomputing.it/pepcomposer/webserver	Template-based docking procedure: (i) search for regions structurally similar to region of predefined binding site in database of experimentally solved monomeric proteins; (ii) retrieve continuous backbone fragments in contact with region of binding site; and (iii) design peptide sequence
Rosetta FlexPepDock http://flexpepdock.furmanlab.cs.huji.ac.il and standalone version	Local docking procedure: Monte Carlo-based optimization of fully flexible peptide within binding pocket. Receptor flexibility is limited to side-chains, but can be extended to full receptor. Clustering and scoring according to Rosetta energy function
DynaDock Not available publicly	Local docking procedure: (i) rigid-body optimization of peptide orientation within binding site, followed by (ii) refinement of fully flexible peptide receptor with Optimized Potential Molecular Dynamics procedure (using soft-core potentials for implicit receptor flexibility)
PepCrawler http://bioinfo3d.cs.tau.ac.il/PepCrawler/	Local docking procedure: (i) fully flexible peptide docked with Rapidly-exploring Random Trees algorithm, followed by (ii) clustering-based scoring. Receptor flexibility limited to side-chains
HADDOCK peptide docking	Local docking procedure: (i) generation of peptide

http://milou.science.uu.nl/services/HADDOCK2.2/haddock.php	structures by threading peptide sequence onto three peptide conformations (alpha-helix, polyproline-II or extended); (ii) rigid-body docking of peptide structures within binding pocket; (iii) scoring based on binding free energy (calculated using dampened Molecular Mechanics Poisson–Boltzmann Surface Area); (iv) flexible refinement of model; peptide and interacting residues of receptor are fully flexible
PEP-FOLD 3 http://bioserv.rpbs.univ-paris-diderot.fr/services/PEP-FOLD3	Local docking procedure: (i) generation of starting poses; (ii) Monte-Carlo-based sampling of peptide conformation; (iii) RMSD-based clustering of resulting models
AutoDock Vina Standalone version	Local docking procedure: Monte-Carlo-based sampling of peptide conformations within binding pocket. Receptor flexibility is by default limited to side-chains, but can be extended to include backbone
DINC 2.0 http://dinc.kavrakilab.org	Local docking procedure: based on AutoDock 4 for docking long peptides, in which a peptide is divided into segments of increasing length. During docking, receptor structure remains rigid
Gold Standalone version	Local docking procedure: Monte-Carlo-based sampling of peptide conformations within binding pocket. Receptor flexibility either limited to side-chains or implicit (ensemble docking)
pepATTRACT http://bioserv.rpbs.univ-paris-diderot.fr/services/pepATTRACT/	Global docking procedure: (i) generation of peptide structures by threading peptide sequence onto three peptide conformations (alpha-helix, polyproline-II or extended); (ii) global rigid-body docking of peptide structures within binding pocket; (iii) scoring with ATTRACT score; followed by (iv) flexible refinement of models with iATTRACT [137]. Both peptide and interacting residues of receptor are fully flexible
CABS-dock http://biocomp.chem.uw.edu.pl/CABSdock and as a standalone version	Global docking procedure: (i) explicit fully flexible docking simulation; and (ii) clustering-based scoring. Receptor flexibility limited by default to small backbone fluctuations, but can be increased to include selected receptor fragments
ClusPro PeptiDock https://peptidock.cluspro.org/	Global docking procedure: (i) motif-based prediction of peptide conformation; (ii) PIPER [138] rigid-body docking; (iii) scoring according to structural clustering; and (iv) minimization of final structures
PIPER-FlexPepDock	Global docking procedure: (i) prediction of peptide

http://piperfpd.furmanlab.cs.huji.ac.il	conformation using Rosetta fragment picker; (ii) PIPER-based rigid-body docking [138]; (iii) refinement using Rosetta FlexPepDock [139] and (iv) clustering and scoring according to Rosetta energy function

Table 7. **Non exhaustive list of available protein-peptide docking software and their summarized protocol**

First column corresponds to the software name, the second column correspond to the availability and the third column is the brief description of the protocol (Adapted from Ciemny *et al.* 2018[124])

2. Evaluation of docking prediction

As protein structure prediction can be assessed by the CASP experiment, it is also the case for protein complexes prediction. Indeed, modeled on CASP, a special experiment called CAPRI (Critical Assessment of PRedicted Interactions) was created at the beginning of this century with a first publication in 2003. It is a community wide experiment with the aim to assess the capacity of actual protein docking methods to predict protein-protein interactions [140]. Today, the variety of interesting complexes such as protein-peptide interaction leads to a more open assessment with protein-protein, protein-peptide, protein-RNA, protein-DNA and large assemblies assessments [87,140–148]. The CAPRI experiment is based on Rounds, where predictor groups can participate to predict complexes called targets. Since 2005, not only predictors can participate in CAPRI Rounds, but also prediction scorers with the aim to see if the selected models by the scoring algorithm are the closest to reality. These targets have been proposed by experimentalists who have resolved a complex structure and provided it to CAPRI with a non-disclosure agreement. To be able to propose a new target to CAPRI, the structure has to be experimentally resolved and not yet published in a database so there is no available data to orient the prediction. Once there is enough target for a Round, predictors and scorer groups can register to participate in the prediction or the scoring of these targets. A Round is processed as the following (and as it can be seen on Figure 29):

- Once the predictors have registered, the sequences of the chains are provided to predictors
- Predictors provide two sets: the first one called P-set which contains 10 of the best models produced by a group according to their algorithm; the second one is the U-set which is a larger set consisting of 100 models, including the 10 best.

- The combined U-set from all participants is then proposed to the scoring groups and these groups will select the 10 best models according to their methods and become a S-set.
- Then every set is assessed according to the experimental structure

The assessment is based on different criteria (see Figure 11):

- $\int_{nat}$ and $\int_{non-nat}$: which are the fraction of receptor-ligand residue contacts found in the model and the experimental structure and the fraction of contacts which have been predicted in the model and which are not present in the target structure[149].
- I–rms, L-rms and S-rms: which are criteria to assess the quality of the predicted interface and are based on RMSD. I-rms is the rmsd at the interface backbone atoms and S-rms is the same for the interface side-chain atoms.

Models can have, then depending on these criteria, 4 different qualities from worst to best: incorrect, acceptable, medium and high quality. The criteria to define the quality are summarized in the Table 8 above:

Ranking	**Conditions based on Capri computed parameters**
High	$f_{nat} \geq 0.5$ AND (L_rms ≤ 1.0 OR I_rms ≤1.0)
Medium	($f_{nat} \geq 0.3$ AND $f_{nat} \leq 0.5$) AND (L_rms ≤ 5.0 OR I_rms ≤2.0) OR $f_{nat} \geq 0.5$ AND L_rms>1.0 AND I_rms >1.0
Acceptable	($f_{nat} \geq 0.1$ AND $f_{nat} < 0.3$) AND (L_rms ≤ 10.0 OR I_rms ≤4.0) OR $f_{nat} \geq 0.3$ AND L_rms>5.0 AND I_rms >2.0
Incorrect	$f_{nat}<0.1$ OR (L_rms>10.0 AND I_rms >4.0)

Table 8. **Assessment criteria condition for model quality**

(Adapted from Méndez *et al.* 2005)

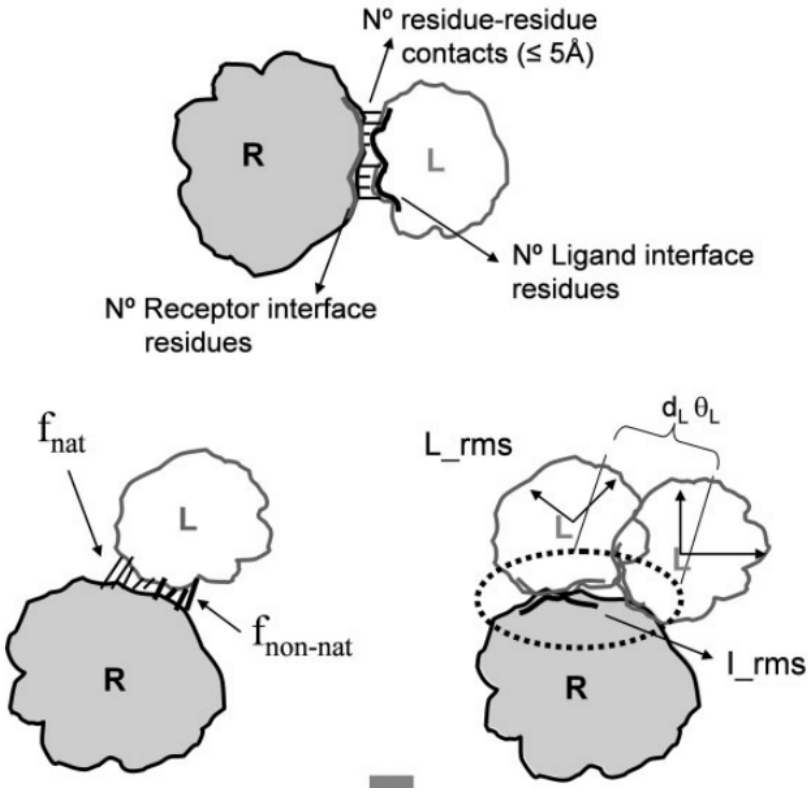


Figure 11. **Schematic illustration of the quality measures used to evaluate the predicted models**

For each target, we computed the number of residue–residue contacts between the receptor (R) and the ligand (L), and for each of the components, the number of interface residues. See text for the definition of the interface in each case. For each model, we computed the fractions f_{nat} of native and $f_{non\text{-}nat}$ of non-native contacts in the predicted interface. In addition we computed the RMSD of the backbone atoms of the ligand (L_rms), the misorientation angle L and the residual displacement dL of the ligand center of mass after the receptor in the model and experimental structures were optimally superimposed. We also computed I_rms, the RMSD of the backbone atoms of the interface residues after they have been optimally superimposed (from Méndez *et al.* 2005).

Today, CAPRI is grouped to the CASP experiment and in December 2022, the CASP15-CAPRI will take place in Turkey. So for this manuscript, the results and state of the art will stop before this event. For CAPRI we denombrate 180 targets including some different interfaces between two same chains. Most of these targets are retrieved inside a new website available at this link: www.scoreset.org.

3. Modification sites: the case of Post-Translational Modifications

Post Translational Modifications (PTMs) are involved in cell cycle and have been shown to be involved in pathogenicity. The prediction of such PTMs is a key research to win time and money so many software have been developed. As prediction is a theoretical result there is a need to evaluate such prediction software. To this we can use different statistical measurements such as sensitivity (Sn), specificity (Sp), accuracy (Acc), Area Under Curve (AUC) and precision (Pr). These measurements are calculated according to four kinds of results: True Positive (TP) which are results predicted to be positive (here modified) and which are actually modify; False Positive are results predicted to be positive but are not in reality; True False (TF), results predicted to be negative (or non modified) by the algorithm and which are indeed negative; False Negative (FN) which are results predicted to be negative but are actually positive. These statistical measurements are explained just below:

- Sensitivity (Sn) is the percent of positive results found by the algorithm which are actually positive: $Sn = \frac{TP}{TP+FN}$
- Specificity (Sp) is the percent of negative sites which are actually predicted to be negative by the algorithm: $Sp = \frac{TN}{TN+FP}$
- Accuracy (Acc) is the statical measurement to see how well a prediction predict both positive and negative results: $Acc = \frac{TP+TN}{TP+FP+TN+FN}$
- Precision (Pr) is the characteristic of a software to correctly predict positive results without without too much false positive results: $Pr = \frac{TP}{TP+FP}$

Also, to be able to compare results of predictive software it can be interesting to plot the ROC and precision-recall curves. As the first one shows the rate of True Positive in function of the Rate of False Positive and allows to calculate the AUC which is a good revealer of the predictive power of a tool. The precision-recall curve shows the precision of a prediction tool in function of the accuracy also called recall. It allows one to determine the best ratio between precision and accuracy.

As some have well-known mechanisms regarding their substrate recognition like *N*-GlcNAcylation with a N-X-S/T motif (where X can be any residue except a proline) others are still difficult to identify. It is the case for phosphorylation and its competitive modification the *O*-GlcNAcylation. If the first modification can be divided regarding the super family of kinases to identify the mechanism of substrate recognition, this is not the case of the second which has only one enzyme to add the sugar.

Phosphorylation site prediction is a well-studied problem and the high number of experimental sites and the improved knowledge coupled to machine learning algorithms tends to show good results. Nowadays, the number of phosphorylation predictions is over 40 and the latest ones show good results with a sensitivity of 47.80% and a precision for 82.70% for a specificity of 90% and a sensitivity of 33.86 and a precision of 87.13% for a specificity of 95% [25].

Like phosphorylation, the prediction of *O*-GlcNAcylation sites is also a long-time challenge with a first prediction tool in 2002, Yin-O-Yang based on the crosstalk between *O*-GlcNAcylation and the phosphorylation and only 40 experimentally proven *O*-GlcNAcylation sites. Experimentally proven sites were difficult to obtain and its number hardly increased at the beginning and the first software showed a AUC of 74.3% according to their article for OGlcNAcScan [150], to counter the lack of data, some algorithms used training sequences with homology sequences and others used phosphorylation prediction to help thanks to the fact of the competition between these two modifications. Even with a bigger number of *O*-GlcNAcylated sites, the number of false positive sites is still too high to be really helpful for experimentalists. The reason for these low quality results can be explained by the diversity of the sites regarding the amino acid composition around them. Today, to counter these problems different supervised machine learning algorithms have been proposed like neural networks (NN), Support Vector Machine (SVM), Random Forest (RF) but also different kinds of data.

Random Forest (RF) is a well-known and well-used algorithm, with a major advantage which is that the features used to discriminate results can be explained which is not always the case with machine learning. It is in the family of tree ensembles algorithm with the

Gradient Boosting Tree (GBT) also called Gradient Boosting Machine (GBM) algorithm (see Figure 12) . A decision tree is a flowchart that will divide the interest population into subgroups that differ according to a feature analyzed by the tree [151]. A good decision tree will separate the total population into subgroups which have a high similarity inside each group and a high variability between-groups.

In principle, RF builds an ensemble of decision trees called a forest and this forest will combine all the tree results to give an overview result [152]. Every tree of the ensemble is trained on a subset of the dataset and its result will be combined with all the other trees and the final answer will take the majority of the votes. The random came from the cutting of the dataset into random subsets and each one is used to train a model.

GBT is also based on trees and takes the mean of all these trees which are trained on re-weighted subsets of the data. The first trees produced errors and these errors contribute to learning more optimal trees in the next iteration [153]. RF is used as a classifier in the *O*-GlcNAcylation predictor called *O*-GlcNAcPRED-II [34]. This predictor also uses another classifier called Support Vector Machine (SVM). SVM is a regression algorithm initially created to separate two distinct groups. But it has been adapted to perform regression on multiple groups even if it is not the most adapted algorithm.

SVM algorithm is based on a graphical interpretation of data: each element of the training set will be plotted in a multidimensional graph according to its features or variables. Then based on this plot, the algorithm will find a way to separate thanks to a dimensional hyperplane. This hyperplane will then rely on the nearest data, called support vector, and will try to optimize by enlarging the distance, called margin between the hyperplane and the dots called support vectors. Once the function is trained, the new data will be separated according to the model by the hyperplane and the confidence will be calculated by the distance of the new data with the hyperplane (see Figure 13). The hyperplane was first linear but the optimization of the algorithm now allows radial basis, polynomial and sigmoid function to have a better fit to the training data. This method is used to predict *O*-GlcNAcylation as in OGTSite, OGlcNAcScan and O-GlcNAcPRED-II [34,36,150]

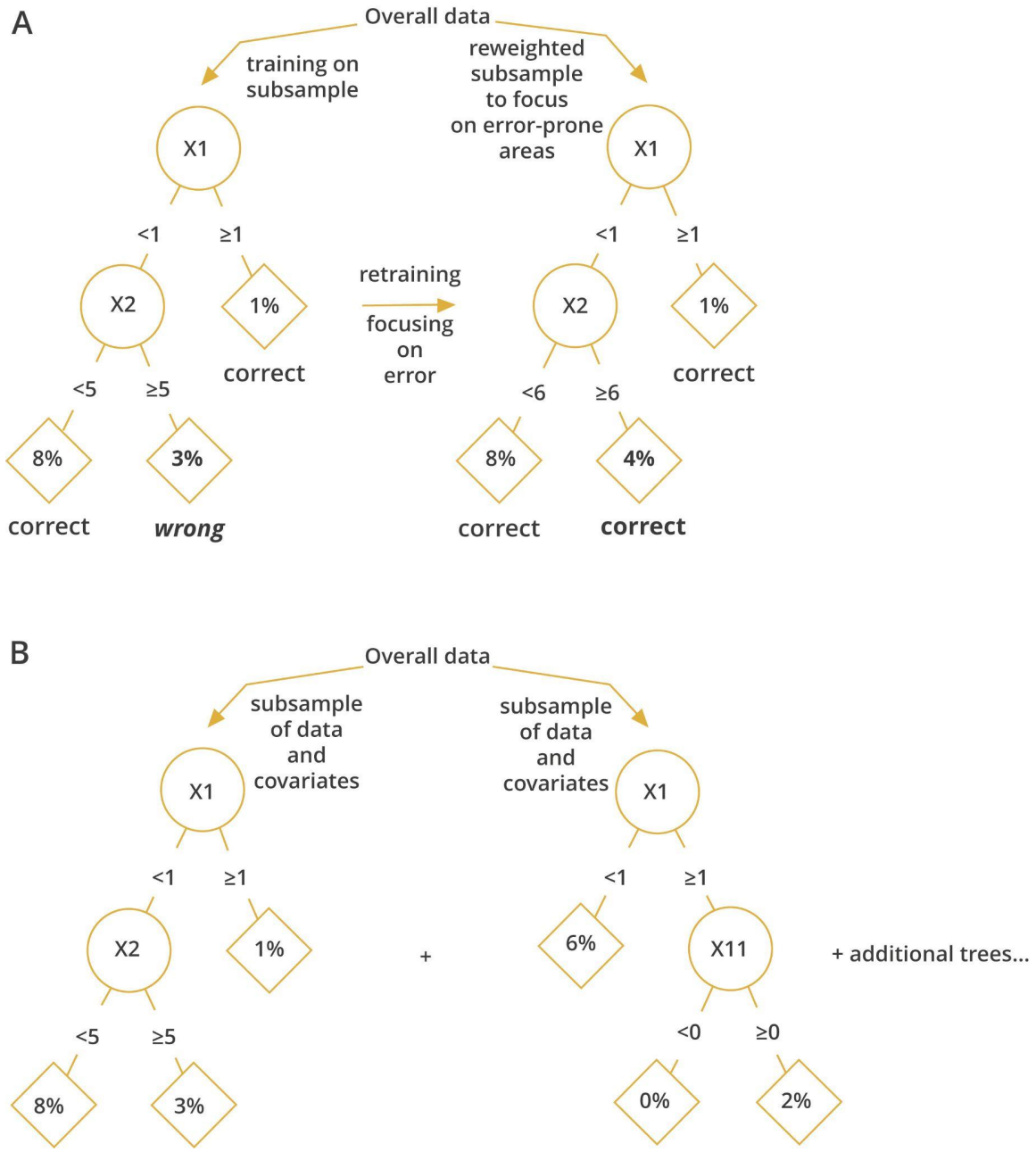


Figure 12. **Decision trees**

(A) Gradient boosting machines (GBM) and (B) random forests (RF). The circles display the covariates (X variables) whose values determine each branch point, whereas the diamonds provide the tree-predicted probability of the outcome under study. (from Doupe *et al.* 2019)

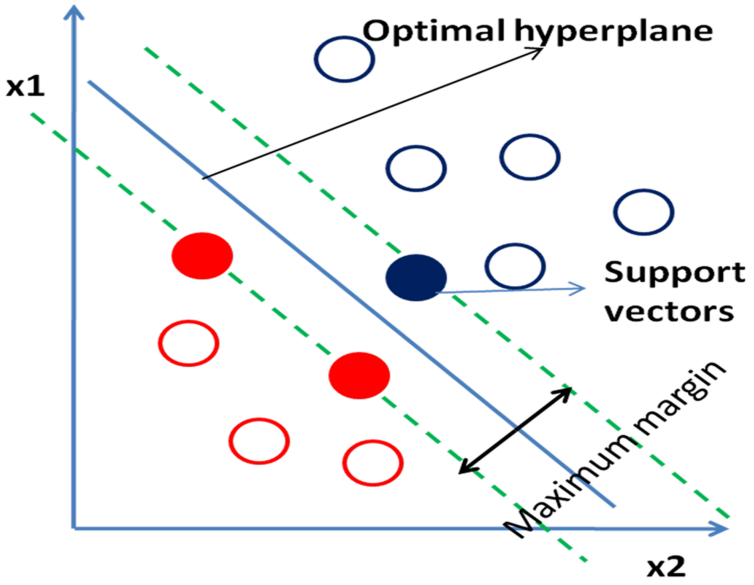


Figure 13. **Hyperplane (blue line) representation in SVM**

Red and blue circles represent data points from two different classes. Solid filled circles denote support vectors (from Chauhan *et al.* 2014) [154]

Neural network (NN) is the most complex machine learning algorithm and is very close to what is called deep learning. The main principle of NN consists in the capture of complex relationships in the associations between the outcome and the input variables. These associations are represented by multiple hidden layers with prespecified functions and the goal is to estimate the weights through input and outcome data to minimize the average error between the outcome and their predictions [155]. The layers are composed of formal neurons inspired from biological neurons where they retrieve information from previous neurons, make a pondered sum of the information and give it as an outcome. The number of neurons inside a layer and the number of layers will vary according to the neural network. This technique is very powerful but the amount of data to train a big neural network needs to be very important. This method is used by YinOYang [35].

In 2021, a big database called The *O*-GlcNac Database, retrieves every *O*-GlcNAcylated sites known in the litterature and has emerged with the highest number of sites making it is the biggest database currently available [156]. Each site is categorized by a score calculated according to the number of time it has been found in different publications. This high number of data may be a solution to finally have a good prediction model for *O*-GlcNAcylation. The database can be found at this address: www.oglcnac.mcw.edu [157].

II. Thesis objectives

This thesis can be divided into three main parts. All these parts are connected in different manners so they will be described chronologically. The second part is divided in two sub parts (B,C)

A. Prediction of *O*-GlcNAcylated sites

O-GlcNAcylation is a main Post-Translational Modification (PTM) with several challenges, whether to understand the mechanism of *O*-GlcNAcylation or for curative purposes in certain metabolic pathways involved in different diseases. Already available *O*-GlcNAcylation prediction tools produce a high number of false positives in their results and this is a big problem for biologists. All of these algorithms are based on sequences with a low amount of data. Even if the methods they are based on are different, the results remain particularly poor. These software statistical measurements are also based on different dataset, it could be interesting to create a new dataset to test them on the same data. Despite the analyses of available tool results, *O*-GlcNAcylation prediction is still a hard task. Thus, as *O*-GlcNAcylation is known to be a dynamical Post-Translational Modifications, our idea was first to see the structural aspect of *O*-GlcNAcylated sites to see if the sites are accessible for example. So in this first part of my PhD thesis, the objectives were to build a dataset to test current *O*-GlcNAcylation prediction tools and then use it to train and test a new prediction software based on sequences, structure and accessibility to improve the current methods. Some new features will be extracted from the dataset and used by different machine learning algorithms. These results have been added in this manuscript in the form of an article which has been published. Some supplemental results obtained during the PhD have been added in a second section of this main objective.

B. Development of an assessment method for complexes in the context of CoViD-19

During my PhD, the world was hit by the CoronaVirus Disease-19 (CoViD-19) health crisis. As it became a pandemic, researchers joined their effort to understand the mechanisms of this virus infection, to highlight molecules to block the propagation or find a

vaccine. In this optic, an ensemble of researchers highlighted 332 high confidence interactions between human and viral proteins using affinity-purification–mass spectrometry (AP-MS) on HEK-293T/17 cell line infected with the SARS-CoV-2 virus [158]. From this study, the CAPRI committee selected some of the interactions based on the available structure of the partners of the interactions, giving rise to a special CoViD prediction Round. For this Round, 5 Targets have been proposed to the CAPRI community in the purpose to understand the mechanisms under the interaction with good models. But as CAPRI always assesses models with an experimentally found template and as there is none for the 5 targets, the problem was to be able to determine from all these models the best models and be able to determine the quality of a model. The objective of my work was to verify the different set of models and then find a way to determine good models without templates or at least find a way to highlight good quality models or find a way to determine if a model can be likely or not.

C. Testing adjacency overlap scoring method

In the previous section, a scoring method has been highlighted called Adjacency Overlap (AO). But the validation set for this method was quite small, so we decided to test on a new and larger benchmark, using data from the CAPRI community. This benchmark allows us to test our method on various sets and compare it to other scoring methods. If this method performs well it will allow us to conclude on the results obtained in the previous part.

D. Modeling of the interaction between OGT and beta-catenin

The crosstalk between phosphorylation and *O*-GlcNAcylation is well known and it has been shown to be involved in many diseases. This is the case in the Wnt pathway, involved in cell proliferation and migration, where a proto-oncoprotein protein called beta-catenin (β-catenin or CTNB1) is impacted by this competition. Indeed, β-catenin is widely expressed in many tissues. Mutations and over-regulation of this protein are associated with several forms of cancer and notably ColoRectal Cancer (CRC). β-catenin is usually phosphorylated to be degraded, but it has been shown that the β-catenin can also be *O*-GlcNAcylated which

inhibits its proteasomal degradation and leads to the activation of the transcription of target genes. Although the interaction between CTNB1 and *O*-GlcNAc Transferase (OGT) is established, the molecular details of this interaction remain unknown. My objective here was to use a combination of computational techniques as molecular modeling, protein-protein and protein-peptide docking and dynamics to understand the recognition process. The understanding of the interaction at a molecular level may lead to the identification of hot spots of the interaction on both proteins. These key spots may lead to the identification of pharmacological molecules capable of inhibiting this interaction. The *O*-GlcNAcylation/phosphorylation are already identified on the N-terminal part of the β-catenin, in a region which has been named destruction box. This region is called like this because it allows a molecular mechanism leading to its degradation by the proteasome. This information means that this unstructured part is able to go into the catalytic pocket of the OGT. As OGT and β-catenin have both a recognition domain, respectively called TPR domain and Armadillo domain one hypothesis can be that these two interact. Also, as the TPR domain of the OGT is a super helix, another hypothesis can be the interaction of the N-terminal segment of the β-catenin inside the TPR. The last hypothesis of this interaction which is less likely to happen is that the C-terminal segment of the β-catenin (which is also unstructured) interacts with the TPR. These three hypotheses are retrieved in Figure 14. Another hypothesis which is hard to verify is the presence of a chaperone protein to bring the two partners together.

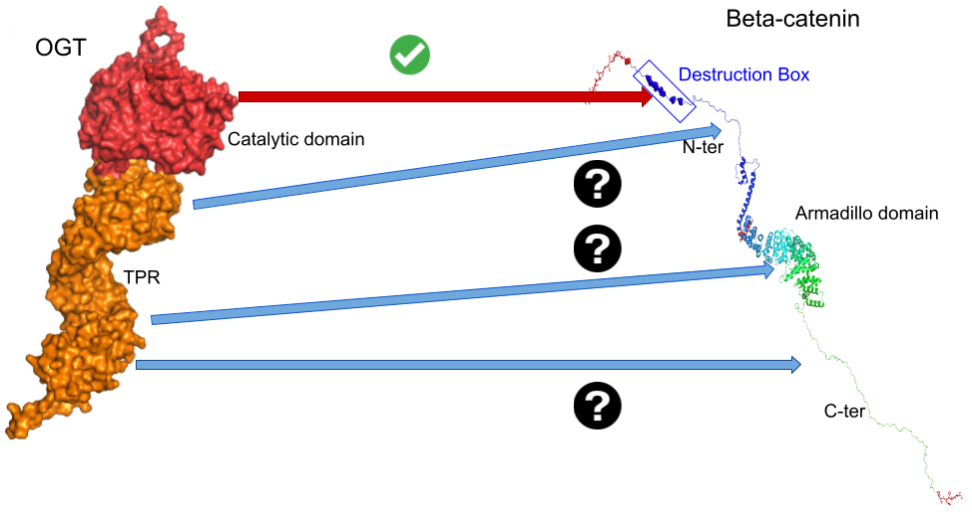


Figure 14. **Schematic representation of the three hypotheses of interaction between OGT and the TPR domain**

III. Specific Interaction prediction: the specific case of *O*-GlcNAcylation

A. The *O*-GlcNAcylation Prediction: An Unattained Objective

Advances and Applications in Bioinformatics and Chemistry Dovepress
open access to scientific and medical research

Open Access Full Text Article ORIGINAL RESEARCH

O-GlcNAcylation Prediction: An Unattained Objective

Theo Mauri[1]
Laurence Menu-Bouaouiche[2]
Muriel Bardor[2]
Tony Lefebvre[1]
Marc F Lensink[1]
Guillaume Brysbaert[1]

[1]Univ. Lille, CNRS; UMR8576 - UGSF - Unité de Glycobiologie Structurale et Fonctionnelle, Lille, F-59000, France; [2]Normandy University, UNIROUEN, Laboratoire Glyco-MEV EA4358, Rouen, 76000, France

Background: *O*-GlcNAcylation is an essential post-translational modification (PTM) in mammalian cells. It consists in the addition of a *N*-acetylglucosamine (GlcNAc) residue onto serines or threonines by an *O*-GlcNAc transferase (OGT). Inhibition of OGT is lethal, and misregulation of this PTM can lead to diverse pathologies including diabetes, Alzheimer's disease and cancers. Knowing the location of *O*-GlcNAcylation sites and the ability to accurately predict them is therefore of prime importance to a better understanding of this process and its related pathologies.

Purpose: Here, we present an evaluation of the current predictors of *O*-GlcNAcylation sites based on a newly built dataset and an investigation to improve predictions.

Methods: Several datasets of experimentally proven *O*-GlcNAcylated sites were combined, and the resulting meta-dataset was used to evaluate three prediction tools. We further defined a set of new features following the analysis of the primary to tertiary structures of experimentally proven *O*-GlcNAcylated sites in order to improve predictions by the use of different types of machine learning techniques.

Results: Our results show the failure of currently available algorithms to predict *O*-GlcNAcylated sites with a precision exceeding 9%. Our efforts to improve the precision with new features using machine learning techniques do succeed for equal proportions of *O*-GlcNAcylated and non-*O*-GlcNAcylated sites but fail like the other tools for real-life proportions where ~1.4% of S/T are *O*-GlcNAcylated.

Conclusion: Present-day algorithms for *O*-GlcNAcylation prediction narrowly outperform random prediction. The inclusion of additional features, in combination with machine learning algorithms, does not enhance these predictions, emphasizing a pressing need for further development. We hypothesize that the improvement of prediction algorithms requires characterization of OGT's partners.

Keywords: machine learning, glycosylation, *O*-GlcNAc, post-translational modification, dataset, OGT

Correspondence: Theo Mauri; Guillaume Brysbaert
UGSF Campus CNRS, Parc de la haute-borne, 50 Avenue de Halley, BP 70478, 59658 Villeneuve d'Ascq Cédex, Lille, France
Tel +33 3 62 53 17 32
Fax +33 3 62 53 17 01
Email theo.mauri@univ-lille.fr; guillaume.brysbaert@univ-lille.fr

Introduction

O-GlcNAcylation (*O*-linked β-*N*-acetylglucosaminylation) is a dynamic post-translational modification (PTM) occurring in cytosol, nucleus and mitochondria under the supervision of two antagonist enzymes: the *O*-GlcNAc Transferase (OGT) and the *O*-GlcNAcase (OGA).[1,2] The target proteins are modified by the addition of a *N*-Acetylglucosamine (GlcNAc) residue onto serines (S) or threonines (T), which derives from UDP-GlcNAc supplied by the hexosamine biosynthesis pathway (HBP). The OGT structure can be divided into two parts: the N-terminal tetratricopeptide repeats (TPR) domain which binds to the substrate, and the C-terminal catalytic domain

that first recruits UDP-GlcNAc and then adds the GlcNAc moiety on the target protein. Three isoforms of OGT are currently known: ncOGT, mOGT and sOGT. The ncOGT, located in the nucleus and cytoplasm, contains 13.5 TPR repeats, while mOGT, located in the mitochondria, exhibits 9 TPRs. The sOGT (small OGT) is detected in the nucleus and the cytosol like the ncOGT, but contains only 2.5 TPR repeats.[3,4] *O*-GlcNAcylation occurs on thousands of proteins involved in many different pathways and dysregulation of its cycling leads to many pathologies such as cancers, diabetes and Alzheimer's disease.[5] The accurate prediction of *O*-GlcNAcylation sites would constitute a significant advance as this major PTM is involved in many vital pathways. Unlike *N*-glycosylation for which the consensus site is well known (N-X-S/T/C with X any residue except proline) and conserved, no specific pattern is currently known for the *O*-GlcNAcylation.

Therefore, developing efficient prediction tools represent a challenge. Few prediction tools such as YinOYang, *O*-GlcNAcPred-II and OGTSite are already available.[6–8] They implement algorithms such as Random Forest, Neural Networks or Principal Component Analysis and are based on sequence data. They advocate to show good prediction results with sensitivity up to 81.05% and specificity up to 95.91% for *O*-GlcNAcPred-II. However, these numbers depend on the underlying test set which is different for each tool, making the results hard to compare. In order to be able to properly evaluate the performances of these predictors, we decided to build a large dataset of currently available experimentally proven *O*-GlcNAcylated sites and to test the performance of the three tools on this new dataset.

We show here that the predictions are not as efficient as expected. Thus, we decided to use the dataset to develop a new prediction tool for *O*-GlcNAcylation sites, expanding upon the use of sequence information by including structural parameters. We investigated the primary, secondary and tertiary structure environment of every experimentally proven *O*-GlcNAcylation site in order to define a set of parameters that could be further used by machine learning algorithms. Here, we show that none of the available tools correctly predict *O*-GlcNAcylated sites, even when we attempted to improve the parameters for machine learning algorithms.

Materials and Methods

Dataset Creation

The dataset was constructed with data from experimentally proven *O*-GlcNAcylated sites of 236 mammal proteins from the UniProt reviewed database (Swiss-Prot) with the following research: "annotation: (type: carbohyd o glcnac evidence: experimental) AND reviewed: yes".[9] To complete this set, we also retrieved experimental data from PTM-ssMP[10] and from the results of Deracinois et -al.[11] In these sets, we rejected the sites found by sequence homology. The negative sites were taken from the same sequences, considering serines and threonines which were not described as *O*-GlcNAcylated.

We curated the full set, removing non-mammal sequences. We also removed redundant sequences. After curation, 565 *O*-GlcNAcylated sites (positive sites) and 40,271 non *O*-GlcNAcylated sites (negative sites) were gathered. We created a second dataset, removing sequences longer than 4000 residues, since two of the three tested tools do not work on such long sequences: this dataset totals 550 *O*-GlcNAcylated and 38,665 non *O*-GlcNAcylated sites.

Evaluation of *O*-GlcNAcylation Prediction Software

We evaluated the performance of three prediction tools, which are YinOYang, *O*-GlcNAcpred II and OGTsite,[6–8] using the sequences of the reduced dataset. The results of predictions of each of them were then compared to experimental data. A prediction was considered as

- false positive (FP) if not identified experimentally,
- true positive (TP) if validated experimentally,
- false negative (FN) if not predicted but proven experimentally,
- true negative (TN) if not predicted and not proven experimentally.

We further calculated the specificity, the sensitivity, the precision (also called Positive Predictive Value (PPV)), the Negative Predictive Value (NPV), the False Discovery Rate (FDR) and the accuracy of each prediction tool:

- sensitivity is the percentage of unmissed positive sites and corresponds to the number of positive sites correctly classified among all positive ones (TP/(TP+FN)),
- specificity is the percentage of unmissed negative sites and corresponds to the number of TN among all negative ones (TN/(TN+FP)),

- precision (or PPV) corresponds to the chance to predict a site as positive and be correct (TP/(TP+FP)),
- False Detection Rate (FDR) is the contrary, namely the chance to be wrong when predicting a positive site (1-PPV),
- NPV is the same as PPV but for negative sites (TN/(TN+FN)),
- accuracy corresponds to the proportion of correctly predicted sites whether they are positive or negative ((TN+TP)/(TN+FN+TP+FP)).

In OGTSite results, only serines and threonines which are predicted as *O*-GlcNAcylated are shown. Thus, to calculate the number of TN, we calculated the total number of serines and threonines in our dataset and subtracted FN, TP and FP from it. The total number is 39,215. For the other tools, the "show all serine and threonine" option was available.

Features

Sequence – Structural and Polarity Classification

First of all, each amino acid around the sites in a window of -/+10 residues in the sequence were translated into size (Table 1A) and polarity (Table 1B) classes with Python v3.6 scripts. For the size class, we chose to focus on the nature and the length of each residue totaling 8 classes, whereas for the polarity class, we considered physico-chemical properties of amino acids, totaling 9 classes.

We calculated the proportions of each class at each position in the -/+10 windows and compared them to a random composition of residues of all mammal sequences (reviewed only) retrieved from UniProt (82,495 sequences).

The Chi Square tests were performed after calculating the number of individuals from the proportions. For the case of the random set, these values are theoretical and correspond to the number of individuals that would have been observed.

Sequence – Flexibility Prediction

The flexibility of each site was predicted with the DynaMine tool that only requires a sequence as input.[12] A S^2 score is provided for each residue. It is lying between 0 and 1 where score inferior to 0.69 is considered as flexible, superior to 0.8 considered as rigid and between these two values there is a twilight zone called context dependent. The results were parsed with a homemade Python v3.6 script to extract the flexibility score of each site, for positive and negative data, and depending on the nature of the site (serine or threonine).

Table 1 Definition of the Classes of Amino Acids

A.	
Sidechain Size Class	**Residues**
No residue (Empty) (E)	NA
Glycine (G)	Gly
Very Small (V)	Ala, Val
Small (S)	Ser, Thr, Ile, Leu, Cys
Normal (N)	Asp, Asn, Glu, Gln, Met
Long (L)	Arg, Lys
Aromatic (A)	Phe, Trp, Tyr, His
Proline (P)	Pro
B.	
Polarity Class	**Residues**
Polar uncharged with hydroxyl group (A)	Ser, Thr
Polar uncharged with amide (B)	Asn, Gln
Positively charged polar (C)	Arg, Lys, His
Negatively charged polar (D)	Asp, Glu
Non-polar suffered (E)	Met, Cys
Non-polar aromatic (F)	Tyr, Phe, Trp
Non-polar aliphatic (G)	Ala, Val, Leu, Ile, Pro
Glycine (H)	Gly
No residue (I)	NA

Abbreviations: Gly, Glycine; Ala, Alanine; Val, Valine; Ser, Serine; Thr, Threonine; Ile, Isoleucine; Leu, Leucine; Cys, Cysteine; Asp, Aspartic Acid; Asn, Asparagine; Glu, Glutamic Acid; Gln, Glutamine; Met, Methionine; Arg, Arginine; Lys, Lysine; Phe, Phenylalanine; Trp, Tryptophan; Tyr, Tyrosine; His, Histidine; Pro, Proline.

Sequence – Secondary Structure and Angles Predictions

For every site, secondary structures were predicted using two software: SPIDER3 and PSSpred, run locally.[13,14] Because the predictions of both were very close to each other, the results were only retrieved from SPIDER3 and stored in a file with all the other parameters.

φ and ψ angles of residues in a −3/+2 window were also retrieved from SPIDER3 predictions, classified as follows: β-strands (−160° < phi < −50° and 100° < psi < 180°), α-helix (−160° < phi < −50° and −60° < psi < 20°) and other.

Structure – Models

To calculate the accessibility of each site, a structure for each protein was needed. As all proteins do not have an available structure in the Protein Data Bank or structures at site locations were missing, we modeled all of them with I-TASSER v5.1 (default parameters),[15] installed and run locally on the HPC cluster of the Mesocenter of the University of Lille. We chose I-TASSER, which combines threading and *de novo* modeling, because it was ranked as the best structure modeling server according to the Critical Assessment of Techniques for Protein Structure Prediction (CASP) for rounds 7 to 14.[16–18] Because the tool is limited to 1500 amino acids, we used a sequence window of 1500 residues for those cases where the sequence is longer, ensuring the maximum number of residues at either side of the *O*-GlcNAcylation site.

Structure – Accessibility to Solvent

NAccess v2.1.1[19] was used with default parameters to compute the accessibility to solvent of the hydroxyl group of serines or threonines. ElNémo (v10/18/2018)[20] was used to create 10 normal modes, from −100 to 100 perturbations by step of 20, from each model obtained by I-TASSER: the 10 largest modes with 10 structures for each of them were kept to render the elasticity of the mode, which produced 100 structures per protein. Then, a homemade Python 3.6 script was written to launch NAccess on each model and their modes to calculate the accessibility. The maximum of accessibility for the 100 structures was conserved for each site.

Machine Learning

To train and test our model, we divided our dataset into two sets: a training set which represents 80% of the dataset and a testing set which represents the remaining 20% of the dataset. As the amount of positive data is much lower than the amount of negative data, we performed an oversampling of the training set and an undersampling of the testing set. Oversampling and undersampling were done with an algorithm for non-continuous data called Random Over Sampling Examples (R package "ROSE", v0.0.3).[21] We also tested on real proportions data (1.4% *O*-GlcNAcylated vs 98.6% non *O*-GlcNAcylated) in order to compare the results with the undersampled testing set (Figure 1). We used three types of machine learning algorithms:

- Random Forest (R package "randomForest", v4.6.14).[22] To optimize the parameterization, we trained data on different numbers of trees: once a plateau was found, the first value of this plateau was chosen. Each tree has nodes to test different features. The number of variables tested at each node was chosen to get the best predictions. The finally selected parameters are: ntree = 200 (number of trees), mtry = 3 (number of variables tested at each division).
- Gradient Boosting Tree (GBT) ("xgboost" R package "xgboost", v1.1.1.1)[23] that uses decision trees like Random Forest but including a new variable which is residuals.[24] The difference between residuals and the real value to predict is used in the algorithm.
- Support Vector Machine (SVM) is an algorithm essentially based on prediction of two classes (R package "e1071", v1.7.3).[25] Each data value is set in a matrix with as many dimensions as features. Then the algorithm tries to find a plan to separate the positive from the negative data. The more the values are away from this plan, the better the prediction is. We ran four SVM algorithms based on different functions: Linear, Polynomial, Radial basis and Sigmoid which are used to create the plan to classify the two classes. We first used default parameters for all SVM algorithms; then, we investigated hyperparameter tuning for the four algorithms with the "tune" function of the "e1071" package, which gave the hyperparameters cost = 4 and gamma =1. We only present the results of the hyperparameterized sigmoid because this is the one that showed the best results.

Details about each algorithm can be found in.[24]

To be able to run the machine learning algorithms, the features listed before were transformed to numeric values. All the transformed parameters are listed below:

- Side Chain length: 0, 1, 2, 3, 4, 5, 6 or 7 where 0 is No Residue, 1 is Glycine, 2 Very Small, 3 Small, 4 Normal, 5 Long, 6 Cycle and 7 Proline from positions −1 to +5
- Non-polar aliphatic amino acids from positions −3 to −1: 0, 1, 2 or 3
- Polar positively charged residues from positions −7 to −5: 0, 1, 2 or 3
- Number of serines and threonines in the -/+10 residue window
- Flexibility: continuous value from 0 to 1 where 0 is flexible and 1 rigid

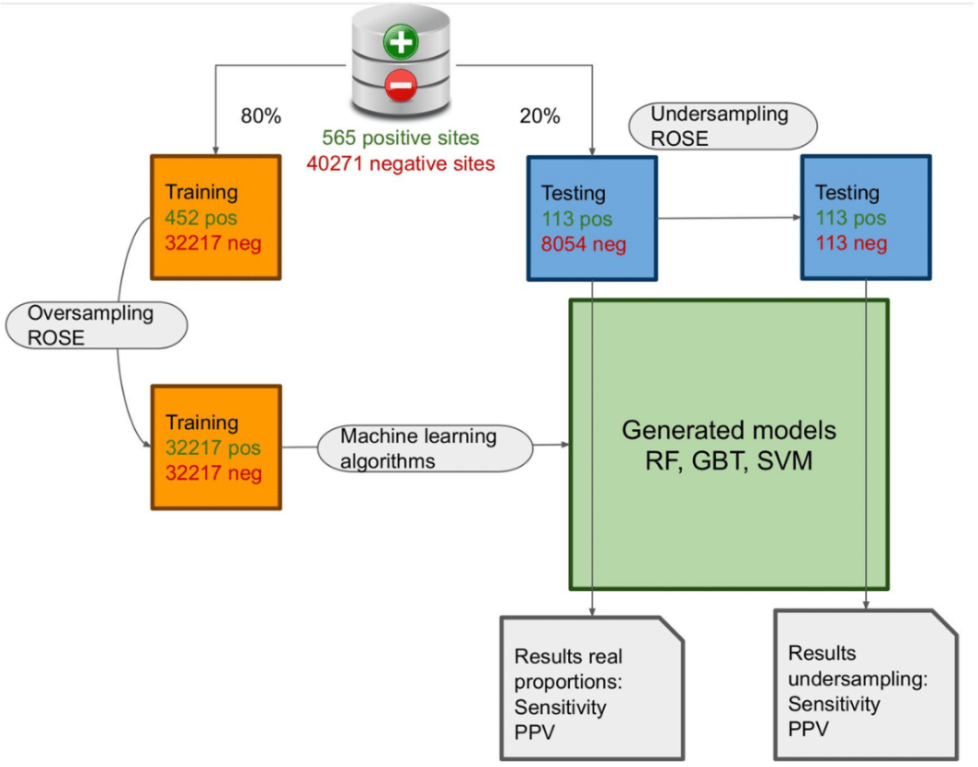


Figure 1 Steps of the machine learning training and testing. Machine learning process pipeline representation with over and undersampling to create the training/testing data and the various models from the different algorithms.

- Secondary structure: 0, 1 or 2 where 0 is not structured, 1 is alpha helix and 2 is beta strand
- Presence of a proline in +1: 0 or 1 (no or yes)
- Secondary structure according to phi and psi angles (0, 1 or 2)
- Nature of the site: 0 or 1 where 0 is serine and 1 threonine

To test the different algorithms we compared the methods according to the sensitivity and the PPV. We ran each algorithm ten times with ten randomly shuffled datasets (80% to train and 20% to test) and computed the statistics for these ten runs.

To ensure the added value of each feature, MRMD V3.0[26] was used regarding its five methods: PageRank, LeaderRank, TrustRank, Hist_a and Hist_h.

Everything in this section was done with R v3.6.3.

OGT Partners Analysis

The OGT partners were retrieved from the IMex database through the PSICQUIC[27] service inside the Cytoscape[28,29] v3.7 software for network visualization and analysis. The enrichment analysis in molecular function of the Gene Ontology was performed with ClueGO v2.5.7[30] and the EBI GOA (v23/07/2020). The selection criteria were:

- Statistical Test Used = Enrichment/Depletion (Two-sided hypergeometric test)
- Correction Method Used = Bonferroni step down
- Min GO Level = 2
- Max GO Level = 6
- Cluster #1
- Sample File Name = Network selection: ManuallyAddedOrModifiedIDs
- Min number of Genes = 3
- Min Percentage of Genes = 3.0
- GO Fusion = false
- GO Group = true
- Kappa Score Threshold = 0.4
- Over View Term = SmallestPValue
- Group By Kappa Statistics = true
- Initial Group Size = 1

- Sharing Group Percentage = 50.0

Availability of Data

Code and data used in the manuscript are all available in the GIT repository: https://gitlab.in2p3.fr/cmsb-public/OGP.

Results

Evaluation of Available Prediction Tools on a Newly Built Dataset

We built a new dataset with only experimentally proven *O*-GlcNAcylated sites, ignoring sites identified by homology in order to avoid inclusion of false positives (see Materials and Methods). We obtained a dataset of 565 *O*-GlcNAcylation sites and 40,271 serine or threonine residues that are not *O*-GlcNAcylated (data set provided in the git repository). This means that only ~1.4% of all S and T residues of *O*-GlcNAcylated proteins are *O*-GlcNAcylated. We refer to *O*-GlcNAcylated sites as positive data and non *O*-GlcNAcylated sites as negative data.

Because the YinOYang and *O*-GlcNAcPred-II prediction tools are limited in terms of protein sequence size, we built a reduced dataset that we used to run each tool. The results of our evaluation are listed in Table 2.

These results show that, although acceptable values are obtained by the tools for the specificity, the sensitivity and the accuracy, the values are lowered with respect to those previously published[6–8] when applied to our new dataset. However, these criteria remain limited in determining which sites are truly *O*-GlcNAcylated. Another indicator such as the precision, also called Positive Predictive Value (PPV), is more useful. PPV is the chance that a positive prediction is right. In our analysis, we observed that the PPV is very low for any tool, the best one showing only 8.68% (most stringent YinOYang), which means that a site predicted as positive has less than 9% chances to be really *O*-GlcNAcylated.

The Negative Predictive Value (NPV) is the chance that a negatively predicted site is factually not *O*-GlcNAcylated. For each tool, values around 99% are found. However, considering that the percentage of non *O*-GlcNAcylated serines and threonines is ~98.6% (100 − 1.4), the tools perform only marginally better than a random prediction.

In conclusion, the more relevant criterion in a prediction tool is its capacity to identify the true positive sites, quantified in the PPV. Since these are found to be so low for any of the currently available tools, we attempted to improve the predictions using new features. So far, all tools are based on the protein primary structure (sequence) around each site only. Consequently, we decided to first characterize the primary structure of all the

Table 2 Evaluation of Commonly Used Methods for *O*-GlcNAcylated Sites Prediction on Our Dataset

	YoY +	YoY ++	YoY +++	YoY ++++	OGP-II	OGT Site
TP	267	172	79	21	358	270
FP	8158	3233	1068	221	8830	4084
TN	30507	35432	37597	38444	29835	34581
FN	283	378	471	529	192	280
Sensitivity (%)	48.55	31.27	14.36	3.82	65.09 (81.05)	49.09 (85.4)
Specificity (%)	78.97	91.67	97.25	99.43	77.16 (95.91)	89.44 (84.1)
Precision (PPV) (%)	3.17	5.05	6.89	8.68	3.90	6.20
NPV (%)	99.08	98.95	98.78	98.65	99.36	99.20
Accuracy (%)	78.55	90.47	96.04	98.32	76.99 (91.43)	88.87 (84.7)
FDR (%)	96.83	94.95	93.11	91.32	96.10	93.80
Total	39215	39215	39215	39215	39215	39215

Notes: Table showing the statistical measures of YinOYang (with different stringency thresholds, the higher number of "+", the more stringent), *O*-GlcNAc-Pred II and OGTSite. When available, published performances of software on their data are put in brackets.

Abbreviations: YoY, YinOYang; OGP-II, *O*-GlcNAcPred II; TP, True Positives; FP, False Positives; TN, True Negatives; FN, False Negative; PPV, Positive Predictive Value; NPV, Negative Predictive Value; FDR, False Detection Rate.

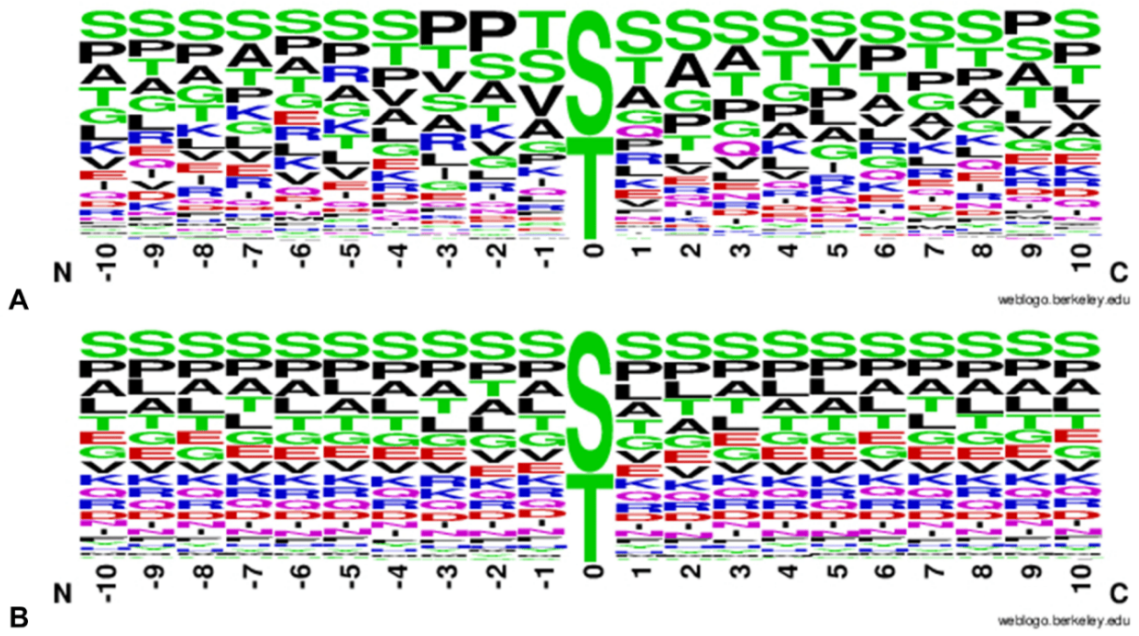


Figure 2 WebLogo representing the proportion of amino acids around sites. WebLogo representing the proportion of each amino acid in a -/+ 10 frame around (**A**) *O*-GlcNAcylated sites and (**B**) non *O*-GlcNAcylated sites.

O-GlcNAcylated sites and then to expand our analysis to secondary and tertiary structures in order to determine relevant features for new prediction tools.

Analysis of Sequences Around the Sites

We analysed first the *O*-GlcNAcylated site sequences over a window of −10 to +10 amino acid residues around each *O*-GlcNAcylated site to keep the maximum of available information without exaggeration. For all the windows, we compared the composition in residues between positive and negative sites in order to highlight over- and under-represented residues in both sets.

Figure 2 shows the proportion of amino acids at each position of the window in the positive (Figure 2A) and negative (Figure 2B) sets. Despite some tendencies, no clear patterns are discernable, although most of the residues around positive sites do show a small side chain. Also, as already described by Leney et al,[31] we can see a slightly lower amount of proline residues in +1 for *O*-GlcNAcylated sites. The authors explain this observation by the crossplay between *O*-GlcNAcylation and phosphorylation. Indeed, some kinases are proline-directed and the presence of a proline at +1 favors phosphorylation over *O*-GlcNAcylation. We cannot also exclude the fact that a proline induces a steric hindrance due to its cycle that could hinder the transfer of the *O*-GlcNAc by the OGT. However, this hypothesis is unfavoured as a proline residue is frequently found at −1 and −2 of an *O*-GlcNAcylation site.

To assess if the size of the side chain is a pertinent criterion for prediction of *O*-GlcNAcylation, we classified residues in function of the size of their side chain. We also classified the residues depending on their polarity and evaluated this criterion as well. In both cases, we compared the proportions to those observed in a random sequence. These classes are listed in Table 1A and B.

Figure 3 shows that the proportions of all the positions in the negative set look more homogeneous between them (Figure 3B) than the proportions of all the positions of the positive set (Figure 3A). Comparing the proportions of the positive set to the negative set and to the proportions of a random sequence, we can see that *O*-GlcNAcylated sites show a light tendency towards the shorter amino acids in their immediate vicinity. The area −1 to +5 in particular is of special interest as the sum of the proportions of the classes E, G, V and S at each of these positions in the positive set is at least 5% higher to the negative set or the random set. In addition, this figure shows that a serine or threonine close to N-terminal or C-terminal positions has a higher probability to be *O*-GlcNAcylated as the percentage of the empty class is overall double for the positive data compared to the negative data.

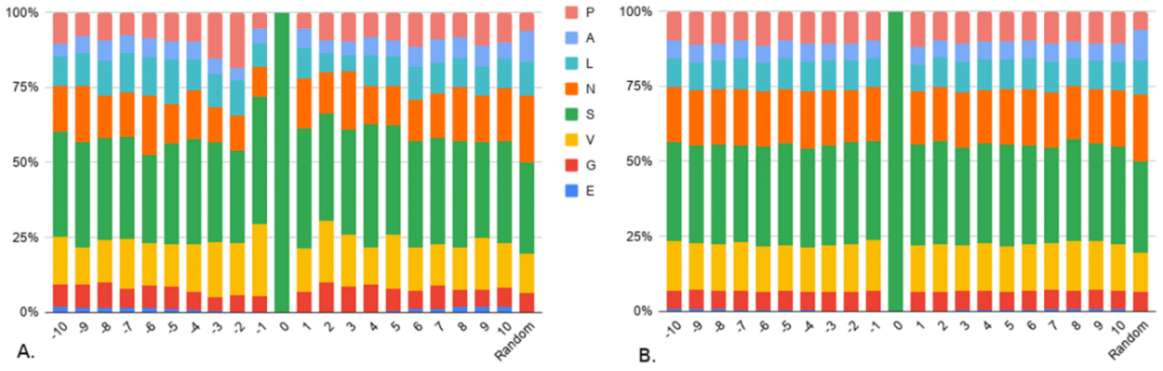


Figure 3 Composition in side chain size classes. Composition of side chain size classes around (**A**) *O*-GlcNAcylated sites and (**B**) non *O*-GlcNAcylated sites. Classes are detailed in Table 1A. Random corresponds to the composition of any position in a random sequence from UniProt.

When considering polarity, Figure 4 also shows that the proportions of all the positions in the negative set (Figure 4B) look more homogeneous than in the positive set (Figure 4A). Here, two classes are over-represented in the positive set compared to the negative set in two areas: the non-polar aliphatic residues (G), in positions −3 to −1 (respectively, 13.1%, 10.5% and 3.5% higher) and the polar positively charged residues (C) in positions −7 to −5 (respectively, 2.7%, 2.3% and 4% higher). But even if over-represented, not all the sites follow these distributions, which means that these criteria show slight tendencies but are not sufficient to discriminate positive from negative data. The majority of the classes of the non *O*-GlcNAcylated sites look very close to the random composition. However, running a Chi-square test of proportions between the mean of all the positions of the negative set and the random set shows that they are significantly different (p-value<0.0001). It is essentially due to the amino acid A class (10% higher on average in the negative set), which is the one gathering together serine and threonine residues. The difference in proportions is even higher for the positive set compared to random, also mainly (but not only) due to the A class (14.1% higher on average).

Therefore, globally a −10/+10 window around each serine or threonine in the positive and negative sets contains more serine and threonine residues than random, which means that they tend to be clustered. Intriguingly, the number of serines and threonines seem to be higher around positive than negative sites. Thus, we counted the number of serines and threonines around the sites (without the proper site) for the two classes (Figure 5).

Even if the mean and median of S/T is higher in the *O*-GlcNAcylated sites (Mann–Whitney test with a p-value <2.2e-16), the standard deviation is high and the distributions show a large overlap, which means that positive sites, like negative sites, can show poor or high densities of S/T, which makes this criterion alone not stringent enough to differentiate between positive and negative sets.

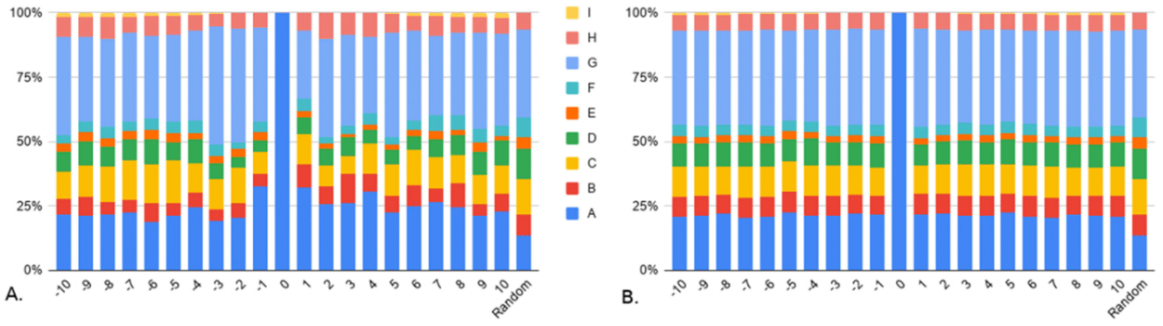


Figure 4 Composition in polarity classes. Composition of polarity classes around (**A**) *O*-GlcNAcylated sites and (**B**) non *O*-GlcNAcylated sites. Classes are detailed in Table 1B. Random corresponds to the composition of any position in a random sequence from UniProt.

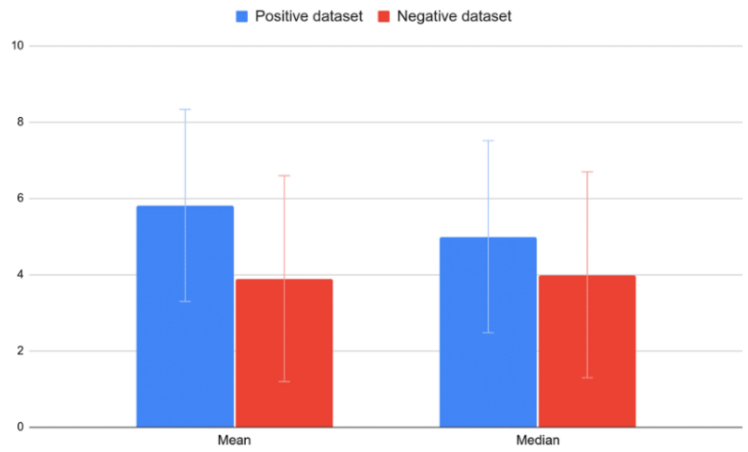


Figure 5 Number of serine and threonine residues around positive and negative sites. Histograms representing the mean and median of the number of serine and threonine residues around positive (blue) and negative (red) sites.

Unfortunately, these criteria show tendencies, but are not discriminative. Additional features will be required in order to enhance the predictive power of our approach. As *O*-GlcNAcylation is a dynamic modification, we hypothesised that the sites should be flexible and accessible. We therefore investigated the secondary structure, flexibility, tertiary structure and solvent accessibility.

Analysis of Secondary and Tertiary Structures Around the Sites

To predict the backbone flexibility, we used the DynaMine software.[12] Figure 6 shows that only 15.27% of all the *O*-GlcNAcylated sites are predicted in rigid regions, 43.7% in flexible regions and 41.03% in context-dependent regions. These proportions are very similar to

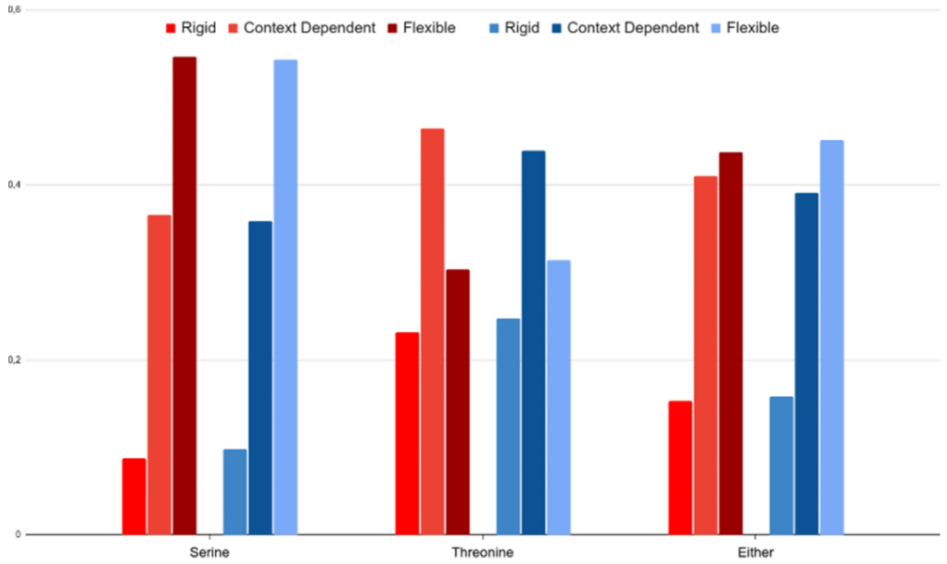


Figure 6 Predictive flexibility of *O*-GlcNAcylated sites and non *O*-GlcNAcylated sites. Flexibility predicted with DynaMine for positive (red) and negative (blue) datasets depending on the nature of the site: serine, threonine, or both.

the negative dataset with, respectively, 15.78%, 45.13% and 39.09%. By making the distinction between serine and threonine, we find that serine residues are more often found in flexible regions than the threonine ones, independently of whether a site is (positive) or is not (negative) glycosylated and on the contrary that less serines are found in rigid regions. But here, the positive and negative distributions are so close that this feature is not discriminative enough. We further anyway conserved it for machine learning prediction in a first round, in case the very light differences between the positive and negative flexibility predictions could be useful in combination with other features.

Secondary structures were predicted with SPIDER3, which also computes φ and ψ angles.[13] Around 85% of *O*-GlcNAcylated sites were predicted on unstructured portions of proteins, 10% on α-helix and 5% on β-strands. For non *O*-GlcNAcylated sites, the percentage of each class is 64%, 28% and 8% for unstructured, α-helix and β-sheet, respectively, which means that *O*-GlcNAcylated sites are found preferentially in unstructured parts. These results also support the hypothesis that the structure of the substrate should have a small steric hindrance to enter the catalytic pocket of the OGT. But, obviously, this does not allow for a prediction of *O*-GlcNAcylated sites only based on secondary structure predictions. According to Pathak et al, the backbone symmetry of *O*-GlcNAcylated sites is like β-strands in the residue range −3 to +2.[32] We used the SPIDER3 predictions to check whether the angles actually correspond to β-strands. The most frequent class for each window of 6 residues was kept to classify a site. Table 3 shows that *O*-GlcNAcylated backbone sites from −3 to +2 are not that different from non *O*-GlcNAcylated sites. The proportion of β-like backbone in the −3/+2 area is slightly higher for the positive sites but here again, the signal is not strong enough to differentiate these two classes.

We hypothesised that the sites should be able to dive into the catalytic pocket of the enzyme. Accordingly, modified sites should be globally accessible to solvent. As a consequence, we computed the accessibility to solvent of the *O*-GlcNAcylated sites. This calculation required the three-dimensional structure of each protein. However, not all proteins of our dataset possess a structure listed in the Protein DataBank (PDB), and even for those for which a structure was available, the area which contained the sites was often missing, undoubtedly due to the intrinsic flexibility of sites prone to *O*-GlcNAcylation. Therefore, we decided to build modeled structures. Because for a part of the proteins any good template was available, we used I-TASSER. We then calculated the accessibility of all sites in the models using NAccess.[19] We also computed normal modes for each structure with the ElNémo software to take into account the potential elasticity of the molecule, which may improve the accessibility of each site.[20] But despite some elasticity of the proteins taken into account, the accessibility of each site was very heterogeneous. Figure S1 shows the accessibility for all the sites; ie, the sites of the human CR2 (P20023) and AQP1 (P29972) proteins are totally accessible while others like the mouse Psma7 (Q9Z2U0) and Psma5 (Q9Z2U1) protein sites are not accessible at all. Therefore, in contrast to what we initially thought, we had to discard the use of accessibility for prediction altogether.

Table 3 Percentage of β-Like, α-Like and Other Backbone Angles from −3 to +2 of *O*-GlcNAcylated Sites (Positive) and Non *O*-GlcNAcylated Sites (Negative)

	Positive	Negative
Beta-like	61,91%	59,36%
Alpha-like	29,65%	27,11%
Other	8,44%	13,53%

Note: Predictions made with SPIDER 3.

Integration of the Features in Machine Learning Algorithms

We studied different aspects of the sites, comparing parameters between the *O*-GlcNAcylated sites and non *O*-GlcNAcylated sites of our dataset, from the primary to tertiary structure. None of these parameters are sufficient to differentiate positive from negative data despite some tendencies. Consequently, to find a way to predict *O*-GlcNAcylation sites, we used them as features in three different types of machine learning (ML) algorithms. We decided not to include the accessibility since it appeared to be unreliable and moreover relies on models of tertiary structures and not on resolved ones. The three types of algorithms we chose are Random Forest (RF), Gradient Boosting Tree (GBT) and Support Vector Machine (SVM),[22,24,33] the latter being split into 4 variants (linear, polynomial, radial basis and sigmoidal). We initially chose a RF algorithm because algorithms based on decision trees are well adapted to treat a mixture of numerical and categorical features and to deal with fuzzy input data (outliers, irrelevant inputs). We then evaluated GBT, which is also based on decision trees and is sometimes found to outperform RF for hard classification problems.[24]

We further tried SVM because it is not based on trees and it is a well-known classifier in bioinformatics that has already been used for *O*-GlcNAcylation prediction.[8,34]

Our dataset was divided into two: one for training and one for testing. Following Box and Meyer (1986),[35] we used 80% of the data for training and the remaining 20% for testing. As the amount of positive and negative data in the set differs by three orders of magnitude, this could bias the training step. To counter this fact, we used an oversampling method, which consists of adding points in the space representation (created by the features) of positive data to make the amount of positive data equal to the amount of negative data[5] for the training (32,217 sites). And we used an undersampling method which consists of choosing randomly the same number of negative data in the set as positive data for the testing (113 sites). The process is described in Figure 1.

