Functional molecular motions in drug design : application to nicotinic acetylcholine receptors

Mariana González Medina

Université Paris Cité
ED 474 : Frontières de l'Innovation en Recherche et Éducation
Unité de Bioinformatique Structurale

Functional molecular motions in drug design: application to nicotinic acetylcholine receptors.

Par Mariana GONZÁLEZ MEDINA

Thèse de doctorat de Biologie Moleculaire et Structurale et Biochimie, Biophysique Moleculaire

Dirigée par Dr. Arnaud Blondel
Et par Dr. Pierre-Jean Corringer

Présentée et soutenue publiquement le 1er Décembre 2022

Devant un jury composé de :

Arnaud Blondel	Dr.	Université Paris Cité	Directeur de thèse
Pierre-Jean Corringer	Dr.	Université Paris Cité	Co-Directeur de thèse
Véronique Stoven	Pr.	Université Paris Sciences et Lettres	Rapporteur
Gwenaëlle André-Leroux	Dr.	Universite Paris-Saclay	Rapporteur
Catherine Etchebest	Pr.	Université Paris Cité	Examinateur
Annick Dejaegere	PR.	Université de Strasbourg	Examinateur
Guillaume Bouvier	Dr.	Université Paris Cité	Invité

Acknowledgements

I would like to thank all the colleagues at Pasteur that got involved in the development of this thesis, in particular I would like to thank Dr. Guillaume Bouvier, Dr. Arnaud Blondel, Dr. Pierre-Jean Corringer and Dr. Max Bonomi for the scientific discussions that helped me to define and develop this project. I am sincerely thankful to Dr. Laura Ortega, Dr. Isaac Filella, Dr. Yoann Dufresne, Dr. Hélène Munier-Llehmann, Dr. Susanna Celli and Dr. Nadia Izadi-Pruneyre for their kindness and guidance through very difficult times.

I am always grateful to Prof. Robert Rice, Prof. José Luis Medina, Dr. Oscar Méndez, MPhil. John Owen and Prof. Jürgen Bajorath, for being the mentors that inspired and guided my academic path from the first years of the School of Chemistry until now. Everything you teached me and your support was fundamental for the completion of this work.

I strongly believe nothing would ever get done without the unconditional love of my family, that keeps me focused on the things that matter and are important in life; the laughs and tears I shared with Cami, Manu and Sophie and the love and patience with which Sylvain has encouraged me over the past years. This work is also dedicated to all of you.

Résumé

Mouvements moléculaires fonctionnels dans la conception de médicaments : application aux récepteurs nicotiniques del'acétylcholine.

L'étude des changements conformationnels se produisant au sein d'une protéine est essentielle pour comprendre sa fonction et la façon dont elle peut être modulée. Ainsi, ce projet de recherche a pour but d'étudier les changements conformationnels du récepteur nicotinique $\alpha_4\alpha_5\beta_2$, de proposer de nouvelles façons de moduler ce récepteur et de développer une méthode pour trouver de nouveaux modulateurs. L'intérêt pour le récepteur $\alpha_4\alpha_5\beta_2$ provient d'études d'association pangénomique associant le polymorphisme d'un seul nucléotide (PSN D398N) de la sous-unité α_5, à la dépendance à la nicotine et au cancer du poumon. En l'absence de structures expérimentales du récepteur nicotinique $\alpha_4\alpha_5\beta_2$, je ne disposais pas des informations structuralles nécessaires pour comprendre ses changements conformationnels, informations essentielles pour concevoir des médicaments ciblant des sites effecteurs spécifiques. Ainsi, comme première étape de ce projet, j'ai procédé à la modélisation in-silico du récepteur $\alpha_4\alpha_5\beta_2$ dans les états de repos, activé et désensibilisé. Après avoir obtenu des modèles satisfaisants, j'ai calculé le chemin de transition conformationnel entre les états modélisés et j'ai utilisé ce chemin conformationnel pour effectuer une analyse des cavités. L'analyse des cavités sur le chemin de transition a pré-validé les modèles en confirmant la présence de sites orthostériques et allostériques connus et j'ai identifié de nouveaux sites effecteurs plausibles qui peuvent être exploités pour moduler le récepteur nicotinique $\alpha_4\alpha_5\beta_2$. Afin de proposer des modulateurs pour ces nouveaux sites, j'ai entrepris de collecter des données de ligands et leur poche de liaison sur leur protéine cible, puis j'ai entraîné une méthode d'apprentissage profond pour générer des ligands complémentaires pour une poche de protéine donnée. En outre, pour aborder la question de la synthèse des composés générés in-silico, j'ai intégré une série de scores d'accessibilité synthétique qui m'ont permis de définir des seuils et des filtres pour poursuivre l'étude uniquement sur les composés générés in-silico ayant une voie synthétique prédite. Le modèle génératif entraîné peut être utilisé pour générer de nouveaux composés complémentaires aux poches du récepteur $\alpha_4\alpha_5\beta_2$ sélectionnés à partir de l'analyse des cavités, mais son utilisation peut également être généralisé à la génération de ligands ciblant d'autres protéines dans des projets de recherche futurs.

Mots clefs: Récepteur nicotinique, modélisation in-silico, chemin de transition conformationnel, apprentissage profond.

Abstract

Functional molecular motions in drug design: application to nicotinic acetylcholine receptors.

The study of the conformational changes occurring in a protein structure is essential to understand its function and how a protein can be modulated. This research project was started with the aim of studying the gating cycle of the $\alpha_4\alpha_5\beta_2$ nAChR, propose new ways of modulating this receptor and to develop a method to find novel modulators. The interest on the $\alpha_4\alpha_5\beta_2$ receptor comes from GWAS studies that have associated a Single Nucleotide Polymorphism (SNP D398N) on the α_5 subunit, to nicotine addiction and lung cancer. Without experimental structures of the $\alpha_4\alpha_5\beta_2$ nAChR, I lacked the structural information to understand its gating mechanism, information essential to perform drug design in specific effector sites. Therefore, as the first step in the development of this project I performed the in-silico modeling of the $\alpha_4\alpha_5\beta_2$ nAChR in resting, activated and desensitized functional states. Once I obtained the nAChR models, I computed the conformational transition path between the modeled functional states and used this conformational path to do cavity analysis. The cavity analysis on the transition path pre-validated the models by confirming the presence of known orthosteric and allosteric sites and I identified novel plausible effector sites that can be exploited to modulate the $\alpha_4\alpha_5\beta_2$ nAChR. As an initial approach to propose modulators for these new proposed cavities, I collected and curated the structural data of protein-ligand pairs and trained a deep learning method to produce ligand shapes and chemical compounds, complementary to protein pockets. In addition, to tackle the issue of how to synthesize in-silico generated compounds, I integrated a series of synthetic feasibility scores that allowed me to set thresholds and filters to select in-silico generated compounds with an hypothetical synthetic route available. The trained generative model could be used to generate novel compounds complementary to the selected $\alpha_4\alpha_5\beta_2$ nAChR pockets extracted from the cavity analysis, but it could also be used to generate ligands for other proteins in future research projects.

Keywords: Nicotinic receptor, in-silico modeling, transition path, deep learning.

Résumé substantiel

L'étude des changements conformationnels se produisant dans la structure d'une protéine est essentielle pour comprendre sa fonction et comment une protéine peut être modulée. Ce projet de recherche a été conçu dans le but d'étudier le cycle d'ouverture/ fermeture du récepteur nicotinique de l'acétylcholine $\alpha_4\alpha_5\beta_2$, de proposer de nouvelles façons de moduler ce récepteur et de développer une méthode pour trouver de nouvelles petites molécules qui pourraient agir comme modulateurs allostériques positifs (PAM) et être utilisés pour traiter la dépendance à la nicotine. L'intérêt pour le récepteur $\alpha_4\alpha_5\beta_2$ provient des études d'association pangénomique qui ont associé un polymorphisme d'un seul nucléotide (SNP) sur la sous-unité α_5 résultant en un échange d'aspartate par de l'asparagine (SNP D398N), à un risque plus élevé de développer une dépendance à la nicotine et un cancer du poumon.

Sans les structures expérimentales du récepteur nicotinique de l'acétylcholine $\alpha_4\alpha_5\beta_2$, il me manquait les informations structurales pour comprendre son mécanisme d'ouverture/ fermeture, informations essentielles pour réaliser la conception de médicaments dans des sites effecteurs spécifiques. Par conséquent, comme première étape dans le développement de ce projet, j'ai effectué la modélisation in-silico du récepteur nicotinique $\alpha_4\alpha_5\beta_2$ dans les états fonctionnels au repos, activé et désensibilisé en utilisant comme modèles les structures homologues publiées de $\alpha_3\beta_4$, $\alpha_4\beta_2$, α_7 ainsi que le récepteur nicotinique de type musculaire. Les modèles obtenus pour le récepteur nicotinique $\alpha_4\alpha_5\beta_2$ sont conformes aux observations structurales rapportées pour les structures expérimentales des récepteurs nicotiniques homologues et d'autres canaux ioniques pentamériques à libération par ligand. En outre, le profil du pore ionique délimité par l'hélice M2 du domaine transmembranaire (TMD) indique que les modèles se trouvent dans les états fonctionnels souhaités : le modèle de récepteur activé a un pore ionique ouvert d'un diamètre minimal de 7,7 Å, le modèle de récepteur au repos a un pore ionique fermé au milieu du TMD d'un diamètre minimal de 3,0 Å et le modèle désensibilisé, a un pore partiellement fermé à l'extrémité inférieure du TMD d'un diamètre minimal de 4,7 Å.

Ensuite, j'ai utilisé une méthode appelée Path Optimization Exploration (POE) pour échantillonner la surface du paysage d'énergie potentielle et trouver les chemins d'énergie minimale reliant les trois états fonctionnels modélisés à faible énergie. Ces chemins d'énergie minimale comportaient 67 conformations qui décrivent un modèle du cycle d'ouverture/ fermeture du récepteur nicotinique $\alpha_4\alpha_5\beta_2$ lorsqu'il est activé, désensibilisé ou désactivé, et c'est la première fois que l'état désensibilisé est inclus dans ce type d'analyse. Avec le cycle d'ouverture/ fermeture du récepteur nicotinique $\alpha_4\alpha_5\beta_2$, j'ai pu décrire et comparer les changements structurels qui ont été rapportés pour d'autres récepteur nicotiniques et qui sont connus pour être essentiels pour réguler et initier l'ouverture du pore ionique : la compaction du domaine extracellulaire (ECD) et la torsion entre l'ECD et le TMD. Lorsque le récepteur nicotinique $\alpha_4\alpha_5\beta_2$ passe de la conformation activée à la conformation au repos ou désactivée, la boucle-C s'ouvre et le ECD se relâche. J'ai également observé la torsion globale attendue dans le sens inverse des aiguilles d'une montre de l'ECD par rapport à le TMD, lorsque le récepteur nicotinique $\alpha_4\alpha_5\beta_2$ devient activé. Pour déterminer si le relâchement de l'ECD précède la torsion entre l'ECD et le TMD qui conduit à l'ouverture du pore ionique, j'ai analysé la synchronisation entre ces changements de conformation. J'ai observé qu'il y a une corrélation inverse entre la torsion et l'épanouissement : lorsque le récepteur est complètement relâché, la torsion entre les domaines est à sa valeur la plus basse. Les mouvements de relâchement et de torsion semblent se produire simultanément, lorsque le

récepteur passe d'une conformation activée à une conformation de repos. Cependant, le relâchement et la torsion ne semblent pas être synchronisés entre l'état de désensibilisation et l'état de repos, car le récepteur commence d'abord par se détordre, puis la ECD se contracte. Dans l'ensemble, le relâchement et la torsion de la ECD sont conformes à ce qui a été décrit pour les structures expérimentales homologues de récepteur nicotiniques utilisées comme modèles.

Le cycle de ouverture/ fermeture a été utilisé comme entrée à mkgridXf, pour faire une analyse de cavité sur le récepteur nicotinique $\alpha_4\alpha_5\beta_2$. L'analyse des cavités sur le cycle d'ouverture/ fermeture a pré-validé les modèles en confirmant la présence du site de liaison orthostérique connu dans l'interface ECD entre les sous-unités α_4 et β_2 et du site allostérique dans l'interface ECD entre les sous-unités α_5 et α_4, les deux sites étant recouvert par la boucle-C. J'ai obtenu 255 cavités qui sont formées au moins une fois le long de la trajectoire. Sur ces 255 cavités, seules 52 avaient au moins une fois au cours de la trajectoire, un volume supérieur à 150 Å3 et ont été sélectionnées une l'analyse typologique. L'objectif de l'analyse typologique était de classifier des types d'évolutions géométrique en fonction de l'état fonctionnel et donc d'identifier de nouvelles cavités allostériques potentielles qui pourraient être modulées avec un PAM. Les cavités ont été regroupées en utilisant le regroupement hiérarchique de leurs volumes pour chaque conformation du cycle d'ouverture/ fermeture/ désenssibilisation comme descripteurs. J'ai sélectionné les cavités qui montrent un volume plus élevé sur l'état activé et les cavités qui étaient situées dans la sous-unité α_5 ou à l'interface entre la sous-unité α_5 et une autre sous-unité. Cette sélection a été faite avec l'hypothèse qu'en modulant ces cavités et en conservant leurs volumes élevés, il devrait être possible de maintenir le récepteur nicotinique $\alpha_4\alpha_5\beta_2$ dans une conformation active et si ces cavités sont situées près de ou dans α_5, les composés conçus pour les moduler devraient être sélectifs aux récepteurs avec cette sous-unité. Le premier site de liaison nouvellement identifié est situé dans la ECD de α_5 et fait face au pore ionique. Les deux autres nouveaux sites sont dans le TMD de α_5 et dans l'interface entre le TMD de α_5 et de β_2. J'ai pris les résidus délimitant ces cavités comme définition de poches et utilisé les informations sur ces poches comme entrée pour la génération de composés de novo.

Comme approche pour proposer des modulateurs pour ces poches nouvellement proposées, j'ai entraîné une méthode d'apprentissage profond pour générer des composés avec des propriétés moléculaires complémentaires aux propriétés des poches protéiques. Le modèle a été publié par le groupe de Gianni de Fabritis et est formé de deux composantes : un réseau de génération de formes et un réseau annotateur. Le réseau bicycle antagoniste génératif (bicycleGAN) conçu pour empêcher le réseau générateur entraîné de rester sur un minimum local ou de souffrir d'un effondrement sur un seul mode tout en produisant une distribution diversifiée de sorties. Ce réseau de génération de formes est entraîné avec une base de données de paires protéine-ligand que j'ai extraites, classées et arrimées (docked) et qui incluait plusieurs modulateurs de récepteur nicotiniques. La base de données comprend 68 enzymes, 35 protéases, 32 kinases, 27 GPCRs, 17 récepteurs nucléaires et 10 canaux ioniques parmi lesquels : α_7, $\alpha_4\beta_2$, $\alpha_3\beta_4$ et le récepteur GABA-A. Une fois le réseau de génération de formes entraîné, j'ai pu l'utiliser pour générer des formes de ligands complémentaires aux poches des protéines. Les formes décrivent les propriétés des atomes présents dans le ligand telles que la localisation en 3D des carbones hydrophobes aliphatiques, des carbones hydrophobes aromatiques, des donneurs de liaisons hydrogène et des accepteurs de liaisons hydrogène. La deuxième partie du modèle génératif de novo est un réseau an-

notateur composé de deux auto-encodeur variationnels (VAE). Le premier VAE est utilisé exclusivement pendant l'entraînement pour ajouter du bruit aux composés chimiques 3D dont les coordonnées sont traduites en une représentation en grille utilisée pour entraîner le second VAE. L'encodeur du deuxième VAE est un réseau de neurones convolutif (CNN) qui codera la forme du ligand dans une représentation latente vectorielle. Ensuite, le décodeur, qui est un réseau "longue mémoire à court terme" (LSTM), utilisera la représentation codée de la forme du ligand et la traduira en une séquence de SMILES chimiquement correcte. Pour entraîner ce réseau, j'ai utilisé une base de données 3D interne du laboratoire, déjà prétraitée pour le docking, qui contient des structures moléculaires extraites de la banque MolPort et compte 5 600 000 composés qui ont été traités et stockés dans un format de données binaires HDF5.

Avant de générer des composés moléculaires pour les cavités sélectionnées, j'ai effectué quelques tests d'évaluation sur les modèles entraînés. J'ai évalué si les formes des ligands générés pour une même poche protéique étaient différentes et si le réseau annotateur était capable de décoder efficacement les formes générées. Pour ce faire, j'ai sélectionné quatre protéines pour lesquelles je disposais d'au moins 900 composés ayant une activité biologique annotée (que nous appellerons ligands) : le récepteur de la progestérone, l'anhydrase carbonique 2, le transporteur de sérotonine sodium-dépendant et le récepteur intracellulaire sigma non-opioïde 1 . Les protéines ont été traduites en une représentation en grille et données en entrée au réseau de génération de formes avec un vecteur échantillonné aléatoirement à partir d'une distribution normale standard. 100 formes de ligands complémentaires ont été générées pour chacune des 4 poches protéiques et comparées par paires en utilisant le coefficient de corrélation de Pearson. J'ai trouvé que le bicycleGAN générait une distribution de formes de ligands qui étaient très similaires les unes aux autres. Cette constatation suggère que le générateur de bicycleGAN a souffert d'un effondrement de mode ou que, puisque la complémentarité entre les propriétés moléculaires des protéines et des ligands est limitée (les donneurs de liaisons hydrogène interagissent avec les accepteurs de liaisons hydrogène), il n'est pas nécessaire d'utiliser un modèle génératif aussi complexe puisqu'il n'est pas possible d'obtenir une large distribution de résultats. J'ai observé que les formes de ligands générés avaient beaucoup d'atomes hydrophobes aliphatiques et seulement quelques atomes hydrophobes aromatiques et ce comportement a été correctement appris par le réseau annotateur pour lequel les SMILES générés avaient principalement des cycles aliphatiques et un nombre inférieur de cycles aromatiques. Ensuite, une forme de ligand de chaque protéine a été décodée en 10 000 SMILES différents par le réseau annotateur. J'ai calculé le nombre de donneurs de liaison hydrogène, d'accepteurs de liaison hydrogène, le nombre de cycles aromatiques, le nombre de cycles aliphatiques, le nombre de liaisons rotatives, le nombre d'halogènes, le poids moléculaire, la surface polaire topologique (TPSA) et le coefficient de partage d'une molécule entre les phases aqueuse et lipophile (LogP), pour les composés générés et les liants et j'ai comparé leurs distributions. De cette analyse, j'ai observé que les composés générés ont en moyenne un poids moléculaire plus élevé, plus de cycles aliphatiques, d'halogènes, d'accepteurs de liaisons hydrogène, de donneurs de liaisons hydrogène et de liaisons libre en rotation. Ce qui montre que ces composés sont plus gros que les liants réels annotés. Lorsque j'ai comparé la distribution des longueurs des SMILES canoniques générés à la longueur des SMILES canoniques des ligands connus, j'ai observé que les composés générés ont des SMILES avec plus de symboles. En moyenne, les SMILES ont une longueur comprise entre 60 et 70 symboles.

Pour filtrer les composés générés, j'ai décidé d'utiliser la faisabilité synthétique comme critère principal. J'ai utilisé trois scores pour classer les composés générés : le score d'accessibilité synthétique (SA), le score d'accessibilité rétrosynthétique (RA) et le score de complexité synthétique (SC). Je n'ai sélectionné que les composés dont le score RA était supérieur à 0,7 (70% de probabilité de trouver une voie rétrosynthétique avec AiZynthFinder) et les scores SA et SC inférieurs à 3,5 (un score de 1 pour les deux méthodes indique que le composé est facile à synthétiser). Ces seuils de score ont démontré que la plupart des composés générés ne pouvaient pas être synthétisés et se sont avérés être un critère de filtrage très strict. Après ces étapes de filtrage, les composés générés sélectionnés ont été traités avec AiZynthFinder et ceux pour lesquels une voie rétrosynthétique a été trouvée ont été dockés et sont présentés dans ce manuscrit de thèse. A partir des composés dockés, il est possible de noter que les modèles génératifs entraînés performent beaucoup mieux en générant des composés pour des poches larges et exposées aux solvants, comme c'est le cas pour l'anhydrase carbonique 2 ou sur l'α_5, face au pore ionique. Globalement, j'ai eu plus de difficultés à trouver des ligands adéquats pour les cavités hydrophobes et plus petites dans le TMD de l'α_5. Cela est probablement dû au fait que le réseau génère des composés avec, en moyenne, aucun ou seulement un cycle aromatique par molécule. Pour résoudre ce problème, il serait possible de mettre en place une procédure pour continuer à générer des composés jusqu'à ce qu'un nombre de suffisant composés avec les propriétés souhaitées aient été générés.

Ce travail de recherche a produit des modèles de récepteur nicotinique $\alpha_4\alpha_5\beta_2$ dans trois états fonctionnels : état de repos, état activé et état désensibilisé. Je présente et décris également le cycle d'ouverture/ fermeture/ déssensibilisation du récepteur nicotinique $\alpha_4\alpha_5\beta_2$ qui n'avait pas été calculé ou décrit auparavant. Les modèles et les cycles d'ouverture/ fermeture/ déssensibilisation peuvent être utilisés dans de futurs projets de recherche de médicaments, pour étudier si la SNP D398N a un effet sur les changements structurels du récepteur nicotinique $\alpha_4\alpha_5\beta_2$ et pour analyser les interactions électrostatiques et hydrophobes qui ont lieu dans l'interface ECD-TMD et qui conduisent à l'ouverture du pore ionique. Une autre contribution de ce travail est la prédiction des trois cavités présentes sur la sous-unité α_5, uniquement présente à l'état activé. Les cavités dans le TMD sont particulièrement intéressantes, car il existe des preuves que les modulateurs allostériques positifs pourrait se lier dans un site de liaison allostérique dans le TMD. Enfin, le réseau génératif pourrait être utilisé pour générer des petites molécules de novo avec une voie rétrosynthétique qui pourraient être synthétisés et testés dans le cadre d'autres projets de découverte de médicaments.

Contents

List of Figures

List of Tables

List of Abbreviations

ADME-Tox absorption, distribution, metabolism, excretion and toxicity 11

AI Artificial Intelligence 11

CE Cross Entropy 59

CHL cholesterol 25

CNNs Convolutional Neural Network 11

cLR-GAN Conditional Latent Regressor GAN 14

COPD Chronic Obstructive Pulmonary Disease 5

COS chain-of-states 10

CPR Conjugate Peak Refinement 10

cVAE-GAN Conditional Variational Autoencoder - GAN 14

D discriminator 12

dmin minimum diameter 32

E encoder 12

ECD extracellular domain 2

G generator 11

GANs Generative Adversarial Network 11

GWAS genome-wide association studies 5

ICD intracellular domain 2

Ł generated ligand xvii

L real ligand xvii

LSTMs Long Short-Term Memory cells 12

MD Molecular dynamics 39

ML Machine learning 11

MSE Minimum Squared Error xvii

nAChRs nicotinic acetylcholine receptors 1

NAMS negative allosteric modulators 4

P protein xvii

PAMS positive allosteric modulators 4

PCA Principal Component Analysis 41

PDB Protein Data Bank 7

PLGICs Pentameric Ligan Gated Ion Channel 1

PLIP protein–ligand interaction profiler 70

POE Path Optimization Exploration 39

POPA 1-palmitoyl-2-oleoyl-phosphatidic-acid 25

Chapter 1

Introduction

In this chapter I introduce the general topic of pentameric ligand gated ion channels Pentameric Ligan Gated Ion Channel (PLGICs), the sub-types and their therapeutic applications as well as the experimental structures that have been resolved and their functional states. It was important to introduce them since I used some of these structures as templates to validate my work on the $\alpha_4\alpha_5\beta_2$ receptor. Then I focus on nicotinic acetylcholine receptors (nAChRs), their subunit composition and the orthosteric binding site. This led to the introduction of the concept of allosteric modulation, defined by the Monod Wyman-Changeux model that explains how the activation of nAChRs takes place and also all kinds of modulators that can bind and modulate these PLGICs. I describe how the $\alpha_4\alpha_5\beta_2$ nAChRs was chosen as a potential therapeutic target to treat nicotine addiction, from genome-wide association studies that related a Single Nucleotide Polymorphism (SNP) on α_5 to lung cancer and nicotine dependence. I present some of the cell, mice and structural studies that suggest this SNP could be producing a loss of function that leads to less activation of the receptor and the behavioral studies performed on mice showing that they self-administer more nicotine when they present this SNP. These observations lead to the hypothesis that if less activation of the receptor led to nicotine addiction, the receptor could be selectively activated to treat nicotine addiction. Since this research project was purely in-silico, I finished the introduction chapter explaining some of the computational methods that were key to the development of this work.

1.1 Pentameric Ligand Gated Ion Channels (PLGICs)

PLGICs are membrane proteins that function as modulators of electrochemical signals in the peripheral and central nervous system. [1, 2] PLGICs are protein systems of 150 to 300kDa that can be found in mammals, invertebrate insects and fish as well as in bacteria. [2, 3] They share an architecture of five subunits, with their rotational axis centered on an ion channel. [2] These five subunits can be homomeric and symmetrical or heteromeric and pseudo-symmetrical. Each subunit has an N-terminal extracellular domain (ECD), with the binding site for agonists that is made of 10 β strands, a transmembrane domain (TMD) composed by four membrane-spanning α helices (M1-M4), with the pore being formed and delineated by TM2 and an intracellular domain (ICD) that contributes to many different functions including: channel conductance and desensitization, receptor trafficking, assembly and anchoring. [4] The ICD is present only in eukaryotes and it is composed by a disordered loop with highly variable sequence and length that connects the TM3 and TM4 helices (Figure 1.1). [5–7] They respond to the binding of an activator, with structural rearrangements that shift the receptor from the resting functional state (or closed ion pore), to the opening of the ion pore and the charge selective influx of ions to the cell. After prolonged exposure to an activator the receptor changes into a conformation with high affinity to the agonist and reduced ion conductance called desensitization. [3] The binding site of agonists or antagonists is called the orthosteric binding site. The orthosteric binding site is located on the ECD, in the solvent accessible area between the interface of two subunits. It is formed by the loops A-C, from the principal subunit (+) and D-F from the complementary subunit (-) and is capped by the loop-C. Once an effector binds to the orthosteric site, the loop-C closes, the ECD contracts and these structural changes initiated on the the ECD are transmitted to the TMD by a series of hydrophobic and electrostatic interactions forming and evolving in the interface between the ECD and TMD. These structural and chemical changes translate into rearrangements of the TMD, including the side chains of TM2 and the opening and closing of the ion pore. [2, 7, 8] In resting state, many studies performed on the experimental structures of PLGICs have reported a gate in the middle of the pore at positions 9' and 13', in the notation that counts from the N-terminus of the TM2. These studies suggest that hydrophobic residues on those positions constitute the main permeation barrier in the ion pore of most PLGICs. [9]

PLGICs are important therapeutic targets to treat addiction, anxiety, pain, schizophrenia as well as Alzheimer's disease and several mutations affecting their function have been associated with congenital pathologies including epilepsy and autism. [2] Scientist are well aware of their therapeutic relevance and big efforts have been invested into the study and determination of their experimental structures, which presents several technical challenges including the difficulty to overexpress and then solubilize or reconstitute in detergent or nanodiscs, large quantities of functional proteins. [7, 10] The first insights into the full length structure and function of PLGICs came from the cation selective prokaryotic receptors *Erwinia chrysanthemi* (ELIC) and *cyanobacterium Gloeobacter violaceus* (GLIC), which were the first high resolution structures to be experimentally resolved with resolutions of 3.3 Å and 3.1 Å in activated and resting states. [11, 12] In addition, a locally closed conformation of GLIC that occurs during allosteric gating transitions was also published. [13] Among eukaryotes, the first structure to be published was that of the anion selective C.

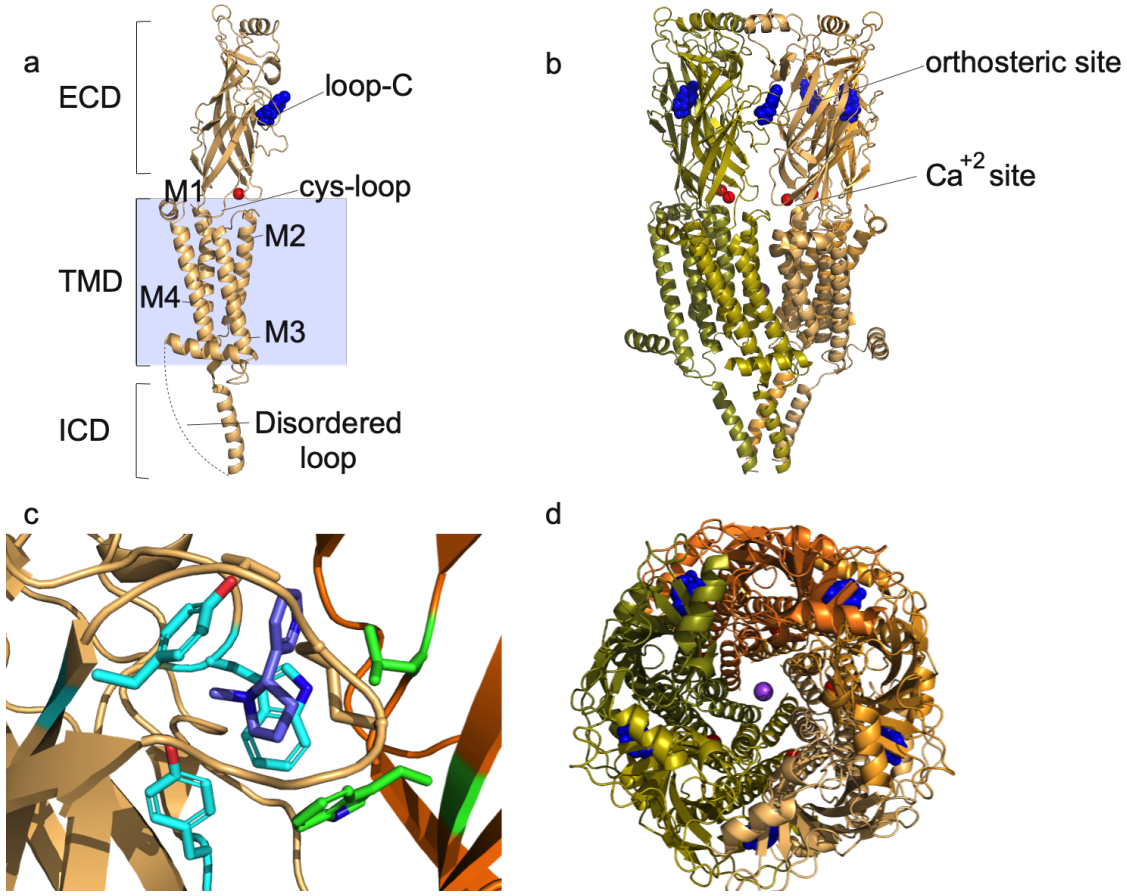

Figure 1.1: **Structure of a Pentameric Ligand Gated Ion Channel (PLGIC), composed of 5 nicotinic α_7 subunits.** a) PLGICs have an Extracellular Domain (ECD) , a Transmembrane Domain (TMD) and an Intracellular Domain (ICD). b) Pentameric assembly of PLGICs and upper view showing the ion pore. c) Orthosteric binding site between two subunits in the ECD.

elegans (GluCl) receptor in activated state, at a resolution of 3.3 Å. [14]. At the time I started this research project and up to this date, several other eukaryotic structures have been resolved, including the anion selective glycine receptor (GlyR) α_1 in open, resting and desensitized states, [15–17] GlyR α_3 in resting and partially open or desensitized states [18, 19]. Some GlyR structures have also been published in pre-open states, these structures provide insights into the reaction pathway of the receptor. [17] The gamma-aminobutyric acid receptor ($GABA_nR$) β_3 in desensitized and resting states [20, 21], the cation selective serotonin receptor ($5\text{-}HT_{3A}R$) in closed and open states as well as other structures with intermediate profiles that could be described as: pre-active closed state or a desensitized state with a closed pore occurring downstream from the activated state [22–25], several nAChRs including neuronal $\alpha_3\beta_4$ [26] and $\alpha_4\beta_2$ [27–29] in desensitized stated and α_7 [8, 30], the first nicotinic receptor resolved in open, resting and desensitized conformations, as well as the muscle type *Torpedo* nAChR in resting state. [31, 32]

1.2 Nicotinic Acetylcholine Receptors (nAChRs)

nAChRs regulate the neuroexcitation of synaptic membranes by converting the chemical signal of the neurotransmiter acetylcholine, into an electrical signal controlled by the entrance of cations (sodium, potassium and calcium), after the opening of the channel pore. [7, 33] They are important therapeutic targets pursued to treat addiction, depression and pain, as well as to improve cognition and neuroprotection for Parkinson's and Alzheimer's disease. [34]. Their function and location, in the central and peripheral nervous system, is determined by their subunit composition, which can include different combinations of 17 known subunits: $\alpha_1 - \alpha_{10}$, $\beta_1 - \beta_4$, γ, δ and ϵ. Neuronal types can be heteromeric and constituted by a combination of α and β subunits, or homomeric with only α subunits like α_7 and α_9. Muscle types are formed by α_1, β_1 as well as γ, δ and ϵ subunits. [35] nAChRs with heteromeric composition have 2 orthosteric binding sites, located in the interface between α and β subunits. Homomeric nAChRs have 5 orthosteric or acetylcholine binding sites, in the interface of $\alpha_7 - \alpha_7$ and $\alpha_9 - \alpha_9$ subunits. [36, 37] These acetylcholine binding sites in the ECD, are largely separated and at the same time, functionally linked to the ion pore in the TMD, located at a 60 Å distance from the orthosteric sites. This distinct regulation of the ion pore by topographically different and distant acetylcholine binding sites, is why we recognize nAChRs as allosteric machines. [33, 37–39] The Monod Wyman-Changeux model has been fundamental to our understanding of acetylcholine's allosteric modulation of nAChRs. [40] Following this model statements, and from experimental data, [41] we know that nAChRs can spontaneously shift between resting and activated conformations and that each conformation has a distinct affinity to the allosteric modulator. More specifically, the activated conformations have a higher affinity to acetylcholine or any other activator, as compared to the resting conformation, and the probability of the channel being in an ion-permeable conformation increases as the occupation of the 2 to 5 different orthosteric sites gets fulfilled. [38, 40]

In addition to the orthosteric site, nAChRs contain distinct sites that can be allosterically modulated to either help stabilize an active conformation of the receptor (positive allosteric modulators (PAMS)) or to stabilize a resting conformation of the receptor (negative

allosteric modulators (NAMS)). [39] Both PAMS and NAMS have no direct effect on the receptors, instead they increase the affinity of the receptor to orthosteric agonists or antagonists. Several PAMS have been identified for the α_7 and $\alpha_4\beta_2$ nAChRs. These PAMS present diverse structures and functionality, which suggests there must be different binding sites in the receptor. Some binding sites are located in the ECD, in the interface between $\alpha_4 - \alpha_4$ (PAMS NS9283 and Zn^{2+}) and $\beta_2 - \beta_2$ subunits (PAM HEPES) and in the TMD, on the α_4 subunit (PAM NS206). In addition, mutations done on α_7 suggest that the PAM PNU-120596, binds in the TMD. [42–44] A new class of activators ago-PAMS, primarily reported for α_7 (compound GAT107), are compounds capable of potentiating orthosteric activation as well as activating the receptor without an orthosteric agonist. These compounds are believed to bind in a TMD allosteric binding site, as well as in a site in the ECD that connects to the vestibule of the ion channel and produces a direct allosteric activation. [45–47] Allosteric binding sites and modulators offer some significant advantages over the classical orthosteric sites. The first of them being the possibility to find selective effectors, since allosteric binding sites can show a higher degree of sequence diversity, as compared to orthosteric binding sites which have evolved to bind the same neurotransmitter. In addition, allosteric modulation can be achieved both with small molecules and antibodies specific to an allosteric conformation, which increases the possibilities of therapeutic approaches that can be pursued. Another advantage that will benefit the development of drugs to treat addiction, is that there is no competition between allosteric effectors and endogenous neurotransmitters, therefore they should not have reinforcing effects that could lead to either dependence or abuse. [39, 48, 49]

Several drugs targeting nAChRs have been approved, such as mecamylamine and varenicline and succinylcholine. Mecamylamine acts as a non-selective nAChRs antagonist and is used to treat severe hypertension, it binds to nAChRs in the peripheral and central nervous system. [50] Varenicline is the only selective $\alpha_4\beta_2$ partial agonist, approved as therapeutic treatment to treat nicotine addiction and smoking cessation. [51] Succinylcholine binds to post-synaptic motor nAChRs and it used to be used for general anesthesia. However, succinylcholine present adverse effects such as cardiac arrest, which have forced researchers to find replacements. [52]

1.3 nAChR $\alpha_4\alpha_5\beta_2$ as a therapeutic target to aid smoking cessation

nAChR $\alpha_4\alpha_5\beta_2$ is the main focus of this work, years of research efforts have focused on this PLGIC to find novel and selective PAMS, to be used to aid smoking cessation and treat nicotine dependence. The interest in this receptor emerged from genome-wide association studies (GWAS) which have highlighted a link between a SNP to nicotine dependence, Chronic Obstructive Pulmonary Disease (COPD), and lung cancer susceptibility, independently of smoking behavior. This SNP results in an exchange of the amino acid aspartate (D) to asparagine (N) at position 298 on the α_5 submit, an amino acid highly conserved in several species. [53–55] α_5 is an accessory subunit that forms part of functional nAChRs $\alpha_3\beta_4$ and $\alpha_4\beta_2$. $\alpha_4\beta_2\alpha_5$ receptors have the highest affinity to nicotine and are expressed in the midbrain dopamine system, implicated in stress, mood disorders, learning and drug

reward. Therefore, $\alpha_4\beta_2\alpha_5$ receptors have been identified as the main contributors to nicotine's addictive effects. [56, 57]

Sophisticated experiments on HEK cells and *Xenopus oocytes* have been carried on, to determine how (or if) the α_5SNP was affecting the intrinsic functional properties of the receptor. [53] The results suggested the α_5SNP was producing a partial loss of function and therefore, less activation of $\alpha_4\alpha_5\beta_2$ receptors by endogenous acetylcholine. [53, 54, 56] Structural studies on $\alpha_4\alpha_5\beta_2$ receptors, did not observe this loss function and suggest the α_5SNP is rather affecting the biosynthesis or trafficking and export of the receptor to the cell surface, since the SNP is located in the ICD of the receptor.[58] Studies on neurons isolated directly from mice nervous tissue, measured the increase of neuronal intracellular free Ca^{2+} concentration, that should be modulated by receptors with the α_5 subunit. Their results also showed a loss of function for receptors with the α_5SNP which had a decreased ability to translate acetylcholine chemical signals, to Ca^{2+} entry to the cells and subsequent depolarization. [59] In vivo analyses of the role of α_5SNP on mice, showed that mice with the α_5SNP self-administered higher doses of nicotine, compared to wild type mice. [57, 60] A hypothesis derived from these observations is that if the α_5SNP produces a loss of function that translates into less activation and increased nicotine consumption, it would be desirable to find specific PAMS for the receptors with the α_5 subunit, to be used as a therapeutic treatment to aid smoking cessation.

