Computer-aided design of tubulin polymerization modulators

Maxim Shevelev

UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE DES SCIENCES CHIMIQUES

Chimie de la matière complexe – UMR 7140

en cotutelle avec Université de Barcelone

THÈSE présentée par :

Maxim SHEVELEV

soutenue le : **18 septembre 2023**

pour obtenir le grade de : **Docteur de l'Université de Strasbourg**

Discipline/ Spécialité : Chimie / Chémoinformatique

Conception assistée par ordinateur des modulateurs de polymérisation de la tubuline

THÈSE dirigée par :

M. VARNEK Alexandre	Professeur, Université de Strasbourg
M. RUBIO Jaime	Professeur, Université de Barcelone

THÈSE co-dirigée par :

M. HORVATH Dragos	Directeur de Recherche, CNRS, Université de Strasbourg
Mme. CASCANTE Marta	Professeur, Université de Barcelone

RAPPORTEURS :

M. KOLB Peter	Professeur, Université de Marbourg
M. MORELLI Xavier	Directeur de Recherche, CNRS, Université d'Aix-Marseille

---

AUTRES MEMBRES DU JURY :

Mme. DEJAEGERE Annick	Professeur, Université de Strasbourg

Maxim SHEVELEV

Conception assistée par ordinateur des modulateurs de polymérisation de la tubuline

Résumé

La protéine tubuline, cruciale pour la division cellulaire et le transport intracellulaire, est une cible clé dans la recherche sur le cancer et la neurodégénérescence. Les difficultés de synthèse et les propriétés pharmacologiques médiocres des agents existants ciblant la tubuline nécessitent de nouvelles découvertes. L'objectif de cette thèse était d'utiliser la conception de médicaments assistée par ordinateur pour identifier de nouvelles molécules qui ciblent des sites de liaison moins explorés et qui sont plus accessibles. La thèse a ciblé les sites peu étudiés de la maytansine, de la pironétine et du todalam avec des approches de criblage virtuel basées sur les ligands et la structure, et a conçu de nouvelles molécules pour le site de la colchicine en utilisant des technologies avancées d'apprentissage profond. La recherche a permis d'obtenir un total de 28 agents déstabilisateurs de microtubules nouveaux et structurellement diversifiés, ciblant les sites todalam, maytansine et colchicine. En outre, un logiciel d'analyse automatisée des images de microscope provenant d'expériences de diffraction de fibres de microtubules a été développé.

Mots clés : criblage virtuel, apprentissage profond, les agents antitubulines

Résumé en anglais

The tubulin protein, crucial for cell division and intracellular transport, is a key target in cancer and neurodegeneration research. Synthetic challenges and poor pharmacological properties of existing tubulin-targeting agents necessitate new discoveries. The goal of this thesis was to use computer-aided drug design to identify novel molecules that target less explored binding sites and are more synthetically accessible. The thesis targeted the understudied maytansine, pironetin, and todalam sites with ligand- and structure-based virtual screening approaches, and designed new molecules for the colchicine site using advanced deep learning technologies. The research yielded a total of twenty-eight structurally diverse and novel microtubule-destabilizing agents targeting the todalam, maytansine, and colchicine sites. Moreover, a software for automated analysis of microscope images from microtubule fiber diffraction experiments was developed.

Keywords: virtual screening, deep learning, microtubule-targeting agents

Acknowledgements

First and foremost, I would like to express my heartfelt appreciation to my supervisors, Prof. Alexandre Varnek and Dr. Dragos Horvath, whose patient guidance and shared wealth of experience have been instrumental in my journey. Their unwavering support and encouragement were a beacon during the more challenging periods of my work. Profound gratitude is extended to Prof. Jaime Rubio and Prof. Marta Cascante, my supervisors from the University of Barcelona, whose assistance, both scientifically and personally, proved invaluable during difficult times.

I am thankful to the jury members, Prof. Peter Kolb, Dr. Xavier Morelli, and Prof. Annick Dejaegere, for their time and expertise in reviewing and evaluating my work.

I would also like to thank my colleagues at the Laboratory of Chemoinformatics, University of Strasbourg. The opportunity to work and grow among such a gifted and innovative group of scientists has been an invaluable experience. In particular, I would like to thank Dr. Olga Klimchuk, Dr. Fanny Bonachera, Dr. Gilles Marcou, and Mme. Soumia Hnini, for their constant support in many ways during these past three years.