Table 4 shows the sensitivity and the PPV for ten runs of each algorithm for testing on undersampled data on one side and not undersampled data (with real proportions of positive vs negative) on the other side. We chose these two measures to be able to compare them with those of YinOYang, *O*-GlcNAc-Pred II and OGTSite. For the testing set based on undersampled data (50% positive/50% negative), this table demonstrates that Random Forest is the best algorithm based on the mean values of sensitivity followed by the GBT algorithm. The radial basis SVM algorithm follows in third place. Following the PPV, it shows the best prediction algorithm to be GBT while looking at the maximum. Nevertheless, the GBT gives

Table 4 Sensitivity and PPV of the Three ML Algorithms Tested on Undersampled (Equal) and Not Undersampled (Real) Data

	Min (Equal/Real)	Max (Equal/Real)	Mean (Equal/Real)	Median (Equal/Real)	Standard Deviation (Equal/Real)
RF (Sensitivity %)	97.35 **97.35**	99.12 **100**	98.58 **98.67**	98.58 **98.58**	0.62 **1.04**
RF (PPV %)	47.46 **1.35**	51.61 **1.39**	48.98 **1.37**	48.69 **1.37**	1.36 **0**
GBT (Sensitivity)	13.64 **30.97**	82.30 **47.79**	48.76 **39.56**	49.11 **39.82**	32.18 **5.71**
GBT (PPV)	13.51 **2.52**	86.11 **3.41**	47.04 **3.06**	46.08 **3.10**	27.47 **0.29**
SVM Linear (Sensitivity)	29.20 **32.74**	51.33 **47.79**	37.70 **39.73**	38.94 **38.94**	6.82 **5.54**
SVM Linear (PPV)	31.13 **0.81**	43.94 **1.08**	36.53 **0.90**	37.33 **0.90**	4.13 **0.01**
SVM Polynomial (Sensitivity)	8.90 **8.90**	98.23 **23.48**	36.81 **12.69**	15.93 **10.93**	42.11 **5.53**
SVM Polynomial (PPV)	11.11 **0.20**	48.88 **1.37**	30.79 **0.88**	29.16 **0.92**	14.62 **0.45**
SVM Radial basis (Sensitivity)	34.51 **34.51**	53.10 **50.44**	42.74 **40.97**	42.92 **39.38**	5.99 **5.47**
SVM Radial basis (PPV)	33.05 **0.71**	43.48 **1.01**	37.76 **0.82**	37.59 **0.78**	3.70 **0.09**
SVM Sigmoid (Sensitivity)	28.32 **33.63**	53.10 **49.56**	39.12 **39.65**	38.94 **38.94**	7.56 **6.14**
SVM Sigmoid (PPV)	28.32 **0.81**	53.10 **1.11**	39.12 **0.88**	38.94 **0.85**	4.34 **0.10**

Notes: Undersampled testing data contains the same number of positive vs negative data (50%/50%) whereas not undersampled data contains real proportions (1.4%/98.6%). Statistics that correspond to real data are set in bold. Blue background contains results for sensitivity, white background for PPV. Values are indicated in %.
Abbreviations: PPV, Positive Predictive Value; ML, Machine Learning; RF, Random Forest; GBT, Gradient Boosting Tree; SVM, Support Vector Machine.

greater heterogeneous results with a standard deviation of 32.18% for the sensitivity and 27.47% for the PPV and, on average, the Random Forest performs better on both tables. We then ran the same algorithms without the features which result from a prediction, namely flexibility, secondary structure and φ/ψ angles, to avoid the potential background noise inherent to any prediction. The results are equivalent to the previous results or less good, showing that including all features is preferred (Table S1).

The results we get when testing on undersampled data are better than the ones of any other tools (Table 2). Nevertheless, this testing set is not representative of the reality where we have less than 2% of serine and threonine residues that can be *O*-GlcNAcylated. Thereby, we ran ten times the same algorithms but tested them on real proportions of positive and negative data (113 positive sites, ie, 1.4%, and 8054 negative sites, ie, 98.6%), thus without undersampling.

The results presented for real proportions are in bold in Table 4. They are not as good as the previous ones. The sensitivity decreased except for the Random Forest algorithm but looking at the PPV values, they all decreased drastically to around 1% except for the GBT algorithm which is around 3%, a value close to other already existing *O*-GlcNAcylation prediction tools. For the sensitivity, RF gives really good results but the amount of false positives is very high, which explains the low PPV. Once again, we ran the predictions with the three algorithms on data without the features based on predictions, and obtained similar or slightly worse results (Table S1).

Yet, we tuned the SVM sigmoid algorithm, which gave the best results among the SVM algorithms, with hyperparameterization. Here, the best hyperparameters were cost equals 4 and gamma equals 1. Once the hyperparameters set, the results of the SVM sigmoid algorithm slightly increased but stayed lower than GBT and Random Forest (Table S1).

Thus, currently available tools are not as efficient as claimed, considering the PPV (best PPV is lower than 9%). As we tried to improve predictions with the various machine learning algorithms, optimizing the features, we failed to get better results when running them on data with real proportions (our best PPV is around 3%).

Discussion

Predicting *O*-GlcNAcylation is a tricky task: unlike other PTMs, such as phosphorylation or *N*-glycosylation, there is no common pattern or consensus sequences as well as limited experimentally validated data. Much effort has been spent attempting to develop software to solve this problem but we showed here that their predictive power is disappointing. Subsequently, we tried to improve the predictions using machine learning algorithms by adding characterized features based on primary to tertiary structure information. Although they show better results on undersampled data, they fail to give good results on data with realistic proportions of *O*-GlcNAcylated serine and threonine residues (~1.4% *O*-GlcNAcylated/98.6% non *O*-GlcNAcylated).

The published tools are all based on algorithms trained on protein primary sequences. Because there is only a limited amount of experimentally proven data available, the training of an algorithm to predict sites is very hard. Some of the currently available tools therefore use the prediction of *O*-GlcNAcylated sites by homology. This provides more data for the training, but the data are intrinsically biased, which may explain the poor results we get when running them on our dataset. Precisely, regarding the results of the different tools, we pointed out that the statistical quantities used to compare their efficiency are not as significant as they should be. Looking at the biological problem, the statistical measurements used here must be improved by taking into account the number of false positives. Sensitivity is a major statistical measurement to see if some positive sites are missed which is interesting when data are balanced. Yet, it is a quantity to take carefully into account because if a tool simply predicted all the sites as positive, this statistic would be maximized, but the tool would be pointless. Looking at the Positive Predictive Value (PPV) (also called precision) is more relevant because it shows the proportion of correct positive predictions compared to all the positive predictions, which is exactly what a researcher wants to know when he/she performs such a prediction. However, currently, no available software succeeds to show a correct PPV value.

All these tools are only primary sequence based. In our sequence analysis, we only showed slight tendencies when comparing various features in a -/+10 window around serines/threonines between positive and negative sets, which makes it hard to classify sites into the two categories. We showed a similarity of composition between non *O*-GlcNAcylation sites and random. The positive data are different from this composition but the difference could result from the low number of data and having a bigger amount could lead to a homogenisation of the composition. A significantly higher number of positive

data may bend the composition to a composition close to the random set.

To bypass this problem of lack of discriminating information from sequences and in order to improve the classification, we characterized additional structural and dynamic parameters derived from the sequence to enrich the features list for machine learning algorithms. Unfortunately, these new features either showed slight tendencies again (secondary structure, dihedral angles) or were without real interest (accessibility), and integrating them into machine learning algorithms did not enhance the predictive power. Derived features like these ones can sometimes improve the training but may also generate a bias due to their predictive aspect, which is why we trained our algorithms also without them. The results obtained thus were close to those obtained with predicted features, but never better. For a better evaluation of the feature selection, we used the MRMD3.0 tool to rank and reduce the features for machine learning.[32] It provides 5 different ranking algorithms. Two of five did not reduce the number of features and the three others reduced by two or three the features. The Hist_h and the LeaderRank methods removed the flexibility and the number of threonines. LeaderRank scoring also removed the presence of proline at +1 while the TrustRank removed the feature length class at position +5. But these removals resulted in a very low ranking score difference with the other kept features (a difference of 0.0001 point), meaning the results would not significantly change when removing them. Therefore, removing features does not improve predictions.

The data retrieved to build our dataset are all experimentally proven by tandem mass spectrometry, which is a conventional method currently used to identify *O*-GlcNAcylated sites, and/or by site-mutagenesis. Although experimentally proven, we cannot exclude that false positives and negatives may exist, which adds noise to the training. The FP and FN may be related to technical issues or experimental conditions, in particular considering that *O*-GlcNAcylation in an *invivo* environment can be more or less efficient than in an *invitro* environment.

The lack of positive data which leads to the limitation of the unbalanced data (which add bias in prediction) can explain the low PPV obtained by our method but also by the previously published tools. We pointed out that the use of equal amounts of positive and negative data in our and other works like *O*-GlcNAcPred-II showed really better results. Still, these proportions are not realistic because only 1% to 2% of the total serine and threonine residues of proteins are truly *O*-GlcNAcylated and, using our best models on realistic proportions, the PPV dropped significantly from 86% for the best GBT model to 3%. Also, considering the sensitivity, a criterion mostly used in the publications of the currently available tools, we obtained better values than the other tools with a Random Forest algorithm. We do, however, consider this difference insignificant because of the large amount of false-positive data. Indeed, our RF gives a sensitivity close to 100% but with a precision ~50% for balanced data and ~1% for real proportionate data, which basically means it classifies all the sites as *O*-GlcNAcylated, making the predictor irrelevant. This problem is linked to the oversampling of trained data but training on undersampled data is not possible in this case because of the lack of positive data (we tried to train RF models on undersampled data, which gave on average a sensitivity = 22.12% and a PPV= 0.54%). It highlights the fact that the amount of available experimental data is crucial for any machine learning algorithm. Thus, it will be worthwhile to apply our methods again when significantly more positive experimental data are available. Here, we showed that the PPV of any tool is currently so low (<9%) that there is a necessity to develop a new and efficient tool based on a bigger dataset of positive data. The RF, GBT and SVM classifier methods we used showed in the past to be very efficient for other predictions,[36–38] and we are convinced now that the problem resides in the data and associated features used, rather than in the algorithms themselves. That is the reason why we did not explore more algorithms. At present, our conclusion is that protein sequence or any sequence-derived information is simply not sufficient for a good prediction and that the amount of positive data is too low for an efficient prediction.

In contrast to its competitive relationship with phosphorylation, *O*-GlcNAcylation is catalyzed by only two antagonistic enzymes, OGT and OGA. A hypothesis is that the OGT requires scaffold partners to be addressed to and to bind its substrates, but also to unfold them (some sites being predicted to be buried in the structures we modeled). To preliminary explore this way, we performed an enrichment of these partners in Gene Ontology - Molecular Functions to identify the proteins which are currently known to be in interaction with OGT.

Figure 7 shows that partners of OGT are involved in protein binding but also that some are involved in unfolded protein binding, which may relate to the

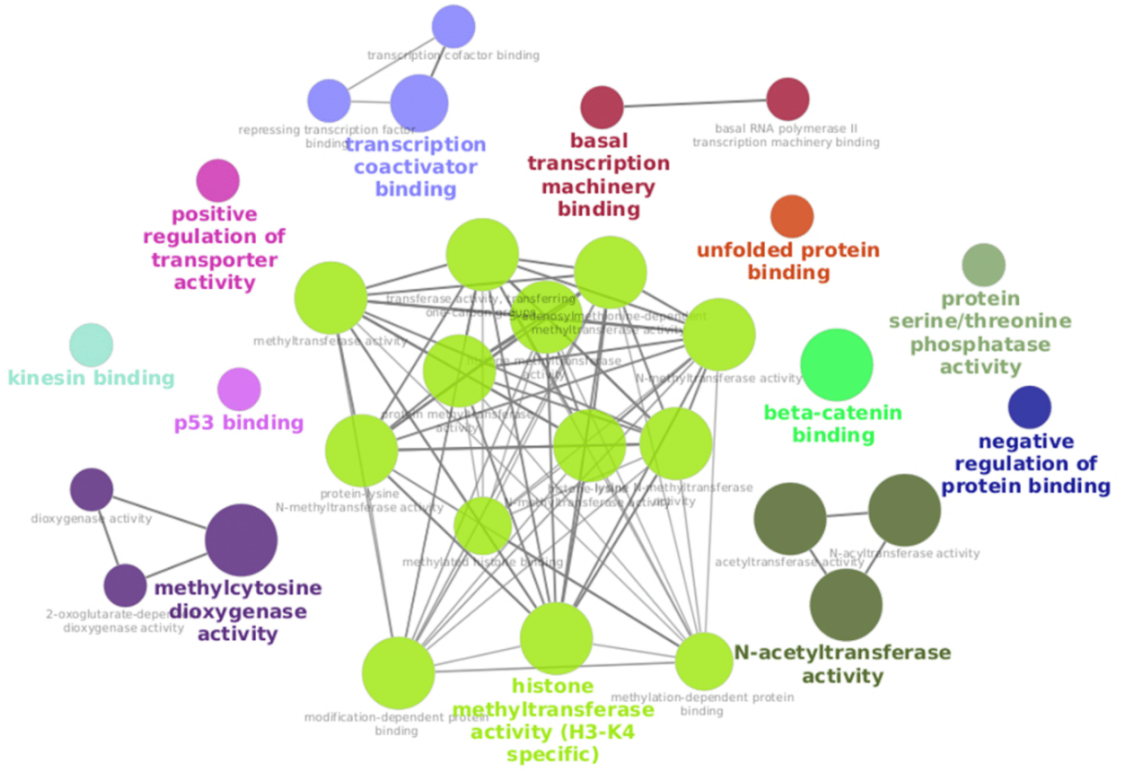


Figure 7 Network of GO terms (Molecular Functions) of partners of OGT. Network visualisation of GO terms (Molecular Functions) of proteins known to interact with the human OGT from the IMex interaction database - enrichment performed with ClueGO (see Material and methods for parameters).

existence of buried *O*-GlcNAcylated sites. This GO class contains 4 proteins: Chaperonin Containing TCP1 (T-Complex Protein 1) Subunit 2 (CCT2), 3 (CCT3) and 5 (CCT5) (a chaperone involved in TCP1),[39] and a Heat Shock Protein Family D Member 1 complex (HSPD1)[40] (a mitochondrial chaperone of imported proteins in the mitochondria). Indeed, these chaperones can bind unfolded proteins to fold them. Their interaction with OGT could participate in the *O*-GlcNAcylation of target proteins before, during or after folding; this subject area would deserve to be deeply studied in a future work for a better understanding of *O*-GlcNAcylation processes. Another intriguing point that could help to improve the predictions if taken into account is that *O*-GlcNAcylation is a reaction that can occur co-translationally,[41] meaning that protein folding and *O*-GlcNAcylation can occur at the same time. These sites should be analysed separately from the post-translational ones to be able to improve predictions but such annotation is currently lacking in databases. Therefore, predicting an *O*-GlcNAcylation as co-translational is an even more difficult task.

OGT presents several isoforms exhibiting different numbers of TPRs in the N-terminal domain. This variation in the number of repetitions could play a role in the enzyme capacity to recognize and modify sites which are on structured or unstructured parts of proteins, with unstructured sections that would be able to enter the lumen of the TPR superhelix. For instance, the structure of human *O*-GlcNAc Transferase (PDB ID: 4N3B)[42] harbors a modified peptide from HCF-1 inside the superhelix of the first TPR repeat, which indicates that OGT is able to accept unfolded structures inside its TPR domain. The ncOGT may thereby be able to accept longer unfolded structures inside its longer TPR repeats domain compared to sOGT. Furthermore, these *O*-GlcNAcylation sites could be presented to OGT with the help of its partners. These interactors could help the enzyme to discriminate between sites to *O*-GlcNAcylate or not. The key for an efficient prediction may thereupon require a detailed study of OGT's partners

and the annotation and classification of each experimentally validated data with its specific context (partners, subcellular compartment of the cell, function of the target …) for a training of the algorithm specific to the context.

Conclusion

Currently available software tools for prediction of sites that can be *O*-GlcNAcylated do not show relevant statistics because sensitivity and specificity do not reflect the capacity of a predictor to provide an unambiguously positive and correct answer. To this purpose, the precision is more adapted and we showed that these tools are less efficient than expected because of the high amount of false positives in their predictions. We tried to improve the prediction methods by characterizing structural and dynamic features such as flexibility, accessibility and secondary structure prediction, but also amino acid composition through the classification of amino acids around positive or negative sites in function of their nature. We found that these features only showed tendencies and that none could be given discriminatory powers. We have combined them in machine learning algorithms, but none of the algorithms succeeded in enhancing the precision. The highest precision currently reached by any algorithm lies below 9%, which makes the *O*-GlcNAcylation prediction an as of yet unattained objective.

Acknowledgments

TM is a recipient of a fellow from the "Ministère de l'Enseignement Supérieur et de la Recherche". We thank the "Centre de Formation par Alternance" of the University of Rouen, the CNRS (Centre National de la Recherche Scientifique) and the University of Lille for their support. We acknowledge support from the High Performance Computing Mesocenter of the University of Lille and the bioinformatics platform bilille for providing access to cluster and cloud computing resources. We thank Thierry Lecroq (University of Rouen, Normandie, France) and the "*O*-GlcNAcylation, signalisation cellulaire et cycle cellulaire" team of the UGSF lab for their advice and discussion in this project.

Disclosure

All the authors report no conflicts of interest in this work.

References

1. Vercoutter-Edouart A-S, El Yazidi-Belkoura I, Guinez C, et al. Detection and identification of O-GlcNAcylated proteins by proteomic approaches. *Proteomics*. 2015;15(5–6):1039–1050. doi:10.1002/pmic.201400326
2. Yang X, Qian K. Protein O-GlcNAcylation: emerging mechanisms and functions. *Nat Rev Mol Cell Biol*. 2017;18(7):452–465. doi:10.1038/nrm.2017.22
3. Aquino-Gil M, Pierce A, Perez-Cervera Y, Zenteno E, Lefebvre T. OGT: a short overview of an enzyme standing out from usual glycosyltransferases. *Biochem Soc Trans*. 2017;45(2):365–370. doi:10.1042/BST20160404
4. Vocadlo DJ. O-GlcNAc processing enzymes: catalytic mechanisms, substrate specificity, and enzyme regulation. *Curr Opin Chem Biol*. 2012;16(5–6):488–497. doi:10.1016/j.cbpa.2012.10.021
5. Bond MR, Hanover JA. A little sugar goes a long way: the cell biology of O-GlcNAc. *J Cell Biol*. 2015;208(7):869–880. doi:10.1083/jcb.201501101
6. Gupta R, Brunak S. Prediction of glycosylation across the human proteome and the correlation to protein function. *Pac Symp Biocomput Pac Symp Biocomput*. 2002;310–322.
7. Jia C, Zuo Y, Zou Q, Hancock J. O-GlcNAcPRED-II: an integrated classification algorithm for identifying O-GlcNAcylation sites based on fuzzy undersampling and a K-means PCA oversampling technique. *Bioinforma Oxf Engl*. 2018;34(12):2029–2036. doi:10.1093/bioinformatics/bty039
8. Kao H-J, Huang C-H, Bretaña NA, et al. A two-layered machine learning method to identify protein O-GlcNAcylation sites with O-GlcNAc transferase substrate motifs. *BMC Bioinform*. 2015;16 (Suppl 18):S10. doi:10.1186/1471-2105-16-S18-S10
9. Pundir S, Martin M, O'Donovan C. UniProt tools. *Curr Protoc Bioinforma Ed Board Andreas Baxevanis Al*. 2016;53:1.29.1–1.29.15. doi:10.1002/0471250953.bi0129s53
10. Liu Y, Wang M, Xi J, Luo F, Li A. PTM-ssMP: a web server for predicting different types of post-translational modification sites using novel site-specific modification profile. *Int J Biol Sci*. 2018;14(8):946–956. doi:10.7150/ijbs.24121
11. Deracinois B, Camoin L, Lambert M, et al. O-GlcNAcylation site mapping by (azide-alkyne) click chemistry and mass spectrometry following intensive fractionation of skeletal muscle cells proteins. *J Proteomics*. 2018;186:83–97. doi:10.1016/j.jprot.2018.07.005
12. Cilia E, Pancsa R, Tompa P, Lenaerts T, Vranken WF. The DynaMine webserver: predicting protein dynamics from sequence. *Nucleic Acids Res*. 2014;42(W1):W264–W270. doi:10.1093/nar/gku270
13. Heffernan R, Yang Y, Paliwal K, Zhou Y, Valencia A. Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility. *Bioinforma Oxf Engl*. 2017;33(18):2842–2849. doi:10.1093/bioinformatics/btx218
14. Yan R, Xu D, Yang J, Walker S, Zhang Y. A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction. *Sci Rep*. 2013;3:2619. doi:10.1038/srep02619
15. Yang J, Zhang Y. I-TASSER server: new development for protein structure and function predictions. *Nucleic Acids Res*. 2015;43(W1):W174–181. doi:10.1093/nar/gkv342
16. Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J. Critical assessment of methods of protein structure prediction (CASP)—round XIII. *Proteins Struct Funct Bioinforma*. 2019;87 (12):1011–1020. doi:10.1002/prot.25823
17. Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A. Critical assessment of methods of protein structure prediction (CASP) — round x. *Proteins*. 2014;82(02):1–6. doi:10.1002/prot.24452
18. Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A. Critical assessment of methods of protein structure prediction (CASP)—round XII. *Proteins Struct Funct Bioinforma*. 2018;86 (S1):7–15. doi:10.1002/prot.25415
19. Hubbard SJ, Thornton JM. *NACCESS*. London: Department of Biochemistry and Molecular Biology, University College; 1993.

20. Suhre K, Sanejouand Y-H. ElNémo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement. *Nucleic Acids Res*. 2004;32(suppl_2):W610–W614. doi:10.1093/nar/gkh368
21. Lunardon N, Menardi G, Torelli N. ROSE: a package for binary imbalanced learning. *R J*. 2014;6:79–89. doi:10.32614/RJ-2014-008
22. Liaw A, Wiener M. Classification and regression by random forest. *Forest*. 2001;23.
23. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. *ACM*. 2016;785–794. doi:10.1145/2939672.2939785
24. Hastie T, Tibshirani R, Friedman J. Boosting and additive trees. In: *The Elements of Statistical Learning. Springer Series in Statistics*. New York: Springer;2009:337–387. doi:10.1007/978-0-387-84858-7_10
25. Dimitriadou E, Hornik K, Leisch F, et al. The e1071 package. 2006.
26. He S, Guo F, Zou Q, Ding H. MRMD2.0: a python tool for machine learning with feature ranking and reduction. *Curr Bioinforma*. 2021;15(10):1213–1221. doi:10.2174/1574893615999200503030350
27. Orchard S, Kerrien S, Abbani S, et al. Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nat Methods*. 2012;9(4):345–350. doi:10.1038/nmeth.1931
28. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 2003;13(11):2498–2504. doi:10.1101/gr.1239303
29. Su G, Morris JH, Demchak B, Bader GD. Biological network exploration with cytoscape 3. *Curr Protoc Bioinforma*. 2014;47(1):8.13.1–8.13.24. doi:10.1002/0471250953.bi0813s47
30. Bindea G, Mlecnik B, Hackl H, et al. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics*. 2009;25(8):1091–1093. doi:10.1093/bioinformatics/btp101
31. Leney AC, El Atmioui D, Wu W, Ovaa H, Heck AJR. Elucidating crosstalk mechanisms between phosphorylation and O-GlcNAcylation. *Proc Natl Acad Sci U S A*. 2017;114(35):E7255–E7261. doi:10.1073/pnas.1620529114
32. Pathak S, Alonso J, Schimpl M, et al. The active site of O-GlcNAc transferase imposes constraints on substrate sequence. *Nat Struct Mol Biol*. 2015;22(9):744–750. doi:10.1038/nsmb.3063
33. Cortes C, Vapnik V. Support-vector networks. *Mach Learn*. 1995;20(3):273–297. doi:10.1007/BF00994018
34. Wang J, Torii M, Liu H, Hart GW, Hu -Z-Z. dbOGAP - an integrated bioinformatics resource for protein O-GlcNAcylation. *BMC Bioinform*. 2011;12(1):91. doi:10.1186/1471-2105-12-91
35. Box GEP, Meyer RD. An analysis for unreplicated fractional factorials. *Technometrics*. 1986;28(1):11–18. doi:10.1080/00401706.1986.10488093
36. Huang S, Cai N, Pacheco PP, Narandes S, Wang Y, Xu W. Applications of Support Vector Machine (SVM) learning in cancer genomics. *Cancer Genom Proteomics*. 2017;15(1):41–51. doi:10.21873/cgp.20063
37. Fan C, Liu D, Huang R, Chen Z, Deng L. PredRSA: a gradient boosted regression trees approach for predicting protein solvent accessibility. *BMC Bioinform*. 2016;17(S1):S8. doi:10.1186/s12859-015-0851-2
38. Hou Q, De Geest PFG, Vranken WF, Heringa J, Feenstra KA. Seeing the trees through the forest: sequence-based homo- and heteromeric protein-protein interaction sites prediction using random forest. *Bioinformatics*. 2017;33(10):1479–1487. doi:10.1093/bioinformatics/btx005
39. Hauri S, Comoglio F, Seimiya M, et al. A high-density map for navigating the human polycomb complexome. *Cell Rep*. 2016;17(2):583–595. doi:10.1016/j.celrep.2016.08.096
40. Fasci D, van Ingen H, Scheltema RA, Heck AJR. Histone interaction landscapes visualized by crosslinking mass spectrometry in intact cell nuclei. *Mol Cell Proteomics*. 2018;17(10):2018–2033. doi:10.1074/mcp.RA118.000924
41. Zhu Y, Liu T-W, Cecioni S, Eskandari R, Zandberg WF, Vocadlo DJ. O-GlcNAc occurs cotranslationally to stabilize nascent polypeptide chains. *Nat Chem Biol*. 2015;11(5):319–325. doi:10.1038/nchembio.1774
42. Lazarus MB, Jiang J, Kapuria V, et al. HCF-1 is cleaved in the active site of O-GlcNAc transferase. *Science*. 2013;342(6163):1235–1239. doi:10.1126/science.1243990

B. Additional results

After this publication, new questions bring new hypotheses. As demonstrated in the publication of Newaz *et al.* (2020), the variety of codons available for the same amino acid might change the conformation of the protein backbone at the output of the ribosome[159]. From this information, we hypothesized the possibility that a certain conformation induced by the nature of the codon impacts the *O*-GlcNAcylation of a serine or threonine. In parallel, results from the analyses of the specific interaction between OGT and beta-catenin described in the next section created a new hypothesis regarding the asparagine ladder inside the TPR domain of the OGT. This ladder has been shown to be important for the catalytic activity of the OGT. Indeed the mutation of the five asparagine into alanine reduces the enzyme activity (see Figure 15 [37]. We can hypothesize that the hydroxyl group of the serine and threonine modified by the addition of the GlcNAc moiety interact with the asparagine from this ladder and that these asparagines pull and push the modification site into the catalytic domain. Hypothesis is also supported by the proximity of a lot of serine and threonine in the neighborhood of the sites [160]. Another perspective highlighted with this publication is the need of chaperon proteins for the OGT to interact with its substrate. This point has been discussed with the interactome of the catalytic enzyme where we identified three chaperon proteins.

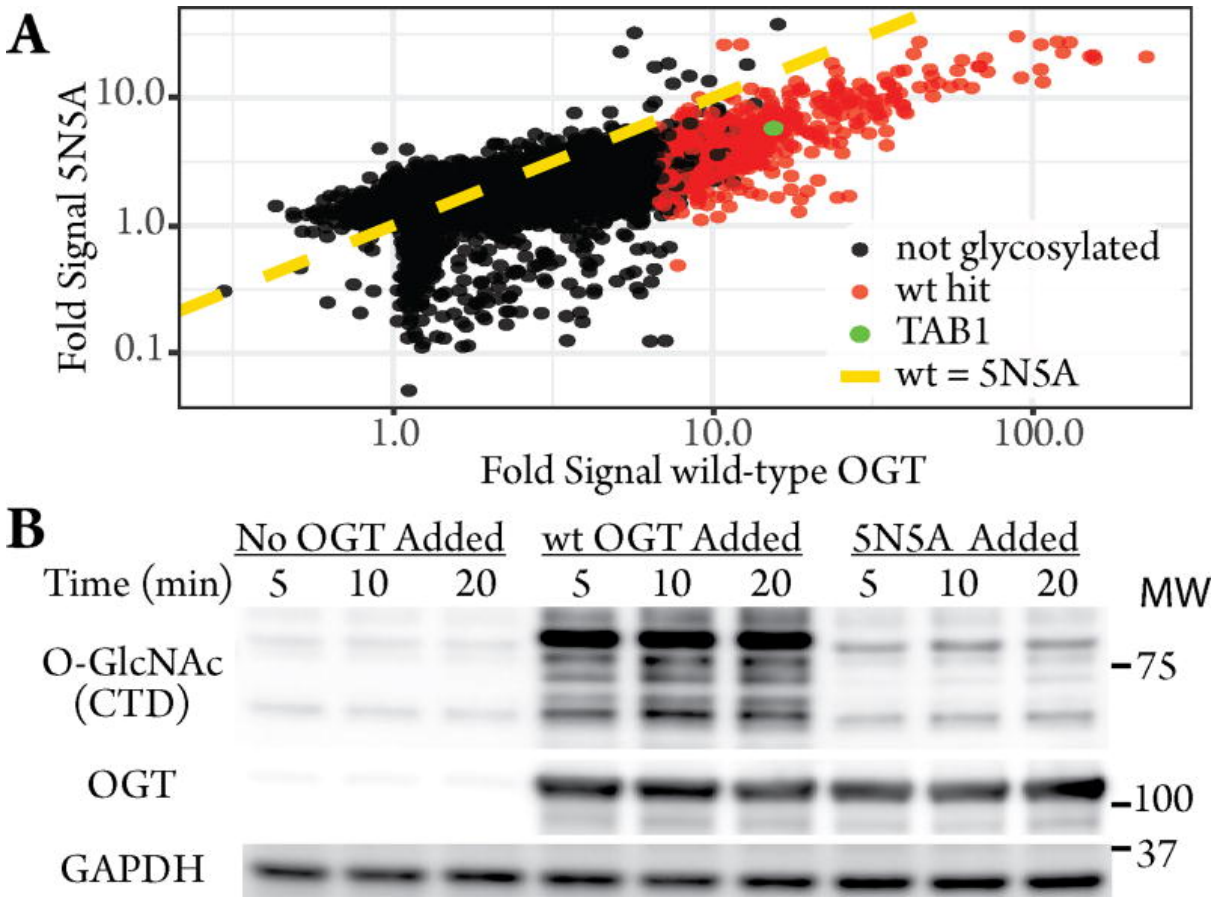


Figure 15. **The asparagine ladder in the TPR lumen is critical for recognition of OGT substrates**

A) Fold signal above control for every protein for wild-type OGT (x-axis) and 5N5A (y-axis) based upon median normalized data. The dashed line represents equivalent activity between wild-type and 5N5A enzymes. Red circles are hits for wild-type enzyme; black circles are proteins that do not score as glycosylated. TAB1, a known poor 5N5A substrate, is high-lighted in green. B) Western blot of HeLa S3 cell extracts incubated with UDP-GlcNAc for indicated times with or without added OGT. OGT and GAPDH blots show enzyme and extract loading; CTD stains for O-GlcNAc. Molecular weight markers are indicated on the right (from Levine *et al.* 2018)

1. Distribution analyses of codons for *O*-GlcNAcylated Sites

a. Material and Method

Datasets:

For this complementary study, two different datasets have been analyzed separately: the one from our study for the *O*-GlcNAcylation prediction site and the second from the new *O*-GlcNAc database developed after the publication [156,160]. For each dataset, protein sequence IDs were retrieved to extract the CoDing Sequence (CDS) if available. If not, the

positive and negative sites associated were removed from the two datasets. Finally, the first dataset was composed of 422 positive sites and 30,760 negative sites, and the second contained 11,745 positive sites and 478,833 negative sites. In parallel, we retrieved all the sequences of reviewed mammal proteins in the Swiss-prot database for a total amount of 3,663,890 serines and threonines. This extraction was done with the searching option of Uniprot and the taxonomy_id:40674 [161].

Extracting positive and negative site from *O*-GlcNAcylated proteins:

The first dataset provided by Mauri *et al.* was in a csv format where each row is an experimentally proven *O*-GlcNAcylated site [160]. This file was constructed as follows: the first column is the Uniprot ID of the protein, the second the index of the positive site of the corresponding sequence and the third one is the protein sequence. From this file it was easy to retrieve the positive sites with a handmade Python script. For the negative sites, this script retrieves every serine and threonine from the protein sequence and if the amino acid does not have the same index as the *O*-GlcNAcylated site it was classified as a negative site.

The second dataset, from the *O*-GlcNAc Database (OGD), was also a csv file, but with more information as follows:

| Uniprot ID | Entry name | Organism | full name | oglcnacscore | oglcnac sites | phosphorylation site | PMIDS | sequence |

From this file it was also possible to retrieve positive and negative as previously with the index of the *O*-GlcNAcylated sites and the protein sequences.

Coding sequence retrieval and codon extraction:

To extract the CoDing Sequences (CDS) of proteins of interest, the first step was to identify the good isoform sequence IDs or the canonical sequence if no isoform was specified. With these IDs, we were able to make requests to the Uniprot server and get the CDS IDs called CCDS ID for Consensus CDS. These new IDs allow us to make requests to the Ensembl server with its Rest API and get the coding nucleotide sequence [162]. From these sequences, with the site indices (positive or negative), the associated codon was retrieved. An additional step has been set up to verify if the codon corresponds correctly to the amino

acid. In total, there are ten different codons: six for serines (AGC, AGT, TCA, TCC, TCG and TCT) and four for threonine (ACA, ACC, ACG and ACT).

Flexibility and secondary structure prediction:

As in our publication at the beginning of this section, we wanted to see the flexibility and secondary structure in function of the different codons.

To predict the flexibility, we used the prediction software Dynamine (available on Bio2Byte tools), which uses a linear regression approach based on experimental Nuclear Magnetic Resonance data. These data highlight chemical shifts of residues in proteins [163]. Dynamine only needs a protein sequence as input and will provide a file with the residue and the flexibility score. This score, called S^2, estimated from experimental data content (NMR chemical shifts) [164]. As described in Mauri *et al.*, this score is between 0 and 1. Between 0 and 0.68 the residue is considered as flexible [160]. Above 0.8 it is predicted to be rigid and between these two scores there is a twilight zone considered as context dependent.

The secondary structure prediction software has been updated since the publication. Here, the SPIDER3-Single has been used[165]!!. This software uses deep learning algorithms based on neural networks and has been trained on 11,192 protein sequences. This software can be used locally and only needs a protein sequence as input. The output is very similar to SPIDER3 used in our publication: the residue, the secondary structure, the Accessibility to Solvent Area (ASA) and the phi, psi and theta angles.

b. Results

Distribution of the codon between *O*-GlcNAcylated and non *O*-GlcNAcylated sites:

The hypothesis was that one or even several particular codons of a serine and threonine would be favored for *O*-GlcNAcylation. To investigate this, every codon was retrieved for positive and negative sites from two different datasets called “Mauri’s dataset” and “*O*-GlcNAc Database dataset”. As the first one has a smaller amount of data but contains only experimentally proven sites, we decided to analyze the two datasets separately, also to

see if there is a difference of distribution between *O*-GlcNAcylated protein and random mammal proteins. The results are summarized in Figure 16. This figure shows a very high distribution similarity of codons between the six possible codons of serine, whether it is modified or not or a random mammal protein sequence. For the distribution of threonine codons we can see also similarity between the 3 categories but a little rise of the propensity of ACC codons for *O*-GlcNAcylated sites from Mauri's dataset. However, this increase is to be explained by the lower number of sites and not significant.

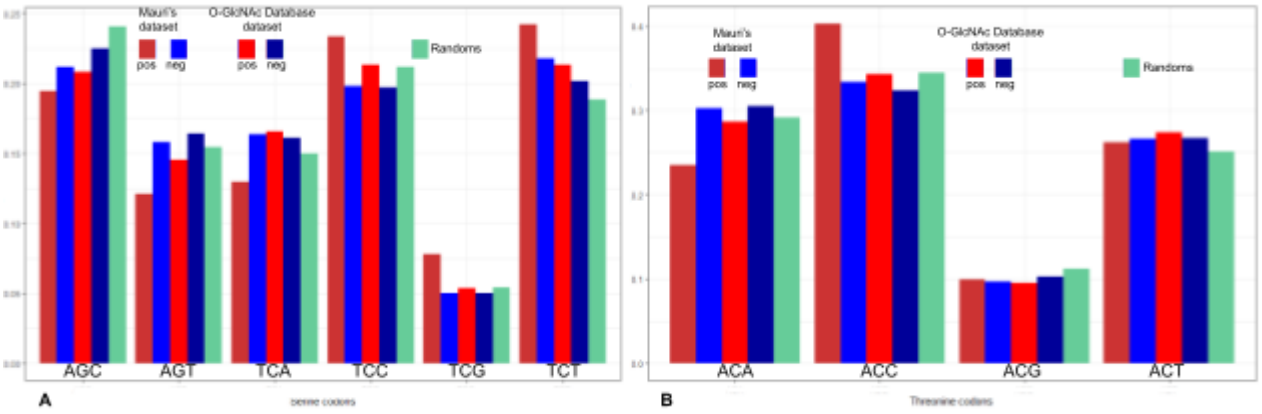


Figure 16. **Distribution of codons of serine and threonine for *O*-GlcNAcylated and non *O*-GlcNAcylated sites regarding two datasets and random mammal protein sequences**

A: Serine codons; B: Threonine codons. (red columns are for positive data, blue ones for negative sites and green for random sequences)

Flexibility of the different codons:

As no significant difference had been found for *O*-GlcNAcylated sites, we decided to analyze the flexibility of the site depending on these codons to see if the results differ from the ones we obtain in our article. As shown in Figure 17, the flexibility of *O*-GlcNacylated sites is above the non *O*-GlcNacylated sites which is also higher than random serine or threonine, in particular serine which are quite flexible according to this software. Threonines are more context dependent in terms of flexibility for positive sites than others where they are more rigid so this still shows a higher flexibility for *O*-GlcNAcylated sites. But we can see that the proportion is quite similar whether it is for positive sites or negative sites so it is still hard to discriminate *O*-GlcNAcylation sites from non *O*-GlcNAcylation sites. We can also see that a random serine or threonine is more rigid than a random serine or threonine from a protein which is modified somewhere by the OGT. This induces that the flexibility favors the

protein to go into the TPR domain or at least the catalytic domain and be modified. Regarding the proportion of the flexibility categories, there is no difference between all the various codons whether it is for threonine or for serine. According to Figure 17, serines are generally found more in flexible regions than threonines.

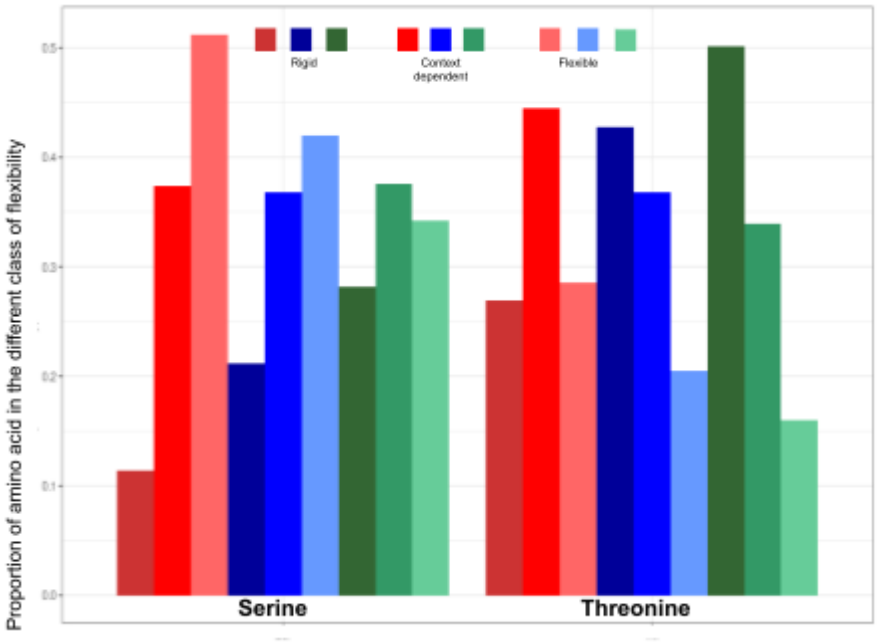


Figure 17. **Proportion of residue for each category of flexibility according to Dynamine**

The flexibility category of a residue is calculated using Dynamine. The color corresponds to the dataset as in Figure 16.

Secondary structure prediction:

Secondary structure was predicted with the SPIDER3-Single software [165]. The results (see Figure 18) show that threonine are more often found in a structured area than serines which coincide with the previous flexibility prediction showing that threonine flexibilities are more often context dependent. But in any case the two residues are more often found in random coils according to the prediction software. But compared to the negative set there is no significant difference; the only one is regarding threonines which are a more on beta sheet for the *O*-GlcNAcylated ones. But once again when regarding the secondary structure separately according to the specific codons no difference is noticeable.

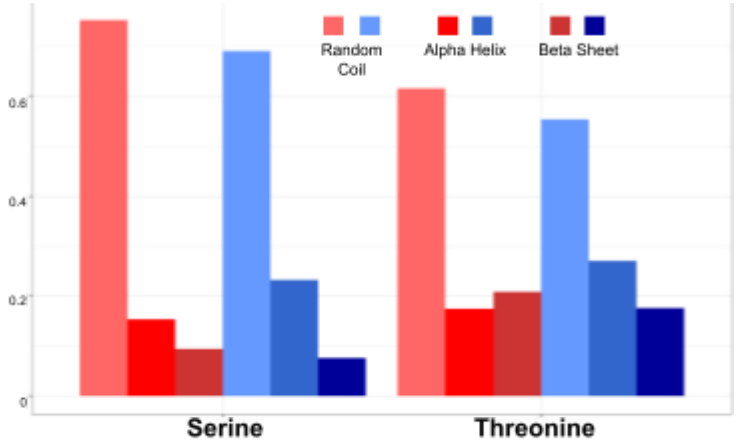


Figure 18. **Prediction of secondary structure between *O*-GlcNAcylated sites and non *O*-GlcNAcylated sites with SPIDER3-Single**

2. Accessibility to the solvent

a. Material and methods:

Protein modeling

To obtain models of *O*-GlcNAcylation proteins, we chose to use AlphaFold (V2.2.0) as it shows very good solutions for protein modeling in CASP13 and CASP14 [8,9]. This version of AlphaFold is an update of the published software. It uses deep learning based on an available database of protein structures and more specifically the interaction between residues inside a structure. It then uses multiple sequence alignments (MSA) to predict the structure models. These MSAs are created thanks to alignment from evolutionarily related proteins thanks to JackHMMR or HHBlits [166,167]. This information is coupled to 3D atom coordinates of a small number of homologous structures when available [10]. The model building starts from one of the five random seeds. Each random seed has been trained differently as with or without a template. For each random seed, it is possible to select the number of models we want it to construct. Here we modeled 79 *O*-GlcNAcylated proteins with 1 model per seed meaning five models were predicted for each protein. Each model gets a predicted local distance difference test (plddt) score which is essentially a confidence score between 0 and 100. Here we only selected the best model out of the five based on this score. AlphaFold was run on a gpu machine from our lab with the following components:

- GPU: Nvidia RTX A5000 with 24 GB of RAM

- RAM: 128 GB
- CPU: Intel(R) Core(TM) i9-10900 CPU @ 2.80GHz (20 CPUs)
- Storage: 18TB + 4TB (SSD which contains the database)

Accessibility calculation:

The goal here was to calculate the accessibility of the *O*-GlcNAcylated sites based on the models predicted by AlphaFold (v2.2.0). We used NAccess with default parameters to calculate the solvent accessibility [168]. This software calculates the accessibility of the protein surface to solvent, with output at atom, residue and protein level. In total, the accessibility of 149 *O*-GlcNAcylated sites was analyzed. The accessibility is provided as a surface area ($Å^2$).

b. Results

As *O*-GlcNAcylation is a dynamic PTM, the accessibility of a substrate site must be sufficient for a quick modification and it is with this idea in mind that we first hypothesized that the site must therefore be (solvent)-accessible. In our previous article about the *O*-GlcNAcylation prediction, we tried to validate this hypothesis with I-tasser which was one of the most powerful *de novo* protein modeling tools available as a server [169]. But the quality of the models was very low according to their confidence score so we decided to try it again with AlphaFold (v2.2.0)

First, the qualities of the 79 models can be seen in Figure 19 with the distribution of plddt score. The density curve shows a bimodal distribution: one between 50 and 70 and another between 70 and 80. The accessibility of the different *O*-GlcNAcylated sites have been studied in function of the quality score of the model. The results are visualized in Figure 22. In this Figure 22 we can see that the accessibility varies a lot with some residues being fully accessible and some completely buried. But regarding the codons specifically we can see that the TCA codon has a higher accessibility (above 20 $Å^2$) than the others. For the models with a plddt score above 75, the TCC codons also have a high accessibility (>35$Å^2$) while for AGC the mean accessibility score decreases. As some models have a lower quality score, its depression may be due to unstructured regions. To be sure the quality does not affect the accessibility values, we analyzed the correlation between the accessibility of the residue and the plddt score of the corresponding model. We obtained a correlation score of

-0.1083939 which means there is no correlation between these two variables (see Figure 20). Also, the analysis of the density regardless of the model score (Figure 21), shows us that according to our models, *O*-GlcNAcylated sites can be buried inside the protein but the majority of them are solvent accessible.

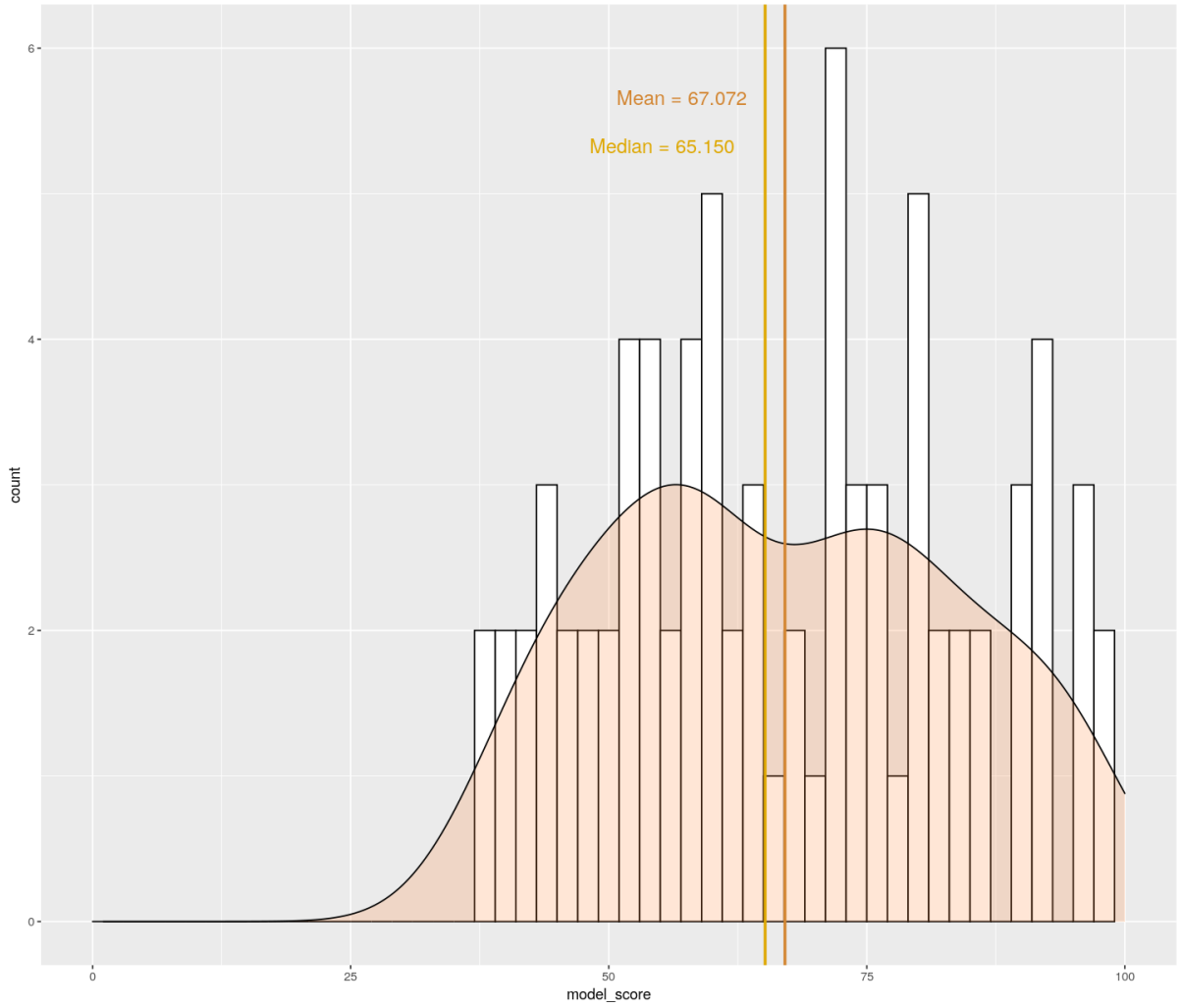


Figure 19. **Barplot representing the quality of the different models**

The orange curve is a density curve and the median and mean are in yellow and dark orange respectively.

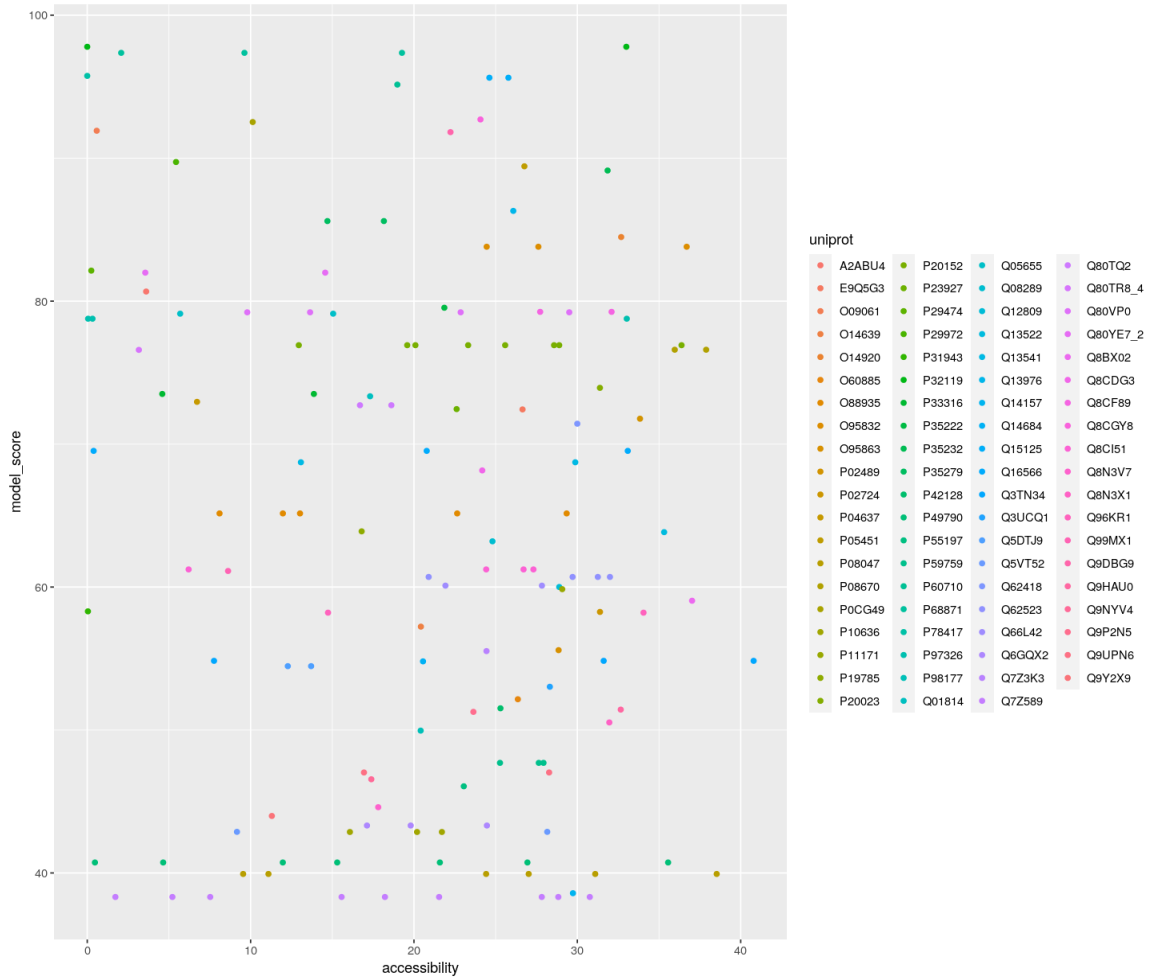


Figure 20. **Dotplot between the quality score of the models and the accessibility of their *O*-GlcNAcylated sites**

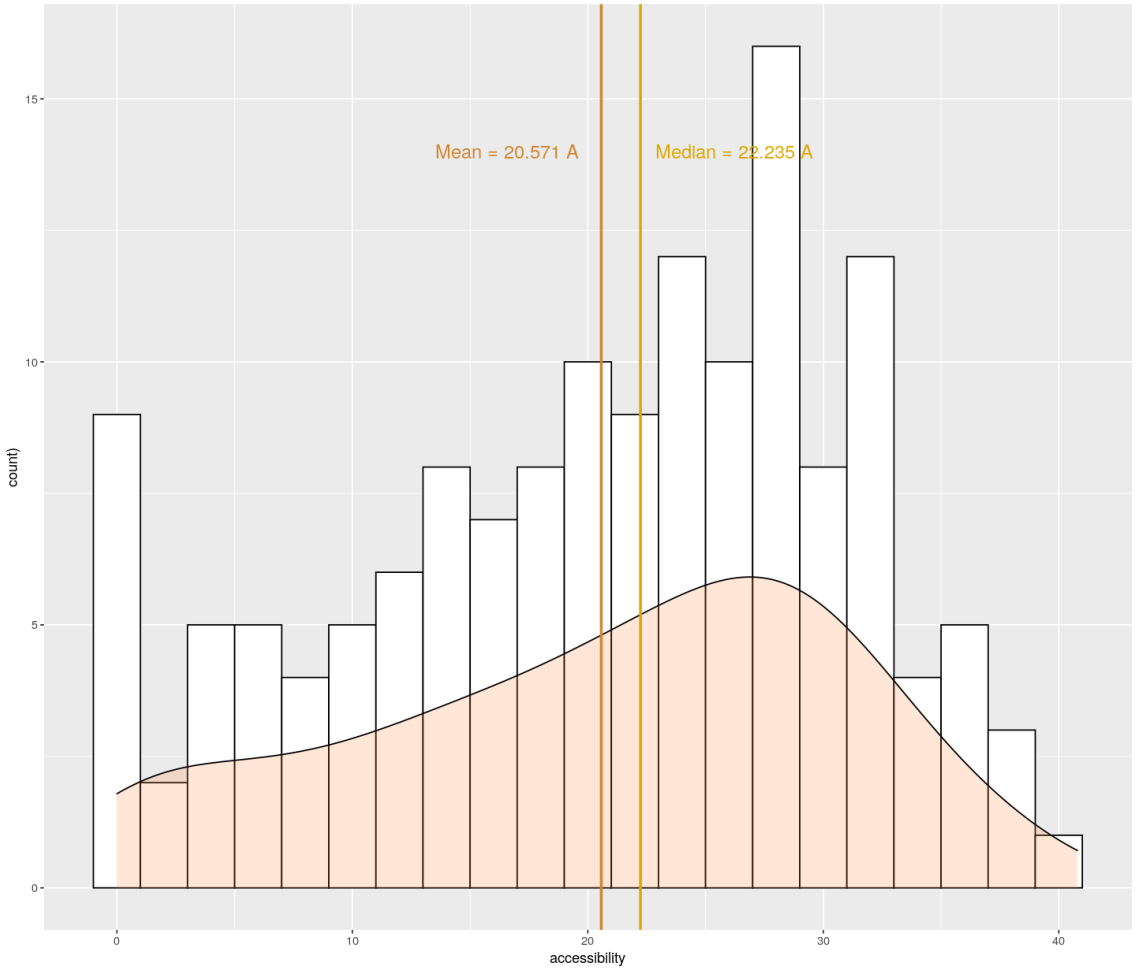


Figure 21. **Barplot representing the density of the accessibility for all the modeled *O*-GlcNAcylated sites**

The orange curve is a density curve and the median and mean are in yellow and dark orange respectively.

The accessibility of *O*-GlcNAcylated sites is very heterogeneous, as some sites are really accessible, around 10% of the sites are not accessible according to the prediction models. This absence of accessibility could be explained by the possibility of co-translational activity of the OGT for protein folding. Another hypothesis would be that a positive site can be *O*-GlcNAcylated while being in a peptide but not when the protein is fully constructed. The majority of the sites analyzed are at least accessible and it could be interesting to try adding a GlcNAc on the site in the structure and see if a link can really happen.

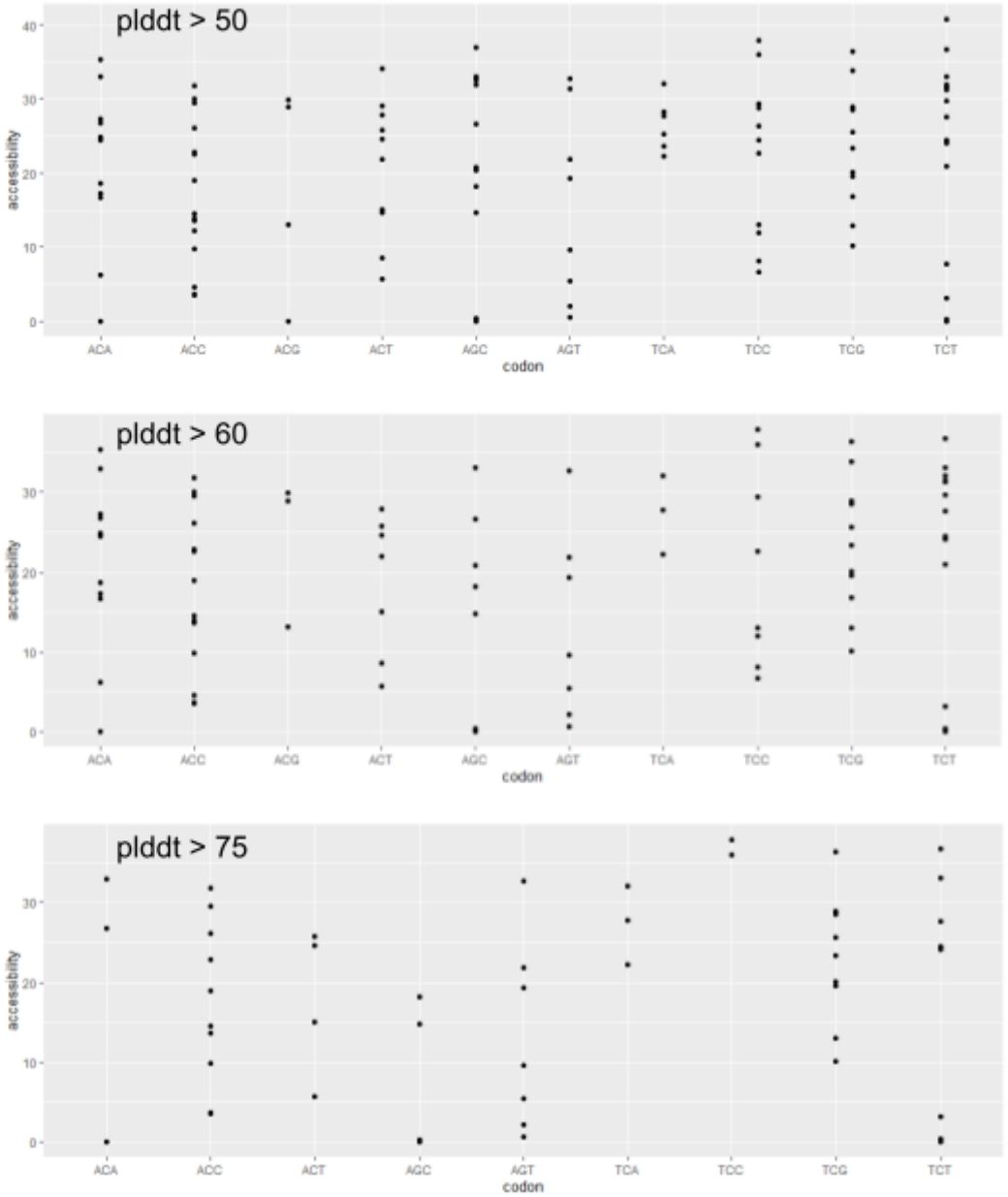


Figure 22. **Scatter plots of accessibility (Å²) depending on codon computed on models with different score thresholds**

The plddt score corresponds to the confidence score according to AlphaFold algorithm. ACA, ACC, ACG and ACT are threonine codons; AGT, AGC, TCA, TCC, TCG and TCT are serine codons.

3. Do *O*-GlcNAcylated peptides interact with the asparagine ladder of OGT TPR?

As seen in Levine *et al.* (2018), the asparagine ladder of the recognition domain of the OGT called TPR is important for the *O*-GlcNAcylation activity [37]. This ladder may be involved in substrate recognition. In the results from the third part of this PhD manuscript, describing the special interaction between beta-catenin and OGT, we showed that the unstructured N-terminal segment of the β-catenin (known to be *O*-GlcNAcylated) interacts with the lumen of the TPR and more precisely between the two threonines (T40 and T41) and asparagines from this ladder. In addition, in our team we have studied the Tau protein, which is also well known to be *O*-GlcNAcylated when cut into peptides. Otherwise the number of *O*-GlcNAcylation revealed is lower [170]. Some of the Tau peptides also have unstructured conformations and are similar to the N-terminal segment of the β-catenin. To support our hypothesis, we decided to analyze the interaction of these peptides with the TPR domain of the OGT. The modeling of Tau peptides with OGT shows similar interactions between the *O*-GlcNAcylated sites and the asparagine ladder. And this, each time the peptide was unstructured and had a length of 51 amino acids (±25 around the *O*-GlcNAcylated site). From these two observations, we wanted to see if we can use interaction modeling between unstructured peptides and OGT to predict *O*-GlcNAcylation.

a) Material and methods

Dataset:

To be able to model peptides containing *O*-GlcNAcylation sites, we selected proteins retrieved in our previous publication with a total amount of 536 experimentally proven modified sites (the "Mauri" set) [160]. According to Figure 23, the peptide should be 51 residues long with ±25 amino acids around the *O*-GlcNAcylated site, the protein sequences were processed to remove sites which were too close to the beginning or end of the protein sequence. In parallel, to be able to compare results with negative data, serines and threonines not proven to be *O*-GlcNAcylated were also retrieved with ±25 residues at either side. But if we found a negative site to be close to an *O*-GlcNAcylated site (in the ±25 window) it was removed from the data to avoid unexpected interactions between the *O*-GlcNAcylated site. Once positive and negative peptide sequences had been extracted,

they were analyzed through SPIDER3 as previously described and the percentage of coil structure was calculated for each. As we wanted to analyze only unstructured peptides, only sequences with a coil rate equal to 1 (i.e. fully coil) were selected. As a result, from the first dataset, the number of positive and negative sites decreased to 175 and 8,775 peptide sequences, respectively. The peptides in the positive data set have been classified in function of the number of positive sites inside the sequence stretch, the site itself (at the index of 26) included.

Complexes modeling:

To model the different complexes, AlphaFold (v2.2.0) was chosen. For modeling complexes DeepMind provides AlphaFold-Multimer which has not been published yet but already many results have been made available on BioRXiv as shown in the Introduction section. Briefly, AlphaFold-Multimer has been trained on experimentally resolved structures of complexes and showed good results because of the effect of co-evolution of interacting residues in a protein complex. As for the protein modeling, AlphaFold-Multimer predicts a model from five different random seeds. The number of models generated per seed was set to three, meaning the total number of models produced for an interaction was fifteen. Again we selected only the first ranked model according to the plddt score for the whole complex. To reduce the time spent per calculation we tried different sizes of the TPR domain for the OGT by reducing the number of TPRs but we also tried different lengths for the peptide. As shown in Figure 23, we can see that the optimal number of TPR repeats is 8, as it shows the best quality score for models.

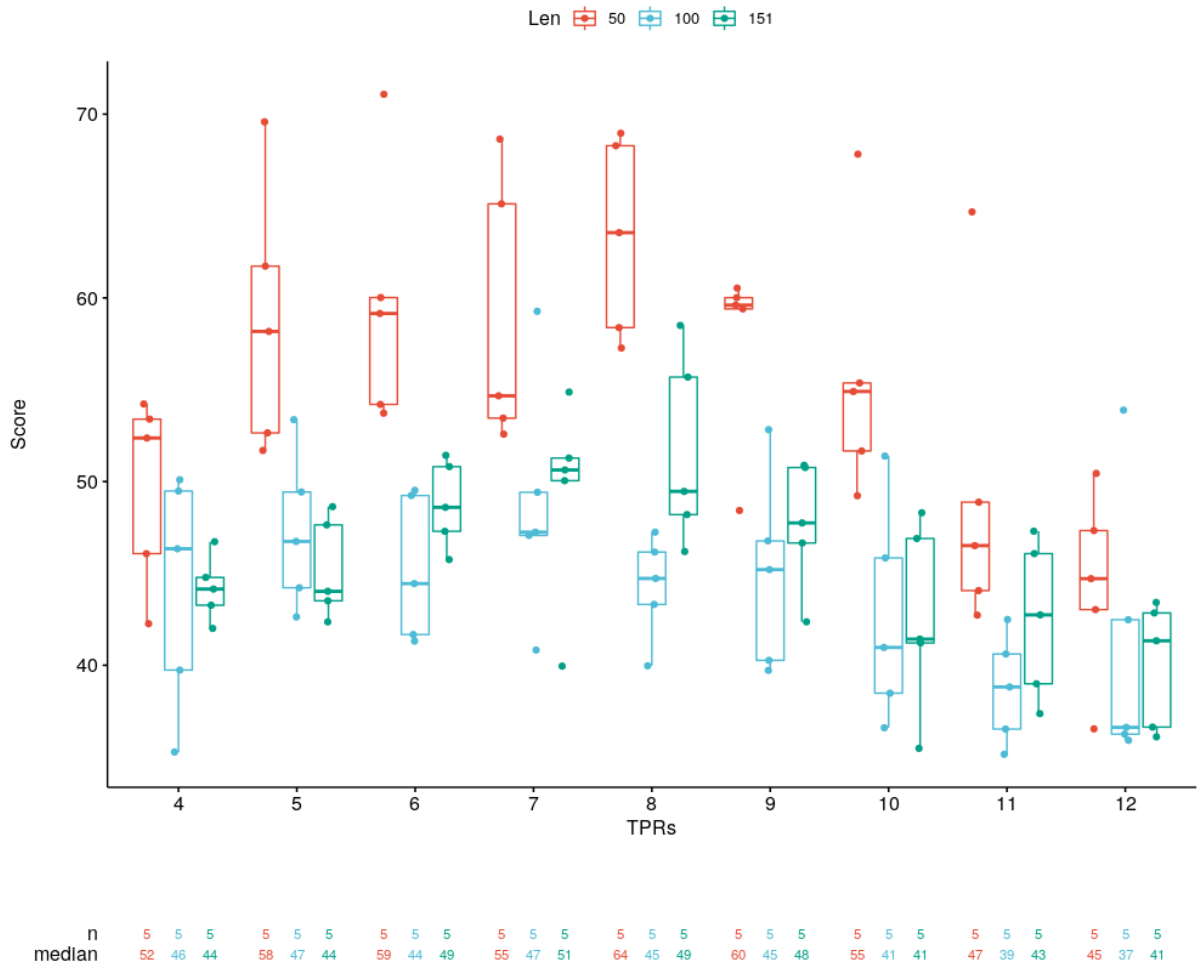


Figure 23. **Distribution of plddt score of complexes as function of the size of peptides and the number of TPR repeats**

The plddt score is the confidence score of a model given by AlphaFold-Multimer. For each configuration five models have been produced. The median corresponds to the median score for a configuration. Colors correspond to the size of the peptide: red = 50 residues long, blue=100 and green = 151 residues.

Extraction of interactions between asparagines from the ladder and *O*-GlcNAcylated sites:

To investigate the interaction between residues from the asparagine ladder and our sites of interest, the PyMOL software was used with a Python script to automate the process with the cmd function from the PyMOL module[171]. The script extracted, for each asparagine from the ladder, the amino acid environment within a sphere of 5Å. Every residue was then stored in a list where we are looking for the *O*-GlcNAcylated or non *O*-GlcNAcylated site. If the site was found in one of the asparagine lists, its score was increased by one. As there are 6 different asparagines (N321, N322, N325, N356, N390, N424), a peptide can have a score between 0 and 6.