Smoking is a risk factor associated with the development of chronic respiratory and cardiovascular diseases as well as cancer and diabetes. [61] In Europe alone, despite observed declining numbers of smoking rates, tobacco use remains a major preventable cause of cancer [62] and over six million people will die every year due to tobacco use. Statistical predictions have shown that the improvements and implementation of tobacco consumption control policies, including offering people help to quit, have the potential to reduce future lung cancer incidence in Europe. [63] Nowadays, it is of great relevance to find more effective treatments, with less side effects, to reduce nicotine consumption in the form of smoking.

1.4 In silico methods for drug discovery

In silico methods have been conceived with the aim of designing novel and safe drugs, to reposition marketed drugs and to narrow down the number of compounds to synthesize or assay on in vivo or in vitro systems. [64, 65] The first step in a drug design project, is to identify a target known to be involved in a disease, for which there is no ideal medical treatment. Once a target has been identified, we must strive to find a number of small molecules that produce the desired effect and that can be optimized in subsequent and iterative steps. In an ideal drug discovery process, the final selected molecules that could turn into drugs, should have an optimal profile of good affinity, selectivity, solubility, permeability and non-toxicity. [66] The in silico approaches to rational drug design, can be divided in two categories: ligand-based design and structure-based design. For the methods classified as ligand-based design, the structure of the target is not known and can't be determined but there are compounds known to bind to the target. New ligands are designed or found based on the similarity property principle, coined by Gerald Maggiora and Mark Johnson,

which states that structurally similar compounds should have similar properties (or biological activity). [67] As its name implies, for structure-based methods, the structure of the target is known or can be determined by comparative modeling and selected ligands should be complementary in terms of steric fit as well as hydrophobic and electrostatic properties to the binding site.

1.4.1 Comparative modeling

If the experimental structure of the chosen therapeutic target is not known, its 3D structure can be predicted using comparative modeling. To do so, we need at least one other protein with known structure (template), that shares a significant sequence similarity with our target. [68] When we do comparative modeling, we assume that the 3D structure is determined by the primary sequence of amino acids and that evolutionary related sequences will have the same structure. [69]

The first step for comparative modeling is to select the template(s) or reference structure(s). The chosen template(s) should have the highest possible sequence similarity or identity to our target. To find template structures the sequence of the target (in our case the sequence of each subunit) can be downloaded in FASTA format from UniProt [70] and the BLASTp [71] algorithm can be used to fetch from the RCSB Protein Data Bank (PDB) [72] those structures with the highest sequence similarity. Once the template(s) have been chosen the next step is to do sequence alignment between the template(s) and the target. This alignment will be used to determine which parts of the template(s) will be used for the model, therefore, this step is determinant for the model's quality. Structurally conserved regions must be well aligned and structurally variable regions must be identified. [68] The alignment can be done with tools like Clustal Omega [73], MUSCLE [74] or T-coffee. [75] Afterwards, there are many different methods available to determine the 3D structure of the target, based on the sequence alignment with the templates, [69] here I introduce MODELLER. MODELLER measures torsional angles and atom distances on the template(s) and uses them as spatial restraints over the target protein. These restraints are derived from the sequence alignment and represented as probability density functions which are combined into an objective function that will be optimized using molecular dynamics (Figure 1.2). [76, 77] Finally, the produced models can be evaluated and refined iteratively until a satisfactory model is obtained [68]. Several scores exist to evaluate comparative models: PROCHECK [78], zDOPE[79] and Molprobity [80]. QMEANBrane[81] was also included, as it is a knowledge-based score that checks local quality of membrane protein models.

The quality of the models will be mostly limited by the choice of templates and the structural alignment. Templates with low resolution are not suitable for comparative modeling since the placement of some side chains could be uncertain. The degree of homology or similarity is another important factor. Domains with sequence identity values lower than 25% will be difficult to model.[69, 82] However, for two pairs of proteins with a sequence identity higher than 50%, 90% of their backbone will have a similar 3D structure. [69, 82] The sequence alignment is used to generate the backbone of the protein, therefore, errors in the sequence alignment will be handed-down to the model backbone and side chains will not be positioned correctly. [83]

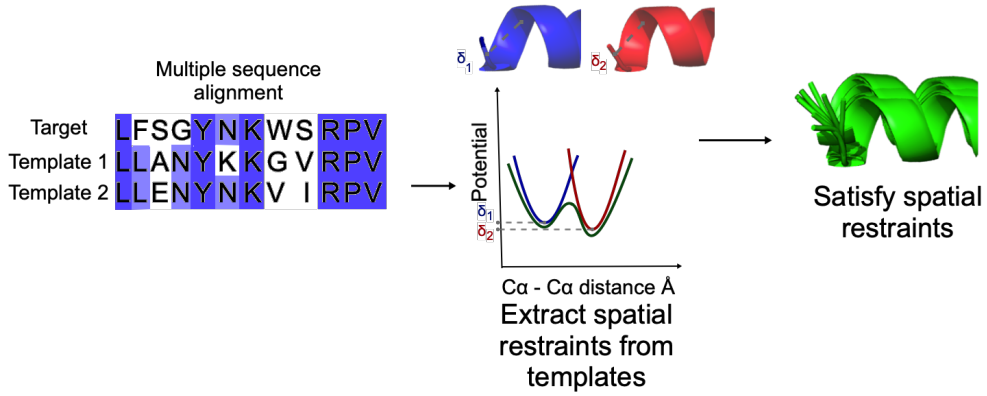

Figure 1.2: **Comparative modeling with MODELLER.** The first step for comparative modeling is the multiple sequence alignment between the target and the template sequences. Then MODELLER extracts spatial restraints from the templates and applies them on the target structure. Those models that satisfy most of the spatial restraints are selected.

1.4.2 Molecular dynamics and transition path calculation

Molecular Dynamics methods predict the movements of the atomic coordinates of proteins and other molecular systems as a function of time. These methods allow us to investigate biological processes that would otherwise be challenging to study experimentally, including ligand binding, conformational changes or protein folding. [84] The main concept behind molecular dynamics implies that given a biomolecular system with its atoms positions, we can calculate the forces applied on each atom by all other atoms to repeatedly update their velocities and positions. In the end, we obtain a movie in which each frame depicts in 3D the atomic conformation of the biomolecular system at every point of the simulated time interval. [85] A molecular dynamics algorithm has three different components: 1) The molecular topology files containing the description of the bonds, bond angles and other simulation parameters. 2) The molecular mechanics force fields, which integrate results of quantum mechanical calculations and other experimental measurements to estimate the forces between atoms or more precisely the estimated potential energy as a function of atom coordinates. 3) Integrators to integrate Newton's equations of motion. Integrators will allow us to determine the position of an atom or object at any given time (t), given its previous time (t-1) to calculate the next time point position (t + 1) and velocities efficiently. 4) If the total system energy were constant, the kinetic energy would increase as the potential energy of the structure decreases, which would make the system overheat. To avoid this we have a thermostat, which will be added to the integrator to control the temperature. Something similar is observed for the pressure of the system, that will be adjusted by a barostat by scaling the box size. [86, 87] The potential force field functions are divided in two terms: Bonded and non bonded interactions (Equation 1.1). Bonded interactions include covalent bond stretching, torsion potential or rotation around bonds, angle bending, and improper torsion potentials as well as correction terms like Urey-Bradley [88] and CMAP [89] for CHARMM36m [90]. Non bonded interactions between atoms include a

term for the Lennard-Jones potential and Coulomb electrostatic potential. [86]

Bonded terms:

$$\begin{aligned} U(r) = & \sum_{bonds} K_r(r - r_0)^2 + \sum_{angles} K_\theta(\theta - \theta_0)^2 \\ & + \sum_{dihedrals} K_\Phi(1 + cos(n\Phi - \delta)) \\ & + \sum_{\substack{improper \\ dihedrals}} K_\phi(\phi - \phi_0)^2 + U_{corr}(r) \end{aligned} \tag{1.1}$$

Nonbonded terms:

$$+ \sum_{nonbonded} \frac{q_i q_j}{4\pi D r_{ij}} + \epsilon_{ij} \left[\left(\frac{R_{min,ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{R_{min,ij}}{r_{ij}} \right)^{6} \right]$$

In the bonded term, K's are the force constants, r_0, θ_0, Φ_0 and ϕ_0 are the bond length, torsion angle and improper angle values. n and δ are the dihedral multiplicity and phase. $U_{corr}(r)$ is a correction term. The first part of the non bonded terms describe electrostatic interactions where q_i and q_j are the partial atom charges of atoms i and j. The second term describes van der Waals interactions, where ϵ_{ij} is the well depth, $R_{min,ij}$ is the radius in the Pauli repulsion and Lennard Jones potential term and r_{ij} is the distance between atoms i and j. [91]

The energy landscape is a 3N-dimensional function that describes the free energy of the system in terms of the coordinates of its N atoms. [92] In this landscape each point has a particular conformation that the protein adopts and the conformational changes are described by the displacement from conformation A to conformation B, where A and B are low energy favorable conformations, depicted as basins, and transition states are intermediate states with higher energy, shown as saddle points. The free energy (G) estimates the stability of a system and governs the direction of its structural changes. [93] The free energy variations ΔG are determined by an enthalpic ΔH and entropic term ΔS proportional to the temperature of the system (T), as shown on Equation 1.2:

$$\Delta G = \Delta H - T\Delta S \tag{1.2}$$

When we assume an adiabatic system (T=0), then the free energy of the system is only determined by the enthalpy $\Delta H = \Delta U + P\Delta V$. In addition, if volume changes are negligible, the enthalpy corresponds to the potential energy and can be described by the force fields of molecular dynamics. However, current molecular dynamics are conventionally limited to time scales of microseconds. This time scale is too short for processes like conformational changes and biological reactions. To circumvent this time scale limitation, we can use algorithms for Transition Path Sampling, which will do an efficient search within the highly dimensional space of all possible transition paths between two stable states. [94]

One of these methods is Conjugate Peak Refinement (CPR) [95], which finds minimum energy saddle points on the potential energy landscape of an adiabatic process. The potential energy surface depicts the low energy reactant and product of a conformational change as basins and the path connecting these two basins going through the saddle points. This path is known as the minimum energy path and is a close average of the most probable conformational paths. The CPR algorithm represents a reaction path as a chain of discrete points termed chain-of-states (COS), between the reactant and the product state (end states). The number of states is not fixed and the path is constructed and modified step by step. The CPR algorithm uses a force field to measure the energy of the intermediate states placed on a piecewise linear interpolation between end states. The point with the highest energy is determined and will be minimized in a direction conjugated to the interpolated segment, to approach a lower energy valley. The structure obtained is a new point in the transition path and the same procedure is repeated for all other segments.[92]

1.4.3 Docking

Molecular docking predicts the affinity, orientation and conformation (or binding pose) of a ligand to a target by simulating the molecular interactions between them. [96] To achieve this, the ligand is analyzed based on its state variables: a position given by the translation of the ligand's x,y and z coordinates, its orientation given by its rotation, as well as its flexibility and conformation described by the different torsion angles of each rotational bond. [97] To perform docking we need two things: a searching algorithm to find the binding pose and a scoring function to rank the binding poses. Searching algorithms can either systematically go over the multidimensional search space at predefined intervals or be stochastic and randomly change the state variables until a criteria is met. Another criteria for the searching algorithm is whether they will find a local or nearest minimum energy to the current conformation or a global best minimum energy. Hybrid methods have proven to be more efficient at finding the best binding pose. [97, 98] Scoring functions can also be divided in three categories: force field function, knowledge based and empirical scores. Force field functions measure molecular forces and interactions between ligand and target. It quantifies electrostatic and hydrophobic and Van der Waals interactions. Knowledge based scoring functions use databases of structural data to extract and compare large amounts of data describing the preferred geometries of interacting protein ligand atoms to derive a pseudopotential. Empirical scoring functions use a regression or classification algorithm to correlate experimental affinity data to physically meaningful terms that might be similar to force-field based scoring functions but can also include terms such as hydrophobic and desolvation interactions. [99, 100]

Before using docking the target and ligand should be prepared. In brief, for the target all hydrogen atoms at desired pH should be added, water molecules should be removed (except those important for the protein-ligand binding site) and if needed partial charges should be computed. For the ligand, a 3D low energy conformation with explicit hydrogens will be required, special attention should be put into generating the correct or most likely tautomers, stereoisomers and protonation states of the ligand. [97]

One of the limitations of docking is sampling the correct pose for the ligand. As the

ligand's flexibility increases it becomes more and more complex and time consuming to sample its conformational space. Rigid docking can be more efficient, however, if we start with the wrong ligand conformation our possibilities to find a good ligand-protein binding pose are lower. The second limitation is the choice of the scoring function to accurately rank correct poses. This issue can be tackled by re-scoring each pose with different scores and selecting those poses that receive high scores or by visual inspection of the docked poses. [100, 101]

1.4.4 Artificial intelligence for drug discovery

Artificial Intelligence (AI) has been used for decades on drug discovery to identify new molecular representations that capture two and 3D chemical characteristics and to derive mathematical functions that explain the relationship between these molecular representations and their biological properties. As the amount of structural information for potential and known therapeutic targets increases and diverse as well as specialized databases continue to appear and expand, the applicability and significance of AI for drug discovery continues to become more relevant. [64, 102]

Machine learning (ML) is a subclass of AI.[103] ML methods can be divided into two groups: supervised and unsupervised learning. Unsupervised learning is used for dimensionality reduction, clustering or to find patterns in unlabeled data.[104] Supervised methods usually require large labeled data sets to train a model to learn the relationship between known input and output data. These methods are used to classify or predict continuous variables. Supervised methods like support vector machine (SVM), random forest and XGboost have been used by computational chemists to predict absorption, distribution, metabolism, excretion and toxicity (ADME-Tox), [105–108] molecular properties [109], virtual screening [110], among other applications [109, 111].

Neural networks are computational architectures designed to mimic the connections between neurons in brains. They are composed of multi-layer structures of interconnected nodes which have associated weights and a firing threshold that determines how high the output of the node should be for the node to become activated and send information to the next layer of nodes. The weights are optimized during training to fit the training data and reduce the error loss. [112, 113] When we talk about deep learning we refer to neural networks with many layers, each of them capable of understanding more and more complex levels of abstraction and learn complex functions. [112, 114] Convolutional Neural Networks Convolutional Neural Network (CNNs) were designed to process data in the form of multidimensional arrays and identify increasingly complex patterns while keeping spatial information. A convolutional layer has three components: the input array, a filter or kernel and a feature map. During convolutions, the kernel moves across the receptive field of the input array, to check if a feature is present. [114] CNNs are part of some of the most used Deep Learning models still being used: Generative Adversarial Network (GANs) and Variational Autoencoders (VAEs).

GANs consist of two networks competing against each other: a generator (G) network that is trained to generate images which should look like the training data and fool

a discriminator (D) and the discriminator which is trained to distinguish between generated and real samples. [115] VAEs have two components, an encoder (E) network that will find a distribution over the latent space that will condense and describe the most important features in the training data and a D that will reconstruct the original data from a vector sampled from the latent space distribution provided by the encoder. A VAE is trained to minimize the reconstruction error between the encoded and decoded data (Figure 1.3). [116]

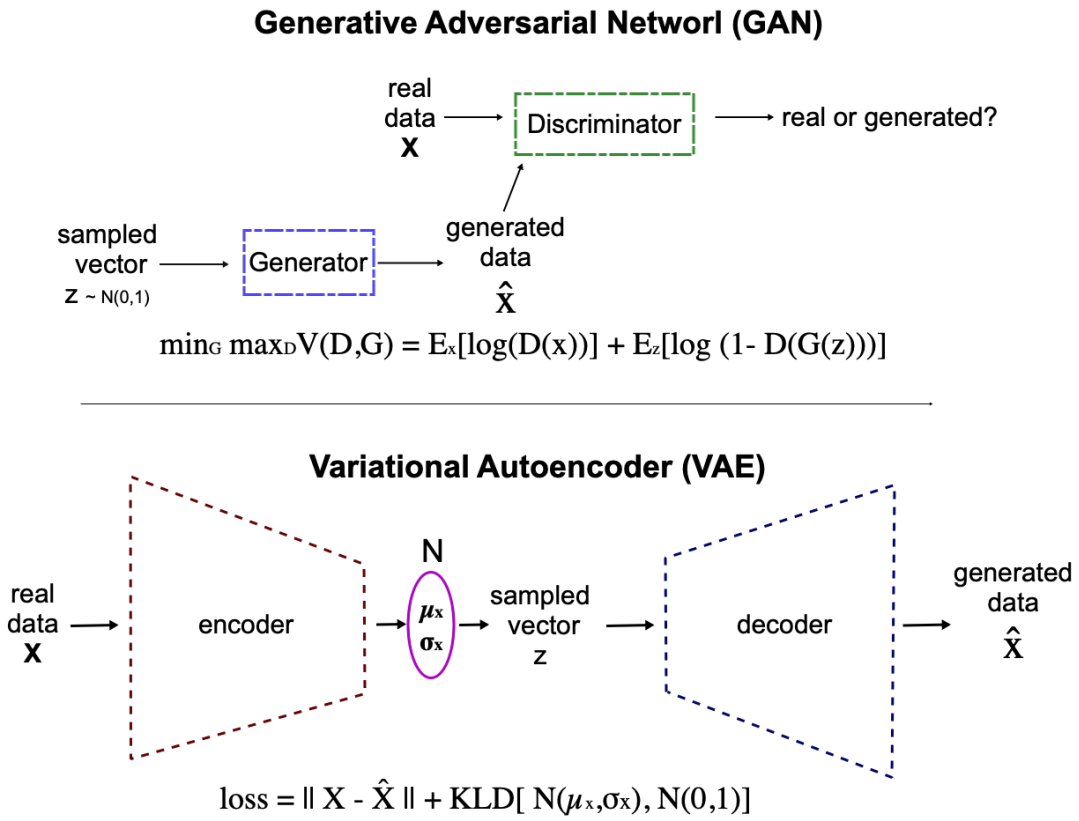

Figure 1.3: **Schema of Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs).** The upper schema depicts the generator and discriminator networks competing against each other, so the generator learns to produce realistic data. In the original implementation, the discriminator tries to maximize the binary cross entropy and the generator tries to minimize it. The lower diagram shows the encoder and decoder networks on VAEs, the encoder encodes the data into a distribution over the latent space and the decoder decodes a sample from this distribution. In the original implementation, the loss function measures how well the original data was reconstructed and the KLDivergence ensures the encoded distribution stays close to a standard normal distribution with $\mu = 0$, $\sigma = 1$.

Long Short-Term Memory cells (LSTMs) networks are the successors of Recurrent Neural Networks (RNNs), developed to handle and propagate long-term information and to

understand sequential patterns. They contain cells with closed loop connections of feedback that allow information to persist and be used from one step to the next as a kind of memory that gets communicated from initial steps until the last step. These cells have different gates: a forget gate, an input gate and an output gate. The forget gate determines which information should be preserved by using a sigmoid activation function that will assign values between 0 and 1, 0 to forget data. The input gate combines the current state X_t with the previous hidden state h_{t-1} and decides what information to keep and scales the data between -1 and 1 with a tanh activation function (Figure 1.4). The output gate determines the values of the next hidden state that will contain the information from all previous inputs. In addition, there is also a cell state that stores the information from the forget gate (if it should be kept) and the input state, giving the network a new cell state.[117]

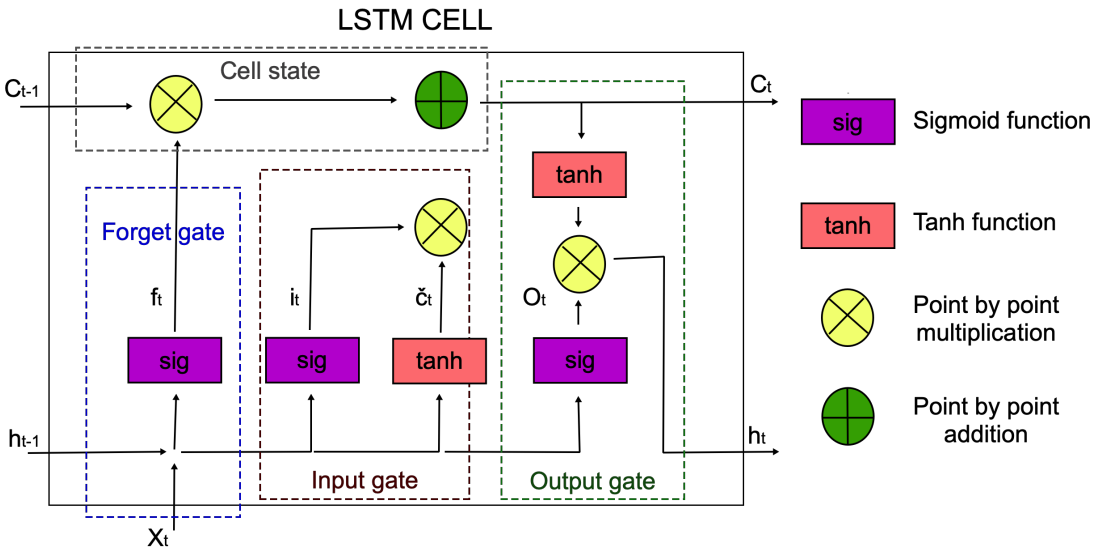

Figure 1.4: **Long Short-Term Memory (LSTM) cell.** LSTMs have three different gates: an input gate, a forget gate and an output gate. Sigmoid activation functions are there to determine which information should be propagated or preserved and tanh activation functions scale the data values. On the schema, t is timestep, X_t is the current input, h_{t-1} the previous hidden state, f_t forget gate, i_t input gate, C_t is the cell state and O_t is the output gate, all at time t.

De novo drug design is being researched with the goal of exploring new regions of the chemical space, containing molecules with desirable drug properties. Exploring new areas of the chemical space will be advantageous, since it will allow chemists to have access to molecules not protected by intellectual property and to exploit their structure activity relationships to uncover new mechanisms of target modulation or even find a novel class of compounds that are more potent and selective. In addition, these molecules can be patented and generate profits to help continue the research of new therapeutics. [118] Over the last years, ligand-based deep learning methods have been implemented to perform molecular generation and optimization. [119–123] Many of these methods train CNNs or RNNs with two and 3D, molecular representations. 2D representations include string based representations such as SMILES, InChi and molecular fingerprints, whilst 3D representations are

mainly grids of voxels that specify atom densities contained in each voxel. [124, 125]

Given that the activity of a ligand is determined by its 3D interactions with a protein, including the 3D structure of both the target's binding site and the ligand should improve the quality of generated compounds. Applications for structure-based *de novo* drug design using deep learning had been limited to the work of Miha Skalic and Gianni de Fabritiis. [126] In this work, they implement a GAN to produce ligand shapes complementary to a protein pocket, or shape generation network, and these ligands shapes are decoded into grammatically correct SMILES by a shape-captioning network. The architecture of the shape generation network is a bicycleGAN shown in Figure 1.5, a successor of GANs conceived to tackle the issue of multimodal image-to-image translation. [127] BicycleGANs were implemented to avoid mode collapse and produce a distribution of different outputs. Its objective is to encourage a bijection between the output and latent space. To achieve this, two tasks are jointly learned during training, the first task is to map the latent code and input to the output and the second task is to learn to go from the output back to the latent space. This way each latent space sample should correspond to only one output. When training the shape generation network as a Conditional Variational Autoencoder - GAN (cVAE-GAN), the encoder provides the latent space from the real data and the generator has the benefit of seeing ground truth input-output pairs. This might lead to a failure to generate new data when sampling random latent noise at test time. In addition the discriminator will not get to see generated results from sampled noise during training. If the generator is also trained as a Conditional Latent Regressor GAN (cLR-GAN) the latent space is sampled from a distribution, circumventing the previous issues and encouraging the bijection between the output and latent space. The captioning network has two components, a VAE and an Encoder-Decoder system. The VAE is only used during training to add noise to the voxelized ligands so that they can resemble the generated ligand shape outputs from the BicycleGAN. The Encoder is a convolutional neural network that will produce a vector representation of the ligand shape for the Decoder, which is a LSTM network and will translate the feature vector into a sequence of SMILES. [126, 127]

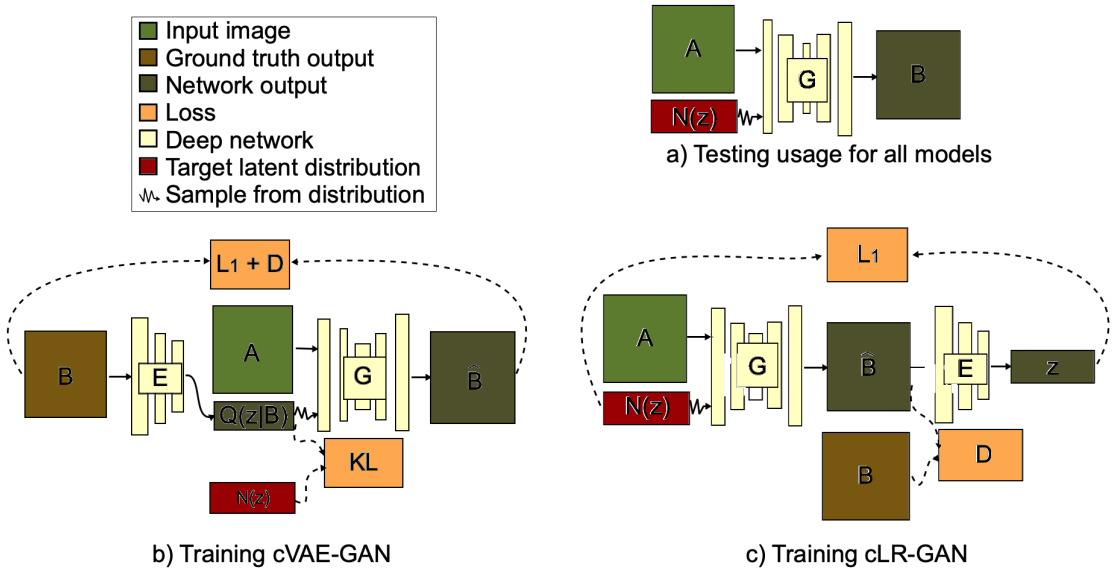

Figure 1.5: **Schema of the original training of a BicycleGAN.** a) Testing of the trained generator (G). To produce a distribution of outputs, the latent code z will be randomly sampled from a standard normal distribution. Then the G will map the input image (A) and z to produce the output $\hat{B}$. b) On the Conditional Variational Autoencoder GAN (cVAE-GAN) the encoder (E) gets as input the ground truth target image B and ecodes it into the latent space Q(x|B). The G will try to map the input image A along with sampled z back into the original image B, and generate $\hat{B}$. The first components to the loss function are the adversarial loss from the discriminator(D), the reconstruction L1 loss between B and $\hat{B}$ and the KLD between a standard normal distribution and the distribution output from the E. c) On the Conditional Latent Regressor GAN (cLR-GAN) a random latent vector N(z) is sampled from a standard normal distribution and is used to map A into the output $\hat{B}$, then the E tries to reconstruct N(z) from the output. The next components to the loss function are: the adversarial loss from the discriminator and the L1 reconstruction loss between sampled N(z) and z reconstructed from output.

1.5 Project objectives

The main goal of this project was to study the structure of the $\alpha_4\alpha_5\beta_2$ nAChR to propose alternatives to modulate it. Since the experimental structure of $\alpha_4\alpha_5\beta_2$ nAChR has not been resolved in any functional state, the first objective of this work was to model the receptor in three functional states (activated, resting and desensitized), using comparative modeling with homologous nAChRs as templates. The models were refined and relaxed in a membrane before moving on to the second objective of this research project which was obtaining the conformational transition path between the modeled states. This transition path was then used to validate the models and to compute cavities that appeared throughout the trajectory and in the three domains of the receptor. The final objective of this project was to train and validate a *de novo* generative model to explore new areas of the chemical space and to propose novel interesting compounds that could be synthesized and bind to

the cavities in the $\alpha_4\alpha_5\beta_2$ nAChR. All the objectives of this project are summarized on Figure 1.6.

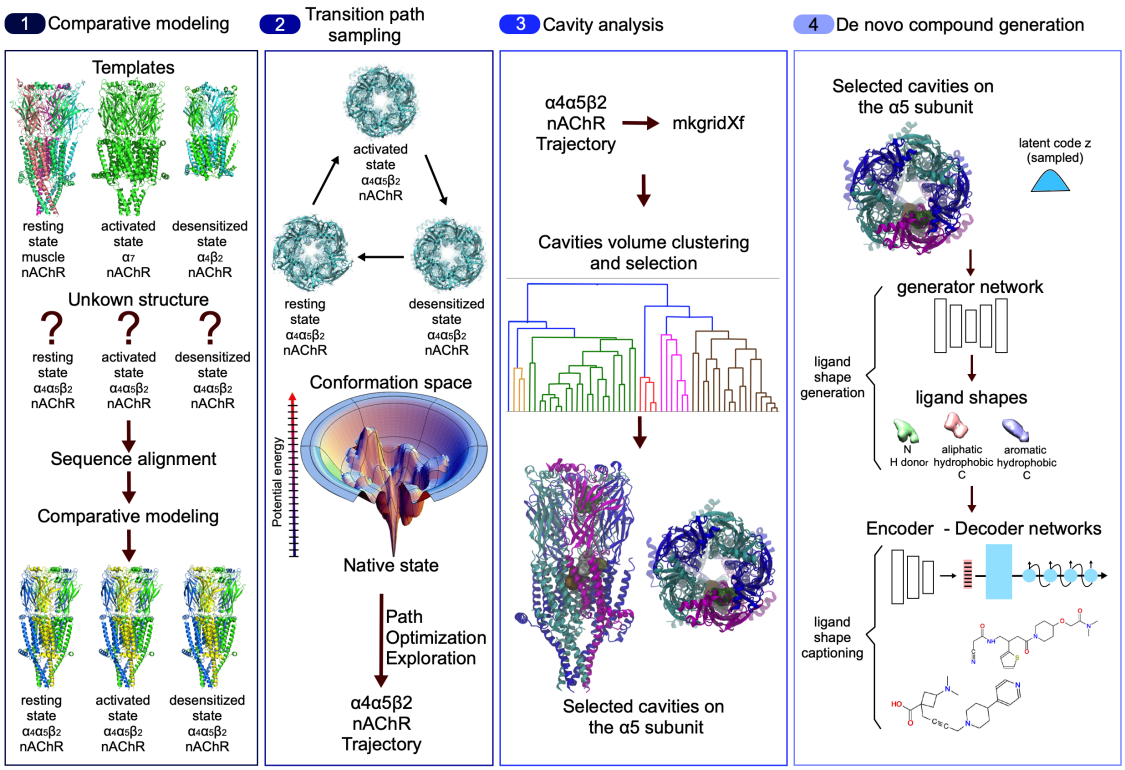

Figure 1.6: **The four objectives of the project are summarized on this figure.** 1) The models of $\alpha_4\alpha_5\beta_2$ nAChR in resting, activated and desensitized states were obtained by comparative modeling. 2) The trajectory between the conformational states was calculated using Path Optimization Exploration (POE). 3) The trajectory was processed by mkgridXf to compute the cavities formed on the receptor. After clustering analysis, the cavities formed on α_5 in the activated conformation were selected. 4) A generative model was trained to generate compounds complementary to the selected cavities on the $\alpha_4\alpha_5\beta_2$ nAChR.

Chapter 2

Comparative Modeling of the functional states of $\alpha_4\alpha_5\beta_2$

The $\alpha_4\alpha_5\beta_2$ nAChR is a pseudo-symmetrical, membrane protein with two α_4 subunits and each of them has 627 amino acids, two β_2 subunits each with 502 amino acids, and one α_5 subunit with 468 amino acids. Overall, the pentamer has a molecular weight of 180 kilodalton. Although several experimental studies have shown the relevance of the $\alpha_4\alpha_5\beta_2$ nAChR as a promising therapeutic target, the experimental structure of this receptor is complicated to obtain and has not been determined in any functional state. Therefore, to be able to study this receptor, I began this project using comparative modeling to obtain models of $\alpha_4\alpha_5\beta_2$ in three functional states: activated state with open ion pore, resting state with closed ion pore and desensitized state with partially blocked ion pore. I started by collecting the information of experimentally resolved structures with good resolution and with an amino acid sequence homologous to the subunits of $\alpha_4\alpha_5\beta_2$. The first challenge was that these experimental homologous structures had to be in the functional states that I wished to model. However, at the beginning of this project, the only experimental structures of cation selective PLGICs available in open and resting state, were those of the serotonin receptor (5-

$HT_{3A}R$). In this chapter I describe the challenges I faced while modeling the $\alpha_4\alpha_5\beta_2$ nAChR and how I obtained the models only after the structures of α_7 were published and used as templates.

2.1 Introduction

The reason behind why comparative modeling can be used to model proteins is that proteins with similar amino acid sequence will share a similar 3D structure. This remains true as long as the protein we wish to model does not go through large conformational changes. If we think of the structural changes occurring on all PLGICs, as the receptor becomes activated, it is clear that the exact same amino acid sequence can have very different 3D structures. Therefore, to be able to use comparative modeling and obtain the $\alpha_4\alpha_5\beta_2$ structure on the desired conformations, I needed homologous experimental structures as templates that should also be in the functional state I wished to model.

At the beginning of the project, while looking for homologous structures with high sequence similarity to $\alpha_4\alpha_5\beta_2$, the cation selective serotonin receptor (5-HT$_{3A}$R) structures appeared to be good template candidates. Several structures of the 5-HT$_{3A}$R had been experimentally resolved with the three domains and in several functional states. In the first published X-ray structure of the 5-HT$_{3A}$R, with a resolution of 3.5 Å, the ICD was truncated and the receptor had to be stabilized by single chain antibodies as crystallization chaperones in detergent. The functional state of this receptor was difficult to assign but given the 4.6 Å hydrophobic constriction at the level of L9 (L260) the structure was determined to be in a non-conductive resting state. This structure showed some unexpected behaviors as the global backbone conformation of the protein resembled what had been observed for open state channels, in addition the loop-C that should remain open in resting state, appeared to be slightly closed. [22] The publication of an apo-structure of 5-HT$_{3A}$R, confirmed the unexpected observations on the crystal structure. This structure was a full length 5-HT$_{3A}$R, resolved by cryo-EM with a closed ion pore and an open loop-C. However, it had a much lower resolution of 4.3 Å. [24] That same year two different groups published other cryo-EM structures of 5-HT$_{3A}$R with bound agonists, antagonists and a PAM. Chakrapani et.al. were able to assign one of the structures to an activated conformation with a wide open ion pore, this structure contains the three domains with a resolution of 3.8 Å and serotonin bound to it. [24] Nury et.al. could also assign one of the structures to an activated state, however, this structure has a low resolution 4.1 Å and the ICD had been truncated. A resting state, stabilized by the antagonist tropisetron, with a resolution of 4.5Å, showed a higher similarity to the X-Ray structure (r.m.s.d. of 0.6 Å) than to the apo structure (r.m.s.d. of 1.15 Å). [128] The 5-HT$_{3A}$R stabilized in a resting state by the antagonist granisetron, was resolved by cryo-EM to a resolution of 2.92 Å. What was surprising about this structure was that its conformation is slightly different to what was observed on the structures of other 5-HT$_{3A}$R with the antagonist tropisetron bound and that being granisetron such a potent competitive antagonist, it would be expected the stabilized confirmation would be similar to the apo-conformation. However, it was observed that the loop-C in the ECD is slightly closed and overall the TMD structure appears to be in a conformation between serotonin bound structures and apo structures. [129] It is challenging to assign 5-HT$_{3A}$R structures to a functional state due to the limited resolution, the influence of the detergent, crystal packing, different receptor engineering methods and the diverse agonists and activators used to stabilize the structures.

An heteromeric $(\alpha_4)_2(\beta_2)_3$ nAChR was solved by X-ray diffraction in detergent, with

nicotine bound to the orthosteric site and without the ICD. This structure has a resolution of 3.94 Å and is described to be in desensitized state with a constriction in the transmembrane domain, at the cytosolic end of the pore, formed by glutamate (E) in the -1' position of the TM2 alpha helices, that has a diameter of 3.8 Å. [27, 29]. A few years later, two structures of $(\alpha_3)_2(\beta_4)_3$, were determined by electron microscopy, with resolutions of 3.34 and 3.87 Å [26]. These structures offered information on the ICD, since only the disordered loop connecting the MX and M4 helices was replaced by a linker (BRIL). The structure with the highest resolution was resolved with nicotine bound to the orthosteric site and in a lipid nanodisc. The pore architecture of $\alpha_3\beta_4$ is identical to $\alpha_4\beta_2$, therefore, the structures were determined to be in desensitized conformation with the ion pore constricted by glutamate (E) -1' of the M2 alpha helices and with a diameter of 3.4 Å. The first nAChR in resting conformation and with a resolution of 2.69 Å, was determined by electron microscopy for the native muscle-type, $\alpha\gamma\alpha\delta\beta$ [32]. The structure was reconstructed into nanodiscs and with the polypeptide antagonist a-bungarotoxin, bound to the orthosteric site. In this structure, the hydrophobic leucine (L) 9', located in the middle of the pore, constricts the pore with a diameter of 2.8 Å. The structures of α_7 in resting, activated and desensitized states were resolved by cryo-EM in a lipidic-nanodisc [8] and in detergent. [130] The activated state was stabilized by the agonist epibatidine and a PAM PNU-120596, which allowed to obtain the first nAChRs structure in activated state with a resolution of 2.70 Å. The resting state was resolved with the antagonist α-bungarotoxin, with a resolution of 3.00 Å and the desensitized state with a resolution of 3.60 Å, was obtained in complex with epibatidine. All these structures are an outstanding contribution to our understanding of the gating cycle of nAChRs and its discrepancies with other pLGICs.