The TubInTrain International Training Network has my deepest appreciation for the financial support that made this thesis possible. I am grateful to all the supervisors from the TubInTrain consortium for their scientific support, the collaborative atmosphere they fostered, and for their role in making this endeavor a reality. I am particularly grateful to Prof. Daniele Passarella, Dr. Benedetta Santini, Dr. Fernando Diaz, and Dr. Andrea Prota for setting a remarkable example of scientific excellence.

My colleagues and dear friends from TubInTrain, namely Dr. Helena Perez-Peña, Sai Prashanth Shantapuri, Anne-Catherine Abel, Dr. Zlata Boiarska, Francesca Bonato, Óscar Fernandez Blanco, and Dr. Ahmed Soliman, deserve my warmest thanks. Without their support and significant scientific contribution, this thesis would not have been possible.

A special note of gratitude goes to all my friends from the Laboratory of Chemoinformatics, University of Strasbourg, both students and alumni, specifically Dr. Tagir Akhmetshin, Dr. Dmitry Zankov, Dr. Yuliana Zabolotna, Dr. Arkadii Lin, Karina Pikalyova, Regina Pikalyova, Farah Asgarkhanova, Shamkhal Baybekov, Polina Oleneva, Pierre Llompart and others.

I am also thankful to Anastasiia Delova, my sweetheart, whose unwavering support throughout this journey has been nothing short of extraordinary.

Lastly, I owe a deep sense of gratitude to my family, whose continuous support and encouragement have been a pillar of strength throughout this journey. Thank you for always believing in me.

Contents

Résumé en français

Introduction

Cette thèse de doctorat fait partie du programme européen de formation doctorale appelé TubInTrain, qui réunit 13 doctorants de différents profils scientifiques pour étudier les microtubules (MT) et leur rôle dans les maladies neurodégénératives et la neurotoxicité. La complexité croissante de la recherche scientifique nécessite la collaboration de spécialistes de nombreux domaines. Au sein de ces collaborations interdisciplinaires, la chimie computationnelle joue un rôle crucial car elle rationalise l'incorporation de modèles computationnels, l'analyse des données et la vérification expérimentale, reliant ainsi le travail théorique et empirique.

L'objectif central de cette thèse était de créer de nouveaux ligands de petites molécules qui modulent la polymérisation de la tubuline. Ce processus a fait appel à des méthodologies de conception de médicaments assistée par ordinateur pour faire avancer le processus de recherche et favoriser la collaboration entre les chimistes de synthèse, les biochimistes et les biologistes. Cette coopération interdisciplinaire a permis de trouver des solutions innovantes à des défis scientifiques complexes.

Dans cette thèse de doctorat, le Chapitre 1 présente la biologie de la tubuline et des microtubules, en détaillant leurs attributs structurels, leur nature dynamique et leurs rôles fonctionnels. Il souligne également leur importance en tant que cibles thérapeutiques potentielles dans le traitement de maladies telles que les maladies neurodégénératives et le cancer. De plus, le Chapitre 1 offre un aperçu des principales techniques de conception et de modélisation moléculaire assistées par ordinateur utilisées dans le cadre de cette recherche. Le Chapitre 2 décrit les efforts de criblage virtuel effectués pour trouver de nouvelles petites molécules qui ciblent le site de liaison de la maytansine de la protéine tubuline. Dans ce chapitre, le criblage pharmacophore, le docking protéine-ligand et la conception de médicaments *de novo* basée sur la structure ont été utilisés pour concevoir et découvrir des petites molécules qui inhibent la polymérisation de la tubuline, qui peuvent être davantage exploitées pour concevoir de nouveaux agents anti-tubuline ciblant le site de liaison de la maytansine. Le Chapitre 3 décrit les efforts de criblage virtuel réalisés pour trouver de nouvelles petites molécules qui ciblent le site de liaison de la pironétine. Dans ce chapitre, en plus des techniques déjà mentionnées, nous avons mis en œuvre le concept de criblage virtuel basé sur le docking protéine-ligand et piloté par l'apprentissage automatique de grandes bases de données. Le Chapitre 4 décrit l'investigation du site de todalam, récemment découvert dans la tubuline α. Ce chapitre détaille comment les approches de conception moléculaire assistées par ordinateur ont été utilisées pour identifier des chémotypes alternatifs ciblant le site de todalam