Extraction of the interaction between positive or negative site inside the OGT/peptide complex:

Another interesting material would be to know which OGT residues are interacting with the sites of interest. To that, a Python script automated the process of retrieving the types of residue that the modified serines and threonines are or are not in contact with, considering a distance between 2.5 and 5 Å. The type of residue is stored in a list where redundancy was removed. Each element of the list is then added to a dictionary to get the number of times a type of amino acid is found in contact for positive or negative data.

b) Results

Analyses of the different complexes:

The confidence score of every model according to the fact that if a serine or a threonine is in contact with at least one of the asparagine from the TPR asparagine ladder was plotted in Figure 24. We can see that models with a peptide interacting with the asparagine ladder globally have a better model score than others. As we now want to see if the quality of the site of interest improves when it interacts with asparagines, we plot the plddt score of the 26 residues of the peptides in function of whether it is interacting or not in Figure 25. We can see that the scores are higher when interacting. These results may tend to think that the interaction with asparagine ladder improves the quality of the model but here we do not know if the site if it is *O*-GlcNAcylated improves the confidence in the model. The model and the site of interest scores regarding if the serine or threonine is *O*-GlcNAcylated or non have been plotted in Figure 26. Unfortunately, we can see in this figure that models have a slightly lower score for *O*-GlcNAcylated sites than negative data and this is also the case considering the residue-specific plddt score.

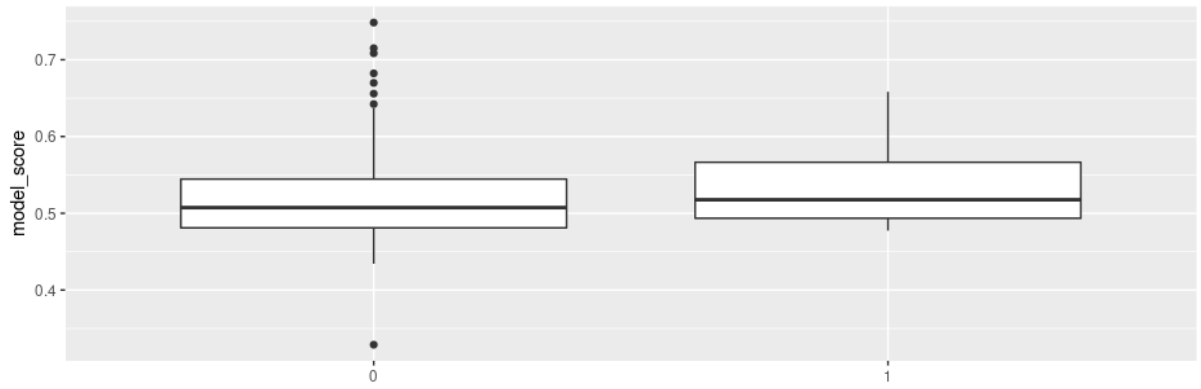


Figure 24. **Box plot representing the different model plddt scores depending if the serine or threonine are interacting at least once with an asparagine from the asparagine ladder**

0 corresponds to no interaction with asparagine from the ladder; 1 corresponds to model with at least one interaction between the serine or threonine and asparagine from the ladder

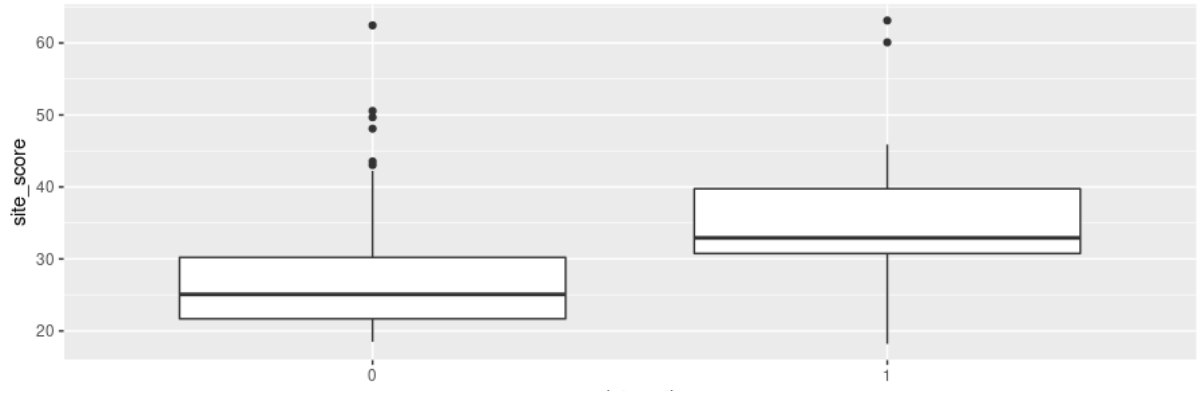


Figure 25. **Box plot representing the different residue specific plddt scores depending if they are interacting at least once with an asparagine from the asparagine ladder**

0 corresponds to no interaction with asparagine from the ladder; 1 corresponds to model with at least one interaction between the serine or threonine and asparagine from the ladder

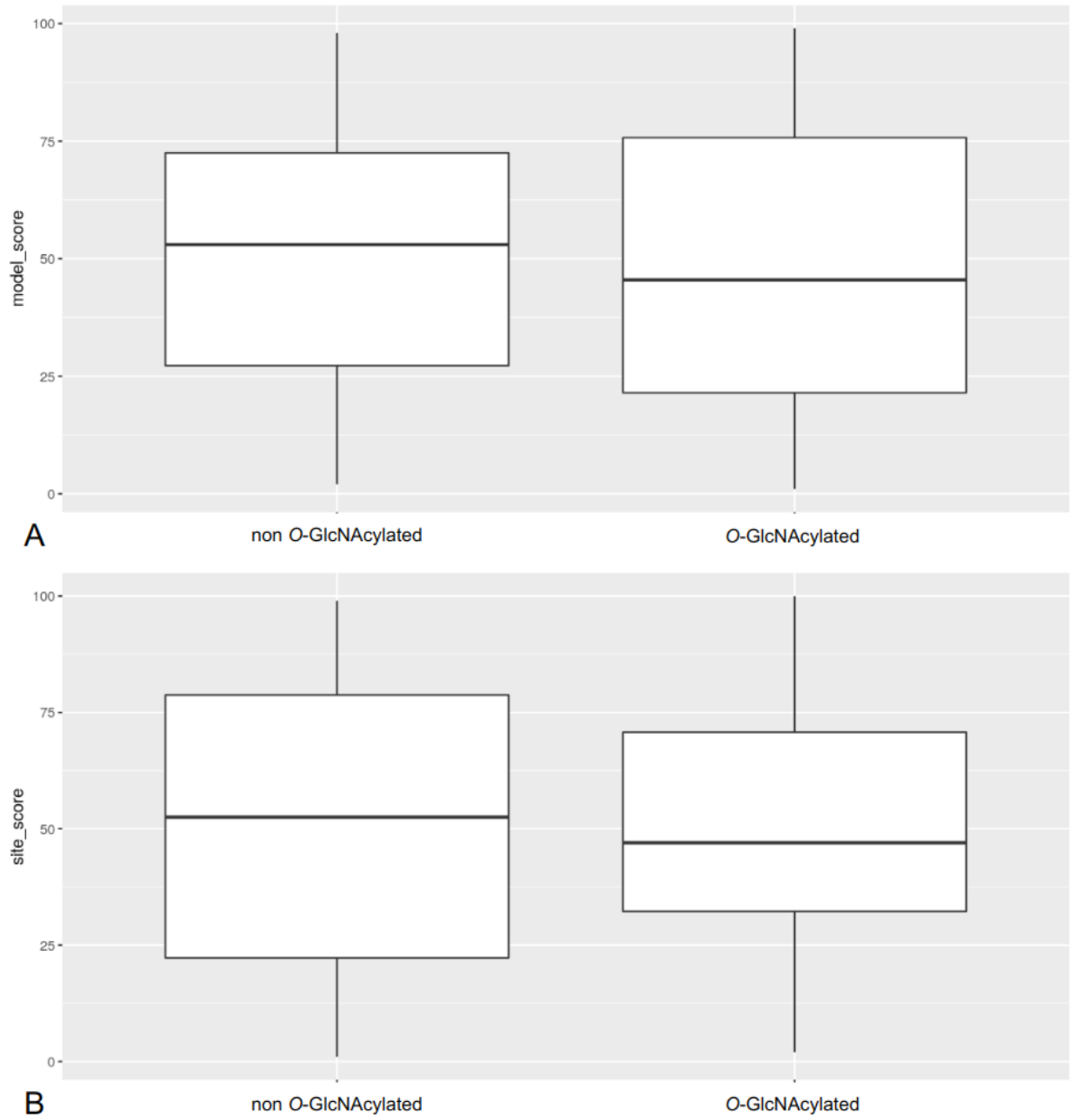


Figure 26. **Boxplot of model and site scores according to *O*-GlcNAcylation of the non *O*-GlcNAcylation of the peptide**

A: model plddt scores whether it is *O*-GlcNAcylated or not. **B:** Site of interest plddt score whether it is *O*-GlcNAcylated or not.

Study of the the interaction between *O*-GlcNAcylated sites and the asparagine ladder:

In this study, we wanted to see if AlphaFold-Multimer would preferentially create an interaction between *O*-GlcNAcylated sites and asparagine of the asparagine ladders. In that case, it could therefore be used as a prediction tool for *O*-GlcNAcylation prediction of unstructured parts of proteins. But as the time of calculation is still high the number of

interactions we produced with AF-Multimer is 100 (58 negatives and 42 positives). For each of these models we looked if at least one interaction had been created between the serine or threonine and the asparagine ladder of the TPR. Regarding the positive data we counted only 6 models with at least one interaction, i.e. 14.29%. We also counted 6 models for the negative set, i.e. 10.35%. Even if the results are slightly better for the *O*-GlcNAcylated peptides the low percentage does not allow us to think that it can be used to predict *O*-GlcNAcylation sites on disordered parts of proteins.

Analysis of the kind of residues that interact with positive and negative data:

As the specific interaction between the asparagine ladder and an *O*-GlcNAcylated site does not seem to be relevant for *O*-GlcNAcylation prediction, we hypothesized that the hydroxyl group of serine or threonine might engage in a specific interaction with certain types of amino acids of the OGT. For this purpose, the environment of this hydroxyl at 5 Å was analyzed thanks to PyMOL. These results have been retrieved in Table 9. We can see that, regardless of whether the site is *O*-GlcNAcylated or not, the hydroxyl group of the site is predicted by AlphaFold-Multimer to interact more often with asparagines, aspartic acids and lysines.

Amino acid	Percentage of the total amount of interaction with positive sites	Percentage of the total amount of interaction with negative sites
Asparagine	34.45%	29.77%
Aspartic acid	15.56%	11.31%
Lysine	10.00%	10.78%
Glutamic acid	7.78%	2.98%
Phenylalanine	6.67%	5.96%
Serine	4.45%	3.58%
Cysteine	4.45%	3.58%
Arginine	3.34%	0.60%
Valine	3.34%	6.55%

Glycine	2.23%	1.79%
Alanine	2.23%	1.79%
Histidine	2.23%	5.90%
Tyrosine	2.23%	4.77%
Isoleucine	1.12%	2.39%
Leucine	0%	2.98%
Proline	0%	1.79%
Glutamine	0%	1.79%
Threonine	0%	1.79%
Methionine	0%	0%

Table 9. **Percentage of specific interaction between *O*-GlcNAcylated and non *O*-GlcNAcylated sites with the different kinds of amino acids**

In Figure 27, we represented all the best peptide models for positive and negative data in the OGT. We can see that all the peptides have the same conformation inside the TPR. We also drew the central serines and threonines. Unfortunately the sites are not central at all whether it is known to be *O*-GlcNAcylated or not.

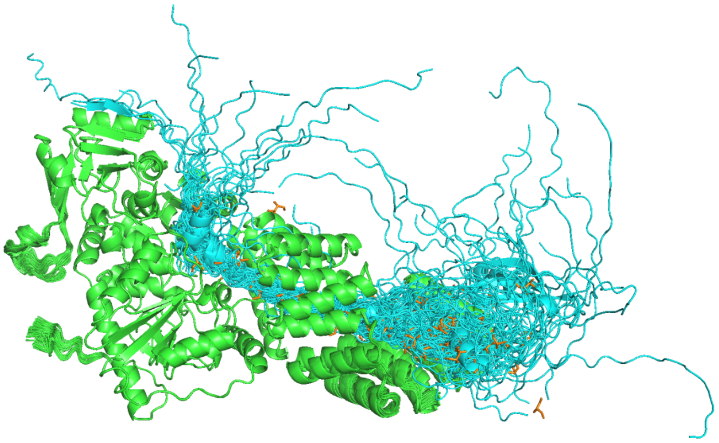


Figure 27. **Representation of all the peptides in complex with OGT with 8 TPRs modeled by AlphaFold-Multimer (v2.2.0)**

In green is represented the OGT and in blue all the different peptides. The orange residues are the serine and threonine of interest.

Analyses of the chaperone proteins found in the interactome :

As AlphaFold was available in our laboratory, the modeling of interaction between the OGT and the four different proteins has been done. For each possible complex a total of 25 models have been predicted and the different scores have been extracted and summarized in Table. As these scores are very low the investigation was stopped here but with bigger capacities of calculation it could be interesting to model the full CCT and then create the complex CCT-OGT. An another possibility is to cut the proteins into interface areas an analyze step by step the full complex

Complex	min score	max score	average score	median score
CCT2 - OGT	0.2381	0.2682	0.2478	0.2459
CCT3 - OGT	0.2388	0.2685	0.2536	0.2530
CCT5 - OGT	0.2367	0.3056	0.2545	0.2506
HSPD1 - OGT	0.2405	0.3833	0.3040	0.3058

Table 10. **Summary of the different complex scores between OGT and chaperone proteins**

C. New discussion and conclusion

At the end of our *O*-GlcNAcylation article, we accused the lack of positive data against the high number of negative data to hamper the ability to predict *O*-GlcNAcylated sites. In addition, new hypotheses have been proposed leading to new studies such as the role of the codons in the selection of serine and threonine by the *O*-GlcNAclTransferase (OGT), the accessibility to the solvent and the possibility to predict *O*-GlcNAcylation thanks to the asparagine ladder of the TPR domain [156,159].

The emergence of the *O*-GlcNAc database helps us to get more data in particular for the analysis of the potential role of the amino acid codons. But this analysis has not helped to discriminate between positive and negative sites. Even if the codons could play a role in the torsion of the protein backbone at the outside of the ribosome, the global structure of

the protein should be preponderant for the substrate recognition and the catalytic activity of the OGT.

The accessibility of O-GlcNAcylated sites has been calculated again thanks to the breakthrough created by AlphaFold which shows a very good capability to predict *de novo* structures which was not the case when we first studied it. Even if the majority of sites are at least accessible, more than 10% are buried and therefore not accessible at all. These sites may have been modified co-translationally as *O*-GlcNAcylation is not always a post-translational modification. It could be interesting to go back to the source and look at the determination method. Indeed, some serines or threonines can be *O*-GlcNAcylated when cut into peptides meaning that they are accessible in that case but not if the full protein is folded [170]. Also if only the OGT and the two substrates are present and the *O*-GlcNAcylation occurs, it would mean that the site is accessible. Otherwise, the presence of other unknown participants could explain the *O*-GlcNAcylation. But the high amount of experiments whether by computational methods or experimentales ones make it extremely difficult to obtain answers. Furthermore the accessibility according to the different codons does not bring any information except for the TCA codon which shows only good accessibility. These results have to be handled with care as the low amount of this codon can add a bias to the results. Additional data with this codon would show a low surface accessibility and give similar results to the other codons.

The asparagine ladder of the TPR domain has been described to play an essential role in the substrate recognition [37]. Thus, the hypothesis that this ladder interacts with any *O*-GlcNAcylated site is relevant but the poor results we obtained here did not support that hypothesis. However, as the models confidence of the positive and negative complexes are low. The poor results can be explained by these low qualities. Although we looked at the specific interaction of the asparagine with serines and threonines, it could be interesting to go wider and look at interaction around the sites. The models predict a position of the peptide inside the TPR domain at a given time which may change with dynamics. Indeed, with time it could have been interesting to perform molecular dynamics on the peptide to see if this ladder interacts with positive sites to pull the peptide inside the catalytic domain. Also, to win time, we decided to reduce the number of TPRs. But removing TPR repeats may

lead to a loss of information. That is why we tried to get the highest model qualities with the lowest TPR repeats but the results may have a bias. As AlphaFold-Multimer is trained on available structures from the Protein Data Bank (PDB), structures of truncated OGT with a peptide are available. These structures may have led to redundant results as we can see that all the peptides are predicted with the same conformation (Figure 27). The predicted position of the different peptides could be explained by the presence of OGT structures with a peptide in this location. However, the difference of sequence composition may explain the different scores at this position.

In our study, we highlighted four proteins (CCT2, CCT3, CCT5 and HSPD1) which are all chaperone proteins. The idea that they play a role in *O*-GlcNAcylation is a major hypothesis and its confirmation would be a big step forward in the comprehension of the *O*-GlcNAcylation process. In conclusion, *O*-GlcNAcylation site prediction, even with more data and more powerful software, is still an unattained objective. The heterogeneity of the data coupled with the large number of *O*-GlcNAcylated proteins for only one enzyme support the need of chaperon proteins to bring the substrate to the *O*-GlcNAc Transferase. Theory supported by the cleavage activity of the OGT. Unfortunately, at the time of writing, the ones highlighted by our article are too big to be modeled in complex with the enzyme. To counteract the size limitation, only keeping potential interaction surfaces can be a solution. This would also increase plddt scores due to lower uncertainty of prediction in the remaining regions. But generating big complex structures could be possible in the very near future.

IV. Protein-protein interaction related to CoVid-19: how to define new methods to assess prediction

A. Introduction

1. CoViD-19

The viral family of Corona Viruses (CoVs) is well studied and has caused three different international outbreaks in the last twenty years. These diseases, called severe acute respiratory syndromes (SARS), Middle Eastern Respiratory Syndromes (MERS) and the most recent COronaVIrus Disease 2019 (COVID-19) showed a capability to spread quickly over the world [172]. COVID-19 is caused by SARS-CoV-2, coming from a new *Betacoronavirus* linked to SARS-CoV. COVID-19 was declared a pandemic by the World Health Organization (WHO) in March 2020 [173,174]. Following the quick and world-wide spread of the virus, and the high number of deaths, a worldwide effort was undertaken to understand the different mechanisms of infection and to find ways to counteract this pandemic. Early in 2020, a publication bringing together 100 authors had been deposited on bioRxiv entitled “A SARS-CoV-2 protein interaction map reveals targets for drug repurposing”, later published in Nature in late July. This publication highlighted the viral proteins that could be physically associated to human proteins by affinity-purification mass spectrometry. This resulted in the identification of 332 high-confidence protein-protein interactions (Figure 28).

The aim of this research was to find druggable proteins targeted by 69 different compounds. But from this study, specific protein-protein interactions between SARS-CoV-2 and human proteins can be studied or predicted in order to have a better understanding of infection and replication methods. To this, CAPRI (Critical Assessment of PRotein Interactions) based on this PPI network, selected interesting and possibly resolvable interactions. This will be further described later in this manuscript.

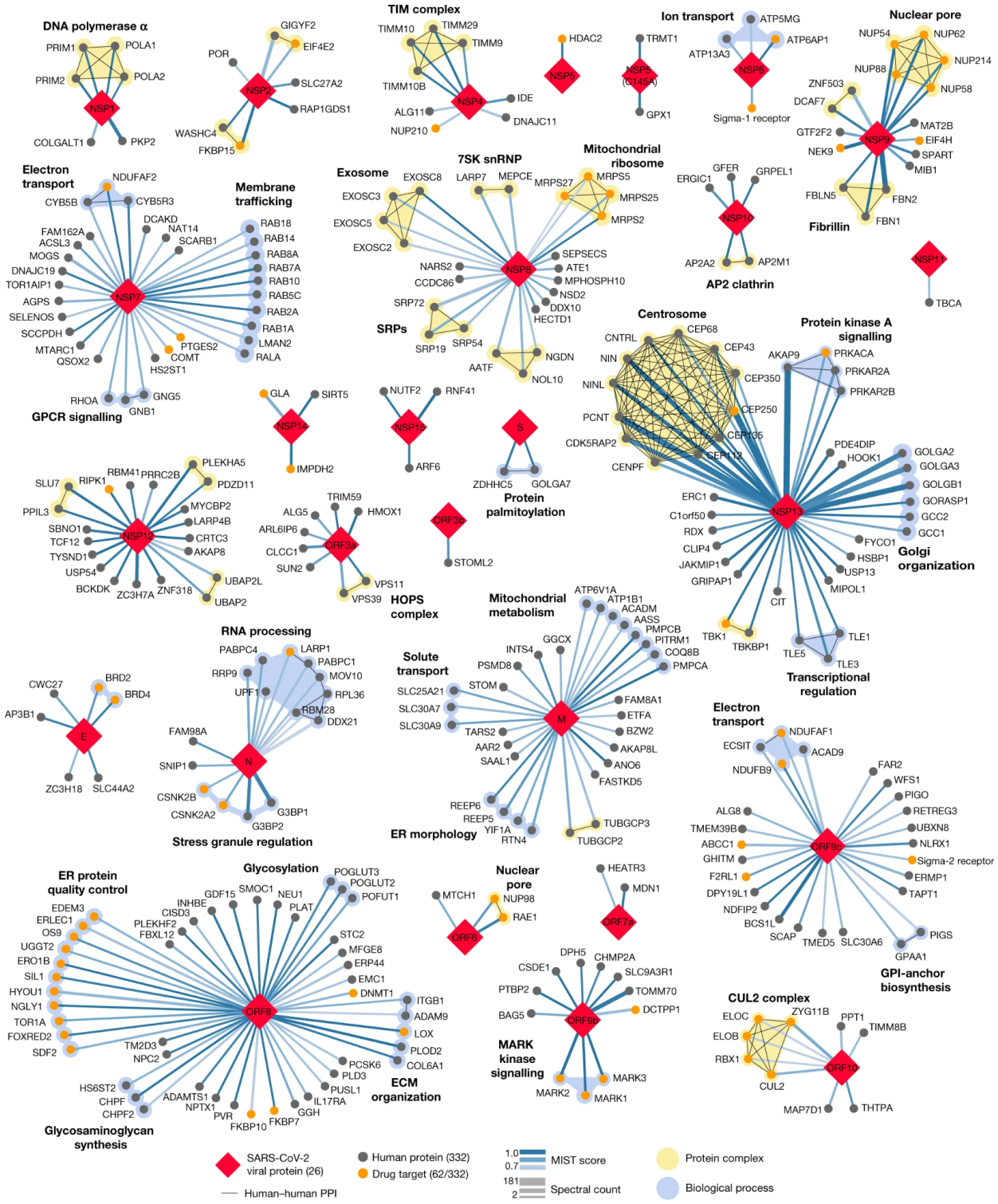


Figure 28**. SARS-CoV-2 protein–protein interaction network**

332 high-confidence interactions between 26 SARS-CoV-2 proteins (red diamonds) and human proteins (circles; drug targets: orange; protein complexes: yellow; proteins in the same biological process: blue). Edge colour proportional to MiST score; edge thickness proportional to spectral counts. Physical interactions among host proteins (thin black lines) were curated from CORUM, IntAct, and Reactome. An interactive protein–protein interaction map can be found at kroganlab.ucsf.edu/network-maps. ECM, extracellular matrix; ER, endoplasmic reticulum; snRNP, small nuclear ribonucleoprotein. n = 3 biologically independent samples (from Gordon *et al.*, 2020) **.**

2. CAPRI Community

a) Summarized modeling methods

The aim of PPI modeling is to predict a detailed atomic-level interaction between two or more molecules [149]. This not trivial goal has been used since the early 80's when the human genome was far away from being fully sequenced [175]. The sequencing of our genome highlighted the contrast between the low number of genes and the number of proteins and isoforms involved in different metabolisms inside a living organism. This implies the essential roles of the PPI and the need to understand those. Several methods have been developed according to different theories, organisms or molecules. Indeed, some algorithms will be species specific or accurate for the docking of small molecules, peptides or proteins. As there is a lot of different software, it is complicated to describe all of them so the following paragraphs will describe the most commonly used methods for the different types of molecules.

Docking of proteins and small ligands is very effective in terms of medical research. Small molecules are often used as drugs and this method is used for drug discovery or design. Indeed, the process of drug discovery is very long (between 10 and 15 years) and costs around 2.5 billion USD[60,176,177]. The use of computer-assisted drug discovery (CADD) techniques in early-stage studies by leading pharmaceutical companies and research groups has accelerated the drug discovery and development process by minimizing costs and late-stage failure [60,178].

To perform Structure-Based Drug Design (SBDD), the availability of the therapeutic target structure and its catalytic pocket are the two bases [179]. Structure-Based Virtual Screening (SBVS) is one the most common methods for SBDD with molecular docking and Molecular dynamics (MD) simulations. SSBD can be sub processed in 7 steps starting with preparation of the target protein followed by the identification of the ligand binding site. Then the preparation of the compound library, molecular docking and scoring, molecular dynamic simulation, and binding free energy calculation [60]. The target protein structure can be obtained by browsing structure databases as PDB or by experimental data or even with homology modeling based on sequence similarity. Defining a ligand binding site can be also performed according to different experimental methods such as studies by directed

mutagenesis or X-ray crystallography of proteins co-crystallised with substrates or inhibitors but sometimes it is too complicated. For this purpose many software of web servers have been made available like CASTp[180], DoGSite Scorer[181], NSiteMatch[182], DEPTH[183], MSPocket[184], MetaPocket[185], and Q-SiteFinder[186].

CADD is improving each year and its efficient use is a major contributor to human health. Recently, the potential of eighteen repurposed drugs in clinical development against SARS-CoV–2 M^{pro} have been explored thanks to combined molecular docking and molecular dynamics techniques. This led to the identification of TMC-310911 and ritonavir as promising drugs for the treatment of COVID-19[187].

Protein-protein modeling is another main subject in the analyses of interaction inside different pathways of metabolism involving the presence of different pattern proteins. This modeling can be divided into 3 categories: protein-protein docking, template based modeling and hybrid methods[188]. Each approach has its own advantages and disadvantages.

Modeling peptide-protein interactions is also a main research domain which has grown a lot recently. Indeed, some interactions are mediated through a peptide [189]. These peptides are really important in many cellular processes, mediating 40% of protein-protein interactions; they therefore constitute attractive drug candidates[126].

The main next step once the binding site is found is the peptide docking. Like protein-small molecule and protein-protein interaction modeling, this step can be realized by different methods , algorithms and tools.

One of the most common methods is template-based prediction. As its name suggests, this method will use already resolved protein-peptide interaction structures. As the number of such structures increases in the PDB, this method will be increasingly accurate as the probability to find a similar complex is higher[190]. This method is used by the GalaxyPepDock server which is part of the GalaxyWEB server and allows to perform template-based protein-peptide modeling online[136,191]. This tool uses, after finding a template, a Galaxy energy to do some minimization. This energy combines physicochemical

energy terms derived from the force field of molecular mechanics, knowledge-based energy terms derived from the statistics of the interactions between pairs of atoms in the structure database, and restraint energy terms derived from the information on interactions found in homologous complexes [190].

On the other hand, for high-throughput modeling, the need for software which can model peptide-protein interaction without prior information is a very important aspect [192] .

b) CAPRI Rounds

CAPRI (Critical Assessment for PRotein Interactions) is an experiment created on the CASP (Critical Assessment of protein Structure Prediction) model. The aim of this experiment is to be a catalyst of protein-protein docking. The chosen way is to perform blind docking. Blind docking principle is easy and consists of proposing protein complex sequences with the stoichiometry associated. The complexes are resolved experimentally by wet laboratories and then provided to CAPRI before being published. This means that the predictors can not find a solution in existing databases. They have to perform blind docking and provide the best models according to their own knowledge. Each complex is called a “Target” and many targets are regrouped into one “Round”. Targets were first protein-protein interactions but as the research field progressed new kinds of targets appeared such as protein-peptide, protein-RNA, protein-DNA, protein with small ligands. As predictors have to register to participate in a Round, the new kind of target inside a round creates new problems to resolve. That is why CAPRI is a catalyst in protein interaction modeling. But not only predictors are participating in CAPRI Rounds. Indeed, CAPRI is also a catalyst experiment for model scoring. Being able to predict models is for sure very important for biologists but the capacity to identify good models is even better. CAPRI provides the possibility for scorers to test their scoring methods on the targets. Rounds are the perfect way to test, try and train every modeling and scoring algorithm.

Rounds can be divided into different steps as shown on Figure 29. The first is to retrieve target complex structures to create a Round. Then predictors can register to participate in this round while signing a confidentiality agreement regarding the data. When the Round begins, predictors are being sent the sequences of the binding partners together

with their stoichiometry. From these sequences each participant is asking to provide two sets of models. The first one, called “P-set”, consists of the ten best models according to their own algorithm. The second one retrieves one hundred models (ranked or not, but including the ones from the P-set) and it is called “U-set”. These two sets are then sent to the CAPRI Assessment members where they are concatenated with the other P and U-sets from the different predictor groups and randomized (shuffled). Once these two sets are completed, the Round is opened to scorer participants. The scorers will be provided the whole U-set and have to select according to their scoring functions their ten best models. These models are collected into a “S-set”.

The assessment step briefly consists of comparing every model provided for every set to the experimental solution with different assessment criteria. These criteria are f_{nat}, $f_{non\text{-}nat}$, i-rms,l-rms and s-rms and allow to rank a model between four possibilities. Quality ranks are incorrect, acceptable, medium and high quality and they have been described in Section 1.C.b) with Table 8 while criteria are shown in Figure 11.

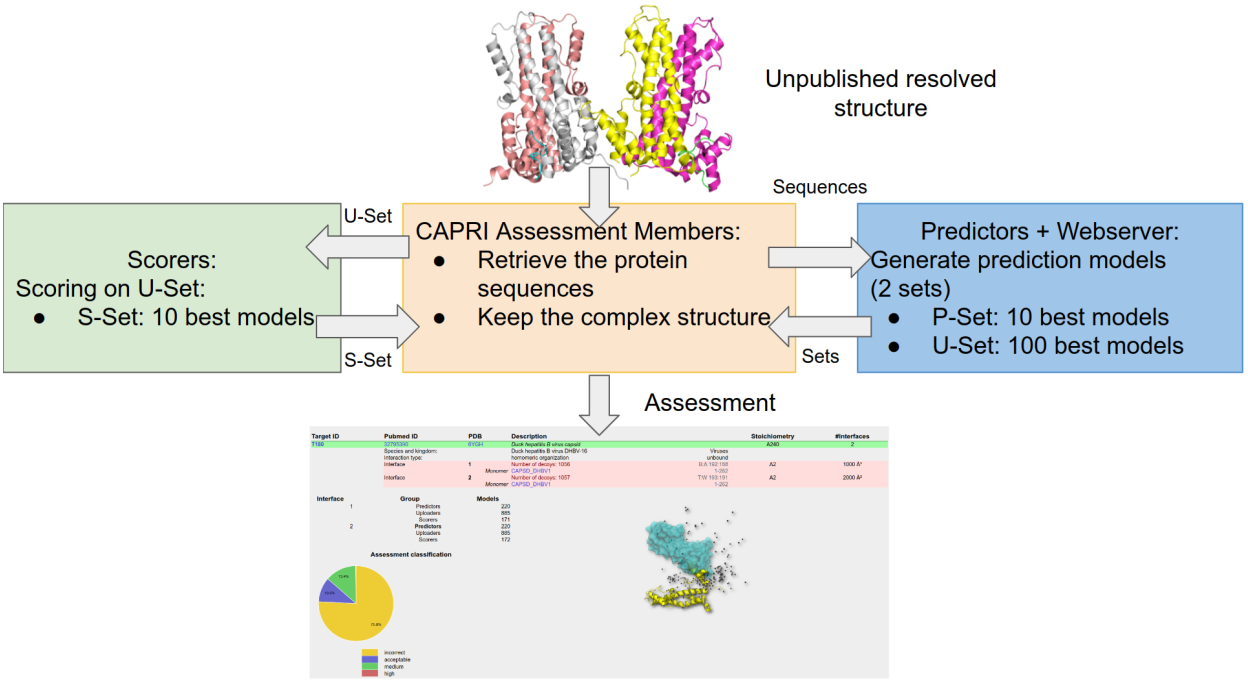


Figure 29. **Schematic representation of a CAPRI Round**

For every published target, models, solution and assessment are available on a dedicated website: scoreset.org.

3. Round 51: A CAPRI-COVID special round

With the CoViD-19 crisis, the researcher community gathered to find solutions to block the spreading, find drugs or at least understand the underlying mechanisms of virus replication. In this optic, CAPRI community wanted to use its knowledge and power in protein-protein interaction. From the publication of Gordon *et al.*, four interaction complexes have been selected because of the availability of the component structure in PDB[158]. These complexes are called T182, T183, T184 and T185 and are described in the Table in the Material and Method section. In addition to these four targets, another virus-host interaction proposed during CASP14-CAPRI has been added called T181 also described in the Table[87]. These five targets compounded the Round 51 of CAPRI. Contrary to the usual CAPRI Rounds, the solutions here were not resolved yet. The goal also differs from the other Rounds because the aim here is to be able to find good models which should be the most likely. To this, this Round was an Open Science Initiative meaning that every result was publicly available so as to be usable for other participants to have better predictions. Also, as no solution is available at this moment, the need to find a way to select the best models is also a major step in this special CoVid-19 Round. This model selection needs a validation set to be approved.

B. Material and Methods

1. The targets

For this special Round, a selection of five different targets (T181 to T185) was proposed.

- Target 181 (CASP ID: H1103) is a complex between the SARS-COV-2 protein Orf3a and the human heme oxygenase HMOX1 with a standard stoichiometry A1B1
- Target 182 is a complex between the viral protein Nsp15 and the human NUTF2. Nsp15 is known to have two main functions: endoribonuclease and interfering with dsRNA-interacting proteins. NUTF2 is a small hub protein related to nuclear import. The complex has a probable A1B2 stoichiometry.

- Target 183 is a complex between the human EXOSC8 which is a non-catalytic component of the RNA exosome complex (9 subunits) and Nsp8 is known to bind elements from this RNA exosome complex.

  The probable stoichiometry of this complex is A1B1.
- Target 184 is a complex between the viral Nsp7 and human RhoA with a probable A1B1 stoichiometry. RhoA is a small GTPase that has GTP-bound and GDP-bound forms. Lots of PTM are known to occur on this protein.
- Target 185 is a big complex with 19 chains: one Nsp7 octamer, one Nsp8 octamer, one big viral protein Nsp12 and 2 chains of RNA.

Supplementary information is available on capri-docking website on a dedicated space written by us. All the information has been summarized in Table 11.

CAPRI ID	CASP ID	Components	Uniprot IDs	PDB templates
T181	H1103	Orf3a / HMOX1	P0DTC3 / P09601	6XDC / 1N3U
T182	NA	Nsp15 / NUTF2	P0DTD1 / P61970	6WLC / 1GY5
T183	NA	Nsp8 / EXOSC8	P0DTD1 / Q96B26	2NN6 / 3UB0:D, 2AHM:G, 6XIP
T184	NA	Nsp7 / RhoA	P0DTD1 / P61586	6XIP:C, 3UB0:C, 6M5I:A / 5C2K:A, 4LHW:E, 2J1L:A190
T185	NA	NSP7/ NSP8/ NSP12/ RNA strands: 1,2,3,4	P0DTC1(3860-3942)/ P0DTC1(3943-4140)/ P0DTD1(4393-5324)/ RNA : (1)CAUGCUACGCGUAG (2)CAUGCUACGCGUAG (3)UGCUACGCGUAG (4)CAUGCUACGCGUAG	2AHM.A-D, 6YYT.C, 7D4F.C / 2AHM.E-H, 6YYT.BD, 7D4F.BG / 6YYT.A , 7D4F.A / 6YYTP-T

Table 11. **Summarized information about T181 to T185 of CAPRI Round 51**

2. The different sets

For every set, the models have been curated following the template provided by CAPRI. For each target, its template indicates the IDs of the chain, the protein sequences and the stoichiometry. All information has been verified and if needed the sequences have been renumbered. If a model was submitted to the wrong Target, it was manually reattributed to the correct target. But if a model belongs to no target (if the sequence does not match for

example or the stoichiometry is wrong) it was removed. For this special Round a maximum of 30 predictor groups and 19 scorers participated. The total number of models for each category of set are described in Table 12 below.

Target ID	number of predictors	U-set number models	Number of Scorers	S-set number of models
T181	26	1257	19	185
T182	30	1972	19	181
T183	27	1523	19	164
T184	30	1811	19	190
T185	NA	878	NA	164

Table 12. **Number of models for every CARPI Round51 targets regarding the predictors and the scorers**

As shown in previous CAPRI's works, score sets (S-set) have overall better models than P or U-sets [149]. Since our goal was to find the best model, we henceforth only focus on the S-sets.

3. Analyses of interface residues

a) Interface residue composition

To compare the residue composition at the interface for the different models, we transformed all models into RINs (Residue Interaction Networks). These networks establish an edge (interaction) between pair residues if their distance falls between 2.5 and 5 Å [193,194]. RINs text files are constructed as followed:

Residue1	ContactType	Residue2	Distance
Met1.B	Residue:IntraContact:Residue-Residue	Glu2.B	2.81
Met1.B	Residue:IntraContact:Residue-Residue	Ser92.A	4.86

From this information, the residues at the interface are retrieved and then the number of times a residue has been predicted to be at the interface in the different models

calculated thanks to a home-made Python script. A special attention has been paid when the structures are dimers to get interactions between the receptor and the ligands specifically as the dimer structure is known and so are the interactions between the two chains. From these networks three different types of information were retrieved, which are listed below from less to more detailed :

- Interface residues which are amino acids between 2.5 and 10 Å from the other chain.
- Contact residues are residues within 5 Å from the complex partner
- Specific contacts are the contacts between entities that were determined using a 5 Å distance threshold.

Clashes (contacts below 2.5 Å), if any, were ignored. Only inter-chain contacts were considered.

b) Residue conservation

Conservation of residue has been calculated thanks to the Rate4Site algorithm. It provides a score where a low result means a high conservation. This software is based on the maximum likelihood principle and maps the rate of evolution among homologous proteins onto the surface of the molecule of one of the homologous proteins with a known three-dimensional structure [195].

c) Visualization

To easily compare residue conservation and the one predicted to be in contact, molecular visualization is a powerful tool. For each target, a PyMOL session has been set with the two same molecules in the same orientation. For one molecule, the bfactor of the pdb files have been replaced with the conservation and the other molecule with the number of times the residue have been predicted to be in contact normalized by the number of models [171,196].

Non-default PyMOL parameters:

- solvent radius = 2
- surface quality = 2

To avoid visual bias from overrepresented residues, we capped their number of contact at the 90th percentile

Also, to visualize where the location of the interactions are, for every model, the center of mass of the different elements have been calculated. These coordinates have been set to a sphere visible.

Another visualization has been realized with barplots. On these barplots are representing the contact hits and the Rate4Site score multiplied by 100 to a better visualization. The plots have been created with Google Sheets.

4. Clustering

To find a consensus between all the models, we performed a clustering on all the models based on different features: interface residues (residues which are up to 10 Å from the other chain of model), contact residues (up to 5 Å) and specific contacts. As clustering is a method to see the distribution of elements according to their distances, we defined the distance (*D*) between two models as followed and illustrated in Figure 30:

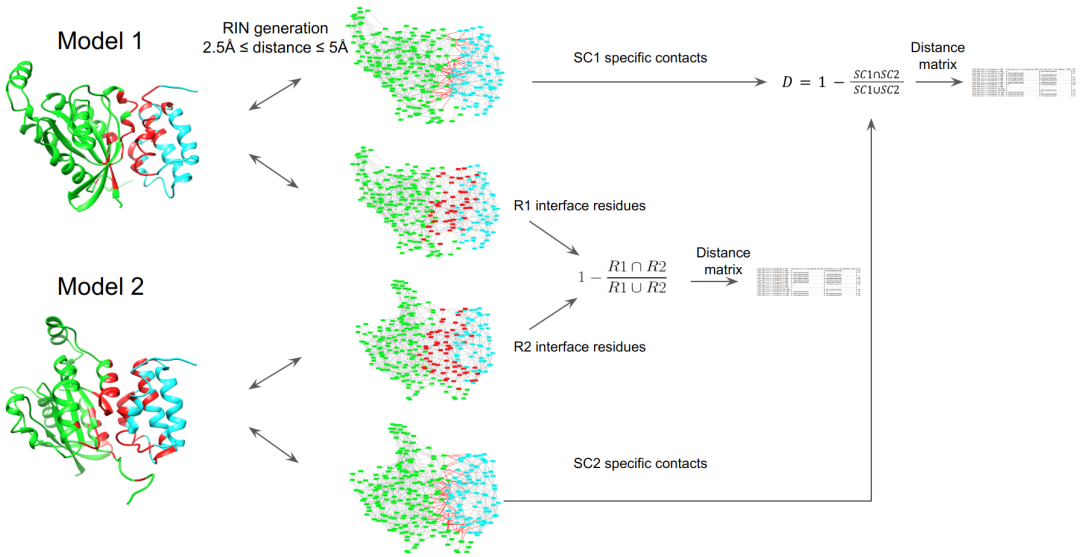


Figure 30. **Schematic representation of distance calculation from models through RINs**

- Interface and contact residues:

  $D = 1 - \frac{R1 \cap R2}{R1 \cup R2}$ with *R1* interface or contact Residue from model1 and *R2* from model2

- Specific contacts:

  $D = 1 - \frac{SC1 \cap SC2}{SC1 \cup SC2}$ with *SC1* specific contacts from model1 and *SC2* from model2

For the S-set, if two different models have a distance of 0, it is considered as redundant because it can be a model selected by two different scorers. In that case the second one is removed from the list of models. To perform the clustering, we used the R software with agnes (Agglomerative Nesting) R package to be able to cluster the solutions.

Different clustering methods have been compared with the linkage criteria: Average, Single, Complete and Ward. In principle, Hierarchical clustering produces an amount of clusters starting with the number of elements and decreasing iteratively to one big cluster [197]. To set a representative number of clusters to represent every different solution, we decided to use a cut-off value. This value corresponds to the similarity inside a cluster, so a cut-off of 0.25 means that inside a cluster all models have at least 75% of similarity or a distance ⩽ 0.25 pair-wise. In this study we selected a cut-off of 0.25 after trying different ones like 0.4 and 0.6. To this, a R script has been written to get the good number of cluster (as shown in Figure 31).

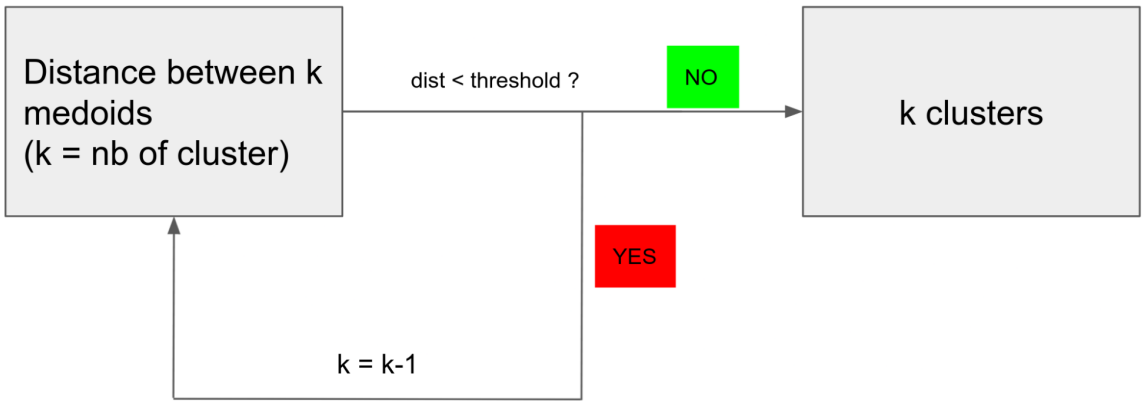


Figure 31. **Schematic representation of the algorithm to define the ideal number of cluster**

5. Meta-clustering

To have a better point of view of the results, we decided to perform some meta-clustering with Cytoscape using Markov clustering (MCL) based on the Jaccard index (1-Distance) between the most representative structure of each cluster (the medoid) [198,199]. This allows us to merge clusters where the clusters are close. For the validation set, this will allow us to see if the largest meta-cluster contains the best solution.

6. Adjacency overlap method

To see how well a model matches the consensus of all models, we calculated the adjacency overlap. For each model an adjacency matrix was constructed from the interaction network. If a specific interaction is detected between two residues in a model, this very contact is equal to 1, otherwise the value is 0. Once the matrix is set for all the models, a meta matrix is then constructed by overlapping all the matrices. Every contact is summed and then the values are normalized with the number of models. Once the metamatrix is calculated, every model is compared to this matrix to see how well this model matches the matrix. To calculate the overlap, we calculate the square root of the sum of all specific interactions multiplied by the value of this interaction in the adjacency matrix squared (*M*):

$$\sqrt{\sum_{i,j}(m_{ij} * M_{ij})^2}$$

where *m* is the contact matrix for one model and *i, j* the residues we are looking for the interaction.

7. Validation dataset

We selected the S-set of 4 targets from previous CAPRI Rounds where the different models have been assessed to different quality according to the CAPRI assessment criteria: f_{nat} and $f_{non-nat}$ which are the fraction of receptor-ligand residue contacts found in the model and the experimental structure and the fraction of contacts which have been predicted in the model and which are not present in the target structure [149]. In addition, to assess the quality of the predicted interface, two other quantities based on Root Mean Square Deviation (RMSD) are used: I-rms which is the rmsd at the interface backbone atoms and S-rms which is the same for the interface side-chain atoms. The different qualities are the following: incorrect, acceptable, medium and high.

The selected targets are listed below with their subjective difficulty according to the number of acceptable or above provide by their scorer sets:

- T039: considered here as a hard complex to predict
- T041: considered here as an easy complex to predict
- T050: considered as a medium difficulty target to predict with a A1B2 stoichiometry
- T053: considered as a medium difficulty target to predict

The number of models and the quality of the different targets are summarized in the following Table 13:

Target ID	Number of models	Incorrect models	Acceptable models	Medium quality models	High quality models
T039	120	120	0	0	0
T041	120	46	54	18	2
T050	140	105	21	14	0
T053	130	98	25	13	0

Table 13. **Information regarding the quality of the models from the S-set of the four CAPRI previous targets chosen to validate the methods**

For each model, an F-score is available for the ligand and the receptor. It corresponds to a rate between precision and sensitivity: $Fscore = 2 * \frac{fIR*(1-fOP)}{fIR+(1-fOP)}$, where fIP is the sensitivity and fOP the overall Precision. Basically, we looked at how many residues at the interface are found for the ligand and the receptor and how many are missing and how many are true.

C. Results

1. Target analyses

For each target and for each component of the complex, the number of hits for interface residues, contact residues and specific contacts have been calculated on each model of the S-set. But for the following the target representations only contact residues (residues which are within 5 Å from the other partner) have been taken into account.

Target 181 (HMOX1/Orf3a):

The number of times a residue has been predicted to be at the close interface has been retrieved for the 185 models and their conservation according to Rate4Site can be seen in Figure 32. In parallel, to have a better representation of the interface area, the number of hits has been represented by color shades on the protein surfaces as the conservation. It can be seen on Figure 33. From the two barplots we can see that the residues often predicted to be in contact are not well conserved. For the viral protein, some residues with a high number of hits are more conserved. According to the surface representation we can see that indeed there is a big surface patch often predicted on the HMOX1 protein which is not well conserved. Contrary to the viral Orf3a protein which has many places predicted to be in contacts, areas which are more conserved. According to these results, we can say that the viral protein has a consensus for the binding area which is not the case for the human one. But regarding the center of mass of the different models, even if there is an area of consensus, the position of the different models seems to change a lot. For the human representation we can still see some really close dots meaning that some models are really close.

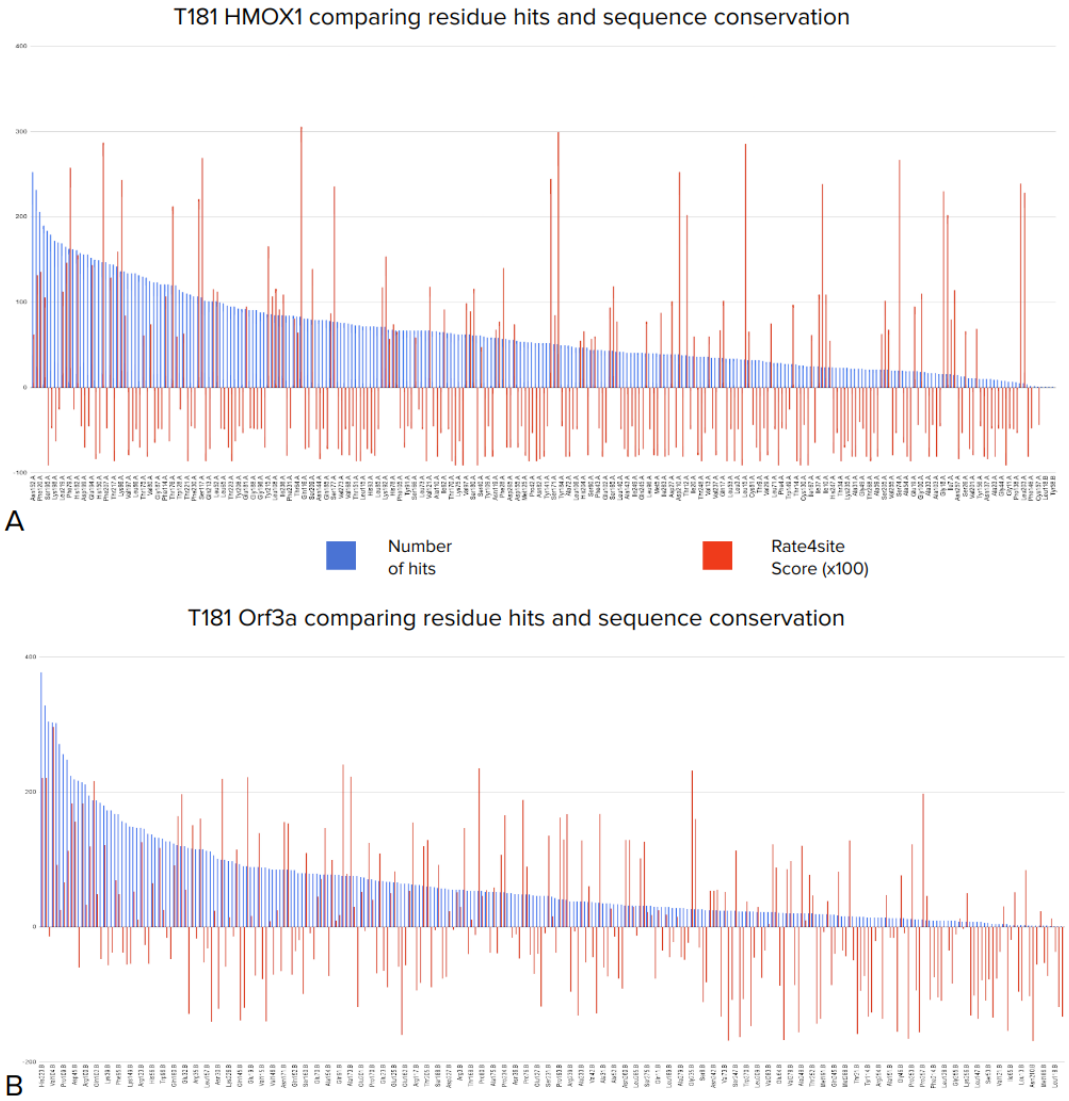


Figure 32. **Barplot of contact hits and residue conservation for human protein HMOX1 (A) and viral protein Orf3a (B) in a complex for the Target 181**

Blue bars are the number of hits and red ones are Rate4Site scores multiplied by 100. The lowest the conservation score is, the highest is the conservation.

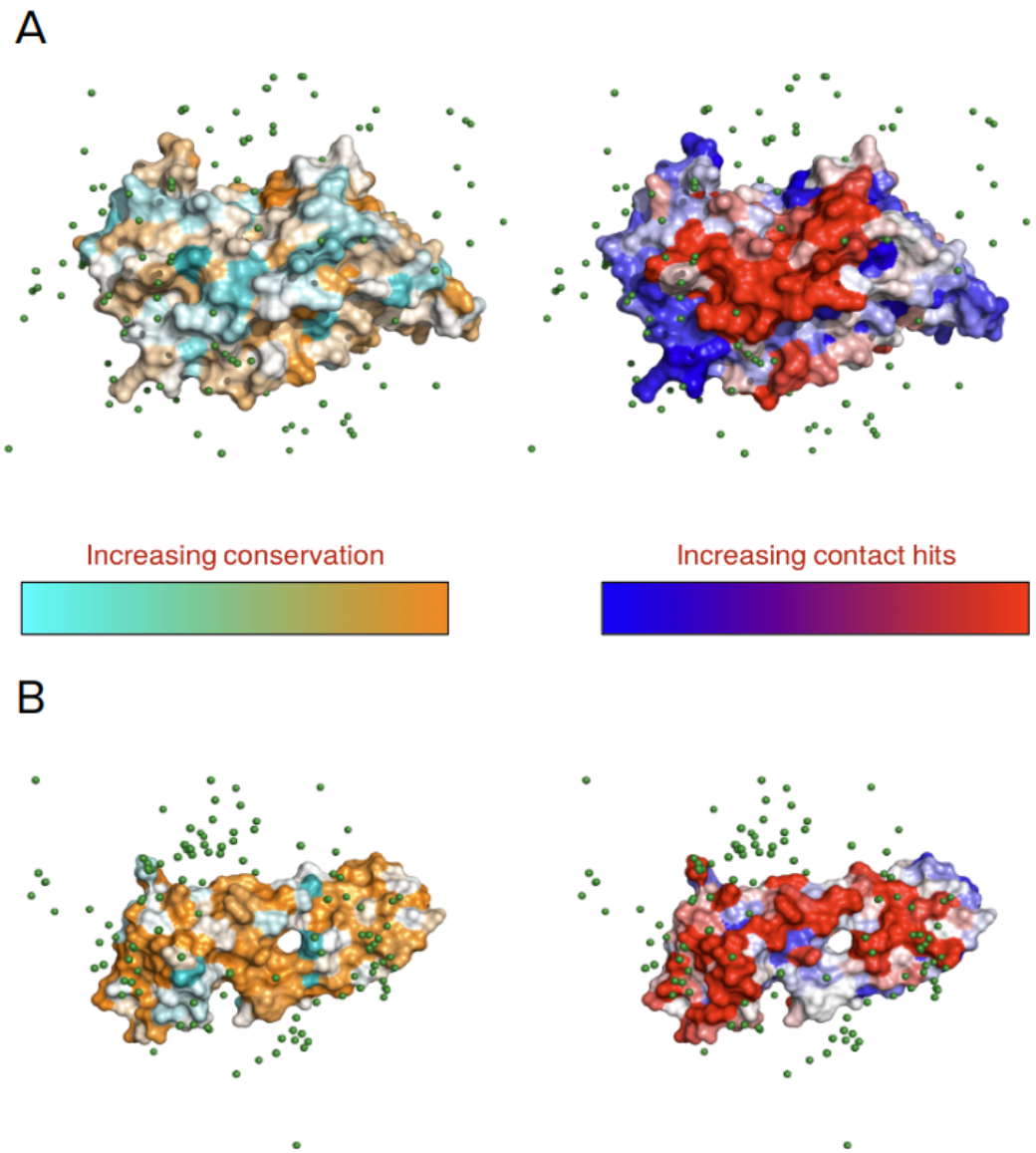


Figure 33. **Surface representation of human protein HMOX1 (A) and viral protein Orf3a (B) with coloration according to residue conservation and contact hits**

Left corresponds to the residue conservation from cyan to orange and Right to the contact hits where red is for high occurrence, capped at the 90th percentile for a better visualization to blue with few occurrences. Green spheres around are the centers of mass of every partner predicted by the different models.

Target 182 (NUTF2 / Nsp15):

This target was a little bit different from the others as one of the complex components is a dimer. There is one barplot per chain for NUTF2: one called NUTF2.A eand the second called NUTF2.B. We can see on the Figure 35 that contact hits and obviously conservation of amino acids are very similar for the two different chains of NUTF2. The most

conserved residues for the human protein are also well conserved. For the viral protein, the results are more heterogeneous with residues well conserved and others often mutated. Regarding Figure 34, NUTF2 dimer surface representation shows as for T181 a big patch of ten predicted to be at the interface. This is supported by the many green spheres very close showing some consensus for the different models. The surface representation of Nsp15 shows more conservation with the highlight of two possible interaction faces even if one is more often predicted (lot of green spheres in this area).

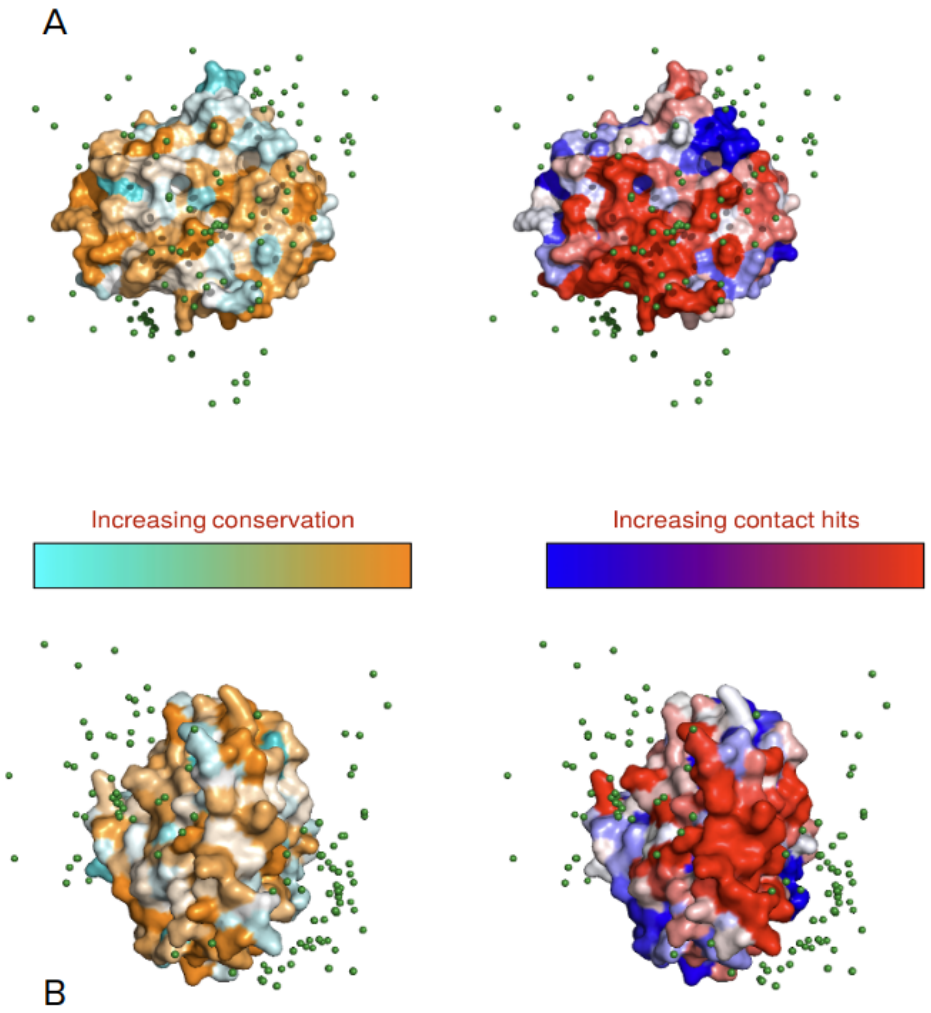


Figure 34. **Surface representation of human protein NUTF2 (A) and viral protein Nsp15 (B) with coloration according to residue conservation and contact hits**

Left corresponds to the residue conservation from cyan to orange and Right to the contact hits where red is for high occurrence, capped at the 90th percentile for a better visualization to blue with few occurrences. Green spheres around are the centers of mass of every partner predicted by the different models.

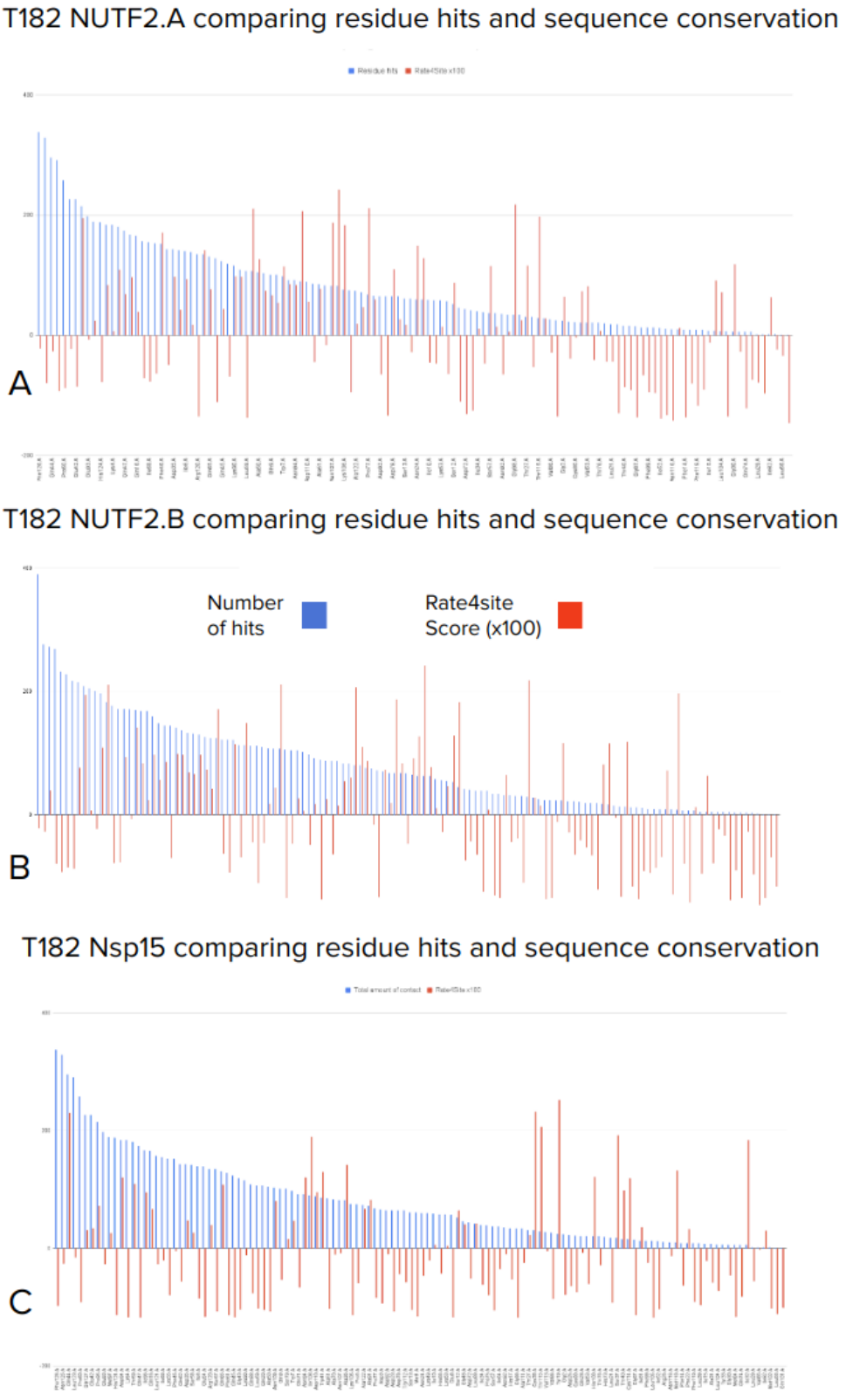


Figure 35. **Barplot of contact hits and residue conservation for human dimer NUTF2 (A,B) and viral protein Nsp15 (c) in a complex for the Target 182**

Blue bars are the number of hits and red ones are Rate4Site scores multiplied by 100. The lowest the conservation score is, the highest is the conservation.

Target 183 (EXOSC8 / Nsp8):

The EXOSC8 most redundant residues at the interface are mostly conserved and very often predicted to be at the interface as Leucine 42 is found almost 400 times at the interface in 165 different models (see Figure 36 A) But this is also the case for Nsp8 where a consensus regarding the interface residue can be seen on the barplot with more than 500 occurrences for the three main residues (see Figure 36 B). While looking at the surface representation on Figure 37, we can find the big patch on the Nsp8 cavity and also a big patch on the EXOSC8 protein. But as this last one is a thin protein, the possibilities of interaction are very high explaining the disparity of the green spheres representing the different center of mass. Contact residues, here, show a consensus but if we are looking at the specific contact we can see that they differ from the different models.

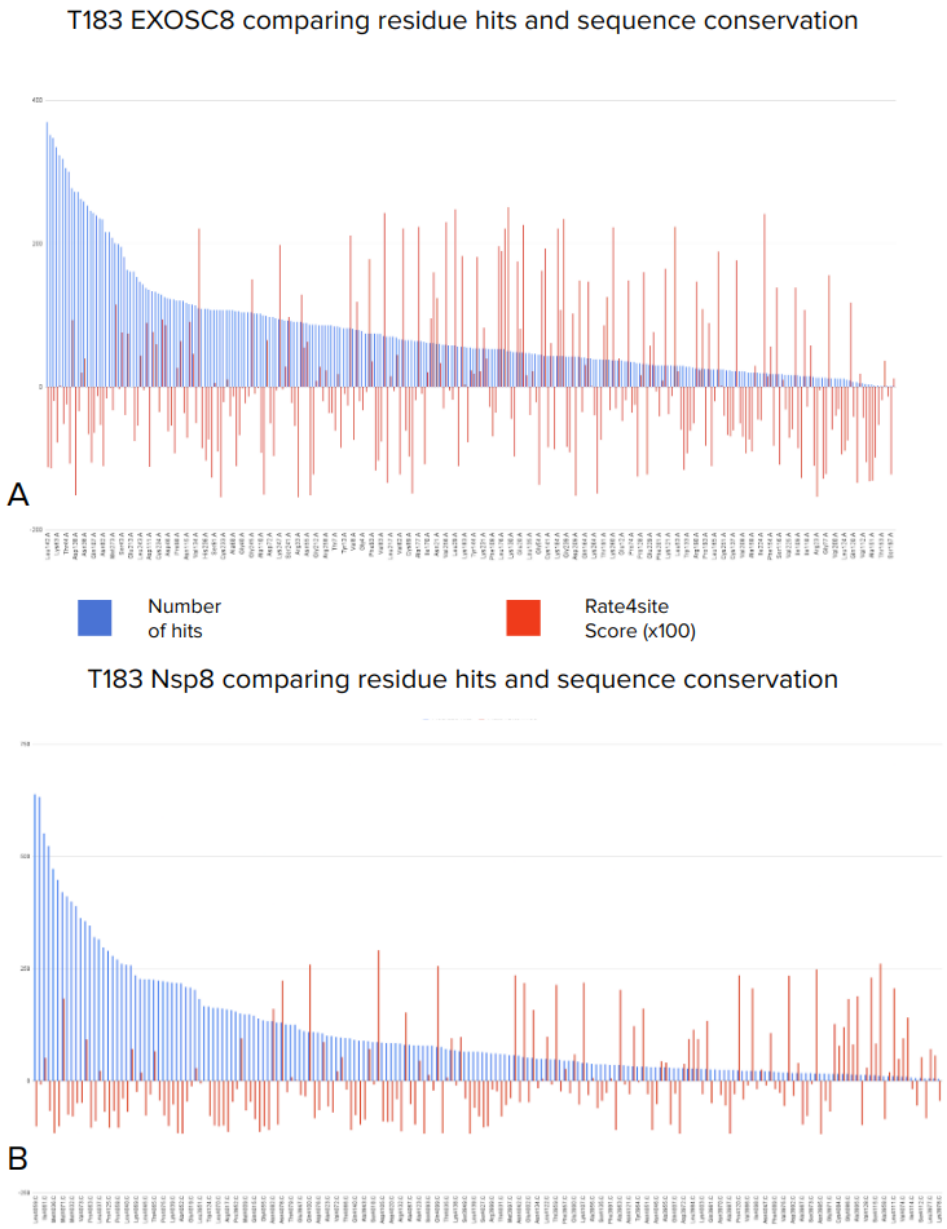


Figure 36. **Barplot of contact hits and residue conservation for human protein EXOSC8 (A) and viral protein Nsp8 (B) in a complex for the Target 183**

Blue bars are the number of hits and red ones are Rate4Site scores multiplied by 100. The lowest the conservation score is, the highest is the conservation.