In this chapter I explain the method and templates used to model $\alpha_4\alpha_5\beta_2$ nAChR, as well as the differences observed between the experimentally solved structures of nAChRs and 5-HT_{3A}Rs. I present and discuss the ion pore profile used to characterize the functional state of the models and compare some structural features of $\alpha_4\alpha_5\beta_2$ nAChR to other PLGICs.

2.2 Methods

2.2.1 Definitions

Activated state / open-pore: the activated functional state of the receptor can be described as activated or as in open ion pore state.

Resting state /closed-pore: the apo or resting functional state of the receptor can be described as resting or as closed ion pore state.

Desensitized state: This is a holo state of the receptor where the ECD conformation is similar to the activated state but the TMD shows a blocked ion pore. The ion pore is blocked at the lower end of the pore, with a different profile from the resting state.

Template: I use this term to name the selected homologous structures that were used to model $\alpha_4\alpha_5\beta_2$ by comparative modeling.

Target: the target sequences are those of α_4, α_5 or β_2 whose 3D structure is to be obtained by comparative modeling.

Model: Here the structures of $\alpha_4\alpha_5\beta_2$ obtained by comparative modeling in resting, activated and desensitized states are called models.

Restraint: The restraints defined by the algorithm of MODELLER, used for comparative modeling, are the probability density functions describing the spatial arrangement of groups of atoms in the target protein (distance between atoms, angles, dihedral angles). The restraints are calculated from the structural templates and their multiple sequence alignment with the target sequence.

2.2.2 Selection of homologous proteins as templates

To find homologous sequences with experimental structures, I performed a blastp search on the NCBI BLAST [71] service, comparing the sequence of α_5 against the PDB. [72] I filtered out those templates with a query cover lower than 70% and I gave priority to X-ray structures or cryo-EM structures with the best resolution, structures that were clearly assigned to a specific functional state (resting, activated, desensitized) and structures with the three domains (ECD, TMD, ICD).

Before the structures of α_7 were published, I attempted to model $\alpha_4\alpha_5\beta_2$ with the structures of mice 5-HT$_{3A}$R as templates for the activated and resting states. To simplify the description of the structures, I will name them by their 4 letter identifier from the PDB. To model the activated state I chose 5-HT$_{3A}$R 6DG8 [23] and 6HIN,[128] both activated by the agonist serotonin and assigned to an open-pore state. The resting state was modeled with mice 5-HT$_{3A}$R 6NP0 [129] inhibited by the competitive antagonist granisetron, electric ray muscle-type nAChR 6UWZ inhibited by α-bungarotoxin [32] and an unpublished acetylcholine binding protein AChBP$_{\alpha_5-\alpha_4}$ chimera, where the interface between subunits was mutated to the sequence of the interface between $\alpha_5\alpha_4$, determined to be an allosteric binding site by in-house data. Both the data and the chimera were provided by the Institut Pasteur Receptor Channels Unit (Pierre-Jean Corringer and experimental data from Akos Nemecz). The desensitized state was modeled with nAChRs $\alpha_4\beta_2$ (5KXI) [27] and $\alpha_3\beta_4$ (6PV7), [26] both assigned to a desensitized state and activated by nicotine.

After a year of unsuccessful model production, the structures of α_7 in resting, activated and desensitized states [8] were published and the templates were changed (Figure 2.1). The activated state was modeled with nAChR α_7 7KOX stabilized in activated state by the agonist epibatidine and the PAM PNU120596. The resting model with the AChBP$_{\alpha_5-\alpha_4}$ chimera, electric ray muscle-type nAChR 6UWZ and α_7 7KOO inhibited by α-bungarotoxin. The templates to model the desensitized state were nAChRs $\alpha_4\beta_2$ 5KXI, $\alpha_3\beta_4$ 6PV7, as well as α_7 7KOQ, with bound epibatidine.

PDB ID	7KOX	7KOO	6UWZ		7KOQ	5KXI	6PV7
Technique	Cryo-EM	Cryo-EM	Cryo-EM	X-ray	Cryo-EM	X-ray	Cryo-EM
Resolution	2.7	3.0	2.69		3.6	3.9	3.3
Ligand	Epibatidine and PNU120596	α-bgt	α-bgt	MLA	Epibatidine	nicotine	nicotine
Detergent / nanodisc	nanodisc	nanodisc	nanodisc	----------	nanodisc	detergent	nanodisc
composition	α7	α7	αγαδß	AChBP α5-α4	α7	α4ß2	α3ß4
State	open	resting	resting	resting	desensitized	desensitized	desensitized

Figure 2.1: **Templates used to model $\alpha_4\alpha_5\beta_2$ in resting, activated and desensitized states.** The figure presents the templates information regarding the experimental technique, resolution, bound ligand(s), composition and reported functional state of each structure and whether the protein was stabilized in nanodisc or detergent.

2.2.3 Multiple sequence alignment

The sequences of human α_4, α_5 and β_2 were extracted from UniProtKB [131] (ACHA5 gene, entry: P30532; ACHA4 gene, entry: P43681; ACHB2 gene, entry: P17787). Since I wanted to model the three domains of the protein excluding the disordered loop in the ICD, I used JPred4 [132] to predict the parts of the sequence that should have an α helix secondary structure and belong to the MX and MA helices and those that should be part of the disordered loop. The parts of the sequence not covered by available homologous structures were also removed, that is: 33 amino acids in the N-terminal part and 203 amino acids from the disordered loop were removed from the α_4 sequence, 41 amino acids in the N-terminal and 24 amino acids on the disordered loop from the α_5 sequence were also removed and 25 amino acids in the N-terminal side and 72 amino acids from the disordered loop were erased from the β_2 sequence. It should be noted that PLGICs without the disordered loop remain functional and with agonist binding profiles similar to those of wild type receptors. [22] Since the disordered loop in the ICD was removed, all the subunits were split into two chains, the first chain going from the N-terminal ECD to the MX helix and the second chain with MA and TM4 helices. The truncated sequences were concatenated to generate a decamer and are reported in the appendix Figure 6.1.

Since I was modeling a heteromeric protein, special attention had to be put on the order of the sequences and subunits. To preserve the orthosteric binding site, the loop-C in the main subunit (+) must come from α_4 and the complementary subunit (-) must be β_2. At the same time, on the allosteric site the main subunit (+) must be the α_5 subunit and the complementary subunit (-) α_4 (Figure 2.2).

The model and the templates have 5 subunits, 10 chains and the multiple sequence alignment was done per chain and per functional state in three- and two-dimensions. Initially the sequences were aligned using Clustal Omega [73] with default parameters. This

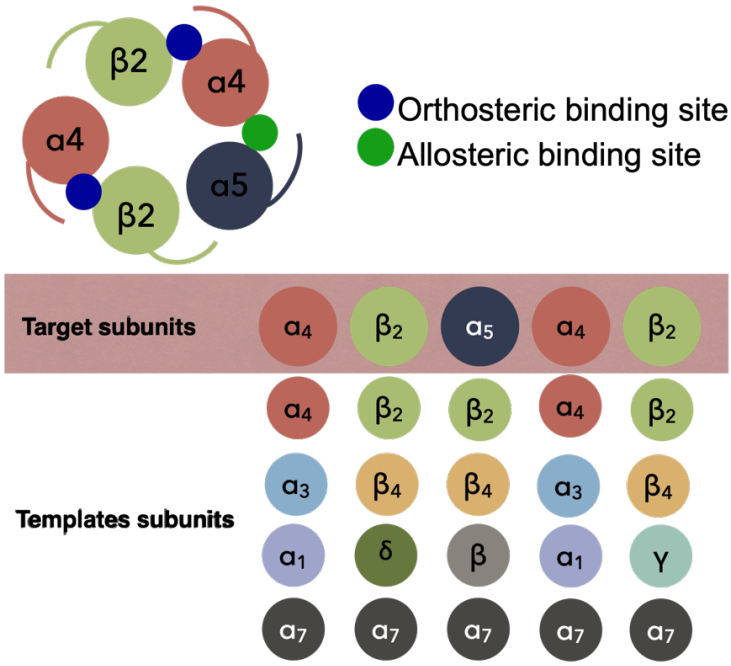

Figure 2.2: **Subunit alignment of templates used to model $\alpha_4\alpha_5\beta_2$.** The known allosteric and orthosteric sites are shown in green and blue, respectively, and capped by the loop-C. The figure indicates which subunits from the templates were aligned with each subunit in $\alpha_4\alpha_5\beta_2$.

sequence alignment was refined using VMD's 3D MultiSeq alignment. [133] For example, the multiple sequence alignment of α_5 in desensitized state was performed only with the sequences and structures of β_2 from 5KXI, α_7 from 7KOQ and β_4 from 6PV7. This 3D alignment allowed us to refine the gaps from the initial Clustal Omega alignment. For each functional state, a pir file with the multiple sequence alignments, in the format required by MODELLER, [76, 77] was created.

2.2.4 Comparative modeling with MODELLER

I used MODELLER for comparative modeling, since it allows to control certain structural characteristics of the receptors and to model the side chains in the orthosteric binding site with the information from the bound ligand in the pdb of the templates. MODELLER takes as input a multiple sequence alignment of the target sequence with all the template sequences in .pir format as well as the structures of the templates (cartesian coordinates as pdb files). To compute the 3D structure of the target, MODELLER derives from the multiple sequence alignment a set of structural restraints to be imposed on the atoms of the target protein. In addition, I included the following restraints: a restraint to keep the proline from the conserved sequence FPF and numbered 136 on α_4, 137 on α_5 and 138 on β_2 in cis isomerization (ω angle 0) and a restraint to maintain sulfur atoms on cysteines at a distance shorter than 2.0 Å to form a disulfide bond. These disulfide bonds are in the ECD of all subunits and are formed by cysteines 128 and 142 on $\alpha4$, 129 and 143 on $\alpha5$ and 130

and 144 on β_2. In addition α subunits have a disulfide bond in the loop-C formed between cysteines 199 and 200 on α_4 and 193 and 194 on α_5. α_4 and β_2 subunits were also restrained to keep them symmetrical (Figure 2.3).

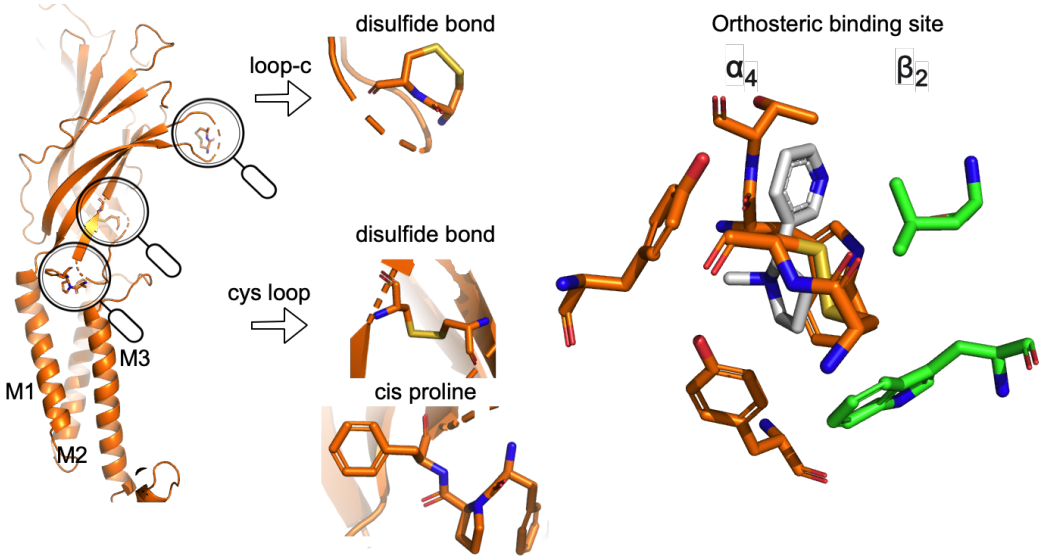

Figure 2.3: **Structural restraints imposed on the structure of $\alpha_4\alpha_5\beta_2$.** On the left side the restraints set to preserve the disulfide bonds on the ECD and the proline on cis conformation. The right side shows the ligands in the orthosteric binding site whose information was used to model the side chains between subunits.

I also used the information of ligands bound on the experimental structures of the templates, to model the side chains in the ECD on the allosteric and orthosteric binding sites. Both the allostric and orthosteric binding sites present the conserved hydrophobic box where ligands bind. For the activated model I used the agonist epibatidine in the orthosteric binding site of the structure of α_7 7KOX. For the closed model I used the ligand in the AChBP chimera and for the desensitized model I took nicotine in $\alpha_4\beta_2$ and epibatidine in α_7.

Afterwards, MODELLER optimizes the placement in space of the atoms in the initial 3D structure of the target, to reduce the violation of structural constraints. The parameters included on this step are "library_schedule = autosched.slow" and "md_level = refine.slow".

2.2.5 Evaluation and selection of models

I used MODELLER's setup to produce 1000 models of each functional state. In the initial process of template selection and sequence alignment validation, the structures were scored

to measure their likelihood to be correct with Modeller's zDOPE score [76]. Once the templates and sequence alignment were selected, I also included QMEAN, [81] Procheck, [78] Verify3D [134] and ProsA scores. [135] zDOPE is the Z-score of a protein's DOPE score, a statistical potential based on atomic distances, calculated from native protein structures to estimate the energy of the model. Negative zDOPE values suggest a lower energy and the modeled protein is more likely to be in a native conformation. [76] Similar to zDOPE, ProSA is a statistical potential based on C_α-C_α distances, that measures the energy of the system. If this value is outside a range, characteristic for native structures with a certain number of residues, the measured protein might have an erroneous structure. [135] Verify3D measures the compatibility of a 3D structure with its sequence. For each amino acid a statistical preferred environment was established (described by its secondary structure, fraction of side chains covered by polar atoms and how buried a residue is) and the protein is scored depending on what percentage of its amino acids agree with the expected environment. Higher percentages indicate models with better quality. [134] Molprobity is a log-weighted combination of the clash score, percentage of Ramachandran outliers and percentage of bad side-chain rotamers, that yields one number which reflects the crystallographic resolution at which those values would be expected. [80] Therefore, for a modeled protein, the lower its Molprobity scores is, the better its structures should be. QMEANBrane ranges from 0 (poor quality model) to 1 (good quality model) and is a linear combination of several molecular descriptors assessing: the local geometry, long-range interactions, burial status of residues, agreement between predicted and calculated secondary structure and solvent accessibility, to assess the quality of the TMD. [81]

In addition the 3D structures of the templates were also scored and compared to the models. The models with the best scores were selected for further optimization.

2.2.6 Relaxation and equilibration of models with GROMACS

Once I selected the models with the best scores for each functional conformation, I added the hydrogens with Maestro and the protonation state of histidines and other residues that could be ionized at physiological pH were selected using reduce. [136] The receptor without the ICD disordered loop has 31511 atoms and 1926 residues. A fast routine was set up in collaboration with Max Bonomi from the structural Bioinformatics Unit at the Institute Pasteur, to minimize and equilibrate the three conformational states of $\alpha_4\alpha_5\beta_2$ using a collection of python [137, 138] and gromacs-2020.4 scripts. The first step of this procedure is to use the Membrane Builder [139–141] from the CHARMM-GUI web server [142] to generate an heterogeneous lipid bilayer and $\alpha_4\alpha_5\beta_2$ protein input files for gromacs. The components of the lipid bilayer include: 1-palmitoyl-2-oleoyl-phosphatidylcholine (POPC), 1-palmitoyl-2-oleoyl-phosphatidic-acid (POPA) and cholesterol (CHL) with a stoichiometry of 3POPC, 1POPA and 1CHL. Previous studies have shown that this composition can stabilize the receptor in an active conformation [143–145]. The upper layer and lower layer had 180 and 175 lipids, respectively. The lipid bilayer was centered at 0 on the z axis and a gap with a radius of 2.0 nm was created at the center of the lipid bilayer to fit the protein. The protein was solvated and two disulfide bonds were set for each α subunit (128-142, 192-193 in α_4 and 129-143, 193-194 in α_5) and a single disulfide bond for the β_2 subunits

(130-144). The protein was inserted into the lipid bilayer using gromacs insert-molecules. The system was solvated using gromacs command solvate and excess waters in the lipid bilayer were removed, afterwards ions were inserted using gromacs command genion. A new gromacs topology file was created considering the final number of lipids and the correct number of waters and ions. The steepest-descent algorithm for energy minimization was used for 50000 steps or until the maximum force was smaller than 1000 KJ/molηm. The system was equilibrated for seven steps at 303.15K, with position restraints on the heavy atoms to prevent the activated receptor from collapsing.

2.2.7 Validation of $\alpha_4\alpha_5\beta_2$ models functional state

During the study and characterization of ion channels, the minimum pore diameter is routinely used to assign an experimental structure to a functional state. In this work, I computed the pore profile and pore radius of the $\alpha_4\alpha_5\beta_2$ models using the program HOLE. The configuration of HOLE that I used was defined by the following commands on the input file: radius, I used simple.rad to specify the van der Waals radii of each atom with the definition given by the AMBER potential energy function; cvect 0 0 1 to specify the channel pore lies along the Z axis and enrad of 14Å, which is the radius above which the end of the pore should be found. [146]

2.3 Results and discussion

The latest blastp results yield 33 homologous sequences with e-values between $2e^{-51}$ and $1e^{-142}$ and identity percentages between 56.13% and 26.44%. The results include human nAChRs $\alpha_4\beta_2$, $\alpha_3\beta_4$, the recently published α_7 structure, electric ray muscle-type nAChR $\alpha_1\gamma\delta\beta_1$ and mice 5-HT$_{3A}$R. The percentage of sequence identity between the α_4, α_5, β_2 subunits and the templates subunits are shown in Table 2.1. α_4, α_5 and β_2 have the lowest sequence identity with 5-HT$_{3A}$R. With values between 27% and 31%, the 5-HT$_{3A}$R could still be used as template but the least conserved regions would be difficult to model. The sequences of α_5 and α_4 are most similar to α_3 and β_2 has a high sequence similarity to β_4. Overall, I was surprised to see that α_7, being a nAChR in the central neural system, like $\alpha_4\beta_2$ and $\alpha_3\beta_4$ had the lowest percentage of sequence identity among α subunits, even lower than α_1 in the electric ray muscle-type nAChR.

Table 2.1: α_4, α_5 **and** β_2 **subunits percentage of sequence identity.** This table shows the percentage identity of the α_4, α_5 and β_2 subunits with the subunits of the templates used to create the models.

Identity %	α_5	α_3	α_4	β_4	α_1	β_2	α_7	β_1	δ	γ	5-HT$_{3A}$
α_5	–	56.13	54.74	49.85	45.56	41.95	38.12	36.48	33.05	30.84	27.04
α_4	54.74	67.66	–	56.8	56.97	58.28	47.6	46.27	42.57	42.73	31.37
β_2	41.95	54.65	58.28	76.99	41.94	–	43.95	42.95	38.6	39.63	27.53

At the beginning of the project, my objective was to model $\alpha_4\alpha_5\beta_2$ with the three domains and in three functional states, so I did not have a lot of choices to make while selecting the templates, as the structures of α_7 had not been published. I chose the structures of the 5-HT$_{3A}$R as templates, because these were the only structures available in resting and activated functional states, among all eukaryotic and cation selective PLGICs, and with the highest homology to $\alpha_4\alpha_5\beta_2$ nAChR. For the activated state, since the 5-HT$_{3A}$R structure 6HIN does not have an ICD, I decided to include the 5-HT$_{3A}$R structure 6DG8 to be able to model it. For the resting state there were several 5-HT$_{3A}$R structures to choose from but 6NP0 was chosen since it had the best resolution (2.92Å). The other 5-HT$_{3A}$R structures with a closed pore had resolutions between 3.5 to 4.5Å. The structure of 6NP0 was described to have a closed-pore profile similar to what is observed in the apo 5-HT$_{3A}$R [22] but the overall structure is believed to be in a conformation between the serotonin activated 5-HT$_{3A}$R and the apo structure. In addition the loop-C expected to be extended in resting structures appears to be slightly closed. These unexpected structural features are likely to come from the interactions established by the agonist granisetron [129] A few months later the structure of muscle-type nAChR was also published and determined to be in a resting, closed pore conformation so I decided to include it in the templates, to get structural information from its α and β subunits. For the desensitized state, I chose only 5KXI among the other $\alpha_4\beta_2$ structures in desensitized state because it is the X-ray structure that served for the building and refinement of subsequent cryo-EM structures (6CNJ and 6CNK).[29] Since none of these structures have an ICD, I also decided to include the structure of $\alpha_3\beta_4$ 6PV7 to model the desensitized state. I chose 6PV7 instead of 6PV8 because the structure of the second could not be solved in lipidic nanodisc, the receptor had to be purified in detergent and the authors suggest there could be conformational differences produced from detergent artifacts. [26]

The first score I used to decide if I was obtaining reasonably good models was MODELLER's zDOPE score. Before α_7 was published, I used 5-HT$_{3A}$ structures as templates for the resting and activated states. The zDOPE scores I obtained from MODELLER for the model in the activated stare were consistently bad (zDOPE = -0.090). In addition I had difficulties modeling the disulfide bonds in the loop-c of α subunits and the cis-proline in the cys-loop. Once the structures of α_7 were published and used as templates, the zDOPE improved significantly for the activated states and I was able to continue the following modeling steps (Figure 2.4).

One hypothesis to explain why the models in activated state had such bad scores, compared to the resting and desensitized models, relies on the quality and disagreement between the experimental templates. The 6DG8 and 6HIN structures have low resolution, 4.10 Å for 6HIN and 3.89 Å for 6DG8. Within the structures, the highest B-factors (a measure of the uncertainty of the position of the atoms) are located in the TMD and ICD. The B-factors for 6HIN range between 35.76 to 154.02 Å^2, with an average of 57.80 Å^2 and for 6DG8 between 102.82 to 659.00 Å^2, with and average of 140.60Å ^{2}. Some studies on X-ray structures have estimated that at resolutions worse than 3.3 Å the average maximum B-factor to consider that the structure was constructed based on experimental evidence rather than an over-interpretation of the atom position, should be 80 Å^2. [147] Which makes me question how reliable the atom positions in the structure 6DG8 actually are, although it most be noted that these structures were obtained by cryo-EM and not by X-ray crystal-

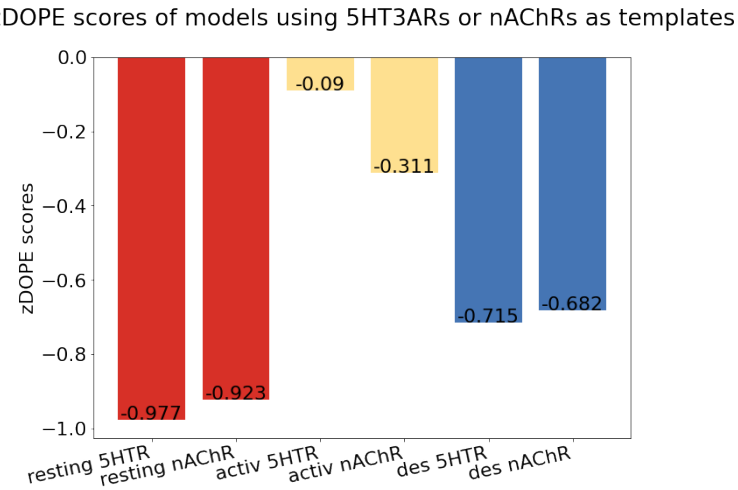

Figure 2.4: **Comparison of the model's zDOPE scores with different templates.** zDOPE scores assigned by MODELLER to the models in the three functional states. The red bars are the scores for the models in resting state, the yellow bars for the activated state and the blue bars the desensitized state. The scores changed when the 5-HT$_{3A}$ template structures were changed for α_7 nAChR structures.

lography. The ECDs of 6HIN and 6DG8 appear to have considerable structural differences, although they have similar pore profiles with an open pore showing a minimum radius of 3.0 and 3.3 Å. The structural alignment of the TMD of these receptors has an RMSD of 3.12 Å and the TM1-TM4 helices seem to be positioned differently. For comparison, the structures of the TMD of $\alpha_4\beta_2$ in desensitized state 5KXI, 6CNJ and 6CNK were aligned and are shown in Figure 2.5. These structures have an average RMSD of 1.11 Å with well aligned TM1-TM4 helices. It is not clear to me whether 6DG8 and 6HIN are two 5-HT$_{3A}$Rs in two different activated conformations or if the limited resolution of the structures does not allow us to define with enough certainty the atom positions in the TMD. The resting and desensitized models benefited from the structural information of other nAChRs and the models had acceptable zDOPE scores.

Another structural feature, present in other experimental structures of nAChRs, that was difficult to model using 5-HT$_{3A}$Rs as templates, was the cis-proline located in the cys-loop of the ECD, that interacts with the TM2-TM3 loop in the interface between the ECD and TMD and the disulfide bond in the loop-C of the ECD. Figure 2.6 There is no agreement on whether or not this proline should be in cis or trans conformation, or its exact role in the gating cycle of the ion chanel. However, NMR studies show that peptides with a phenylalanine followed by a proline are more likely to present a proline as a cis isomer. [148] In addition, the $\alpha_4\beta_2$ crystal structure and all other nAChRs have this proline is in cis conformation. As it was described in the methods, restraints were integrated during the modeling process for both features. However, in the multiple sequence alignment between the target and the templates the phenyl-proline-phenyl (FPF) sequence is conserved but the 5-HT$_{3A}$Rs have a trans-proline isomer. Since a cis conformation was imposed but the templates had this same proline in trans conformation the resulting models had a very deformed semi cis-proline that was difficult to fix during posterior refinement steps. Simi-

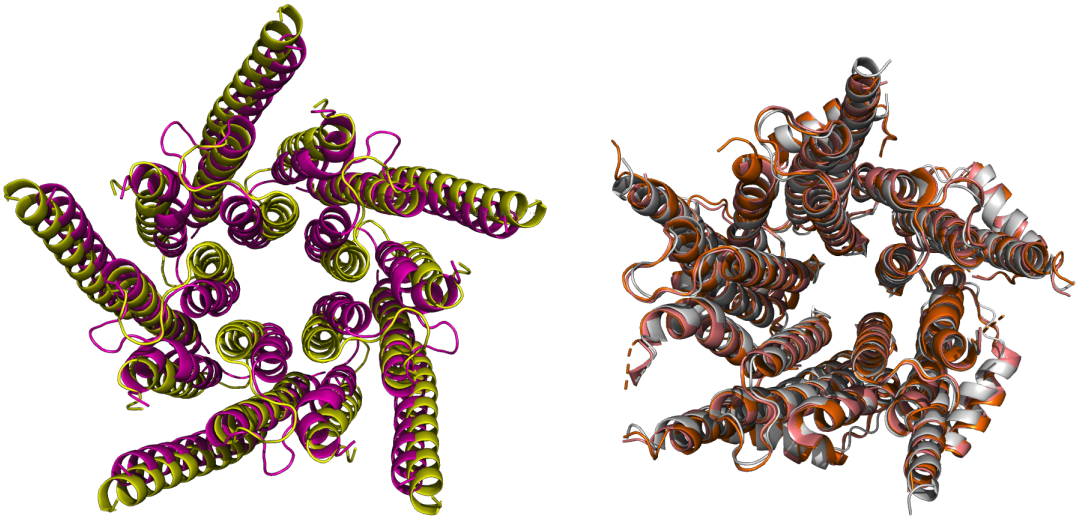

Figure 2.5: **Structural alignment of the activated 5-HT$_{3A}$R templates and desensitized structures of $\alpha_4\beta_2$.** On the left, the alignment of the TMD between the 5-HT$_{3A}$R structures 6HIN and 6DG8 has an RMSD of 3.12 Å. These structures are of the same receptor and have been determined to be in an activated functional state with similar pore radius but show different TMD conformations. For comparison, on the right the structural alignment of the TMD between $\alpha_4\beta_2$ nAChRs has an average RMSD of 1.11 Å, these structures are in desensitized state.

larly, the cysteines in the loop-C of the α subunits are not present in 5-HT$_{3A}$Rs loop-C and I often observed that models had these cysteines positioned at a distance or conformation that did not allow the conserved disulfide bridge to be formed.

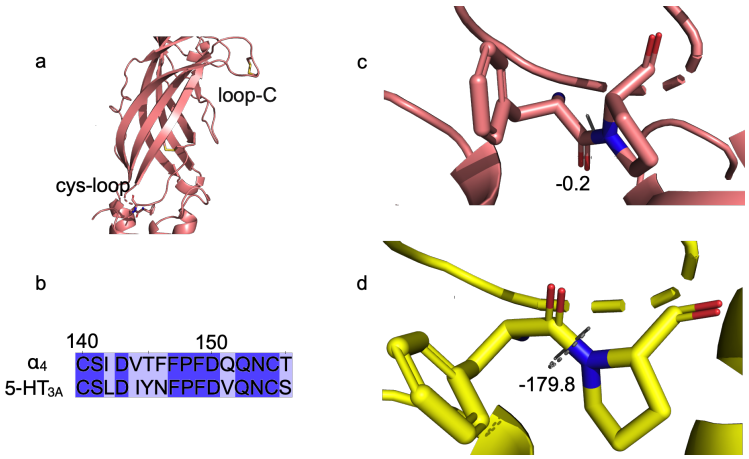

Figure 2.6: **Proline isomerization in 5-HT$_{3A}$ and nAChRs templates.** a) The cis-proline isomer is located in the cys-loop of some PLGICs. Here α_4 has a cis-proline isomer in the cys-loop. A disulfide bond in the loop-c, only present in the α subunits of nAChRs, is also shown. b) Conserved phenyl-proline-phenyl (FPF) amino acids in 5-HT$_{3A}$ and nAChRs. c) Cis-proline in the cys-loop of α_4. d) Trans-proline in the cys-loop of 5-HT$_{3A}$.

Once the α_7 structures were published I decided to use only nAChRs to generate the models, so I removed the 5-HT$_{3A}$ templates and did a new multiple sequence alignment. By modeling $\alpha_4\alpha_5\beta_2$ with nAChRs, the isomerization of the prolines in the cys-loop as well as the disulfide bond in the loop-C could be modeled and the modeled activated receptor improved significantly. These models were minimized and relaxed in the membrane, as described in the methods. To have an idea of what scores I would obtain for the experimental structures used as temples and to compare these values to the minimized models, the templates and models where scored by the zDOPE as well as ProSA, QMEAN, Verify3D and molprobity (Figure 2.7). I was surprised to see 6DG8's positive zDOPE score, if a model got a positive score this would suggest it is not a native-like model. In addition only 37% of the residues in 6DG8 are in a statistically expected environment as described by the Verify3D score. Both zDOPE and ProSA agree that the 5-HT$_{3A}$R structures have the least native-like structures among all the templates. Overall the 5-HT$_{3A}$R structures in activated state, score worse than the nAChR templates with the zDOPE, ProSA, QMEAN and Verify3D scores. The relaxed models zDOPE scores improved significantly from the scores I had before minimization and relaxation in membrane: resting -1.27, activated -1.00, desensitized -1.05. In addition, the models done without 5-HT$_{3A}$R templates had better zDOPE and Verify3D scores in active and resting states than the 5-HT$_{3A}$R structures. Overall, the models have worse scores than the nAChR structures, however for most scores (except molprobity and ProSA) their scores are still within the range of the scores obtained for the experimental structures used as templates.

The modeling of $\alpha_4\alpha_5\beta_2$ in its functional states proved to be a very challenging and time consuming step of this research project. No amount of pages I could write will be

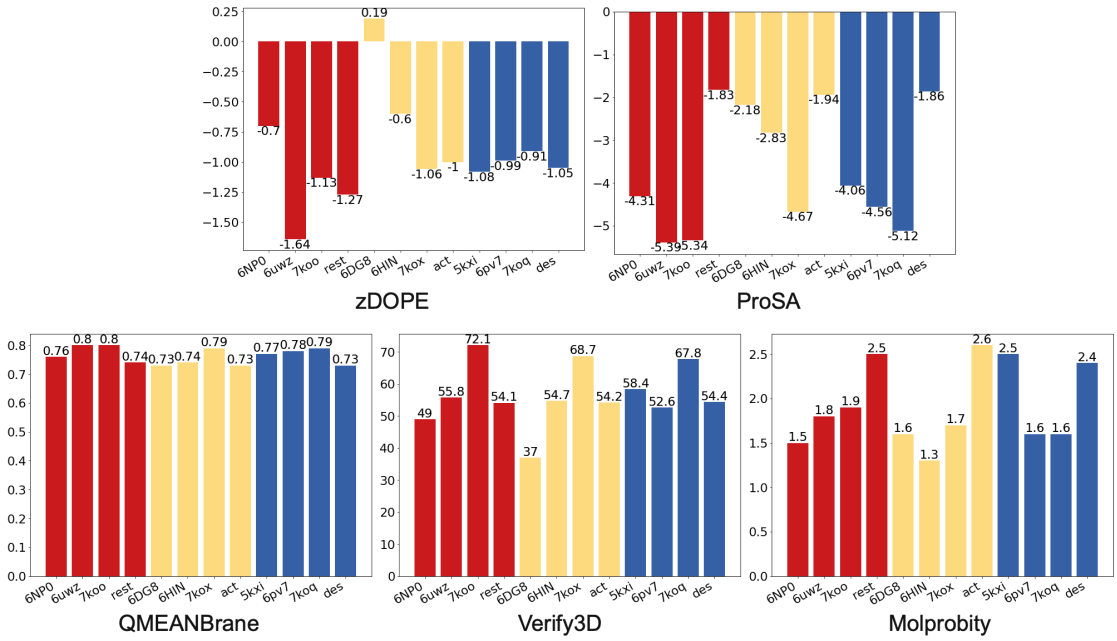

Figure 2.7: **Models evaluation using different score functions.** The models in resting, activated and desensitized states made with α_7 nAChR as template are compared to all the structures used as templates. For zDOPE, good models have the lowest negative scores. QMEAN is in the range of 0-1, good models have values close to 1. Verify3D is a percentage, good models have higher percentages. ProSA is a Z-score and the value shown is the average of the subunits with (400 amino acids), proteins with 400 amino acids have values between -13 and -3. Molprobity the lower the score, better the predicted quality should be.

representative of the actual number of multiple sequence alignments or combination of templates and template subunits that I tested to come to the point where the models were acceptable. Of course, the final choices I made will affect all the following steps of this research project.

One of the first questions raised by the models appeared after the transition path between the functional states was computed. It was observed that the loop-C of the α_5 subunit remains slightly open in the desensitized state, while it would be expected to remain closed, as it is the case for the orthosteric binding site between α_4 and β_2 in the experimental structures. After the cavity analysis, I was able to notice that the volume of the cavity capped by the loop-C of α_5 follows the expected extended loop-C in resting state (larger volume), closed loop-C in activated state (smaller volume) pattern, but in desensitized state this loop-C has a volume similar to the volume in the interface between $\beta_2-\alpha_4$ and $\beta_2-\alpha_5$. When I looked at the volumes in the desensitized states of the heteromeric $\alpha_4\beta_2$ and $\alpha 3\beta_4$ receptors, I noticed that non-orthosteric interfaces show variable cavity volumes and only orthosteric sites have a low volume, closed loop-C profile. This made me realize that while doing the multiple sequence alignment, for the desensitized state α_5 was aligned to the β_4, β_2 and α_7 subunits. It is likely that because α_5 has a higher sequence identity to β_4 and β_2 their structures had a higher weight during the modeling than α_7 structure. In addition the loop-C is shorter in α_5, and therefore, the conserved cysteines in all α subunits could not be modeled from $\alpha 7$ and some gaps had to be inserted which leads to α_5 having a better alignment to the β subunits. Figure 2.8

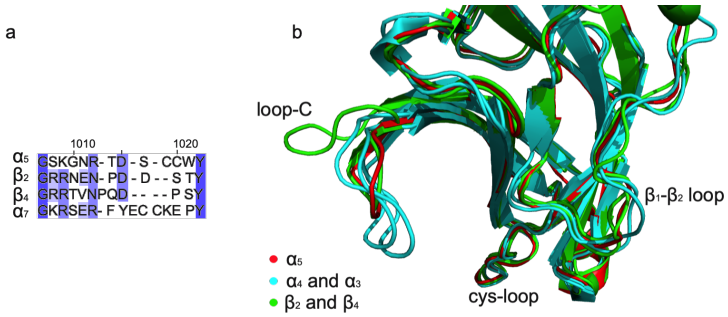

Figure 2.8: **Alignment of modeled loop-C in α_5 with desensitized templates.** a) Multiple sequence alignment between the loop-C of α_5, β_2, β_4 and α_7. α_5 is the target sequence while the rest of the subunits are part of the templates to model the desensitized state 5KXI, 6PV7 and 7KOQ. b) Structural alignment between the modeled α_5 and α_4, α_3, β_2, β_4 and α_7. In green β_2 and β_4, the second has a very extended loop-C in desensitized state. In red α_5's loop-C, with a structure similar to β_2 and shorter than the loop of α_4 and α_3, shown in cyan.

To confirm the functional states of the models before the transition path analysis, I compared the pore profile of the models to the published data of PLGICs. In particular the pore radius and the side chains of the residues of TM2 that delineate the ion pore (Figure 2.9).