afin de les exploiter davantage pour concevoir des ligands covalents supposés à ce site. Le Chapitre 5 décrit l'application d'une approche de modélisation de relation quantitative structure à activité (modélisation QSAR) inverse pour la conception *de novo* de nouvelles petites molécules ciblant le site de liaison de la colchicine. Le Chapitre 6 décrit nos efforts pour enquêter sur l'applicabilité de l'approche d'apprentissage par transfert à l'apprentissage de représentations moléculaires utiles à partir de données moléculaires non étiquetées, et leur utilité dans la tâche de modélisation QSAR en aval. Le Chapitre 7 décrit l'étude de la dynamique conformationnelle de la protéine tubuline à l'aide de simulations de dynamique moléculaire accélérées dans le but de trouver d'éventuelles poches de liaison cryptiques à la surface de la tubuline. Enfin, le Chapitre 8 décrit le développement d'une application logicielle avec une interface utilisateur graphique, qui facilite et automatise l'analyse des résultats expérimentaux obtenus à partir d'expériences de diffraction des fibres de microtubules.

Les études computationnelles décrites dans ce travail ont été menées à la fois à l'Université de Strasbourg, en France, et à l'Université de Barcelone, en Espagne. La synthèse organique, la cristallographie par diffraction des rayons X et les essais biologiques décrits dans cette thèse de doctorat ont été réalisés par d'autres collègues de TubInTrain dans d'autres institutions, notamment à l'Université de Milan, en Italie, à l'Institut Paul Scherrer, en Suisse, et au Consejo Superior De Investigaciones Cientificas, en Espagne.

Chapitre 1. Contexte biologique et méthodologique

La protéine tubuline est un complexe globulaire dimérique composé de deux sous-unités distinctes, connues sous le nom d'α-tubuline et de β-tubuline[1]. Ces sous-unités sont structurellement similaires et sont maintenues ensemble, tête-bêche, par des interactions longitudinales non covalentes. Cet hétérodimère constitue le composant fondamental des microtubules, un élément clé du cytosquelette de toutes les cellules eucaryotes[1]. Les microtubules sont essentiels pour une multitude de fonctions cellulaires, telles que la signalisation cellulaire, le maintien de la forme de la cellule, la facilitation des mouvements cellulaires, la division cellulaire et le contrôle du trafic intracellulaire sur de longues distances[2].

Les microtubules se développent en ajoutant un hétérodimère de tubuline avec une molécule de guanosine triphosphate (GTP) liée aux deux sites de liaison des nucléotides. Les microtubules se développent tête-bêche, en ajoutant toujours de l'α-tubuline à la β-tubuline exposée (Figure R-1). Cela conduit à une structure de microtubule polaire avec la β-tubuline à l'extrémité qui s'étend (l'extrémité plus du microtubule). L'incorporation du dimère de tubuline dans la structure du microtubule entraîne un changement de conformation (d'une géométrie courbe

à une géométrie droite), suivi de l'hydrolyse du GTP dans le monomère de β-tubuline. Une "coiffe de GTP" située à l'extrémité plus du microtubule, composée de dimères avec du GTP dans les deux sites, stabilise l'extrémité, empêchant la dépolymérisation. Lorsque le taux d'ajout de dimères dépasse celui de l'hydrolyse du GTP, il y a extension du microtubule ; dans le cas contraire, le microtubule subit une dégradation, également connue sous le nom de "catastrophe"[1,2].