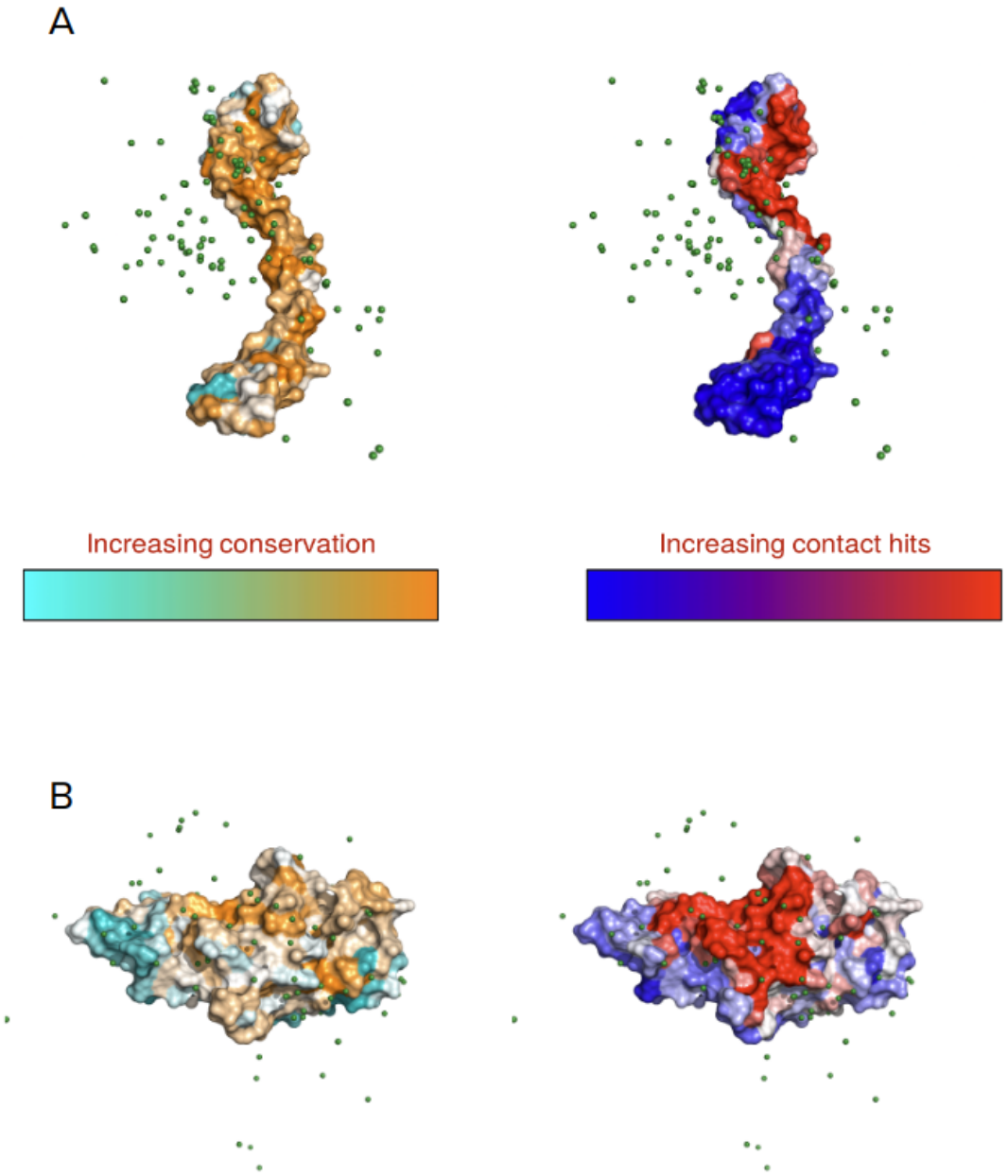


Figure 37. **Surface representation of human protein EXOSC8 (A) and viral protein Nsp8 (B) with coloration according to residue conservation and contact hits**

Left corresponds to the residue conservation from cyan to orange and Right to the contact hits where red is for high occurrence, capped at the 90th percentile for a better visualization to blue with few occurrences. Green spheres around are the centers of mass of every partner predicted by the different models

Target 184 (RhoA / Nsp7):

Regarding the Figure 38, the most predicted to be in contact residues are more conserved for the viral protein. Also these residues are well predicted to be in contact (between 400 and 500 times for the first four) and as it can be seen on Figure 39 they form an area in front of which we can see a lot of green spheres meaning the predictions are very consensual. Regarding the human protein three-dimensional representation, we can see thanks to the different centers of mass two mains area where the interaction is predicted. This target seems to be the one with the more consensus.

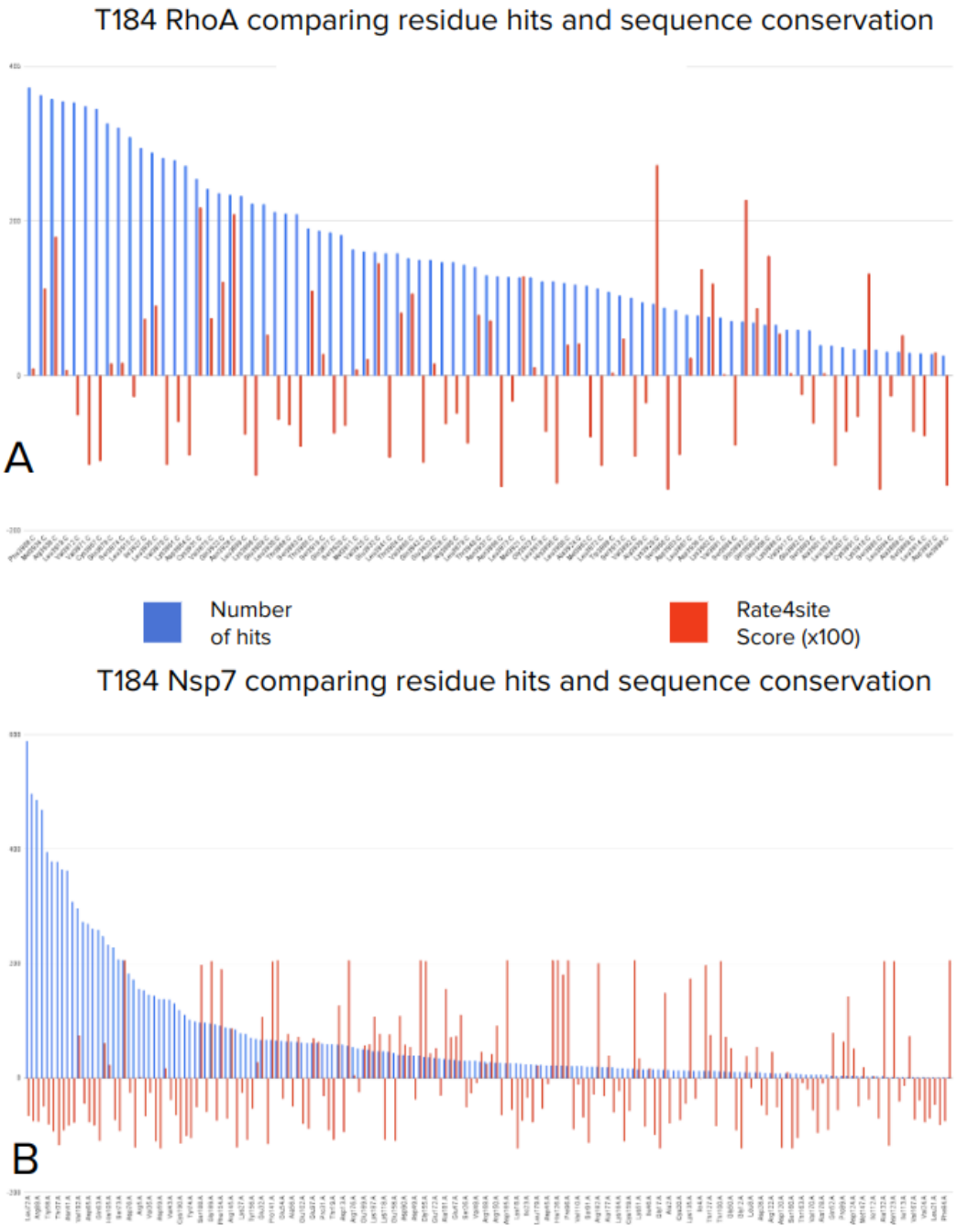


Figure 38. **Barplot of contact hits and residue conservation for human protein RhoA (A) and viral protein Nsp7 (B) in a complex for the Target 184**

Blue bars are the number of hits and red ones are Rate4Site scores multiplied by 100. The lowest the conservation score is, the highest is the conservation.

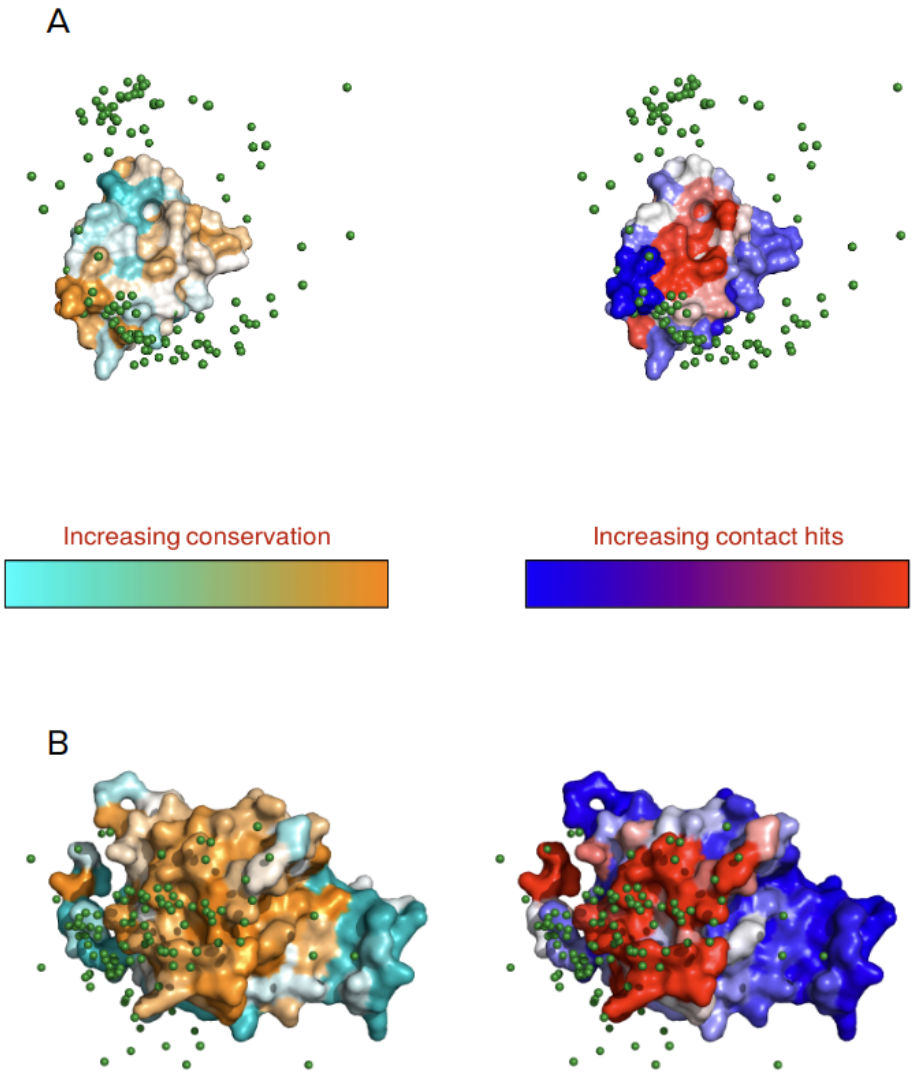


Figure 39. **Surface representation of human protein RhoA (A) and viral protein Nsp7 (B) with coloration according to residue conservation and contact hits**

Left corresponds to the residue conservation from cyan to orange and Right to the contact hits where red is for high occurrence, capped at the 90th percentile for a better visualization to blue with few occurrences. Green spheres around are the centers of mass of every partner predicted by the different models

All these results in a csv format and the associated PyMOL sessions can be found on the capri-docking website.

Target 185 (Nsp7/Nsp8/Nsp12/RNA):

This target was a particular target with 17 proteins and two double stranded RNA leading to a very big complex which can not be analyzed as the four others. To this, we split the complex into binary problems: interactions between the two octamers of Nsp7 and

Nsp8, but also interactions between Nsp7 and RNA, Nsp 12 and RNA. Nevertheless most of the complex structures selected by the score are very similar even if when fitted on the double stranded RNA we can see the symmetry is different. Regarding the full complex, we can assume two lines of approach for the construction of this multimeric structure. The first one, represented in Figure 40 is first a complex between the Nsp7 octamer and RNA, then consolidated with the Nsp8 octamer followed by Nsp12 at the bottom. The second story that we hypothesize is first the creation of the complex between the two Nsp7 and Nsp8 octamers which will then surround the RNA and the complex may be stabilized by the Nsp12 (see Figure 41). For each story, each step has been analyzed separately to provide specific contact information. As a lot of interacting residues are buried inside the complex, the surface representation was not informative enough explaining the cartoon representation for this target. These two figures show us how strong is the interaction inside the Nsp7/Nsp8 complexes but also their contact with the RNA. The Nsp12 which is the distant part seems to interact essentially with the RNA and a little with Nsp8.

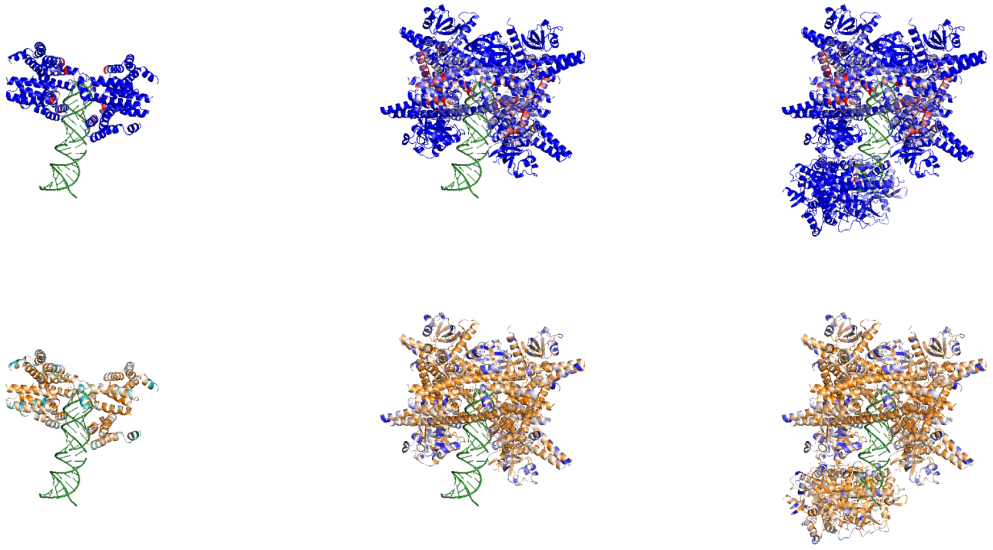


Figure 40. **Creation of the T185 complex (Story 1)**

The color chart corresponds to the previous figures. For each step (from left to right) the colors of the residue hits (blue and red) are recalculated to correspond to the whole complex. The first step is Nsp7 with RNA, the second Nsp8 is added and then Nsp12.

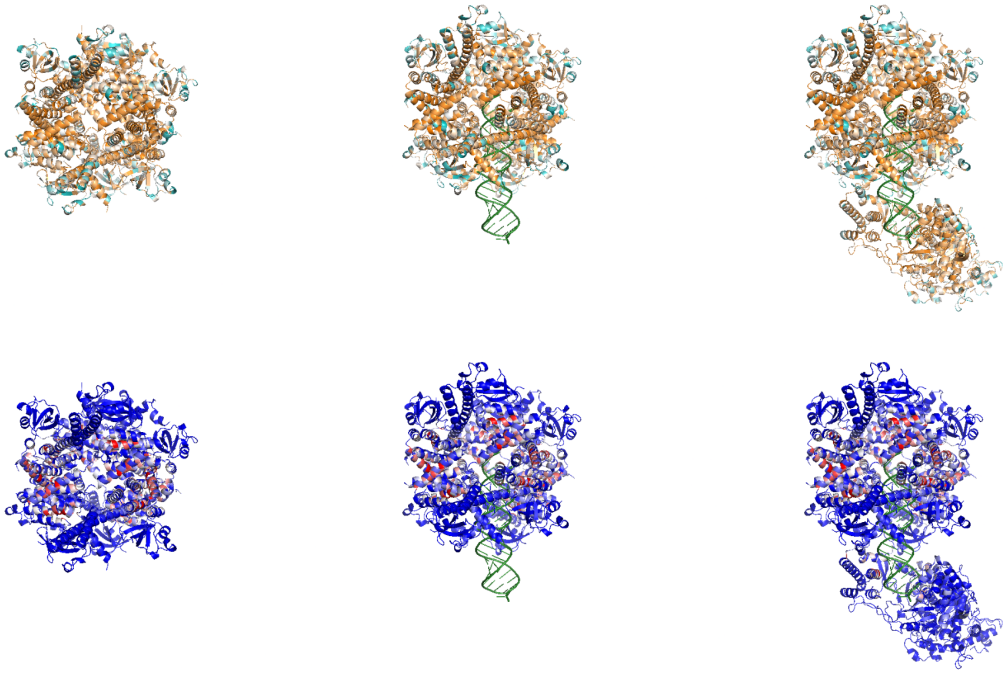


Figure 41. **Creation of the T185 complex (Story 2)**

The color chart corresponds to the previous figures. For each step (from left to right) the colors of the residue hits (blue and red) are recalculated to correspond to the whole complex. The first step is Nsp7 with Nsp8, the second RNA is added and then Nsp12.

As this target is very different from the others, it has been set aside while the other targets are analyzed.

For Targets 181 to 184, the interface composition coupled to conservation did not bring any consensus and it is hard to find a good model with such information. Notably for T181 and 183 which it seems to have no consensus at all according to the different center of mass. We decided then to perform clustering on these models to see if we will be able to find a consensus for each target.

2. Clustering and meta-clustering

To select the hierarchical clustering, the agglomerative coefficient of the four methods have been tested and are retrieved in the Table 14. Regarding these scores, the Ward method is the one that has to be used for the following results.

Target ID	Average	Single	Complete	Ward
T181	0.8006	0.7738	0.8377	0.9365
T182	0.9398	0.9115	0.9571	0.9964
T183	0.9191	0.8801	0.9555	0.9947
T184	0.8916	0.8023	0.9301	0.9945

Table 14. **Agglomerative coefficient of the four hierarchical clustering methods (”Average”, ”Single”, ”Complete”, ”Ward”) applied on the fourth targets T181, T182, T183 and T184**

The clustering has been first performed on validation datasets to see if this method works. Clustering has been done regarding interface residues, contact residues and specific contacts. Looking only at residues gives less accuracy and ends in one big cluster which is not helpful to discriminate good from bad models. Thus, it has been decided to only continue with specific contacts clustering.

Clustering on validation dataset:

Target 039:

Ward clustering has been performed on distances calculated on common specific contacts between the two chains. T039 was defined as difficult as no acceptable or better solution has been found in the S-set and only 4 (1 acceptable and 3 medium quality models) have been predicted on the full U-set consisting of 1400 models. To avoid redundancy, similar models have been removed resulting in a validation set of 120 models.

In total, with a threshold of 0.25, the clustering results in 108 groups with a maximal size of 4 models coming from 4 different scorers. This means a heterogeneity of the specific contacts in the different model which can show that there is no consensus model.

Target 041:

This target is considered as an easy target with more than 60% of acceptable or better quality models. The U-set is composed of 1199 models with 22.5% of acceptable models, 11.9% medium quality and 1.8% of high quality models. The complex only consists of two different interacting chains. The Ward clustering has been processed on the 10

models of the S-set resulting in 105 clusters with a biggest cluster size of 3 obtained two times.

Target 050:

The target 050 is an interesting table as validation set because the complex is made with a AB:C stoichiometry which can be similar to Target 182 because AB can be considered as one as NUTF2. The U-set is composed of 1451 models with 7.9% of acceptable models and 3.4% of medium quality models. The S-set consists in 140 models with 25% of acceptable or better models (without any high quality models). These models have been clusterized into 113 clusters with 6 biggest clusters with a size of 3.

Target 053:

This target is similar to T050 in terms of difficulty with around 75% of incorrect models but this target has a standard stoichiometry A1B1. This target U-set is composed of 1400 models where only 9% has acceptable quality and 2.1% has medium quality. All the 130 predicted models of the S-set are gathered into 107 clusters with a maximum size of 3 for three clusters.

Finally the clustering on specific contacts regarding the validation set is too stringent to find a consensus even for “easy” targets such as T041. An interesting point is the variety of the scorer groups inside a cluster. We can see in the clustering results that the biggest clusters have models coming from different scorer groups while the predictors groups are often repeated. This means that different assessment methods lead to highlighting similar models predicted by similar methods. The discrimination by the specific contacts is too stringent and some models are isolated in a cluster while not that far from the other. From this observation we wanted to perform a meta-clustering to regroup models/clusters close from each other.

Target 039:

Meta-clustering of the 108 clusters from this Target 039 determined by Ward hierarchical clustering with Markov Clustering (MCl) results in many meta-clusters as it can be seen in Figure 42 (A). Indeed, from the 108 clusters we finally obtain 15 meta-clusters

plus two singletons. An interesting thing is that the main cluster (C1) is not in the first meta-cluster. As there is no acceptable data it is difficult to say if a meta-cluster contains better results than others and if it is the biggest one but other quantities are available. These quantities are interface recall and precision and were calculated thanks to the available correct structure.

Target 041:

The meta-clustering of the 105 clusters results in a very lower number of meta-clusters contrary to T039. Indeed, we can see on Figure 42. (B) that all the models are dispatched into 4 meta-clusters plus one singleton. The first metal-cluster, which is the biggest, is mostly composed of the clusters with correct models except for 3 clusters which are incorrect. Looking at the Fscore of these clusters (or the size of the node) they are very close to be acceptable and this explains why they are near acceptable models. We can also remark that every cluster has a lot of links with the other clusters even between the metacluster meaning that all the models are very close regarding the specific contacts and a consensus should be easily findable.

Target 050:

Meta-clustering for Target 050 shows a profile similar to the one from T039 (Figure 42 (C)). Indeed, the 107 clusters are classified in 13 meta-clusters plus one singleton. But as this target has few models which are at least acceptable, we can find them in the first and biggest metacluster. As T041, inside this meta-cluster, clusters with incorrect solutions are found. But once again, regarding the F Score the models are close to be acceptable. Meta-clusters have many links between them but mostly between the two first and between the first and the last metacluster with only clusters inside. This last one seems to also have better solutions than the other small meta-clusters.

Target 053:

This target shows a very particular profile with only one meta-cluster with every solution except three singletons as shown in Figure 42 (D). Inside this meta-cluster we can see good solutions which share a big link between them but it is also the case for incorrect models. Regarding the F Score we can see that very poor quality solutions are found inside

this metacluster and without this information it would be difficult to separate good from bad models.

Meta-clustering seems to be a good solution to highlight good solutions in a Target according to this validation set but it is dependent on its difficulty. Indeed, the first cluster seems to contain good solutions but it does not mean that there is a good solution as it is the case for T039. We can hypothesize that a low amount of meta-clusters is synonymous to having an easy target but T050 and T039 have similar profiles while one has good solutions whether the other has no good answers inside. With these results we decided to apply clustering and meta-clustering to the CAPRI Round51 targets, hoping we will be able to highlight good results or consensus as T041.

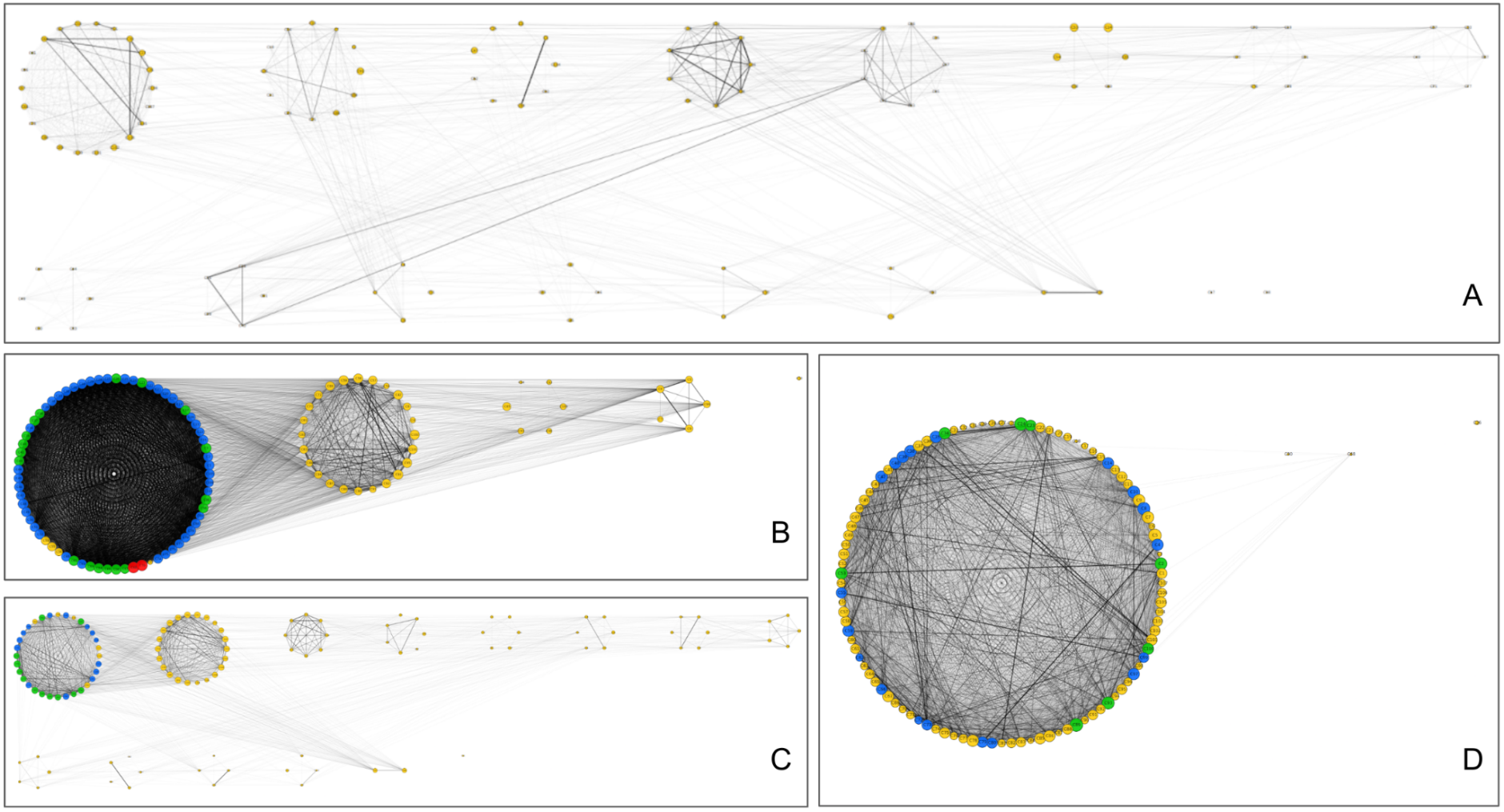


Figure 42. **Meta-clustering representation for Targets T039 (A), T041 (B), T050 (C) and T053 (D)**

Colors represent the quality of the best model inside a cluster. Yellow = incorrect, blue = acceptable, green= medium quality and red= high quality. The size of the nodes corresponds to the highest mean F-score inside the cluster. Links between clusters correspond to the Jaccard Index (1-Distance)

Target 181:

From the 185 models, 44 models which were redundant have been removed resulting in a total of 141 models. The hierarchical clustering performed on these 141 models regrouped them in 120 clusters with two biggest clusters of 3 models which is very similar to results obtained with the validation set. Meta-clustering has been performed of these clusters and the results can be observed in Figure 43 (A). Results will be described and discussed after with all the CAPRI CoVID Round targets.

Target 182:

After redundancy curation, the number of models drops from 181 to 135. After hierarchical clustering, we count 120 clusters as T181 with one biggest cluster composed of 3 models from different scorers. These 120 clusters have been clustered with Markov clustering and the results are shown in Figure 43 (B). As T181, the results will be described and discussed soon.

Target 183:

Target 183 is the target with the lowest number of models (164). After curation, this number decreased to 127 models clustered in 102 groups with a biggest cluster containing 4 models selected from 4 different scorers. The meta-clustering results of this target can be seen on Figure 43 (C) and will be described and discussed after T184.

Target 184:

This target counts a total of 190 models reduced to 146 after redundancy curation. These 146 models have been regrouped in 115 clusters with the Ward Hierarchical clustering. This target shows good consensus with 3 biggest clusters of 4 models selected by a different scorer each time. Meta-clustering of these 115 clusters is shown in Figure 43 (D).

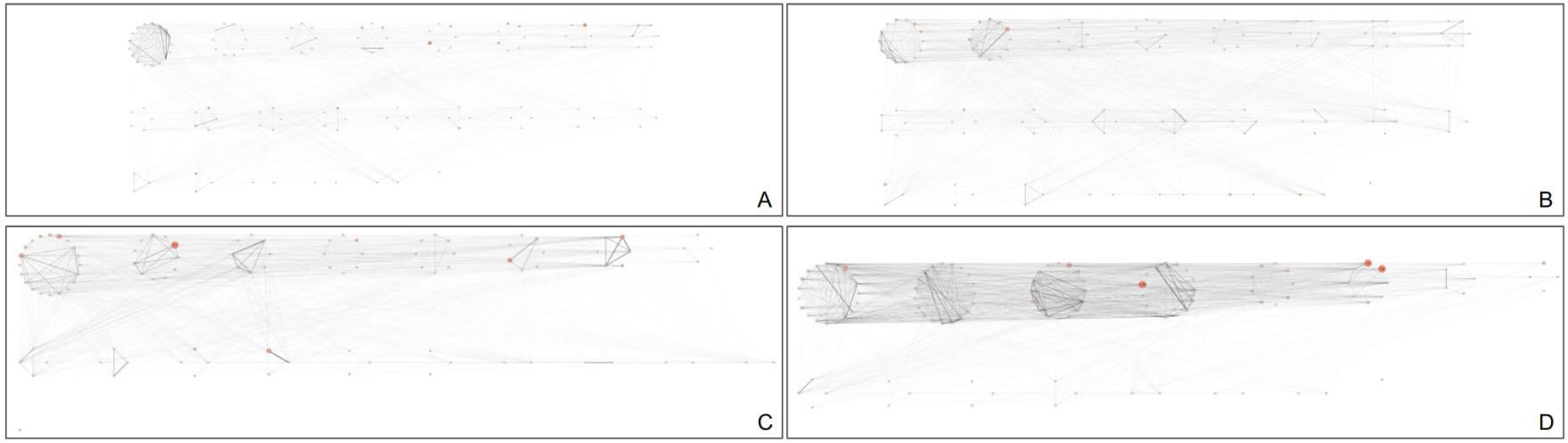


Figure 43. **Meta-clustering representation for Targets T181 (A), T182 (B), T183 (C) and T184 (D)**

Colors represent the number of models inside a cluster. The size of the nodes corresponds to the number of different scorers inside a cluster. Links between clusters correspond to the Jaccard Index (1-Distance).

Regarding the Figure 43 in general, we can see that all CAPRI-Covid Round targets have similar profiles with a high number of meta-clusters which is a result close to T039 where no good solution is available. Regarding the links between clusters we can see that T182 and T184 have more consensus than T181 and T183 meaning that good solutions may be available in these targets but can't be found with this method. We need to define a new method to find good solutions based on consensus.

3. Adjacency overlap scoring

Regarding clustering and meta-clustering, we can see that models share specific contacts and these methods do not allow highlighting the models which are sharing the most of them. To be able to find the model which is the most consensus-based to every model, adjacency overlap has been performed on the validation set to see if this method can separate good from bad models and tell if a model can be correct. This method consists of retrieving every specific contact for every model to create a big pattern of interactions. Then each model is compared to this pattern and ranked. Applied to the validation set we obtain the following results:

Target 039:

The output of the adjacency overlap regarding the target gave models with a low score while the best model tends to hardly reach a score of 0.04. Regarding the distribution of all the scores and ranks of this difficult target in Figure 44, the score can be divided in three parts. The first group consists of the seven best models according to our scoring method. After retrieving these models i-rms and l-rms we can see that theur i-rms scores are between 15 and 20 which is slightly better than the other models but nothing significant.

Target 041:

This target is the easiest target of the validation set and all the good models were found in the same meta-cluster. Here, with the adjacency overlap, we can see on Figure 45 that all the good models (acceptable or better) are ranked first with a best score of 0.0875. Regarding the distribution of the results on the graph, the different dots can be divided in three parts. The first part consists of models with a score between 0.0875 and 0.0600 which

are classified as acceptable or better. But an interesting result is that the two models which have been classified as having high quality are not ranked in the first models according to our developed method. The second part consists of a majority of incorrect models except for one acceptable model. The model scores are between 0.0300 and 0.0600. According to the DockQ scores these models also have a score below the ones from the first part but 3 times superior to one from the third part. The last part is composed of incorrect models with an adjacency overlap score between 0.0200 and 0. Regarding these results it could be interesting to compare dockQ with our method to see if there is a correlation.

Target 050:

The stoichiometry of this target was particular as an AB:C one. The chains A and B are considered as one and only interactions between the C chain and the AB complex are taken into account. Regarding the result of the adjacency overlap on this target (Figure 46), we can see as for T041 good results in the first ranked with a majority of medium then acceptable models in the first part of the graph. Then the incorrect models are mixed with acceptable models in the following part. On this target we can see that medium quality models are well discriminated.

Target 053:

This target, very similar to the Target 050 regarding the percentage of acceptable or better models but different in terms of stoichiometry. Looking a the Figure 47, as for T041 we can find good quality models in the best ranked model according to the adjacency overlap scoring. These good models have a score between 0.046 and 0.060 which is lower than incorrect models from T041. Also medium quality models are in average better ranked than acceptable models. In addition the presence of 3 models considered by the CAPRI criteria as incorrect is an interesting result. Regarding these models l-rms and i-rms scores, we can see that they are close to be assessed as acceptable with scores of 11.4135 and 6.0368 for the best incorrect model and for 11.6438 and 4.7386 the second for a maximum of 10 and 5 respectively to be acceptable.

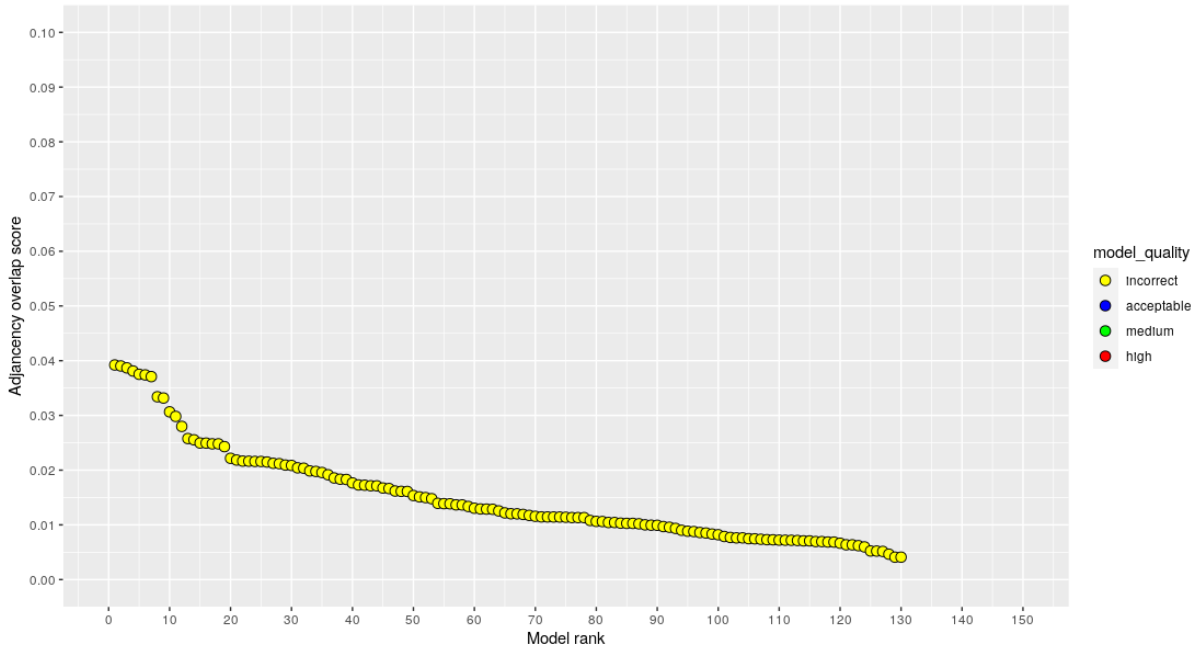


Figure 44. **Plot of Validation set - T039 models ranked with their adjacency overlap scores**

Colors corresponds the model qualities: Yellow = incorrect, blue = acceptable, green = medium quality and red = high quality

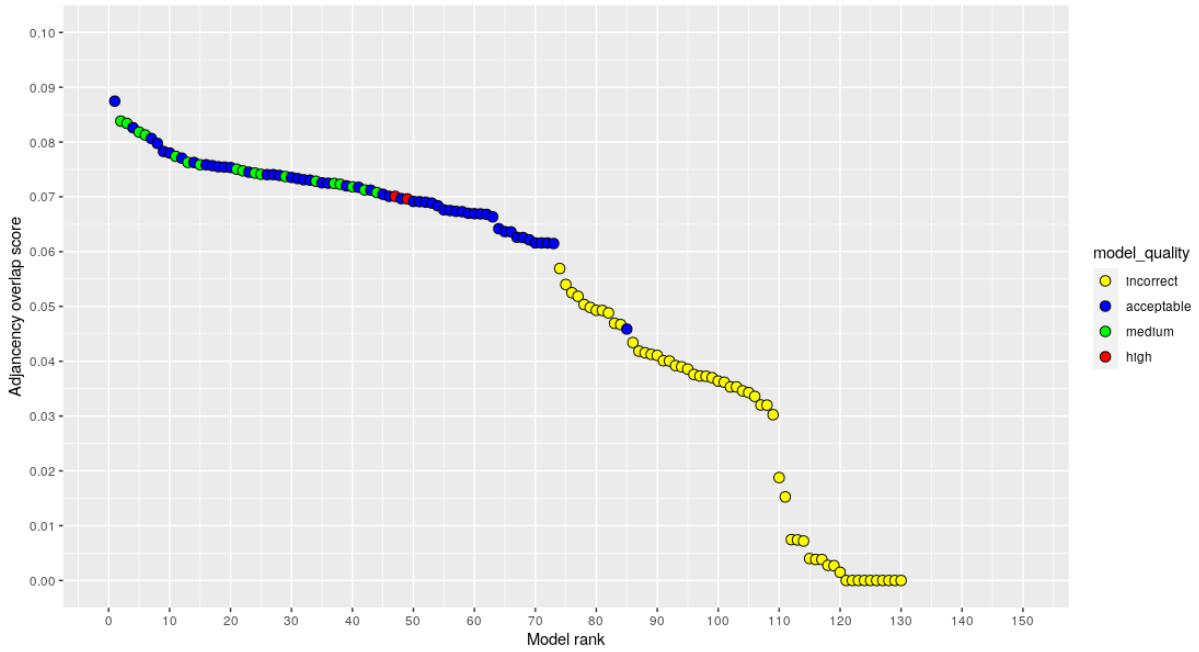


Figure 45. **Plot of Validation set - T041 models ranked with their adjacency overlap scores**

Colors corresponds the model qualities: Yellow = incorrect, blue = acceptable, green = medium quality and red = high quality

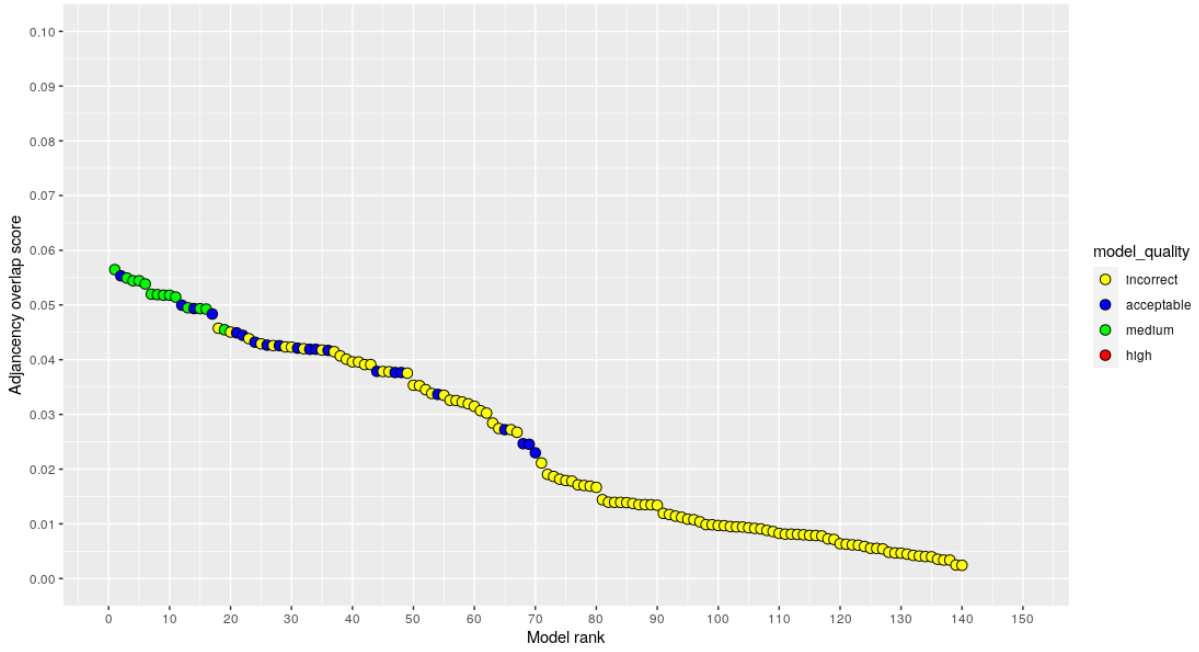


Figure 46. **Plot of Validation set - T050 models ranked with their adjacency overlap scores**

Colors corresponds the model qualities: Yellow = incorrect, blue = acceptable, green = medium quality and red = high quality

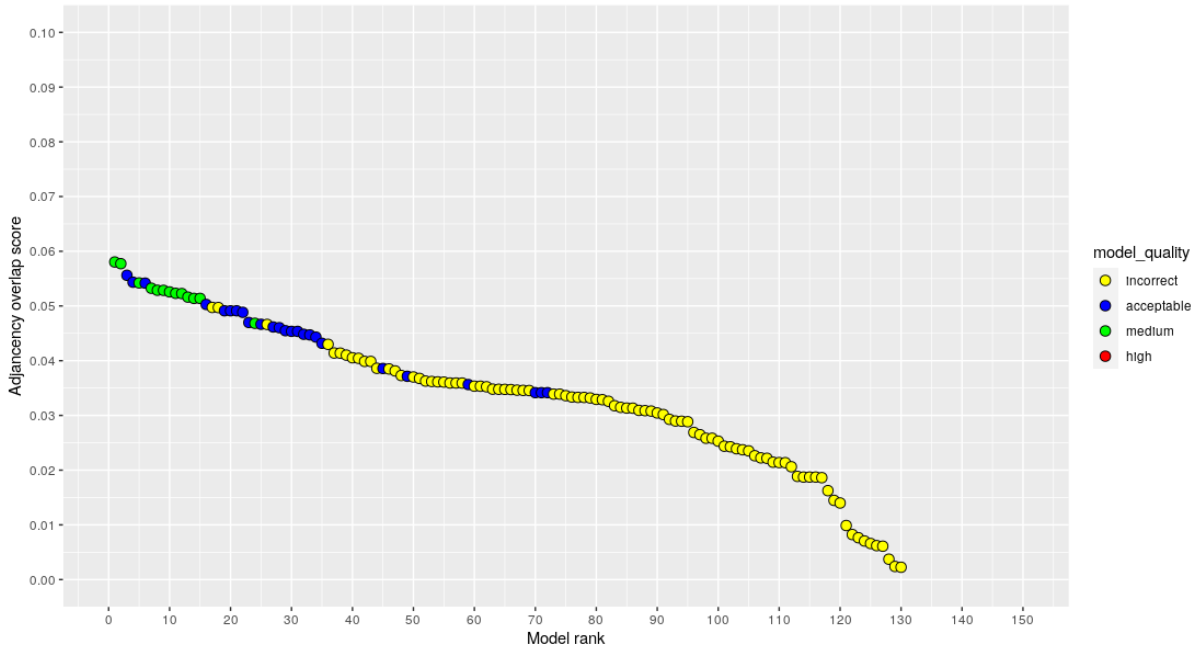


Figure 47. **Plot of Validation set - T053 models ranked with their adjacency overlap scores**

Colors corresponds the model qualities: Yellow = incorrect, blue = acceptable, green = medium quality and red = high quality

According to these results, the adjacency overlap seems to work in discriminating good from bad models when there are good solutions. As the score is normalized by the size of the complex and the number of models it is possible to compare results from the different validation set targets. An interesting point should be to define a threshold from which the probability to have a good model is high. If we look at the validation set results a cut-off around 0.04 could be a good one except for T041 where some models assessed as incorrect are above this threshold. As this adjacency overlap scoring method seems to highlight good models, it has been applied on CAPRI Round 51 targets :

Target 181:

The adjacency overlap score shows a few models above the others with a curve profile similar to the one from T039 but with lower scores. These 8 models have a score between 0.023908 for the lowest and 0.025762 for the greatest which is quite low. Figure 48 also shows the scorer rank attributed to every model to see if the more confident models are well ranked by our scoring method. But the results are very heterogeneous and we can see that none of the models ranked 1st is in the first model according to our method.

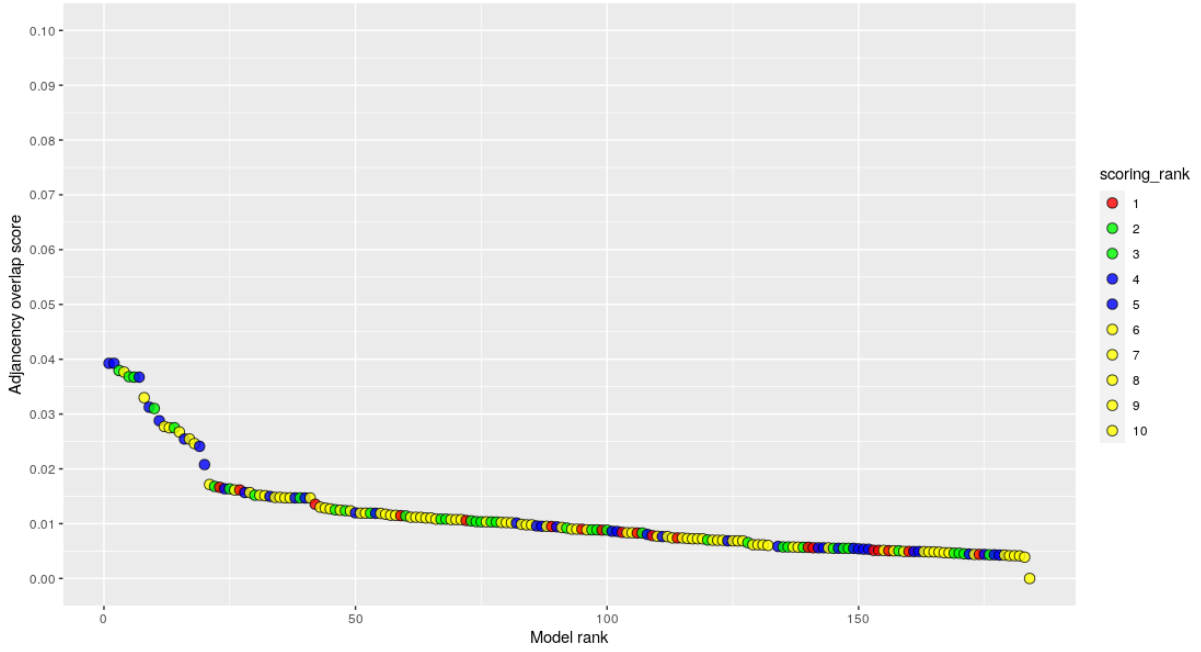


Figure 48. **Plot of the CAPRI COVID Round Target 181 models ranked by their adjacency overlap scores**

Colors corresponds the scorer ranks: Yellow = ranked 6th-10th, blue = ranked 4th-5th acceptable, green = ranked 2nd-3rd and red = ranked 1st

Target 182:

This target seems to have the best results regarding the adjacency overlap scores of the models. Indeed, the best model has a score of 0.030055 which is under the cut-off previously hypothesized at 0.04. The repartition of the model according to the scorer ranks is also very heterogeneous. In Figure 49, we can see that the first five models have the same overlapping score then it decreases slowly for almost forty models to a score of 0.022. Then the score declined sharply for a few models and, after, a linear depreciation to 0.003192.

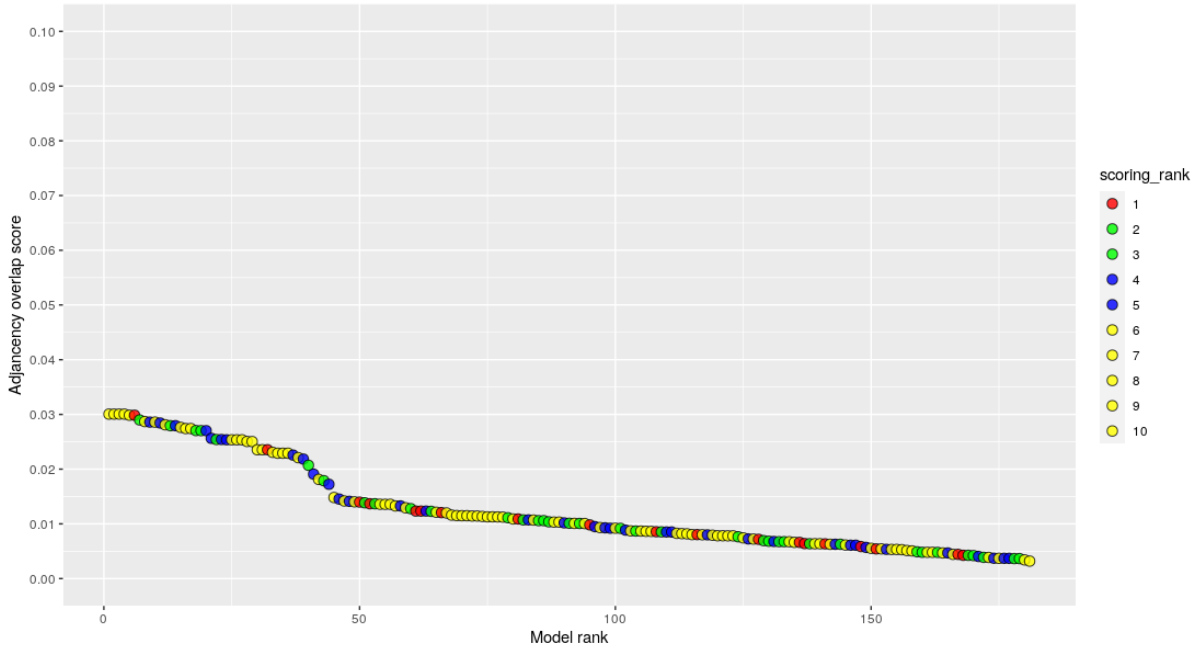


Figure 49. **Plot of the CAPRI COVID Round Target 182 models ranked by their adjacency overlap scores**

Target T183:

The results of this target are very close to the ones of T182. We can observe on Figure 50 a linear decrease of the adjacency overlap score from 0.027841 to 0.003194 in 180 models. One more time, no model has a score above the 0.04 cut-off and the quality of the model according to the scorers is heterogeneous.

Target 184:

This target shows the best results from the CAPRI COVID Round. Indeed, some models have an adjacency overlap score higher than 0.04. Precisely, four models are above this threshold with these scores: 0.044681, 0.041289, 0.040674 and 0.040039. Then as

usual, a linear decline of the model scores down to 0. The Figure 51 shows disparate scorer ranks all along.

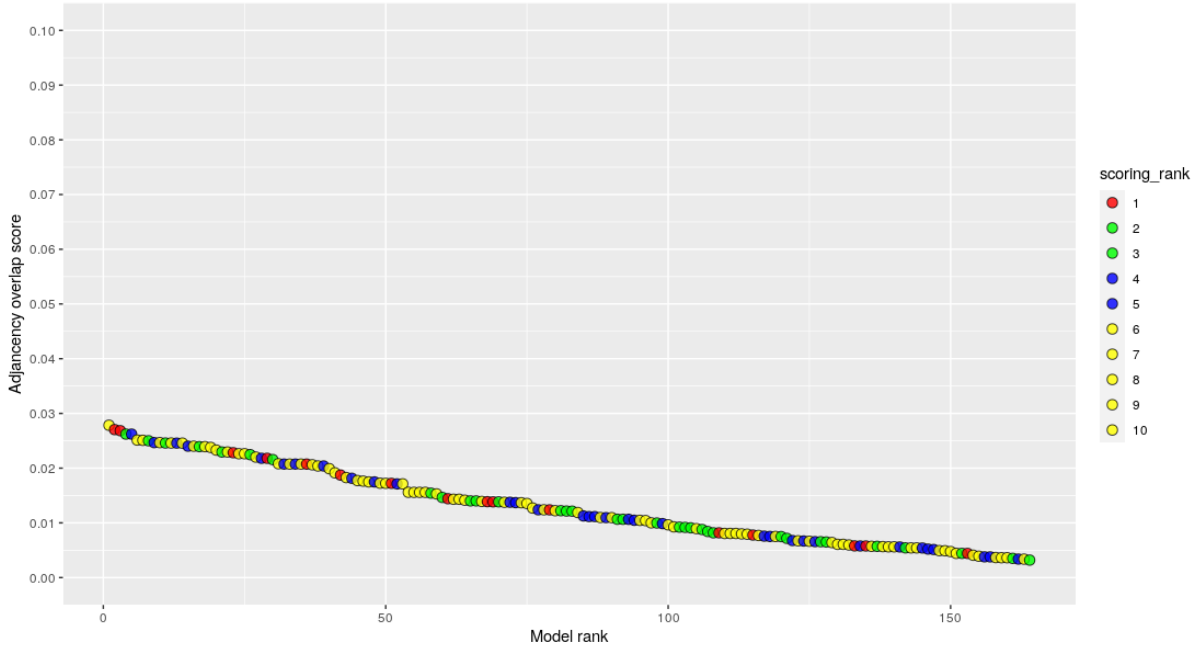


Figure 50. **Plot of the CAPRI COVID Round Target 183 models ranked by their adjacency overlap scores**

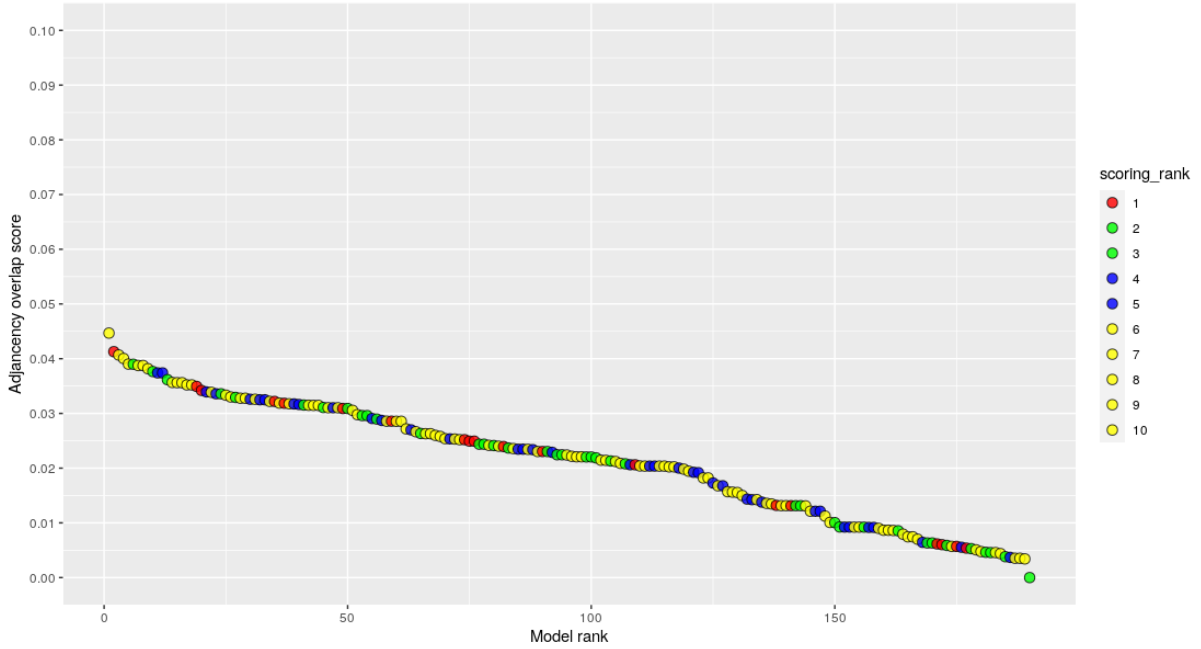


Figure 51. **Plot of the CAPRI COVID Round Target 184 models ranked by their adjacency overlap scores**

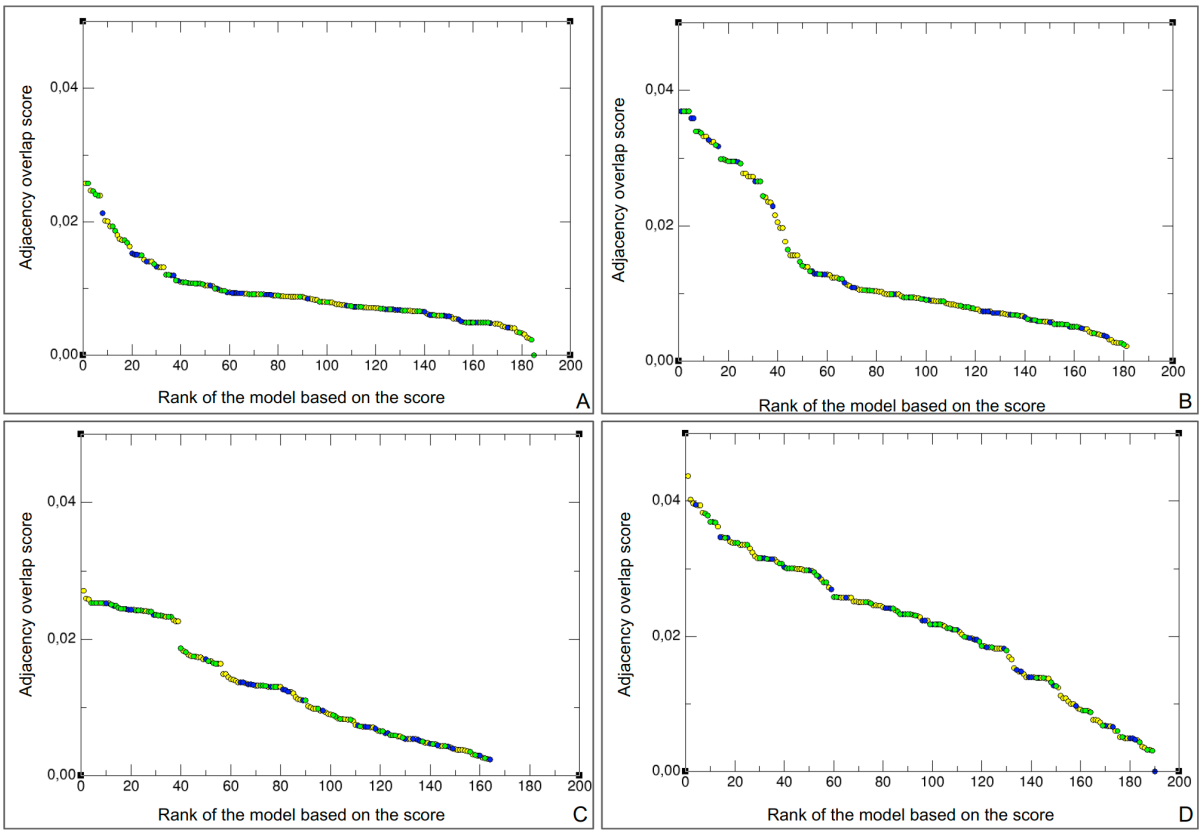


Figure 52. **Plot of CAPRI COVID Round models ranked with their adjacency overlap scores per targets**

A: T181; B: T182; C: T183; D: T184. Colors corresponds the scorer ranks: Yellow = ranked 6th-10th, blue = ranked 3th-5th acceptable and green= ranked 1st-2nd

As the adjacency overlap method shows good results on the validation set, it could be interesting to mix the methods and see if both can be combined to have better results. To that we filtered the meta clustering with the adjacency overlap (see Figure 53). Regarding the easy targets (T041, T050 and T053) we can see that a lot of clusters with incorrect models have been removed but also clusters with high quality models for T041 (Figure 53. B). The filter is interesting when only one meta cluster is available like for T053 (Figure 53. C): most of the clusters with only incorrect models have been rejected and the majority of the remaining clusters are at least acceptable. But, the filtering on a difficult target (T039) shows a removal of clusters reported in different meta-clusters inducing a support for no consensus.

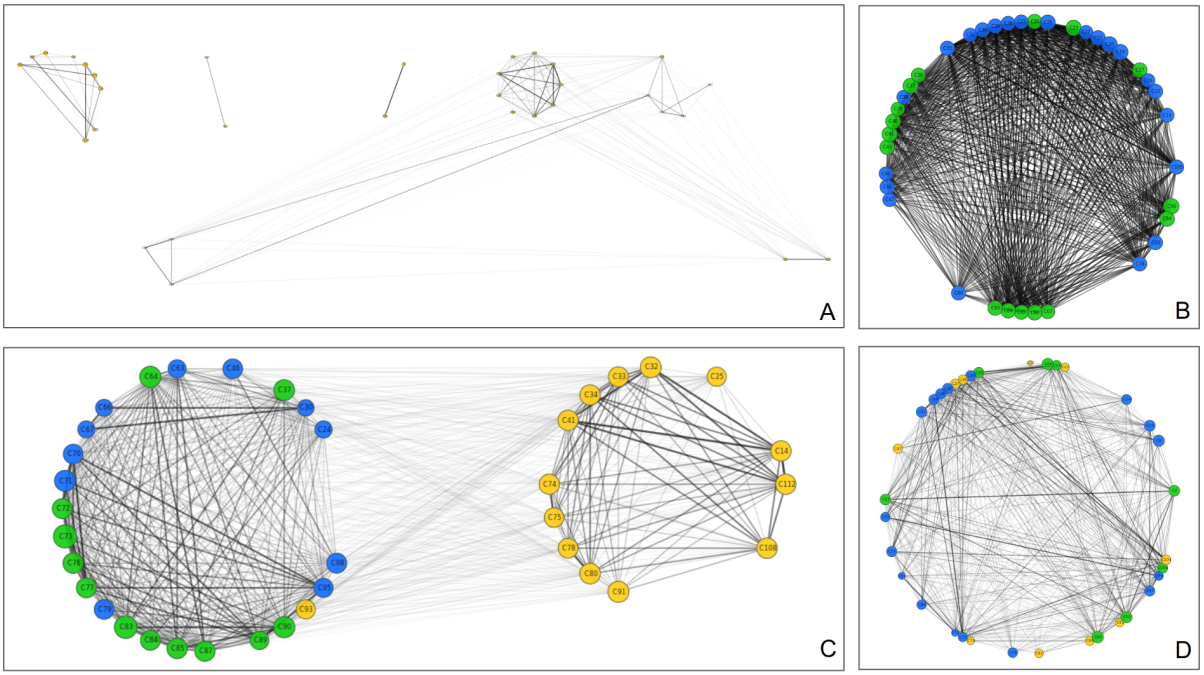


Figure 53. **Meta-clustering results filtered by the top tier models according to the adjacency overlap scoring**

A: T039; B: T041; C:T050; D: T053. Color of clusters correspond to the best model quality inside a cluster: Yellow = incorrect, blue = acceptable, green = medium quality and red = high quality

D. Discussion/Conclusion

Assessing the quality of a model without a solution is a veritable challenge. Today assessment methods have been developed to define model quality thanks to criteria developed by CAPRI and then summarized by DockQ. While these methods have proven successful, the scoring algorithms based on models themselves show different results depending on used features. But in most of the cases there is no solution to select the good model(s). The need to develop a good method has existed for a long time but has been raised by the pandemic crisis. Indeed, the interactions highlighted by Gordon *et al.* at the beginning of the 2020 led to an international effort to model a subset of these interactions [158]. But as the models were predicted, it was necessary to assess their quality. It has been shown in previous works that scoring algorithms are able to catch the majority of the good results and are efficient to rough out on. That is why we decided to analyze models selected by the CAPRI scorers. We then performed hierarchical clustering based on specific contacts on selected models to highlight consensual models like on the validation set. But the results

are not efficient as there are almost the same number of clusters as models. Maybe the threshold was too stringent but has been already used by Rodrigues *et al.* showing good results [200]. In addition, we removed model redundancy to only analyze different ones. But the fact that different scorers may have chosen the same models would have added confidence in such models. The meta-clustering we then performed shows profiles similar to Target 039, a difficult target without any good solutions. As the results did not highlight any consensus based on the specific contacts between chains, a method based on a whole profile determined by every model has been developed and applied. The adjacency overlap method shows very interesting results with the high ranking of good models. Indeed, when a validation set target has good models in its solution they are ranked before incorrect models. But the high quality solution (*i.e.* T041) are not well ranked compared to medium or acceptable models. This can be explained by the fact that there is a very low amount of high quality models and as this method is based on consensus these models are put aside because they have predicted specific contacts that other models do not have.

Residue conservation is good information to predict key areas for a protein, notably the interaction interface. But as we see in this study, viral proteins have a predicted interface with not conserved residues. This can be explained by the virus' facilities to mutate for an easier infection of different organisms. The non conserved residues could, unlike eukaryotes for example, be useful information for viral interface prediction.

Finally, according to the meta-clustering and adjacency overlap method it seems that the prediction of interactions between the viral and human proteins does not give good results. One possible reason could be that it is complicated to predict complexes between viral proteins and human ones because of their ability to quickly mutate. Another reason could stem from the prediction of interaction in the Gordon *et al.* study: indeed, the identification has been made through affinity-purification mass spectrometry but the specific interaction between the different components is not one hundred percent sure and the hypothesis of other proteins involved in the interaction is still present.

Regarding only the different adjacency overlap score plots for the validation set, we could put a threshold of 0.04 to discriminate good from bad models. Unfortunately, with this

cut-off, none of the CAPRI special Covid Round scoring models would be considered as correct models. However, bad incorrect models can still show more or less correct interfaces which is more likely the case for the human proteins in CAPRI COVID-19 targets which is still useful information. But as effective as it seems to be, this method has only been tested on 4 different complexes and it is not sufficient to claim it as an effective new scoring method for protein-protein complexes or to define a correct threshold.

V. Validation of the adjacency overlap method: Analyses of the CAPRI scoreset_2022

A. Introduction

1. Protein-protein interaction scoring methods

Being able to assess a protein-protein complex is a main research field since the improvement of protein-protein docking algorithms in the last two decades [105]. Protein-protein complex predictions being also referred to as protein-protein docking, can be divided into two categories: template-based and *de novo* methods [201,202]. Even if the number of methods and algorithms has grown a lot recently and there is a need to assess the quality of the different methods. To this, predictors developed their own scoring function to propose to users their best models. CAPRI highlighted the need for improvement of scoring as it is a key aspect in protein-protein modeling [149]. These scoring methods can be divided in different categories such as physics-based potentials, interface shape, knowledge-based statistical potentials, machine learning and deep learning methods and evolutionary profiles of interface residues. These categories are often mixed, which is the case of AccuRefiner, MaSIF and GNN-DOVE [105,106,108].

All these methods are trained on available data such as Dockground dataset 1.0 [203], a database containing many structures of complexes. This can add a bias if a particular interaction is not often found in the database. One of the most current and recent scoring methods is GNN-Dove which uses many features combined in a graph neural network. It shows on the 2014 CAPRI Score_set very good results, as did iScore developed by the Bonvin lab[109]. These two methods are based on RINs defining the interaction surface. From these graphs constructed on databases they trained their algorithms to rank a protein-protein complex. iScore also adds a physics-based potential such as Van der Waals, electrostatic and desolvation energies. These methods are trained on a set of experimentally resolved structures (Protein-protein docking benchmark version 4.0 for iScore).

In the previous part of this PhD manuscript, we have developed a new scoring method based on the knowledge of all CAPRI predictors and scorers. More precisely on their

consensus answers regarding the interface. As the two methods highlight before, it is based on the specific contacts at the interface, found thanks to RINs. As it showed good results on 4 previous Targets of CAPRI, we decided to test it on a larger dataset.

2. Protein-protein complexes databases

The majority of experimentally resolved protein-protein structures are available on the PDB. But some databases contain specific quaternary structures. This is the case for BACKGROUND, an online resource that provides various dataset of X-ray and modeled structures. This tool consists of 5 different subsets of protein-protein complexes: bound, unbound, models, docking decoys and docking templates [204]. In total, an amount of 215 363 pairwise complexes is available of the bound-bound subset with 149 416 homo-dimers. Two benchmarks of model complexes are also availables and consist of the first one of arrays of six models for each of the proteins (from 63 complexes). These models have been constructed by template modeling [105,109]. The second one which is larger is also composed of 6 models but on 165 protein complexes.