In the resting model, the ion pore in the TMD is constricted by the hydrophobic Leu side chains at positions 9' with a minimum diameter (dmin) of 3.00 Å and 16' with a dmin = 3.35 Å as well as by Val 13', dmin = 4.34 Å. In addition, the resting state shows a narrow

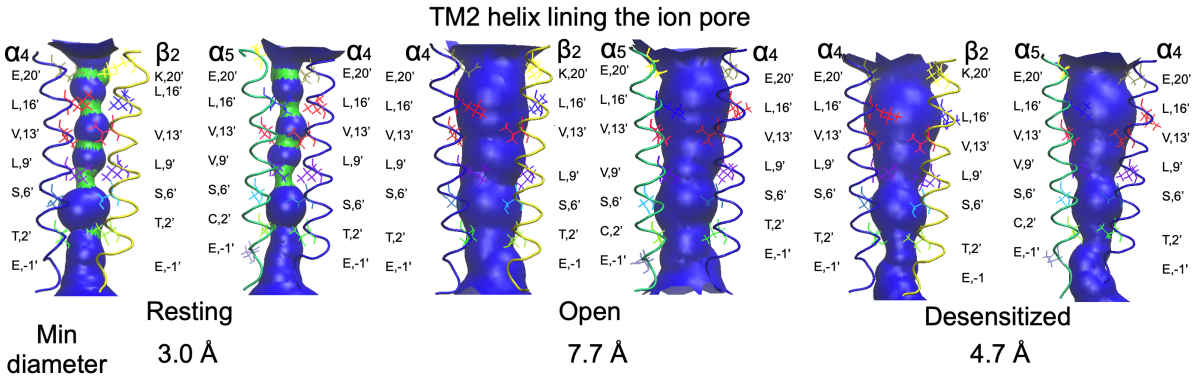

Figure 2.9: **Residues on TM2 helix delineating the pore.** In blue the pore profile of the models in resting, activated and desensitized states as well as the residues delineating the ion pore on the M2 helix of the TMD.

pore in the interface with the cytosol, at the Thr 2' position (dmin = 4.40 Å), a similar behavior was observed from molecular dynamics simulations of the $\alpha_4\beta_2$ receptor [149]. These constrictions are smaller than the estimated radius of a hydrated Na^+ ion, which is 2.76 Å (diameter = 5.52 Å). The same residues contribute to the ion pore blockage in the structures of α_7 [8] and the muscle type nAChR [32] in resting state. On the structure of α_7, is at the level of Leu 9' where the pore is most constricted (dmin = 2.4 Å). The same is observed for the muscle type nAChR with a dmin of 2.8 Å at the Leu 9' position. The Leu 9' constriction is conserved among other PLGICs and is observed both in receptors with a bound antagonist and in apo structures [8]. In the active conformation of the $\alpha_4\alpha_5\beta_2$ nAChR model, the pore diameter near the Leu 9' position increases to 8.14 Å and has a minimum pore radius of 7.70 Å at the Thr 2' position. Similar to the activated structure of α_7, where the ion channel has a minimum pore radius of 7.2 Å at the Leu 9' position. The constriction described in the activated structure of α_7 [8], located in the ECD at E97 and R98 (6.4 A°), within the sequence DER, was also observed in the activated model (dmin = 6.90 Å). The E97 is only present in α_7 where it contributes to ion permeability [8]. In all other subunits there is a glycine and the residues contributing to the constriction are D106, in the sequence DGD of α_4, R101 in the sequence DGR of α_5 and M101 in the sequence DGM of β_2. This constriction in the ECD has not been observed in the open structures of the 5-HT_3R, where the residues at that position are smaller V106, G107 and a K108 which is parallel to the pore axis [128] or on the ECD of the desensitized structure of $\alpha_4\beta_2$ [27, 29] in which the ECD is un-bloomed with the agonist nicotine bound in the orthosteric binding site. Similar to what is observed in the experimental structure of $\alpha_4\beta_2$ [27] and $\alpha_3\beta_4$ [26] on the desensitized model the lower part of the pore is contracted and blocked by T250, 2' and E247, -1' in α_4 and T242, 2' E239, -1' in β_2 while in α_5 there is C242 2' and E230 -1' potentially contributing to the blockage of the pore (dmin = 4.70 Å). The ECD pore has more resemblance to α_7 in desensitized state, with a mild constriction (dmin = 8.44Å) that is not observed in $\alpha_4\beta_2$ or $\alpha_3\beta_4$ (Figure 2.10).

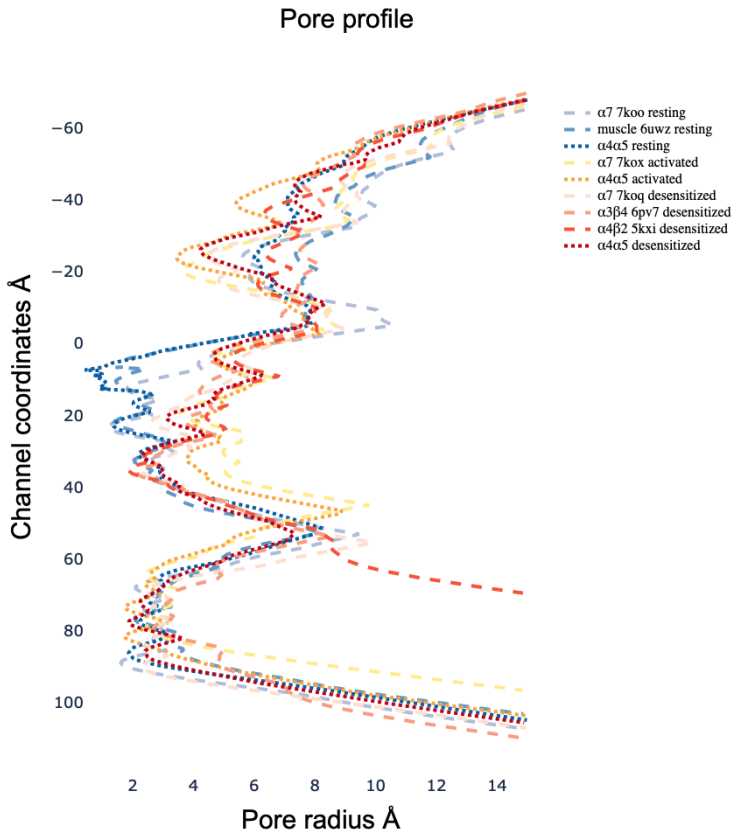

Figure 2.10: **Pore radius along the ion pore of the $\alpha_4\alpha_5\beta 2$ models and templates.** The figure shows on the y axis the channel coordinates starting from the ECD to end on the ICD. The x axis depicts the pore radius in Å. The model and templates in resting state are colored in blue, in orange to yellow the activated state and in reds the desensitized sates. The templates have dashed lines and the models have dotted darker lines.

2.4 Conclusion and perspectives

Without the experimental structure of $\alpha_4\alpha_5\beta 2$ it is difficult to assert the exact structure of this receptor. Based on the analyses I performed and the comparison of the models to the available $\alpha_4\beta_2$ and α_7 structures, we can confirm the orthosteric site in the models shows a conformation that is similar to the experimental structures in all functional states. The pore profile of these models indicates, the models have ion pores blocked in the resting state, open ion pore in the activated state and partially blocked ion pore in the desensitized state. These profiles are similar to what is observed in other experimental structures of nAChRs. The structure of the allosteric binding site between α_5 and α_4 is also similar to the activated and resting experimental structures. The question on what should be the loop-C conformation of α_5 in desensitized state remains debatable.

From the development of these models I have learned which steps to prioritize and how to select the right templates for future projects. I present in this chapter a description and analysis of all the currently available structures that could be used to model an eukaryotic and cation selective PLGICs along with their advantages and disadvantages. In addition the structure of the models of $\alpha_4\alpha_5\beta_2$ in resting, activated and desensitized states, show in overall and agreement with available nAChR structures and can now be used for any other future structural biology studies or drug design projects.

Chapter 3

Calculation of the conformational path of $\alpha_4\alpha_5\beta_2$ and cavity analysis

In this chapter I explain the procedure employed to calculate a series of intermediate conformations describing the gating cycle of the $\alpha_4\alpha_5\beta_2$ nAChR. That is, the conformational changes taking place as the receptor shifts from the resting conformation into an activated functional state and then into a desensitized conformation. This is, to my knowledge, the first time the gating cycle between three functional states has been obtained. Then I visualize and measure these structural changes by following the different cavities formed along the receptor during the gating cycle and I associate these volume changes to a func-

tional state. This analysis allowed me first to validate both the models and the gating cycle and to find new cavities with interesting properties that could be targeted to modulate the receptor.

3.1 Introduction

Proteins are dynamic systems that go through a series of conformational changes associated to biological functions such as protein folding, ligand recognition and catalysis. A detailed understanding of these structural changes is crucial to determine how we can modulate a protein. However, these conformational changes are difficult to study, given the extremely high number of theoretical conformations that could exist when considering all the rotational degrees of freedom each residue in a protein could have. [150] However, not all of these conformations are likely to occur. In fact, we only need to find a small set of low energy favored conformations, connected by intermediate less favorable but accessible states. The low energy favorable conformations can be topically found by experimental techniques like X-ray crystallography and cryo-EM. However, in silico methods remain essential to understand and determine the intermediate states.[7, 151]

In particular, Molecular dynamics (MD) simulations have been fundamental to the study of PLGICs. An atomistic solvent and membrane MD analysis of μs-long, was done on the eukaryotic X-ray structures of PLGICs in activated state. At the end of these long MD simulations the receptor converged to the apo conformation, once ivermectin was removed to abruptly destabilize the active state.This study also described the un-blooming, or closing of the loop-C, and twisting of the ECD with respect to the TMD, which are key conformational changes that drive the ion-pore opening and are observed in all PLGICs. [152] Similar studies have been done on α_7 nAChRs, [153–155] which predicted the ECD and TMD Leu9' constriction in the desensitized state, later on confirmed by the experimental structures. [8] MD simulations on the desensitized X-ray structure of $\alpha_4\beta_2$ showed that by either removing the agonist or replacing it by an antagonist, the loop-C expanded and the receptor was stabilized in a resting conformation. [149, 156] From these studies on $\alpha_4\beta_2$ we can reconstruct the pattern of communication that takes place between the ECD and the TMD to modulate the opening and closing of the ion pore. These process starts at the loop-C in the main subunit (+) at the ECD, the conformational changes happening on this loop when it opens (blooms) or closes (un-blooming) are communicated to the cys-loop and the loop-F of the complementary subunit (-) and then to the M2-M3 linker in the interface of the ECD and TMD domains. This communication occurs through a series of electrostatic and hydrophobic interactions that break and form during the gating cycle. One of the first publications describing a method for transition path sampling between two endpoints was done on the X-ray structure of the GLIC channel in a resting and activated state. The lowest free energy path between both states was found using sting method simulations. Their results contributed to our understanding of the allosteric interactions governing the communication between the ECD and TMD, in particular, the salt bridge (D32-R192) and its influence on the gating cycle. [157]

While current MD simulations remain at the scale of μs, in vitro studies show that the spontaneous opening of the ion pore of nAChRs occurs once every second (1.3 ± 1.4/seg). [41] As it was described in the introduction, different methods to bypass this time frame limitation are focused on transition path sampling.[94] For this project I used an adiabatic path calculation method called Path Optimization Exploration (POE) developed by the Structural Bioinformatics Unit at the Institut Pasteur. POE samples the potential energy landscape sur-

face to find the minimum energy paths connecting low energy favored conformations. To do so it searched for topological shortcuts along a complex transition path generated by CPR and selects shorter paths with lower energy and fewer intermediate points.[158, 159]

A protein's function and regulation is tightly linked to the interactions they can form with other proteins and molecular compounds. The type and strength of these interactions relies on the spatial arrangement and the types of atoms in the protein.[160] The solvent accessible areas between the protein atoms are called cavities. These cavities are delineated by the residue side chains that constitute a protein pocket and determine its molecular properties. [161] The discovery of these pockets is a critical step in the process of drug design. To find the cavities along the $\alpha_4\alpha_5\beta_2$ transition path I used mkgridXf. [162] This method is able to find cavities in the conformational path of a protein by finding the solvent accessible area over the van der Waals atom surface of a protein. We can define the protein pocket of a cavity by extracting the atoms delineating it.

Here I describe and compare the structural changes observed during the gating cycle of $\alpha_4\alpha_5\beta_2$ to the information I have from other in-silico methods applied on experimental or modeled structures of other PLGICs. In addition I analyze the cavities computed by mkgridXf and select a few of them that have large volumes or appear only in the activated conformational frames.

3.2 Methods

3.2.1 Transition path sampling with Path Optimization and Exploration (POE)

After minimizing and equilibrating the receptor in the membrane, it had to be desolvated and extracted from the membrane before the transition path sampling was done with POE in an adiabatic system. POE was executed with CHARMM version 35b2 and the force field for all-atom parameterization CHARMM36m. [90] The contribution of non-bonded and non-local interactions between atoms to the potential energy, was calculated with a sigmoidal distance dependent dielectric function.[163] The CPR algorithm [95] used in POE is implemented in the TReK module of CHARMM. The first iteration of POE constructs an initial path by interpolating between two conformations or end points. Interpolation of the backbone coordinates and of the side chain internal coordinates. This is done using CHARMM Hammer Drill, developed in the Structural Bioinformatics Unit. Then all possible shortcuts along this path are calculated and combined to generate new alternative and shorter transition paths. This process is repeated iteratively and the final path is selected if its curvilinear length (sum of rmsd between adjacent conformations) and its maximum energy are lower than in the previous paths. In order to prevent getting atoms highly displaced while doing the initial guess of a path, the atoms are only authorized to have a displacement of 0.5Å. In addition, to avoid cis- trans isomerizations on the prolines they were restrained to a trans conformation (Ω = 180 °) with the exception of prolines 136 in α_4, 137 in α_5 and 138 in β_2 that were restrained to ensure they would maintain the cis conformation throughout the simulation. I performed 3 POE iterations for the transition path

between the resting-activated states and the activated-desensitized state, and 2 POE iterations for the transition path between the desensitized-resting conformations. I stopped the iteration process once I did not observe a reduction in the potential energy and a reduction in the number of intermediate conformations.

3.2.2 Analysis of the properties of the gating cycle

In order to find the most important motions in the $\alpha_4\alpha_5\beta_2$ gating cycle trajectory needed for conformational changes, I performed Principal Component Analysis (PCA) using the atomic coordinates of each atom and each frame in the trajectory as descriptors. [164]

I also needed to determine if the models and the conformations in the trajectory displayed the conformational changes described to be essential to initiate and regulate the gating cycle of nAChRs. To do so, I compared the transition path to experimental structures using two known descriptors broadly used to characterize PLGICs: expansion of the ECD and the torsion angle between the ECD and TMD. The expansion or spread of the ECD has been used before to describe the structural changes taking place in the ECD as it contracts and expands when complexed with an agonist or antagonist, respectively [149, 154, 157]. This movement can be described using the radius of gyration of the C_α atoms on each chain of the ECD, which quantifies the degree to which each subunit expands inward and outward relative to the central pore. The radius of gyration is defined on Equation 3.1.

$$\frac{\sum_i (x_{mean} - x_i)^2 + (y_{mean} - y_i)^2 + (z_{mean} - z_i)^2}{N} \tag{3.1}$$

Where N is the number of C_α atoms, x_{mean}, y_{mean}, z_{mean} are the center of mass of the C_α atoms and x_i, y_i, z_i are their atom coordinates.

The torsion angle or twisting between the ECD and the TMD measures the torsion of the ECD relative to the TMD around the pore axis. Structural comparisons of resting and activated states of PLGICs indicate that a global twisting or opposite direction rotation of the ECD and TMD around the pore axis is involved in ion channel activation. [149, 153, 154, 157]. The torsion angle was calculated as the average of the dihedral angles obtained for each chain$_i$ between the three following vectors: $\overrightarrow{chain_i ECD, ECD_{com}}$, $\overrightarrow{ECD_{com}, TMD_{com}}$ and $\overrightarrow{chain_i TMD, TMD_{com}}$. Where $chain_i ECD$ is the average of the atomic coordinates of C_α in the ECD of chain i, $chain_i TMD$ is the average of the atomic coordinates of C_α in the TMD of chain i, and TMD_{com} and ECD_{com} are the center of mass of the entire TMD and ECD.

The pocket of the allosteric site between the $\alpha_5\alpha_4$ subunits was extracted and NS9283, a published allosteric modulator [165] as well as nicotine were docked to validate the modeled allosteric site using smina. [99] The molecules used for docking were prepared using a pipeline of in-house scripts that generates the lowest energy tautomers, isomers and conformers. [166] I performed flexible docking allowing the residues PHE95, TRP150 and TYR196 in $\alpha_5(+)$ and TRP55, HIS109 and THR119 in $\alpha_4(-)$ to remain flexible.

3.2.3 Cavity analysis of $\alpha_4\alpha_5\beta_2$

The transition path between resting, closed and desensitized states was combined and used as the input for mkgridXf [162] to determine new and known binding sites throughout the receptor structure. The only mkgridXf parameter that was changed from the default was rou (6 Å), the radius of the large probe that defines the excluded volume considered to be part of the solvent. After cavity calculation, the cavities whose mean volume was not higher than 100 $Å^3$ and the cavities that did not have a volume higher than 150 $Å^3$ at least once during the transition path, were excluded from the analysis.

To find cavities that displayed similar volumes at a given step of the conformational trajectory, hierarchical clustering was performed using Scipy's [167] linkage function, with the following parameters: cosine distance to compute the distance matrix between cavities and the UPGMA algorithm (or average) as the method to combine clusters. These parameters were chosen among all the other options using the cophenetic correlation coefficient implemented in Scipy that measures how well a dendrogram preserves the pairwise distances between the original data points. The cavity volumes were standardized by subtracting the mean and dividing by the standard deviation. The clustered cavities presenting higher volumes during the frames close to and during the activated state, were visually inspected.

As part of mkgridXf's functionalities, I obtained the footprint of the selected cavities, which defines a protein pocket by the atoms or groups of atoms outlining the cavities. From mkgridXf's implementation the delineating residues were encoded by Equation 3.2:

$$fp_g^{real}(c) = \begin{cases} \sigma - \delta(c,g), & if \delta(c,g) < \sigma \\ 0, & otherwise \end{cases} \tag{3.2}$$

where the, σ is 5 Å and $\delta(c,g)$ is the non mathematical distance between the cavity c, with voxel centers c and g a group of protein atoms a, which can be defined by Equation 3.3:

$$\delta(c,g) = \min_{v \in c, a \in g} (d(v,a) - rad(a))) \tag{3.3}$$

The threshold selected for the inner pocket residues was $fp_g^{real}(c) < 3$ Å afterwards, the residues in the outer pocket were those within $radius^2 < 9$ Å from the inner pocket residues.

3.3 Results and discussion

Based on the pore profiles, I assigned the models of $\alpha_4\alpha_5\beta_2$ to a specific functional state and used them as end points to obtain the minimum energy conformational path with POE. The minimum energy conformational paths between end states have a set of 23 frames (or conformations) distributed between the resting and activated states, 23 frames between activated and desensitized states, and 21 frames between desensitized and resting states. These frames represent a progression of the structural changes that occur in a landscape of potential energy, where the end points represent the low energy conformations and the intermediate frames are scattered along the lowest energy saddle points.

3.3.0.1 Global motions analyzed by PCA

PCA Figure 3.1 highlights the most important structural changes occurring in $\alpha_4\alpha_5\beta_2$ gating cycle. The first 3 components have an explained variance of 0.757, 0.218 and 0.014. I observed that the first two principal components represent the structural changes observed as the receptor shifts from being in resting state to becoming activated. That is, the blooming and un-blooming as the loop-C opens and closes and the ECD contracts or expands as well as the global counterclockwise twist of the ECD relative to the TMD. The third component represents smaller changes observed in the TMD along the TM2 helix lining the pore.

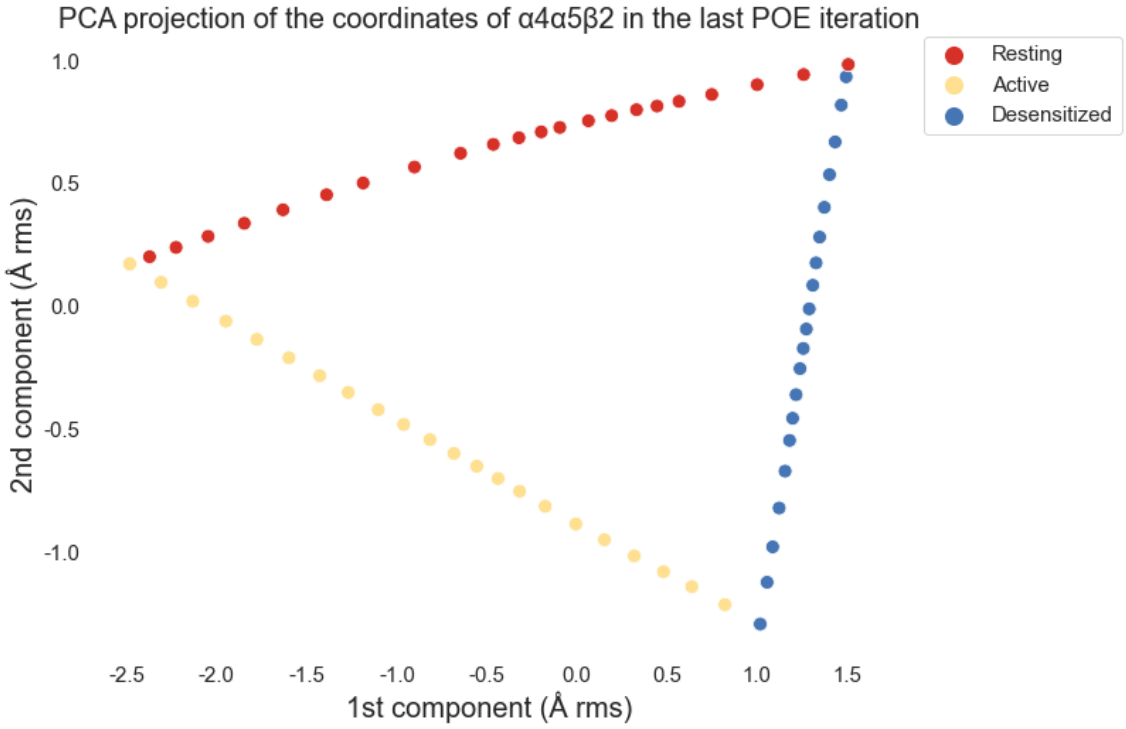

Figure 3.1: **Principal Component Analysis (PCA) of the coordinates on the frames of the $\alpha_4\alpha_5\beta_2$ trajectory.** The x and y axes depict the 1st and 2nd components, each data point is a conformation or frame in the trajectory between the resting-activated-desensitized conformational path.

3.3.0.2 Analysis of the twisting between the ECD and TMD

Afterwards, I compared the twisting of $\alpha_4\alpha_5\beta_2$ to the published data on other PLGICs. As it is shown on Figure 3.2, the twisting of the ECD during the gating cycle of $\alpha_4\alpha_5\beta_2$ is more emphasized as compared to what I observed for α_7 in all three conformations (α_7 resting twisting angle = 17.04°, activated angle= 23.45°, desensitized angle = 16.22°). In the models, the twisting increases from 19.94° in resting state to 26.19° in activated state to then decrease to 24.69° in the desensitized state. α_7 shows a similar pattern of increasing twisting when becoming activated. The difference is observed between the desensitized structures

of α_7 and $\alpha_4\alpha_5\beta_2$. While it seems that for α_7 the opposite direction rotation of the ECD and TMD around the pore axis increases from activated to resting state, in $\alpha_4\alpha_5\beta_2$ the twisting angle decreases to 24.69°. However, I observed a similar twisting angle in the desensitized structure of $\alpha_4\beta_2$ (24.23°). The structure of the desensitized model was probably more influenced by the structure of $\alpha_4\beta_2$ used as template, since comparative modeling methods follow sequence similarity. But overall, the models and the conformations in the trajectory, align with the increasing twisting description between the ECD and TMD after activation, observed on the other PLGICs.

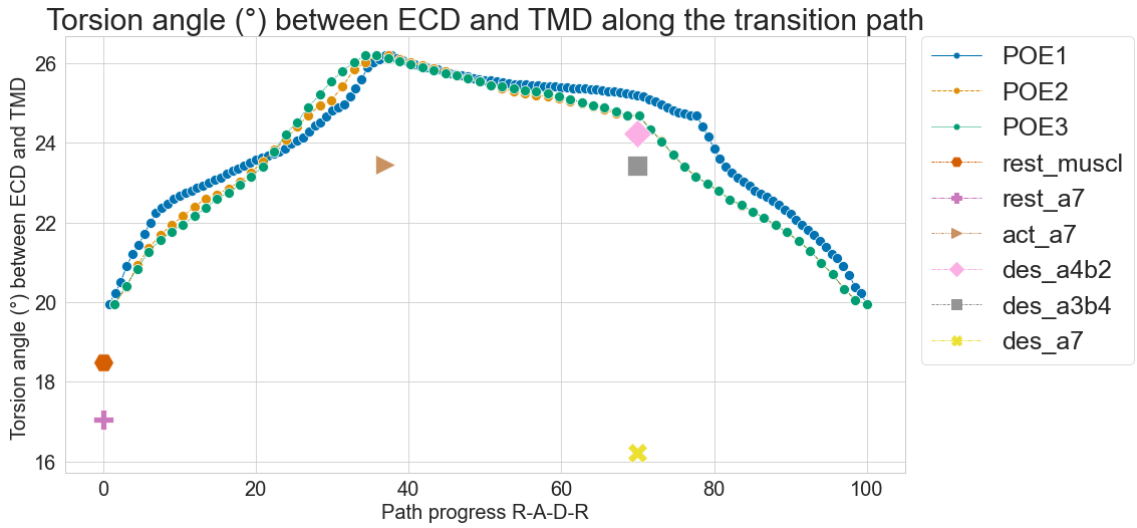

Figure 3.2: **Twisting angle between the ECD and TMD, along the transition path between the resting, activated and desensitized states.** The twisting angles between the ECD and TMD domains for each frame, in the three different POE iterations are plotted in the y axis. The x axis depicts the percentage of progression along the trajectory starting from resting then activated, desensitized and back to resting state. Each dot is a structural conformation in the trajectory between states. For comparison, I included the twisting angle of homologous experimental nicotinic receptors used as templates for modeling.

3.3.0.3 Analysis of the blooming of the ECD

Then I compared the expansion of the ECD during the gating cycle of $\alpha_4\alpha_5\beta_2$, to homologous receptors. On Figure 3.3 the ECD is more expanded (or bloomed) in the resting state (30.28 Å), similar to the structures of α_7 (30.34 Å) and muscle type nAChR (30.37 Å). As the receptor shifts to activated conformation, it un-blooms and the ECD contracts. However, the contraction is less accentuated in $\alpha_4\alpha_5\beta_2$ (29.70 Å) compared to activated α_7 (29.50 Å). When the receptor becomes desensitized the ECD starts blooming (30.40 Å), to the same extent as the resting state. A behavior that is not observed for α_7(30.25 Å), for which the ECD also blooms but stays slightly more compact than in resting state. I believe this accentuated relaxation was derived from the $\alpha_3\beta_4$ template in desensitized state, for which the ECD appears to be even more relaxed or extended than for the resting structures (30.87 Å).

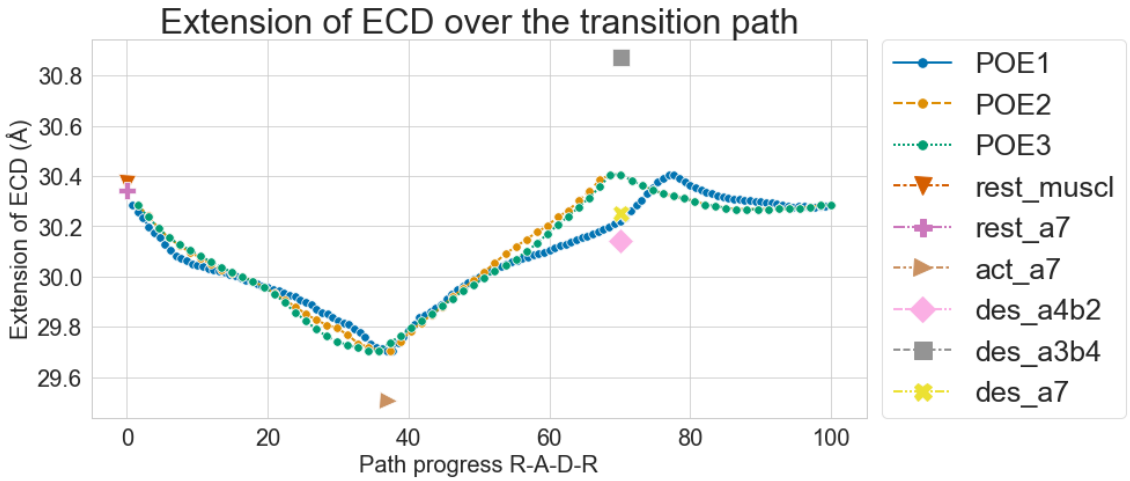

Figure 3.3: **Extension of the ECD along the transition path between the resting, activated and desensitized states.** The extension of the ECD for each frame, in the three different POE iterations is plotted in the y axis. The x axis depicts the percentage of progression along the trajectory starting from resting then activated, desensitized and back to resting state. Each dot is a structural conformation in the trajectory between states. For comparison, I included the ECD extension of homologous experimental nicotinic receptors used as templates for modeling.

3.3.0.4 Cavity analysis

I took the models and templates in the activated functional state and compared the volumes of the cavities capped by the loop-C in the ECD. I observed that in orthosteric binding sites, the loop-C is more tightly closed as compared to the other subunit interfaces (Figure 3.4). The α_7 receptor has an orthosteric site between all subunits, which would explain why its ECD is more compact in the activated and desensitized states, as compared to the $\alpha_4\alpha_5\beta_2$ models. The orthosteric sites along the gating cycle of $\alpha_4\alpha_5\beta_2$, show an open loop-C in resting state with volumes similar to resting α_7 and a compact loop-C in activated and desensitized states similar to open α_7 and desensitized $\alpha_4\beta_2$ and $\alpha_3\beta_4$. During activation, the loop-C on the allosteric site between $\alpha_5\alpha_4$ and the non-orthosteric sites between $\beta_2\alpha_4$ and $\beta_2\alpha_5$ close to a volume comparable to α_7. But as the receptor shifts to a desensitized state the loop C starts to open. A behavior I also observe in the non-orthosteric interfaces of $\alpha_4\beta_2$ and $\alpha_3\beta_4$, although the volumes are smaller in these homologous structures.

3.3.0.5 Correlation between twisting and blooming in the transitional modeling

When I analyzed the synchrony between the twisting and blooming movements in the receptor shown on Figure 3.5, the trajectory of $\alpha_4\alpha_5\beta_2$ indicates there is an inverse correlation between the twisting and the blooming and these two movements appear to occur simultaneously, except to a certain extent between the desensitized and resting states. When the receptor is in its resting state, the twisting between the ECD and TMD is at its lowest angle and the ECD is more extended. As the receptor becomes activated, the ECD is less

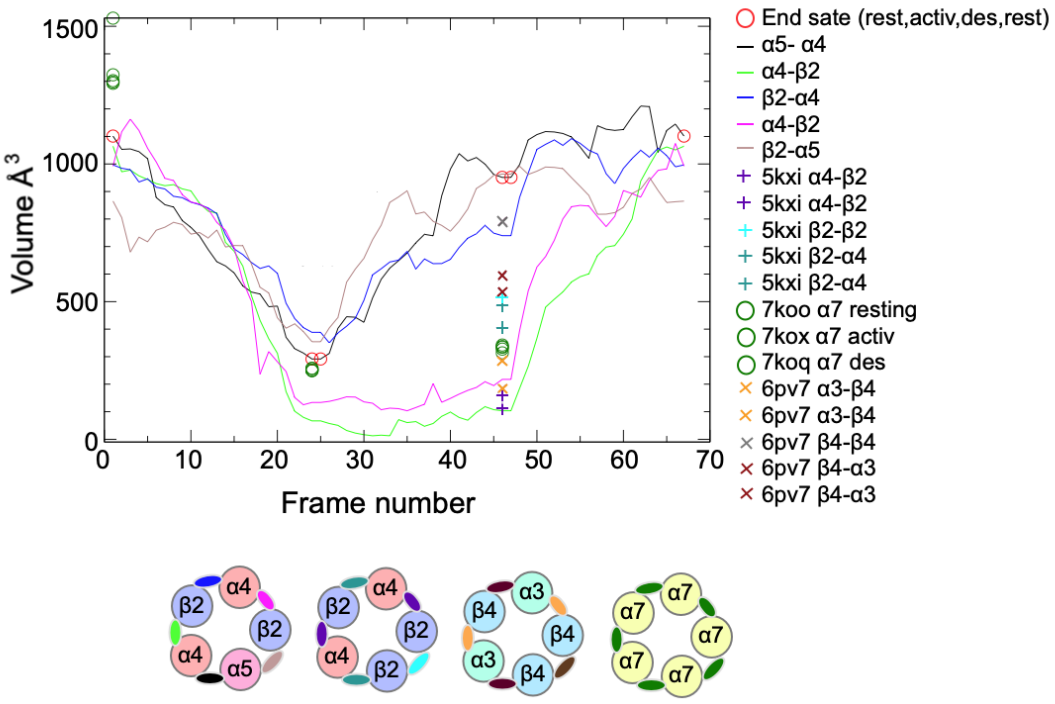

Figure 3.4: **Cavity volumes in allosteric and orthosteric binding sites of $\alpha_4\alpha_5\beta_2$ and templates.** This figure depicts the cavity volume changes in all the interfaces of the $\alpha_4\alpha_5\beta_2$ and compares it to the templates.

extended and the twisting between domains increases to its maximum. From the activated state to the desensitized state, the twisting between domains is preserved but the ECD starts to expand and is more expanded than the resting state. The expansion of the ECD on desensitized state was unexpected to observe but as it had been discussed, it is likely a bias inherited from the multiple sequence alignment of α_5 with β subunits and the presence of only two orthosteric binding sites, which are more tightly closed in activated and desensitized state. As the conformation shifts from desensitized state to resting states, I observed two less synchronized movements, first the receptor starts to un-twist and afterwards, the ECD appears to have a slight contraction.

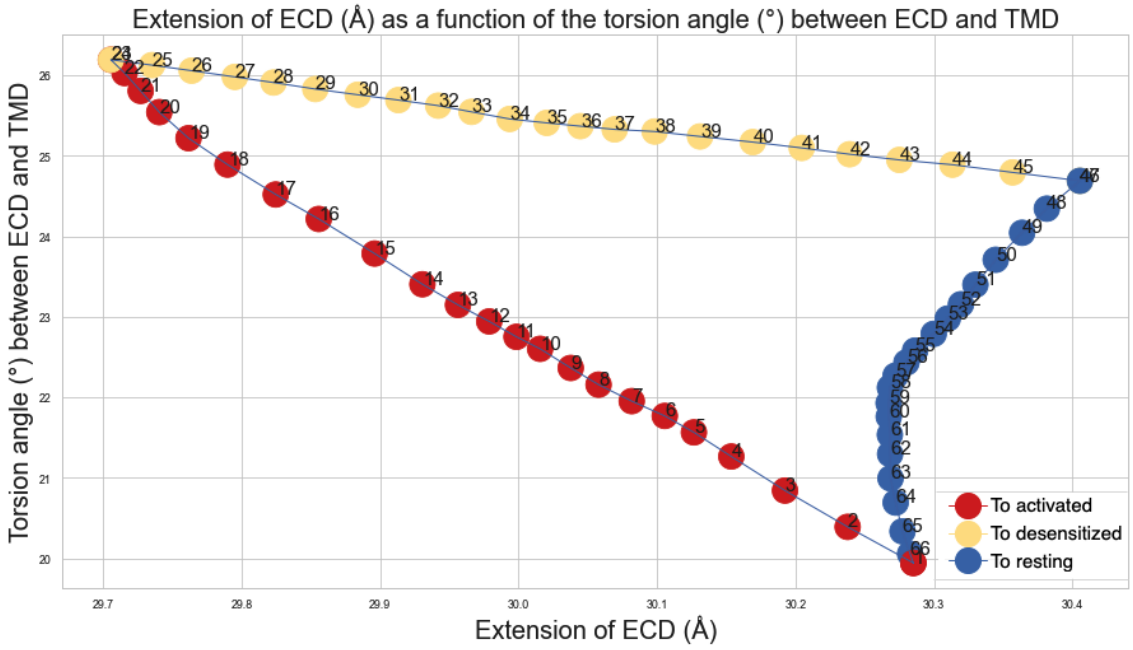

Figure 3.5: **Correlation between the twisting and blooming movements.** The x axis represents the extension of the ECD (Å) and the y axis the torsion angle (°) between ECD and TMD. Each data point represents a conformation in the gating path trajectory. The conformations between resting and activated state are in red and numbered from 1 to 23, from activated to desensitized state are light orange numbered from 23 to 43 and from activated to desensitized are in blue from 43 to 66.

3.3.0.6 Docking on the orthosteric and allosteric site

Experimental data suggests that nicotine does not bind in the interface between $\alpha_5\alpha_4$ [58], this is most likely due to the absence of the TYR100 on α_5, which is present on α_4 and forms a hydrogen bond with the nitrogen on the pyrrolidine of nicotine, as it is shown on panel b of Figure 3.6. Other differences between the orthosteric ($\alpha_4(+)\beta_2(-)$) and allosteric ($\alpha_5(+)\alpha_4(-)$) binding sites are the THR119 and LYS57 on $\alpha_4(-)$. These two residues in the interface of the allosterinc binding site can act as hydrogen bond donors, however, on the orthosteric site with $\beta_2(-)$ as the complementary subunits these residues are substituted by LEU121 and THR59. The leucine could form weaker hydrophobic interactions with a

ligand and threonine has a smaller side chain that is distant form the binding site. Another structural difference I could observe between the orthosteric and allosteric binding site is the length of the loop-C, which is shorter in $\alpha_5(+)$ as compared to $\alpha_4(+)$. This difference appears to keep the disulfide bond away from the binding site and the loop-C slightly more open in activated and desensitized states (panel a of Figure 3.6). From the docking of NS9283 on the orthosteric site in activated state (panel c of Figure 3.6), we observe this compound forms a cation-π interaction between the aromatic 1,2,4-oxadiazole and ARG189, hydrogen bonds between LYS57, THR119, and the cyano group in the benzonitrile, hydrogen bonds between ARG189 and the nitrogen in the 1,2,4-oxadiazole group as well as π stacking between the benzoitrile and TRP150. I also observed the conformation of the side chain of ARG189 at the bottom of the allosteric binding site is quite important to determine the types of interactions that can occur between the ligand and the protein.