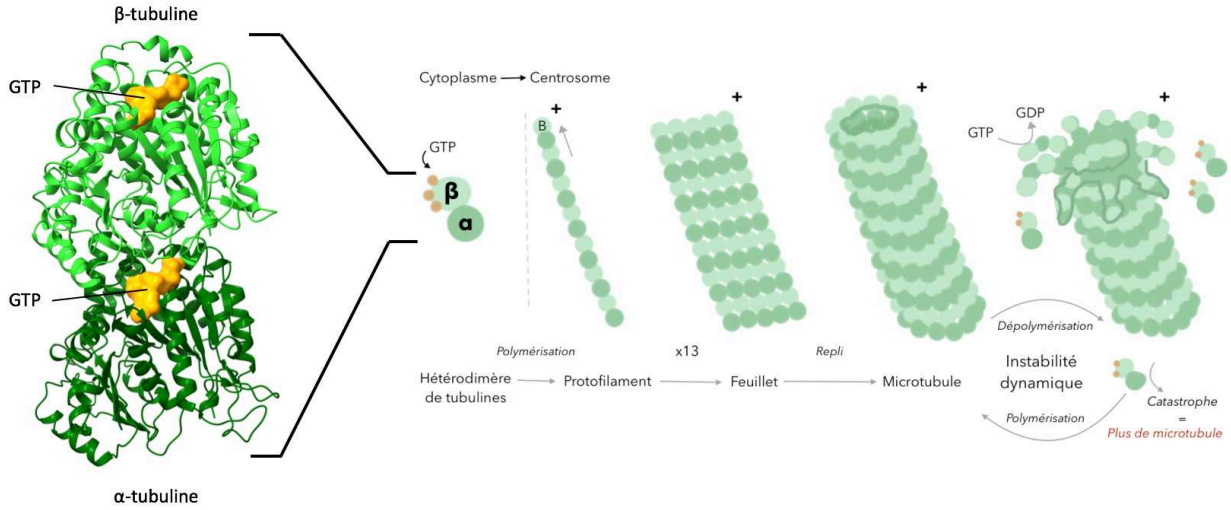


Figure R-1. Schéma général de la formation des microtubules et de leur instabilité dynamique

La dynamique du réseau de microtubules est influencée par les protéines associées aux microtubules (MAP), les modifications post-traductionnelles et les agents ciblant les microtubules (MTA).

Les protéines associées aux microtubules (MAP) sont un ensemble diversifié de protéines qui régissent et guident l'instabilité dynamique des microtubules. Les MAP interagissent avec les microtubules pour réguler leur dynamique, leur stabilité et leur organisation. Leur rôle critique s'étend à de nombreux processus cellulaires. Les cellules eucaryotes comptent généralement plus de 100 protéines différentes qui se lient aux microtubules. La structure de la tubuline comprend un court C-terminal (~20 acides aminés) enrichi en acides glutamiques et aspartiques. Par conséquent, lorsque la tubuline s'assemble en microtubules, la surface de ces derniers porte une charge négative nette. Par conséquent, de nombreux MAPs, qui sont chargés positivement, se lient aux microtubules par le biais d'interactions électrostatiques. Les MAPs peuvent être classés en deux grandes catégories : ceux qui interagissent avec des hétérodimères de tubuline individuels et ceux qui se lient à des microtubules entièrement formés. En outre, les MAP peuvent augmenter ou diminuer la stabilité des microtubules, en favorisant ou en inhibant la polymérisation de la tubuline.

Les MTA à forte concentration ont un impact sur la dynamique des microtubules de différentes manières, ce qui a conduit à les classer en deux groupes : les agents stabilisateurs des

microtubules (MSA), qui favorisent la polymérisation des microtubules et augmentent la stabilité de leur structure, et les agents déstabilisateurs des microtubules (MDA), qui empêchent l'assemblage des dimères dans les microtubules[1].

Les microtubules jouent un rôle essentiel dans le processus de mitose et participent à diverses opérations cellulaires. En raison de leur rôle central dans la division cellulaire, un processus indispensable à l'expansion et à la multiplication des cellules cancéreuses, les microtubules constituent une cible intéressante pour le traitement du cancer. Cependant, le problème des MTA est leur manque de spécificité, qui peut endommager des cellules saines, ce qui entraîne des effets secondaires graves. Dans le contexte des maladies neurodégénératives, les dysfonctionnements de la dynamique des microtubules neuronaux constituent un mécanisme causal fondamental, et les microtubules représentent donc une cible attrayante pour le traitement de ces maladies. La compréhension de la fonction des microtubules dans l'activité neuronale et la maladie peut grandement contribuer à la création de traitements efficaces pour les troubles neurologiques. Les méthodes informatiques peuvent être utilisées pour concevoir et créer des MTA en tant que sondes moléculaires, réduisant ainsi le temps, les dépenses et les risques associés au développement de nouveaux médicaments[1].

Les MTA se lient à la tubuline, entravant sa polymérisation, et peuvent être utilisés comme agents thérapeutiques dans le traitement du cancer et des maladies neurodégénératives. Ils sont également utiles à la recherche pour étudier la structure et la fonctionnalité des microtubules. Huit sites de liaison confirmés pour les MTA ont été identifiés sur la tubuline, dont cinq sur la β-tubuline (sites colchicine, taxane, vinca, peloruside/laulimalide et maytansine), un sur l'α-tubuline (site pironétine), un à l'interface intra-dimère (site gatorbuline) et un à l'interface inter-dimère (site todalam) (Figure R-2).