Another database is the protein-protein docking benchmark version 4.0 [205]. It consists of a set of non-redundant protein-protein complexes as a protein-protein benchmark. In total, this benchmark includes 176 cases with various difficulties (121 easy, 30 medium and 25 difficult complexes). As it is a docking benchmark, the models have to be created and then scored by a scoring method.

The CAPRI Community also provides a dataset called Score_set [149]. A first version published in 2014 provided 19,013 predicted models in total for 15 complexes (called targets). According to CAPRI assessment criteria about 11% of the models have acceptable or better quality. The CAPRI benchmark can be divided into three sets: P, U and S-set. The P-set, as described in the previous section, is the ten best models of a complex according to predictors as the U set a set of 100 models, including the ten best. From this U-set scorers were proposed to select the ten best models according to their method creating the S-Set. In addition the experimental solution is also provided so people can test their docking and/or scoring method on the different dataset. For each model many assessment measures are provided,such as model quality (according to CAPRI criteria), number of clashes, f_{nat} and

$f_{non\text{-}nat}$, i-rms, l-rms, dockQ score. This set is very useful to see the improvement of docking and scoring methods and is well used as a benchmark to compare method results.

Recently, a new version of the CAPRI Score_set now called CAPRI Scoreset v2022 has been made available on the scoreset.org website (and will be published soon). This new dataset incorporates all the publicly released CAPRI targets, increasing significantly the number of models from about 20,000 (for 15 targets) to more than 170,000 models (for 96 targets) which makes it the biggest benchmark available yet for scoring methods.

B. Material and Methods

1. CAPRI Scoreset v2022

The new CAPRI Scoreset V2022 (which will henceforth be called scoreset) has been used in this study to test our adjacency overlap scoring method. This scoreset is composed of 170,310 docking models. These models are made from 148 interfaces coming from 96 CAPRI targets. From these interfaces, 120 have been retrieved for this benchmark. In total, 121,209 models have been analyzed. These target interfaces can be mainly divided into four different kingdoms: Archaea, Bacteria, Viruses and Eukaryotes, but some are interfaces between different kingdoms (like between eukaryotic and bacteria proteins) or artificial (coming from protein design). The repartition of the models in the different kingdoms is shown in Table 15.

Kingdom	Number of interfaces	Number of models	Percentage of total set
Archaea	17	7,890	6.36 %
Bacteria	49	45,908	37.00 %
Eukaryotes	36	50,645	40.82 %
Viruses	6	7,906	6.37 %
Cross species	4	8,918	7.19 %
Other	2	2,800	2.26 %
Total	114	124,067	100 %

Table 15. **Distribution of the Scoreset 2022 Uploader models set in four kingdoms**

When a complex involves a particular stoichiometry such as A1B2 or AB:C for example, it is specified. In addition, two other sets have been created from the full set, one

for homodimers and one for heterodimers. They contained 56 and 48 interfaces respectively (complexes that involve more than 2 chains have not been taken into account).

For each kind of set (P,U and S) the number of interfaces is different. This can be explained by the tight timing that sometimes happens in CAPRI, usually due to impending publication of the target's associated manuscript. In that case the scoring round did not take place so there is no S-set for these targets. The missing U-set targets are because they are still not assessed for the moment. For every analysis of a set the number of interfaces will be indicated in the results section.

2. Scoring method

The adjacency overlap method that was tested on this benchmark is the same as the one developed for the previous manuscript session. To calculate the overlap, we calculate the square root of the sum of all specific interactions (m) multiplied by the value of this interaction in the adjacency matrix squared (*M*):

$$\sqrt{\sum_{i,j}(m_{ij} * M_{ij})^2}$$

where *m* is the contact matrix for one model and *i, j* the residues we are looking for in the interaction.

To apply this method developed in C language, it takes as input a multi pdb file containing all the models for an interface, but also the number of chains, the stoichiometry of the chains with the good chain IDs. In addition, for CAPRI's targets, as the quality of every model is provided, the method can take as input a text file with the model ID and its quality. To automate the process, a Python script has been written to look at the multi pdb file in the REMARK section, the number of the chain and their IDs to write the correct command line in a bash script. The REMARK line is the following:

"REMARK 1 NCHAIN 3 CHAIN A 305 CHAIN B 75 CHAIN C 104"

This means that there are three chains in the models: the chain A with 305 residues, the chain B with 75 residues and the chain C with 104 amino acids. The output of the script is a

text log file. So, to better handle the results, the Python script adds a few lines in the bash script to only retrieve the scoring results in a .txt and a .csv file that can then be analyzed.

3. Efficiency of ranking method

To improve the next part comprehension, the definition and formula of statistical terms are retrieved in the following list:

- TP: True Positive (being correctly predicted as positive)
- FP: False Positive: data which has been predicted to be positive but is negative
- TN: True Negative: Correctly predicted as negative
- FN: False Negative: Predicted as negative while being positive
- Sensitivity (or Recall or True Positive Rate (TPR): It is the rate of positive data correctly predicted as positive.

$$TPR = \frac{TP}{TP+FN}$$

- Specificity (or True Negative Rate (TNR)) : it is the rate of negative data correctly predicted as negative.

$$TNR = \frac{TN}{TN+FP}$$

- Precision (or Positive Predictive Value (PPV)): Corresponds to the ratio of positive data correctly predicted on the total amount of data predicted as positive. Basically it gives the percentage of chance to a positive predictive data to be really positive.

$$PPV = \frac{TP}{TP+FP}$$

- False Discovery Rate: Corresponds to the percent of chance to be wrong when predicting a positive result.

$$FDR = 1 - PPV = \frac{FP}{FP+TN}$$

- Accuracy (ACC): It is the rate to be correct whether it is positive or negative.

$$ACC = \frac{TP+TN}{TP+FP+TN+FN}$$

To be able to calculate the efficiency of our method, the Receiver Operating Characteristic (ROC) curves have been plotted for every subset to see the performance regarding the sensitivity (also called True Positive Rate (TPR)) and the False Positive Rate (FPR) which is 1 − specificity. This curve allows us to calculate the Area Under Curve (AUC), which is a good indicator to see if the method can correctly discriminate data into two

categories. In parallel, a Precision-Recall curve has been drawn to evaluate the threshold to obtain the highest precision in function of the sensitivity. To determine this threshold, the wanted precision is selected and looked at the depreciation of it while increasing the sensitivity.

C. Results

1. U-set

The Adjacency overlap has been run on the full U-set containing 114 interfaces. In total 121,586 models have been used to test our method. To calculate the ROC and the Precision-Recall curves, we determined that a correct model has an assessment quality of acceptable or better. These two curves can be seen in Figure 54. With an AUC of 0.935, our method seems very effective to predict good models. Here the final idea is to retrieve only the good models, meaning that the precision of the method is more important than the sensitivity. So from the precision-recall curve we can say that we wanted to select to the highest precision. According to this analysis, the highest precision is 0.8189 for a sensitivity of 0.2986577. To obtain these statistical measurements, the threshold must be 0.080531.

This set can be separated according to kingdoms as explained in the Material and Method section. We decide to do the same analysis for the four main kingdoms (Archaea, Bacteria, Eukaryotes and Viruses). The results of the ROC AUC, precision, sensitivity and associated thresholds are listed in Table 16. Some of the interfaces have no kingdom information (two for the U-set). Four interfaces are complexes between two different kingdoms (here called “Cross”).

Regarding these table results, we can see that our method is kingdom dependent. The precision is very high when looking at Bacteria and Eukaryotes while very low for Archaea or crossed kingdoms. For Viruses, the precision is very high but with a very low sensitivity as it can be seen on the corresponding supplementary Figure in supplementary data. Bacteria and Eukaryotes sets have better results but also a higher number of models.

Kingdom	Interfaces	Models	AUC	Precision	Sensitivity	Threshold
Archaea	16	5,858	0.775	0.1256	0.624	0.0520
Bacteria	49	45,908	0.943	0.9276	0.2536	0.0811
Eukaryotes	36	50,645	0.880	0.9775	0.1342	0.0919
Viruses	6	7,906	0.782	1	0.0152	0.0764
Cross	4	8,918	0.717	0.0177	0.8429	0.0121

Table 16. **Adjacency overlap score in function of the complex kingdoms for the U-Set**

"Cross" corresponds to the interface between two different kingdoms

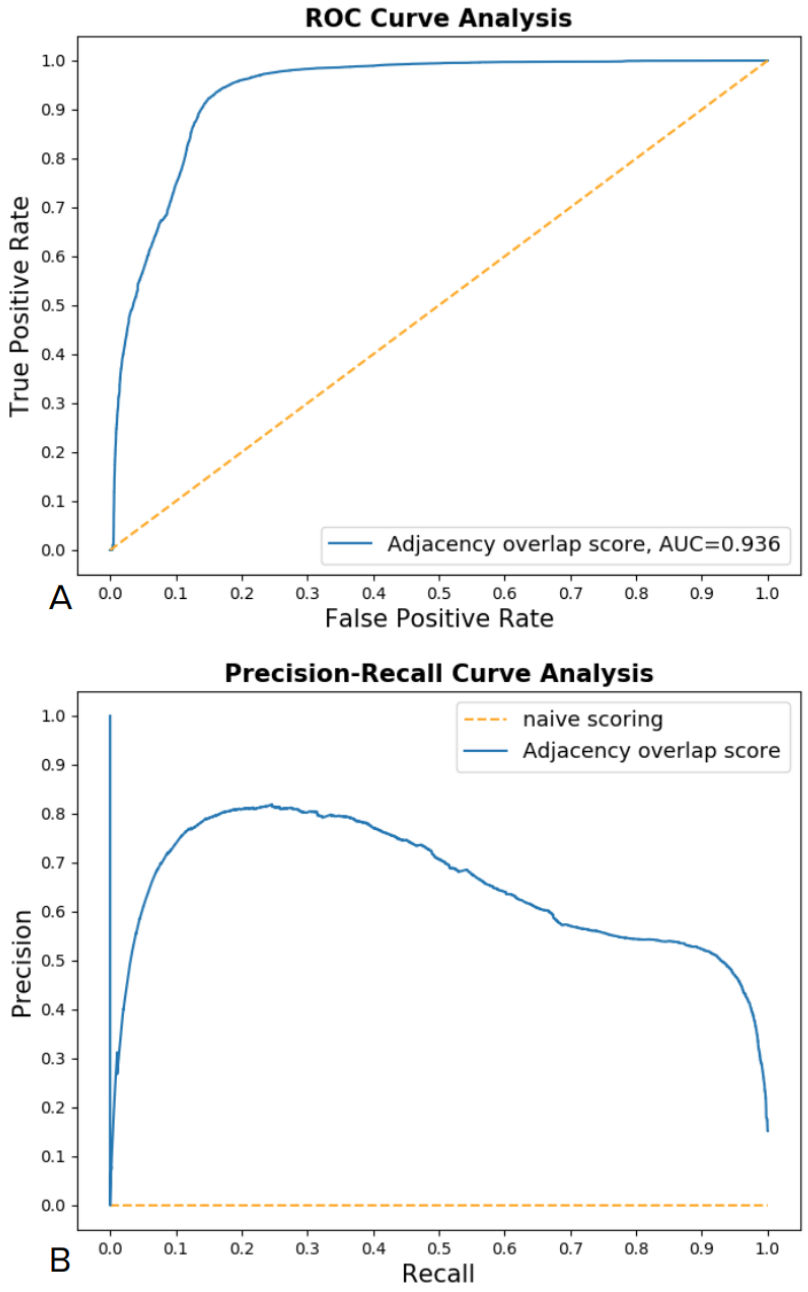


Figure 54. **ROC and Precision-Recall curves of the adjacency overlap scoring method on the U-Set**

Naive scoring corresponds to the theoretical results if we randomly score as a positive model.

We also looked at the difference of results regarding the dimer and more precisely between homo and hetero dimer. The results can be found in the following Table 17.

Dimer type	Interfaces	Models	AUC	Precision	Sensitivity	Threshold
Homo	51	54,987	0.953	0.9893	0.1849	0.0877
Hetero	48	54,929	0.937	0.6565	0.3673	0.0774

Table 17**. Adjacency overlap score in function of the type of the U-set dimer complexes**

These results show that our method performs well both on hetero or homo-dimer. However, for homodimers the precision is higher but with a lower sensitivity.

2. S-Set

The S-set is one corresponding to the validation set used for the CAPRI-Covid Round but with a bigger amount of interfaces. In total, 113 interfaces have been assessed by the CAPRI assessment team for this Set for a total of 15,799 models.

This S-set is composed of fewer models than the U-set but should contain a better percentage of good quality models as shown in previous CAPRI publications [142,146,149]. As for the U-Set, our adjacency overlap method has been performed on the whole dataset and the graphical results are shown in Figure 55. The AUC of 0.880 shows a good capability for our method to distinguish good from bad models also for smaller sets. With a precision of 0.7898 and a sensitivity of 0.2612, the results on this dataset are slightly lower than expected. To obtain these statistical measurements, the threshold is set at 0.0718.

As for the U-set, the different kingdoms have been analyzed separately to see how well our method performs. All the results are summarized into Table 18. Regarding these results we can see the tendency of lower quality results is found in every kingdom except for the Viruses where there is an increase of about 0.11 point.

Again the AUC shows good results for Bacteria and Eukaryotes and poor results for Archaea and Crossed kingdoms.

Kingdom	Interfaces	Models	AUC	Precision	Sensitivity	Threshold
Archaea	7	827	0.673	0.2382	0.9785	0.0220
Bacteria	53	7,172	0.873	0.7963	0.3613	0.1717
Eukaryotes	40	5,762	0.834	0.9247	0.1626	0.0755
Viruses	11	1,778	0.891	1	0.1027	0.0706
Cross	4	650	0.670	0.0413	1	0.0136

Table 18. **Adjacency overlap score in function of the complex kingdoms for the S-Set**

"Cross" corresponds to the interface between two different kingdoms

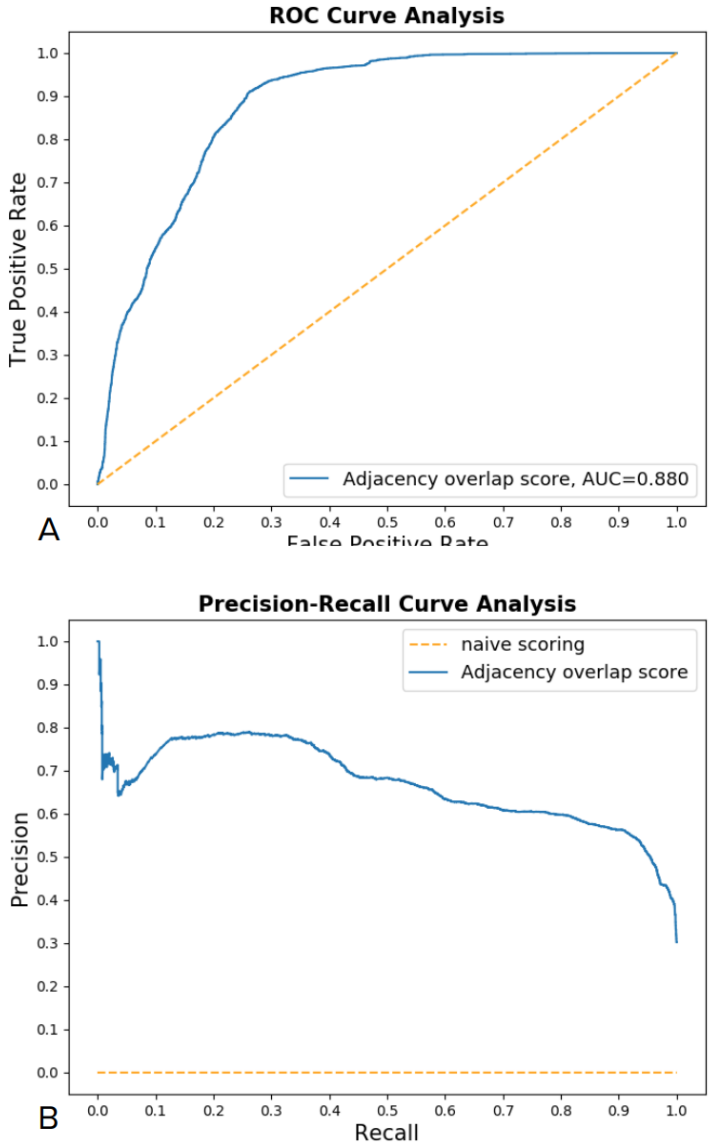


Figure 55. **ROC and Precision-Recall curves of the adjacency overlap scoring method on the S-Set**

Dimer type	Interfaces	Models	AUC	Precision	Sensitivity	Threshold
Homo	55	7,933	0.877	0.8980	0.0939	0.0818
Hetero	48	6,716	0.916	0.7447	0.4407	0.07554

Table 19**. Adjacency overlap score in function of the type of the S-set dimer complexes**

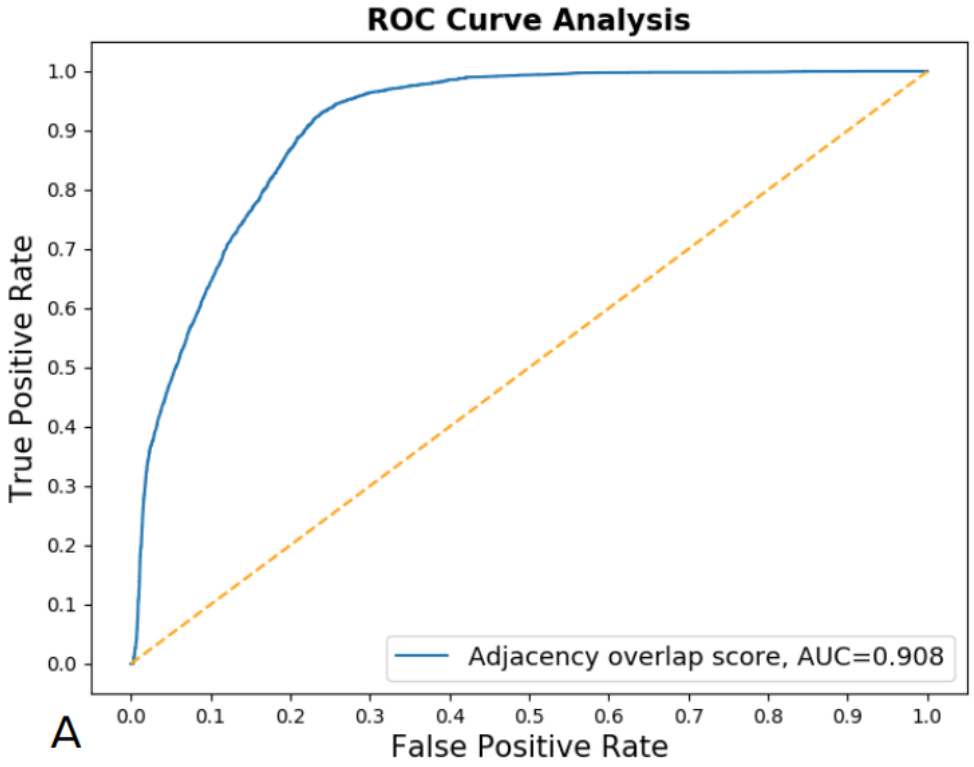
Regarding the difference between the homo and hetero dimer (Table 19) we can see once again a loss of quality compared to the U-Set by comparing the ROC AUC. But for hetero-dimers we notice a better quality regarding the precision and sensitivity.

3. P-Set

As U and S-sets gave different results, we performed our adjacency overlap method on the P-set, the set composed of the 10 best models of each predictor according to their own algorithms. This contains the highest amount of interface compared to the two other sets with a total of 148. As before, the ROC and Precision-Recall curves have been drawn and can be seen in Figure 56. The ROC AUC for this whole set is 0.908 which is higher than for S-Set but lower than U-set. The difference could be explained by the difference of the number of models. The precision is 0.7967 for a sensitivity of 0.2775. The values are for a threshold of 0.075081. These values are very close to the one of the S-set. Once again, these measures have been calculated for every kingdom and retrieved in the Table 20.

Kingdom	Interfaces	Models	AUC	Precision	Sensitivity	Threshold
Archaea	16	4,588	0.756	0.2006	0.8654	0.0439
Bacteria	58	14,306	0.938	0.9231	0.2271	0.0840
Eukaryotes	58	17,292	0.853	0.8547	0.0522	0.1112
Viruses	6	3,273	0.840	0.9959	0.4255	0.0454
Cross	7	1,860	0.964	0.9302	0.3008	0.0719

Table 20. **Adjacency overlap score in function of the complex kingdoms for the P-Set**



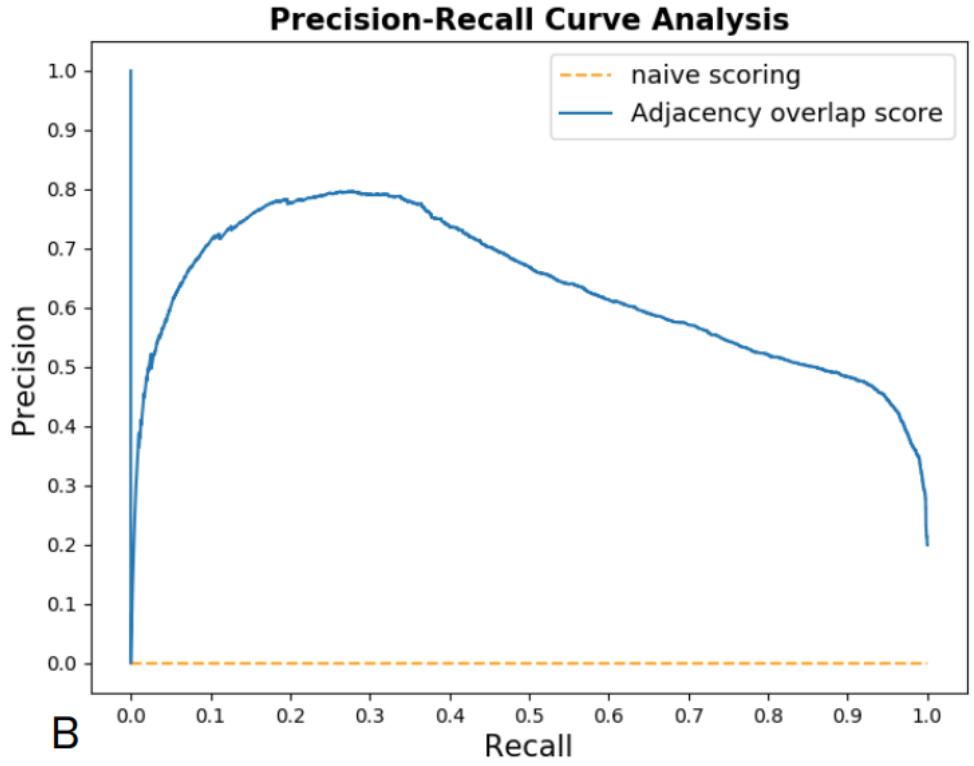


Figure 56. **ROC and Precision-Recall curves of the adjacency overlap scoring method on the P-Set**

An interesting result is regarding the Viruses results for the P-set. The results show a very good precision with a higher sensitivity than usual while the ROC AUC shows lower results than other kingdoms. Then the difference between homo and hetero dimers have been analyzed and are summarized in Table 21:

Dimer type	Interfaces	Models	AUC	Precision	Sensitivity	Threshold
Homo	56	13,635	0.937	0.95118	0.1222	0.0877
Hetero	75	21,934	0.922	0.7257	0.4018	0.0748

Table 21**. Adjacency overlap score in function of the type of the P-set dimer complexes**

Here again results are better for the P-set than the S-set but lower than U-set regarding the AUC. But for the precision of our method on the P-set, we can see that it is easier to retrieve good results without so many false positives, which is more important according to us.

4. CAPRI-Covid

According to the results of the adjacency overlap on the CAPRI Scoreset 2022, our method is able to discriminate good from bad models with a high precision but with a low sensitivity. We also were able to determine a threshold for this method according to the different kingdoms. So we decided to look back at our results to see if a model from the different T181, T182, T183 and T184 is a good model according to the adjacency overlap. To this, the results score of the best model for each target is retrieved in the Table 22 for and compared to the different threshold obtained earlier in the S-set.

Target ID	Best model score	Threshold - Viruses	Threshold - Eukaryotes	Threshold - Cross kingdom
T181	0.0393	0.0706	0.0755	0.0136
T182	0.0301	0.0706	0.0755	0.0136
T183	0.0278	0.0706	0.0755	0.0136
T184	0.0447	0.0706	0.0755	0.0136

Table 22. **Best model scores for the different S-Sets of the CAPRI Covid Round compared to the threshold for the different kinds of kingdoms**

As we can see on these results no models are above the defined threshold except for the Cross-kingdom one. As this latter is very low and induces a precision about 0.04, the probability that a model is actually correct is very low. Regarding the other threshold , we can say that no model has more than 92% to be correct. But it could be interesting to have the percentage of chance for each model to be correct. In this perspective, from the different Precision-Recall curves, the precision corresponding to the closest threshold result has been retrieved and all the results have been written in the Table 23.

Target ID	Best model score	Precision - Viruses	Precision - Eukaryotes	Precision - Cross kingdom
T181	0.0393	0.7101	0.4463	0
T182	0.0301	0.5815	0.4446	0
T183	0.0278	0.5660	0.4384	0.0222
T184	0.0447	0.8216	0.4782	0

Table 23. **Precision of the best models according to their adjacency overlap scores through the different kingdoms**

These results show us the probability for a model to be correct according to the different kingdoms. If we consider these models as interaction between viral proteins, all the good models for each target have more than half chance to be correct, even more for T181 and T184 where the best models have a probability to be a good model of 71% and 82% respectively. But if we consider them like eukaryotes interactions the results are lower with no model with at least 50% of chance to be correct which is even worse if we consider the model as cross kingdom interaction which is actually the case. A strange result is the better precision for the T182 best models compared to the others targets. Indeed, with the lowest adjacency overlap score we could naturally think it would have the lowest precision. This result can be explained with the precision-recall and ROC curves for this kingdom. The precision is very low and regarding the ROC curve, we can see first that there is more false positive than true. As there are few positive models inside this set, the values of the precision can have a quick increase and then a decrease as there are bad models.

5. Comparison with other scoring methods

As our method seems to have a good regression capacity regarding the AUC, it could be interesting to compare our results to the actual scoring methods as iScore and GNN-DOVE which show good results [105,109]. In Wang *et al.* (2021) GNN-DOVE is compared to iScore on previous CAPRI Targets. These 13 targets have been selected and we compared our ten best results obtained with the Adjacency Overlap (OA) on these to the iScore and GNN-DOVE ones. The results are gathered in Table 24.

Target ID	iScore	GNN-DOVE	AO - U-set	AO - P-set
T29	4	2	**10/8****	10/6**
T30	0	**1**	0	0
T32	4/1**	0	0	**8/6***/2****
T35	0	**1**	0	0
T37	4/2**	0	**4/4****	3/2**
T39	0	0	0	0
T40	4/1***	4/4***	10/6***/3**	**10/8***/2****
T41	10/2**	5	10/6**	**10/7****
T46	4	1	0	**9**
T47	10/6***/4**	9/4***/5**	**10/8***/2****	10/2***/8**
T50	**4/3****	6	0	4/2**
T53	5/1**	2	2	**8/2****
T54	0	0	0	0
Total	9/2***/5**	9/2***/1**	6/2***/3**	**9/3***/5****

Table 24. **Comparison of scoring method on the CAPRI Scoring dataset**

The values are in bold when the performance is higher than other scoring methods. The scoring performance for each target is reported as the number of acceptable or better models (hits), followed by the number of high (indicated with ***) or medium quality models (**). The total is reported in a similar way: first number corresponds to the number of targets in which a method found at least an acceptable or better quality model, the second corresponds to the number of targets in which at least a high quality model have been found and the third number for the medium quality.

AO is for Adjacency Overlap

These results show how successful our method is to score models. Indeed, even if the improvement is marginal target-wise, it offers an enrichment of the good models in the top 10 models compared to two very recent scoring methods. More specifically, our adjacency overlap method finds good models in the same targets as iScore but finds higher quality models or at least a higher number of models. GNN-DOVE was able to find a good model in two targets that neither one of our methods nor iScore were able to find. Our method better performs on the P-Set than the U-Set. As it is based on the global consensus, the U-set does have more incorrect model proportions than the P-set.

Our method was also tested on the S-set but the results have not been summarized in the Table as S-set was not tested by iScore and GNN-DOVE. The total results for adjacency overlap is 8/4***/3** which is better than the U-set results. There is one less target found by the other scoring methods but a higher number of high quality models.

D. Discussion

In this study, we tested a new scoring method based on the adjacency overlap. We were able to test it on a full new dataset, ideal for a benchmark with at least 110 different complexes with different stoichiometries. From this benchmark, three sets have been analyzed with different numbers of models. The U-set which contains up to 100 models per predictor, the P-set which is the 10 best model per predictor and the S-set, a specific set with the 10 best models of the U-set selected by scorers. Regarding the U-set, we can see that our method did not perform well for two kingdoms: Archaea and Cross-species. These two subsets have a very low amount of interfaces and models. The Archaea subset is composed of 17 from 4 different targets meaning there are a lot of interfaces for the 3 of the 4 complexes (5 interfaces for T149, T150 and T151). In total, 90.9% of the models of Archaea complexes were incorrect while 4.9% were acceptable, 3% with medium quality and 1.2% with high quality. The knowledge of Archaea protein complexes is very low compared to bacteria or eukaryotes. In PDBe, there are only about 1,300 complexes for Archaea protein interacting with other macromolecules (against 27,740 for Eukaryotes and 9,360 for Bacteria). The low amount of experimental results may have an impact on the predictive model algorithms and quality. This is similar for complexes with different kingdoms involved.

Regarding the homo and hetero-dimers, we can see a better prediction for homo-dimer than hetero-dimers despite a similar number of models. This could be explained by better interaction interface recognition.

The S-set has been then used to test our method. We hypothesized to achieve better results on this set than the previous set. Indeed, our method is based on the overall quality of the models. As previously shown in a CAPRI prediction, the S-set is the one from three available to show the best results [149]. But even if the tendencies are similar to the U-set regarding the kingdoms, there is an overall lower discrimination capacity according to the AUC (a loss of 0.05 to 0.1) except for Viruses where we can see an increase of the AUC of 0.1. This could be explained by the lower amount of models between the U and the S-set.

According to the previous results from the U and S-sets, the results should be under the one from the S-set as there is a lower number model than the U-set and the selected models are not from scorer but predictor scoring. But regarding the global AUC, we can see that our method performs better on the P-set than the S-Set. Regarding the different kingdoms, we can see that the prediction works well for all the kingdoms except for Archaea. The poor results for this kingdom can be explained by the same reason as the U-Set. The better prediction capacity for the P-set than the S-set could come from the higher number of models in the P-set because there are more predictors than scorers which is contrasting with the previous observation. There is a need to determine the number of models for this method to be optimized.

The comparison of our scoring method to the two methods iScore and GNN-DOVE shows better results for our method based on the P-set than the U-set. It could be explained by the fact that there is a pre-selection by the predictors which removes models which may lead to a wrong consensus. Indeed, our method is based on the overall consensus of a number of models. If the majority of models are wrong our method would not be able to retrieve the good models. Contrary to the reference scoring methods, ours is not based on training on experimental data but only uses the available models. Our method has the particularity to consider every model and so every information with the same level of importance, it therefore creates a scoring based on mixed knowledge and algorithms. But a

big constraint is that our method needs a high number of models provided by different predictors. It could be used as a meta clustering, selecting a consensus model from different docking algorithm results.

The benchmark set used by iScore and GNN-DOVE is old and scored models were predicted by old algorithms. It could be interesting to compare results on more recent data that should be available with the release of the CAPRI Scoreset v20200.

Regarding the special CAPRI COVID Round, the predictor could have produced good models if we based our analysis on the Eukaryotes threshold defined on the S-set. Indeed some models have between 70 and 74% of chance to be correct according to this threshold. But these special COVID-19 targets are interaction between human and viral proteins which are cross-kingdom and, based on the S-set cross-kingdom result, there are no good results. Of course, our method may not be able to recognize good models for cross-kingdom but there is also the possibility that the interactions between the proteins in the target are not that binary. Indeed, some other protein may be involved in the complex. Or the prediction that an interaction exists between the proteins may be wrong. As our method shows poorer results on the S-set, it could be interesting to redo this analysis on the P-set and the U-set. There is one CAPRI Target (T165) which involves in its interaction a viral and a human protein. But the number of correct models is very low (4 acceptable models for the U-Set, 1 for the P-set and none for the S-set). This shows the difficulty to predict interaction between eukaryotes and viral proteins. This difficulty could be explained by the higher rate of mutation for viral proteins.

In perspective, we could also determine the ideal number of models for our method to be applied with the highest confidence. An article with these results is in preparation and will be published as soon as possible.

E. Conclusion

We developed an adjacency overlap scoring method based on the overall consensus of a high number of models. This method shows better results than two reference scoring

algorithms when applied to the CAPRI 2014 score_set benchmark. But as we highlight its good prediction rate for bacteria, eukaryotes and viral complexes, we also demonstrate its poor capacity to find good models for Archaea and across kingdoms. As archaea complexes were not yet available as a benchmark it could be interesting to look at the other scoring results regarding this kingdom to compare our method to the others.

VI. Modeling the specific interaction of β-catenin and *O*-GlcNAc Transferase

A. Introduction

1. *O*-GlcNAcylation

a) Pathway

The *O*-GlcNAcylation is a dynamical post-translational modification regulated by the Hexosamine Biosynthesis Pathway (HBP). This pathway is supplied by different nutrients: monosaccharides, glycogen, glutamine, fatty acids, lipids and nucleotides. This will lead to the production of the nutrient sensor, substrate of the *O*-GlcNAc Transferase, called UDP-GlcNAc (UDP-*N*-acetylglucosamine). The glucose is transformed into Glc-6-P by the HexoKinase (HK) which is then transformed into Fru-6-P by the PhosphoGlucose Isomerase. This sugar will be irreversibly transformed into $GlcNH_2$-6-P by the Gln:Fru-6-P AmidoTransferase. Then the metabolism of fatty acid brings the Acetyl-CoA in HBP to produce GlcNAc-6-P from the GlcNH2-6-P thanks to the GlcN-6-P N-AcetylTransferase. From this, the PGM (PhosphoGlucoMutase) transforms the product into GlcN-1-P. Finally, the GlcN-1-P is activated into the UDP-GlcNAc thanks to the UDP-N-Acetylglucosamine Pyrophosphorylase. This UDP-GlcNAc is the first molecule recruited by the OGT, the enzyme responsible for the addition of the GlcNAc on the substrate, then the polypeptide is fixed in a bi-bi mechanism [30]. The sugar addition reaction is made thanks to a nucleophilic attack from the hydroxyl group of the serine or threonine on the anomeric carbon of the sugar. This leads to a β-glycosidic link. Once it is done the remaining nucleotide and the *O*-GlcNAcylated polypeptid are released.

b) The *O*-GlcNAc Transferase

The *O*-GlcNAc Transferase (OGT) is a GlycosylTransferase (GT) in GT41 family according to CAZy (Carbohydrate-Active enZymes) database [206]. It is composed of two main parts called domains: the catalytic domain and the TPR (TetratricoPeptide Repeats) domain linked by an intermediate region called linker [207]. The catalytic domain can be divided into two domains called Catalytic Domain I and Catalytic Domain II (CDI and CDII). These two sub domains are linked by a region called Intermediate Domain (Int-D). This region's function is

still unknown. There are three different isoforms of OGT with a similar catalytic domain but different size of TPR domain in function of the number of TPR: the ncOGT for nuclear and cytoplasmic OGT, the sOGT for small OGT and mOGT for mitochondrial OGT. Their number of TPR is respectively 13.5, 3.5 and 9.5. The ncOGT owns a NLS sequence responsible for the interaction with the importin α5 protein. This sequence is found in the 14th TPR of the ncOGT [208]. This TPR domain known to be the substrate recognition domain also possesses an asparagine ladder. This ladder has been shown to play an important role for the catalytic activity with a loss of the OGT activity when the asparagines are mutated into alanines [37].

OGT is also known to have a protein cleavage activity. Indeed, in 2013, Lazarus *et al.* describe the cleavage of the Host cell factor-1 (HCF-1), a co-regulator of the human cell-cycle [209]. This cleavage is made thanks to the UDP-GlcNAc. The region where this process is called HCF-1rep1 and contains the first HCF-1PRO repeat plus N-terminal HCF-1 sequences containing several *O*-GlcNAc sites [39]. The need of a glutamate on the substrate and the UDP-GlcNAc has been proved to cleave the HCF-1.

Today there are 38 experimentally resolved structures of the OGT or parts of it in the PDBe. The first structure is the OGT TPR and was published in 2004. It then took more than six years to obtain a structure of the OGT catalytic domain with the first TPR repeats. Indeed, 3 structures (PDB IDs: 3PE3, 3PE4 and 3TAX) of the OGT have been published in early 2011 [30]. These structures have the particularity to be complexes of OGT with a peptide substrate. Until the end of 2021 there was no full OGT structure available but with the improvement of experimental methods, notably, cryo-EM it is now longer the case [210]. But the resolution is not as high as for X-ray microscopy with a resolution of 5.32 Å.

c) The asparagine ladder

In 2018, Levine *et al.* showed the impact of the asparagine ladder of the TPR of the OGT by mutating them into alanines [37]. They performed a permutation test approach to determine which proteins were glycosylated by wild-type and 5N5A OGT variants. According to their results (Figure 57), 739 proteins scored as hits when comparing wild-type OGT-treated arrays to controls and from these proteins 736 had a higher signal for wild-type

OGT than 5N5A. Based on these results they concluded that the asparagine ladder is used by the OGT to recognize its substrate.

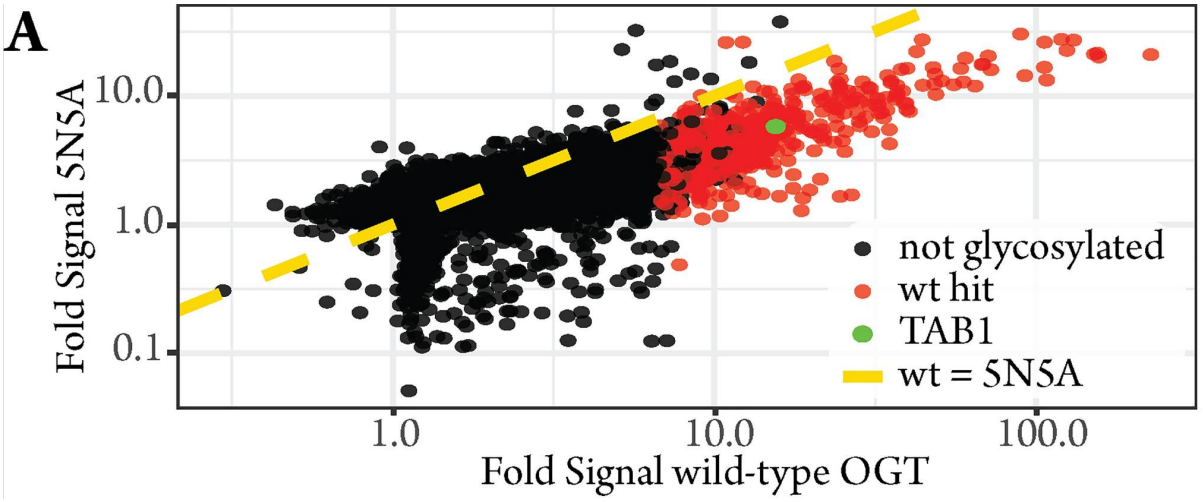


Figure 57. **The asparagine ladder in the TPR lumen is critical for recognition of OGT substrates**

Fold signal above control for every protein for wild-type OGT (x-axis) and 5N5A (y-axis) based upon median normalized data. The dashed line represents equivalent activity between wild-type and 5N5A enzymes. Red circles are hits for wild-type enzyme; black circles are proteins that do not score as glycosylated. TAB1, a known poor 5N5A substrate, is high-lighted in green. From Levine *et al.* (2018) [37]

2. Wnt pathway: Crossplay between *O*-GlcNAcylation and phosphorylation

The Wnt pathway is one of the most important signaling pathways. It has been shown to be crucial in development and growth and its complexity has a major role in it [211,212]. This wnt pathway is the most active during the embryogenesis where its role is to facilitate the cell differentiation, polarization and migration. At the mature age of an organism, this pathway should be knocked out. But it has been shown to be activated during the development of tumors and other diseases. It also can be reactivated in organ injury and regeneration such as kidney injury [213]. The wnt pathway can be categorized as canonical, that is β-catenin dependent and as non-canonical, that is β-catenin-independent signaling pathways [212]. The β-catenin is a protein which is a transcription factor inducing the activation of T cell factor (TCF). Its activation is gene dependent and will happen when the β-catenin is translocated in the nucleus. In a normal state, the β-catenin is usually degraded by a destruction complex. This complex is composed of different kinases: glycogen synthase kinase 3β (GSK3β) and casein kinase 1 α (CK1α). These two kinases are interacting with

adenomatous polyposis coli (APC) and axin. This big protein complexes phosphorylates a part of the N-terminal segment of the β-catenin called the Destruction box (D-box). Its phosphorylation will allow its recognition by an ubiquitin ligase β-transducin repeat containing protein (β-TrCp), the ubiquitination of the β-catenin will lead to the degradation of this protein by the proteasome [214]. The Figure 58 A summarizes this information. The natural degradation of the β-catenin can be countered by the competition between *O*-GlcNAcylation and phosphorylation [215]. The *O*-GlcNAcylation has been shown to modify several residues of the β-catenin D-box such as serine 23, threonine 40, threonine 41 and threonine 112. The competition between phosphorylation and *O*-GlcNacylation on threonine 41 avoids its phosphorylation by the kinases and blocks its degradation. The β-catenin will accumulate into the cytosol leading to its transfer into the nucleus activating the transcription of cell cycle genes thanks to the TCF activation (Figure 58 B).

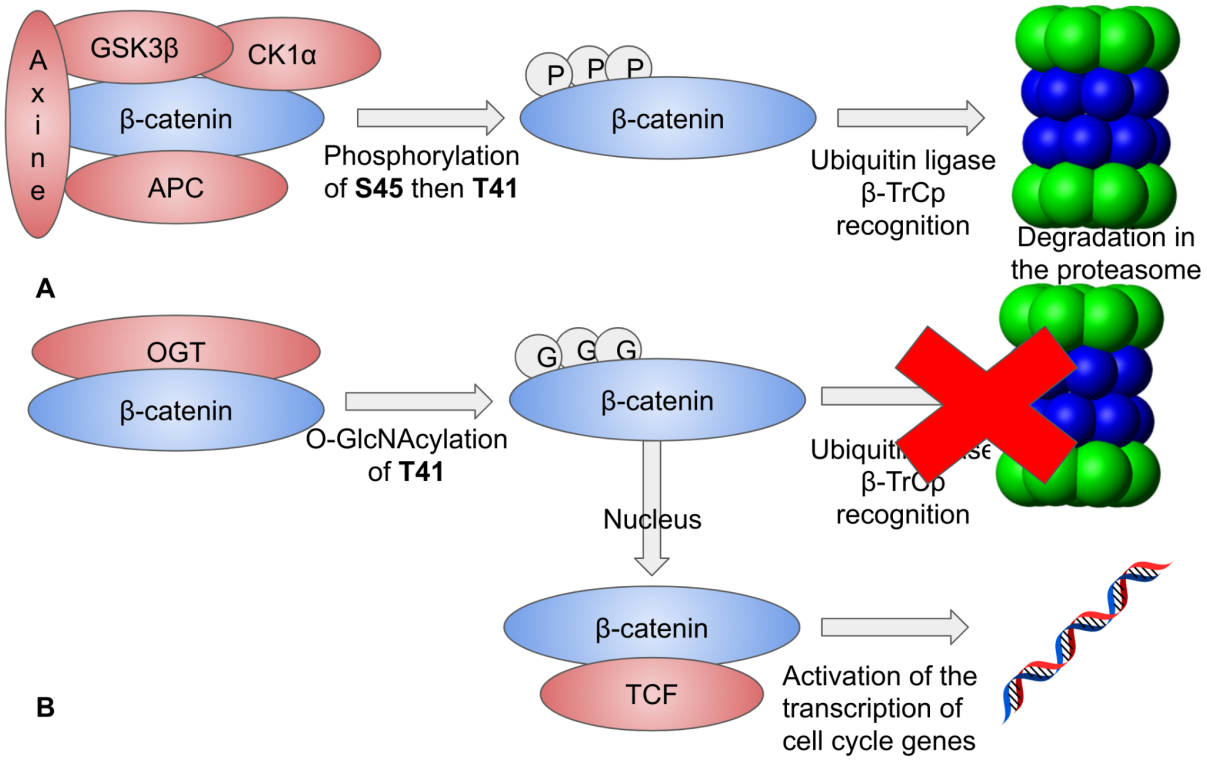


Figure 58. **Schematic representation of the cross talk between phosphorylation and *O*-GlcNAcylation of the β-catenin and its impact on its degradation**

A: the β-catenin phosphorylation leads to its degradation into the proteasome

B: the β-catenin *O*-GlcNAcylation on threonine 41avoid its degradation and leads to the activation of the transcription of cell cycle genes.

The specific interaction between the OGT and the β-catenin is known to increase the risk of colorectal cancer (CRC). Indeed, 90% of CRC cases are an alteration of the Wnt/β-catenin. If the majority of these cases are genetic alteration of the APC, the non degradation of the β-catenin because of its *O*-GlcNAcylation is also implicated. As only the N-terminal segment of the β-catenin is known to be modified by the OGT, we still don't know how this specific interaction happens and filling this lack of knowledge may bring therapeutic treatments.

The beta-catenin is composed of an unstructured N-terminal segment where the D-box is found. Then there is an Armadillo domain known to be a recognition domain which is followed by an unstructured C-part.

3. AlphaFold-Multimer

The prediction of structural interaction between proteins and between protein and peptides is still a major research of interest. Numerous software has already been proposed to predict such complexes. As for protein structure prediction, an improvement of the results has been shown very recently with the involvement of DeepMind and their software AlphaFold-Multimer. As AlphaFold, its multimeric version is trained on a variety of experimental resolved structures. But for this new algorithm, training data is composed of protein complexes. Its score formula has also changed to fit the new kind of data. The old formula only looks at the interaction between residues of the same chain which is no longer the case. Indeed, now the inter-chain interaction is also taken into account with a rate of 0.8 and the intra-chain interaction contribution is now 0.2 [118]. This highlights the quality of the protein interaction rather than the unbound protein structures. But, unfortunately, contrary to AlphaFold, the multimeric version has not been accepted for publication yet and is still on BioRxiv. But its code has been made available and already a benchmarking of complexe structures has been done [216]. The Figure 59 shows how successful AlphaFold-Multimer is. Indeed, regarding only the top model according to its scoring AlphaFold-Multimer is about 50% of the time acceptable. Results are slightly better when regarding the top 5 of results. ColabFold and ZDOCK were also part of this benchmark and showed similar results with few lower results for ColabFOld and very low results for ZDOCK [120,217].

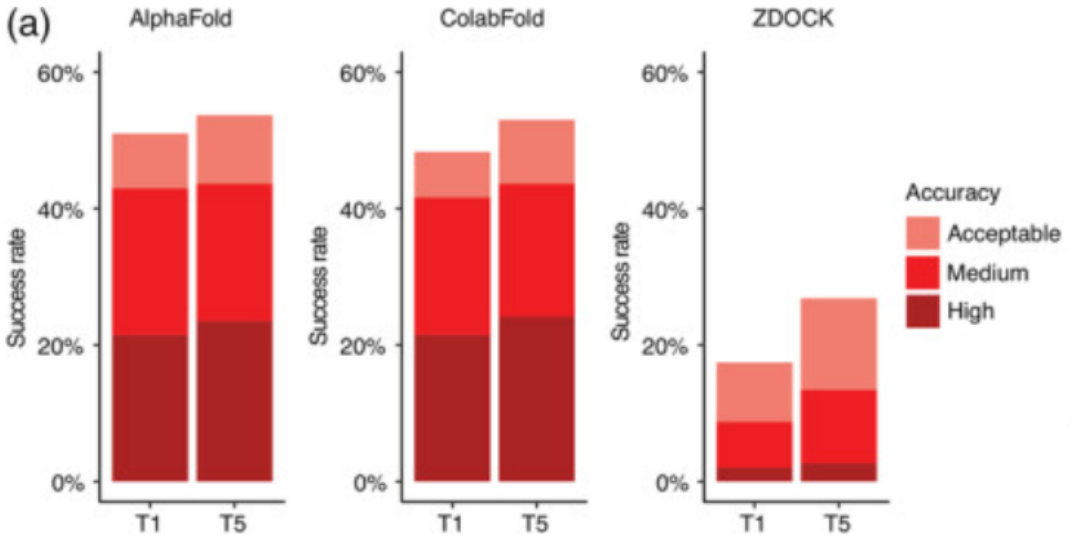


Figure 59. **Complex prediction success of AlphaFold, ColabFold, and ZDOCK for the top 1 (T1) and top 5 (T5) models considered**

Transient protein–protein complex structure prediction success by AlphaFold, ColabFold and ZDOCK. End-to-end modeling using AlphaFold 24 and ColabFold 29 was performed on 152 complex test cases.

AlphaFold failed to generate predictions for three complexes, thus AlphaFold predictions were obtained for 149 complexes; these 149 test cases were used to calculate success rates in this figure. Docking models were also generated with ZDOCK, 33 using unbound protein structures as input. Criteria for quality are based on CAPRI's one. AlphaFold and ColabFold models were ranked by AlphaFold pTM scores, and ZDOCK models were ranked by IRAD scores. 39 The percent success was calculated as the percentage of test cases with a given model accuracy from the top N models considered. Bars are colored according to the CAPRI quality classes.

From Yin *et al.* (2022)

B. Material and Methods

1. Experimental Structures

a) OGT Structures

Until 2021 there was no full structure of OGT in the Protein Data Bank (PDB) but only catalytic domains with the first TPR repeats or the TPR domain alone. In total, 38 structures are available:

- 1 dimer of full OGT (PDB ID: 7NTF)
- 1 structure of the TPR region (PDB ID: 1W3B) + Mutation (6EOU)
- 36 structures of OGT's catalytic domain:
    - 4 with UDP-GlcNAc only (PDB IDs: 3TAX, 3PE3,4GZ5, 4GZ6)
    - 1 with peptide only (PDB ID: 5BNW)

- 28 with UDP-GlcNAc + peptide (PDB IDs: 3PE4, 4AY5, 4AY6, 4GZ3, 4GYY, 4GYW, 4CDR, 4N3A, 4N3B, 4N3C, 4N39, 4XI9, 4XIF, 5C1D, 5LVV, 5LWV, 5NPR, 5NPS, 5VIE, 5VIF, 6MA1, 6MA2, 6MA3, 6MA4, 6MA5, 6IBO, 6E37, 6TKA)
- 3 with mutations (PDB IDs: 5HGV, 6EOU, 6Q4M)

b) β-catenin structures

When the key word “beta-catenin” is used in research for a molecule name inside the PDBe, there are 45 results. In all these results, the only resolved part of the β-catenin is its Armadillo domain. Indeed, the unstructured C and N-parts are too unstable to be crystallized. In total, there are 35 protein-protein complexes with the full Armadillo domain, a part of it or with peptide from the N-terminal segment which is phosphorylated.

2. Docking OGT/ β-catenin

As shown in his PhD objectives, we want to simulate the interactions between the OGT and β-catenin and this regarding the potential link between the two recognition domains (TPR and Armadillo) but also with the unstructured N-terminal segment of β-catenin and the TPR domain. In the end, we want to perform protein-protein docking but also protein-peptide docking.

To this, as AlphaFold-Mutimer shows already good results despite its unpublished article yet, it has been used for predicted both interactions. To model the interaction of the Armadillo domain and the TPR domain of the OGT, the Armadillo domain has been selected between the residues 152 and 663 included of the human β-catenin and the 1^{rst} and 473rd residues of the human OGT (Uniprot IDs:P35222, O15294).

Thanks to AlphaFold-Mutimer installed on our GPU computer, we were able to produce 20 models per random seed. As there are 5 random seeds, it will generate a total amount of 100 models. The same amount of models has been calculated for the docking protein-peptide. To determine the ideal size of the peptide to have enough information without losing quality, we decided to try 3 different sizes of the peptide (50, 100 and 151 which is the maximal size of the β-catenin unstructured N-terminal segment). In parallel we tried a different number of TPR repetitions to avoid losing too much time. The goal was

there to have the best p-TM score without losing too much information and keeping the asparagine ladder.

As seen in Figure 23, the ideal number of TPR to win time and have good pTM results is 8 with a peptide of 50 residue length.

To look at the interaction of the asparagine ladder highlighted in the TPR of the OGT, we selected peptide of 51 residues with experimentally provent *O*-GlcNAcylation sites (S23, T41, T112) and peptides without *O*-GlcNAcylated sites. These later have been selected on the unstructured N-terminal segment and C-terminal segment in order to be the farthest of a modified site possible. For the negative sites, S71 and S718 have been selected. For each site, several peptides have been constructed to be able to move the site from the beginning to the end of the peptide (in position 6, 16, 26, 36, 46). This, in the purpose to see if the asparagine ladder would be able to pull and push the site in the catalytic domain. To see if the asparagines interact with our site, we retrieve, with PyMOL, every residue at 5 Angstrom from the hydroxyl of each group and see if some correspond to the ones we are looking for. The asparagine indexes we are looking at are 321, 322, 325, 356, 390 and 414. We added the 322 to the five ones from Levine *et al.* article as it was very close and oriented inside the TPR lumen.

In order to see if AlphaFold-Multimer was able to predict the interaction of the OGT with the unstructured N-terminal segment of the β-catenin plus the full Armadillo domain, the prediction of 100 hundred models has also be performed with the ncOGT and the 663 first β-catenin residues.

C. Results

1. Docking protein/protein

a) ncOGT vs Armadillo domain

The modeling of the interaction between the full OGT and the Armadillo domain was performed to see if the two recognition domains have a high predicted quality score. To this, we modeled 100 complexes. But despite the number of structures available for these two

domains, the results are not that great. With an average of 0.2734 and a median score of 0.2649, the confidence is low. The best model according to AlphaFold-Multimer scoring function has a score of 0.42823 and is illustrated in Figure 60. It corresponds to the fifteenth prediction of the fourth random seed. In this model, we can see that the C-terminal segment of the Armadillo domain is close to the catalytic domain of the OGT. This result can be surprising as the literature describes the unstructured N-part to be modified by the *O*-GlcNAcylation enzyme. But we can hypothesize that this modified part, thanks to the interaction between the TPR and the Armadillo, is in a good position to go inside the TPR domain lumen. An unstructured part at the end of the Armadillo domain seems to be oriented inside the catalytic domain of the OGT. Regarding the residue inside, an interesting amino acid can be found here: a threonine. To follow the hypothesis of the destruction box going inside the TPR, it could be interesting to model the full Armadillo domain with the additional unstructured N-terminal segment.

b) ncOGT vs Armadillo domain with additional unstructured N-terminal segment

For this experiment, we also produced 100 models to have a high chance of producing a good model. Unfortunately, the average p-TM score is 0.2639 and the median slightly lower with a value of 0.2543. The best score has a p-TM score of 0.3602. It can be seen in Figure 61. Regarding the previous results we can see that the two folded parts have high confidence regarding the plddt scores but also the same interaction between the two different proteins. But contrary to our previous hypothesis, the unstructured N-part with the destruction box is predicted to be in the opposite of the OGT but with a very low score. This low score is similar to scores of unstructured parts. Despite these results we still consider our hypothesis. To this, we decided to perform the modeling of the OGT with a peptide of 50 residue long containing the D-box and see where the N and C-terminal segments of this peptide are predicted to be.

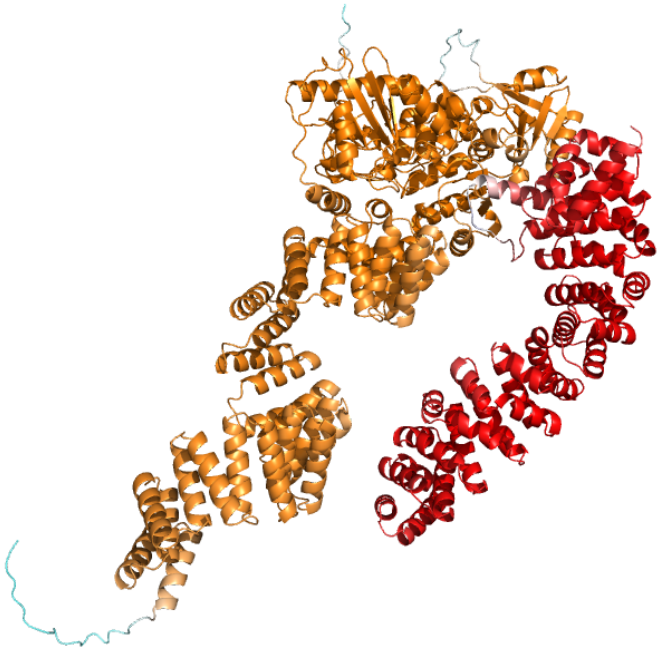


Figure 60. **Cartoon representation of the best interaction model between the full ncOGT and the Armadillo domain of the β-catenin**

Colors represent the confidence score. in cyan to orange is the confidence score of the ncOGT and blue to red represents the confidence score for the β-catenin.

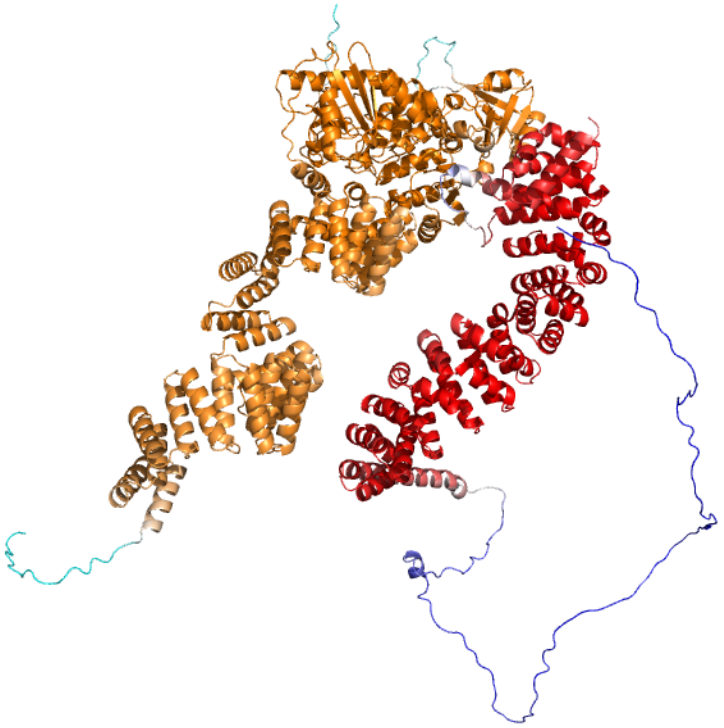


Figure 61. **Cartoon representation of the best interaction model between the full ncOGT and the N-terminal segment and Armadillo domain of the β-catenin**

Colors represent the confidence score. in cyan to orange is the confidence score of the ncOGT and blue to red represents the confidence score for the β-catenin.

2. Docking protein/peptide

Regarding the poor previous results to interpret the interaction between the Armadillo domain and the destruction box and the OGT with 8 TPR repeats, the modeling of only one big peptide known to be *O*-GlcNAcylated in its center has been generated one hundred times. This time, the scores are much higher with a median of 0.6380 and a mean of 0.6248. The best model according to AFM is the fifth model of the second random seed, with a score of 0.7398. It can be seen in Figure 62. These results show a good confidence but, looking at the plddt score of residues, this high score is mostly due to the OGT modeling. Indeed, as it can see in Figure 63, the peptide plddt score is much lower but two parts seem more confident. These parts, in red in Figure 62 (A), are *O*-GlcNAcylated. The first peak corresponds to the residues around the 25th β-catenin amino acid. This area counts an *O*-GlcNAcylated site which is the serine 23. The second peak with the highest confidence corresponds to the center of the peptide where belong the two other *O*-GlcNAcylated residues (threonine 40 and threonine 41). The higher scores for the two parts can be probably explained by a confidence in the interaction of these parts with the OGT. As this type of interaction is the one we are looking for we looked at the composition of the OGT at the interface with the PyMOL software at 5Å radius.

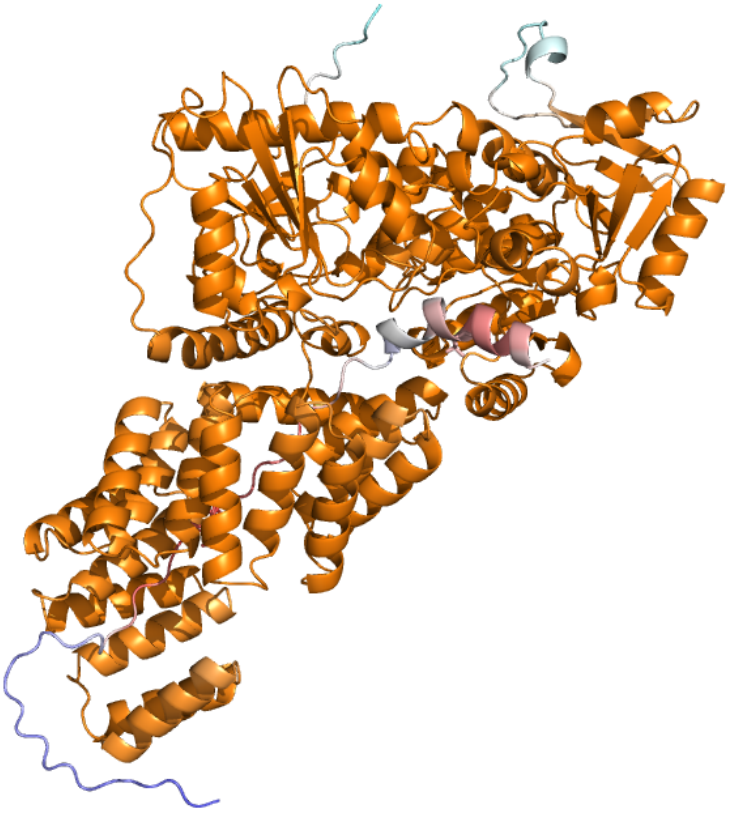


Figure 62. **Cartoon representation of the best interaction model between the ncOGT with 8 TPRs and an *O*-GlcNAcylated N-terminal segment peptide of the β-catenin**

Colors represent the confidence score. in cyan to orange is the confidence score of the ncOGT and blue to red represents the confidence score for the β-catenin.

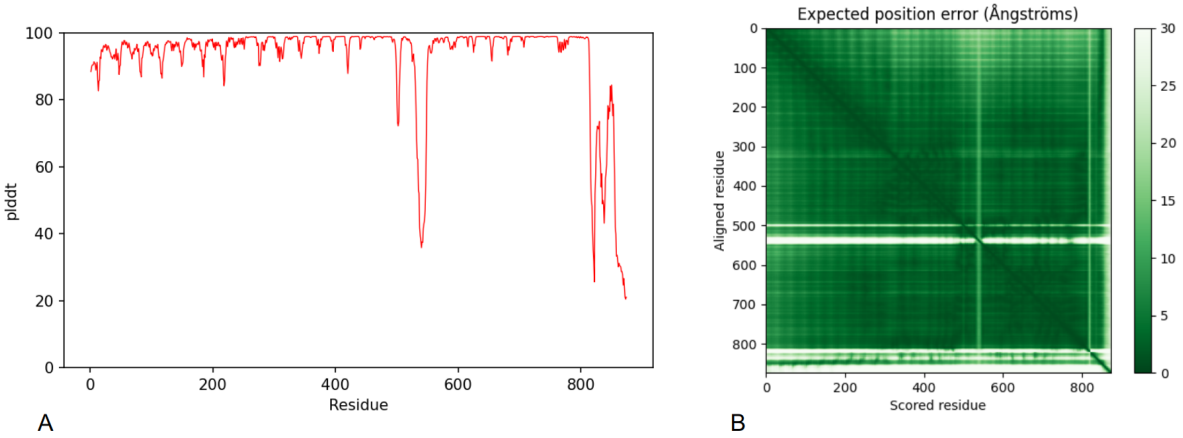


Figure 63. **Confidence score representation of the best interaction model between the OGT and the D-box peptide**

A: plddt score of each residue. B: Position error expected. The two partners sequences follow each other

The OGT residues interacting with the peptide part near the serine 23 and their occurrence is recapped in the Table 25 below:

Res	A	C	E	G	K	N	P	R	T	V	Y
Count	**2**	1	1	1	1	**2**	1	1	1	1	1

Table 25. **Counting of OGT residues that interact with the 3-9 peptide residues**

The numbers in bold have the higher count

These results do not show any favored residues except two asparagines and two alanines. The same operation has been done for the second peak and the results are in the Table 26.

Res	A	C	D	E	F	K	N	Q	R	S	Y
Count	1	2	5	1	4	2	**10**	1	1	5	5

Table 26. **Counting of OGT residues that interact with the 19-31 peptide residues**

The number in bold has the higher count

This table highlights the high number of asparagines interacting with the peptide. Regarding the sequence, a repeated pattern can be found. This pattern illustrated in Figure 64, shows the presence of a tyrosine at 9 residues of a phenylalanine (or an aspartic acid) followed by an aspartic acid (or alanine) two residues after. These residues are then followed by a serine and an asparagine. This repeated motif which is almost identically found 5 times can be explained by the TPR repeats which are made to create a superhelix. The important role of asparagines inside the lumen of the OGT's TPR has been put forward with Levine *et al.* 2018 [37]. But even if some asparagines of these repeats are part of the asparagine ladder, others have not been highlighted. But as the asparagine ladder has been shown to reduce dramatically the OGT efficiency so it could be interesting to focus on this part with all the β-catenin *O*-GlcNAcylated sites which are serine 23, threonine 40, 41 and 112.

An interesting thing should be to superimpose the results of the three last parts (including this one) to see if the predictions are consistent together. The results are shown in Figure 65. The prediction of the position of the Armadillo domain of the β-catenin is very similar from the two different predictions. The orange part on this figure represents the

peptide inside the TPR. According to this image we can reasonably imagine that it can go inside the TPR domain and get the same conformation as in red. But we can see a clash from the peptide modeling and the Armadillo domain which are predicted to be in interaction in the same area at the output of the OGT catalytic domain. This part seems to be often predicted to be in interaction.

Figure 64. **WebLogo representation of the pattern found inside the TPR of the OGT**

X corresponds to any residue and is not in interaction with the peptide. Figure generated by WebLogo [218]

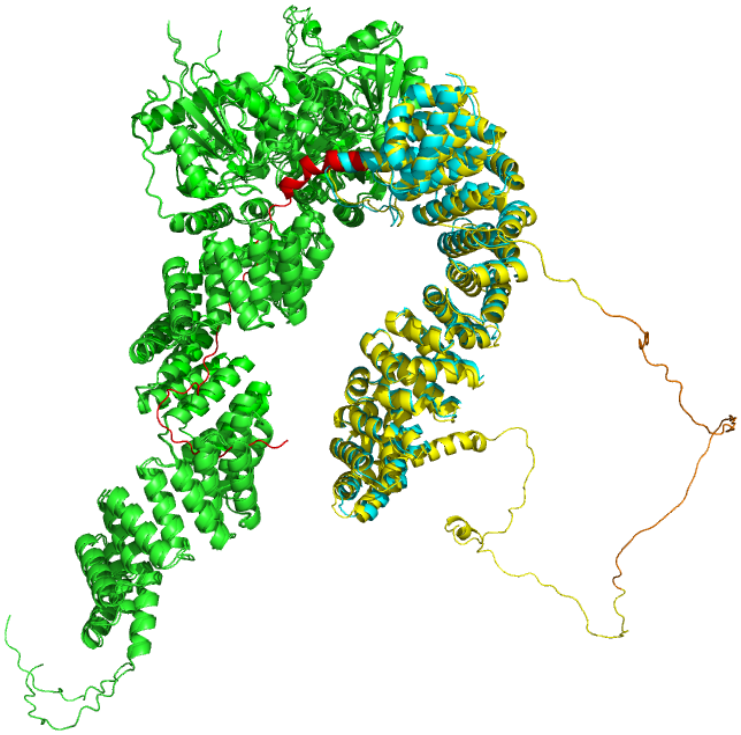


Figure 65. **Concatenation of the 3 best models of the complexe prediction between OGT and β-catenin**

In green is the OGT on which every models have been aligned, the blue corresponds to the model prediction for the Armadillo alone, the yellow is the armadillo domain with the unstructured N-terminal segment and the red part is the 51 long residue peptide with *O*-GlcNAcylation site. The orange part on the unstructured part corresponds to the peptide in red.

3. The asparagine ladder interactions

As shown, in the previous results, we saw an interaction of the *O*-GlcNAcylation site with many asparagine and a previous article (Levine *et al.* 2018) highlights the need of an asparagine ladder [37]. To see if the asparagines interact with the modified sites, we decided to model peptides with *O*-GlcNAcylated sites in different positions in the peptide (6,16,26,36 and 46) and non peptide with a serine non modified as negative data. The peptides are 51 residues long. To be able to see if the interaction exists, PyMOL was used to retrieve all the residues around 5 Å of the hydroxyl group of serine and threonine of interest.

a) Positive sites

For the serine 23, we noticed interaction with the asparagines 390 and 424 only when the site was in position 6 and no interaction when in position 16. We did not go further in the position because of the presence of threonine 40 and 41 which changed the results. The two models have a score of 0.3128 and 0.6398 for position 6 and 16 respectively. Starting from position 16, the threonines 40 and 41 are in the peptide which change the results and can explain the increase of the confidence score despite no interaction with asparagine from the ladder.