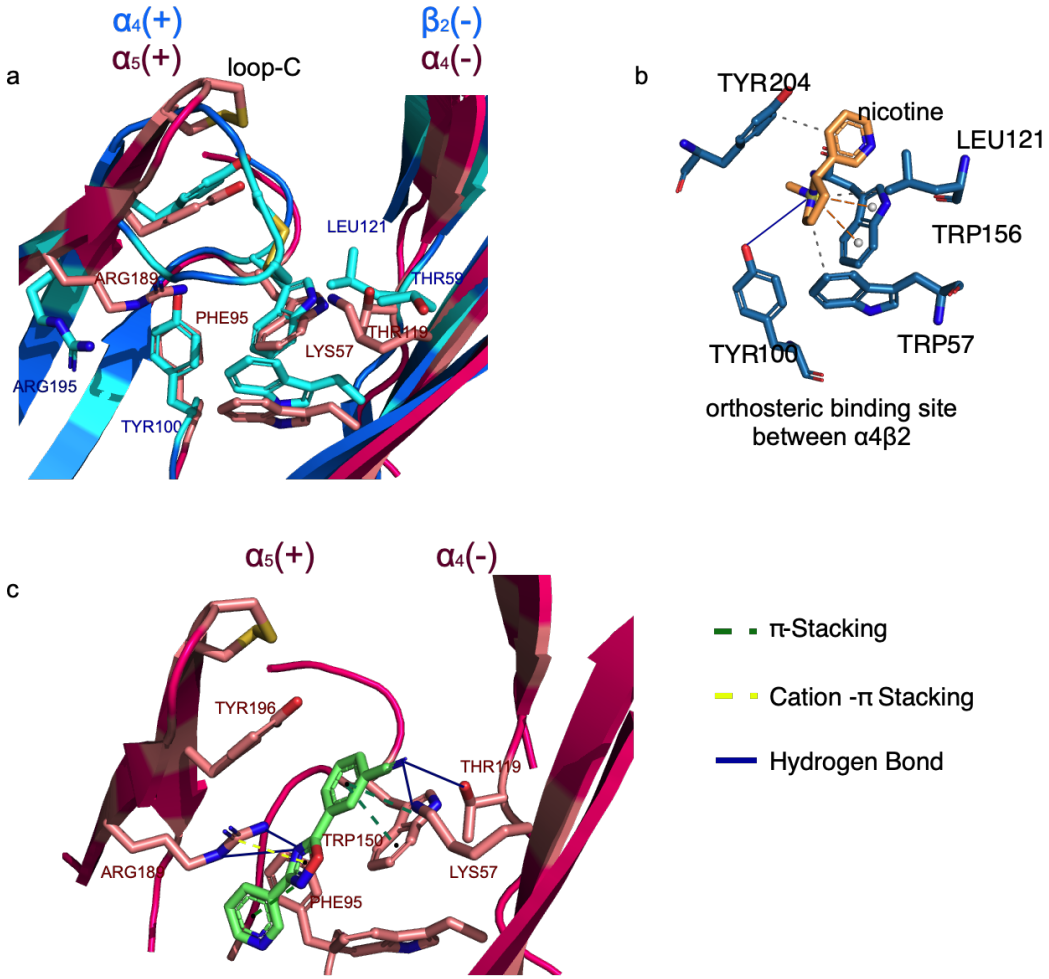

Figure 3.6: **Orthosterc site between $\alpha_4(+)\beta_2(-)$ and allosteric site between $\alpha_5(+)\alpha_4(-)$** a) Alignment between the orthosteric site in the experimental structure of $\alpha_4\beta_2$ (5kxi) and the modeled allosteric site, the residue differences between binding sites are indicated in light pink and cyan. b) nicotine bound to the orthosteric site of $\alpha_4(+)\beta_2(-)$. c) PAM ARG189 bound in the allosteric site between $\alpha5(+)\alpha4(-)$.

3.3.0.7 Clustering of cavities to identify plausible effector sites

The global gating cycle trajectory of $\alpha_4\alpha_5\beta_2$ contains 67 conformations or intermediate structures describing the activation and deactivation of the receptor. This trajectory was aligned to perform cavity detection with mkgridXf. The RMSD between resting and activated conformation is 3.11 Å, between activated and desensitized conformation is 2.46 Å and between desensitized and resting conformations is 2.054 Å. 255 cavities were found to be formed at least once along the transition path. On average, each frame has 30 cavities and the maximum number of cavities formed in a single frame is 67 while the minimum is 1. Only 55 cavities appear in more than 75% of the trajectory, 94 on less than 25% of the trajectory and 41 on less than 10%. In terms of volumes, the maximum average volume along the trajectory is observed in the allosteric site between $\alpha_5 - \alpha_4$ (666.98 Å^3) and in the non-orthosteric sites between $\beta_2 - \alpha_4$ (550.24 Å^3) and $\beta_2 - \alpha_5$ (405.07 Å^3), all of then capped by the loop-C. Only 22 cavities have average volumes higher than 100 Å^3 and 52 cavities have at least once during the trajectory a volume higher than 150 Å^3. For comparison, the volume of a water molecule at 20°C is around 30 Å^3 and its van der Waals exclusion volume about 12 Å^3 .[168] These 52 cavities were selected for the clustering analysis. I obtained 5 clusters and cavity 79, in the interface of an α_4, β_2 subunit near the α-helix in the N-terminal end, as an outlier Figure 3.7. The most populated cluster is cluster 2 with 21 cavities while the least populated clusters is cluster 1 with only 3.

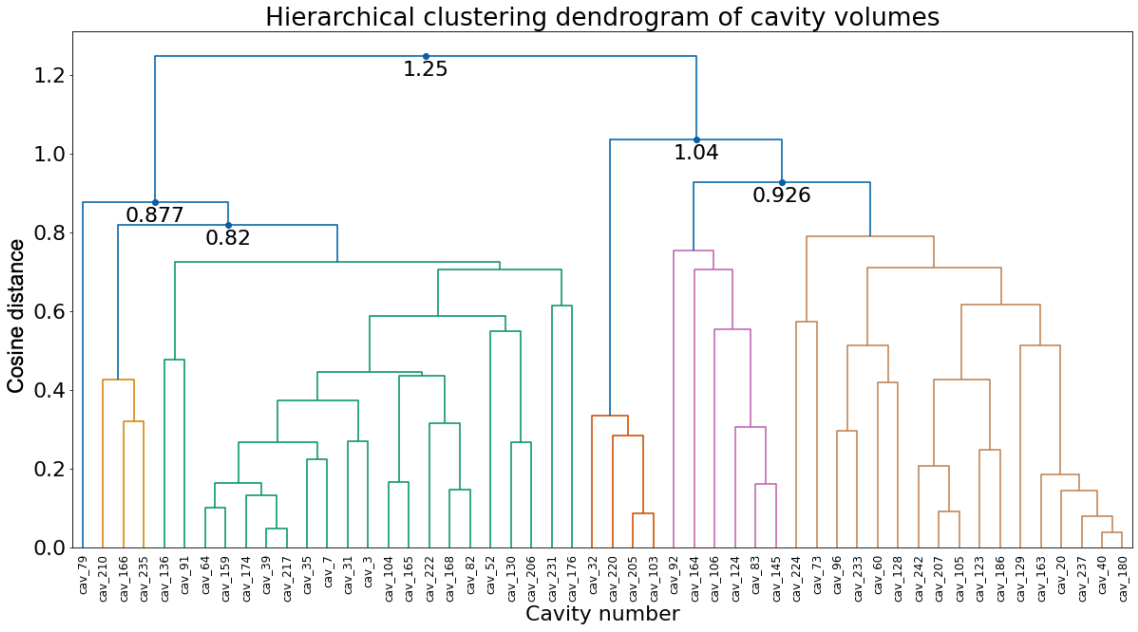

Figure 3.7: **Clusters of cavities using the volume Å^3 as descriptor.** The y axis depicts the cosine distance and the x axis the cavity number as selected by mkgridXf. There is one outlier, cavity 79 and six clusters identified with different colors.

To facilitate the visualization I averaged the volume values within clusters and obtained the volume profile of each of them. On Figure 3.8 cluster 2 with 20 cavities, appears to gather cavities with volumes predominantly higher when the receptor is in activated state, this cluster is interesting because I could try to occupy these cavities with a ligand to preserve

the cavity volume and observe if the receptor is maintained in activated conformation. Cluster 3, with 4 cavities is also interesting, since the cavities appear to have higher volumes only in resting state. Cluster 5 gathers all the cavities capped by the loop-C and shows the lowest volumes during the activated state.

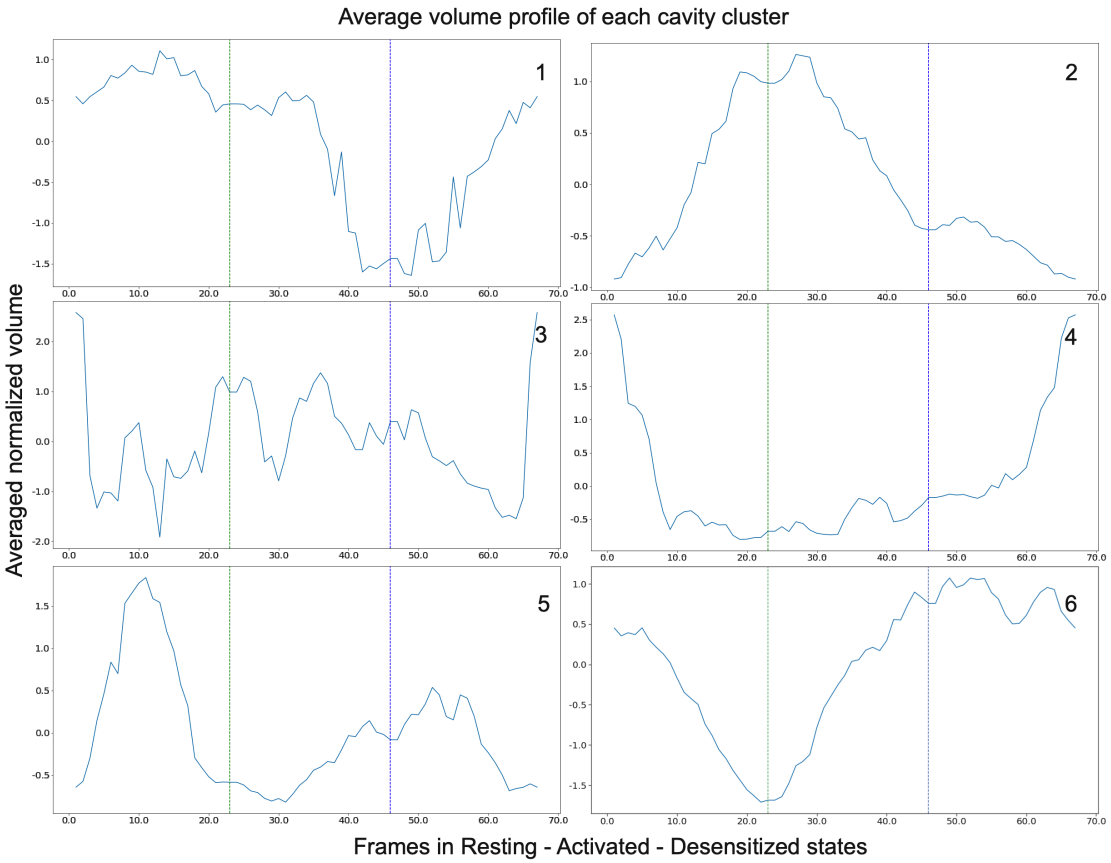

Figure 3.8: **Cavity volume profile for each cluster.** The averaged volumes of all the cavities in each cluster is plotted on the y axis. The x axis shows the ordered frames of the transition path progression resting-activated-desensitized. The number on the top-right corner indicates the cluster number, the first vertical line in green is the frame with the activated state conformation and the second vertical line in blue is the frame with the desensitized conformation.

Among all the cavities in cluster 2, I selected those appearing on the α_5 subunit or in the interface between another subunit and α_5 with the idea that these cavities could be targeted to selectively activate receptors with an α_5 subunit Figure 3.9. In the end I selected a cavity in the ECD (cavity 206, shown in green) located on loop-B facing the ion pore and with a volume between 400-460 Å^3 during the activated state. As well as three cavities in the TMD: a cavity located in the interface between β_2 (TM2-TM3) and α_5 (TM1-TM2), with a maximum volume of 350-420 Å^3 during the activated state (cavity 174), a second cavity located in in the TMD of α_5 between TM1,TM2 and TM3 with a maximum volume in the activated state between 250-290 Å^3 (cavity 104) that in activated state extends and connects to another cavity (cavity 165) in the TMD between α_5 and α_4 with volumes around 170 Å^3

in activated state.

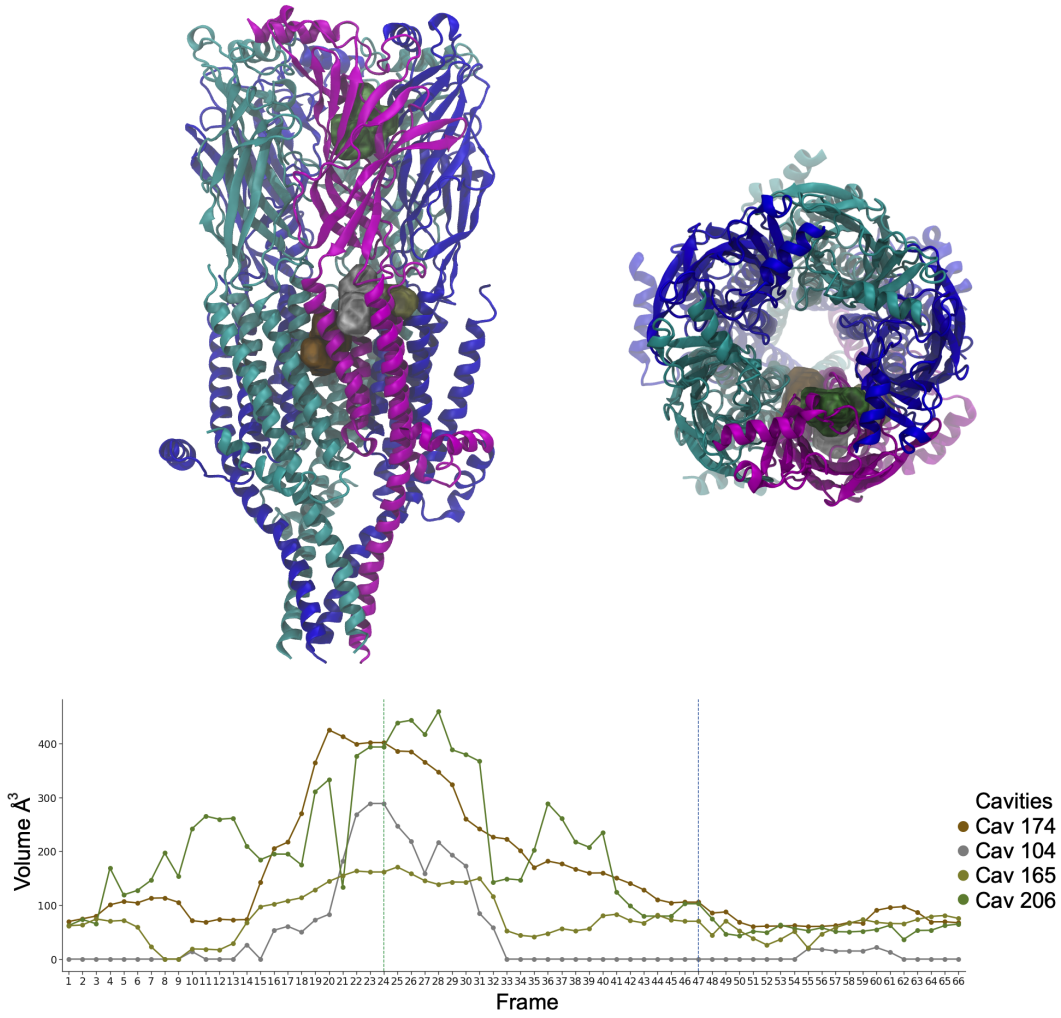

Figure 3.9: **Cavities in α_5 selected from clustering analysis.** The upper left figure depicts the $\alpha_4\alpha_5\beta_2$ nAChR where α_4 subunits are shown in blue, α_5 is in magenta and $beta_2$ subunits are shown in cyan. All the cavities are located in the ECD or TMD of α_5. Upper view of the receptor showing cavity 206 in green in the ECD facing the pore. The plot on the bottom of the figure shows the cavity volume change on each frame and is color coded with the same color as the upper left figure.

3.4 Conclusion and discussion

Here I showed that the gating cycle obtained from the transition path sampling between the models used as endpoints, presents the un-twisting and twisting as well as the bloomed and un-bloomed structural movements essential to the initiation and regulation of the pore opening. The twisting and un-blooming movements appear to happen simultaneously, although I would have expected the un-blooming to start before the twisting of the receptor

as I would expect the loop-c closing to be the first structural change occurring after activation. However, some experimental results show that the closing of loop-C is not necessary to induce the channel opening and that there might not be a consecutive sequence of structural changes leading to the activation of the receptor. [169] In addition, the cavity analysis showed the presence of the allosteric and orthosteric binding sites and demonstrates the orthosteric sites show volume profiles changing according to the loop-C expansion, which is what we observe in experimental structures. I identify and report cavities that can be further studied in future research projects to maintain the receptor in activated state and validate the allosteric site by docking of NS9283 and nicotine.

Chapter 4

De novo compound generation for $\alpha_4\alpha_5\beta_2$ nAChRs

In this chapter, I explain the reasoning behind why I chose this generative model to propose compounds to be tested for $\alpha_4\alpha_5\beta_2$, the data gathered to train the model and the training process. I present ligand shapes generated by the shape generation network and compare the results for proteins in the evaluation and test set. Then I show the captioning network is capable of decoding these shapes into valid SMILES and provide some insights into the relationship between the generated ligands and the generated shapes. I report a predicted synthetic route for the compounds with good synthetic feasibility scores and with the best docking scores.

4.1 Introduction

De novo compound generation means proposing new chemical structures that satisfy a molecular profile. This molecular profile can be a desirable biological activity or a set of molecular properties. [170]

Virtual screening of existing libraries is usually the method of preference to identify or short list compounds that could modulate a target. However, the difference between virtual screening and *de-novo* generation of compounds is that for virtual screening the molecules must be known *a priori*. This is a limiting factor considering the chemical space of potential drug-like compounds is extremely large, with estimates ranging between 10^{23} to 10^{60} molecules. [171] It is unlikely that we would be able to navigate the chemical space and select a subset of molecules that is both representative of all areas of the chemical space and with a number of compounds that is still feasible to evaluate with a reasonable time frame. The goal behind *de-novo* generation of compounds is to navigate the chemical space more efficiently by considering fewer molecules to be evaluated, but that were designed to have a set of desired properties. [172] Nonetheless, the actual usefulness of de-novo drug design, depends on our ability to translate an in-silico generated chemical structure into a molecule that can be experimentally tested. [173] Another limitation of these methods is the inconsistent consideration of the stereoisomers of a molecule, which leaves the end user with the task of deciding the most relevant enantiomer before synthesis. Other difficulties have also been encountered while trying to design an objective function that could allow us to ensure that we will obtain compounds with the desired properties. [170]

A collaborative project between the Channel Receptors Laboratory and the Structural Bioinformatics Unit at the Institut Pasteur as well as the Natural Product Chemistry group, BioCIS at Paris-Saclay University was initiated with the aim to identify potential allosteric modulators of the $\alpha_4\alpha_5\beta_2$ nAChRs. An important part of this project was developed by Dr. Laura Ortega Varga and it involved the design of an in silico nAChR-tailored database containing fragmented known orthosteric and allosteric nAChR ligands, natural alkaloids and commercially available analogs of these compounds obtained from the ZINC12 database. [174] This large database was screened in silico, to select the fragments with the best docking scores to be tested experimentally and 4 hits were obtained (unpublished data). The latest experimental data produced by Gabrielle Dejean, from the Channel Receptors Laboratory, suggests that one of the compounds studied during this research project selectively inhibits $\alpha_4\alpha_5\beta_2$ nAChRs. In order to expand the chemical space that has been previously studied and by suggestion of the group at Paris-Saclay University, we decided to explore generative models to propose new compounds that could be experimentally tested. We were inclined on the method published by Skalic et.al. [126]. We found this method to complement well our project since it is trained on three-dimensional grid representations of protein-ligand pairs, to generate chemical compounds with molecular properties complementary to a protein pocket. Therefore, once the models were trained we would be able to generate compounds for the new pockets extracted from the cavity analysis described on the previous chapter. These compounds could either be synthesized and experimentally tested or be used as starting points to create a new virtual screening library with new commercially available compounds but targeting different pockets in the receptor.

The models implemented by Skalic et.al. [126] are two generative models composed of several neural networks that can be subdivided in two steps. The first step uses a bicycleGAN [127] that is trained with protein ligand pairs to learn to generate ligand shapes complementary to the protein pocket. We will call this first part of the model the shape generation network. The second step combines a convolutional neural network or CNN-Encoder to extract a vector of features representing the ligand shapes and an LSTM-Decoder learns to translate these input features into SMILES strings with correct syntax. We will call the second part of the model the caption network (Figure 4.1).

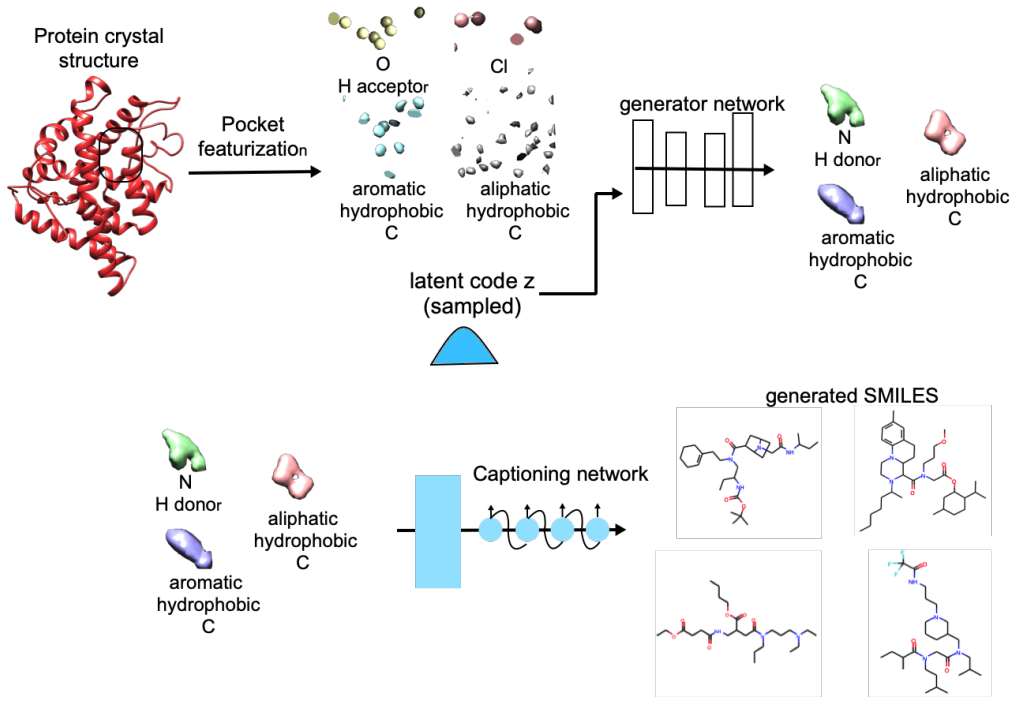

Figure 4.1: **Generation of molecules with the shape generation network and captioning network.** The first component of the generative model is the shape generation network that once trained takes as input a protein pocked voxelized into a grid with 14 channels describing molecular properties and a sampled vector from a standard normal Gaussian distribution. From this input the generator will generate ligand shapes. The captioning network takes as input the generated ligand shapes grids and decodes them into the SMILES representation of molecules.

In this chapter, I will describe the expanded DUD-E database, that I curated in order to have more nAChRs and binders to these nAChRs with reported affinity. In addition I will also describe the difference between the generative models published by Skalic et.al. [126] and the models that I trained. Finally I will evaluate the generative models by generating compounds complementary to the pockets extracted from the cavity analysis of the $\alpha_4\alpha_5\beta_2$ nAChR as well as other proteins in the evaluation and test set.

4.2 Methods

4.2.1 Shape generation network

The shape generation network is a bicycleGAN, that was implemented to avoid mode collapse and produce a distribution of different outputs in the domain of multimodal image-to-image translation.[127] Its objective is to encourage a bijection between the output and latent space, Figure 4.2.

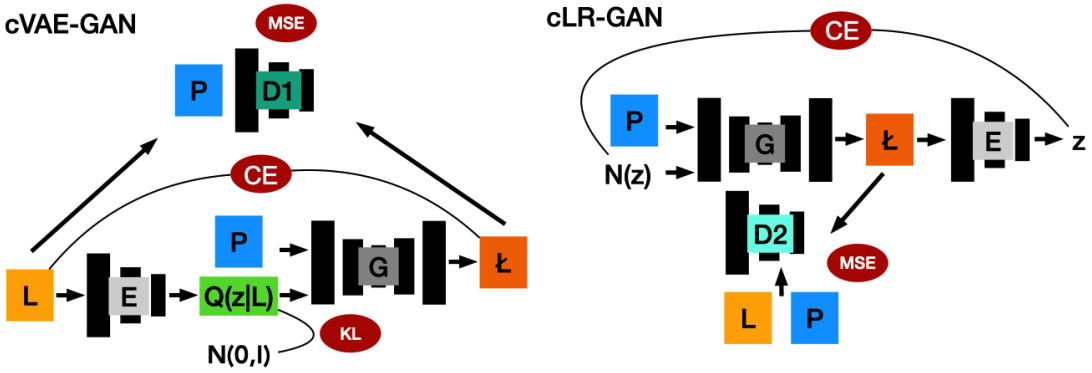

Figure 4.2: **Schema of the shape generation network.** The network architecture is the same as for the bicycleGAN, except the input is the 3D grid of the P and the L and the output is the 3D grid of the Ł. E is the encoder, D is the decoder and G is the generator. CE is the cross entropy between real and generated ligand. The adversarial loss from the discriminators, D1 and D2, between protein real-ligand and protein generated-ligand is the MSE. On the cVAE-GAN schema, Q(z|L) is the distribution of the latent variable z given the ligand shape L, N(0,I) is standard normal Gaussian distribution and KL is the KLDivergence. On the cLR-GAN schema, N(z) is a randomly sampled vector from a standard normal Gaussian distribution.

4.2.1.1 Data collection and subsetting

Similarly to the original publication, [126] all ligands taken from DUD-E [175] were reprocessed for docking, with a pipeline of in-house scripts to obtain the most predominant tautomer conformations, stereoisomers and low energy three-dimensional conformations. I compared the binding sites of the proteins on the DUD-E database to the allosteric binding site in $\alpha_5 - \alpha_4$ using probis [176] and observed that kinases, with very similar binding sites among them, were over represented in the dataset and none of the protein binding sites presented similarities to $\alpha_4\alpha_5\beta_2$ allosteric binding site. Therefore, I decided to expand the database to include nAChRs structures but also distinct proteins with binding sites similar to the allosteric and orthosteric binding sites in nAChRs.

The first step involved taking as reference 17 known nicotinic ligands.[177] and find-

ing all the proteins to which they could bind. I used the swiss target prediction web tool [178] and retrieved all the resulting UniprotIDs. [70] Afterwards I used those UniprotIDs to extract their protein structures from the PDB database [72]. Those protein structures with a bound ligand and the best resolution were manually selected and curated. Afterwards I used the BindingDB [179] API to extract the compounds with reported affinity values. I kept all ligands with annotated activity data (Ki, IC_{50}, EC_{50}) better than 1 μM to their target, molecular weights lower than 600 Da and fewer than 20 rotatable bonds. In the end only the proteins with 30 binders were prepared and kept. All proteins and ligand pairs were prepared for docking with smina, as described before.

The enriched database includes 68 enzymes, 35 proteases, 32 kinases, 27 GPCRs 17 nuclear receptors, 10 ion-channels, among which are α_7, $\alpha_4\beta_2$, $\alpha_3\beta_4$ and the GABA-A receptor and 40 proteins with diverse functions (Figure 4.3). With this database, I was able to train my generative network with the orthosteric binding site of homologous proteins and with a more diverse and balanced group of proteins.

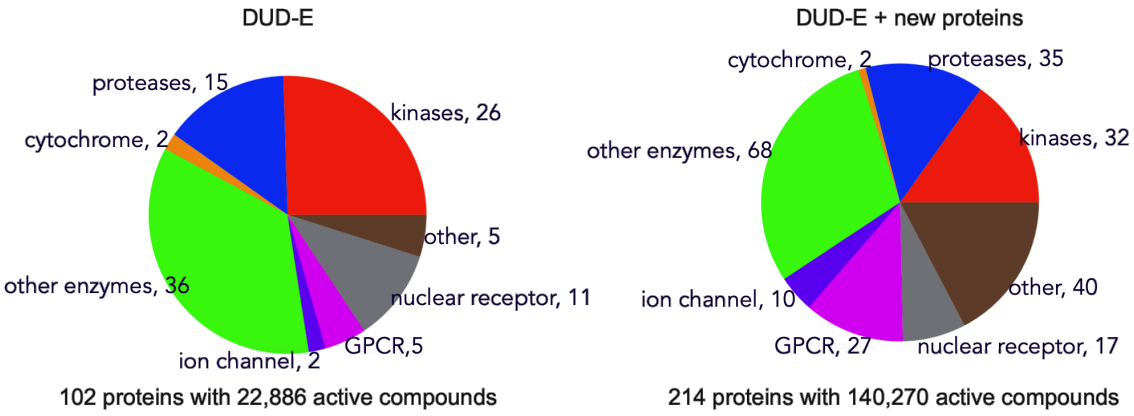

Figure 4.3: **Comparison of the proteins and ligands in DUD-E and the DUD-E + new proteins.** This figure depicts the protein classification and the number of proteins in each group. In the extended database the kinases are no longer over represented

Afterwards, I compared again all the binding sites of the proteins in the new database that included both the previous DUD-E docked ligands and proteins and the new protein-ligand pairs curated to extend DUD-E. I used this comparison to select the training, evaluation and test sets. The training set contains both proteins that have unique binding sites and proteins that have similar binding sites to those in the evaluation set (Figure 4.4). For the test set I kept only the subset of proteins with unique binding sites that had between 300 and 3000 binders. The test set will allow me to test if the network is able to generalize and generate ligands even for protein pockets different from those used for training. The binding sites were again compared with probis; two binding sites were considered to be similar if more than 10 nodes overlapped with an E-value lower than 1x10-4 and a RMSD lower than 2.0 Å.

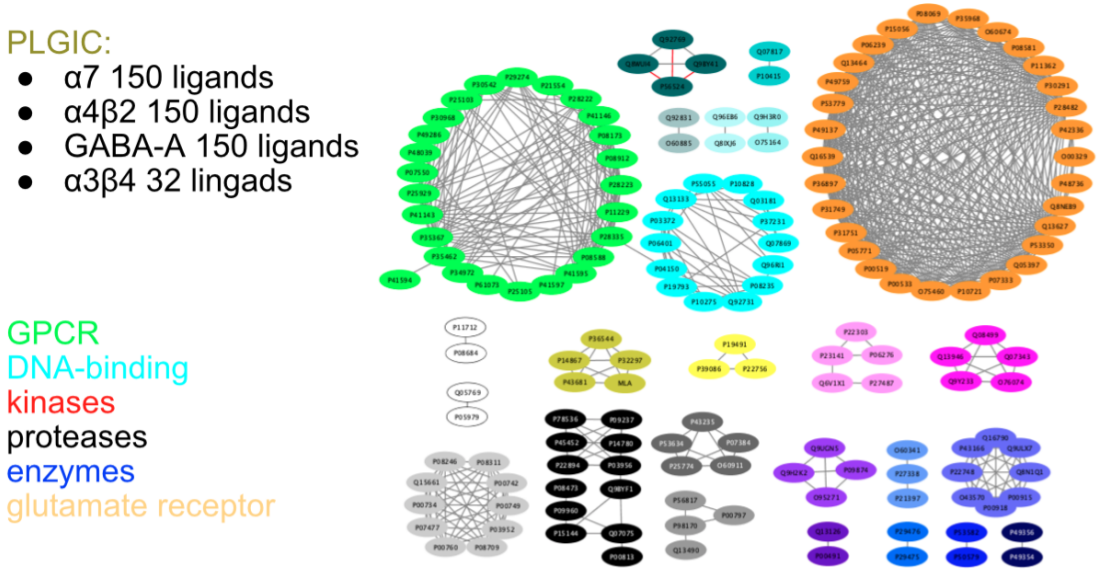

Figure 4.4: **Proteins in the extended-DUDE database grouped by binding site similarity.** The binding sites were compared with probis. Each node represents a protein and an edge was added if two binding sites were similar, with more than 10 nodes overlapped, an E-value lower than 1x10-4 and a RMSD lower than 2.0 Å. The color codes represent different protein families, named on the left of the figure

The final training database includes 148 proteins that have at least another protein with a similar binding site and 39 proteins with no similar binding sites within this database; each protein has between 30 and 150 docked ligands. The test set includes 20 proteins with no similar binding sites within this database and each protein has between 46 and 3000 docked ligands. The rest of the proteins were kept in the evaluation set.

4.2.1.2 Training setup

I used Libmolgrid [125] to translate the three-dimensional coordinates of the protein and ligand pairs into a 3D grid representation. This grid is a vector of three-dimensional grids of atom density called channels and each channel describes an atom property. In my case I had 30 channels, the first 14 channels correspond to protein molecular properties and the last channel is the void channel, which completes for the density of the voxels that had values lower than the summed max density. The next 15 channels correspond to the 14 ligand molecular properties and the void ligand channel:

Protein channel properties: Aliphatic Hydrophobic Carbon, Aliphatic NonHydrophobe Carbon, Aromatic Hydrophobic Carbon, Aromatic NonHydrophobe Carbon, Halogen, Nitrogen Acceptor, Nitrogen HDonor HAcceptor, Oxygen HAcceptor, Oxygen HDonor HAcceptor, Sulfur, Phosphorus, Calcium, Zinc, GenericMetal, protein void.

Ligand channel properties: Aliphatic Hydrophobic Carbon, Aliphatic NonHydrophobe

Carbon, Aromatic Hydrophobic Carbon, Aromatic NonHydrophobe Carbon, Bromine or Iodine, Chlorine, Fluorine, Nitrogen HAcceptor, Nitrogen HDonor HAcceptor, Oxygen HAcceptor, Oxygen HDonor HAcceptor, Sulfur, Phosphorus, GenericMetal, ligand void.

To compute the atom density at a grid point, the function shown on Equation 4.1 and implemented in Libmolgrid, was used. In this function, r is the van der Waals radius and d is the distance from the atom center. The function A(d, r) is a continuous Gaussian from the atom center to the van der Waals radius, that turns to 0 when the distance is 1.5 times the van der Waals radius. [112]

$$A(d,r) = \begin{cases} e^{-\frac{2d^2}{r^2}}, & 0 \leq d \geq r \\ \frac{4}{e^2 r^2}d^2 - \frac{12}{e^2 r}d + \frac{9}{e^2}, & r \leq d \geq 1.5r \\ 0, & d \geq 1.5r \end{cases} \tag{4.1}$$

The grids had a resolution of 1.0 and a dimension along each side of the cube of 24 Å. This process is also called voxelization. To locate the binding site of the protein to be voxelized, the center of the docked ligands was taken as reference to set the grid position. The atom density values of both the protein and the ligand were translated and randomly rotated before being voxelized during the training process.

The network architecture was kept as described by Skalic et.al. [126] However, to train the BicycleGAN the cVAE-GAN and cLR-GAN objective functions were changed and combined into Equation 4.2.

$$\begin{aligned} G, E = argmin_{G,E}\, max_D\, MSE_{GAN}^{VAE}(G, D, E) \\ +\lambda CE^{VAE}(G, E) + MSE_{GAN}(D, G) \\ +\lambda_{latent} L_1^{laltent}(G, E) + \lambda_{KL} L_{KL}(E) \end{aligned} \tag{4.2}$$

Where G is the generator, E the encoder and D the discriminator. $MSE_{GAN}^{VAE}(D, G, E)$ represents the mean squared error adversarial loss of the cVAE-GAN, $CE^{VAE}(G, E)$ is the Cross Entropy (CE) loss measuring the reconstruction of the ligand, $MSE_{GAN}(G, D)$ is the mean squared error adversarial loss for the cLR-GAN, L_1^{latent} is the L_1 latent reconstruction loss and L_{KL} is the KLDivergence distribution loss between the Encoder's distribution and a standard normal Gaussian distribution. The parameters λ, λ_{latent} and λ_{KL} are weights controlling the relative importance of each term and were set to λ =10, λ_{latent}= 0.5 and λ_{KL}=0.01.

As input to the discriminator, I concatenated the grid representation of the protein and ligand. The dimension of the latent vector (Z) was set to 8. Networks were optimized with the Adam optimizer, with a learning rate of 5.10^{-4} that was kept constant during training. The training was done for a total of 5,798 epochs. I measured the Pearson Correlation coefficient per channel, between the training ligand and the generated ligand, to monitor the evolution of the ligand reconstruction. The training was stopped once the Pearson Correlation, cross entropy and the adversarial loss stopped improving.

4.2.2 Captioning network

The captioning network has two components, a VAE and an Encoder-Decoder. The VAE is only used during training, to add noise to the voxelized ligands so that they can resemble the generated ligand shape outputs from the BicycleGAN. The Encoder is a CNN that will produce a vector representation of the ligand shape for the Decoder, which is an LSTM network and will translate the feature vector into a sequence of SMILES, Figure 4.5.

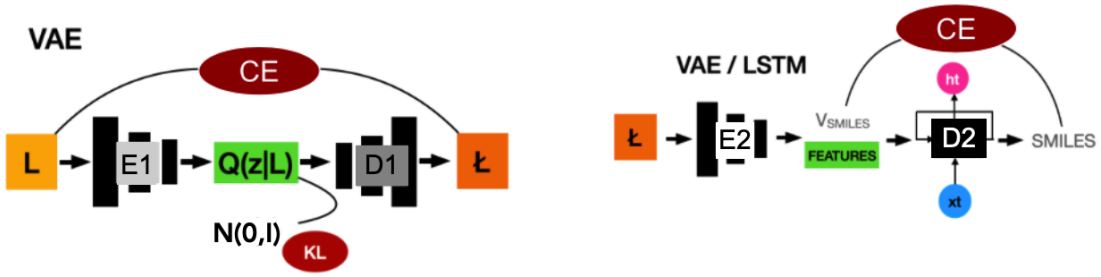

Figure 4.5: **Schema of the captioning network architecture.** The VAE is used during training to add noise to the voxelized ligands that will be used to train the captioning network. It has two CNNs: the encoder E1 and the decoder D1. Q(z|L) is the distribution of the latent variable z given the ligand shape L, N(0,I) is a standard normal Gaussian distribution and KL is the KLDivergence. The captioning network is composed of an encoder (E2), which is a convolutional neural network and a decoder (D2), which is an LSTM with an input state (xt) and a hidden state (ht), at time t. The encoder takes as input the noisy ligand shape generated by the VAE (Ł) and compresses it into a vector that is given as input to D2 with the SMILES vector. D2 outputs a sequence of letters or SMILES encoding a compound.

4.2.2.1 Data collection

To construct the database to train the captioning network I used an in-house database, already pre-processed for docking that contains molecular structures extracted from MolPort. I did a filtering to get those SMILES containing only the following symbols: C, c, N, n, S, s, P, O, o, B, F ,I, Cl, [nH], Br, 1, 2, 3, 4, 5, 6, # ,=, -, (,). Having the low energy three-dimensional structure of the ligands allowed me to reduce the data preparation time during training. In the end I had 5,600,000 compounds stored as sdf that were processed and stored in an HDF5 binary data format using h5py and rdkit scripts, where the keys are the SMILES and the values are the 3D coordinates for each compound.

The input to the captioning network are the grid of three-dimensional coordinates of each ligand, the length of the SMILES and the SMILES. The ligands were rotated before being voxelized which was performed with libmolgrid, as it was already described. [125]

4.2.2.2 Training setup

The VAE architecture was kept as described by Skalic et.al. [180] The VAE objective function is show on Equation 4.3:

$$VAEloss = CE^{VAE}(E, Dec) + L_{KL}(E)\beta \tag{4.3}$$

Where E is the CNN encoder and D is the CNN decoder. CE^{VAE} is the cross entropy loss measuring the reconstruction of the ligand and L_{KL} is the KLDivergence loss to ensure the latent space distribution found by the Encoder is similar to a standard normal Gaussian distribution. The parameter β ranges from 0 to 1 and are values from a sigmoid function $S(x) = \frac{1}{1+e^{-x}}$, were x are numbers between -10 and 10. For the LSTM, both the encoder and decoder doing the captioning, have as objective function the cross entropy loss between the class probabilities that are the output of the Decoder and the actual SMILES symbol indexes.

The VAE was optimized with the Adam optimizer (α= 5.10^{-4}) and the learning rate was reduced by half every 10,000 iterations. I trained the VAE for 5 epochs, until I observed the Pearson Correlation per channel between generated ligand shapes and real ligand shapes stopped improving. Afterwards, the Encoder and Decoder doing the captioning, were trained for 128 epochs until I observed the cross entropy loss stopped improving. They were optimized with the Adam optimizer (α = 1.10^{-3}) and the learning rate was reduced by half if the epoch number was divisible by 2 and the iteration number was divisible by 21000.

4.2.3 Filtering and docking of generated compounds by synthetic feasibility

I decided to use synthetic feasibility, as the criteria to filter the generated compounds. The Synthetic Accessibility (SA) score [181], Retrosynthetic Accessibility (RA)score [182] and the Synthetic Complexity (SC)Score [183] score were calculated for each compound. The SA score ranges from 1 (easy to synthesize) to 10 (very difficult to synthesize) and combines fragment contribution and a complexity penalty. If a given molecule has fragments commonly present in PubChem and not many complex features (large rings, non-standard ring fusions, stereo complexity and large molecule size) it will get a score close to 1. The SCScore ranges from 1 (easy to synthesize) to 5 (difficult to synthesize). It uses reaction knowledge that gives high scores to compounds that would require a lot of reaction steps to be synthesized. The RAScore is a probability that serves as a fast first approach to determine if a synthetic route can be identified for a molecular compound by the CASP tool AiZynthFinder. [184] AiZynthFinder tries to find a synthetic route for a compound by recursively breaking it down into purchasable precursors. AiZynthFinder is very time consuming and can take up to 2 minutes per molecule which is why the RAScore was implemented. I could not find an agreement as to which value to set up as a threshold to keep or not the generated ligands. So I calculated these scores for 10000 generated compounds for the

sodium-dependent serotonin transporter (SC6A4_HUMAN), on my test set, and observed the correlation between scores to select 3.5 as the threshold for SA score and SCScore and 0.7 for the RAscore. Finally, only the compounds that have good synthetic feasibility scores were analyzed with AiZynthFinder and those with the shortest predicted synthetic path were selected to be pre-processed and docked with smina, as described before.

4.2.4 Models analysis

The first thing I wanted to evaluate was how diverse the generated shapes were and whether the captioning network that was trained with a different dataset from the shape generating network, was able to decode the generated shapes efficiently. To evaluate this, I selected two proteins from the evaluation set: the progesterone receptor (P06401, PRGR_HUMAN) and the carbonic anhydrase 2 (P00918, CAH2_HUMAN); and two proteins from the test set: the sodium-dependent serotonin transporter (P31645, SC6A4_HUMAN) and the sigma non-opioid intracellular receptor 1 (Q99720, SGMR1_HUMAN). These proteins had more than 900 binders or binders in the binding database. The docked poses of the binders were voxelized and given as input to the captioning network. Then the protein was also voxelized at the ligand center and given to the shape generation network with a vector z (randomly sampled from a standard normal distribution) as input, to generate complementary shapes. The ligand shapes were decoded into 10,000 SMILES by the captioning network.

The generated shapes were pairwise compared with the Pearson correlation per channel. I randomly sampled as many generated shapes and SMILES as the number of binders I had for each protein. I added all the values per channel for the generated shapes and the voxelized binders and compared the distribution of these values. To study the captioning network I calculated the hydrogen-bond donors, hydrogen-bond acceptors, number of aromatic rings, number of aliphatic rings, number of rotatable bonds, number of halogens, molecular weight, topological polar surface area (TPSA) and the partition coefficient of a molecule between aqueous and lipophilic phases (LogP), for generated compounds and binders and compared their distributions. In addition I also compared the distribution of the number of characters (or lengths) of generated SMILES with the distribution of the lengths of the binders.

4.3 Results

It is possible to generate 10,000 ligand shapes for a protein pocket in less than 4 minutes. Afterwards 25,000 SMILES can be decoded in less than 3 minutes and from these generated SMILES nearly 40% will be chemically correct.

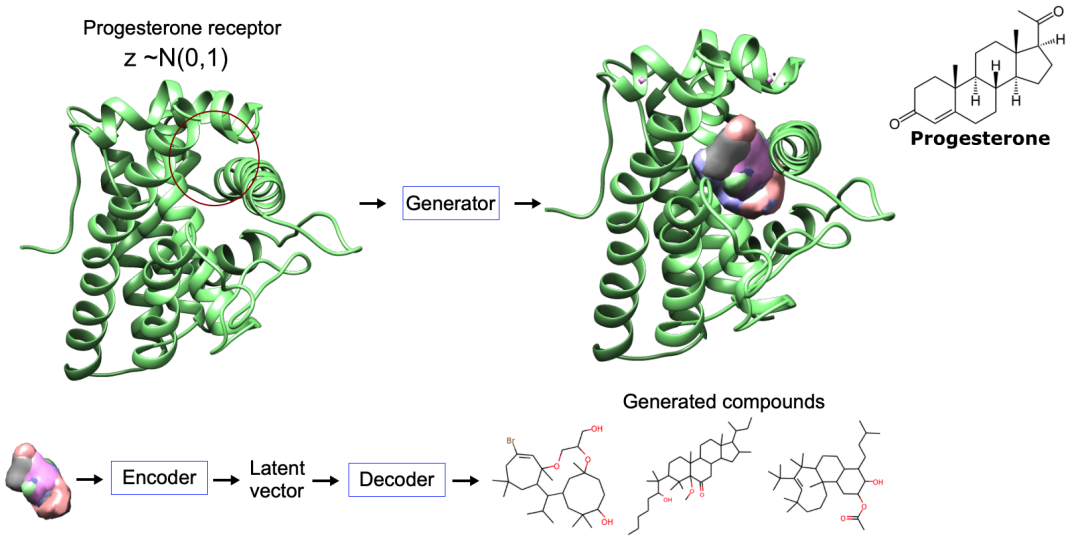

Figure 4.6: **Example of generated shape and compounds for the progesterone receptor.** The generator on the shape generation network takes as input the voxelized progesterone binding site and outputs a ligand shape. Then the encoder takes this ligand shape and encodes it into a latent vector that is the input to the decoder which will generate strings of SMILES.

Pairwise comparisons of the generated ligand shapes for each protein indicate that the generated shapes are very similar to each other with pearson correlation per channel average values ranging between 0.95-0.98, with most of the variability in the least populated channels encoding halogens, S, and P. In addition, a narrow distribution of the channel sums, over all the channels shown in the boxplots on Figure 4.7 and on Figure 4.8 suggests the generated shapes might not be very different from each other. Overall, larger density sums per channel on the generated shapes as compared to binders shapes, might indicate the shapes of generated ligands are bigger than the ligand shapes of voxelized binders. In particular for the channels with aliphatic hydrophobic atom densities. Generated shapes are most likely covering the entire protein pocket and contain in a single shape, all the chemical properties complementary to the protein pocket. In this case, it does not make sense to expect the generator to produce different shapes with each randomly sampled z, since the complementary chemical properties of the pocket should not change. The diversity of ligands is obtained in the captioning step since different atoms can be associated to the same property. No difference was observed between the generated shapes for the proteins in the evaluation and the test sets which suggests the shape generation network performs well for protein with binding sites not similar to those observed during training. One interesting difference between generated compounds and binders is observed in the channel with sulfur. This atom underrepresented in the trained data is less likely to be present in generated compounds, the same could be the case for P and metals. However, compounds could be quickly generated until a desired minimum of S, P or metals are present on the generated chemical structures. The small densities on aromatic channels are interesting and unexpected to see.

Afterwards the distribution of the molecular properties of the generated compounds

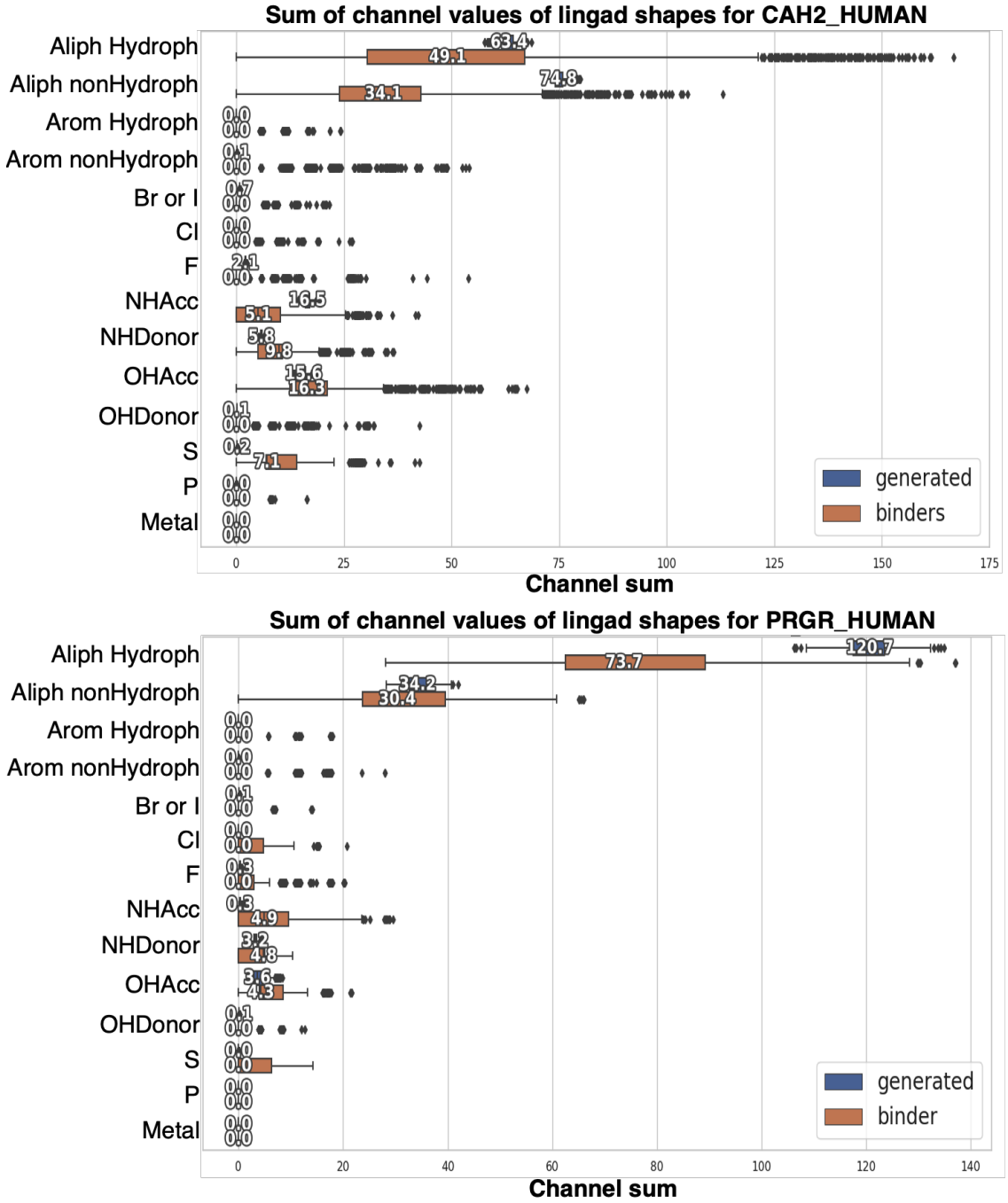

Figure 4.7: **Distribution of the summation of each channel values on the generated molecules.** These box plots depict the distribution of the channel values sum on the x axis and the channel descriptor on the y axis, for the two proteins in the evaluation set.

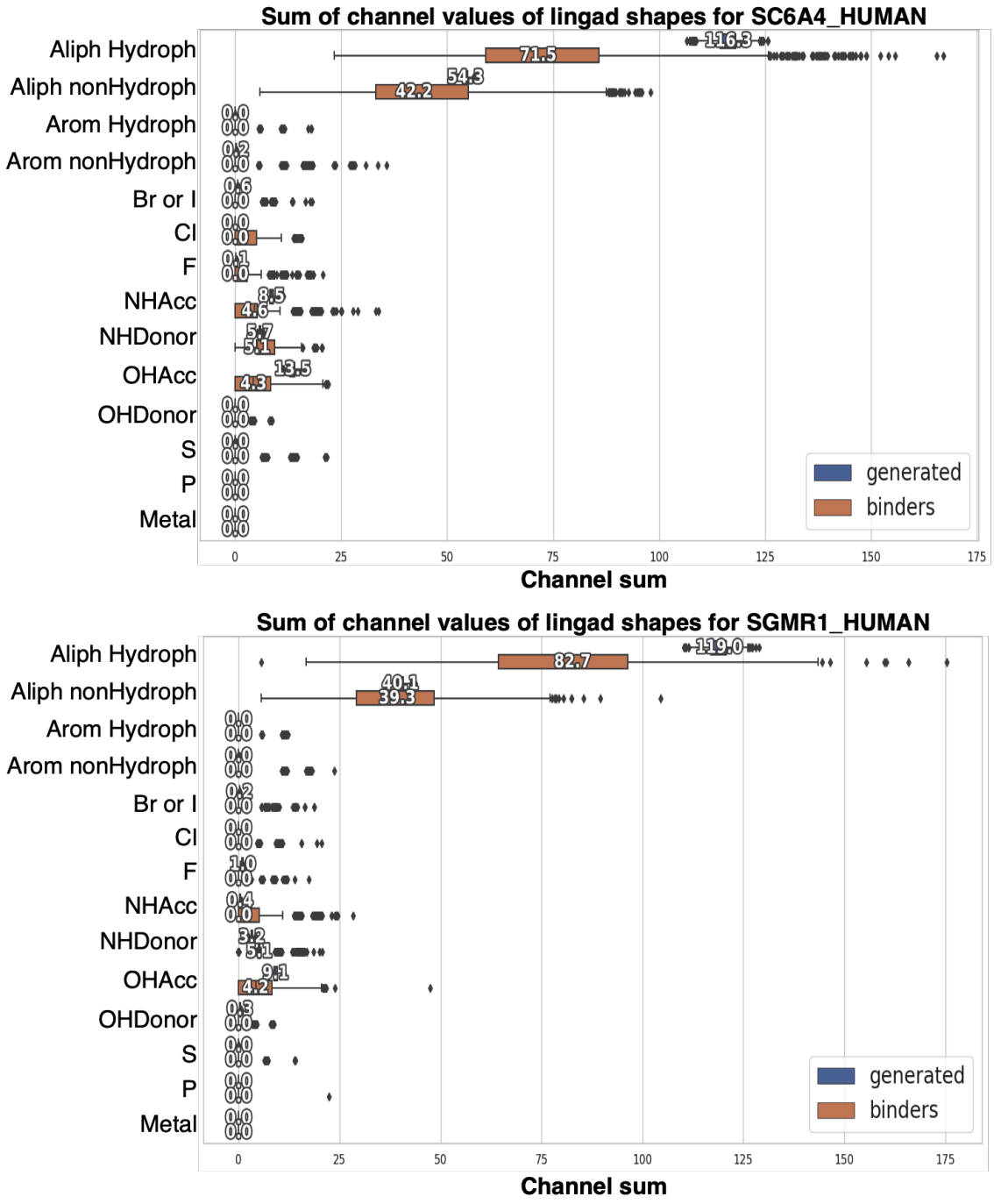

Figure 4.8: **Distribution of the summation of each channel values on the generated molecules.** These box plots depict the distribution of the channel values sum on the x axis and the channel descriptor on the y axis, for the two proteins in the test set.

and binders was compared. I observed that generated compounds have on average more aliphatic rings, halogens, hydrogen bond acceptors, hydrogen bond donors and rotatable bonds. Which again suggests that these compounds might be bigger than the binders. The most striking difference between generated compounds and binders is the number of aromatic rings which are between 0 to 1 for generated compounds and 2 to 4 for binders. However, aromatic compounds are still generated and since thousands of molecules can be produced in a matter of seconds a rule can be applied on the generation step to produce new molecules until N molecules with 2,3, or 4 aromatic rings are generated. A possible explanation for this behavior is that the network stayed in a local minima where it is more likely to get a good score when it generates rings in uppercase, which encodes aliphatic toms and rings in SMILES representation, instead of switching to lowercase letter which encode aromatic rings in the canonical SMILES representation used to train the model.

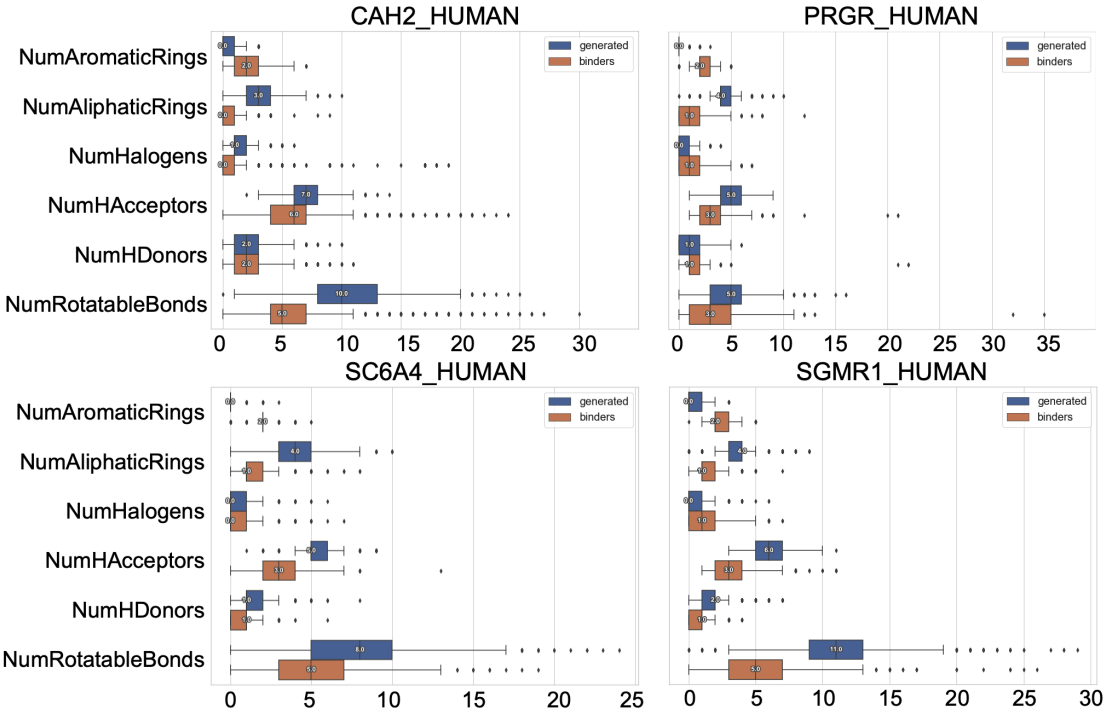

Figure 4.9: **Distribution of ring, bond and atoms counts.** These boxplots show the distribution of the number of aliphatic and aromatics rings, hydrogen bond donors and acceptors as well as the number of halogens and rotatable bonds on known binders and generated compounds.

It is surprising to see that the estimated LogP of the generated compounds follows a similar distribution to the estimated LogP of the known binders (Figure 4.10). The higher molecular weights between 500-600 Da also suggest that shapes and compounds are generated to cover the entire voxelized protein pocket and suggests most generated molecules will be larger than the majority of known drug-like compounds. The TPSA values for most of the generated compounds are below 140 A^2 which suggests that generated compounds could still be able to permeate cell membranes. It is also interesting to see that the TPSA values for the generated compounds of proteins in the evaluation set have a narrow and

similar distribution to those of the known binders for CAH2_HUMAN and PRGR_HUMAN, less so for SGMR1_HUMAN and SC6A4_HUMAN.

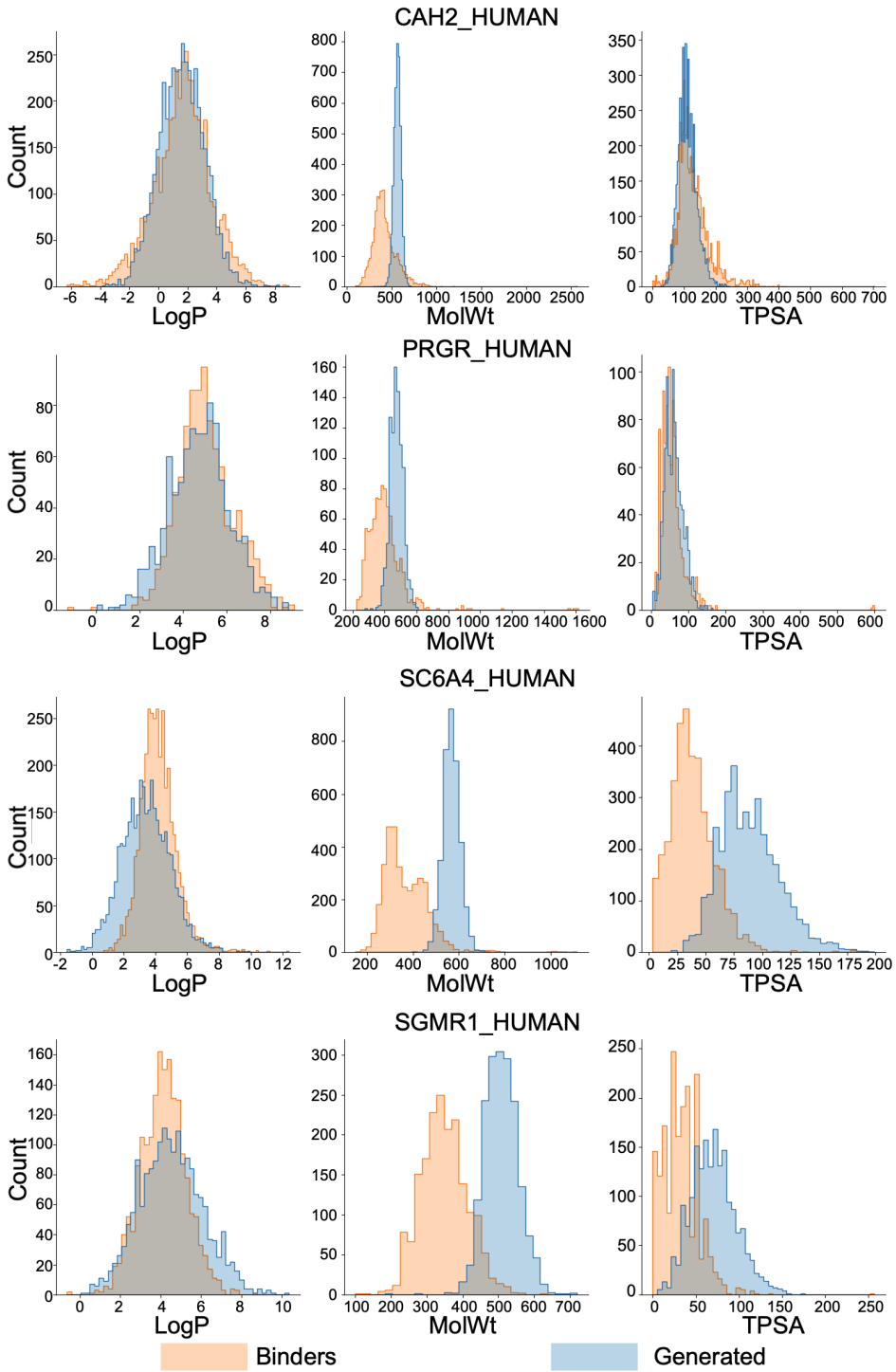

Figure 4.10: **Distribution of LogP, TPSA and MW of generate compounds and binders.** This figure compares the distribution of the LogP, TPSA and MW of the generated compounds and binders for each protein.

When I compared the distribution of the lengths of the generated canonical SMILES to the length of the canonical SMILES of known binders I observed that generated compounds have SMILES with more symbols. Which is not surprising considering the generated molecules are bigger. On average, the SMILES have lengths between 60-70 symbols. It is important to point out that the maximum number of symbols we allowed each SMILES string to have during training is 72. This is the length of the captioning embedding. During early stages of training I observed the generated SMILES had between 20 and 30 characters but the network optimized towards SMILES with the maximum number of symbols. This behavior raises the question of whether or not the models are capable of generating molecules that will adapt to protein pockets with different sizes.

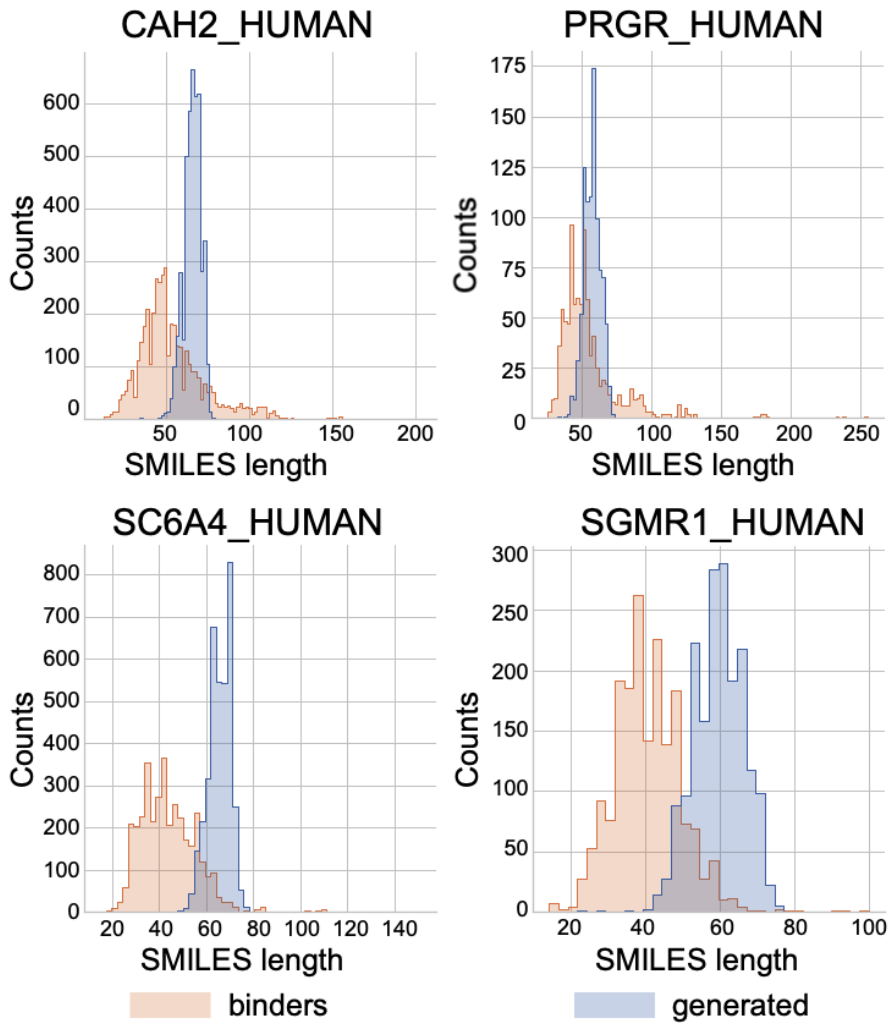

Figure 4.11: **Distribution of SMILES lengths of generated compounds and binders.** This figure compares the distribution of the SMILES length of the generated compounds and binders for each protein.

I observed the generated compounds had SAscores between 4.5 and 5.4. This score suggests that not many of the fragments from generated compounds are present in the

fragments found on PubChem. The generated compounds might be penalized due to their higher molecular size. The SCscore for CAH2_HUMAN and SC6A4_HUMAN indicates the generated compounds would require many reaction steps to be synthesized. From their RAScores the median probability to find a predicted retrosynthetic route by AiZynthFinder for generated compounds is 40%. The generated compounds for PRGR_HUMAN and SGMR1_HUMAN have both SCscores and RAscores suggesting it will be difficult to obtain compounds with a predicted retrosynthetic route. After applying the synthetic feasibility thresholds (SAscore, SCScore, RAscore) on the 10,000 compounds generated for each protein 2 % of the generated compounds for CAH2_HUMAN, 0.8 % of the generated compounds for PRGR_HUMAN, 20.3 % of SC6A4_HUMAN and 13 % of SGMR1_HUMAN generated compounds were selected. Finally, AiZynthFinder was able to propose a retrosynthetic route for 101 compounds for CAH2_HUMAN, 16 compounds for PRGR_HUMAN, 1,852 for SC6A4_HUMAN and 241 for SGMR1_HUMAN. For these compounds I used rdkit's quantitative estimation of drug likeness (QED) as the last filter to select generated compounds to be docked. QED evaluates the number of favorable drug like properties the compounds have and ranges from 0 (all properties unfavourable) to one (all properties favourable). I selected only those compounds with QED scores higher than 0.5.

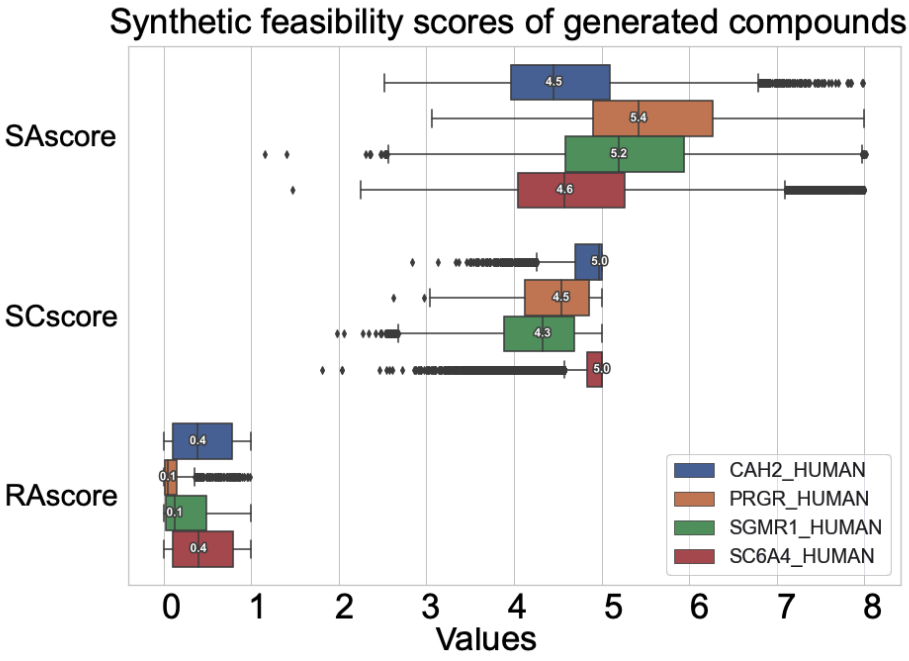

Figure 4.12: **Distribution of the synthetic feasibility scores chose to filter the generated compounds for each protein.** This figure compares the three synthetic feasibility scores for all generated compounds. The scores are: the SA score (1 easy to synthesize, 10 very difficult to synthesize), the SCScor (1 easy to synthesize, 5 difficult to synthesize) and the RAScore (probability to get a synthetic route with AiZynthFinder).

4.3.1 Application of the models on CAH2_HUMAN, SC6A4_HUMAN and $\alpha_4\alpha_5\beta_2$ nAChRs

The compounds that passed the synthetic feasibility filters and had the best drug-likeness scores were docked. On Figure 4.13 two of the generated compounds with the best docking scores are shown. I used the protein–ligand interaction profiler (PLIP) [185] to visualize the interactions between the docked generated ligand and the proteins. The generated molecule shown bound to CAH2_HUMAN, is not a pan assay interference compound, as predicted by the PAINS remover [186] and the ADME-Tox properties predicted by the SwissADME tool [178] indicate this compound is likely to be a P-glycoprotein substrate which could pump this compound out of the cell to be cleared from the body. In addition, this molecule has a high molecular weight (564 Da) and TPSA (147.04 Å^2) which makes it moderately soluble but with low gastrointestinal absorption and unlikely to cross the blood brain barrier. With its large size and many heavy hetero atoms the molecule can form two hydrogen bond between the oxygen of the amide in the pyridine and THR199, a hydrogen bond with the nitrogen of this amide and HIS94, another hydrogen bond between the nitrogen of the piperazine group and THR200 as well as a hydrogen bond between the pyrymidine and PRO201. I obtained 5 different predicted reaction paths proposed by AiZynthFinder [184] for this compound. All paths have similar steps to the reaction path shown on Figure 4.14, which includes a series of nucleophilic aromatic substitutions with Cl as leaving group and either piperazine or piperidine as nucleophile and the conversion of a nitrile into an amide.

The generated compound for SC6A4_HUMAN has a lower molecular weight and TPSA (529 Da, 73.71 Å^2), does not contain any functional groups that would suggest it could be a pan-assay interference compound. This molecule has moderate solubility, good gastrointestinal absorption and is likely to be able to cross the blood brain barrier. It forms salt bridge with the interaction between the tertiary amine and ASP98, a hydrogen bond between the carbonyl following the piperidine and THR497, a π stacking interaction between benzene and TYR176 and other weaker hydrophobic interactions. To synthesize this compound, AiZynthFinder predicted 11 different hypothetical synthetic paths like the one shown on Figure 4.15. The first steps of this hypothetical reaction path include the halogenation of the carbon atom bound to the carbonyl group and the reduction and deoxygenation of the carbonyl group on the N-Methylformamide, followed by the nucleophilic substitution of Br, then the conversion of a ketone into amine using reductive amination, the reaction between the amine and the carboxylic acid to form an amide and finally the nucleophilic aromatic substitution with Br as the leaving group and the piperidine as the nucleophile.

Finally I decided to see how the generative model would perform on the four pockets extracted from the trajectory of the $\alpha_4\alpha_5\beta_2$ nAChRs after the cavity analysis. The first cavity that I will call cavity 206, is located in the ECD of α_5 behind the loop-C and facing the ion pore. The selected generated compound is shown on Figure 4.16. This compound was not marked as a pan assay interference compound. Its ADME-Tox properties imply, this compound is very soluble, with a high gastrointestinal absorption and not likely to cross the blood brain barrier. It has 12 rotatable bonds which could increase the entropic penalty of the ligand and reduce its affinity to the receptor (if it could reach it). From the selected docking pose we can observe this compound forms a hydrogen bond between

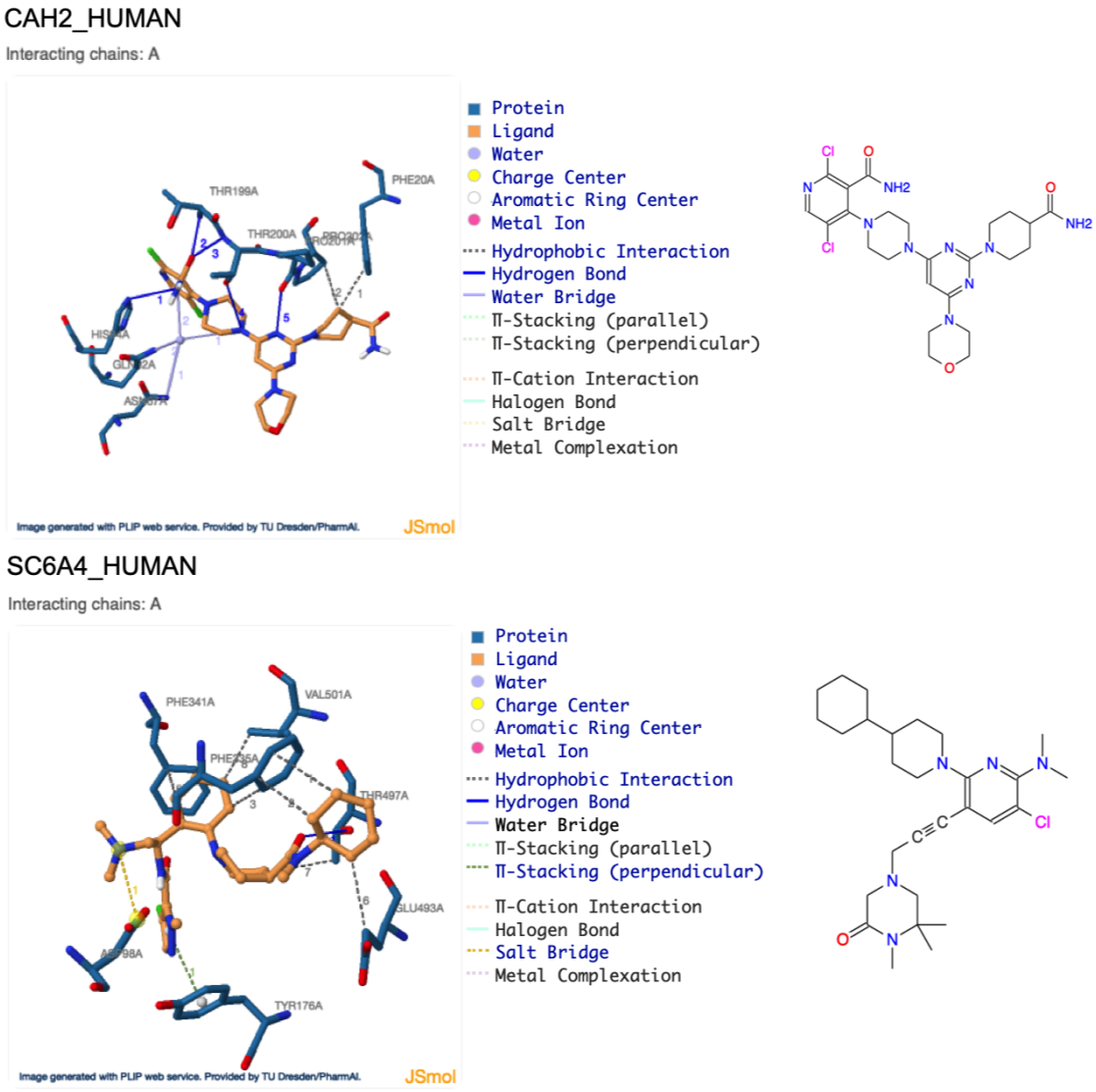

Figure 4.13: **Selected generated compounds with the best docking poses to CAH2_HUMAN and SC6A4_HUMAN.** The left side shows all the protein-ligand interactions obtained with PLIP and the types of interactions. On the right side the structure of the generated compound.