For the threonine 40 at the different positions, we saw interactions with different asparagines. These interactions are summarized in the Table 27. These results show a better confidence in the model with a highest score of 0.7231. We can see the number of asparagine in interaction is reduced while increasing the index of the *O*-GlcNAcylated site and we can see a move of the interaction from the N-terminal segment to the C-terminal segment of the TPR. This may mean that the peptide is predicted to be at more or less the same place and only the site is moving. There are no results for the position 46 because the results would be distorted because of the presence of the serine 23.

Asparagin Indexes	Position of threonine 40			
	6	16	26	36
321				
322				
325	x			
356	x	x		
390	x	x	x	x
424		x	x	x
Score	0.72306	0.57208	0.65129	0.67357

Table 27. **Interaction between asparagines from the ladder and threonine 40 in different position of 51 residues long peptide**

Regarding threonine 41, the results are different from the threonine 40 which is surprising as they are neighbors in the peptide sequence. Indeed, only interaction with asparagines 390 and 424 have been retrieved for position 6, 16 and 36. The confidence of the models are also lower with a best confidence score of 0.6879. This difference can be explained by the conformation of the peptide imposed by the interactions with the threonine 40.

Finally, the threonine 112 is only predicted to interact with the asparagines 321, 322, 325 and 356 in position 16. Otherwise no interaction has been found. The confidence score for the 5 positions is lower than for the other complexes with a best score of 0.4666 (which is the model with the interactions)

b) Negative sites

Results for serine 72 show no interaction between its hydroxyl group and the asparagine from the ladder. The model scores are close from the one of threonine 112 for the position 6, 16, and 26 with values between 0.4435 and 0.4903. The scores are higher for positions of 36 and 46 with a score of 0.7125 for this last position. This can be explained by

the presence on the peptide of the threonines 40 and 41 which are interacting with 321, 322, 325 and 356 for the first one and with asparagines 325,356 and 390 for the second one.

Interestingly, the results for the serine 718, not known to be *O*-GlcNAcylated, are very similar to positive data with even more interactions. These interactions are retrieved in the Table 28. But the confidence scores of the models are still lower than the ones of threonine 40 and 41. According to these results; the peptide seems to be in different conformation. After regarding in the different complex models, there is indeed a changement of conformation: the peptide is inside the TPR with a loop at the N-terminal segment or C-terminal segment of the TPR depending on the positions. This can be explained by the low confidence score and the complex structure found in the PDB. The peptide is predicted to be here but without a clear conformation.

Asparagine indexes	**Position of the serine 718**				
	6	**16**	**26**	**36**	**46**
321		x	x		x
322		x	x		x
325		x	x		x
356		x			
390	x			x	
424	x				
Score	0.5370	0.5430	0.6041	0.5603	0.6675

Table 28. **Interaction between asparagines from the ladder and serine 718 in different position of 51 residues long peptide**

To be able to compare the differents conformations, every positive model have been analyzed and for every models the N-part of the peptide is predicted to go out of the catalytic domain, except for T112 which is predicted to be outside of the TPR where the rest of the TPR should be if complete in fuzzy conformation. For T112 in position 16, the peptide

is predicted to create an interaction between alpha helices (one from the peptide) and the rest of the peptide go into the TPR domain. For the serine 23, when there is no threonine 40 and 41 the peptide is also conformed as negative data but once there are these two sites it has the same conformation as we look for threonine 40 and 41.

D. Discussion / Conclusion / Perspectives

Regarding the protein-protein modeling between full ncOGT and Armadillo domain, the complex scores are low in general, with an average of 0.26 or 0.27. Selected models have higher scores, with values of 0.43 and 0.36, but these are still low values, even when the confidence of the individual monomer structures are higher; these low scores can be explained by the little area of predicted interaction. Indeed, as AFM uses 80% of the inter-chain prediction value to create its score, the interaction area at the output of the catalytic domain of the OGT is not enough to have a high confidence score. The lower results for the complex between the full ncOGT and the Armadillo with the unstructured N-part can be explained by the bigger size of the substrate with the same interaction area leading to a lower ratio of interaction. As shown in the two complex figures (Figure 60 and Figure 61), the unstructured part predicted to interact with the OGT contains a threonine which is an amino acid that can be *O*-GlcNAcylated. It could be interesting to add an UDP-GlcNAc on the histidine 498 in the catalytic pocket and see if this threonine is close to it and could be *O*-GlcNAcylated.

Protein-peptide docking performed on the ncOGT reduced to 8 TPRs with an *O*-GlcNAcylated peptide from the unstructured N-part of the beta-catenin showed better results. Indeed, the average score was around 0.63 and the best model produced a score of 0.74. The higher scores come probably by the higher interaction between the two entities. It is interessant to notice a repeated pattern inside the TPR lumen involved in the interaction with the peptide. This motif with many asparagine could have a “pull and push” purpose to bring the peptide inside the catalytic domain. It could be interesting to see, as it was done for the asparagine ladder, if some residues of this motif mutate leads to the loss of the activity. A decrease of the OGT activity could show new substrate recognition mechanisms. But it could be very difficult to show here a specific interaction with the unstructured part of

the beta-catenin as every unstructured part could interact with this repeated motif as the asparagine ladder.

Also, for these results, we reduced the complex by reducing the number of TPR repeats. To see if the results are the same with the full ncOGT, the calculations need to be redone which is actually the case as they are running on our GPU.

Regarding the concatenation of the 3 best models obtained for the interaction of the OGT and beta-catenin (Figure 65), we can see a predilection for the OGT to interact at the output of its catalytic domain and not with its TPR which is its recognition domain. This also creates a clash between two models as the peptide is predicted to go out from the OGT in the same area as it is predicted to interact with the Armadillo domain. From this model, we can hypothesize the need of a third protein to present the unstructured part of the beta-catenin to the lumen of the TPR supporting the hypothesis of chaperone proteins to help the OGT to its substrate recognition.

For the investigation of the asparagine ladder, we can see the low interaction between the asparagine 321 and 322 and the different peptides except for S718. But as one is part of the asparagine ladder highlighted by Levine *et al.* this results is surprising. It can be explained by the reduction of TPR repeats. The new prediction with the full ncOGT may make up for this absence of interaction. The interaction of these two asparagine with the peptide around the serine 718 can be explained by the fuzzy conformation inside the TPR which takes more place. Even if the scores are lower for this peptide the high number of interactions can make us think of an *O*-GlcNAcylated site which has not been experimentally proven yet.

The predicted models show a higher confidence score when the hydroxyl group of serine and threonine interacts with the asparagine ladder. In Section II, we have tried to use this information to predict *O*-GlcNAcylated sites. Unfortunately such interactions are not often found for *O*-GlcNAcylated sites. But the score has not been taken into account. Maybe a threshold coupled with this information may help predict such modified sites but without much conviction. The conformation of the peptide at the output of the OGT catalytic domain

may be another line of research. Indeed, for positive data the peptides have this conformation except for T112. This last site peptide is predicted to be outside the TPR where the TPR domain should continue.

Also, it could have been interesting to test our scoring method to select models despite the AFM confidence score.

Finally, despite all the efforts in the prediction of the interaction complex between the OGT and the beta-catenin, we were not able to find any specific interaction that could have helped us prevent its formation.

VII. General conclusion

Proteins are large molecules with more or less complex structures. They play many essential roles in the living world. They are responsible for most of the actions in the cells, inducing structure, function and regulation of tissues. But these protein activities are regulated by many complex interactions. Protein-protein interactions study is a main field of research. Indeed, proteins influence different metabolic pathways via interactions with other proteins from other pathways. Understanding the mechanisms inside these numerous crosstalk may help in different actuel health issues. Some diseases have been shown to be provoked by a misregulation of proteins as a worldwide pandemic crisis can emerge as a result of protein mutations enhancing the host-pathogen interaction.

Various experimental methods were created to be able to identify, visualize or prevent protein-protein interactions. Unfortunately, experiments are consuming in terms of money and time or even not being able to characterize specific interactions.
That is why, in parallel, the number of computational methods has risen to help researchers in the identification of interactions. But these methods still are predictions with more or less a good success rate. The need to evaluate such tools with their results and to improve techniques is still a major key in research.

During this PhD thesis, the objectives were cut in different problematics.

First the need to compare already existing *O*-GlcNAcylation prediction sites on a benchmark. The poor results of this analysis lead to the demand to develop a new method to improve this prediction. Second, the emergence of the COVID-19 crisis leads to the need of identifying interaction to the atomic levels to encounter the virus propagation. The objective was here to develop a method to select the more likely models proposed by protein-protein three dimensional interaction predictors. Third, as with any new method, it must be tested on a sufficiently large data set and compared to other current tools. Finally, atomistic interaction between two proteins involved in the colorectal cancer needed to be elucidated.

A. Prediction of *O*-GlcNAcylation sites

The first objective of this part was to determine the efficiency of available prediction tools for the *O*-GlcNAcylation prediction. Indeed, this post-translational modification has been shown to be implicated in various diseases such as cancers, diabete and Alzheimer' disease and is estimated to modify more than 1000 other proteins. To that, a new dataset has been built to test three actual tools. But even if these tools argued a good sensitivity, we highlighted their poor capacity to predict *O*-GlcNAcylated while having a high rate of false positives. The need to focus on precision (or positive predictive value) rather than sensitivity has been discussed. To improve current methods to predict this post-translational modification, we added structural features of experimentally proven substrates. Unfortunately, our results were not able to show a better prediction efficiency and demonstrate how far we are to predict such modification. In addition, we hypothesize the need of chaperone proteins to help the only enzyme capable of the UDP-GlcNAc addition to recognize its substrate.

B. Analyses of SARS-CoV-2 and human protein interactions

The worldwide COVID-19 crisis brings quick and joint efforts to stop the virus spreading. This is the case of a study of Gordon *et al.* where they highlighted more than 300 specific interactions between SARS-CoV-2 and human proteins. The CAPRI experiment with its worldwide community proposed prediction models for 5 of these interactions. But there was still a need to find the most likely ones. We showed the difficulty to find a consensus based on specific contacts with hierarchical clustering and meta-clustering with Markov Clustering (MCl). To counter this problem, we developed a method based on the adjacency overlap. This method shows the capacity to identify good models on a validation set of 4 resolved complexes but it was difficult to define a threshold to determine if a model produced by the CAPRI community for the COVID-19 special Round. We did point out the difficulty to model interaction between SARS-CoV-2 and human proteins and the need to validate our method on a bigger dataset with the second objective to define a threshold from which we could see if models are more likely than others.

The number of residues predicted to be at interfaces and their mutation rate have been analyzed pointing out the consensus area for the human protein interface.

C. Validation of the adjacency overlap method

In this part of the PhD, we described a brand new dataset from the CAPRI community called Scoreset already available on the scoreset.org website. This dataset is composed of several interfaces between proteins. These interfaces can be from eukaryotes, archaea, bacteria or virus proteins and can be divided into three different sets, Uploaders, Predictors and Scorers. We tested our Adjacency Overlap method on the different subsets and kingdoms. The results show the good capacity for our method to discriminate good from bad models, with better results for the Uploaders and Predictors sets and Eukaryotes, Bacteria and Virus sets. The scoring method also performed well on dimers with a better precision for the homo-dimer than hetero-dimer. Unfortunately we highlighted its poor ability to score models of Archaea complexes or complexes with proteins coming from two different kingdoms.

We were also able to define a threshold maximizing the precision (in depreciation of the sensitivity) for our method. Unfortunately, our method did not score as acceptable models from the CAPRI-COVID Round.

In a second time, we compared our method to two of the most recent scoring tools which are iScore and GNN-DOVE. To perform this comparison, the adjacency overlap has been run on a previous CAPRI benchmark also analyzed by the two scoring methods. This benchmark is composed of 13 targets. If our method was not able to highlight a higher number of good models target-wise, it shows a higher capacity to find good models inside a target.

D. Modeling of interaction between OGT and β-catenin

The β-catenin is an oncoprotein involved in the Wnt pathway which is usually phosphorylated on an area of its N-terminal segment called Destruction box (D-box). Its phosphorylation leads to its proteasomal degradation. But the cross-talk between

phosphorylation and *O*-GlcNAcylation on this D-box and more precisely on the T41 has been demonstrated to improve the risk of colorectal cancer. Our objective here was to use modeling methods to predict the specific interface to possibly depress this interaction. The β-catenin can be divided into three main parts: the N and C-terminal segments which are unstructured and the Armadillo domain which is a recognition domain composed of a super helix. As the N-terminal part is modified, the hypothesis of this part goes into the TPR lumen while Armadillo stabilizes the complex by interaction with the outside of the TPR has been submitted.

The first objective was to predict the interaction between the TPR domain of the OGT and the Armadillo of the β-catenin, both described as recognition domains. For this purpose, AlphaFold-Multimer has been used to produce one hundred models. Unfortunately, the confidence of the models was low and the predicted interaction area between the two domains can be discussed as it is at the output of the OGT catalytic domain. The N-terminal part of the Armadillo domain was pointing towards the TPR domain so we decided to model the Armadillo domain with the N-terminal segment of the β-catenin. Unfortunately the results show a N-terminal segment going in the opposite direction. But the size of this unstructured region is sufficient to maintain our hypothesis of the destruction-box going inside the TPR.

The second objective was to model the N-terminal segment with *O*-GlcNAcylated sites with the OGT. The results show a high number of interactions of the peptide with asparagines. From these results, supported by the study of Levine *et al.* on the presence of an asparagine ladder impacting the OGT activity, we hypothesized the role of asparagine to push and pull the unstructured substrate inside the catalytic domain of the OGT. We also highlighted a residue pattern repeated inside the TPR.

Unfortunately we were not able to computationally determine a specific interaction interface and this study deserves more time to obtain more results.

VIII. General discussion and associated perspectives

A. *O*-GlcNAcylation prediction: a lack of information

The *O*-GlcNAcylation prediction is today an unreached objective. From the available software using protein sequences to the structural features used in Mauri *et al.*, no one provides a good precision. The heterogeneity in the positive data and its low proportion compared to the negative make this post-translational modification a very hard task without consensus. The results may be different with a larger amount of data and the creation of the *O*-GlcNAc Database may provide a sufficient set to improve the prediction. In my opinion, the heterogeneity of the *O*-GlcNAcylated sites coupled to the fact that only one enzyme is able to catalyze the sugar addiction, there are unknown mechanisms (notably with chaperone protein) which help the OGT to recognize the substrate. The TPR domain has been shown to help substrate recognition but the point is this domain helps the substrate to go into the catalytic domain rather than recognize it. Analyzing the interactome of OGT and its substrates may highlight proteins involved in this mechanism and it is a thing to keep in mind while looking for experimental sites of *O*-GlcNAcylation. Experiments to find OGT partners could help us to understand the underlying information and provide data useful for its prediction.

B. CAPRI-COVID Round: interaction between viral and human proteins

As shown in this results part, the prediction of interaction between human and viral proteins is a stuff task with low consensus. The low consensus may be due to the cross kingdom interactions where two different types of protein interact. As mentioned before, viruses are known to have a high mutation rate leading to the survival of the virus or interaction with new hosts. It is common for docking algorithms to predict interaction area thanks to the co-evolution of the amino acids. This information gives good results as shown in InterEvDock or more recently in AlphaFold-Multimer. Here we can hypothesize that the mutation of the viral proteins help them to interact with human proteins without these latter being mutated. In that case there is no co-evolution and even if we can find consensus on human protein it is hard task to predict the same with viral proteins. It could be interesting to evaluate the

mutations which allow the virus to infect a new host to direct the interaction interface recognition.

Clustering and meta-clustering can give a good idea to see if the prediction of a complex was a hard task or not. Coupled to the adjacency overlap ranking, we can be even more precise. Unfortunately it seems that this method alone can't score models. It could be interesting to perform these clustering and meta-clustering on a bigger dataset as we did for the adjacency overlap.

At the beginning of this project AlphaFold-Multimer was not released and it was the very beginning of the colab script with a linker to produce multimer form AlphaFold. But we also compared the results of AlphaFold nowadays and the results showed for each target heterogeneous results.

Target 185 was pushed aside because of its difficulty. It could be interesting as it is only composed of viral proteins to test our method on it. Also, it could be interesting to perform the same analysis with every of the 332 interactions determined by Gordon *et al.* [158].

Adjacency overlap results will be discussed in the next part.

C. Adjacency overlap: a new scoring method?

Adjacency overlap has been tested on a big datatest with thousands of models and showed really good results which were different depending on the several subset we look at. To be able to really compare the results according to the kingdom it could be interesting to have the same amount of data in each category. Nevertheless, the tendencies should not change and adjacency overlap should be more efficient with eukaryotes, bacteria and viruses. It could be interesting also to compare the efficiency of this method in each kingdom and their general difficulty. The difficulties of scoring other kingdoms can be explained by the lack of knowledge, the low amount of data and the reduced need to model such complexes.

Comparing the adjacency overlap method on the same benchmark as the already available scoring methods shows similar results target-wise but enlarges the number of good models in the top 10. It could be also interesting to compare the first 1 and 5 as in CAPRI assessment. Contrary to the two other compared methods, our tool is not based on machine learning and therefore depends on a training set. Meaning, there is no need to train our model again; it will only rely on the methods developed by predictors. This is a good point but can also be a bad one as it is depending on other knowledge.

It could be interesting to test this method in the next CAPRI Scoring Round to compare it to recent scoring methods. Scoring can also be performed by humans and could be a good evaluation to compare our method results to the human ones.

The most urgent thing to determine now is the number of models required for our method to perform well, to fix the number of different algorithms needed and also to see the impact of the diversity of the models. We showed in our results the better results of our method on the P-set than the U-set where the number of models is different suggesting that a too high number of models is not a good thing. But this result is contradicted by the lower results in the S-set composed of fewer models. We can hypothesize that the S-set is composed of more similar models than the P-set thanks to the scoring and our method, based on the overall consensus, could be not able to discriminate the good models from the all set if its majority is too similar. To have a better understanding and a more useful use of this method it could be interesting to see how it reacts in function of acceptable or better quality rate inside a set.

Regarding the better results of our method in the P-set than the U-set, we could consider our method as a meta scoring method retrieving all the information from different models and thus knowledge.

D. Modeling of the OGT and β-catenin interaction

The modeling of the interaction between the *O*-GlcNAcTransferase and the β-catenin is not a trivial task. Indeed, β-catenin is a protein which can be divided in three main parts

but here we only focused on the N-terminal segment which is unstructured and the Armadillo domain. The hypothesis of a strong interaction between the TPR domain of the OGT and the Armadillo domain could not be validated as no model shows a strong interaction interface. The only interaction found by AlphaFold-Multimer is at the output of the catalytic domain of the OGT where the *O*-GlcNAcylated substrate should come out from. This part of the OGT has been predicted to interact with alpha helices in different predictions as in the model of the peptide containing the threonine 41 of the β-catenin. It could be interesting to make an experimental directed mutagenesis to see if residues are important for the substrate stabilization whether it is for peptide or structured proteins.

We also highlighted the role of asparagine already emphasized by Levine e*t al.* but others found a repeated motif inside the TPR [37]. The role of the tyrosine at the beginning of the pattern followed by asparagine could also be interesting as a candidate for directed mutagenesis to see if the OGT activity can be decreased. If not, looking at the specific interaction with β-catenin could maybe highlight a specific interaction with it and could be a way to develop an OGT/β-catenin interaction inhibitor.

IX. Other projects

A. What is the potential impact of genetic divergence of ribosomal genes between Silene nutans lineages in hybrids? An *in silico* approach.

This parallel project was done in collaboration with Zoé Postel (PhD student) and Pascal Touzet (Professor of Lille University), her supervisor. We also co-supervised a group of master students (Andréa Bouanich, Marion Liotier, Zinara Lidamahasolo) in bioinformatics which greatly helped us with the first analysis (results in table 2). Pascal Touzet and Zoé Postel formulated the biological questions about the plastid-nuclear interactions in lineages of *S. nutans*. I, under the supervision of Marc Lensink, conducted the analyses, made the figures and wrote the manuscript (Material and methods + results & discussion). Zoé participated in the interpretation of the results and writing of the manuscript (introduction + results & discussion).

1. Introduction

Plastids are ancient cyanobacteria that integrated the eukaryotic cells as endosymbionts roughly a billion years ago (1). After this integration, this organelle transferred a certain amount of its genes to the nucleus, ending up encoding only a few of the original gene set (2). These remaining 120 or so genes are involved in photosynthesis and housekeeping function in the plastid (3). Due to these transfers, the essential plastid protein complexes are encoded both by plastid and nuclear genes whose gene products are targeted to the plastid (later called nuPt). Plastid and nuPt genes need to interact with one another for correct protein complex function (4–6). Nuclear and plastid genomes have contrasting features, such as differences in mutation rate that is much lower in the plastid (7) or different inheritance patterns with biparental inheritance for the nuclear genome and maternal one for the plastid (8). As so, any mutation occurring in one of the two partners will generate strong selective pressure for fixation of compensatory mutation in the other one (4). Tight co-adaptation between interacting plastid and nuclear genes are then required and enforced (9). Independent accumulation of mutations in both plastid and nuclear genes can occur in isolated lineages or populations (10). If hybridization occurs between these isolated lineages, co-adaptation between nuclear and plastid genes will be disrupted in hybrids (2). Indeed, hybridization will bring together a plastid genome mismatched with a part of the hybrid nuclear background, leading to potential hybrid breakdown (i.e. decrease in fertility and survival) through creation of plastid-nuclear incompatibilities (PNIs) (11). These incompatibilities are thought to be part of the first post-zygotic reproductive barrier to emerge as they can lead to reproductive isolation between lineages through decrease in hybrid fitness (12). When such incompatibilities are involved in speciation (*i.e.* the process leading to reproductive isolation (13)), reproductive isolation is asymmetric in reciprocal crosses, depending on the lineage that is the plastid donor (14,15). Molecular mechanisms and identification of co-adapted pairs of genes is still largely missing, even though few studies identified PNIs likely involved in reproductive isolation (16–19). Contrastingly, patterns of co-evolution between plastid and nuclear genes have been extensively studied especially in plant species exhibiting accelerated rate of plastid genome evolution (20–24).

PNIs were also potentially involved in reproductive isolation between lineages of *Silene nutans* (Caryophyllaceae) (25). This species is composed of several genetically differentiated lineages in France, based on plastid sequences and nuclear microsatellite markers and their geographic distribution in Europe reflects colonization from past glacial refugia (26,27). Diallel crosses between four of these lineages, an eastern one E1 and three western one (W1, W2, W3) revealed strong and asymmetric reproductive isolation between them (28)(Van Rossum et al., in prep). Analysis of plastid genetic diversity and nuPt genes in these four lineages uncovered lineage specific coevolution patterns between plastid and nuclear genes that could result in PNIs in hybrids (25). Candidate gene pairs for

PNIs were identified in the plastid ribosomes (25), a plastid complex composed of a large and a small subunit and encoded both by nuclear and plastid genes (29). Plastid and nuPt genes encoding this complex exhibited the largest amount of lineage specific non-synonymous (NS) mutations (i.e. mutation leading to a change of the encoded amino-acid) and elevated d_N/d_S (i.e. proportion of non-synonymous (N) and synonymous (S) mutations on the total number of N and S sites) (25). Elevated d_N/d_S was thought to be the result of positive selection on the plastid genes and one some nuclear genes (25). Regarding the nuclear genes, d_N/d_S was significantly higher compared to nuclear genes encoding the cytosolic ribosome (*i.e.* gene products not targeted to the plastid), suggesting this increase in number of NS mutations might be the result of plastid-nuclear coevolution (25). Some of the NS mutations identified in plastid and nuclear genes encoding the large and small plastid ribosomal subunit were directly located at protein residue contact position, suggesting structurally mediated co-evolution between plastid and nuclear genes within the plastid ribosome (25,30).

Disruption of co-adaptation in the plastid ribosome can have dramatic consequences, especially if it concerns essential plastid ribosomal genes (31,32). For example, a missense mutation in the essential plastid gene *rps4* causes defaults in plant development and photosynthetic performances in *Brassica campestris* ssp. *pekinensis* (33). More generally, any mutation in plastid or nuPt essential ribosomal genes can have dramatic impact on photosynthesis, as all of the plastid-encoded photosynthetic genes are translated by the plastid ribosome (32). Many plastid-nuclear gene pairs encoding subunits of the plastid ribosome were identified as potential candidates for PNIs between lineages of *S. nutans* (25). To further identify which of these pairs could be responsible for PNIs, we used the crystallographic structure of the spinach plastid ribosome (34) to assess the potential impact of the NS mutations identified in each plastid and nuPt genes of these pairs on the residue contact interactions between plastid and nuclear proteins within the large and small plastid ribosomal subunit. To do this, we modeled the different NS mutations for each lineage and each nuclear and plastid candidate proteins on these subunits to further narrow down the list of PNIs candidates in the plastid ribosome. Models were then transformed into graphs called Residue Interactions Networks (RINs) and from these networks centralities of the residues were calculated (35). Centrality of a residue represent its amount interactions with other residues. The centralities of the modified residue contact interactions can be used to see if the mutations of interest may have a role in modifying a central residue contact interaction between plastid/nuclear gene pairs and potentially disrupt plastid ribosome structure, resulting in PNIs in inter-lineages hybrids. This method has been shown to highlight residues important for protein structure and function (del Sol *et al*., 2006; Hu *et al.*, 2014; Trouvilliez *et al.*, 2022) (36–38).

Table 1 – List of the nuclear and plastid genes encoding the small ribosomal subunit, selected as candidate for PNIs in Postel et al, 2022 and analyzed in the present study.

Genome	Gene name	UniProt identifier	Chain's name	Entitled	*Spinacia oleracea*		*S. nutans* lineages				
					Position	Amino Acid	Position	Amino acid			
								E1	W1	W2	W3
Nuclear	***rpl13***	P12629	K	S	132	A	152	A	S	A	A
	rpl19	P82413	Q	Y	229	L	235	F	L	L	L
	rpl21	P24613	S	AA	32	P	61	R	K	K	K
	rpl27	P82190	X	FA	20	L	20	L	L	V	L
	rpl3	P82191	D	L	35	S	32	S	F	F	F
Plastid	***rpl14***	P09596	L	T	49	N	49	N	H	H	H
					104	R	104	G	R	R	R
	rpl16	P17353	N	V	26	R	24	N	N	T	N
					78	P	76	P	P	S	P
	rpl22	P09594	T	BA	9	K	6	R	G	R	R
					114	V	92	L	F	L	L
					115	K	93	K	N	N	N
					121	R	99	R	H	H	H
	rpl32	P28804	1	B	22	K	22	K	K	M	K
					28	A	28	A	A	V	A
					49	R	49	L	L	P	L

Highlighted green: different amino-acid between *S. oleracea* and *S. nutans* lineages
Highlighted purple: different amino-acid for one or more lineages of *S. nutans* compared to *S. oleracea*
Highlighted orange: different amino-acid for one lineage of *S. nutans* compared to the others and to *S. oleracea*
* : mutations identified as under positive selection in Postel et al, 2022.

2. Material and Methods

a) Identification of mutations

Mutation identification was previously done in (25). Briefly, we searched for all mutations differently fixed between lineages of *S. nutans*, in plastid and nuclear gene sequence alignments of the plastid ribosome, using an in-house biopython script (https://github.com/ZoePos/Variants-dectections) (25). We then aligned the plastid and nuclear gene sequences of *S. nutans* with the one of *S. oleracea*, used as reference. The spinach structure and the associated protein sequences were available in PDBe (european Protein Data Bank) (PDB id: 5MMM) and contains 60 chains corresponding to the different rps and rpl subunits and some RNA strands (39). After aligning *S. nutans* and *S. oleracea* sequences, we compared the encoded amino-acid between *S. nutans* and the spinach, at each position containing mutations between lineages of *S. nutans*. We reported the different mutations identified in lineages of *S. nutans* and the corresponding amino acid in the *S. oleracea* in Tables 1 and 2.

Table 2 – List of the nuclear and plastid genes encoding the large ribosomal subunit, selected as candidate for PNIs in Postel et al, 2022 and analyzed in the present study.

Genome	Gene name	UniProt identifier	Chain's name	Entitled	*Spinacia oleracea*		*S. nutans* lineages				
					Position	Amino Acid	Position	Amino acid			
								E1	W1	W2	W3
Nuclear	***rps10***	P82162	j	TA	77	D	82*	E	E	E	K
					142	Y	147	F	F	F	Y
	rps13	P82163	m	WA	127	E	125*	E	E	E	Q
	rps21	P82024	u	EB	88	V	36	I	V	V	V
					89	L	37	F	L	L	S
					91	Q	39	N	D	D	D
					133	H	81	S	A	S	S
					157	E	106	K	E	E	E
					121	Y	51	H	H	Y	H
					127	E	57	E	D	D	D
					156	E	85	D	D	A	D
					62	G	69	S	T	T	S
					135	E	134	E	D	D	D
	rps5	Q9ST69	e	OA	77	K	67	R	R	R	Q
					141	S	131	T	T	T	S
					172	M	183*	L	L	M	L
					198	V	209	V	V	V	I
	rps6	P82403	f	PA	62	A	48	A	T	T	T
					149	V	134	L	V	V	V
					157	K	142	N	N	N	I
					162	A	147	A	E	E	E
Plastid	***rps11***	P06506	k	UA	6	P	6	L	P	P	P
					13	N	13	K	N	N	Y
					76	A	76	T	A	A	A
					78	N	78	D	D	N	D
					82	T	82	T	T	P	T
					98	P	95	P	S	S	P
					98	P	96	P	S	S	P
					98	P	98	P	S	S	P
					104	A	104	A	A	A	G
					108	A	108	V	A	A	A
					114	I	114	I	L	L	L
					116	L	116	L	L	L	V
	rps18	Q9M3K7	r	BB	13	R	14	R	R	R	Q
					18	R	17	H	R	R	R
					50	R	49	R	R	Q	R
					81	E	80	-	R	G	G
					81	E	82	-	R	G	G
					94	A	93	-	I	Q	I
	rps19	P06508	s	CB	19	I	19	M	M	M	I
					33	T	33	T	T	T	N
					65	R	65*	R	R	Y	D
					91	R	91	R	R	R	Q
	rps2	P08242	b	LA	24	T	24	I	I	T	I
	rps3	P09595	c	MA	79	G	79	G	A	G	G
					94	D	94	D	D	A	D
					103	L	103	L	L	F	L
					117	I	117	I	I	I	L
					213	I	213	I	I	I	L
	rps7	P82129	g	QA	151	F	151*	F	L	L	F

Highlighted green: different amino-acid between *S. oleracea* and *S. nutans* lineages
Highlighted purple: different amino-acid for one or more lineages of *S. nutans* compared to *S. oleracea*
Highlighted orange: different amino-acid for one lineage of *S. nutans* compared to the others and to *S. oleracea*
* : mutations identified as under positive selection in Postel et al, 2022.

b) Identification of impacted interactions between the subunits

To identify interactions between subunits inside the plastid ribosome, the structure was transformed into a RIN. The RIN is a graphical representation of the structure where nodes represents the residues and the edges the interactions between residues. To defined an interaction, one atom of a residue A must be at a distance between 2.5 and 5 Ångström (Å) of one atom of the residue B. Detected interactions are then exported into a text file with the two amino acids involved and the minimal distance between these two. From this file, only interactions between plastid and nuclear genes within each of the large and small ribosomal subunits were analyzed to identify potential impacting mutations.

c) Identification of the interaction type and possible modifications

To identify the type of interactions and the potential impact of the mutations on the interaction, mutations were modeled based on the spinach reference structure. To do that, the PyMOL software with the mutagenic tool was used (Delano *et al.*, 2002; Schrödinger *et al.* 2020) (40,41). This tool can replace an amino acid by another one by transforming the lateral chain. From this, an atomic point of view of the interaction can be deduced and the different types of interaction determined. The type of interaction can be one of the following: Hydrogen bond, hydrophobic, salt bridge and polar. The mutation can lead to a change in the type of interaction, a creation of a new one or a loss of the interaction. The type of interaction has been determined manually on the PyMOL interface. Given the results, we chose to focus on one plastid-nuclear gene pair: *rps11* (plastid encoded) - *rps21* (nuclear encoded) (cf Results).

d) Creation of the different models

As there are 4 different lineages (E1, W1, W2 and W3), we created 16 different models called E1_E1, E1_W1, E1_W2…W3_W2 and W3_W3 (*i.e.* one model per cross type and direction). Each model contained the associated mutations on the genes *rps11* and *rps21* described in table 2. The models were minimized using the YASARA software with YASARA minimization (Land *et al.* 2018) (42) (figure 1).

e) Creation of the Residue Interaction Networks (RINs) from the models and centrality analyses

RINs were created for each model for a total of 16 RINs using ringraph, an in-house C program which calculates distances between amino acids as described above (figure 1). From these networks, it is possible to calculate centralities of nodes thanks to graph theory. Centrality of a residue will represent residue that connect other residues together within a protein network (here within protein RPS11,

Table 3 – Summary of the types of centrality calculations used and their method.

CENTRALITY MEASURE	METHOD
Betweenness Centrality Analysis (BCA)	BCA highlights residues often found in the minimal path between every residue.
Closeness Centrality Analysis (CCA)	CCA is calculated as the reciprocal of the sum of the length of the shortest paths between the node and all other nodes in the graph
Degree Centrality Analysis (DCA)	DCA calculates centrality based on the number of nodes connected to the residue analyzed.
Eigenvector Centrality Analysis (ECA)	ECA calculates the centrality of nodes based on the centrality of other nodes meaning that a node connected to a high centrality node will have a higher centrality .
PageRank centrality Analysis (PRA)	PCA was first an algorithm developed for Google and its output is a probability distribution used to represent the likelihood that a person randomly clicking on links will arrive at any particular page

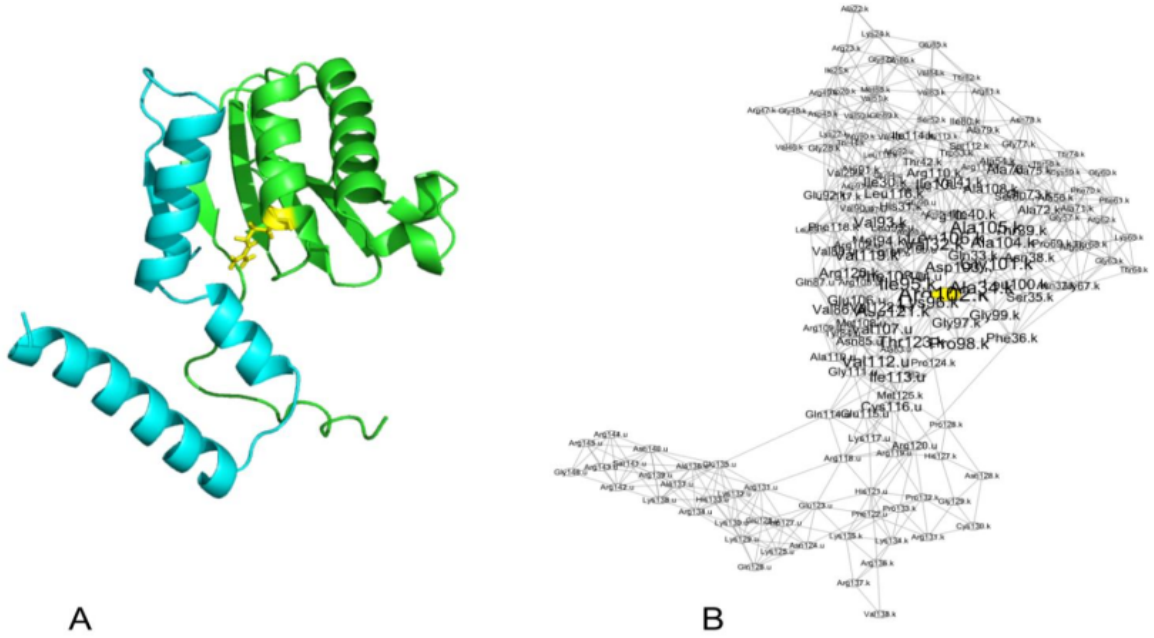


Figure 1 – Representation of the *rps11-rps21* genes as a structure (A) and as a RIN (B). The blue chain is *rps21* and the green chain is *rps11*. The yellow residue corresponds to the yellow node in the networks and corresponds to a residue with a centrality >= to 2. Visualization of the network has been made with Cytoscape after running RINspector (Shannon *et al.*, 2003; Brysbaert *et al.*, 2018) (44,45).

RPS21 and their interactions). The more a residue contributes to residue connection within a structure, the more it is central and has an important structural role. Different types of centrality can be calculated by the same in-house program. In the present study, we calculated the five centralities listed in table 3. We then calculated a centrality score, (i.e. a Z-score) which is normalized with the size of the network. A residue with a high Z-score ($\geq$2) is considered as central. The results of all centralities for the 16 models were retrieved and imported into a csv file. To have an overall comprehension, the ten more central residues in the interaction network of *rps11* and *rps21* were represented and analyzed (Figures 1 & 3).

f) Principal Component Analysis (PCA) on centralities

To see if modification of residue centralities associated with the cross type can explain the different outcomes of inter-lineages crosses and explain the differences in hybrids mortality, PCA were conducted using the centralities values of *rps11-rps21* residues for each cross type and the five different measure of centrality. Results being similar for the five centrality measures, we only reported results of the degree centrality measure. PCA has been run through an R script (R version 3.6.3) with RStudio and R packages (table.data V1.2.0 for data analysis and factoextra V1.0.7 for representation). PCA was calculated with the "prcomp" command (43).

3. Results and discussion

a) Modification of interactions due to mutation

Lots of lineage-specific NS mutations have been identified in interacting plastid and nuclear genes encoding the plastid ribosomes within *S. nutans* lineages (25). Mutation selection leads to a subset of 28 mutations with the associated modified interaction (Table 4). In total, we observed 8 losses of interaction, 4 gains of interaction and 16 mutations without a change in interaction type (Table 4). Some of these changes of interaction might be responsible for inter-lineage hybrid breakdown through disruption of co-adaptation between these plastid-nuclear gene pairs within the subunits of the plastid ribosome. Indeed, these modifications of residue contact interaction between plastid and nuclear genes within the plastid ribosome could alter the whole structure of the plastid ribosome. A large majority of the mutations inducing a change of the interaction between plastid and nuclear genes were located on genes *rps11* (plastid encoded) and *rps21* (nuclear encoded). This gene pair is also the one that contained most of the mutations (i.e. 28 in total) (table 2, table 4). We focused on this gene pair in subsequent analyses. For each *S. nutans* lineage cross direction, we build a three-dimensional model based on the spinach ribosome structure complex resolved in 2007. In total, 16 models have been created with the associated lineage mutation then minimized.

Table 4 – Detail of the interactions found between candidate gene pairs, with the impact of the mutation in one of the two partners on the interaction.

Interaction N°	Partner 1			Partner 2			Distance (Å)	Interaction type	
	Gene	Residue	Amino acid of E1 W1 W2 W3	Gene	Residue	Amino acid of E1 W1 W2 W3		Before mutation	After mutation
1	*rpl32*	Arg49	L L P L	*rpl17*	Tyr122	x	4.37	H bridge	Ø
2	*rpl14*	Arg104	G _ _ _	*rpl19*	Val155	x	4.17	Ø	Ø
3					Glu157	x	4.85	Ø	Ø
4					Ser163	x	2.98	polar	Ø
5					Tyr165	x	2.99	polar	Ø
6	*rps3*	Lys146	x	*rps5*	Val198	_ _ _ I	3.86	Ø	Ø
7	*rps11*	Pro98	_ S S _	*rps21*	Ile113	x	3.47	Ø	Ø
8					Cys116	x	3.86	Ø	Ø
9		Leu116*	_ _ _ V		Val90*	x	3.34	Ø	hydrophobic
10					Glu94	x	3.76	Ø	Ø
11					Leu99	x	3.53	hydrophobic	hydrophobic
12		Ser117	x		Val88	I _ _ _	4.57	polar	polar
13					Leu89	F _ _ S	3.41	Ø	Ø
14		Phe118*	x		Val88*	I _ _ _	3.73	Ø	hydrophobic
15					Leu89*	F _ _ S	3.44	hydrophobic	hydrophobic, Ø
16		Val119	x		Val88	I _ _ _	3.03	polar	polar
17		Pro132	x		Tyr121*	H H _ H	3.67	Ø	Ø
18		Pro133	x				3.49	Ø	Ø
19		Lys134*	x				3.51	Ø	polar
20		Lys135*	x				3.79	Ø	H bridge
21		Lys135	x		Glu127	_ D D D	3.12	salt bridge	salt bridge
22	*rps11*	Arg136	x		Tyr121	H H _ H	3.30	H bridge	H bridge
23	*rps13*	Arg124	x	*rps19*	Arg65	_ H Y D	2.64	H bridge	H bridge
24		Glu127	_ _ _ Q				2.75	salt bridge	Ø
25		Ile128	x				4.50	H bridge	Ø
26	*rps18*	Arg50	_ _ Q _	*rps21*	Arg139	x	3.60	H bridge	Ø
27					Asn140	x	4.12	H bridge	H bridge
28					Arg143	x	3.36	H bridge	Ø

Highlighted blue: no change of the type of the interaction with the mutation
Highlighted green: loss of the interaction with the mutation
Highlighted yellow: creation of a new interaction with the mutation
* : residue of *rps11-rps21* gene pairs for which centrality was calculated

b) RINs analysis of mutations for the *rps11-rps21* gene pair

We looked at the centralities of the mutated residues in *rps11-rps21* genes for the lineages of *S. nutans* (E1, W1, W2 and W3) to see if the mutations could impact the stability of the different subunits of the ribosome. The mutations were simulated from a visualization tool called PyMOL with a mutagenesis tool allowing to change the lateral chain of amino acids. Since these *in silico* mutations can change the conformation of the complex or at least the interface area, we performed energy minimization of the models, this allows us to have more realistic models but these are still predictive models which can add a bias to our analyses. To be sure we did not miss information, we decided to calculate 5 different kinds of accessibility: betweenness, closeness, degree, eigenvector and PageRank. We looked at the difference of centrality for the mutated residues inducing a change of the interaction between *rps11* and *rps21* (i.e. residue marked with a * in table 4). For each residue and each centrality measure, we retrieved a Z-score and looked at the difference of this score according to the different lineage cross type (Figure 2). The ten more central residues in the interaction network of *rps11* and *rps21* (i.e. not only the mutated ones) are represented in figure 3.

In most of the cases, the lineage which is the most impacting in terms of residue centrality is the lineage E1. Most of the time when the lineage E1 is involved in a cross, we can see a variation of the centralities of the residues notably with a change of centrality of the residue 117 (serine) (figure 2). When W3 is the maternal parent, we can see that the cysteine 116 loss its centrality (figure 2). We can also see with these results that the crosses between the lineages E1 and W3 impacted the centrality of the *rps11* alanine 34 (figure 3). Centrality of the residue Arginine 118 of *rps11* was also modified for the cross between lineages E1_W2, W1_E1, W3_E1 and W3_W2 (figure 3). For the same crosses, the residues glutamine 114 and acid glutamic 135 of the *rps21* exhibited a decrease in centrality (figure 3). Cross-specific modification of centrality is residue of genes *rps11* and *rps21* were identified. Especially when lineage E1 is involved and in plastid gene *rps11*. This lineage is also the one resulting in highest percentage of hybrid mortality. Though *rps21* does not seem to be an essential gene in the function of the plastid ribosome, *rps11* is (32). Modification of residues centralities in this essential gene in cross with lineage E1 might contribute to a modification of the protein network interaction and to an elevated amount of hybrid morality. We also observed differences in centrality when lineage W3 is the mother, again in gene *rps11.* Though W3 does not lead to high percentage of hybrid mortality when

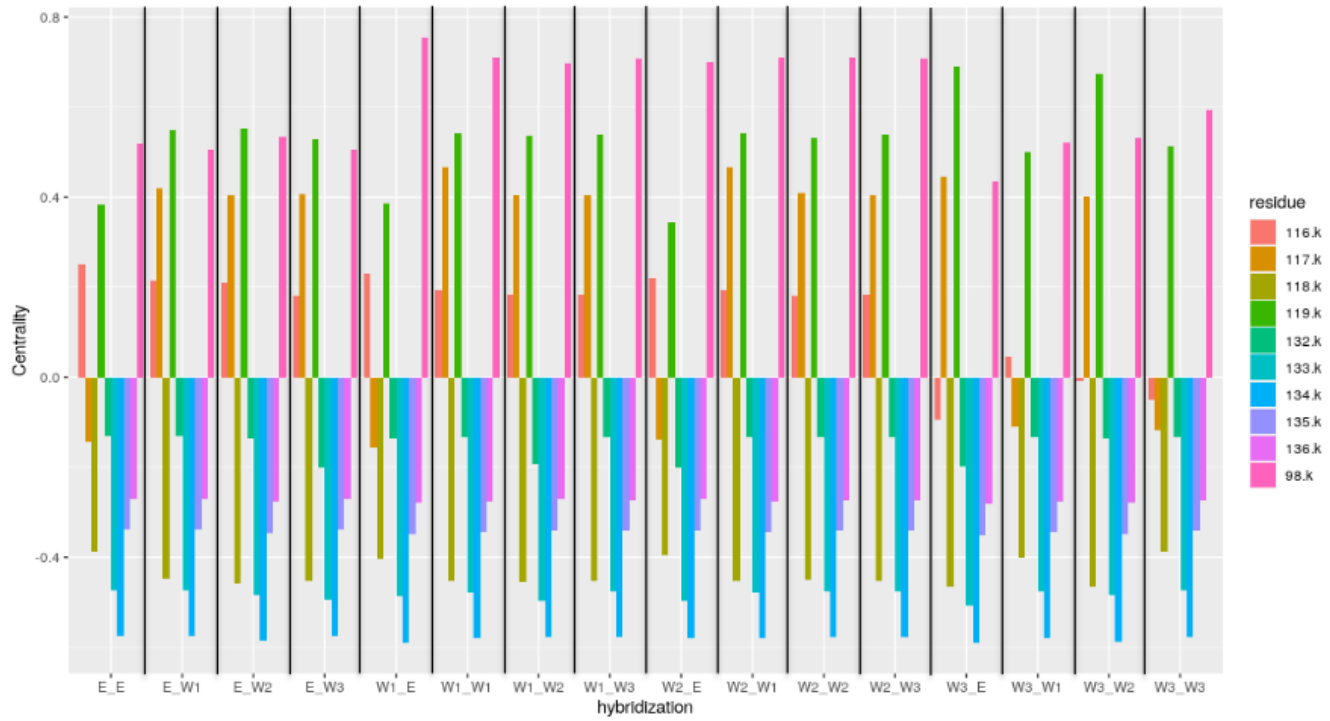


Figure 2 - BCA Centrality score of mutated residues of *rps11* according to the 16 different hybridizations. "k" corresponds to the *rps11*

used in inter-lineages crosses, it could nevertheless impact the whole small subunit ribosomal structure and impact its function. Overall, the different centrality calculations showed a loss or a gain of centrality according to the different mutations and crosses, but there is no clear signal. Especially, this does not seem to follow the associated amount of hybrid mortality. One hypothesis to explain that might be that the centrality moves to another residue not listed in the mutations identified here or that we only focused on one gene pair, the most mutated one while other plastid-nuclear gene pair, even though less mutated, could also generate modification of interaction and centrality that could impact the structure and function of the plastid ribosome. Also, not all of these centrality measures are the most used in this kind of non-oriented networks (i.e. we do not consider the impact of one residue on another one, directionally but interactions between these two), in particular PageRank which is more used for oriented graphs. This might explain the lack of power and strong signals when looking at residues' centrality.

c) Component Analysis (PCA)

To go further and see whether changes in centrality in genes *rps11-rps21* might correlates with cross level of hybrid mortality, we looked at the distribution of the centrality of every residue of this gene

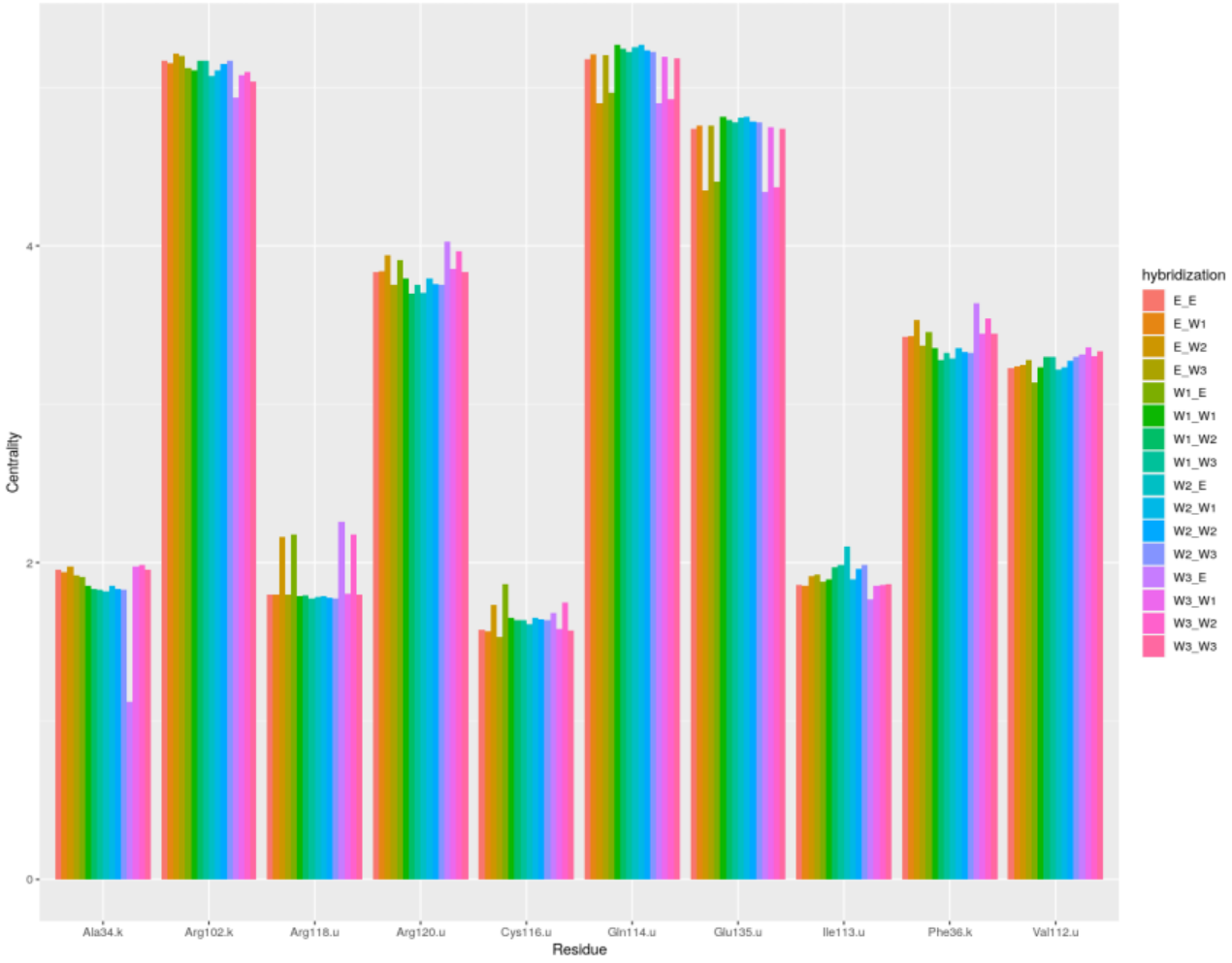


Figure 3 - Betweenness centrality (BCA) of the ten most central residues in function of the different hybridization between the four lineages. "k" corresponds to the *rps11* and "u" to the *rps21*.

pairs for the 16 models with Principal Component Analysis. We only show the results with on measure of centrality : degree centrality as the results are similar with the other four.

The two first dimensions of the PCA explained between 51% and 83.9% of the variance which are good results (figure 4 – A). Centralities associated with the different cross type can be discriminated on these two dimensions (figure 4 – B). However, regarding the mean point for the three different classes (lethal, medium and non-lethal) we can see that they are near the center of the axes meaning that they are not well discriminating on these two PCA dimensions. If we look more closely, the amount of mortality associated with these crosses cannot be discriminated between lethal, medium or non-lethal hybridization on these two dimensions (figure 4 – B). Three medium mortality and one high mortality can be observed in the top right of the plot but there is some other elevated mortality cross type are

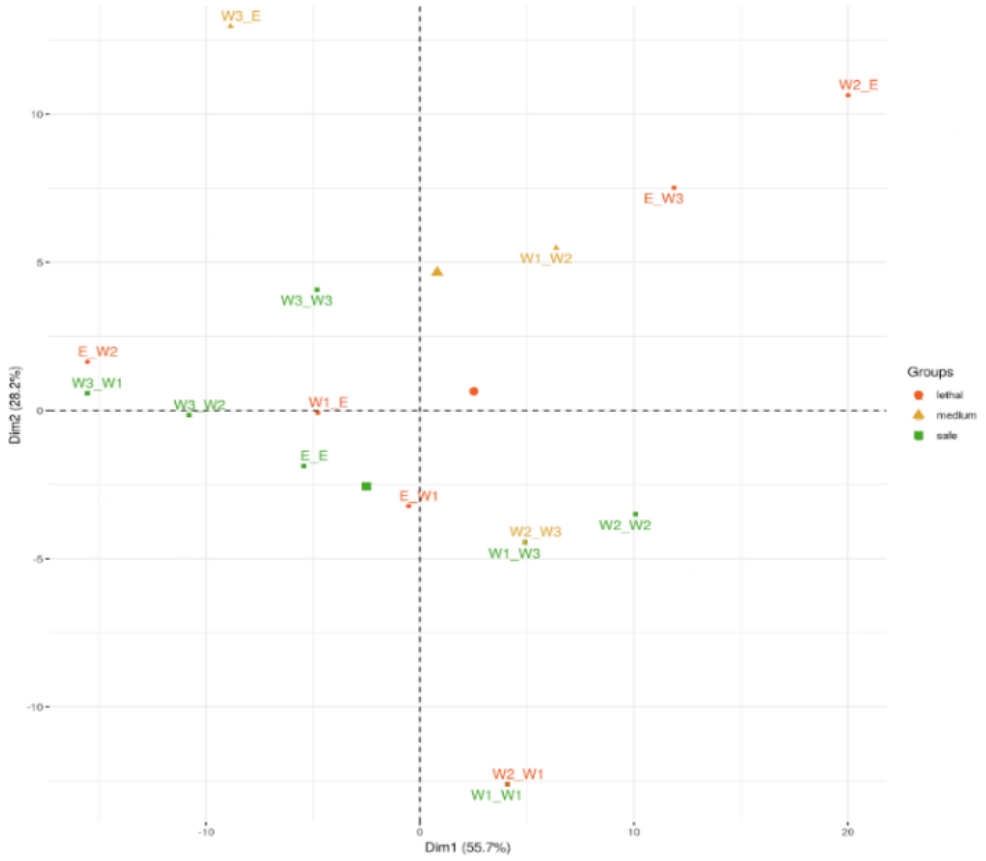


Figure 4 – Principal Component Analysis of Degree Centralities of residues in the *rps11-rps21* genes. Representation of the 16 cross types on the two main dimensions. In green are cross outcomes with a low percent of hybrid mortality (<10%), in orange the one with a medium hybrid mortality (10% >= and <= 80 %) and in red the ones with high hybrid mortality (> 80%). The three biggest points represent the mean coordinates of a group.

also found among the low mortality crosses (figure 4 – B). With the four other centrality measures, the strongly lethal cross types are also mixed with non-lethal ones. Yet, these two dimensions seem to discriminate reciprocal crosses. For example, when looking at cross between lineage E1 and W2 (E1/W2 and W2/E1), the centrality associated with these two directions are not found on the same sides of the PC1 and PC2: E1/W2 is on the top quarter left while W2/E1 is on the top quarter right. This is also true when looking at centrality associated with crosses W3/W1 and W1/W3: the former is near PC1 on top quarter left while the later is near PC2 on bottom quarter right. More generally, we can observe this kind of asymmetry in location along the PC1 and PC2 for all the reciprocal crosses, suggesting modification of centrality differently depending on the cross direction.

Lack of signal to discriminate between strongly lethal and less lethal cross type might (1) suggest these two genes might not be the main driver of plastid ribosome non function in hybrids as modification of centrality associated with the cross type does not seem to correlate with the crosses' degree of hybrid mortality and/or (2) might come from the fact that we are only looking at one plastid-nuclear gene pairs and not the whole plastid ribosome, yet even though centrality modification associated with mutations in *rps11* and *rps21* is not discriminating how lethal are the crosses, it can influence the stability of the whole structure of the plastid ribosome and have an impact on its function.

Conclusion

The results of this study showed that some mutations impacted the interactions between proteins in the plastid ribosome, potentially modifying the whole structure of the plastid ribosome and its function in inter-lineages hybrids. Several mutations modified the interactions between plastid and nuclear genes, either within the large or small subunits of the plastid ribosome. We focused on the most mutated gene pair: *rps11-rps21*. We showed that mutations associated with lineage E1 impacted the centrality of several residues potentially leading to a change of the interaction between these genes and driving the high hybrid mortality when use in inter-lineages cross. Some residues of *rps11* and *rps21* reacted the same way for four other cross type (E_W2, W1_E, W3_E and W3_W2). Modification of structure through mutations in one lineage could result in drastic changes of, if not the whole ribosome complex, interactions with its normally interacting nuclear gene *rps21*. If key interactions are disrupted, this could have subsequent consequences on the translation of photosynthetic proteins, which are essential to plastid function and plant development (3). Overall, centrality modification of residues association with each cross type cannot explain the differences in hybrid mortality observed for each cross.

In the present study, to gain time, we only focused on one plastid-nuclear gene pairs that contained most of the modified interactions. Yet the other mutations might also modify and disrupt the ribosomes structure, and the strength of the functional impact of a mutation and its associated structure modification might not follow a linear tendency: few mutations impacting essential or central residues might also have strong functional consequences. It could be interesting to perform the same analyses but using the whole set of ribosomal mutations, inducing changes of interactions between plastid-nuclear gene pairs identified in Table 3 and see if (i) these mutations change the centrality of highlighted residues in the whole structure and can explain better the lethality during hybridization and (ii) if they generate an impact on the interaction with the RNAs.

References

1. Gray MW. Evolution of organellar genomes. Curr Opin Genet Dev. 1999;9:678–87.

2. Sloan DB, Warren JM, Williams AM, Wu Z, Abdel-Ghany SE, Chicco AJ, et al. Cytonuclear integration and co-evolution. Nat Rev Genet. 2018;19(10):635–48.
3. Zoschke R, Bock R. Chloroplast Translation: Structural and Functional Organization, Operational Control, and Regulation. Plant Cell. 2018;30:745–70.
4. Greiner S, Bock R. Tuning a menage a trois: Co-evolution and co-adaptation of nuclear and organellar genomes in plants. BioEssays. 2013;35:354–65.
5. Zhang J, Ruhlman TA, Sabir J, Blazier JC, Jansen RK. Coordinated Rates of Evolution between Interacting Plastid and Nuclear Genes in Geraniaceae. Plant Cell. 2015;27(3):563–73.
6. Rand DM, Haney RA, Fry AJ. Cytonuclear coevolution: the genomics of cooperation. Trends Ecol Evol. 2004;19(12):645–53.
7. Smith DR. Mutation Rates in Plastid Genomes: They Are Lower than You Might Think. Genome Biol Evol. 2015;7(5):1227–34.
8. Greiner S, Sobanski J, Bock R. Why are most organelle genomes transmitted maternally? Bioessays. 2014;37:80–94.
9. Forsythe ES, Williams AM, Sloan DB. Genome-wide signatures of plastid-nuclear coevolution point to repeated perturbations of plastid proteostasis systems across angiosperms. Plant Cell. 2021;33(4):980–97.
10. Levin DA. The Cytoplasmic Factor in Plant Speciation. Syst Bot. 2003;28:8.
11. Greiner S, Rauwolf U, Meurer J, Herrmann RG. The role of plastids in plant speciation. Mol Ecol. 2011;20(4):671–91.
12. Barnard-Kubow KB, So N, Galloway LF. Cytonuclear incompatibility contributes to the early stages of speciation. Evolution. 2016;70(12):2752–66.
13. Matute DR, Cooper BS. Comparative studies on speciation: 30 years since Coyne and Orr. Evolution. 2021;75(4):764–78.
14. Burton RS, Barreto FS. A disproportionate role for mtDNA in Dobzhansky-Muller incompatibilities? Mol Ecol. 2012;21(20):4942–57.
15. Turelli M, Moyle LC. Asymmetric Postmating Isolation: Darwin's Corollary to Haldane's Rule. Genetics. 2007;176(2):1059–88.
16. Barnard-Kubow KB, McCoy MA, Galloway LF. Biparental chloroplast inheritance leads to rescue from cytonuclear incompatibility. New Phytol. 2017;213:1466–76.
17. Zupok A, Kozul D, Schöttler MA, Niehörster J, Garbsch F, Liere K, et al. A photosynthesis operon in the chloroplast genome drives speciation in evening primroses. Plant Cell. 2021;33(8):2583–601.
18. Ellison CK, Burton RS. DISRUPTION OF MITOCHONDRIAL FUNCTION IN INTERPOPULATION HYBRIDS OF *TIGRIOPUS CALIFORNICUS*. Evolution. 2006;60(7):1382–91.
19. Bogdanova VS, Galieva ER, Kosterin OE. Genetic analysis of nuclear-cytoplasmic incompatibility

in pea associated with cytoplasm of an accession of wild subspecies *Pisum sativum* subsp. *elatius* (Bieb.) Schmahl. Theor Appl Genet. 2009;118:9.

20. Rockenbach K, Havird JC, Monroe JG, Triant DA, Taylor DR, Sloan DB. Positive Selection in Rapidly Evolving Plastid–Nuclear Enzyme Complexes. Genetics. 2016;204(4):1507–22.
21. Sloan DB, Triant DA, Forrester NJ, Bergner LM, Wu M, Taylor DR. A recurring syndrome of accelerated plastid genome evolution in the angiosperm tribe *Sileneae* (Caryophyllaceae). Mol Phylogenet Evol. 2014;72:82–9.
22. Postel Z, Touzet P. Cytonuclear Genetic Incompatibilities in Plant Speciation. Plants. 2020;9(487).
23. Sharbrough J, Conover JL, Tate JA, Wendel JF, Sloan DB. Cytonuclear responses to genome doubling. Am J Bot. 2017;104(9):1277–80.
24. Weng ML, Ruhlman TA, Jansen RK. Plastid–Nuclear Interaction and Accelerated Coevolution in Plastid Ribosomal Genes in Geraniaceae. Genome Biol Evol. 2016;8(6):1824–38.
25. Postel Z, Poux C, Gallina S, Varré JS, Godé C, Schmitt E, et al. Reproductive isolation among lineages of *Silene nutans* (Caryophyllaceae): A potential involvement of plastid-nuclear incompatibilities. Mol Phylogenet Evol. 2022;169:107436.
26. Martin H, Touzet P, Van Rossum F, Delalande D, Arnaud JF. Phylogeographic pattern of range expansion provides evidence for cryptic species lineages in *Silene nutans* in Western Europe. Heredity. 2016;116(3):286–94.
27. Van Rossum F, Martin H, Le Cadre S, Brachi B, Christenhusz MJM, Touzet P. Phylogeography of a widely distributed species reveals a cryptic assemblage of distinct genetic lineages needing separate conservation strategies. Perspect Plant Ecol Evol Syst. 2018;35:44–51.
28. Martin H, Touzet P, Dufay M, Godé C, Schmitt E, Lahiani E, et al. Lineages of *Silene nutans* developed rapid, strong, asymmetric postzygotic reproductive isolation in allopatry. Evolution. 2017;71(6):1519–31.
29. Bieri P, Leibundgut M, Saurer M, Boehringer D, Ban N. The complete structure of the chloroplast 70S ribosome in complex with translation factor pY. EMBO J. 2017;36(4):12.
30. Havird JC, Whitehill NS, Snow CD, Sloan DB. Conservative and compensatory evolution in oxidative phosphorylation complexes of angiosperms with highly divergent rates of mitochondrial genome evolution. Evolution. 2015;69(12):3069–81.
31. Tiller N, Weingartner M, Thiele W, Maximova E, Schöttler MA, Bock R. The plastid-specific ribosomal proteins of *Arabidopsis thaliana* can be divided into non-essential proteins and genuine ribosomal proteins. Plant J. 2012;69(2):302–16.
32. Tiller N, Bock R. The Translational Apparatus of Plastids and Its Role in Plant Development. Mol Plant. 2014;7(7):1105–20.

33. Tang X, Wang Y, Zhang Y, Huang S, Liu, Z, Fei D, et al. A missense mutation of plastid RPS4 is associated with chlorophyll deficiency in Chinese cabbage (*Brassica campestris ssp. pekinensis*). BMC Plant Biol. 2018;18(130):11.

34. Sharma MR, Wilson DN, Datta PP, Barat C, Schluenzen F, Fucini P, et al. Cryo-EM study of the spinach chloroplast ribosome reveals the structural and functional roles of plastid-specific ribosomal proteins. Proc Natl Acad Sci. 2007;104(49):19315–20.

35. Brysbaert G, Lorgouilloux K, Vranken WF, Lensink MF. RINspector: a Cytoscape app for centrality analyses and DynaMine flexibility prediction. Bioinforma Oxf Engl. 2018 Jan 15;34(2):294–6.

36. del Sol A, Fujihashi H, Amoros D, Nussinov R. Residue centrality, functionally important residues, and active site shape: analysis of enzyme and non-enzyme families. Protein Sci Publ Protein Soc. 2006 Sep;15(9):2120–8.

37. Hu G, Yan W, Zhou J, Shen B. Residue interaction network analysis of Dronpa and a DNA clamp. J Theor Biol. 2014 May 7;348:55–64.

38. Trouvilliez S, Cicero J, Lévêque R, Aubert L, Corbet C, Van Outryve A, et al. Direct interaction of TrkA/CD44v3 is essential for NGF-promoted aggressiveness of breast cancer cells. J Exp Clin Cancer Res CR. 2022 Mar 28;41(1):110.

39. Sharma MR, Wilson DN, Datta PP, Barat C, Schluenzen F, Fucini P, et al. Cryo-EM study of the spinach chloroplast ribosome reveals the structural and functional roles of plastid-specific ribosomal proteins. Proc Natl Acad Sci U S A. 2007 Dec 4;104(49):19315–20.

40. DeLano, W.L. The PyMOL Molecular Graphics System. Delano Scientific, San Carlos.; 2002.

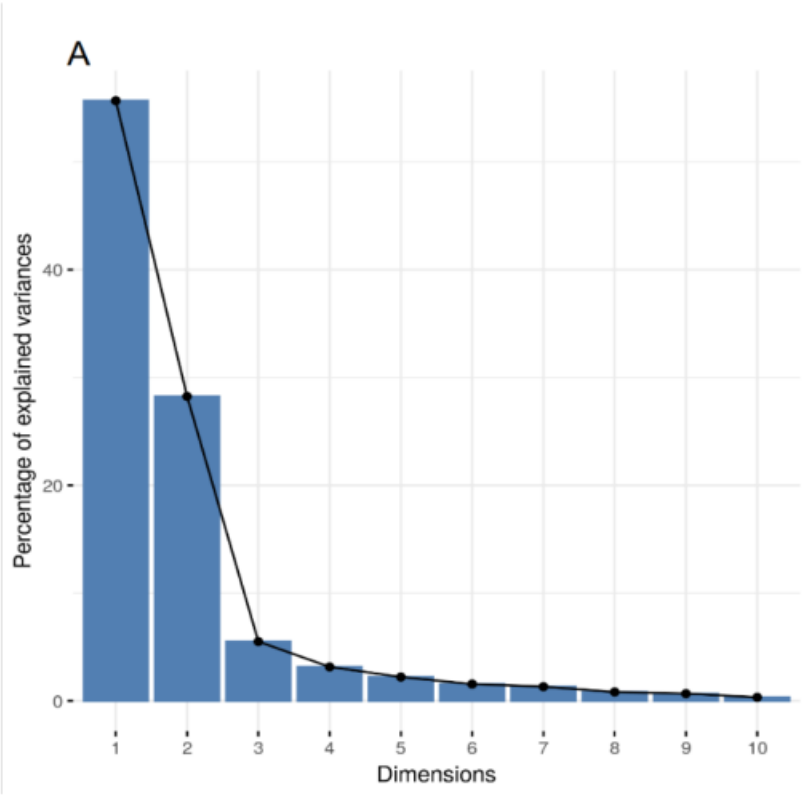
41. Schrödinger, LLC, Warren DeLano. PyMOL [Internet]. 2020. Available from: http://www.pymol.org/pymol

42. Land H, Humble MS. YASARA: A Tool to Obtain Structural Guidance in Biocatalytic Investigations. Methods Mol Biol Clifton NJ. 2018;1685:43–67.

43. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2018. Available from: https://www.R-project.org/

44. Brysbaert G, Mauri T, Lensink MF. Comparing protein structures with RINspector automation in Cytoscape. F1000Research. 2018;7:563.

45. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Res. 2003 Nov 1;13(11):2498–504.



Suppl. Figure 1 – Principal Component Analysis of Degree Centralities of residues in the *rps11-rps21* genes.
Percentage of variance explained by the different dimensions.

B. Xperium

All along my second and third year of PhD I had the chance to participate in Xperium. Xperium is a showcase of research made in Lille University. With 8 stands, Xperium allows middle and high school students to see what university research looks like and help them understand the importance of fundamental research to application in society. Xperium also opens its doors to the general public in some events as “Fête de la Science”. Xperium is divided into seasons of two years. The one I participated was named “Kaleidoscope” and

talked about the vision in general in many scientific disciplines, going from mathematics to law while passing through the study of the death penalty under the old regime.

I participated in the stand creation with researchers of my lab and then had a 64h contract per year to present vulgarized biological science to the students. My subject was on the use of chemicals to visualize the living and more precisely the use of fluorescence to determine the composition of complex molecules such as lignin. A short footage of 5 minutes is available at this link. Usually, a presentation was 15 minutes long with an additional 5 minutes for questions. During a session, about 4 presentations are performed.

This parallel project enhanced my pleasure for the vulgarization and the knowledge sharing with students and people out of the research universe, a pleasure I discovered during the several instances of the “Fête de la Science” I participated in.

X. References

1. Eisenberg D. The discovery of the alpha-helix and beta-sheet, the principal structural features of proteins. *Proc Natl Acad Sci U S A*. 2003;100(20):11207-11210. doi:10.1073/pnas.2034522100
2. Shapovalov M, Vucetic S, Dunbrack RL. A new clustering and nomenclature for beta turns derived from high-resolution protein structures. *PLoS Comput Biol*. 2019;15(3):e1006844. doi:10.1371/journal.pcbi.1006844
3. Romero P, Obradovic Z, Dunker AK. Natively Disordered Proteins: Functions and Predictions. *Appl Bioinformatics*. 2004;3(2):105-113. doi:10.2165/00822942-200403020-00005
4. Sjölin L. 3D-structural elucidation of biologically important macromolecules. *Drug Des Discov*. 1993;9(3-4):261-276.
5. Zhou ZH. Towards atomic resolution structural determination by single-particle cryo-electron microscopy. *Curr Opin Struct Biol*. 2008;18(2):218-228. doi:10.1016/j.sbi.2008.03.004
6. Schaefer A, Naser D, Siebeneichler B, Tarasca MV, Meiering EM. Methodological advances and strategies for high resolution structure determination of cellular protein aggregates. *J Biol Chem*. Published online June 24, 2022:102197. doi:10.1016/j.jbc.2022.102197
7. Pereira J, Simpkin AJ, Hartmann MD, Rigden DJ, Keegan RM, Lupas AN. High-accuracy protein structure prediction in CASP14. *Proteins Struct Funct Bioinforma*. 2021;89(12):1687-1699. doi:10.1002/prot.26171
8. Xu J, Wang S. Analysis of distance-based protein structure prediction by deep learning in CASP13. *Proteins Struct Funct Bioinforma*. 2019;87(12):1069-1081. doi:10.1002/prot.25810
9. Jumper J, Evans R, Pritzel A, et al. Applying and improving AlphaFold at CASP14. *Proteins*. 2021;89(12):1711-1721. doi:10.1002/prot.26257
10. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596(7873):583-589. doi:10.1038/s41586-021-03819-2
11. Rabbani G, Baig MH, Ahmad K, Choi I. Protein-protein Interactions and their Role in Various Diseases and their Prediction Techniques. *Curr Protein Pept Sci*. 2018;19(10):948-957. doi:10.2174/1389203718666170828122927
12. Sun J, Zhao Z. A comparative study of cancer proteins in the human protein-protein interaction network. *BMC Genomics*. 2010;11 Suppl 3:S5. doi:10.1186/1471-2164-11-S3-S5
13. Rao VS, Srinivas K, Sujini GN, Kumar GNS. Protein-Protein Interaction Detection: Methods and Analysis. *Int J Proteomics*. 2014;2014:1-12. doi:10.1155/2014/147648
14. Aderinwale T, Christoffer CW, Sarkar D, Alnabati E, Kihara D. Computational structure modeling for diverse categories of macromolecular interactions. *Curr Opin Struct Biol*. 2020;64:1-8. doi:10.1016/j.sbi.2020.05.017
15. Leutert M, Entwisle SW, Villén J. Decoding Post-Translational Modification Crosstalk With Proteomics. *Mol Cell Proteomics MCP*. 2021;20:100129. doi:10.1016/j.mcpro.2021.100129
16. Walsh CT, Garneau-Tsodikova S, Gatto GJ. Protein posttranslational modifications: the chemistry of proteome diversifications. *Angew Chem Int Ed Engl*. 2005;44(45):7342-7372. doi:10.1002/anie.200501023
17. Inagi R. Organelle stress and glycation in kidney disease. *Glycoconj J*. 2021;38(3):341-346. doi:10.1007/s10719-021-09989-5
18. Jaisson S, Pietrement C, Gillery P. Carbamylation-Derived Products: Bioactive Compounds and Potential Biomarkers in Chronic Renal Failure and Atherosclerosis. *Clin Chem*. 2011;57(11):1499-1505. doi:10.1373/clinchem.2011.163188
19. Akagawa M. Protein carbonylation: molecular mechanisms, biological implications, and

analytical approaches. *Free Radic Res*. 2021;55(4):307-320. doi:10.1080/10715762.2020.1851027
20. Kang HJ, Baker EN. Intramolecular isopeptide bonds: protein crosslinks built for stress? *Trends Biochem Sci*. 2011;36(4):229-237. doi:10.1016/j.tibs.2010.09.007
21. Singh V, Ram M, Kumar R, Prasad R, Roy BK, Singh KK. Phosphorylation: Implications in Cancer. *Protein J*. 2017;36(1):1-6. doi:10.1007/s10930-017-9696-z
22. Hunter T. Protein kinases and phosphatases: the yin and yang of protein phosphorylation and signaling. *Cell*. 1995;80(2):225-236. doi:10.1016/0092-8674(95)90405-0
23. Cohen P. The regulation of protein function by multisite phosphorylation--a 25 year update. *Trends Biochem Sci*. 2000;25(12):596-601. doi:10.1016/s0968-0004(00)01712-6
24. Laarse SAM, Leney AC, Heck AJR. Crosstalk between phosphorylation and O-GlcNAcylation: friend or foe. *FEBS J*. 2018;285(17):3152-3167. doi:10.1111/febs.14491
25. Luo F, Wang M, Liu Y, Zhao XM, Li A. DeepPhos: prediction of protein phosphorylation sites with deep learning. *Bioinforma Oxf Engl*. 2019;35(16):2766-2773. doi:10.1093/bioinformatics/bty1051
26. Wang Z, Gucek M, Hart GW. Cross-talk between GlcNAcylation and phosphorylation: site-specific phosphorylation dynamics in response to globally elevated O-GlcNAc. *Proc Natl Acad Sci U S A*. 2008;105(37):13793-13798. doi:10.1073/pnas.0806216105
27. Yang X, Qian K. Protein O-GlcNAcylation: emerging mechanisms and functions. *Nat Rev Mol Cell Biol*. 2017;18(7):452-465. doi:10.1038/nrm.2017.22
28. Biwi J, Biot C, Guerardel Y, Vercoutter-Edouart AS, Lefebvre T. The Many Ways by Which O-GlcNAcylation May Orchestrate the Diversity of Complex Glycosylations. *Mol Basel Switz*. 2018;23(11):E2858. doi:10.3390/molecules23112858
29. Very N, Vercoutter-Edouart AS, Lefebvre T, Hardivillé S, El Yazidi-Belkoura I. Cross-Dysregulation of O-GlcNAcylation and PI3K/AKT/mTOR Axis in Human Chronic Diseases. *Front Endocrinol*. 2018;9:602. doi:10.3389/fendo.2018.00602
30. Lazarus MB, Nam Y, Jiang J, Sliz P, Walker S. Structure of human O-GlcNAc transferase and its complex with a peptide substrate. *Nature*. 2011;469(7331):564-567. doi:10.1038/nature09638
31. Aquino-Gil M, Pierce A, Perez-Cervera Y, Zenteno E, Lefebvre T. OGT: a short overview of an enzyme standing out from usual glycosyltransferases. *Biochem Soc Trans*. 2017;45(2):365-370. doi:10.1042/BST20160404
32. Varki A, Cummings RD, Esko JD, et al., eds. *Essentials of Glycobiology*. 2nd ed. Cold Spring Harbor Laboratory Press; 2009. Accessed July 18, 2022. http://www.ncbi.nlm.nih.gov/books/NBK1908/
33. Bond MR, Hanover JA. A little sugar goes a long way: the cell biology of O-GlcNAc. *J Cell Biol*. 2015;208(7):869-880. doi:10.1083/jcb.201501101
34. Jia C, Zuo Y, Zou Q. O-GlcNAcPRED-II: an integrated classification algorithm for identifying O-GlcNAcylation sites based on fuzzy undersampling and a K-means PCA oversampling technique. *Bioinforma Oxf Engl*. 2018;34(12):2029-2036. doi:10.1093/bioinformatics/bty039
35. Gupta R, Brunak S. Prediction of glycosylation across the human proteome and the correlation to protein function. *Pac Symp Biocomput Pac Symp Biocomput*. Published online 2002:310-322.
36. Kao HJ, Huang CH, Bretaña NA, et al. A two-layered machine learning method to identify protein O-GlcNAcylation sites with O-GlcNAc transferase substrate motifs. *BMC Bioinformatics*. 2015;16 Suppl 18:S10. doi:10.1186/1471-2105-16-S18-S10
37. Levine ZG, Fan C, Melicher MS, Orman M, Benjamin T, Walker S. O-GlcNAc Transferase Recognizes Protein Substrates Using an Asparagine Ladder in the Tetratricopeptide Repeat (TPR) Superhelix. *J Am Chem Soc*. 2018;140(10):3510-3513. doi:10.1021/jacs.7b13546
38. Bhuiyan T, Waridel P, Kapuria V, Zoete V, Herr W. Distinct OGT-Binding Sites Promote HCF-1 Cleavage. *PloS One*. 2015;10(8):e0136636. doi:10.1371/journal.pone.0136636

39. Capotosti F, Guernier S, Lammers F, et al. O-GlcNAc transferase catalyzes site-specific proteolysis of HCF-1. *Cell*. 2011;144(3):376-388. doi:10.1016/j.cell.2010.12.030
40. Wodak SJ, Janin J. Structural basis of macromolecular recognition. *Adv Protein Chem*. 2002;61:9-73. doi:10.1016/s0065-3233(02)61001-0
41. Lo Conte L, Chothia C, Janin J. The atomic structure of protein-protein recognition sites. *J Mol Biol*. 1999;285(5):2177-2198. doi:10.1006/jmbi.1998.2439
42. Richards FM. The interpretation of protein structures: total volume, group volume distributions and packing density. *J Mol Biol*. 1974;82(1):1-14. doi:10.1016/0022-2836(74)90570-1
43. Finney JL. Volume occupation, environment and accessibility in proteins. The problem of the protein surface. *J Mol Biol*. 1975;96(4):721-732. doi:10.1016/0022-2836(75)90148-5
44. Janin J, Chothia C. Stability and specificity of protein-protein interactions: the case of the trypsin-trypsin inhibitor complexes. *J Mol Biol*. 1976;100(2):197-211. doi:10.1016/s0022-2836(76)80148-9
45. Wodak SJ, Janin J. Computer analysis of protein-protein interaction. *J Mol Biol*. 1978;124(2):323-342. doi:10.1016/0022-2836(78)90302-9
46. Nussinov R, Papin JA, Vakser I. Computing the Dynamic Supramolecular Structural Proteome. *PLoS Comput Biol*. 2017;13(1):e1005290. doi:10.1371/journal.pcbi.1005290
47. Crippen GM. Conformational analysis by energy embedding. *J Comput Chem*. 1982;3(4):471-476. doi:10.1002/jcc.540030404
48. S.Schelstraete, Schepens W, Verschelde H. Energy minimization by smoothing techniques: a survey. In: *Theoretical and Computational Chemistry*. Vol 7. Elsevier; 1999:129-185. doi:10.1016/S1380-7323(99)80038-7
49. Noguti T, Go N. Efficient Monte Carlo method for simulation of fluctuating conformations of native proteins. *Biopolymers*. 1985;24(3):527-546. doi:10.1002/bip.360240308
50. Morris GM, Huey R, Lindstrom W, et al. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J Comput Chem*. 2009;30(16):2785-2791. doi:10.1002/jcc.21256
51. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem*. 2010;31(2):455-461. doi:10.1002/jcc.21334
52. Jones G, Willett P, Glen RC, Leach AR, Taylor R. Development and validation of a genetic algorithm for flexible docking. *J Mol Biol*. 1997;267(3):727-748. doi:10.1006/jmbi.1996.0897
53. Wu G, Robertson DH, Brooks CL, Vieth M. Detailed analysis of grid-based molecular docking: A case study of CDOCKER-A CHARMm-based MD docking algorithm. *J Comput Chem*. 2003;24(13):1549-1562. doi:10.1002/jcc.10306
54. Kramer B, Rarey M, Lengauer T. Evaluation of the FLEXX incremental construction algorithm for protein-ligand docking. *Proteins*. 1999;37(2):228-241. doi:10.1002/(sici)1097-0134(19991101)37:2<228::aid-prot8>3.0.co;2-8
55. Jain AN. Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. *J Med Chem*. 2003;46(4):499-511. doi:10.1021/jm020406h
56. Friesner RA, Banks JL, Murphy RB, et al. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J Med Chem*. 2004;47(7):1739-1749. doi:10.1021/jm0306430
57. Allen WJ, Balius TE, Mukherjee S, et al. DOCK 6: Impact of new features and current docking performance. *J Comput Chem*. 2015;36(15):1132-1156. doi:10.1002/jcc.23905
58. Grosdidier A, Zoete V, Michielin O. SwissDock, a protein-small molecule docking web service based on EADock DSS. *Nucleic Acids Res*. 2011;39(Web Server issue):W270-277. doi:10.1093/nar/gkr366
59. De Vivo M, Masetti M, Bottegoni G, Cavalli A. Role of Molecular Dynamics and Related Methods in Drug Discovery. *J Med Chem*. 2016;59(9):4035-4061. doi:10.1021/acs.jmedchem.5b01684

60. Gurung AB, Ali MA, Lee J, Farah MA, Al-Anazi KM. An Updated Review of Computer-Aided Drug Design and Its Application to COVID-19. Alatas B, ed. *BioMed Res Int*. 2021;2021:1-18. doi:10.1155/2021/8853056
61. Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJC. GROMACS: fast, flexible, and free. *J Comput Chem*. 2005;26(16):1701-1718. doi:10.1002/jcc.20291
62. Case DA, Cheatham TE, Darden T, et al. The Amber biomolecular simulation programs. *J Comput Chem*. 2005;26(16):1668-1688. doi:10.1002/jcc.20290
63. Brooks BR, Brooks CL, Mackerell AD, et al. CHARMM: the biomolecular simulation program. *J Comput Chem*. 2009;30(10):1545-1614. doi:10.1002/jcc.21287
64. Rackers JA, Wang Z, Lu C, et al. Tinker 8: Software Tools for Molecular Design. *J Chem Theory Comput*. 2018;14(10):5273-5289. doi:10.1021/acs.jctc.8b00529
65. Smith W, Yong CW, Rodger PM. DL_POLY: Application to molecular simulation. *Mol Simul*. 2002;28(5):385-471. doi:10.1080/08927020290018769
66. Heideman M, Johnson D, Burrus C. Gauss and the history of the fast fourier transform. *IEEE ASSP Mag*. 1984;1(4):14-21. doi:10.1109/MASSP.1984.1162257
67. Van Loan C. *Computational Frameworks for the Fast Fourier Transform*. Society for Industrial and Applied Mathematics; 1992. doi:10.1137/1.9781611970999
68. Wang C, Wei Y, Zhang H, et al. Constructing effective energy functions for protein structure prediction through broadening attraction-basin and reverse Monte Carlo sampling. *BMC Bioinformatics*. 2019;20(S3):135. doi:10.1186/s12859-019-2652-5
69. Porter KA, Desta I, Kozakov D, Vajda S. What method to use for protein-protein docking? *Curr Opin Struct Biol*. 2019;55:1-7. doi:10.1016/j.sbi.2018.12.010
70. Andrusier N, Mashiach E, Nussinov R, Wolfson HJ. Principles of flexible protein-protein docking. *Proteins Struct Funct Bioinforma*. 2008;73(2):271-289. doi:10.1002/prot.22170
71. Desta IT, Porter KA, Xia B, Kozakov D, Vajda S. Performance and Its Limits in Rigid Body Protein-Protein Docking. *Structure*. 2020;28(9):1071-1081.e3. doi:10.1016/j.str.2020.06.006
72. Huang SY, Zou X. MDockPP: A hierarchical approach for protein-protein docking and its application to CAPRI rounds 15-19. *Proteins*. 2010;78(15):3096-3103. doi:10.1002/prot.22797
73. Glashagen G, de Vries S, Uciechowska-Kaczmarzyk U, et al. Coarse-grained and atomic resolution biomolecular docking with the ATTRACT approach. *Proteins*. 2020;88(8):1018-1028. doi:10.1002/prot.25860
74. Weng G, Wang E, Wang Z, et al. HawkDock: a web server to predict and analyze the protein–protein complex based on computational docking and MM/GBSA. *Nucleic Acids Res*. 2019;47(W1):W322-W330. doi:10.1093/nar/gkz397
75. Ylilauri M, Pentikäinen OT. MMGBSA as a tool to understand the binding affinities of filamin-peptide interactions. *J Chem Inf Model*. 2013;53(10):2626-2633. doi:10.1021/ci4002475
76. Moal IH, Bates PA. SwarmDock and the Use of Normal Modes in Protein-Protein Docking. *Int J Mol Sci*. 2010;11(10):3623-3648. doi:10.3390/ijms11103623
77. Torchala M, Moal IH, Chaleil RAG, Fernandez-Recio J, Bates PA. SwarmDock: a server for flexible protein–protein docking. *Bioinformatics*. 2013;29(6):807-809. doi:10.1093/bioinformatics/btt038
78. Yan Y, Zhang D, Zhou P, Li B, Huang SY. HDOCK: a web server for protein–protein and protein–DNA/RNA docking based on a hybrid strategy. *Nucleic Acids Res*. 2017;45(W1):W365-W373. doi:10.1093/nar/gkx407
79. Dominguez C, Boelens R, Bonvin AMJJ. HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *J Am Chem Soc*. 2003;125(7):1731-1737. doi:10.1021/ja026939x
80. Roel-Touris J, Don CG, V. Honorato R, Rodrigues JPGLM, Bonvin AMJJ. Less Is More: Coarse-Grained Integrative Modeling of Large Biomolecular Assemblies with HADDOCK. *J Chem Theory Comput*. 2019;15(11):6358-6367. doi:10.1021/acs.jctc.9b00310

81. Christoffer C, Chen S, Bharadwaj V, et al. LZerD webserver for pairwise and multiple protein–protein docking. *Nucleic Acids Res*. 2021;49(W1):W359-W365. doi:10.1093/nar/gkab336
82. Kozakov D, Hall DR, Xia B, et al. The ClusPro web server for protein-protein docking. *Nat Protoc*. 2017;12(2):255-278. doi:10.1038/nprot.2016.169
83. Jiménez-García B, Pons C, Fernández-Recio J. pyDockWEB: a web server for rigid-body protein-protein docking using electrostatics and desolvation scoring. *Bioinforma Oxf Engl*. 2013;29(13):1698-1699. doi:10.1093/bioinformatics/btt262
84. Cheng TMK, Blundell TL, Fernandez-Recio J. pyDock: Electrostatics and desolvation for effective scoring of rigid-body protein-protein docking. *Proteins Struct Funct Bioinforma*. 2007;68(2):503-515. doi:10.1002/prot.21419
85. Ramírez-Aportela E, López-Blanco JR, Chacón P. FRODOCK 2.0: fast protein–protein docking server. *Bioinformatics*. 2016;32(15):2386-2388. doi:10.1093/bioinformatics/btw141
86. Quignot C, Rey J, Yu J, Tufféry P, Guerois R, Andreani J. InterEvDock2: an expanded server for protein docking using evolutionary and biological information from homology models and multimeric inputs. *Nucleic Acids Res*. 2018;46(W1):W408-W416. doi:10.1093/nar/gky377
87. Lensink MF, Brysbaert G, Mauri T, et al. Prediction of protein assemblies, the next frontier: The CASP14-CAPRI experiment. *Proteins*. Published online August 28, 2021. doi:10.1002/prot.26222
88. Christoffer C, Kihara D. IDP-LZerD: Software for Modeling Disordered Protein Interactions. *Methods Mol Biol Clifton NJ*. 2020;2165:231-244. doi:10.1007/978-1-0716-0708-4_13
89. Basu S, Wallner B. DockQ: A Quality Measure for Protein-Protein Docking Models. Levy YK, ed. *PLOS ONE*. 2016;11(8):e0161879. doi:10.1371/journal.pone.0161879
90. Baek M, Baker D. Deep learning and protein structure modeling. *Nat Methods*. 2022;19(1):13-14. doi:10.1038/s41592-021-01360-8
91. Dapkūnas J, Olechnovič K, Venclovas Č. Modeling of protein complexes in CASP14 with emphasis on the interaction interface prediction. *Proteins Struct Funct Bioinforma*. 2021;89(12):1834-1843. doi:10.1002/prot.26167
92. Holm L. DALI and the persistence of protein shape. *Protein Sci Publ Protein Soc*. 2020;29(1):128-140. doi:10.1002/pro.3749
93. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol*. 1993;234(3):779-815. doi:10.1006/jmbi.1993.1626
94. Ritchie DW, Kemp GJ. Protein docking using spherical polar Fourier correlations. *Proteins*. 2000;39(2):178-194.
95. Ritchie DW, Grudinin S. Spherical polar Fourier assembly of protein complexes with arbitrary point group symmetry. *J Appl Crystallogr*. 2016;49(1):158-167. doi:10.1107/S1600576715022931
96. Dapkūnas J, Olechnovič K, Venclovas Č. Modeling of protein complexes in CAPRI Round 37 using template-based approach combined with model selection. *Proteins*. 2018;86 Suppl 1:292-301. doi:10.1002/prot.25378
97. Dapkūnas J, Olechnovič K, Venclovas Č. Structural modeling of protein complexes: Current capabilities and challenges. *Proteins*. 2019;87(12):1222-1232. doi:10.1002/prot.25774
98. Olechnovič K, Venclovas Č. VoroMQA web server for assessing three-dimensional structures of proteins and protein complexes. *Nucleic Acids Res*. 2019;47(W1):W437-W442. doi:10.1093/nar/gkz367
99. Eastman P, Swails J, Chodera JD, et al. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput Biol*. 2017;13(7):e1005659. doi:10.1371/journal.pcbi.1005659
100. Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, Simmerling C. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins*. 2006;65(3):712-725. doi:10.1002/prot.21123

101. Onufriev A, Bashford D, Case DA. Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins*. 2004;55(2):383-394. doi:10.1002/prot.20033
102. Jalili V, Afgan E, Gu Q, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update. *Nucleic Acids Res*. 2020;48(W1):W395-W402. doi:10.1093/nar/gkaa434
103. Park T, Woo H, Baek M, Yang J, Seok C. Structure prediction of biological assemblies using GALAXY in CAPRI rounds 38-45. *Proteins Struct Funct Bioinforma*. 2020;88(8):1009-1017. doi:10.1002/prot.25859
104. Lee J, Liwo A, Scheraga HA. Energy-based de novo protein folding by conformational space annealing and an off-lattice united-residue force field: application to the 10-55 fragment of staphylococcal protein A and to apo calbindin D9K. *Proc Natl Acad Sci U S A*. 1999;96(5):2025-2030. doi:10.1073/pnas.96.5.2025
105. Wang X, Flannery ST, Kihara D. Protein Docking Model Evaluation by Graph Neural Networks. *Front Mol Biosci*. 2021;8:647915. doi:10.3389/fmolb.2021.647915
106. Akbal-Delibas B, Farhoodi R, Pomplun M, Haspel N. Accurate refinement of docked protein complexes using evolutionary information and deep learning. *J Bioinform Comput Biol*. 2016;14(3):1642002. doi:10.1142/S0219720016420026
107. Degiacomi MT. Coupling Molecular Dynamics and Deep Learning to Mine Protein Conformational Space. *Structure*. 2019;27(6):1034-1040.e3. doi:10.1016/j.str.2019.03.018
108. Gainza P, Sverrisson F, Monti F, et al. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nat Methods*. 2020;17(2):184-192. doi:10.1038/s41592-019-0666-6
109. Geng C, Jung Y, Renaud N, Honavar V, Bonvin AMJJ, Xue LC. iScore: a novel graph kernel-based function for scoring protein-protein docking models. *Bioinforma Oxf Engl*. 2020;36(1):112-121. doi:10.1093/bioinformatics/btz496
110. Kingsley LJ, Esquivel-Rodríguez J, Yang Y, Kihara D, Lill MA. Ranking protein–protein docking results using steered molecular dynamics and potential of mean force calculations. *J Comput Chem*. 2016;37(20):1861-1865. doi:10.1002/jcc.24412
111. Lu H, Lu L, Skolnick J. Development of Unified Statistical Potentials Describing Protein-Protein Interactions. *Biophys J*. 2003;84(3):1895-1901. doi:10.1016/S0006-3495(03)74997-2
112. Huang SY, Zou X. An iterative knowledge-based scoring function for protein-protein recognition. *Proteins Struct Funct Bioinforma*. 2008;72(2):557-579. doi:10.1002/prot.21949
113. Fink F, Hochrein J, Wolowski V, Merkl R, Gronwald W. PROCOS: Computational analysis of protein-protein complexes. *J Comput Chem*. 2011;32(12):2575-2586. doi:10.1002/jcc.21837
114. Nadaradjane AA, Guerois R, Andreani J. Protein-Protein Docking Using Evolutionary Information. In: Marsh JA, ed. *Protein Complex Assembly*. Vol 1764. Methods in Molecular Biology. Springer New York; 2018:429-447. doi:10.1007/978-1-4939-7759-8_28
115. Yu J, Vavrusa M, Andreani J, Rey J, Tufféry P, Guerois R. InterEvDock: a docking server to predict the structure of protein-protein interactions using evolutionary information. *Nucleic Acids Res*. 2016;44(W1):W542-549. doi:10.1093/nar/gkw340
116. Venkatraman V, Yang YD, Sael L, Kihara D. Protein-protein docking using region-based 3D Zernike descriptors. *BMC Bioinformatics*. 2009;10(1):407. doi:10.1186/1471-2105-10-407
117. van Zundert GCP, Rodrigues JPGLM, Trellet M, et al. The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *J Mol Biol*. 2016;428(4):720-725. doi:10.1016/j.jmb.2015.09.014
118. Evans R, O'Neill M, Pritzel A, et al. *Protein Complex Prediction with AlphaFold-Multimer*. Bioinformatics; 2021. doi:10.1101/2021.10.04.463034
119. Ghani U, Desta I, Jindal A, et al. *Improved Docking of Protein Models by a Combination*

of Alphafold2 and ClusPro. Bioinformatics; 2021. doi:10.1101/2021.09.07.459290
120. Mirdita M, Ovchinnikov S, Steinegger M. *ColabFold - Making Protein Folding Accessible to All*. Bioinformatics; 2021. doi:10.1101/2021.08.15.456425
121. Wu R, Ding F, Wang R, et al. *High-Resolution* de Novo *Structure Prediction from Primary Sequence*. Bioinformatics; 2022. doi:10.1101/2022.07.21.500999
122. Li Z, Liu X, Chen W, et al. *Uni-Fold: An Open-Source Platform for Developing Protein Folding Models beyond AlphaFold*. Bioinformatics; 2022. doi:10.1101/2022.08.04.502811
123. Yu D, Chojnowski G, Rosenthal M, Kosinski J. *AlphaPulldown – a Python Package for Protein-Protein Interaction Screens Using AlphaFold-Multimer*. Bioinformatics; 2022. doi:10.1101/2022.08.05.502961
124. Ciemny M, Kurcinski M, Kamel K, et al. Protein–peptide docking: opportunities and challenges. *Drug Discov Today*. 2018;23(8):1530-1537. doi:10.1016/j.drudis.2018.05.006
125. Berman HM, Battistuz T, Bhat TN, et al. The Protein Data Bank. *Acta Crystallogr D Biol Crystallogr*. 2002;58(Pt 6 No 1):899-907. doi:10.1107/s0907444902003451
126. Yan C, Xu X, Zou X. The Usage of ACCLUSTER for Peptide Binding Site Prediction. *Methods Mol Biol Clifton NJ*. 2017;1561:3-9. doi:10.1007/978-1-4939-6798-8_1
127. Bohnuud T, Jones G, Schueler-Furman O, Kozakov D. Detection of Peptide-Binding Sites on Protein Surfaces Using the Peptimap Server. *Methods Mol Biol Clifton NJ*. 2017;1561:11-20. doi:10.1007/978-1-4939-6798-8_2
128. Mattos C, Ringe D. Locating and characterizing binding sites on proteins. *Nat Biotechnol*. 1996;14(5):595-599. doi:10.1038/nbt0596-595
129. Shuker SB, Hajduk PJ, Meadows RP, Fesik SW. Discovering high-affinity ligands for proteins: SAR by NMR. *Science*. 1996;274(5292):1531-1534. doi:10.1126/science.274.5292.1531
130. Brenke R, Kozakov D, Chuang GY, et al. Fragment-based identification of druggable “hot spots” of proteins using Fourier domain correlation techniques. *Bioinforma Oxf Engl*. 2009;25(5):621-627. doi:10.1093/bioinformatics/btp036
131. Ngan CH, Hall DR, Zerbe B, Grove LE, Kozakov D, Vajda S. FTSite: high accuracy detection of ligand binding sites on unbound protein structures. *Bioinforma Oxf Engl*. 2012;28(2):286-287. doi:10.1093/bioinformatics/btr651
132. Kozakov D, Grove LE, Hall DR, et al. The FTMap family of web servers for determining and characterizing ligand-binding hot spots of proteins. *Nat Protoc*. 2015;10(5):733-755. doi:10.1038/nprot.2015.043
133. Lamiable A, Thévenet P, Eustache S, Saladin A, Moroy G, Tuffery P. Peptide Suboptimal Conformation Sampling for the Prediction of Protein-Peptide Interactions. *Methods Mol Biol Clifton NJ*. 2017;1561:21-34. doi:10.1007/978-1-4939-6798-8_3
134. London N, Movshovitz-Attias D, Schueler-Furman O. The structural basis of peptide-protein binding strategies. *Struct Lond Engl 1993*. 2010;18(2):188-199. doi:10.1016/j.str.2009.11.012
135. de Vries SJ, Chauvot de Beauchêne I, Schindler CEM, Zacharias M. Cryo-EM Data Are Superior to Contact and Interface Information in Integrative Modeling. *Biophys J*. 2016;110(4):785-797. doi:10.1016/j.bpj.2015.12.038
136. Lee H, Heo L, Lee MS, Seok C. GalaxyPepDock: a protein-peptide docking tool based on interaction similarity and energy optimization. *Nucleic Acids Res*. 2015;43(W1):W431-435. doi:10.1093/nar/gkv495
137. Schindler CEM, de Vries SJ, Zacharias M. Fully Blind Peptide-Protein Docking with pepATTRACT. *Struct Lond Engl 1993*. 2015;23(8):1507-1515. doi:10.1016/j.str.2015.05.021
138. Kozakov D, Brenke R, Comeau SR, Vajda S. PIPER: an FFT-based protein docking program with pairwise potentials. *Proteins*. 2006;65(2):392-406. doi:10.1002/prot.21117
139. London N, Raveh B, Cohen E, Fathi G, Schueler-Furman O. Rosetta FlexPepDock web server--high resolution modeling of peptide-protein interactions. *Nucleic Acids Res*. 2011;39(Web Server issue):W249-253. doi:10.1093/nar/gkr431

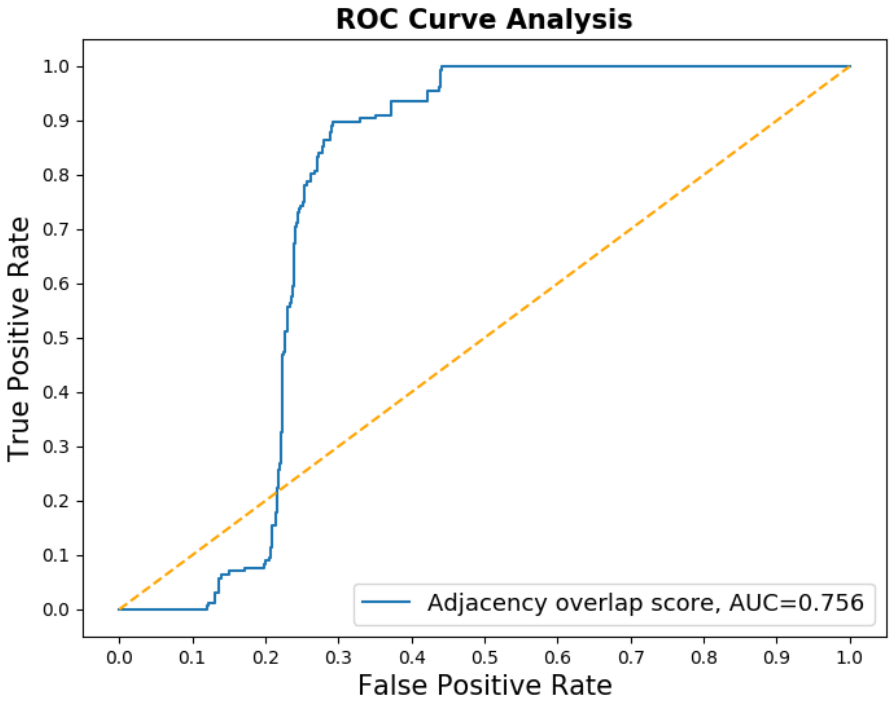
140. Janin J, Henrick K, Moult J, et al. CAPRI: a Critical Assessment of PRedicted Interactions. *Proteins*. 2003;52(1):2-9. doi:10.1002/prot.10381
141. Méndez R, Leplae R, Lensink MF, Wodak SJ. Assessment of CAPRI predictions in rounds 3-5 shows progress in docking procedures. *Proteins*. 2005;60(2):150-169. doi:10.1002/prot.20551
142. Lensink MF, Wodak SJ. Docking and scoring protein interactions: CAPRI 2009. *Proteins*. 2010;78(15):3073-3084. doi:10.1002/prot.22818
143. Lensink MF, Méndez R, Wodak SJ. Docking and scoring protein complexes: CAPRI 3rd Edition. *Proteins*. 2007;69(4):704-718. doi:10.1002/prot.21804
144. Lensink MF, Velankar S, Kryshtafovych A, et al. Prediction of homoprotein and heteroprotein complexes by protein docking and template-based modeling: A CASP-CAPRI experiment. *Proteins*. 2016;84 Suppl 1:323-348. doi:10.1002/prot.25007
145. Lensink MF, Velankar S, Wodak SJ. Modeling protein-protein and protein-peptide complexes: CAPRI 6th edition. *Proteins*. 2017;85(3):359-377. doi:10.1002/prot.25215
146. Lensink MF, Brysbaert G, Nadzirin N, et al. Blind prediction of homo- and hetero-protein complexes: The CASP13-CAPRI experiment. *Proteins*. 2019;87(12):1200-1221. doi:10.1002/prot.25838
147. Lensink MF, Nadzirin N, Velankar S, Wodak SJ. Modeling protein-protein, protein-peptide, and protein-oligosaccharide complexes: CAPRI 7th edition. *Proteins*. 2020;88(8):916-938. doi:10.1002/prot.25870
148. Lensink MF, Velankar S, Baek M, Heo L, Seok C, Wodak SJ. The challenge of modeling protein assemblies: the CASP12-CAPRI experiment. *Proteins*. 2018;86 Suppl 1:257-273. doi:10.1002/prot.25419
149. Lensink MF, Wodak SJ. Score_set: a CAPRI benchmark for scoring protein complexes. *Proteins*. 2014;82(11):3163-3169. doi:10.1002/prot.24678
150. Wang J, Torii M, Liu H, Hart GW, Hu ZZ. dbOGAP - An Integrated Bioinformatics Resource for Protein O-GlcNAcylation. *BMC Bioinformatics*. 2011;12(1):91. doi:10.1186/1471-2105-12-91
151. Quinlan JR. Induction of decision trees. *Mach Learn*. 1986;1(1):81-106. doi:10.1007/BF00116251
152. Doupe P, Faghmous J, Basu S. Machine Learning for Health Services Researchers. *Value Health J Int Soc Pharmacoeconomics Outcomes Res*. 2019;22(7):808-815. doi:10.1016/j.jval.2019.02.012
153. Jerome H. Friedman. Greedy Function Approximation: A Gradient Boosting Machine. *Ann Stat*. 2001;29(5):1189--1232.
154. Chauhan A, Chauhan D, Rout C. Role of Gist and PHOG Features in Computer-Aided Diagnosis of Tuberculosis without Segmentation. Kestler HA, ed. *PLoS ONE*. 2014;9(11):e112980. doi:10.1371/journal.pone.0112980
155. Jiang F, Jiang Y, Zhi H, et al. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol*. 2017;2(4):230-243. doi:10.1136/svn-2017-000101
156. Wulff-Fuentes E, Berendt RR, Massman L, et al. The human O-GlcNAcome database and meta-analysis. *Sci Data*. 2021;8(1):25. doi:10.1038/s41597-021-00810-4
157. Malard F, Wulff-Fuentes E, Berendt RR, Didier G, Olivier-Van Stichelen S. Automatization and self-maintenance of the O-GlcNAcome catalog: a smart scientific database. *Database*. 2021;2021:baab039. doi:10.1093/database/baab039
158. Gordon DE, Jang GM, Bouhaddou M, et al. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature*. 2020;583(7816):459-468. doi:10.1038/s41586-020-2286-9
159. Newaz K, Wright G, Piland J, et al. Network analysis of synonymous codon usage. Ponty Y, ed. *Bioinformatics*. 2020;36(19):4876-4884. doi:10.1093/bioinformatics/btaa603
160. Mauri T, Menu-Bouaouiche L, Bardor M, Lefebvre T, Lensink MF, Brysbaert G. O-GlcNAcylation Prediction: An Unattained Objective. *Adv Appl Bioinforma Chem AABC*. 2021;14:87-102. doi:10.2147/AABC.S294867
161. Pundir S, Martin M, O'Donovan C. UniProt tools. *Curr Protoc Bioinforma Ed Board*

Andreas Baxevanis Al. 2016;53:1.29.1-1.29.15. doi:10.1002/0471250953.bi0129s53
162. Yates A, Beal K, Keenan S, et al. The Ensembl REST API: Ensembl Data for Any Language. *Bioinforma Oxf Engl*. 2015;31(1):143-145. doi:10.1093/bioinformatics/btu613
163. Cilia E, Pancsa R, Tompa P, Lenaerts T, Vranken WF. The DynaMine webserver: predicting protein dynamics from sequence. *Nucleic Acids Res*. 2014;42(W1):W264-W270. doi:10.1093/nar/gku270
164. Cilia E, Pancsa R, Tompa P, Lenaerts T, Vranken WF. From protein sequence to dynamics and disorder with DynaMine. *Nat Commun*. 2013;4(1):2741. doi:10.1038/ncomms3741
165. Heffernan R, Paliwal K, Lyons J, Singh J, Yang Y, Zhou Y. Single-sequence-based prediction of protein secondary structures and solvent accessibility by deep whole-sequence learning. *J Comput Chem*. 2018;39(26):2210-2216. doi:10.1002/jcc.25534
166. Johnson LS, Eddy SR, Portugaly E. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics*. 2010;11(1):431. doi:10.1186/1471-2105-11-431
167. Remmert M, Biegert A, Hauser A, Söding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods*. 2012;9(2):173-175. doi:10.1038/nmeth.1818
168. Hubbard, S.J, Thornton, J.M. NACCESS. Published online 1993.
169. Yang J, Zhang Y. I-TASSER server: new development for protein structure and function predictions. *Nucleic Acids Res*. 2015;43(W1):W174-181. doi:10.1093/nar/gkv342
170. Smet-Nocca C, Broncel M, Wieruszeski JM, et al. Identification of O-GlcNAc sites within peptides of the Tau protein and their impact on phosphorylation. *Mol Biosyst*. 2011;7(5):1420-1429. doi:10.1039/c0mb00337a
171. Schrödinger, LLC, Warren DeLano. PyMOL. Published online May 20, 2020. http://www.pymol.org/pymol
172. Harrison AG, Lin T, Wang P. Mechanisms of SARS-CoV-2 Transmission and Pathogenesis. *Trends Immunol*. 2020;41(12):1100-1115. doi:10.1016/j.it.2020.10.004
173. Peiris J, Lai S, Poon L, et al. Coronavirus as a possible cause of severe acute respiratory syndrome. *The Lancet*. 2003;361(9366):1319-1325. doi:10.1016/S0140-6736(03)13077-2
174. Huang C, Wang Y, Li X, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet*. 2020;395(10223):497-506. doi:10.1016/S0140-6736(20)30183-5
175. Janin J, Wodak SJ. Introduction. In: *Advances in Protein Chemistry*. Vol 61. Elsevier; 2002:1-8. doi:10.1016/S0065-3233(02)61000-9
176. DiMasi JA, Grabowski HG, Hansen RW. Innovation in the pharmaceutical industry: New estimates of R&D costs. *J Health Econ*. 2016;47:20-33. doi:10.1016/j.jhealeco.2016.01.012
177. Song CM, Lim SJ, Tong JC. Recent advances in computer-aided drug design. *Brief Bioinform*. 2009;10(5):579-591. doi:10.1093/bib/bbp023
178. Yu W, MacKerell AD. Computer-Aided Drug Design Methods. *Methods Mol Biol Clifton NJ*. 2017;1520:85-106. doi:10.1007/978-1-4939-6634-9_5
179. Batool M, Ahmad B, Choi S. A Structure-Based Drug Discovery Paradigm. *Int J Mol Sci*. 2019;20(11):E2783. doi:10.3390/ijms20112783
180. Binkowski TA, Naghibzadeh S, Liang J. CASTp: Computed Atlas of Surface Topography of proteins. *Nucleic Acids Res*. 2003;31(13):3352-3355. doi:10.1093/nar/gkg512
181. Volkamer A, Kuhn D, Rippmann F, Rarey M. DoGSiteScorer: a web server for automatic binding site prediction, analysis and druggability assessment. *Bioinforma Oxf Engl*. 2012;28(15):2074-2075. doi:10.1093/bioinformatics/bts310
182. Sun J, Chen K. NSiteMatch: Prediction of Binding Sites of Nucleotides by Identifying the Structure Similarity of Local Surface Patches. *Comput Math Methods Med*.

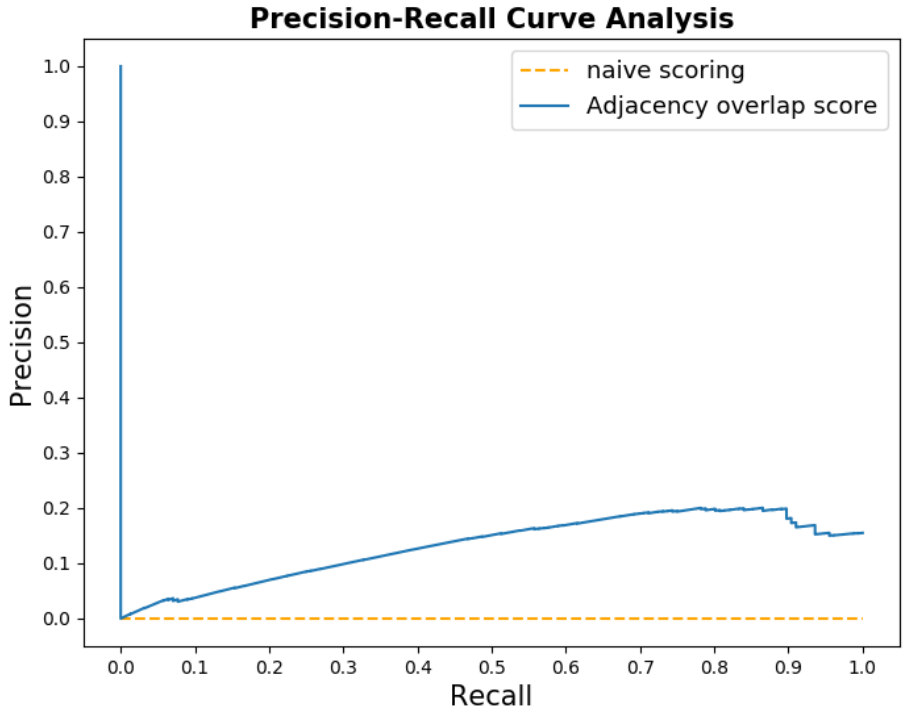
2017;2017:5471607. doi:10.1155/2017/5471607
183. Tan KP, Varadarajan R, Madhusudhan MS. DEPTH: a web server to compute depth and predict small-molecule binding cavities in proteins. *Nucleic Acids Res*. 2011;39(Web Server issue):W242-248. doi:10.1093/nar/gkr356
184. Zhu H, Pisabarro MT. MSPocket: an orientation-independent algorithm for the detection of ligand binding pockets. *Bioinforma Oxf Engl*. 2011;27(3):351-358. doi:10.1093/bioinformatics/btq672
185. Huang B. MetaPocket: a meta approach to improve protein ligand binding site prediction. *Omics J Integr Biol*. 2009;13(4):325-330. doi:10.1089/omi.2009.0045
186. Laurie ATR, Jackson RM. Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinforma Oxf Engl*. 2005;21(9):1908-1916. doi:10.1093/bioinformatics/bti315
187. Ibrahim MAA, Abdelrahman AHM, Allemailem KS, Almatroudi A, Moustafa MF, Hegazy MEF. In Silico Evaluation of Prospective Anti-COVID-19 Drug Candidates as Potential SARS-CoV-2 Main Protease Inhibitors. *Protein J*. 2021;40(3):296-309. doi:10.1007/s10930-020-09945-6
188. Matsuzaki Y, Uchikoga N, Ohue M, Akiyama Y. Rigid-Docking Approaches to Explore Protein-Protein Interaction Space. *Adv Biochem Eng Biotechnol*. 2017;160:33-55. doi:10.1007/10_2016_41
189. Schueler-Furman O, London N, eds. *Modeling Peptide-Protein Interactions: Methods and Protocols*. Vol 1561. Springer New York; 2017. doi:10.1007/978-1-4939-6798-8
190. Lee H, Seok C. Template-Based Prediction of Protein-Peptide Interactions by Using GalaxyPepDock. *Methods Mol Biol Clifton NJ*. 2017;1561:37-47. doi:10.1007/978-1-4939-6798-8_4
191. Ko J, Park H, Heo L, Seok C. GalaxyWEB server for protein structure prediction and refinement. *Nucleic Acids Res*. 2012;40(Web Server issue):W294-297. doi:10.1093/nar/gks493
192. Schindler C, Zacharias M. Application of the ATTRACT Coarse-Grained Docking and Atomistic Refinement for Predicting Peptide-Protein Interactions. *Methods Mol Biol Clifton NJ*. 2017;1561:49-68. doi:10.1007/978-1-4939-6798-8_5
193. Brysbaert G, Mauri T, Lensink MF. Comparing protein structures with RINspector automation in Cytoscape. *F1000Research*. 2018;7:563. doi:10.12688/f1000research.14298.2
194. Amitai G, Shemesh A, Sitbon E, et al. Network analysis of protein structures identifies functional residues. *J Mol Biol*. 2004;344(4):1135-1146. doi:10.1016/j.jmb.2004.10.055
195. Pupko T, Bell RE, Mayrose I, Glaser F, Ben-Tal N. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinforma Oxf Engl*. 2002;18 Suppl 1:S71-77. doi:10.1093/bioinformatics/18.suppl_1.s71
196. DeLano, W.L. The PyMOL Molecular Graphics System. Published online 2002.
197. McLachlan GJ, Bean RW, Ng SK. Clustering. In: Keith JM, ed. *Bioinformatics*. Vol 1526. Methods in Molecular Biology. Springer New York; 2017:345-362. doi:10.1007/978-1-4939-6613-4_19
198. Shannon P, Markiel A, Ozier O, et al. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res*. 2003;13(11):2498-2504. doi:10.1101/gr.1239303
199. Vlasblom J, Wodak SJ. Markov clustering versus affinity propagation for the partitioning of protein interaction graphs. *BMC Bioinformatics*. 2009;10:99. doi:10.1186/1471-2105-10-99
200. Rodrigues JPGLM, Trellet M, Schmitz C, et al. Clustering biomolecular complexes by residue contacts similarity. *Proteins*. 2012;80(7):1810-1817. doi:10.1002/prot.24078
201. Tuncbag N, Gursoy A, Nussinov R, Keskin O. Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM. *Nat Protoc*. 2011;6(9):1341-1354. doi:10.1038/nprot.2011.367
202. Rosell M, Fernández-Recio J. Docking approaches for modeling multi-molecular

assemblies. *Curr Opin Struct Biol*. 2020;64:59-65. doi:10.1016/j.sbi.2020.05.016
203. Liu S, Gao Y, Vakser IA. DOCKGROUND protein-protein docking decoy set. *Bioinformatics*. 2008;24(22):2634-2635. doi:10.1093/bioinformatics/btn497
204. Kundrotas PJ, Anishchenko I, Dauzhenka T, et al. Dockground: A comprehensive data resource for modeling of protein complexes. *Protein Sci*. 2018;27(1):172-181. doi:10.1002/pro.3295
205. Hwang H, Vreven T, Janin J, Weng Z. Protein-protein docking benchmark version 4.0. *Proteins*. 2010;78(15):3111-3114. doi:10.1002/prot.22830
206. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res*. 2014;42(Database issue):D490-495. doi:10.1093/nar/gkt1178
207. Lubas WA, Frank DW, Krause M, Hanover JA. O-Linked GlcNAc transferase is a conserved nucleocytoplasmic protein containing tetratricopeptide repeats. *J Biol Chem*. 1997;272(14):9316-9324. doi:10.1074/jbc.272.14.9316
208. Seo HG, Kim HB, Kang MJ, Ryum JH, Yi EC, Cho JW. Identification of the nuclear localisation signal of O-GlcNAc transferase and its nuclear import regulation. *Sci Rep*. 2016;6:34614. doi:10.1038/srep34614
209. Lazarus MB, Jiang J, Kapuria V, et al. HCF-1 Is Cleaved in the Active Site of O-GlcNAc Transferase. *Science*. 2013;342(6163):1235-1239. doi:10.1126/science.1243990
210. Meek RW, Blaza JN, Busmann JA, Alteen MG, Vocadlo DJ, Davies GJ. Cryo-EM structure provides insights into the dimer arrangement of the O-linked β-N-acetylglucosamine transferase OGT. *Nat Commun*. 2021;12(1):6508. doi:10.1038/s41467-021-26796-6
211. Taciak B, Pruszynska I, Kiraga L, Bialasek M, Krol M. Wnt signaling pathway in development and cancer. *J Physiol Pharmacol Off J Pol Physiol Soc*. 2018;69(2). doi:10.26402/jpp.2018.2.07
212. Schunk SJ, Floege J, Fliser D, Speer T. WNT-β-catenin signalling - a versatile player in kidney injury and repair. *Nat Rev Nephrol*. 2021;17(3):172-184. doi:10.1038/s41581-020-00343-w
213. Edeling M, Ragi G, Huang S, Pavenstädt H, Susztak K. Developmental signalling pathways in renal fibrosis: the roles of Notch, Wnt and Hedgehog. *Nat Rev Nephrol*. 2016;12(7):426-439. doi:10.1038/nrneph.2016.54
214. Liu C, Li Y, Semenov M, et al. Control of beta-catenin phosphorylation/degradation by a dual-kinase mechanism. *Cell*. 2002;108(6):837-847. doi:10.1016/s0092-8674(02)00685-2
215. Olivier-Van Stichelen S, Dehennaut V, Buzy A, et al. O-GlcNAcylation stabilizes β-catenin through direct competition with phosphorylation at threonine 41. *FASEB J Off Publ Fed Am Soc Exp Biol*. 2014;28(8):3325-3338. doi:10.1096/fj.13-243535
216. Yin R, Feng BY, Varshney A, Pierce BG. Benchmarking AlphaFold for protein complex modeling reveals accuracy determinants. *Protein Sci*. 2022;31(8). doi:10.1002/pro.4379
217. Pierce BG, Hourai Y, Weng Z. Accelerating protein docking in ZDOCK using an advanced 3D convolution library. *PloS One*. 2011;6(9):e24657. doi:10.1371/journal.pone.0024657
218. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. *Genome Res*. 2004;14(6):1188-1190. doi:10.1101/gr.849004

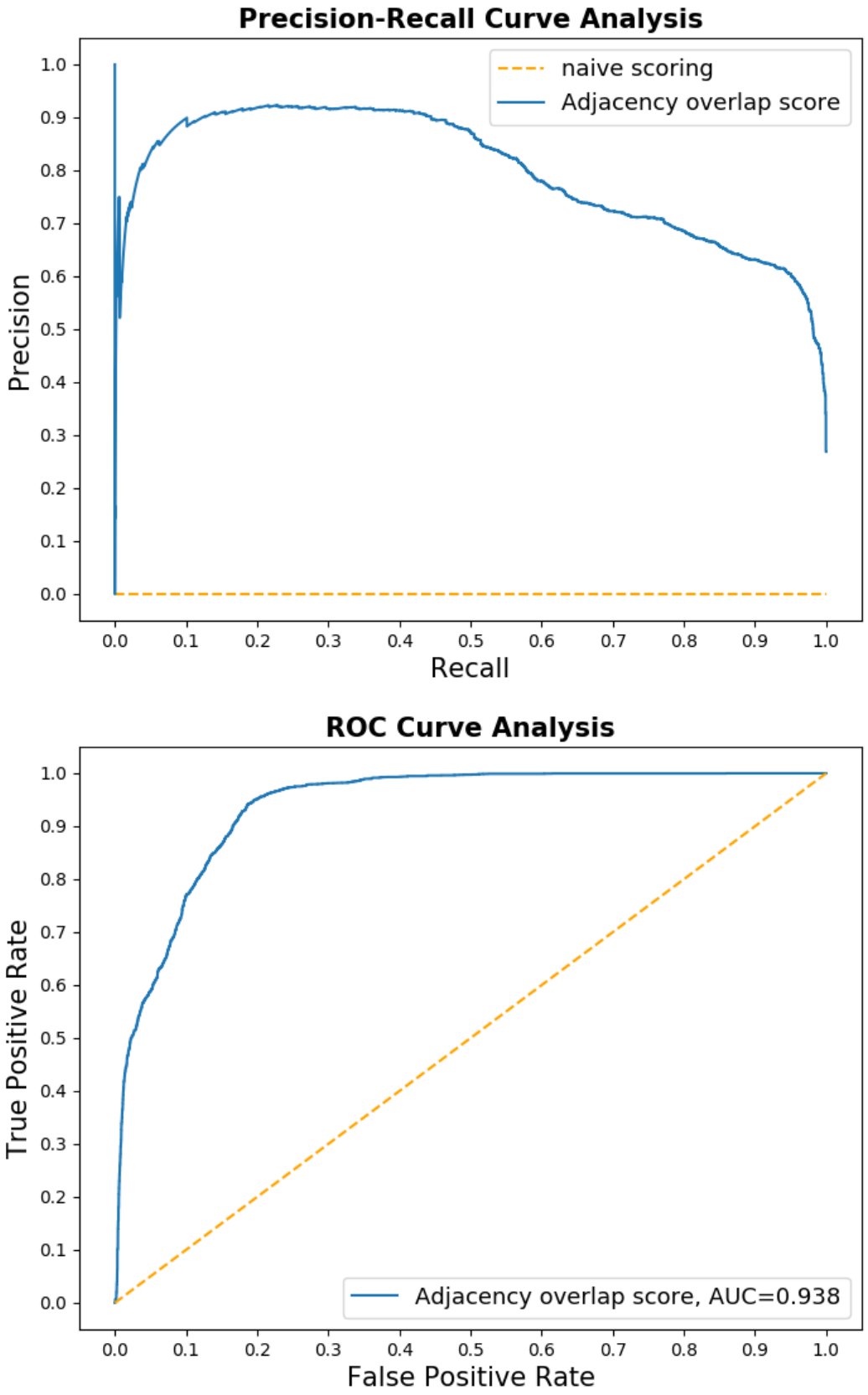
Appendix: Supplementary data



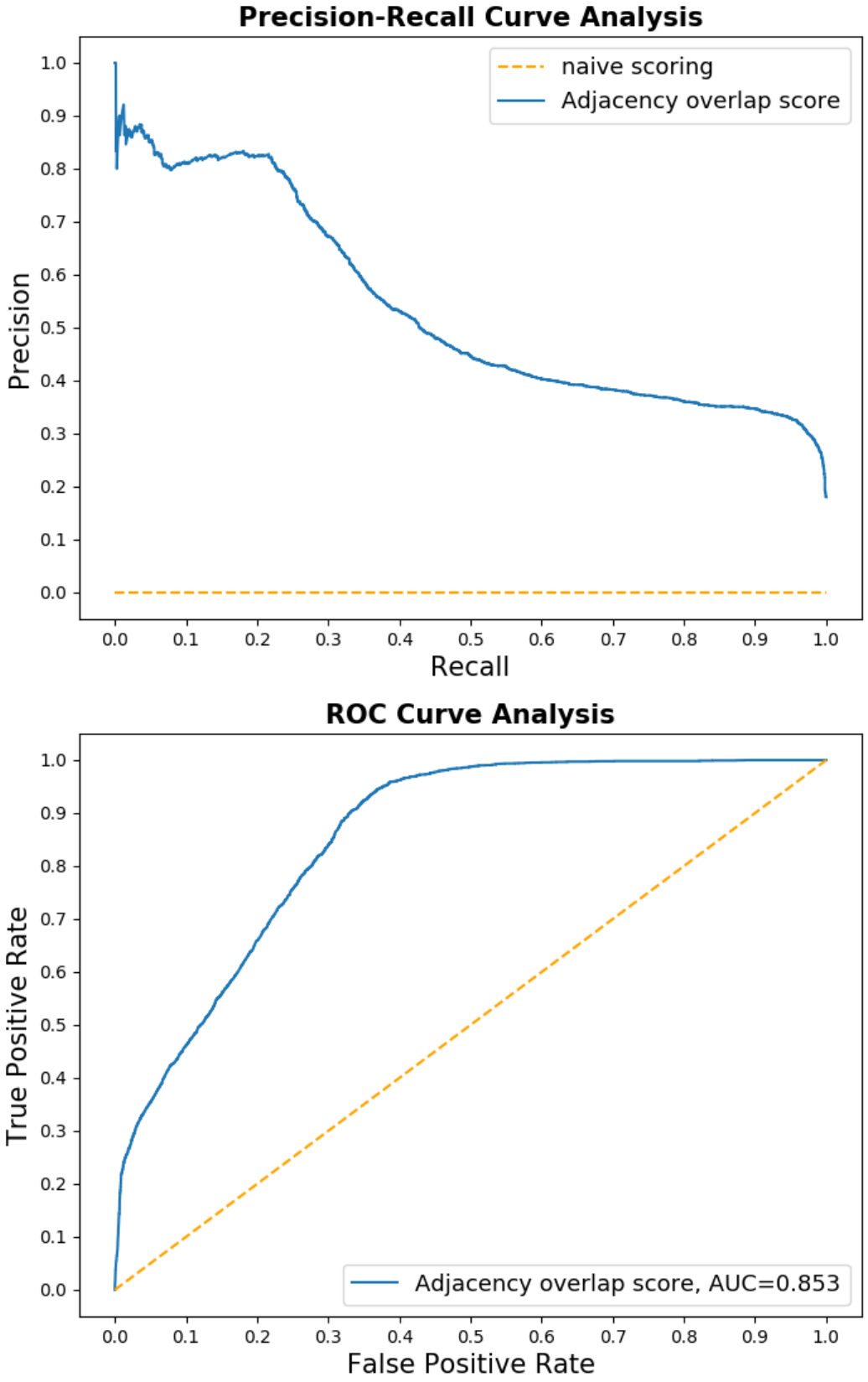
Supplementary Figure. **ROC curve of the adjacency overlap scoring method on the P-Set for Archaea**



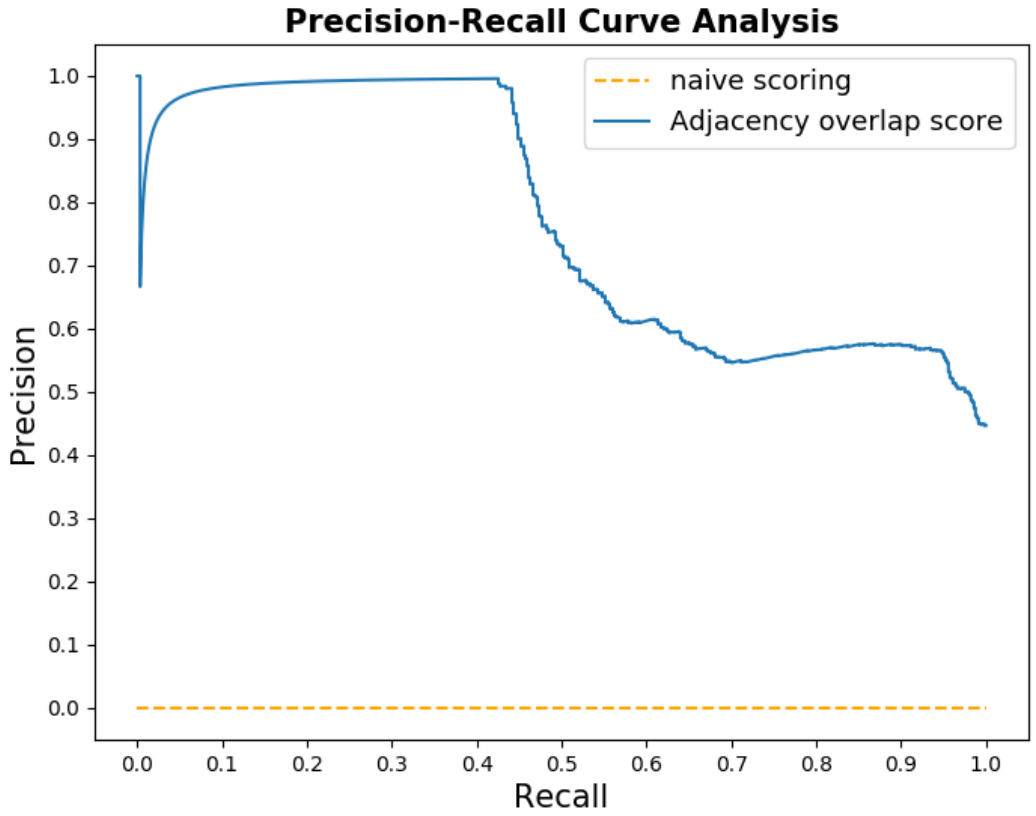
Supplementary figure**. Precision-Recall curve of the adjacency overlap scoring method on the P-Set for Archaea**

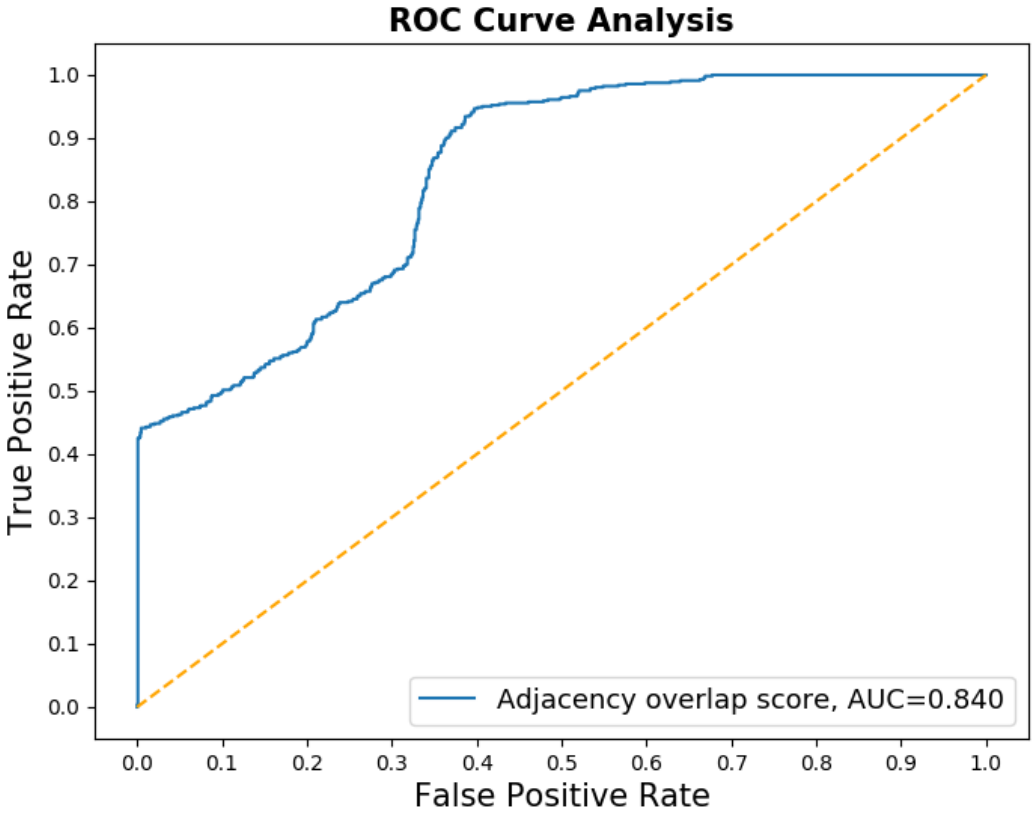


Supplementary Figure: ROC and Precision-Recall of the adjacency overlap scoring method on the P-Set for Bacteria

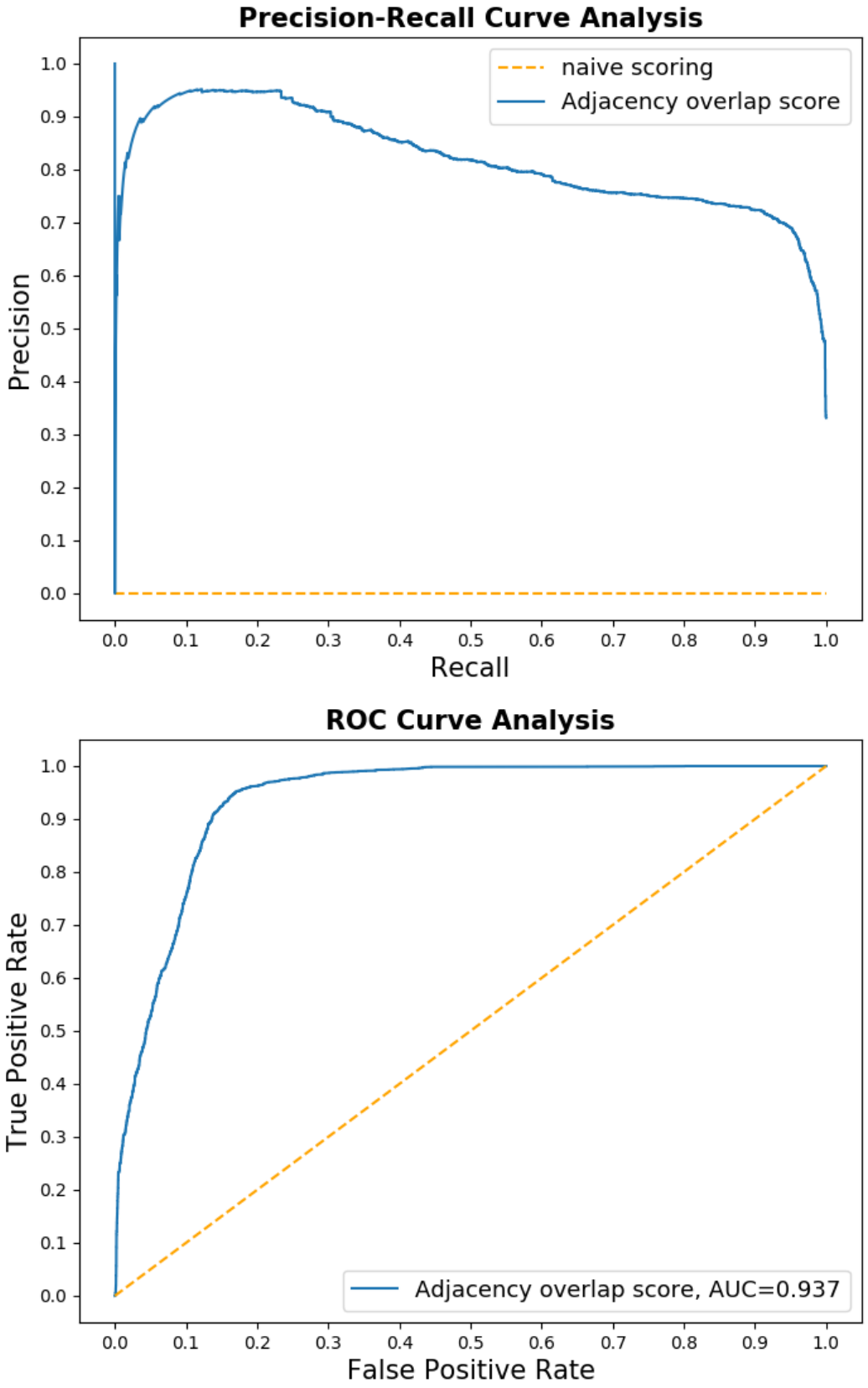


Supplementary Figure: ROC and Precision-Recall of the adjacency overlap scoring method on the P-Set for Eukaryotes