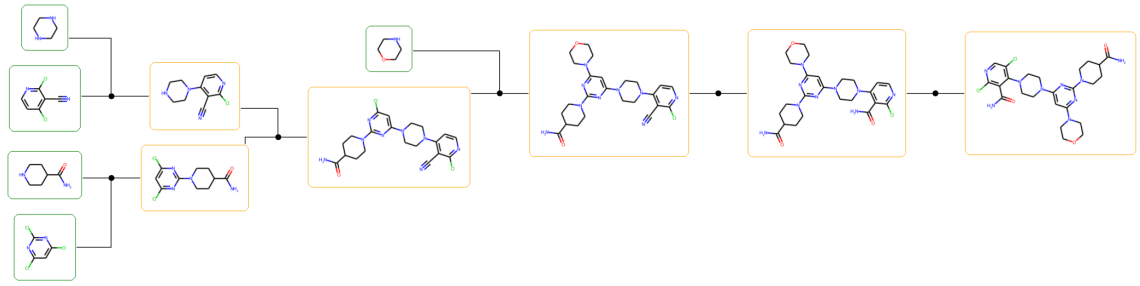

Figure 4.14: **Predicted synthetic route for the generated compound docked to CAH2_HUMAN.** This figure depicts a potential synthetic route, precursors and intermediate reaction steps to synthesize the generated compound for CAH2_HUMAN, as proposed by AiZynthFinder

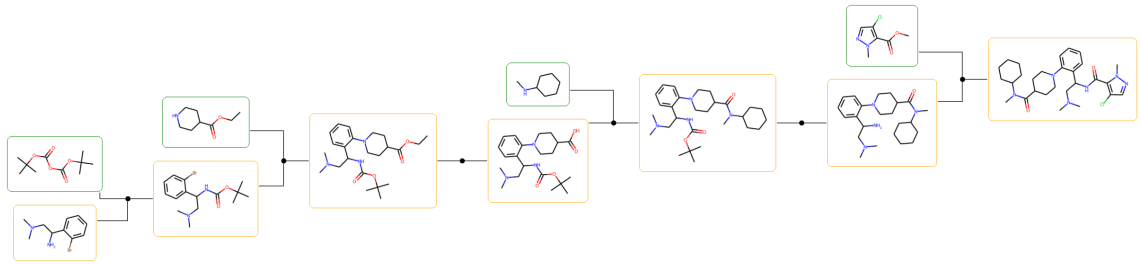

Figure 4.15: **Predicted Synthetic route for the generated compound docked to SC6A4_HUMAN.** This figure depicts a potential synthetic route, precursor and intermediate reaction steps to synthesize the generated compound for SC6A4_HUMAN , as proposed by AiZynthFinder

SER84 and the oxygen of the ether group, a hydrogen bond between the ASP85 and the oxygen of the carbonyl group on the tertiary amide, another hydrogen bond between the secondary amide and ILE92, a hydrogen bond between LEU94 and the nitrile, a hydrogen bond between LYS107 and the oxygen of the carbonyl bound to piperidine as well as π-stacking interaction between HIS104 and the thiophene. The 10 reaction paths proposed by AiZynthFinder look similar to what is shown on Figure 4.17. First the electrophilic Addition of nitromethane on the double bond, followed by the acid medium reduction of nitromethane into a primary amine, then the nucleophilic addition of piperidine to the electrophilic carbon of the carbonyl group and finally a last nucleophilic addition of the primary amine to the electrophilic carbon of the carbonyl group.

Cavity 174 is formed between the TM1 and TM2 helices of α_5 and the TM2 and TM3 heliced of β_2. The compound shown on Figure 4.16 bound to this cavity, was not classified to be a pan assay interference compound and has good predicted ADME-Tox properties since it is water soluble, has good gastrointestinal absorption and is likely to cross the blood brain barrier. It binds to the $\alpha_4\alpha_5\beta_2$ receptor by forming a hydrogen bond between the oxygen on the carboxylic group THR242 and SER246 as well as other hydrophobic interactions. Other compounds generated for this pocked were also forming π-stacking interactions with PHE245. From the hypothetical AiZynthFinder reaction path we observe that: two independent reactions can take place: first, the esterification of the carboxylic acid with methanol and the replacement of the -OH group in the primary alcohol by Br (a reaction that makes sense theoretically but with a very slow rate in reality). Then the reaction between cyclobutane, which has no double bond and is not nucleophilic, with bromine. This reaction will not take place unless head or light is applied to it. Finally the nucleophilic substitution of Cl Figure 4.18.

For the process of shape generation and shape captioning I decided to consider the cavities 104 and 165 as one since they merge in the activated frame of the $\alpha_4\alpha_5\beta_2$ receptor's trajectory. This cavity is located in the TMD of α_5. The chosen ligand bound to 104-165 is shown in Figure 4.16 is not a pan assay interference compound, is also soluble and has a good predicted gastrointestinal absorption but is unlikely to cross the blood brain barrier. It also has 12 rotatable bonds which could affect its binding to the receptor. It forms a hydrogen bond between the tetrahydropyran and ASN216, another hydrogen bond between the nitrogen in the thiocarbonyl and ILE217 as well as several hydrophobic interactions. 5 different reactions were proposed, with most of them having between 2 to 4 steps. The hypothetical reaction shown in Figure 4.19 involves the nucleophilic attack of the primary amine to the electrophilic Cl-carbon bond followed by a second nucleophilic substitution of the -OH group in the carboxylic group by the secondary amine.

4.4 Conclusion and discussion

The training of the bicycleGAN was difficult and took about a month on a single Nvidia GTX 1080TI GPU, once optimal parameters has been identified. We were able to stabilize the training by adding the void channel and by normalizing the tensor so the 15 channels add to one on each voxel. The similarity of the produced ligand shapes for a protein indicates that the generator does not learn to generate a wide distribution of different ligand

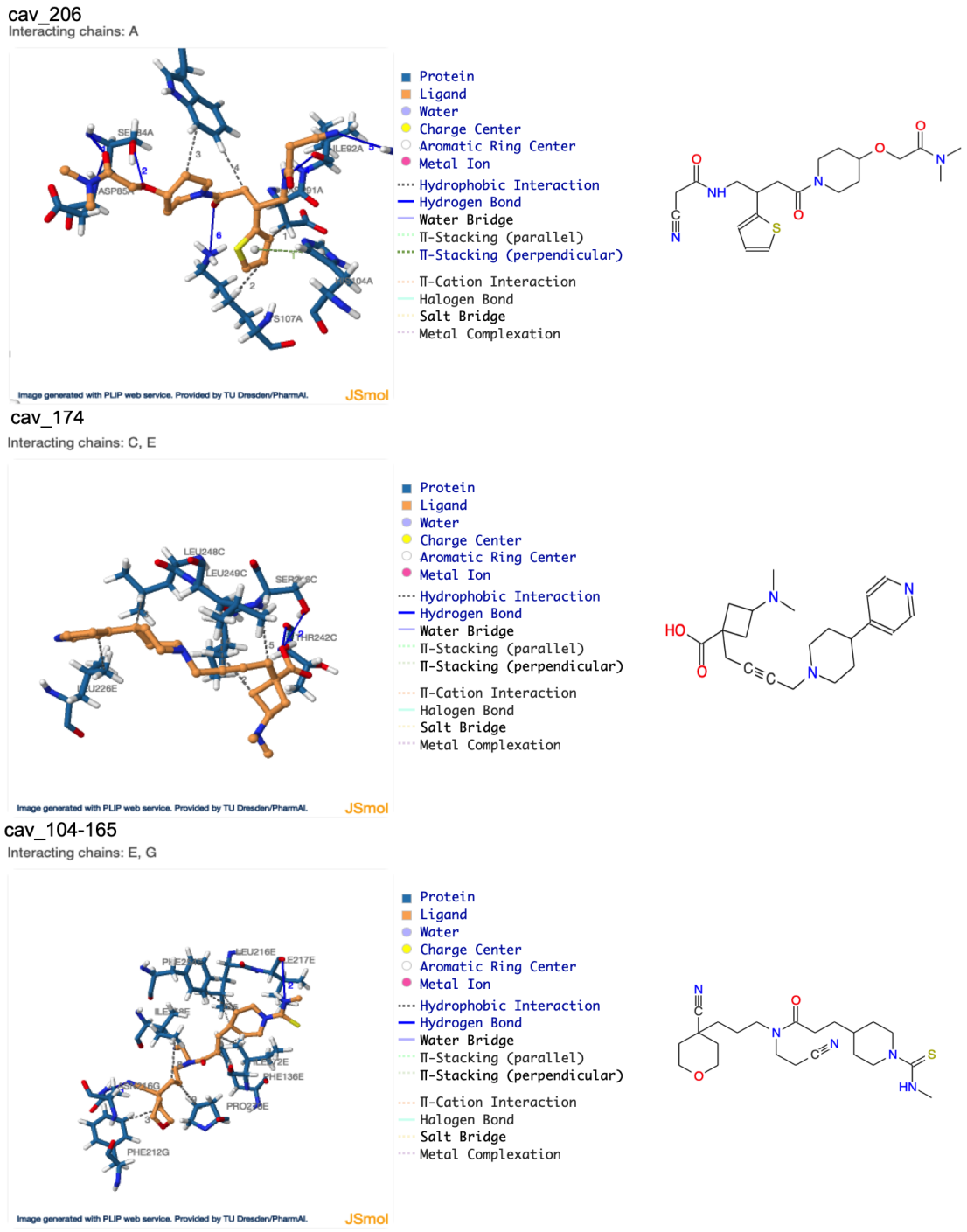

Figure 4.16: **Selected generated compounds with the best docking poses to the 3 pockets selected from the cavity analysis performed on the $\alpha_4\alpha_5\beta_2$ trajectory.** The left side shows all the protein-ligand interactions obtained with PLIP and the types of interactions. On the right side the structure of the generated compound.

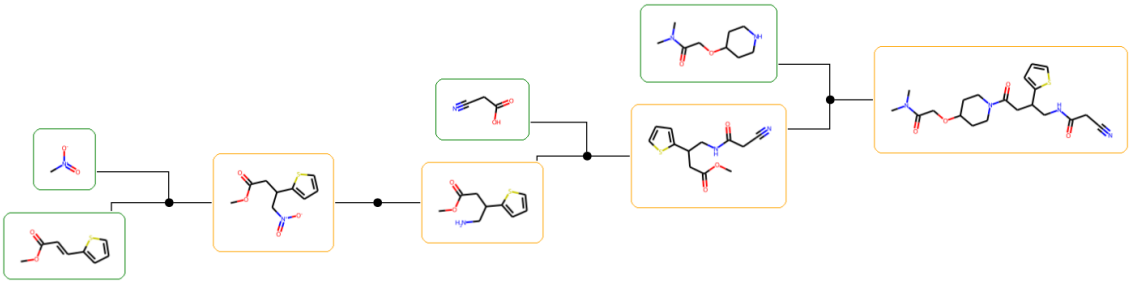

Figure 4.17: **Predicted synthetic route for the generated compound docked to cavity 206 located in the ECD of α_5.** This figure depicts a potential synthetic route, precursors and intermediate reaction steps to synthesize the generated compound as proposed by AiZynthFinder.

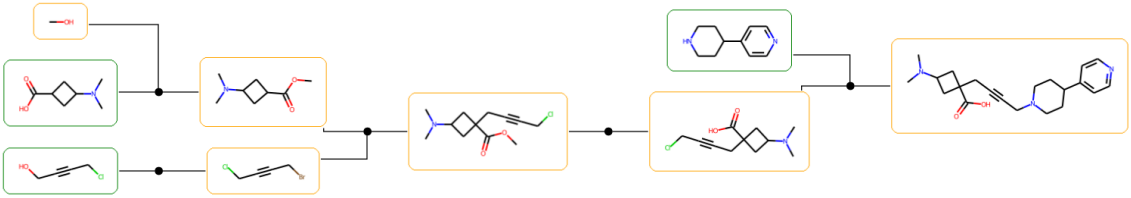

Figure 4.18: **Predicted synthetic route for the generated compound docked to cavity 174 between the TMD of α_5 and β_2.** This figure depicts a potential synthetic route, precursors and intermediate reaction steps to synthesize the generated compound as proposed by AiZynthFinder.

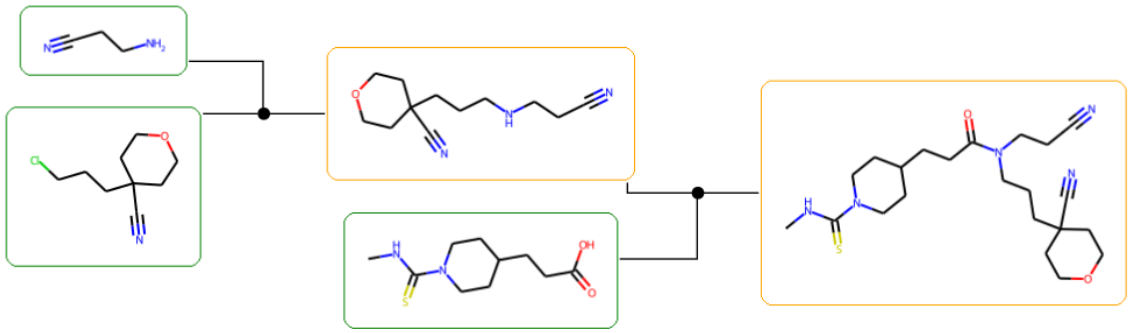

Figure 4.19: **Predicted synthetic route for the generated compound docked to cavities 104 and 164 in the TMD of α_5.** This figure depicts a potential synthetic route, precursors and intermediate reaction steps to synthesize the generated compound as proposed by AiZynthFinder.

shapes complementary to the protein pocket. What I observed from the summation over the values for each channel of generated ligands is that, compared to experimentally known binders, the generated shapes are likely to span and describe the entire protein pocket. The reason why the generated shapes are not significantly different could be that in spite of training the generator as part of the bicycleGAN, a model conceived to avoid mode collapse, the generator gets trapped in a local minimum during training and suffers from mode collapse, producing very similar ligand shapes for a protein pocket. Another explanation to why the generated shapes are not significantly different could be that chemical complementary is actually not that diverse, e.g. if there is a hydrogen bond donor it will most likely interact with a hydrogen bond acceptor or if there is a aromatic hydrophobic side chain it is most likely to interact with other hydrophobic groups. Thus, a ligand shape with complementary properties to a protein pocket might not have a large number of different possible combinations. This raises the question of whether a simpler generative model like a VAE or a VAE-GAN could have been trained to produce the ligand shapes instead of the complex bicycleGAN.

During the training of the captioning network the VAE used to add noise to voxelized ligands was impossible to train until I added a weight to the KLDivergence in the objective function. This weight takes values from a sigmoid distribution so that I could down-weight the contribution of the KLDivergence to the objective function and give priority to the reconstruction of the ligands at the beginning of the training. The captioning network was trained without difficulties and I was quite impressed with the complementarity between the generated ligand shapes and the atoms encoded into the SMILES by the decoder, given that the shape and the captioning networks were trained separately with different data sets. I observed that generated shapes had large channel sums over the aliphatic hydrophobic channel and the captioning network produced mostly SMILES with aliphatic rings instead of aromatic rings. It is also very interesting to observe the diverse functional groups that the captioning network is able to combine for a single ligand shape and still be able to produce correct SMILES. I was also happy to see the presence of less represented atoms in the training data, like halogens and sulfur, being integrated into the generated SMILES.

I noticed the network performs much better generating compounds for large and solvent exposed pockets, as it is the case for the carbonic anhydrase 2 (CAH2_HUMAN) or the cavity 206 on α_5 facing the ion pore. While I had more difficulties finding adequate ligands for the smaller cavities in the TMD of α_5. This is likely due to the network generating compounds with on average, no or only one aromatic ring per molecule. I did observe that the generated ligands for the cavities on α_5 were smaller than those generated for the other 4 proteins I studied.

Another issue I observed was the small number of generated compounds that were kept after the filtering based on synthetic feasibility scores. These numbers were even lower after using AiZynthFinder to select only those compounds with a predicted synthetic route and among these compounds still half of them had to be removed before docking because they did not have drug-like properties and presented a large number of rotatable bonds or high molecular weight. This filtering proved to be hihgly selective. Ideally, it would be more interesting to be able to train generative models to generate only molecules with desirable drug-like and ADME-Tox properties. Something that could be done with the model I have

trained and described in this chaper, is setting up a pipeline to keep generating compounds until N compounds with desired properties have been generated. This is feasible since the process of shape captioning is not time consuming. With this pipeline implemented I would re-generate new small, lipid soluble molecules for the $\alpha_4\alpha_5\beta_2$ nAChR that would be more likely to be able to cross the blood brain barrier by transmembrane diffusion.

Chapter 5

General conclusion and perspectives

During the development of this research project I learned and documented the methodology to produce the $\alpha_4\alpha_5\beta_2$ models in resting, activated and desensitized states. I presented an overview of the currently available experimental structures that can be used as templates to model other eukaryotic cation selective PLGICS, as well as their structural features agreement and disagreement. These models can now be used to study the functional motions of $\alpha_4\alpha_5\beta_2$ and can also be used for drug discovery projects. One aspect that could be improved on the models is the position of the loop-C on the α_5 subunit in desensitized state. This subunit could be remodeled but this time using as templates α subunits with a loop-C that closes completely to cap the orthosteric binding site.

Another important contribution is the first study of a transition path between three possible functional states of a PLGIC. The trajectory between these states provides a model to study the gating cycle of $\alpha_4\alpha_5\beta_2$ receptor and the possible synchrony between the blooming of the ECD and the twisting between the ECD and TMD. In the future it would be interesting to characterize in detail the electrostatic and hydrophobic interactions changing in the interface between the ECD and TMD during the gating cycle. This information is key to understanding how exactly the structural changes happening on each domain are communicated. Another interesting analysis that could be done with the models is, introducing the SNP on the structures of α_5 that has been associated with nicotine dependence and lung cancer. This analysis could allow us to see if the SNP has an impact on the receptor structure or on the gating cycle. Another contribution from this work is the prediction of the three cavities present on the α_5 subunit, only in the activated state. The cavities in the TMD are particularly interesting, since there is evidence that PAMS could bind in an allosteric binding site in the TMD.

I also collected and curated a data set to train a generative model to design novel chemical structures complementary to protein pockets. To tackle the issue of synthetic feasibility that these models present, I integrated several scores to filter generated molecules and to be able to have a synthetic route for all the compounds that can be used in further studies. The network produces chemically correct SMILES from ligand shapes, that integrate a diverse number of functional groups. One of the limitations on the applicability of the model is that most of the generated compounds don't have drug-like properties or don't

have desired ADME-Tox properties. I also observed the network produces less meaningful compounds for small hydrophobic pockets, as is the case for the pockets in the TMD of α_5. To solve this issue, it would be feasible to implement a pipeline that would filter the compounds that don't have desired properties and continue the process of ligand generation until the desired number of generated ligands has been reached. Once this pipeline has been set up specific properties can be sought and a new set of compounds could be generated for the allosteric cavity and the new cavities in the $\alpha_4\alpha_5\beta_2$ receptor.

Bibliography

[1] Angelo Keramidas and Joseph W Lynch. "An outline of desensitization in pentameric ligand-gated ion channel receptors". In: *Cellular and Molecular Life Sciences* 70.7 (2013), pp. 1241–1253. ISSN: 1420-9071. DOI: 10.1007/s00018-012-1133-z.

[2] Akos Nemecz et al. "Emerging Molecular Mechanisms of Signal Transduction in Pentameric Ligand-Gated Ion Channels". In: *Neuron* 90.3 (2016), pp. 452–470. ISSN: 0896-6273. DOI: https://doi.org/10.1016/j.neuron.2016.03.032.

[3] Pierre-Jean Corringer et al. "Structure and Pharmacology of Pentameric Receptor Channels: From Bacteria to Brain". In: *Structure* 20.6 (June 2012), pp. 941–956. ISSN: 0969-2126. DOI: 10.1016/j.str.2012.05.003.

[4] Vasyl Bondarenko et al. "Structures of highly flexible intracellular domain of human α7 nicotinic acetylcholine receptor". In: *Nature Communications* 13.1 (2022), p. 793. ISSN: 2041-1723. DOI: 10.1038/s41467-022-28400-x.

[5] Nelli Mnatsakanyan et al. "Functional Chimeras of GLIC Obtained by Adding the Intracellular Domain of Anion- and Cation-Conducting Cys-Loop Receptors." eng. In: *Biochemistry* 54.16 (Apr. 2015), pp. 2670–2682. ISSN: 1520-4995 (Electronic). DOI: 10.1021/acs.biochem.5b00203.

[6] Mackenzie J Thompson and John E Baenziger. "Structural basis for the modulation of pentameric ligand-gated ion channel function by lipids". In: *Biochimica et Biophysica Acta (BBA) - Biomembranes* 1862.9 (2020), p. 183304. ISSN: 0005-2736. DOI: https://doi.org/10.1016/j.bbamem.2020.183304.

[7] Rebecca J Howard. "Elephants in the Dark: Insights and Incongruities in Pentameric Ligand-gated Ion Channel Models". In: *Journal of Molecular Biology* 433.17 (2021), p. 167128. ISSN: 0022-2836. DOI: https://doi.org/10.1016/j.jmb.2021.167128.

[8] Colleen M Noviello et al. "Structure and gating mechanism of the α7 nicotinic acetylcholine receptor". In: *Cell* 184.8 (2021), pp. 2121–2134. ISSN: 0092-8674. DOI: https://doi.org/10.1016/j.cell.2021.02.049.

[9] Ludovic Sauguet et al. "Structural basis for ion permeation mechanism in pentameric ligand-gated ion channels". In: *The EMBO Journal* 32.5 (Mar. 2013), pp. 728–741. ISSN: 0261-4189. DOI: https://doi.org/10.1038/emboj.2013.17.

[10] Tommy S Tillman et al. "Functional Human α7 Nicotinic Acetylcholine Receptor (nAChR) Generated from Escherichia coli." eng. In: *The Journal of biological chemistry* 291.35 (Aug. 2016), pp. 18276–18282. ISSN: 1083-351X (Electronic). DOI: 10.1074/jbc.M116.729970.

[11] Ricarda J C Hilf and Raimund Dutzler. "X-ray structure of a prokaryotic pentameric ligand-gated ion channel". In: *Nature* 452.7185 (2008), pp. 375–379. ISSN: 1476-4687. DOI: 10.1038/nature06717.

[12] Ricarda J C Hilf and Raimund Dutzler. "Structure of a potentially open state of a proton-activated pentameric ligand-gated ion channel". In: *Nature* 457.7225 (2009), pp. 115–118. ISSN: 1476-4687. DOI: 10.1038/nature07461.

[13] Marie S Prevost et al. "A locally closed conformation of a bacterial pentameric proton-gated ion channel." eng. In: *Nature structural & molecular biology* 19.6 (May 2012), pp. 642–649. ISSN: 1545-9985 (Electronic). DOI: 10.1038/nsmb.2307.

[14] Ryan E Hibbs and Eric Gouaux. "Principles of activation and permeation in an anion-selective Cys-loop receptor". In: *Nature* 474.7349 (2011), pp. 54–60. ISSN: 1476-4687. DOI: 10.1038/nature10139.

[15] Moraga-Cid Gustavo et al. "Allosteric and hyperekplexic mutant phenotypes investigated on an α1 glycine receptor transmembrane structure". In: *Proceedings of the National Academy of Sciences* 112.9 (Mar. 2015), pp. 2865–2870. DOI: 10.1073/pnas.1417864112.

[16] Juan Du et al. "Glycine receptor mechanism elucidated by electron cryo-microscopy". In: *Nature* 526.7572 (2015), pp. 224–229. ISSN: 1476-4687. DOI: 10.1038/nature14853.

[17] Jie Yu et al. "Mechanism of gating and partial agonist action in the glycine receptor". In: *Cell* 184.4 (Feb. 2021), pp. 957–968. ISSN: 0092-8674. DOI: 10.1016/j.cell.2021.01.026.

[18] Xin Huang et al. "Crystal structure of human glycine receptor-α3 bound to antagonist strychnine". In: *Nature* 526.7572 (2015), pp. 277–280. ISSN: 1476-4687. DOI: 10.1038/nature14972.

[19] Xin Huang, Hao Chen, and Paul L Shaffer. "Crystal Structures of Human GlyRα3 Bound to Ivermectin". In: *Structure* 25.6 (2017), pp. 945–950. ISSN: 0969-2126. DOI: https://doi.org/10.1016/j.str.2017.04.007.

[20] Shaotong Zhu et al. "Structure of a human synaptic GABAA receptor". In: *Nature* 559.7712 (2018), pp. 67–72. ISSN: 1476-4687. DOI: 10.1038/s41586-018-0255-3.

[21] Tomasz Uchański et al. "Megabodies expand the nanobody toolkit for protein structure determination by single-particle cryo-EM". In: *Nature Methods* 18.1 (2021), pp. 60–68. ISSN: 1548-7105. DOI: 10.1038/s41592-020-01001-6.

[22] Ghérici Hassaine et al. "X-ray structure of the mouse serotonin 5-HT3 receptor". In: *Nature* 512.7514 (2014), pp. 276–281. ISSN: 1476-4687. DOI: 10.1038/nature13552.

[23] Sandip Basak et al. "Cryo-EM structure of 5-HT3A receptor in its resting conformation". In: *Nature Communications* 9.1 (2018), p. 514. ISSN: 2041-1723. DOI: 10.1038/s41467-018-02997-4.

[24] Sandip Basak et al. “Cryo-EM reveals two distinct serotonin-bound conformations of full-length 5-HT3A receptor”. In: *Nature* 563.7730 (2018), pp. 270–274. ISSN: 1476-4687. DOI: 10.1038/s41586-018-0660-7.

[25] Sandip Basak et al. “High-resolution structures of multiple 5-HT3AR-setron complexes reveal a novel mechanism of competitive inhibition”. In: *eLife* 9 (2020). Ed. by Cynthia M Czajkowski, Olga Boudker, and Cynthia M Czajkowski, e57870. ISSN: 2050-084X. DOI: 10.7554/eLife.57870.

[26] Anant Gharpure et al. “Agonist Selectivity and Ion Permeation in the $\alpha 3\beta 4$ Ganglionic Nicotinic Receptor”. In: *Neuron* 104.3 (2019), pp. 501–511. ISSN: 0896-6273. DOI: https://doi.org/10.1016/j.neuron.2019.07.030.

[27] Claudio L Morales-Perez, Colleen M Noviello, and Ryan E Hibbs. “X-ray structure of the human $\alpha 4\beta 2$ nicotinic receptor”. In: *Nature* 538.7625 (2016), pp. 411–415. ISSN: 1476-4687. DOI: 10.1038/nature19785.

[28] Somnath Mukherjee et al. “Synthetic antibodies against BRIL as universal fiducial marks for singleparticle cryoEM structure determination of membrane proteins”. In: *Nature Communications* 11.1 (2020), p. 1598. ISSN: 2041-1723. DOI: 10.1038/s41467-020-15363-0.

[29] Richard M Walsh et al. “Structural principles of distinct assemblies of the human $\alpha 4\beta 2$ nicotinic receptor”. In: *Nature* 557.7704 (2018), pp. 261–265. ISSN: 1476-4687. DOI: 10.1038/s41586-018-0081-7.

[30] Yue Zhao et al. “Structural basis of human $\alpha 7$ nicotinic acetylcholine receptor activation”. In: *Cell Research* 31.6 (2021), pp. 713–716. ISSN: 1748-7838. DOI: 10.1038/s41422-021-00509-6.

[31] Eleftherios Zarkadas et al. “Conformational transitions and ligand-binding to a muscle-type nicotinic acetylcholine receptor”. In: *Neuron* 110.8 (2022), pp. 1358–1370. ISSN: 0896-6273. DOI: https://doi.org/10.1016/j.neuron.2022.01.013.

[32] Md. Mahfuzur Rahman et al. “Structure of the Native Muscle-type Nicotinic Receptor and Inhibition by Snake Venom Toxins”. In: *Neuron* 106.6 (2020), pp. 952–962. ISSN: 0896-6273. DOI: https://doi.org/10.1016/j.neuron.2020.03.012.

[33] Jean-Pierre Changeux. “The nicotinic acetylcholine receptor: a typical ’allosteric machine’”. eng. In: *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 373.1749 (June 2018), p. 20170174. ISSN: 1471-2970. DOI: 10.1098/rstb.2017.0174.

[34] Hans Rollema, Daniel Bertrand, and Raymond S Hurst. “Nicotinic Agonists and Antagonists BT - Encyclopedia of Psychopharmacology”. In: ed. by Ian P Stolerman. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 887–899. ISBN: 978-3-540-68706-1. DOI: 10.1007/978-3-540-68706-1{_}304.

[35] Anant Gharpure, Colleen M Noviello, and Ryan E Hibbs. "Progress in nicotinic receptor structural biology". In: *Neuropharmacology* 171 (2020), p. 108086. ISSN: 0028-3908. DOI: https://doi.org/10.1016/j.neuropharm.2020.108086.

[36] Hugo Rubén Arias. "Localization of agonist and competitive antagonist binding sites on nicotinic acetylcholine receptors". In: *Neurochemistry International* 36.7 (2000), pp. 595–645. ISSN: 0197-0186. DOI: https://doi.org/10.1016/S0197-0186(99)00154-0.

[37] Marco Cecchini and Jean-Pierre Changeux. "The nicotinic acetylcholine receptor and its prokaryotic homologues: Structure, conformational transitions & allosteric modulation". In: *Neuropharmacology* 96 (2015), pp. 137–149. ISSN: 0028-3908. DOI: https://doi.org/10.1016/j.neuropharm.2014.12.006.

[38] Anthony Auerbach. "Gating of acetylcholine receptor channels: brownian motion across a broad transition state." eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 102.5 (Feb. 2005), pp. 1408–1412. ISSN: 0027-8424 (Print). DOI: 10.1073/pnas.0406787102.

[39] Marco Cecchini and Jean-Pierre Changeux. "Nicotinic receptors: From protein allostery to computational neuropharmacology". In: *Molecular Aspects of Medicine* 84 (2022), p. 101044. ISSN: 0098-2997. DOI: https://doi.org/10.1016/j.mam.2021.101044.

[40] Jacque Monod, Jeffries Wyman, and Jean-Pierre Changeux. "On the nature of allosteric transitions: A plausible model". In: *Journal of Molecular Biology* 12.1 (1965), pp. 88–118. ISSN: 0022-2836. DOI: https://doi.org/10.1016/S0022-2836(65)80285-6.

[41] M B Jackson. "Spontaneous openings of the acetylcholine receptor channel." eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 81.12 (June 1984), pp. 3901–3904. ISSN: 0027-8424 (Print). DOI: 10.1073/pnas.81.12.3901.

[42] Gareth T Young et al. "Potentiation of alpha7 nicotinic acetylcholine receptors via an allosteric transmembrane site." eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 105.38 (Sept. 2008), pp. 14686–14691. ISSN: 1091-6490 (Electronic). DOI: 10.1073/pnas.0804372105.

[43] Jeppe A Olsen et al. "Two Distinct Allosteric Binding Sites at α4β2 Nicotinic Acetylcholine Receptors Revealed by NS206 and NS9283 Give Unique Insights to Binding Activity-associated Linkage at Cys-loop Receptors *". In: *Journal of Biological Chemistry* 288.50 (Dec. 2013), pp. 35997–36006. ISSN: 0021-9258. DOI: 10.1074/jbc.M113.498618.

[44] Morten Grupe et al. "Targeting $\alpha4\beta2$ nicotinic acetylcholine receptors in central nervous system disorders: perspectives on positive allosteric modulation as a therapeutic approach." eng. In: *Basic & clinical pharmacology & toxicology* 116.3 (Mar. 2015), pp. 187–200. ISSN: 1742-7843 (Electronic). DOI: 10.1111/bcpt.12361.

[45] Nicole A Horenstein et al. "Critical Molecular Determinants of $\alpha7$ Nicotinic Acetylcholine Receptor Allosteric Activation: SEPARATION OF DIRECT ALLOSTERIC ACTIVATION AND POSITIVE ALLOSTERIC MODULATION." eng. In: *The Journal of biological chemistry* 291.10 (Mar. 2016), pp. 5049–5067. ISSN: 1083-351X (Electronic). DOI: 10.1074/jbc.M115.692392.

[46] Marta Quadri et al. "Macroscopic and Microscopic Activation of $\alpha7$ Nicotinic Acetylcholine Receptors by the Structurally Unrelated Allosteric Agonist-Positive Allosteric Modulators (ago-PAMs) B-973B and GAT107." eng. In: *Molecular pharmacology* 95.1 (Jan. 2019), pp. 43–61. ISSN: 1521-0111 (Electronic). DOI: 10.1124/mol.118.113340.

[47] Alican Gulsevin et al. "Allosteric Agonism of $\alpha7$ Nicotinic Acetylcholine Receptors: Receptor Modulation Outside the Orthosteric Site." eng. In: *Molecular pharmacology* 95.6 (June 2019), pp. 606–614. ISSN: 1521-0111 (Electronic). DOI: 10.1124/mol.119.115758.

[48] Jean-Pierre Changeux and Arthur Christopoulos. "Allosteric Modulation as a Unifying Mechanism for Receptor Function and Regulation". In: *Cell* 166.5 (2016), pp. 1084–1102. ISSN: 0092-8674. DOI: https://doi.org/10.1016/j.cell.2016.08.015.

[49] Xiu Liu. "Positive allosteric modulation of $\alpha4\beta2$ nicotinic acetylcholine receptors as a new approach to smoking reduction: evidence from a rat model of nicotine self-administration". In: *Psychopharmacology* 230.2 (2013), pp. 203–213. ISSN: 1432-2072. DOI: 10.1007/s00213-013-3145-2.

[50] Anne Catrien Baakman et al. "An anti-nicotinic cognitive challenge model using mecamylamine in comparison with the anti-muscarinic cognitive challenge using scopolamine." eng. In: *British journal of clinical pharmacology* 83.8 (Aug. 2017), pp. 1676–1687. ISSN: 1365-2125 (Electronic). DOI: 10.1111/bcp.13268.

[51] Hans Rollema, Daniel Bertrand, and Raymond S Hurst. "Nicotinic Agonists and Antagonists BT - Encyclopedia of Psychopharmacology". In: ed. by Ian P Stolerman and Lawrence H Price. Berlin, Heidelberg: Springer Berlin Heidelberg, 2015, pp. 1113–1131. ISBN: 978-3-642-36172-2. DOI: 10.1007/978-3-642-36172-2{_}304.

[52] Orlando Hung, Dolores McKeen, and Johannes Huitink. "Our love-hate relationship with succinylcholine: Is sugammadex any better?" In: *Canadian Journal of Anesthesia/Journal canadien d'anesthésie* 63.8 (2016), pp. 905–910. ISSN: 1496-8975. DOI: 10.1007/s12630-016-0664-4.

[53] Uwe Maskos. "The nicotinic receptor alpha5 coding polymorphism rs16969968 as a major target in disease: Functional dissection and remaining challenges." eng. In: *Journal of neurochemistry* 154.3 (Aug. 2020), pp. 241–250. ISSN: 1471-4159 (Electronic). DOI: 10.1111/jnc.14989.

[54] Laura Jean Bierut et al. "Variants in nicotinic receptors and risk for nicotine dependence". eng. In: *The American journal of psychiatry* 165.9 (Sept. 2008), pp. 1163–1171. ISSN: 1535-7228. DOI: 10.1176/appi.ajp.2008.07111711.

[55] Xingxu Yi et al. "The relationship between CHRNA5/A3/B4 gene cluster polymorphisms and lung cancer risk: An updated meta-analysis and systematic review." eng. In: *Medicine* 100.6 (Feb. 2021), e24355. ISSN: 1536-5964 (Electronic). DOI: 10.1097/MD.0000000000024355.

[56] Alexander Kuryatov, Wade Berrettini, and Jon Lindstrom. "Acetylcholine receptor (AChR) $\alpha 5$ subunit variant associated with risk for nicotine dependence and lung cancer reduces ($\alpha 4\beta 2$)$\alpha 5$ AChR function." eng. In: *Molecular pharmacology* 79.1 (Jan. 2011), pp. 119–125. ISSN: 1521-0111 (Electronic). DOI: 10.1124/mol.110.066357.

[57] Silke Frahm et al. "Aversion to Nicotine Is Regulated by the Balanced Activity of $\beta 4$ and $\alpha 5$ Nicotinic Receptor Subunits in the Medial Habenula". In: *Neuron* 70.3 (2011), pp. 522–535. ISSN: 0896-6273. DOI: https://doi.org/10.1016/j.neuron.2011.04.013.