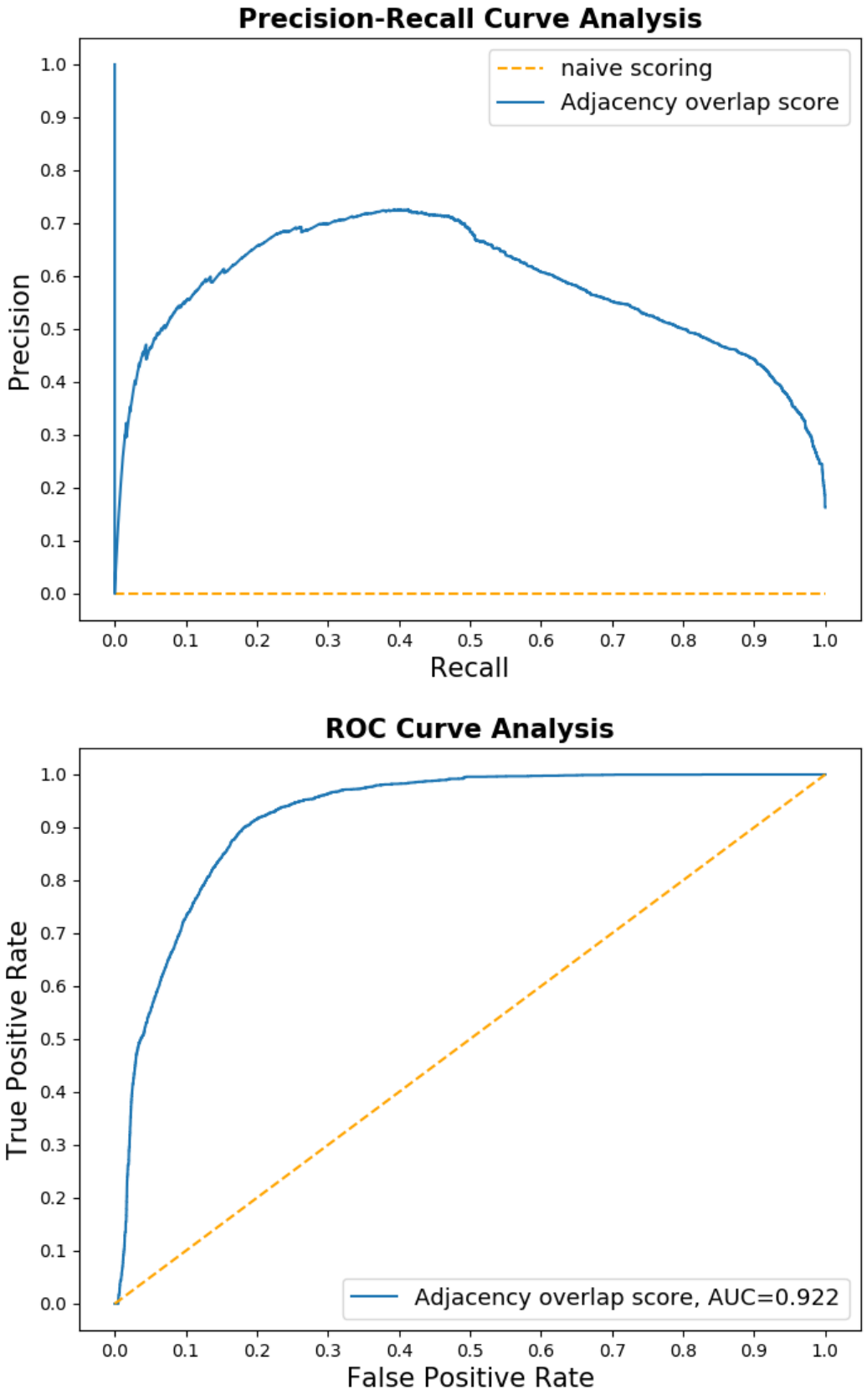




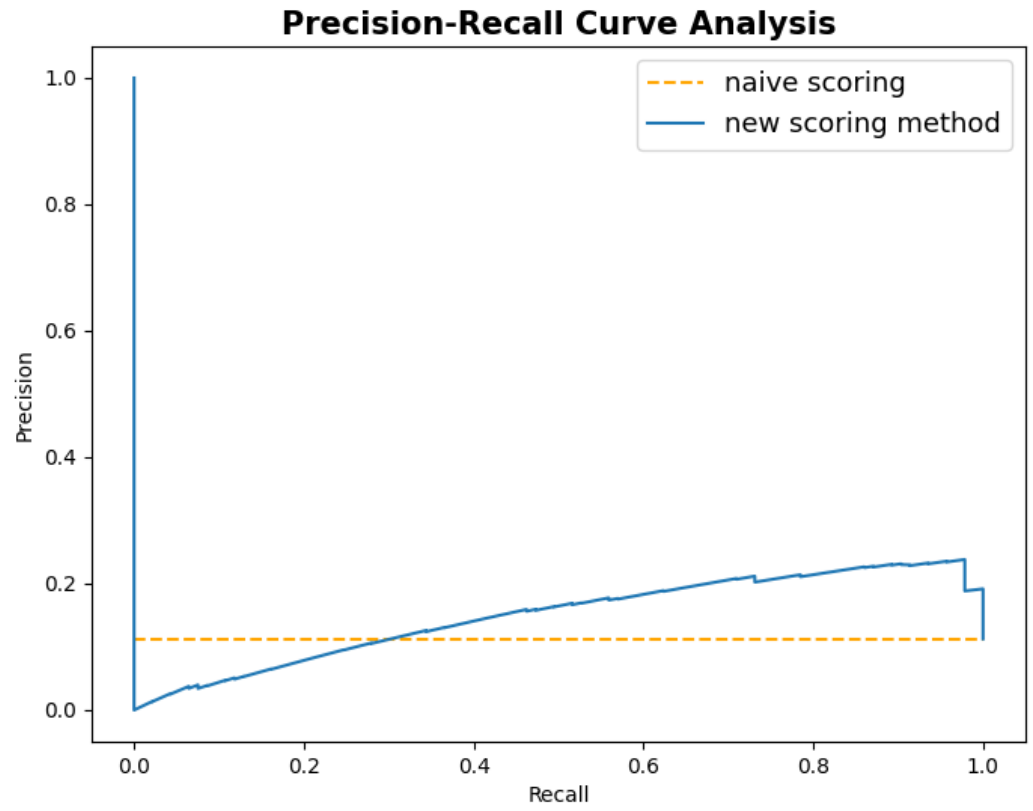
Supplementary Figure: ROC and Precision-Recall of the adjacency overlap scoring method on the P-Set for Viruses

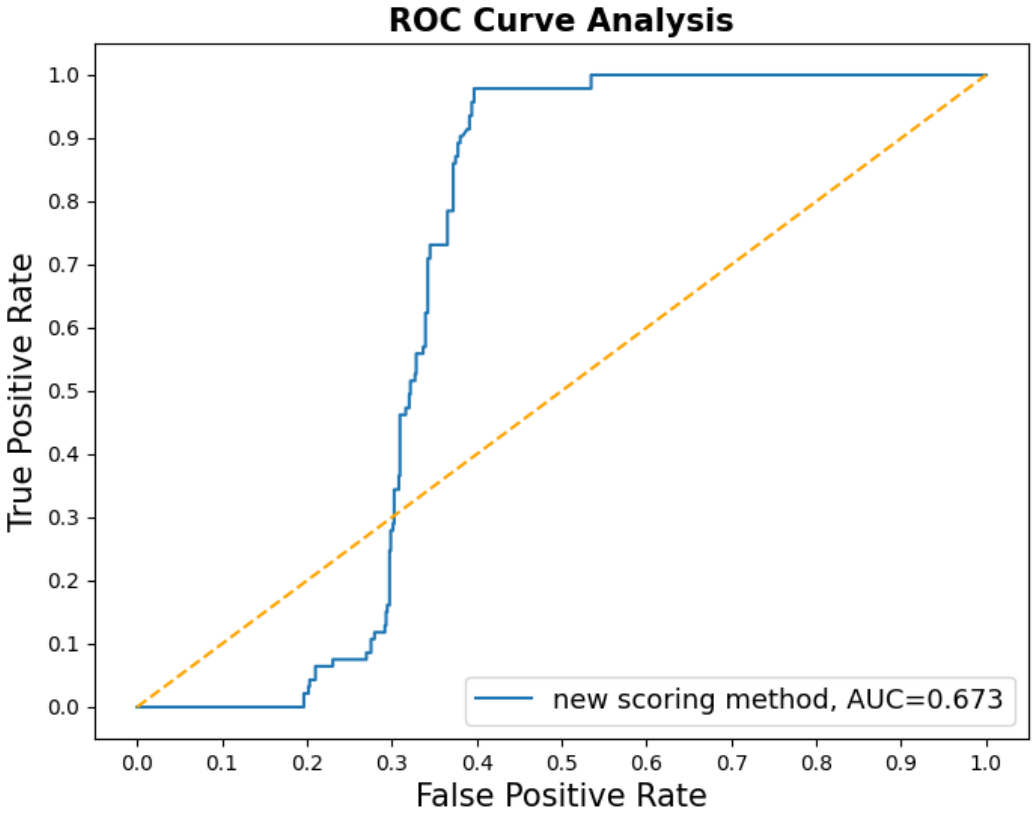


Supplementary Figure: ROC and Precision-Recall of the adjacency overlap scoring method on the P-Set for homo-dimers

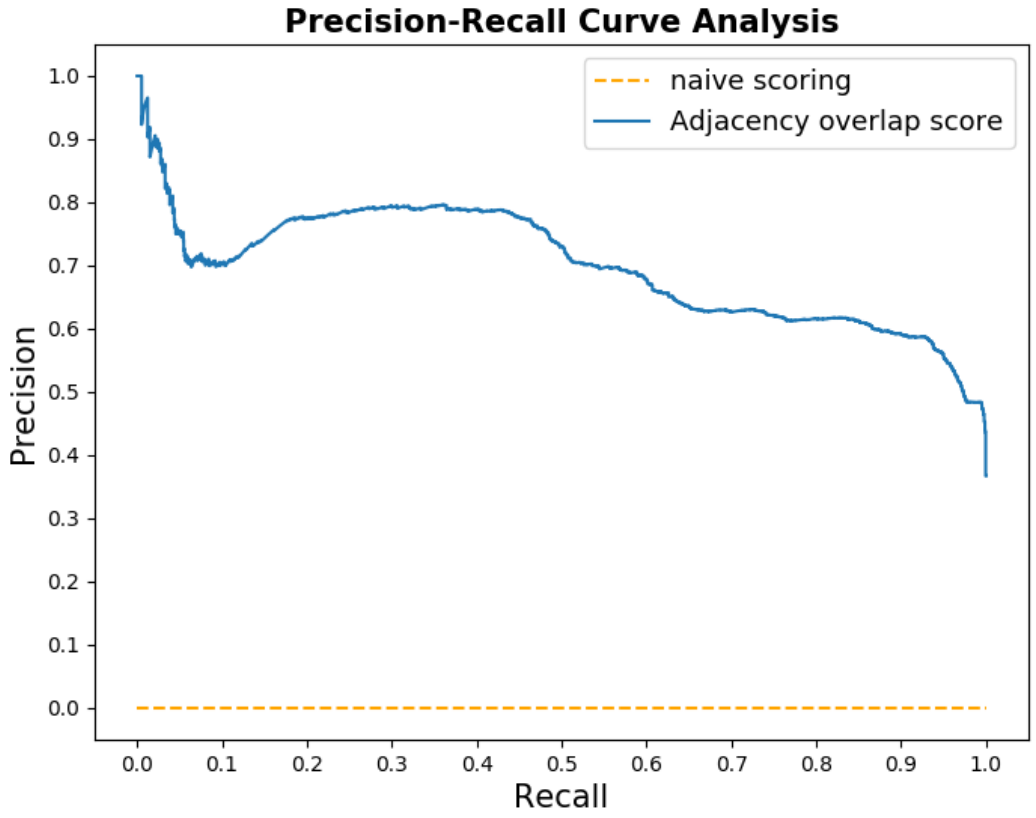


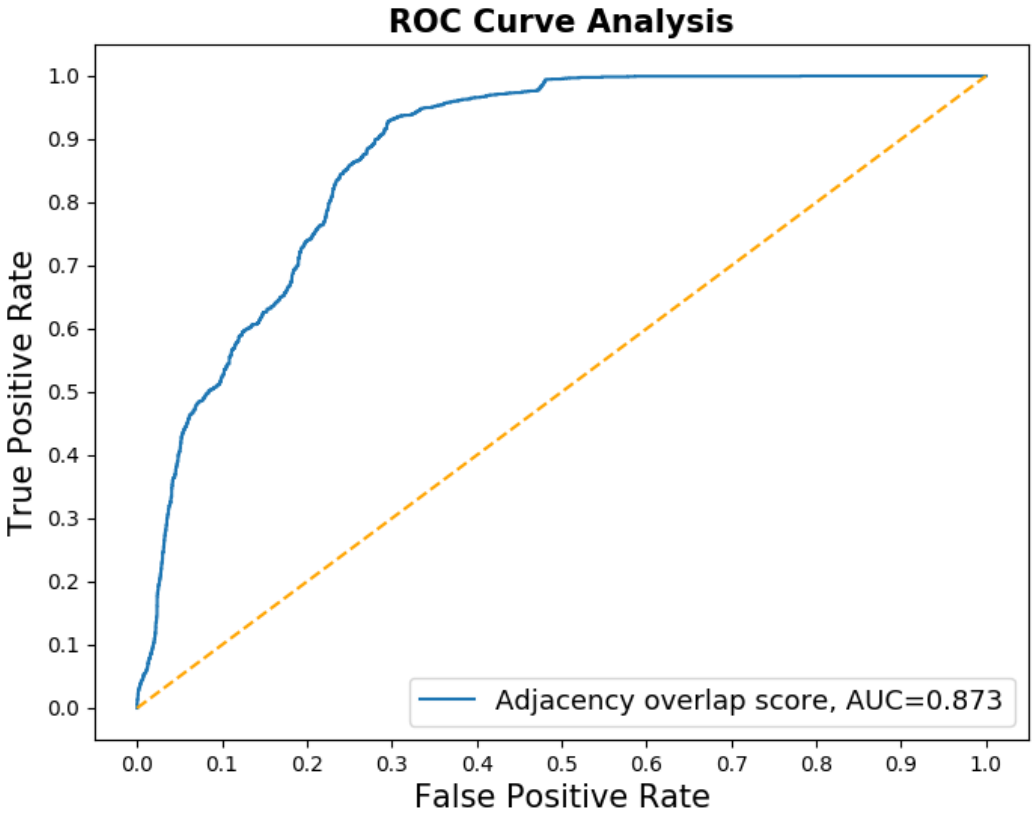
Supplementary Figure: ROC and Precision-Recall of the adjacency overlap scoring method on the P-Set for hetero-dimers



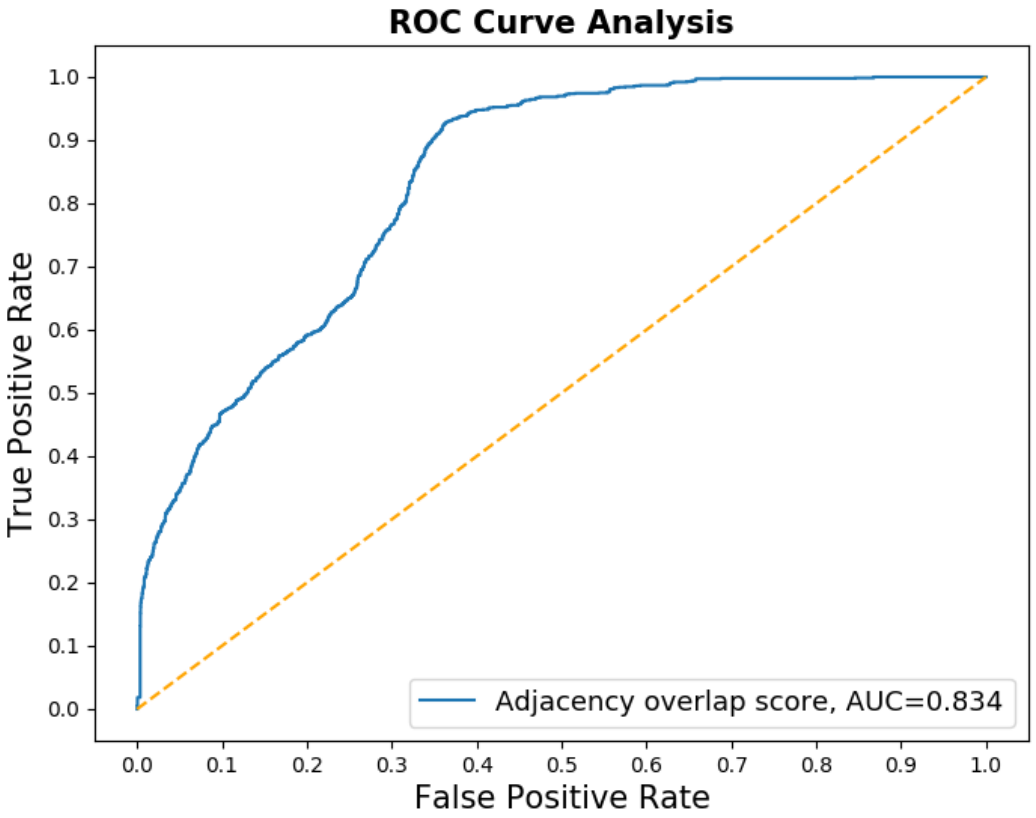


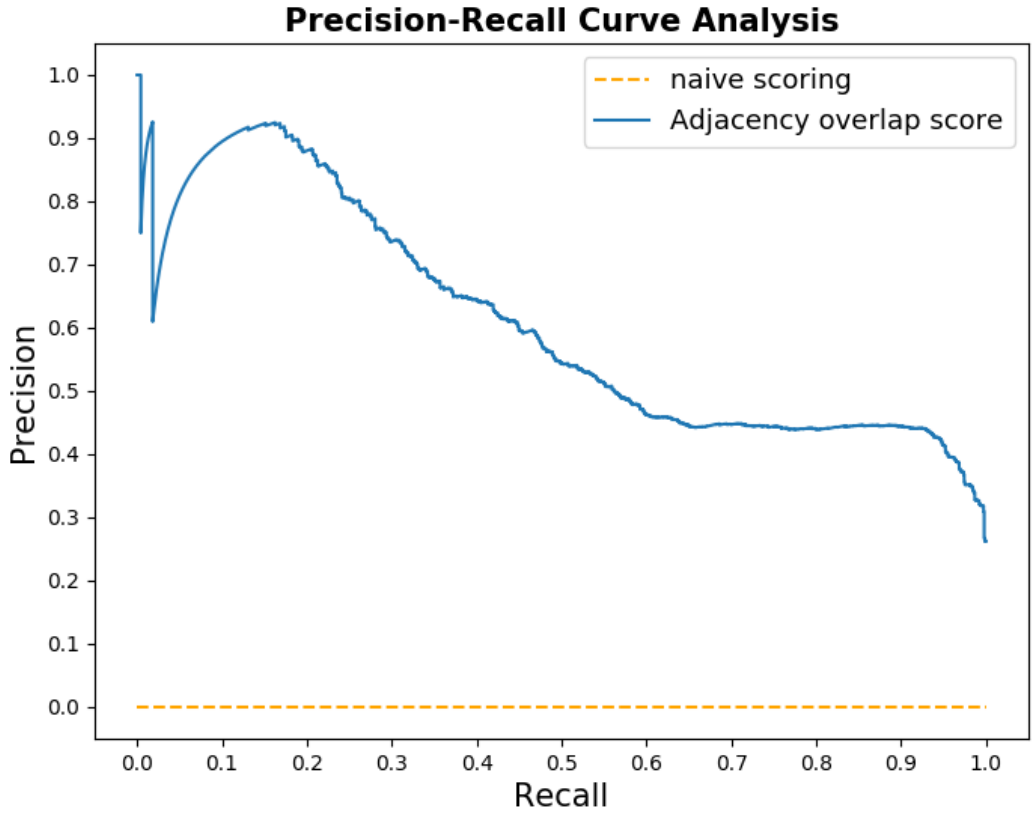
Supplementary Figure: ROC and Precision-Recall of the adjacency overlap scoring method on the S-Set for Archaea



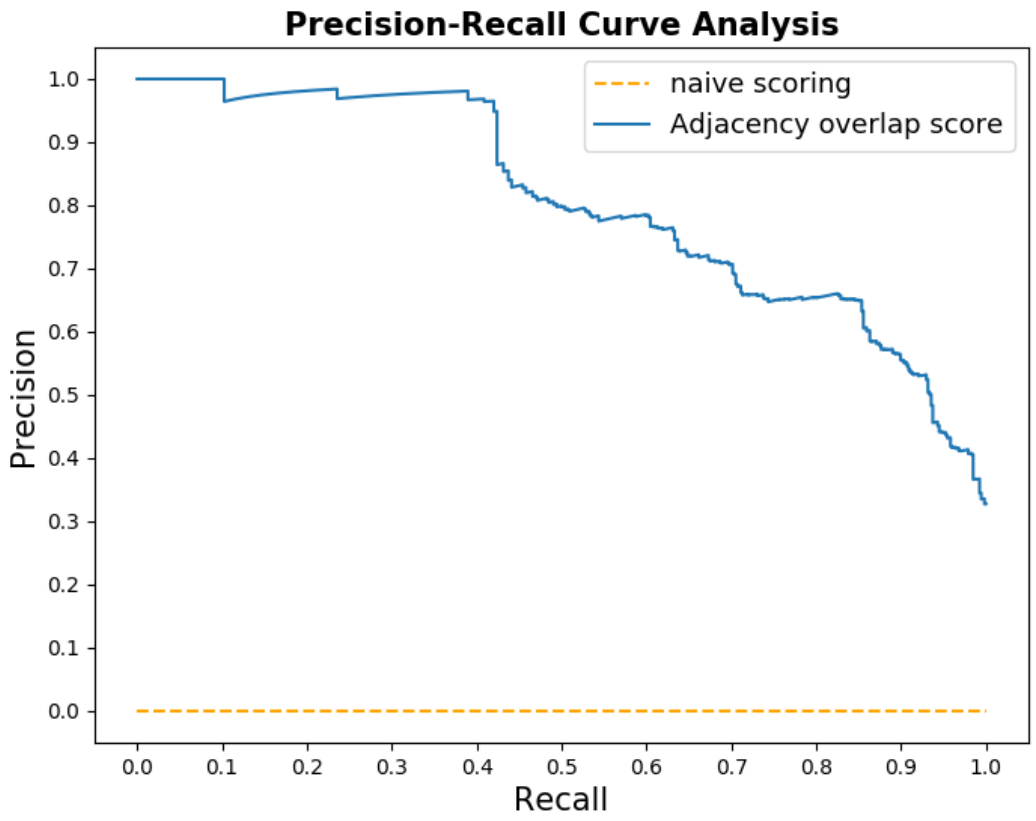


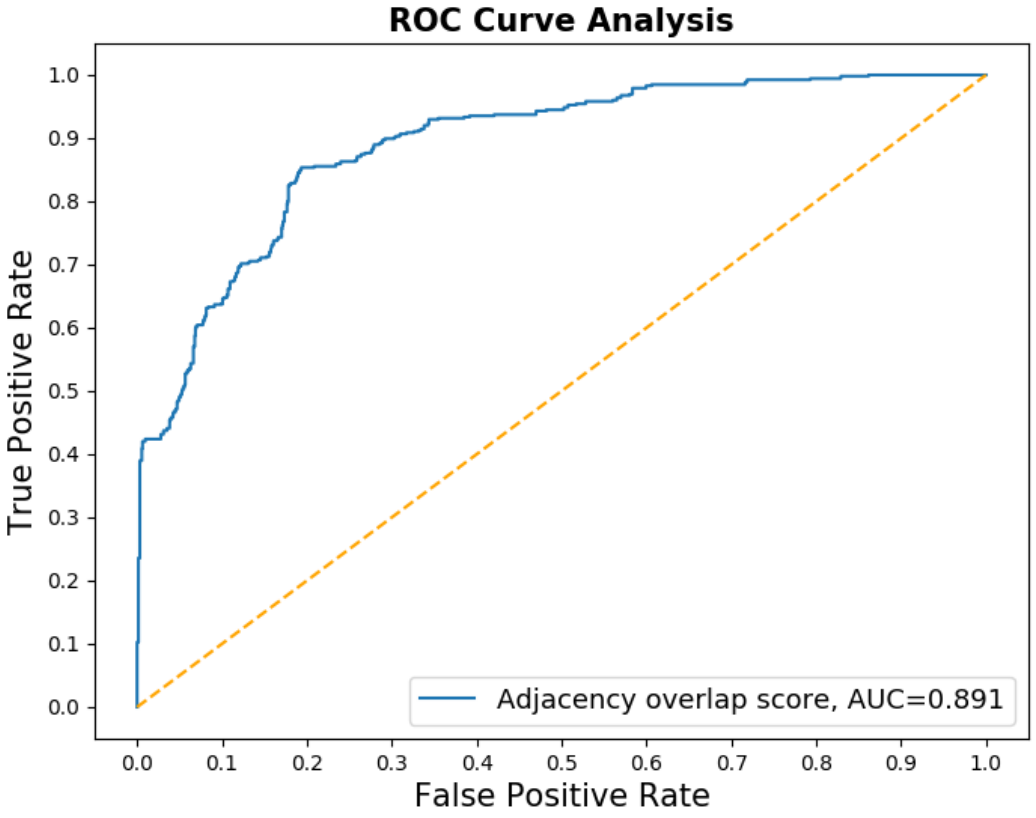
Supplementary Figure: ROC and Precision-Recall of the adjacency overlap scoring method on the S-Set for Bacteria



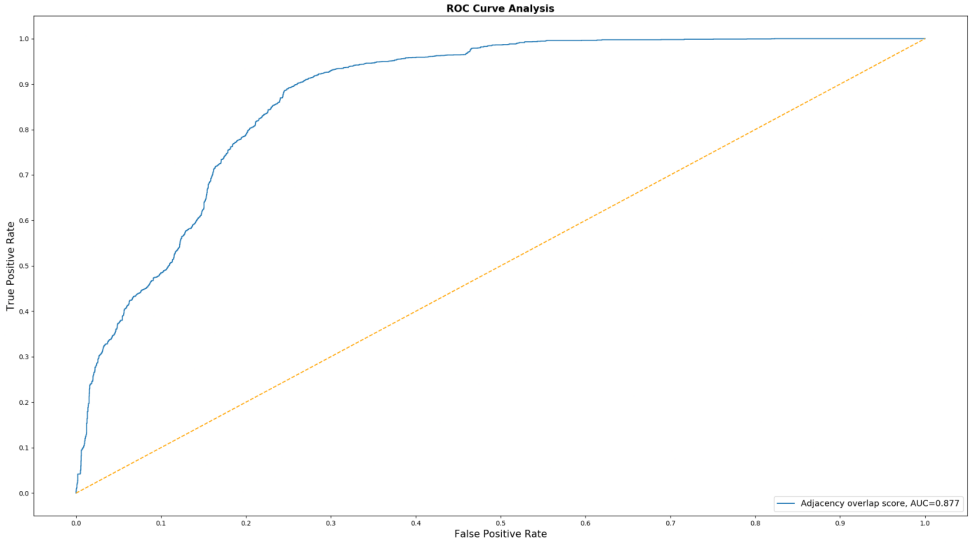


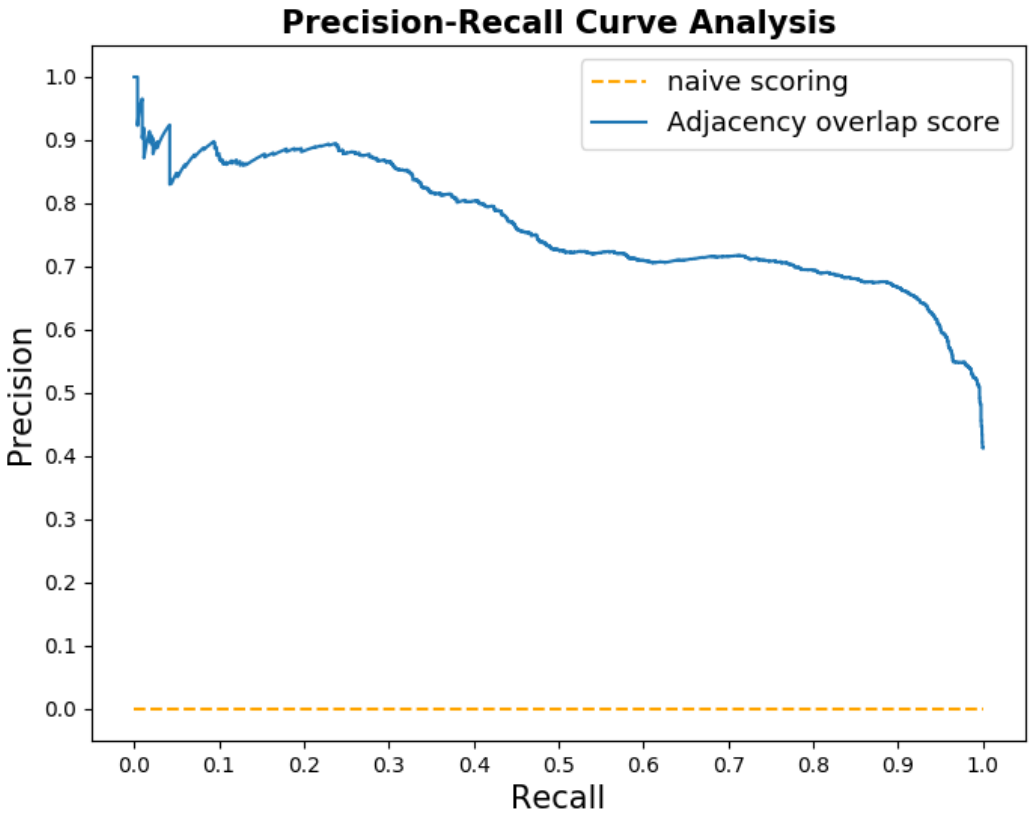
Supplementary Figure: ROC and Precision-Recall of the adjacency overlap scoring method on the S-Set for Eukaryotes



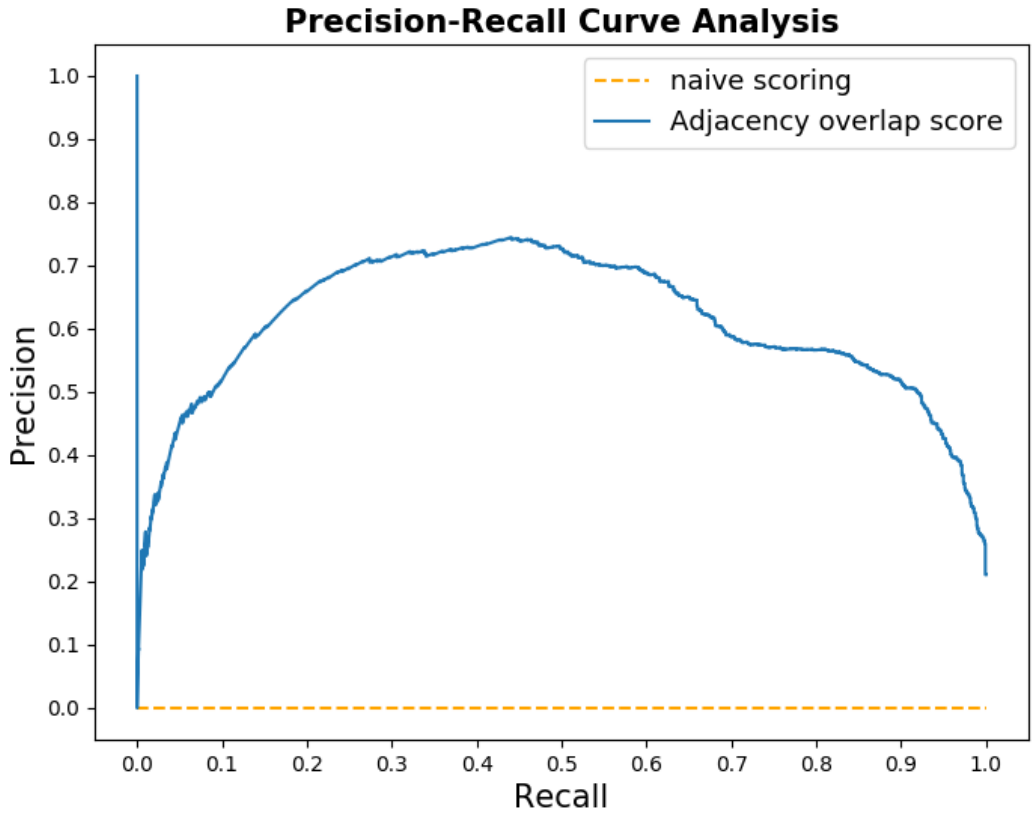


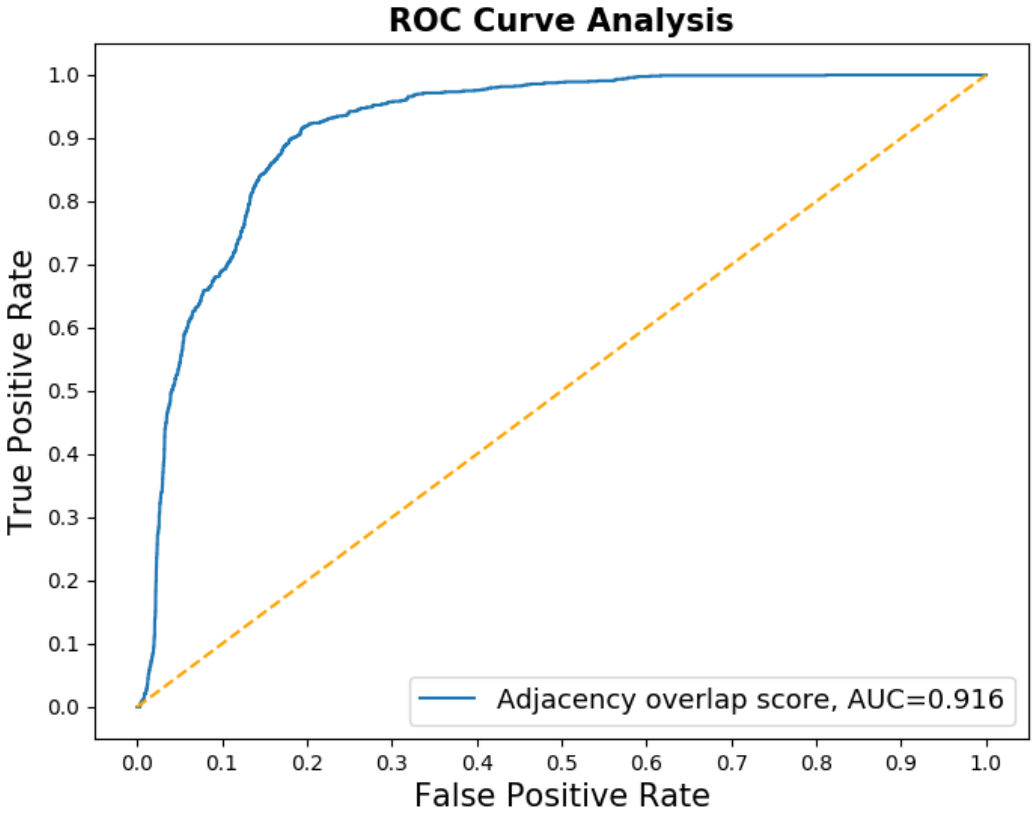
Supplementary Figure: ROC and Precision-Recall of the adjacency overlap scoring method on the S-Set for Viruses



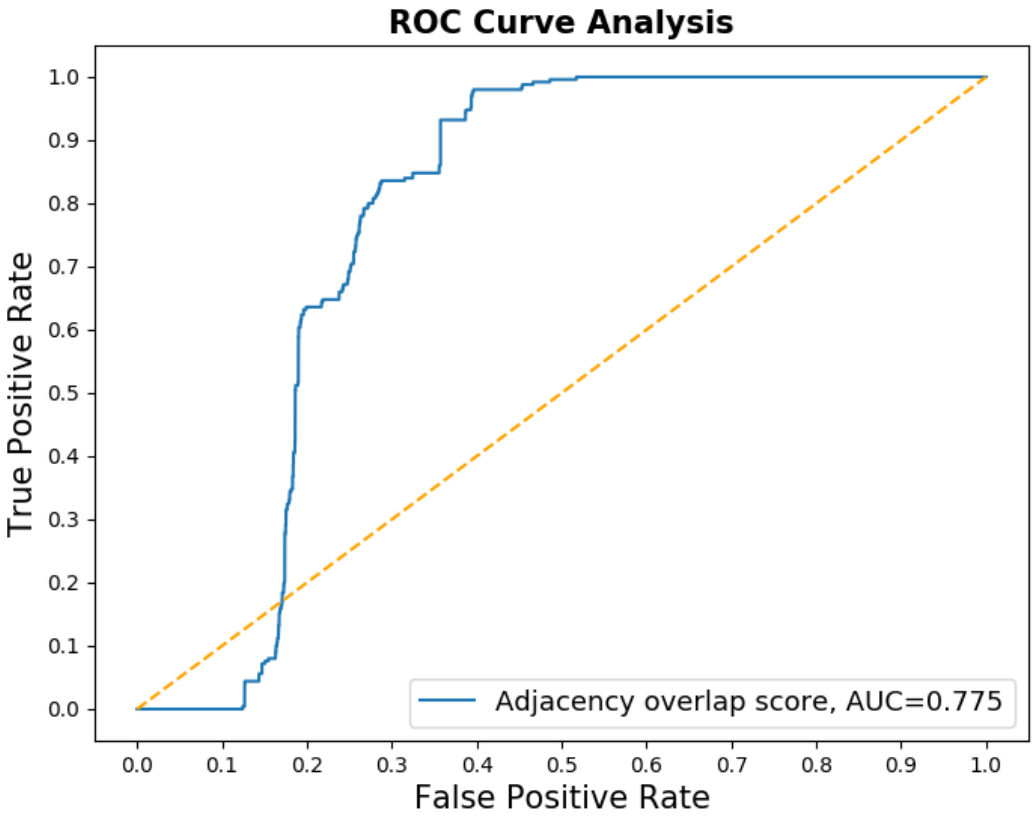


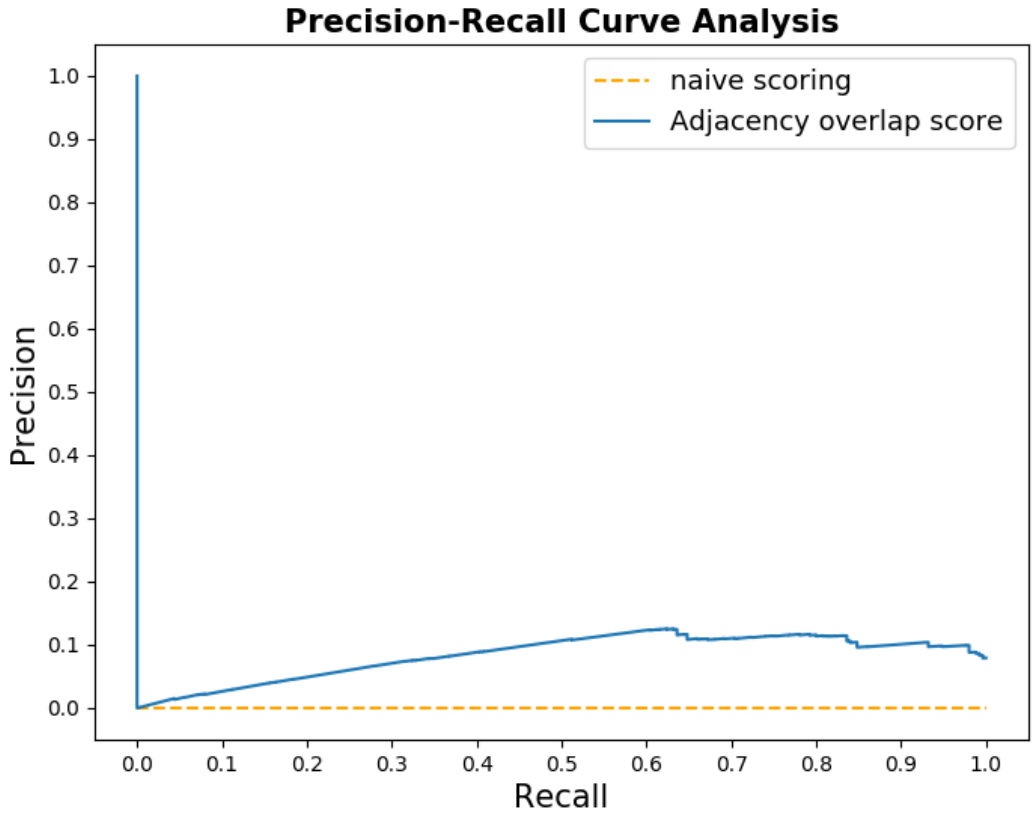
Supplementary Figure: ROC and Precision-Recall of the adjacency overlap scoring method on the S-Set for homo-dimer



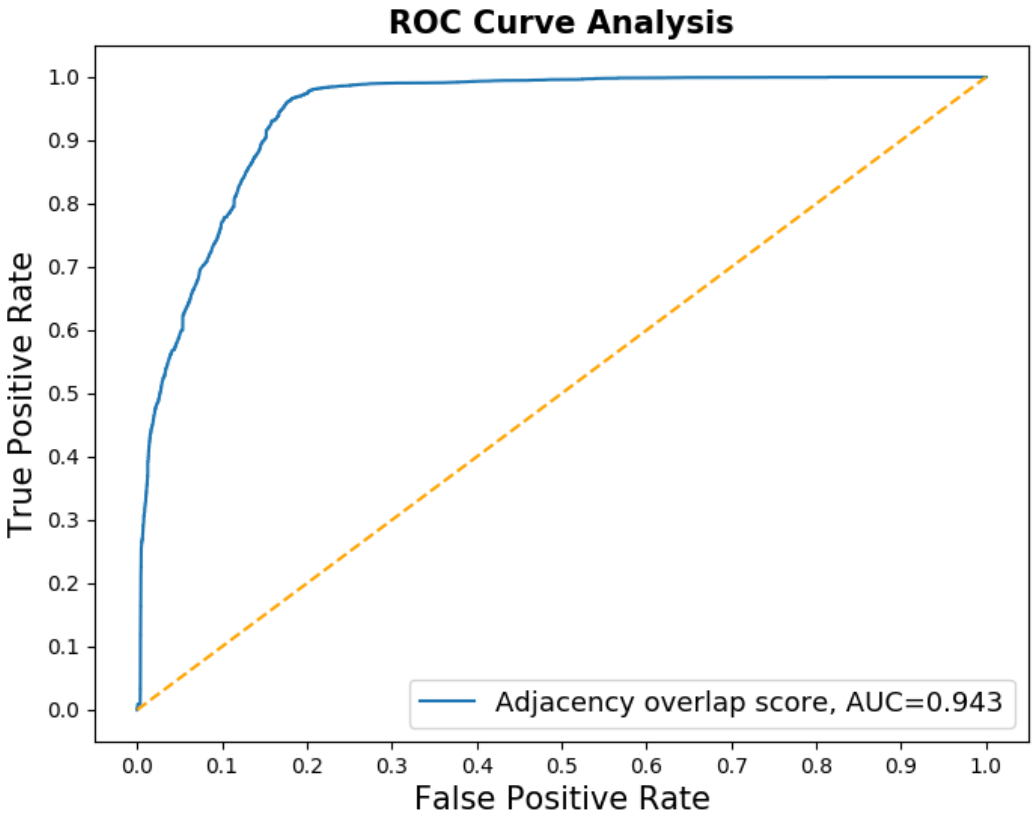


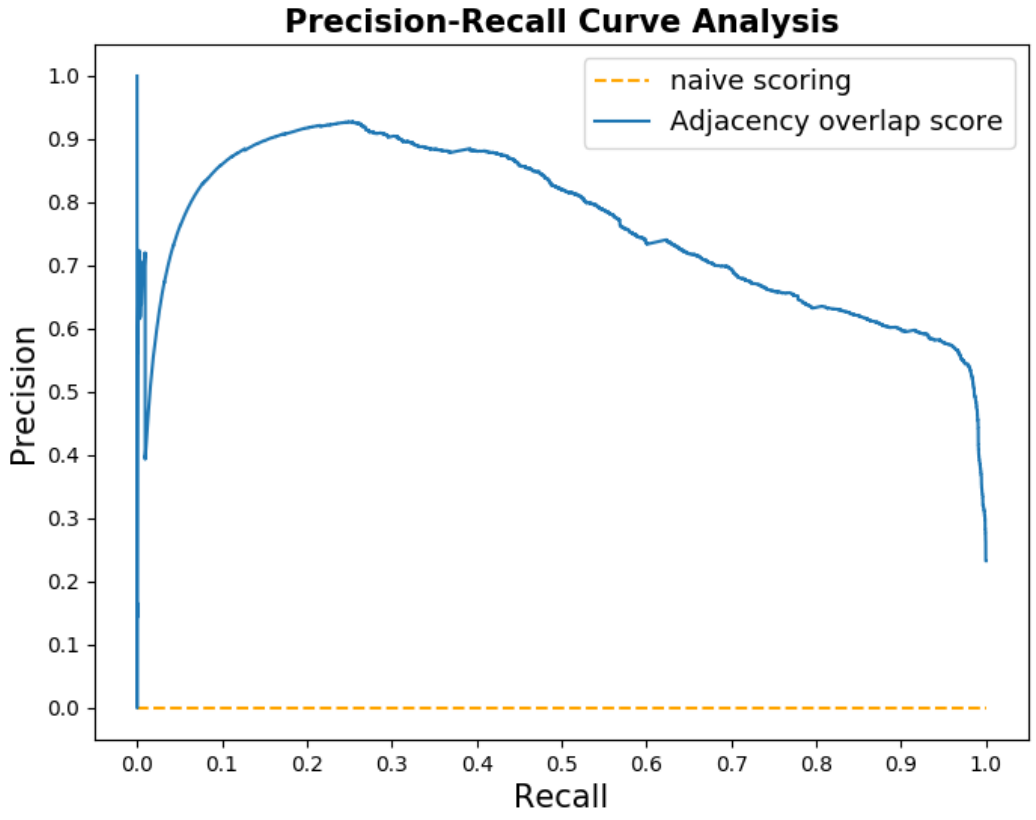
Supplementary Figure: ROC and Precision-Recall of the adjacency overlap scoring method on the S-Set for hetero-set



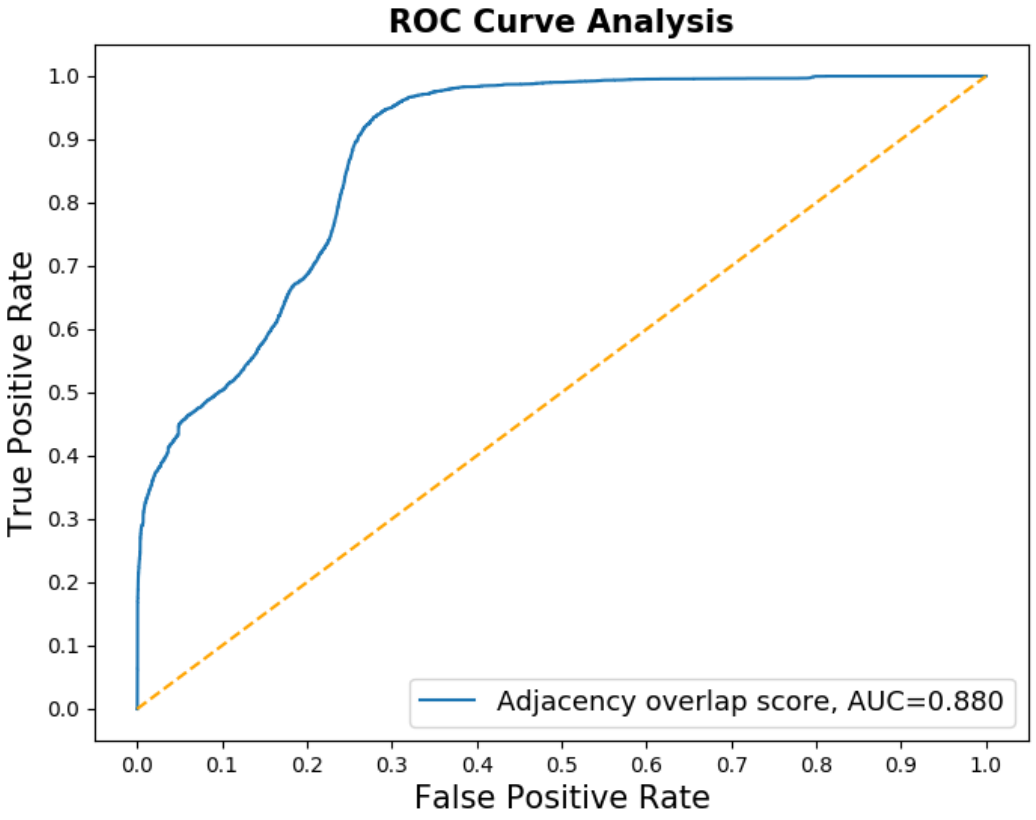


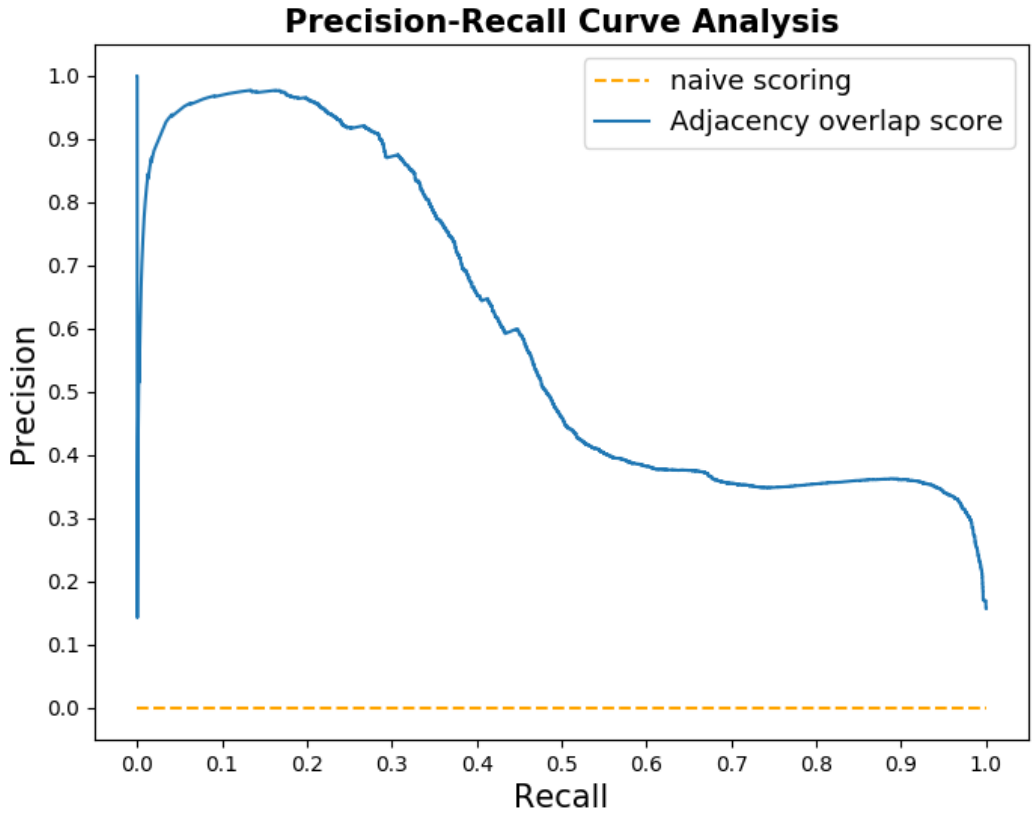
Supplementary Figure: ROC and Precision-Recall of the adjacency overlap scoring method on the U-Set for Archaea



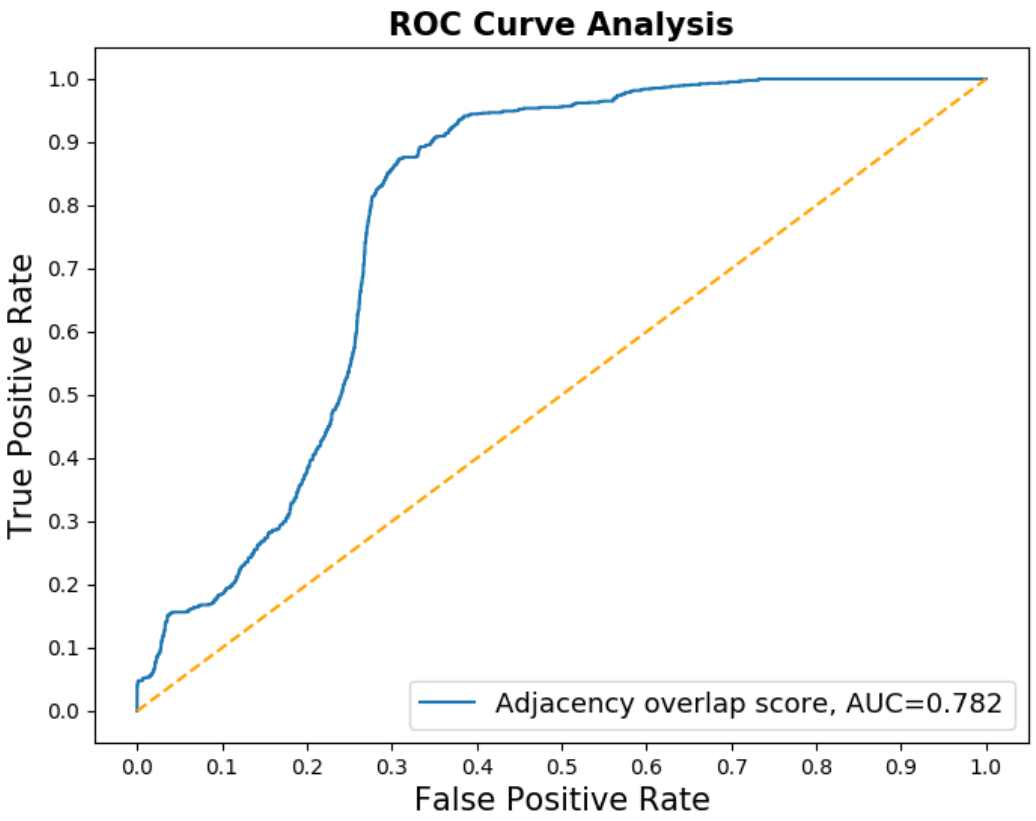


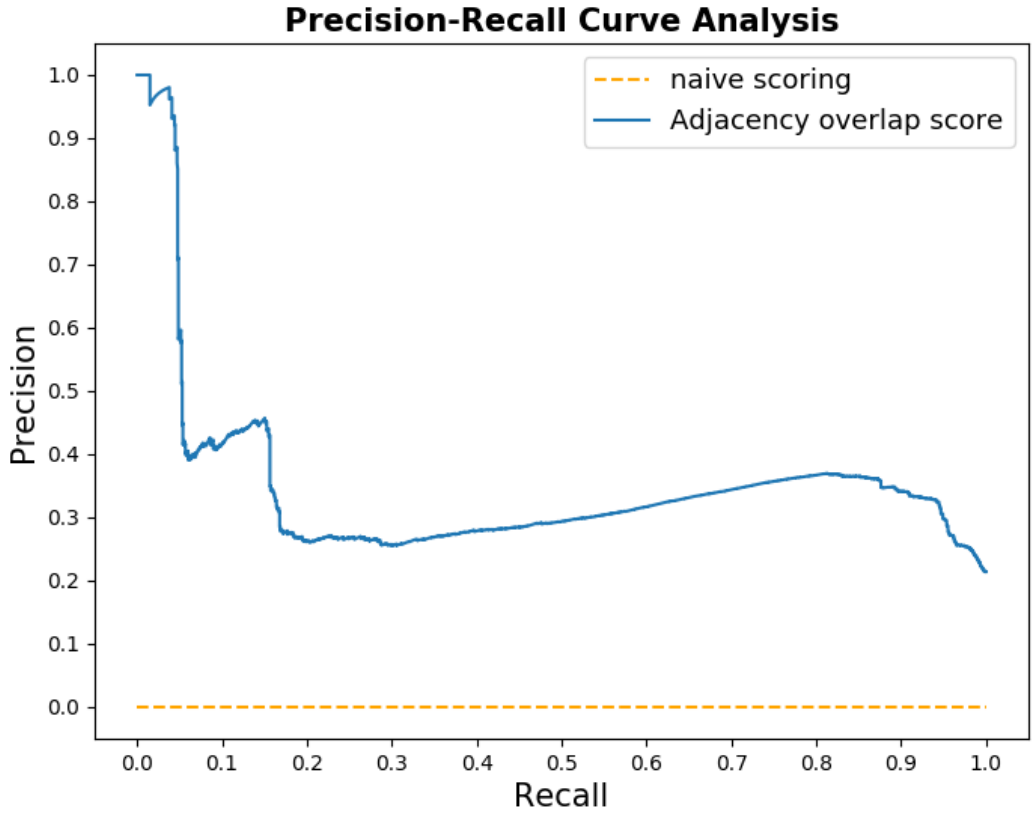
Supplementary Figure: ROC and Precision-Recall of the adjacency overlap scoring method on the U-Set for Bacteria



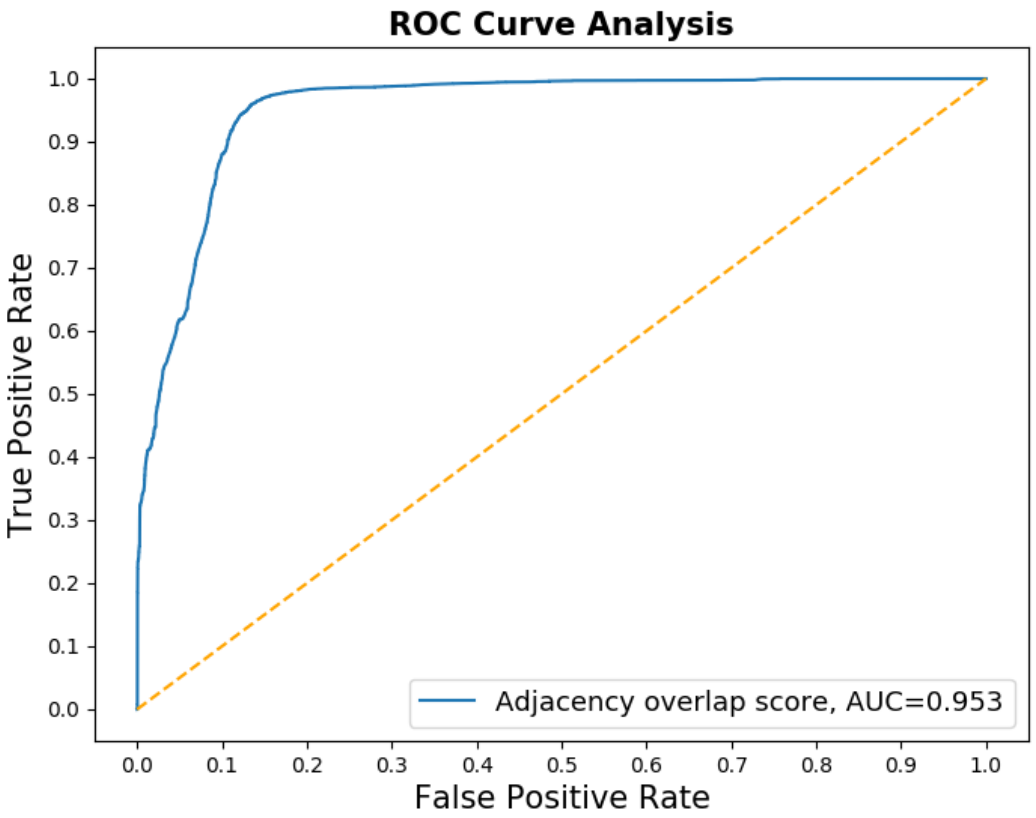


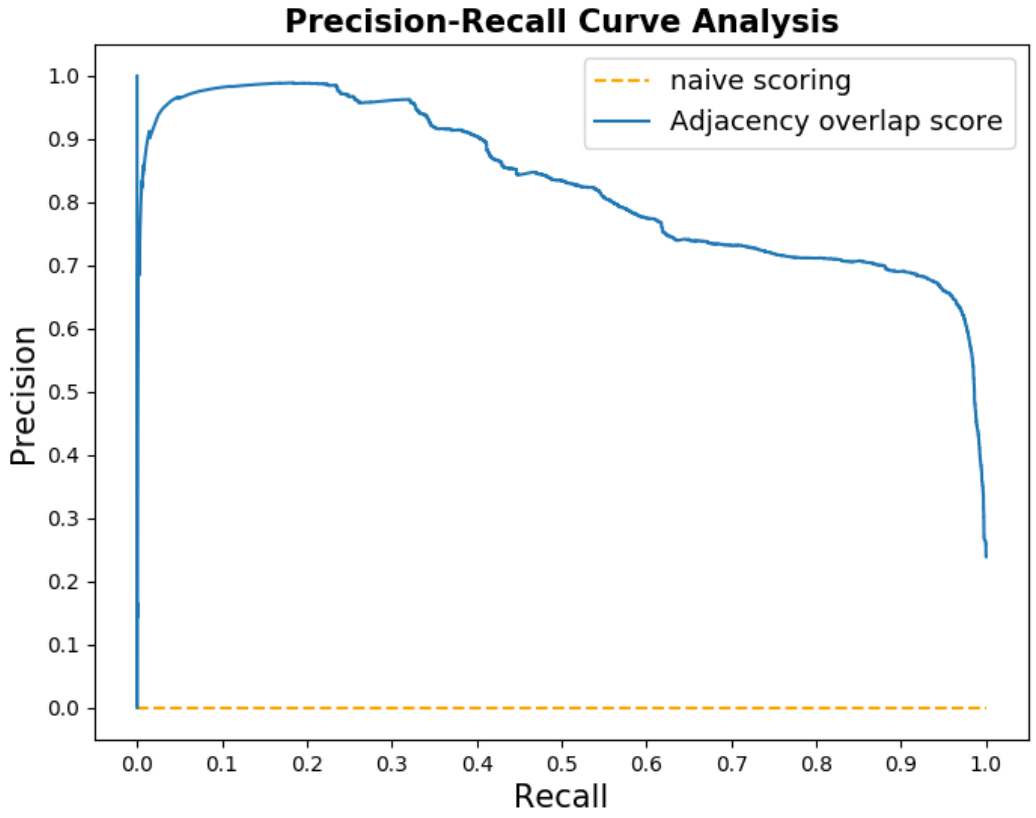
Supplementary Figure: ROC and Precision-Recall of the adjacency overlap scoring method on the U-Set for Eukaryotes



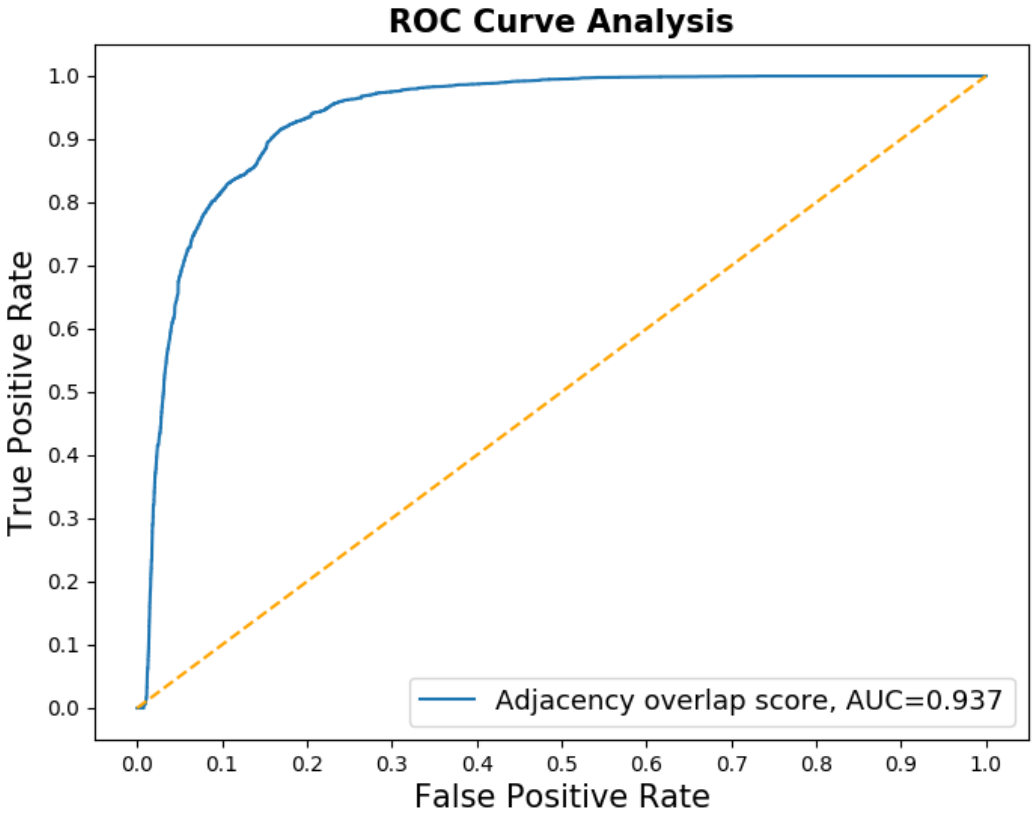


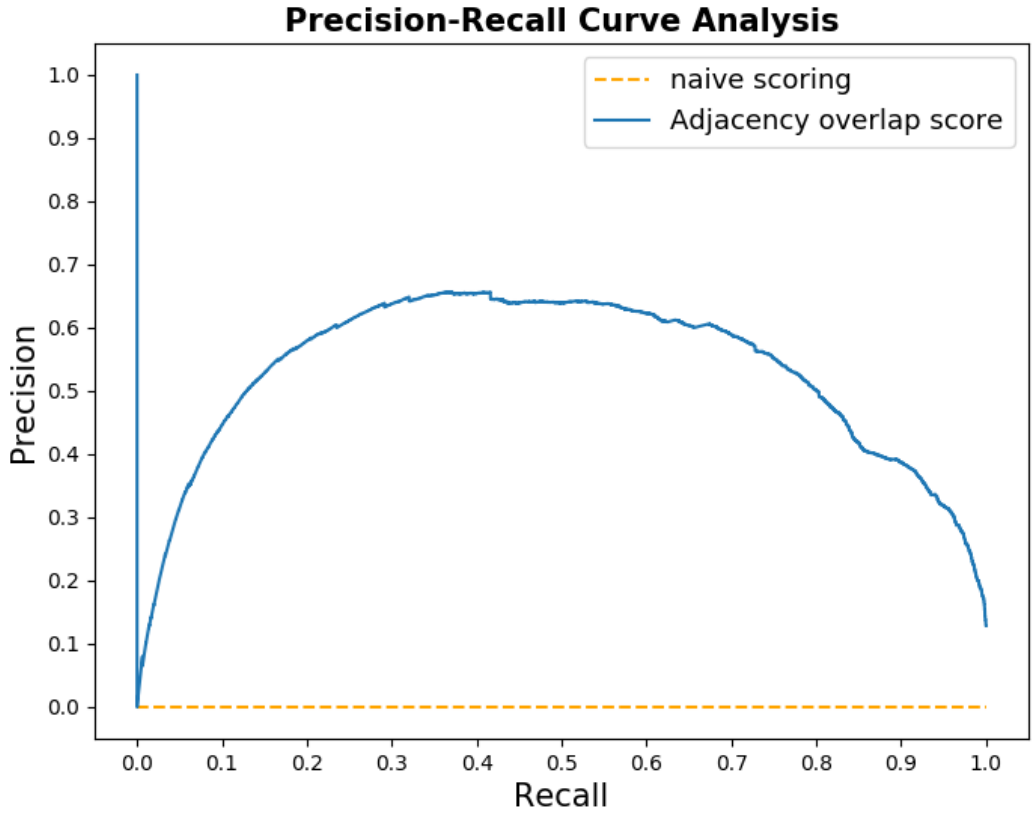
Supplementary Figure: ROC and Precision-Recall of the adjacency overlap scoring method on the U-Set for Viruses





Supplementary Figure: ROC and Precision-Recall of the adjacency overlap scoring method on the U-Set for homo-set





Supplementary Figure: ROC and Precision-Recall of the adjacency overlap scoring method on the U-Set for hetero-set