[58] Marie S Prevost et al. "Concatemers to re-investigate the role of $\alpha 5$ in $\alpha 4\beta 2$ nicotinic receptors". In: *Cellular and Molecular Life Sciences* 78.3 (2021), pp. 1051–1064. ISSN: 1420-9071. DOI: 10.1007/s00018-020-03558-z.

[59] Miriam Sciaccaluga et al. "Crucial role of nicotinic $\alpha 5$ subunit variants for Ca2+ fluxes in ventral midbrain neurons." eng. In: *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* 29.8 (Aug. 2015), pp. 3389–3398. ISSN: 1530-6860 (Electronic). DOI: 10.1096/fj.14-268102.

[60] U Maskos et al. "Nicotine reinforcement and cognition restored by targeted expression of nicotinic receptors." eng. In: *Nature* 436.7047 (July 2005), pp. 103–107. ISSN: 1476-4687 (Electronic). DOI: 10.1038/nature03694.

[61] Sara Kalkhoran, Neal L Benowitz, and Nancy A Rigotti. "Prevention and Treatment of Tobacco Use: JACC Health Promotion Series." eng. In: *Journal of the American College of Cardiology* 72.9 (Aug. 2018), pp. 1030–1045. ISSN: 1558-3597 (Electronic). DOI: 10.1016/j.jacc.2018.06.036.

[62] Alessandro Marcon et al. "Trends in smoking initiation in Europe over 40 years: A retrospective cohort study." eng. In: *PloS one* 13.8 (2018), e0201881. ISSN: 1932-6203 (Electronic). DOI: 10.1371/journal.pone.0201881.

[63] Thomas Gredner et al. “Impact of tobacco control policies implementation on future lung cancer incidence in Europe: An international, population-based modeling study”. In: *The Lancet Regional Health - Europe* 4 (2021), p. 100074. ISSN: 2666-7762. DOI: https://doi.org/10.1016/j.lanepe.2021.100074.

[64] Antonina L Nazarova and Vsevolod Katritch. *It all clicks together: In silico drug discovery becoming mainstream.* eng. Apr. 2022. DOI: 10.1002/ctm2.766.

[65] José L Medina-Franco. “Grand Challenges of Computer-Aided Drug Design: The Road Ahead”. In: *Frontiers in Drug Discovery* 1 (2021). ISSN: 2674-0338. DOI: 10.3389/fddsv.2021.728551.

[66] Irini Doytchinova. *Drug Design-Past, Present, Future.* eng. Feb. 2022. DOI: 10.3390/molecules27051496.

[67] Mark A Johnson and Gerald M Maggiora. *Concepts and applications of molecular similarity.* Wiley, 1990. ISBN: 0471621757.

[68] Krzysztof Ginalski. “Comparative modeling for protein structure prediction”. In: *Current Opinion in Structural Biology* 16.2 (2006), pp. 172–177. ISSN: 0959-440X. DOI: https://doi.org/10.1016/j.sbi.2006.02.003.

[69] Tareq Hameduh et al. “Homology modeling in the time of collective and artificial intelligence”. In: *Computational and Structural Biotechnology Journal* 18 (2020), pp. 3494–3506. ISSN: 2001-0370. DOI: https://doi.org/10.1016/j.csbj.2020.11.007.

[70] The UniProt Consortium. “UniProt: the universal protein knowledgebase in 2021”. In: *Nucleic Acids Research* 49.D1 (Jan. 2021), pp. D480–D489. ISSN: 0305-1048. DOI: 10.1093/nar/gkaa1100.

[71] S F Altschul et al. “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.” eng. In: *Nucleic acids research* 25.17 (Sept. 1997), pp. 3389–3402. ISSN: 0305-1048 (Print). DOI: 10.1093/nar/25.17.3389.

[72] Helen M Berman et al. “The Protein Data Bank”. In: *Nucleic Acids Research* 28.1 (Jan. 2000), pp. 235–242. ISSN: 0305-1048. DOI: 10.1093/nar/28.1.235.

[73] Fabian Sievers et al. “Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega”. In: *Molecular Systems Biology* 7.1 (Jan. 2011), p. 539. ISSN: 1744-4292. DOI: https://doi.org/10.1038/msb.2011.75.

[74] Robert C Edgar. “MUSCLE: multiple sequence alignment with high accuracy and high throughput”. In: *Nucleic Acids Research* 32.5 (Mar. 2004), pp. 1792–1797. ISSN: 0305-1048. DOI: 10.1093/nar/gkh340.

[75] Cédric Notredame, Desmond G Higgins, and Jaap Heringa. “T-coffee: a novel method for fast and accurate multiple sequence alignment11Edited by J. Thornton”. In: *Journal of Molecular Biology* 302.1 (2000), pp. 205–217. ISSN: 0022-2836. DOI: https://doi.org/10.1006/jmbi.2000.4042.

[76] A Sali and T L Blundell. "Comparative protein modelling by satisfaction of spatial restraints." eng. In: *Journal of molecular biology* 234.3 (Dec. 1993), pp. 779–815. ISSN: 0022-2836 (Print). DOI: 10.1006/jmbi.1993.1626.

[77] Narayanan Eswar et al. "Comparative protein structure modeling using Modeller." eng. In: *Current protocols in bioinformatics* Chapter 5 (Oct. 2006), Unit–5.6. ISSN: 1934-340X (Electronic). DOI: 10.1002/0471250953.bi0506s15.

[78] R A Laskowski et al. "PROCHECK: a program to check the stereochemical quality of protein structures". In: *Journal of Applied Crystallography* 26.2 (Apr. 1993), pp. 283–291. DOI: 10.1107/S0021889892009944.

[79] Min-yi Shen and Andrej Sali. "Statistical potential for assessment and prediction of protein structures". In: *Protein science* 15.11 (2006), pp. 2507–2524. ISSN: 0961-8368.

[80] Ian W Davis et al. "MOLPROBITY: structure validation and all-atom contact analysis for nucleic acids and their complexes". In: *Nucleic acids research* 32.suppl_2 (2004), W615–W619. ISSN: 0305-1048.

[81] Pascal Benkert, Silvio C E Tosatto, and Dietmar Schomburg. "QMEAN: A comprehensive scoring function for model quality assessment." eng. In: *Proteins* 71.1 (Apr. 2008), pp. 261–277. ISSN: 1097-0134 (Electronic). DOI: 10.1002/prot.21715.

[82] C Chothia and A M Lesk. "The relation between the divergence of sequence and structure in proteins." eng. In: *The EMBO journal* 5.4 (Apr. 1986), pp. 823–826. ISSN: 0261-4189 (Print). DOI: 10.1002/j.1460-2075.1986.tb04288.x.

[83] Yazan Haddad, Vojtech Adam, and Zbynek Heger. "Ten quick tips for homology modeling of high-resolution protein 3D structures". In: *PLoS computational biology* 16.4 (2020), e1007449. ISSN: 1553-734X.

[84] Martin Karplus and J Andrew McCammon. "Molecular dynamics simulations of biomolecules". In: *Nature Structural Biology* 9.9 (2002), pp. 646–652. ISSN: 1545-9985. DOI: 10.1038/nsb0902-646.

[85] Scott A Hollingsworth and Ron O Dror. "Molecular Dynamics Simulation for All". In: *Neuron* 99.6 (2018), pp. 1129–1143. ISSN: 0896-6273. DOI: https://doi.org/10.1016/j.neuron.2018.08.011.

[86] Pedro E M Lopes, Olgun Guvench, and Alexander D MacKerell. "Current Status of Protein Force Fields for Molecular Dynamics Simulations BT - Molecular Modeling of Proteins". In: ed. by Andreas Kukol. New York, NY: Springer New York, 2015, pp. 47–71. ISBN: 978-1-4939-1465-4. DOI: 10.1007/978-1-4939-1465-4{_}3.

[87] Erik Lindahl. "Molecular dynamics simulations." eng. In: *Methods in molecular biology (Clifton, N.J.)* 1215 (2015), pp. 3–26. ISSN: 1940-6029 (Electronic). DOI: 10.1007/978-1-4939-1465-4{_}1.

[88] Alexander D MacKerell, Michael Feig, and Charles L Brooks. "Improved Treatment of the Protein Backbone in Empirical Force Fields". In: *Journal of the American Chemical Society* 126.3 (Jan. 2004), pp. 698–699. ISSN: 0002-7863. DOI: 10.1021/ja036959e.

[89] Alexander D Mackerell Jr. "Empirical force fields for biological macromolecules: Overview and issues". In: *Journal of Computational Chemistry* 25.13 (Oct. 2004), pp. 1584–1604. ISSN: 0192-8651. DOI: https://doi.org/10.1002/jcc.20082.

[90] Jing Huang et al. "CHARMM36m: an improved force field for folded and intrinsically disordered proteins." eng. In: *Nature methods* 14.1 (Jan. 2017), pp. 71–73. ISSN: 1548-7105 (Electronic). DOI: 10.1038/nmeth.4067.

[91] K Vanommeslaeghe et al. "CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields". In: *Journal of Computational Chemistry* 31.4 (Mar. 2010), pp. 671–690. ISSN: 0192-8651. DOI: https://doi.org/10.1002/jcc.21367.

[92] Florian J Gisdon, Martin Culka, and G Matthias Ullmann. "PyCPR – a python-based implementation of the Conjugate Peak Refinement (CPR) algorithm for finding transition state structures". In: *Journal of Molecular Modeling* 22.10 (2016), p. 242. ISSN: 0948-5023. DOI: 10.1007/s00894-016-3116-8.

[93] Mallamace Francesco et al. "Energy landscape in protein folding and unfolding". In: *Proceedings of the National Academy of Sciences* 113.12 (Mar. 2016), pp. 3159–3163. DOI: 10.1073/pnas.1524864113.

[94] Frauke Gräter and Wenjin Li. "Transition path sampling with quantum/classical mechanics for reaction rates." eng. In: *Methods in molecular biology (Clifton, N.J.)* 1215 (2015), pp. 27–45. ISSN: 1940-6029 (Electronic). DOI: 10.1007/978-1-4939-1465-4{_}2.

[95] Stefan Fischer and Martin Karplus. "Conjugate peak refinement: an algorithm for finding reaction paths and accurate transition states in systems with many degrees of freedom". In: *Chemical Physics Letters* 194.3 (1992), pp. 252–261. ISSN: 0009-2614. DOI: https://doi.org/10.1016/0009-2614(92)85543-J.

[96] Jiyu Fan, Ailing Fu, and Le Zhang. "Progress in molecular docking". In: *Quantitative Biology* 7.2 (2019), pp. 83–89. ISSN: 2095-4697. DOI: 10.1007/s40484-019-0172-y.

[97] Garrett M Morris and Marguerita Lim-Wilby. "Molecular Docking BT - Molecular Modeling of Proteins". In: ed. by Andreas Kukol. Totowa, NJ: Humana Press, 2008, pp. 365–382. ISBN: 978-1-59745-177-2. DOI: 10.1007/978-1-59745-177-2{_}19.

[98] R D Taylor, P J Jewsbury, and J W Essex. "A review of protein-small molecule docking methods". In: *Journal of Computer-Aided Molecular Design* 16.3 (2002), pp. 151–166. ISSN: 1573-4951. DOI: 10.1023/A:1020155510718.

[99] David Ryan Koes, Matthew P Baumgartner, and Carlos J Camacho. "Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise". eng. In: *Journal of chemical information and modeling* 53.8 (Aug. 2013), pp. 1893–1904. ISSN: 1549-960X. DOI: 10.1021/ci300604z.

[100] Isabella A Guedes, Felipe S S Pereira, and Laurent E Dardenne. "Empirical Scoring Functions for Structure-Based Virtual Screening: Applications, Critical Aspects, and Challenges." eng. In: *Frontiers in pharmacology* 9 (2018), p. 1089. ISSN: 1663-9812 (Print). DOI: 10.3389/fphar.2018.01089.

[101] André Fischer et al. "Decision Making in Structure-Based Drug Discovery: Visual Inspection of Docking Results". In: *Journal of Medicinal Chemistry* 64.5 (Mar. 2021), pp. 2489–2500. ISSN: 0022-2623. DOI: 10.1021/acs.jmedchem.0c02227.

[102] Johann Gasteiger. "Chemistry in Times of Artificial Intelligence". In: *ChemPhysChem* 21.20 (Oct. 2020), pp. 2233–2242. ISSN: 1439-4235. DOI: https://doi.org/10.1002/cphc.202000518.

[103] Rishi R Gupta. "Application of Artificial Intelligence and Machine Learning in Drug Discovery." eng. In: *Methods in molecular biology (Clifton, N.J.)* 2390 (2022), pp. 113–124. ISSN: 1940-6029 (Electronic). DOI: 10.1007/978-1-0716-1787-8{_}4.

[104] JaroslawTI Unsupervised Learning in Drug Design from Self-Organization to Deep Chemistry Polanski. *Unsupervised Learning in Drug Design from Self-Organization to Deep Chemistry*. 2022. DOI: 10.3390/ijms23052797.

[105] Jie Shen et al. "Estimation of ADME Properties with Substructure Pattern Recognition". In: *Journal of Chemical Information and Modeling* 50.6 (June 2010), pp. 1034–1041. ISSN: 1549-9596. DOI: 10.1021/ci100104j.

[106] Vladimir Svetnik et al. "Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling". In: *Journal of Chemical Information and Computer Sciences* 43.6 (Nov. 2003), pp. 1947–1958. ISSN: 0095-2338. DOI: 10.1021/ci034160g.

[107] Robert P Sheridan et al. "Extreme Gradient Boosting as a Method for Quantitative Structure–Activity Relationships". In: *Journal of Chemical Information and Modeling* 56.12 (Dec. 2016), pp. 2353–2360. ISSN: 1549-9596. DOI: 10.1021/acs.jcim.6b00591.

[108] Yadi Zhou et al. "Exploring Tunable Hyperparameters for Deep Neural Networks with Industrial ADME Data Sets". In: *Journal of Chemical Information and Modeling* 59.3 (Mar. 2019), pp. 1005–1016. ISSN: 1549-9596. DOI: 10.1021/acs.jcim.8b00671.

[109] Jessica Vamathevan et al. "Applications of machine learning in drug discovery and development". In: *Nature Reviews Drug Discovery* 18.6 (2019), pp. 463–477. ISSN: 1474-1784. DOI: 10.1038/s41573-019-0024-5.

[110] Onat Kadioglu et al. “Identification of novel compounds against three targets of SARS CoV-2 coronavirus by combined virtual screening and supervised machine learning.” eng. In: *Computers in biology and medicine* 133 (June 2021), p. 104359. ISSN: 1879-0534 (Electronic). DOI: 10.1016/j.compbiomed.2021.104359.

[111] Keith T Butler et al. “Machine learning for molecular and materials science”. In: *Nature* 559.7715 (2018), pp. 547–555. ISSN: 1476-4687. DOI: 10.1038/s41586-018-0337-2.

[112] Matthew Ragoza et al. “Protein–Ligand Scoring with Convolutional Neural Networks”. In: *Journal of Chemical Information and Modeling* 57.4 (Apr. 2017), pp. 942–957. ISSN: 1549-9596. DOI: 10.1021/acs.jcim.6b00740.

[113] Raúl Rojas. *Neural networks: a systematic introduction*. Springer Science & Business Media, 2013. ISBN: 3642610684.

[114] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. “Deep learning”. In: *Nature* 521.7553 (2015), pp. 436–444. ISSN: 1476-4687. DOI: 10.1038/nature14539.

[115] Ian Goodfellow et al. “Generative Adversarial Nets”. In: *Advances in Neural Information Processing Systems*. Ed. by Z Ghahramani et al. Vol. 27. Curran Associates, Inc., 2014.

[116] Diederik P Kingma and Max Welling. “Auto-encoding variational bayes”. In: *arXiv preprint arXiv:1312.6114* (2013).

[117] Kamilya Smagulova and Alex Pappachen James. “Overview of Long Short-Term Memory Neural Networks BT - Deep Learning Classifiers with Memristive Networks: Theory and Applications”. In: ed. by Alex Pappachen James. Cham: Springer International Publishing, 2020, pp. 139–153. ISBN: 978-3-030-14524-8. DOI: 10.1007/978-3-030-14524-8{_}11.

[118] Torsten Hoffmann and Marcus Gastreich. “The next level in chemical space navigation: going far beyond enumerable compound libraries”. In: *Drug Discovery Today* 24.5 (2019), pp. 1148–1156. ISSN: 1359-6446. DOI: https://doi.org/10.1016/j.drudis.2019.02.013.

[119] Esben Jannik Bjerrum and Boris Sattarov. “Improving Chemical Autoencoder Latent Space and Molecular De novo Generation Diversity with Heteroencoders”. In: (June 2018). DOI: 10.3390/biom8040131.

[120] Oscar Méndez-Lucio et al. “De novo generation of hit-like molecules from gene expression signatures using artificial intelligence”. In: *Nature Communications* 11.1 (2020), p. 10. ISSN: 2041-1723. DOI: 10.1038/s41467-019-13807-w.

[121] Yibo Li, Liangren Zhang, and Zhenming Liu. “Multi-objective de novo drug design with conditional graph generative model”. In: *Journal of Cheminformatics* 10.1 (2018), p. 33. ISSN: 1758-2946. DOI: 10.1186/s13321-018-0287-6.

[122] Artur Kadurin et al. "druGAN: An Advanced Generative Adversarial Autoencoder Model for de Novo Generation of New Molecules with Desired Molecular Properties in Silico." eng. In: *Molecular pharmaceutics* 14.9 (Sept. 2017), pp. 3098–3104. ISSN: 1543-8392 (Electronic). DOI: 10.1021/acs.molpharmaceut.7b00346.

[123] Evgeny Putin et al. "Adversarial Threshold Neural Computer for Molecular de Novo Design." eng. In: *Molecular pharmaceutics* 15.10 (Oct. 2018), pp. 4386–4397. ISSN: 1543-8392 (Electronic). DOI: 10.1021/acs.molpharmaceut.7b01137.

[124] Daniel C Elton et al. "Deep learning for molecular design—a review of the state of the art". In: *Molecular Systems Design & Engineering* 4.4 (2019), pp. 828–849. DOI: 10.1039/C9ME00039A.

[125] Jocelyn Sunseri and David R Koes. "libmolgrid: Graphics Processing Unit Accelerated Molecular Gridding for Deep Learning Applications". In: *Journal of Chemical Information and Modeling* 60.3 (Mar. 2020), pp. 1079–1084. ISSN: 1549-9596. DOI: 10.1021/acs.jcim.9b01145.

[126] Miha Skalic et al. "From Target to Drug: Generative Modeling for the Multimodal Structure-Based Ligand Design". In: *Molecular Pharmaceutics* 16.10 (Oct. 2019), pp. 4282–4291. ISSN: 1543-8384. DOI: 10.1021/acs.molpharmaceut.9b00634.

[127] Jun-Yan Zhu et al. "Toward Multimodal Image-to-Image Translation". In: *CoRR* abs/1711.1 (2017).

[128] Lucie Polovinkin et al. "Conformational transitions of the serotonin 5-HT3 receptor". In: *Nature* 563.7730 (2018), pp. 275–279. ISSN: 1476-4687. DOI: 10.1038/s41586-018-0672-3.

[129] Sandip Basak et al. "Molecular mechanism of setron-mediated inhibition of full-length 5-HT3A receptor". In: *Nature Communications* 10.1 (2019), p. 3225. ISSN: 2041-1723. DOI: 10.1038/s41467-019-11142-8.

[130] Zhao Yue et al. "Structural basis of human α7 nicotinic acetylcholine receptor activation". In: *Cell Research* 31.6 (2021), pp. 713–716. ISSN: 1748-7838. DOI: 10.1038/s41422-021-00509-6.

[131] Yuqi Wang et al. "A crowdsourcing open platform for literature curation in UniProt". In: *PLOS Biology* 19.12 (Dec. 2021), e3001464.

[132] Alexey Drozdetskiy et al. "JPred4: a protein secondary structure prediction server". In: *Nucleic Acids Research* 43.W1 (July 2015), W389–W394. ISSN: 0305-1048. DOI: 10.1093/nar/gkv332.

[133] John Eargle, Dan Wright, and Zaida Luthey-Schulten. "Multiple Alignment of protein structures and sequences for VMD". In: *Bioinformatics* 22.4 (Feb. 2006), pp. 504–506. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bti825.

[134] David Eisenberg, Roland Lüthy, and James U B T Methods in Enzymology Bowie. "VERIFY3D: Assessment of protein models with three-dimensional profiles". In: *Macromolecular Crystallography Part B*. Vol. 277. Academic Press, 1997, pp. 396–404. ISBN: 0076-6879. DOI: https://doi.org/10.1016/S0076-6879(97)77022-8.

[135] Markus Wiederstein and Manfred J Sippl. "ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins". eng. In: *Nucleic acids research* 35.Web Server issue (July 2007), W407–W410. ISSN: 1362-4962. DOI: 10.1093/nar/gkm290.

[136] J.Michael Word et al. "Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation11Edited by J. Thornton". In: *Journal of Molecular Biology* 285.4 (1999), pp. 1735–1747. ISSN: 0022-2836. DOI: https://doi.org/10.1006/jmbi.1998.2401.

[137] Paul Bauer, Berk Hess, and Erik Lindahl. "GROMACS 2022 Manual". In: (Feb. 2022). DOI: 10.5281/ZENODO.6103568.

[138] Mark James Abraham et al. "GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers". In: *SoftwareX* 1-2 (2015), pp. 19–25. ISSN: 2352-7110. DOI: https://doi.org/10.1016/j.softx.2015.06.001.

[139] Sunhwan Jo, Taehoon Kim, and Wonpil Im. "Automated Builder and Database of Protein/Membrane Complexes for Molecular Dynamics Simulations". In: *PLOS ONE* 2.9 (Sept. 2007), e880.

[140] Sunhwan Jo et al. "CHARMM-GUI Membrane Builder for Mixed Bilayers and Its Application to Yeast Membranes". In: *Biophysical Journal* 97.1 (2009), pp. 50–58. ISSN: 0006-3495. DOI: https://doi.org/10.1016/j.bpj.2009.04.013.

[141] Emilia L Wu et al. "CHARMM-GUI Membrane Builder toward realistic biological membrane simulations". In: *Journal of Computational Chemistry* 35.27 (Oct. 2014), pp. 1997–2004. ISSN: 0192-8651. DOI: https://doi.org/10.1002/jcc.23702.

[142] Sunhwan Jo et al. "CHARMM-GUI: A web-based graphical user interface for CHARMM". In: *Journal of Computational Chemistry* 29.11 (Aug. 2008), pp. 1859–1865. ISSN: 0192-8651. DOI: https://doi.org/10.1002/jcc.20945.

[143] M P McCarthy and M A Moore. "Effects of lipids and detergents on the conformation of the nicotinic acetylcholine receptor from Torpedo californica." eng. In: *The Journal of biological chemistry* 267.11 (Apr. 1992), pp. 7655–7663. ISSN: 0021-9258 (Print).

[144] T M Fong and M G McNamee. "Correlation between acetylcholine receptor function and structural properties of membranes." eng. In: *Biochemistry* 25.4 (Feb. 1986), pp. 830–840. ISSN: 0006-2960 (Print). DOI: 10.1021/bi00352a015.

[145] S E Rankin et al. "The cholesterol dependence of activation and fast desensitization of the nicotinic acetylcholine receptor." eng. In: *Biophysical journal* 73.5 (Nov. 1997), pp. 2446–2455. ISSN: 0006-3495 (Print). DOI: 10.1016/S0006-3495(97)78273-0.

[146] Oliver S Smart et al. "HOLE: A program for the analysis of the pore dimensions of ion channel structural models". In: *Journal of Molecular Graphics* 14.6 (1996), pp. 354–360. ISSN: 0263-7855. DOI: https://doi.org/10.1016/S0263-7855(97)00009-X.

[147] Oliviero Carugo. "How large B-factors can be in protein crystal structures". In: *BMC Bioinformatics* 19.1 (2018), p. 61. ISSN: 1471-2105. DOI: 10.1186/s12859-018-2083-8.

[148] Stephan Lorenzen et al. "Conservation of cis prolyl bonds in proteins during evolution". In: *Proteins: Structure, Function, and Bioinformatics* 58.3 (Feb. 2005), pp. 589–595. ISSN: 0887-3585. DOI: https://doi.org/10.1002/prot.20342.

[149] Rilei Yu et al. "Molecular dynamics simulations of dihydro-β-erythroidine bound to the human $\alpha4\beta2$ nicotinic acetylcholine receptor." eng. In: *British journal of pharmacology* 176.15 (Aug. 2019), pp. 2750–2763. ISSN: 1476-5381 (Electronic). DOI: 10.1111/bph.14698.

[150] Fritz G Parak. "Proteins in action: the physics of structural fluctuations and conformational changes". In: *Current Opinion in Structural Biology* 13.5 (2003), pp. 552–557. ISSN: 0959-440X. DOI: https://doi.org/10.1016/j.sbi.2003.09.004.

[151] Reza Salari, Sruthi Murlidaran, and Grace Brannigan. "Pentameric Ligand-gated Ion Channels : Insights from Computation". eng. In: *Molecular simulation* 40.10-11 (Apr. 2014), pp. 821–829. ISSN: 0892-7022. DOI: 10.1080/08927022.2014.896462.

[152] Nicolas E Martin et al. "Un-gating and allosteric modulation of a pentameric ligand-gated ion channel captured by molecular dynamics". eng. In: *PLoS computational biology* 13.10 (Oct. 2017), e1005784–e1005784. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1005784.

[153] Letizia Chiodo et al. "A Structural Model of the Human $\alpha7$ Nicotinic Receptor in an Open Conformation". In: *PLOS ONE* 10.7 (July 2015), e0133011.

[154] Letizia Chiodo et al. "Closed-Locked and Apo-Resting State Structures of the Human $\alpha7$ Nicotinic Receptor: A Computational Study". In: *Journal of Chemical Information and Modeling* 58.11 (Nov. 2018), pp. 2278–2293. ISSN: 1549-9596. DOI: 10.1021/acs.jcim.8b00412.

[155] L Chiodo et al. "A possible desensitized state conformation of the human $\alpha7$ nicotinic receptor: A molecular dynamics study". In: *Biophysical Chemistry* 229 (2017), pp. 99–109. ISSN: 0301-4622. DOI: https://doi.org/10.1016/j.bpc.2017.06.010.

[156] A Sofia F Oliveira et al. "Identification of the Initial Steps in Signal Transduction in the $\alpha 4\beta 2$ Nicotinic Receptor: Insights from Equilibrium and Nonequilibrium Simulations." eng. In: *Structure (London, England : 1993)* 27.7 (July 2019), pp. 1171–1183. ISSN: 1878-4186 (Electronic). DOI: 10.1016/j.str.2019.04.008.

[157] Bogdan Lev et al. "String method solution of the gating pathways for a pentameric ligand-gated ion channel". In: *Proceedings of the National Academy of Sciences* 114.21 (May 2017), E4158–E4167. DOI: 10.1073/pnas.1617567114.

[158] Paola Conti et al. "Drug Discovery Targeting Amino Acid Racemases". In: *Chemical Reviews* 111.11 (Nov. 2011), pp. 6919–6946. ISSN: 0009-2665. DOI: 10.1021/cr2000702.

[159] Elodie Laine et al. "Use of allostery to identify inhibitors of calmodulin-induced activation of Bacillus anthracis edema factor". In: *Proceedings of the National Academy of Sciences* 107.25 (June 2010), pp. 11277–11282. DOI: 10.1073/pnas.0914611107.

[160] Remo Perozzo, Gerd Folkers, and Leonardo Scapozza. "Thermodynamics of protein-ligand interactions: history, presence, and future aspects." eng. In: *Journal of receptor and signal transduction research* 24.1-2 (Feb. 2004), pp. 1–52. ISSN: 1079-9893 (Print). DOI: 10.1081/rrs-120037896.

[161] Antonia Stank et al. "Protein Binding Pocket Dynamics." eng. In: *Accounts of chemical research* 49.5 (May 2016), pp. 809–815. ISSN: 1520-4898 (Electronic). DOI: 10.1021/acs.accounts.5b00516.

[162] Damien Monet et al. "mkgridXf: Consistent Identification of Plausible Binding Sites Despite the Elusive Nature of Cavities and Grooves in Protein Dynamics". In: *Journal of Chemical Information and Modeling* 59.8 (Aug. 2019), pp. 3506–3518. ISSN: 1549-9596. DOI: 10.1021/acs.jcim.9b00103.

[163] M Toulouse, V Fritsch, and E Westhof. "Rapid Calculation of Any Dielectric Function for Molecular Dynamics Simulations of Biological Macromolecules". In: *Molecular Simulation* 9.3 (Jan. 1992), pp. 193–200. ISSN: 0892-7022. DOI: 10.1080/08927029208047426.

[164] Nathan Desdouits, Michael Nilges, and Arnaud Blondel. "Principal Component Analysis reveals correlation of cavities evolution and functional motions in proteins". In: *Journal of Molecular Graphics and Modelling* 55 (2015), pp. 13–24. ISSN: 1093-3263. DOI: https://doi.org/10.1016/j.jmgm.2014.10.011.

[165] Zhuang Jin et al. "Synthesis and activity of substituted heteroaromatics as positive allosteric modulators for $\alpha 4\beta 2\alpha 5$ nicotinic acetylcholine receptors." eng. In: *Bioorganic & medicinal chemistry letters* 24.2 (Jan. 2014), pp. 674–678. ISSN: 1464-3405 (Electronic). DOI: 10.1016/j.bmcl.2013.11.049.

[166] K Y Chen et al. "A highly sensitive cell-based luciferase assay for high-throughput automated screening of SARS-CoV-2 nsp5/3CLpro inhibitors". In: *Antiviral Research* 201 (2022), p. 105272. ISSN: 0166-3542. DOI: https://doi.org/10.1016/j.antiviral.2022.105272.

[167] Pauli Virtanen et al. "SciPy 1.0: fundamental algorithms for scientific computing in Python". In: *Nature Methods* 17.3 (2020), pp. 261–272. ISSN: 1548-7105. DOI: 10.1038/s41592-019-0686-2.

[168] M Gerstein and C Chothia. "Packing at the protein-water interface." eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 93.19 (Sept. 1996), pp. 10167–10172. ISSN: 0027-8424 (Print). DOI: 10.1073/pnas.93.19.10167.

[169] Prasad Purohit and Anthony Auerbach. "Loop C and the mechanism of acetylcholine receptor-channel gating." eng. In: *The Journal of general physiology* 141.4 (Apr. 2013), pp. 467–478. ISSN: 1540-7748 (Electronic). DOI: 10.1085/jgp.201210946.

[170] Joshua Meyers, Benedek Fabian, and Nathan Brown. "De novo molecular design and generative models". In: *Drug Discovery Today* 26.11 (2021), pp. 2707–2715. ISSN: 1359-6446. DOI: https://doi.org/10.1016/j.drudis.2021.05.019.

[171] Jean-Louis Reymond. "The Chemical Space Project". In: *Accounts of Chemical Research* 48.3 (Mar. 2015), pp. 722–730. ISSN: 0001-4842. DOI: 10.1021/ar500432k.

[172] Xiaochu Tong et al. "Generative Models for De Novo Drug Design". In: *Journal of Medicinal Chemistry* 64.19 (Oct. 2021), pp. 14011–14027. ISSN: 0022-2623. DOI: 10.1021/acs.jmedchem.1c00927.

[173] Wenhao Gao and Connor W Coley. "The Synthesizability of Molecules Proposed by Generative Models". In: *Journal of Chemical Information and Modeling* 60.12 (Dec. 2020), pp. 5714–5723. ISSN: 1549-9596. DOI: 10.1021/acs.jcim.0c00174.

[174] John J Irwin et al. "ZINC: A Free Tool to Discover Chemistry for Biology". In: *Journal of Chemical Information and Modeling* 52.7 (July 2012), pp. 1757–1768. ISSN: 1549-9596. DOI: 10.1021/ci3001277.

[175] Michael M Mysinger et al. "Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking". In: *Journal of Medicinal Chemistry* 55.14 (July 2012), pp. 6582–6594. ISSN: 0022-2623. DOI: 10.1021/jm300687e.

[176] Janez Konc et al. "ProBiS-CHARMMing: Web Interface for Prediction and Optimization of Ligands in Protein Binding Sites". In: *Journal of Chemical Information and Modeling* 55.11 (Nov. 2015), pp. 2308–2314. ISSN: 1549-9596. DOI: 10.1021/acs.jcim.5b00534.

[177] Alvin V Terry and Patrick M Callahan. "Nicotinic Acetylcholine Receptor Ligands, Cognitive Function, and Preclinical Approaches to Drug Discovery." eng. In: *Nicotine & tobacco research : official journal of the Society for Research on Nicotine and Tobacco* 21.3 (Feb. 2019), pp. 383–394. ISSN: 1469-994X (Electronic). DOI: 10.1093/ntr/nty166.

[178] Antoine Daina, Olivier Michielin, and Vincent Zoete. "SwissTargetPrediction: updated data and new features for efficient prediction of protein targets of small molecules". In: *Nucleic Acids Research* 47.W1 (July 2019), W357–W364. ISSN: 0305-1048. DOI: 10.1093/nar/gkz382.

[179] Tiqing Liu et al. "BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities." eng. In: *Nucleic acids research* 35.Database issue (Jan. 2007), pp. 198–201. ISSN: 1362-4962 (Electronic). DOI: 10.1093/nar/gkl999.

[180] Miha Skalic et al. "Shape-Based Generative Modeling for de Novo Drug Design". In: *Journal of Chemical Information and Modeling* 59.3 (Mar. 2019), pp. 1205–1214. ISSN: 1549-9596. DOI: 10.1021/acs.jcim.8b00706.

[181] Peter Ertl and Ansgar Schuffenhauer. "Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions". In: *Journal of Cheminformatics* 1.1 (2009), p. 8. ISSN: 1758-2946. DOI: 10.1186/1758-2946-1-8.

[182] Amol Thakkar et al. "Retrosynthetic accessibility score (RAscore) – rapid machine learned synthesizability classification from AI driven retrosynthetic planning". In: *Chemical Science* 12.9 (2021), pp. 3339–3349. ISSN: 2041-6520. DOI: 10.1039/D0SC05401A.

[183] Connor W Coley et al. "SCScore: Synthetic Complexity Learned from a Reaction Corpus". In: *Journal of Chemical Information and Modeling* 58.2 (Feb. 2018), pp. 252–261. ISSN: 1549-9596. DOI: 10.1021/acs.jcim.7b00622.

[184] Samuel Genheden et al. "AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning". In: *Journal of Cheminformatics* 12.1 (2020), p. 70. ISSN: 1758-2946. DOI: 10.1186/s13321-020-00472-1.

[185] Melissa F Adasme et al. "PLIP 2021: expanding the scope of the protein–ligand interaction profiler to DNA and RNA". In: *Nucleic Acids Research* 49.W1 (July 2021), W530–W534. ISSN: 0305-1048. DOI: 10.1093/nar/gkab294.

[186] Jonathan B Baell and Georgina A Holloway. "New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays". In: *Journal of Medicinal Chemistry* 53.7 (Apr. 2010), pp. 2719–2740. ISSN: 0022-2623. DOI: 10.1021/jm901137j.

Chapter 6

Supplementary data

```
> Neuronal acetylcholine receptor subunit alpha-4 truncated
AHAEERLLKKLFSGYNKWSRPVANISDVVLVRFGLSIAQLIDVDEKNQMM
TTNVWVKQEWHDYKLRWDPADYENVTSIRIPSELIWRPDIVLYNNADGDF
AVTHLTKAHLFHDGRVQWTPPAIYKSSCSIDVTFFPFDQQNCTMKFGSWT
YDKAKIDLVNMHSRVDQLDFWESGEWVIVDAVGTYNTRKYECCAEIYPDI
TYAFVIRRLPLFYTINLIIPCLLISCLTVLVFYLPSECGEKITLCISVLL
SLTVFLLLITEIIPSTSLVIPLIGEYLLFTMIFVTLSIVITVFVLNVHHR
SPRTHTMPTWVRRVFLDIVPRLLL/SPALTRAVEGVQYIADHLKAEDTDF
SVKEDWKYVAMVIDRIFLWMFIIVCLLGTVGLFLPPW

> Neuronal acetylcholine receptor subunit beta-2 truncated
TDTEERLVEHLLDPSRYNKLIRPATNGSELVTVQLMVSLAQLISVHEREQ
IMTTNVWLTQEWEDYRLTWKPEEFDNMKKVRLPSKHIWLPDVVLYNNADG
MYEVSFYSNAVVSYDGSIFWLPPAIYKSACKIEVKHFPFDQQNCTMKFRS
WTYDRTEIDLVLKSEVASLDDFTPSGEWDIVALPGRRNENPDDSTYVDIT
YDFIIRRKPLFYTINLIIPCVLITSLAILVFYLPSDCGEKMTLCISVLLA
LTVFLLLISKIVPPTSLDVPLVGKYLMFTMVLVTFSIVTSVCVLNVHHRS
PTTHTMAPWVKVVFLEKLPALLF/GCGLREAVDGVRFIADHMRSEDDDQS
VSEDWKYVAMVIDRLFLWIFVFVCVFGTIGMFLQPL

> Neuronal acetylcholine receptor subunit alpha-5 truncated
SIAKXEDSLLKDLFQDYERWVRPVEXLNDKIKIKFGLAISQLVDVDEKNQ
LMTTNVWLKQEWIDVKLRWNPDDYGGIKVIRVPSDSVWTPDIVLFDNADG
RFEGTSTKTVIRYNGTVTWTPPANYKSSCTIDVTFFPFDLQNCSMKFGSW
TYDGSQVDIILEDQDVDKRDFFDNGEWEIVSATGSKGNRTDSCCWYPYVT
YSFVIKRLPLFYTLFLIIPCIGLSFLTVLVFYLPSNEGEKICLCTSVLVS
LTVFLLVIEEIIPSSSKVIPLIGEYLVFTMIFVTLSIMVTVFAINIHHRS
SSTHNAMAPLVRKIFLHTLPKLLC/RNTLEAALDSIRYITRHIMKENDVR
EVVEDWKFIAQVLDRMFLWTFLFVSIVGSLGLFVP
```

Figure 6.1: Sequence of the α_4,α_5,β_2 subunits of the modeled receptor