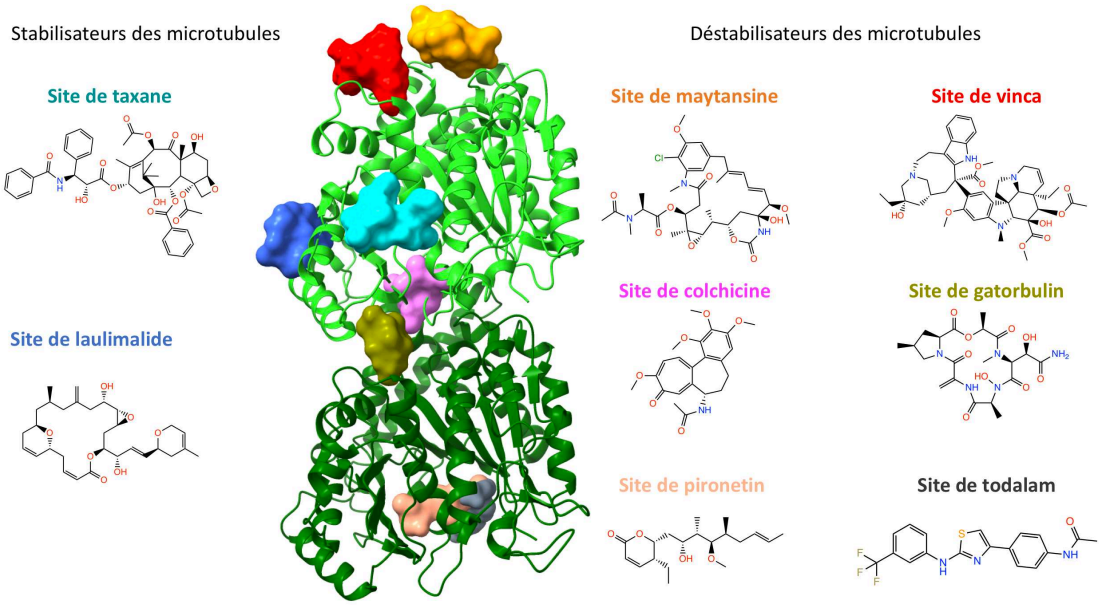


Figure R-2. Différents sites de liaison de la tubuline et exemples de ligands associés

Les MTA induisent divers effets sur la polymérisation de la tubuline et la stabilité des microtubules en fonction du site spécifique auquel ils se lient[1]. Le paclitaxel, un MSA qui se lie au site taxane, renforce la stabilité des microtubules en facilitant la transition courbe-droite. Les ligands du site colchicine empêchent la compaction de la poche formée par les brins βS8 et βS9 ainsi que les hélices βH8 et αH7, bloquant ainsi la transition courbe-droite nécessaire à l'assemblage des microtubules. Les ligands du site peloruside/laulimalide renforcent l'interaction entre les dimères de tubuline dans les protofilaments voisins au sein des microtubules. Les ligands du site vinca introduisent un "coin" à l'extrémité de croissance des microtubules, empêchant ainsi l'ajout de nouveaux hétérodimères de tubuline. Les ligands du site maytansine entravent l'assemblage des microtubules en suivant un mécanisme similaire. La pironétine se lie à une poche enfouie par attachement covalent à la Cys316, perturbant l'hélice αH8 et la boucle αT7. Todalam, le premier ligand de tubuline conçu rationnellement, empêche la formation de microtubules en formant un coin dans la structure de l'oligomère de tubuline. Le mécanisme d'action de la gatorbuline est actuellement à l'étude.

Les MTA ont démontré leur puissance en tant que modulateurs de la croissance cellulaire et ont une importance pratique exceptionnelle (notamment les agents anticancéreux comme le paclitaxel, la vinblastine et la maytansine). Le réseau européen ITN TubInTrain a consacré des efforts considérables à la découverte systématique de nouveaux ligands de la tubuline.

La conception moléculaire assistée par ordinateur est une méthodologie informatique qui exploite la modélisation et les simulations informatiques pour la conception et la découverte de nouvelles molécules[3]. Le criblage virtuel, une technique courante de la conception moléculaire assistée par ordinateur, implique la recherche de composés chimiques ayant une forte probabilité de se lier à une cible thérapeutique, généralement des protéines. Après l'identification de ligands potentiels par des techniques informatiques, il est nécessaire d'utiliser des méthodes expérimentales pour évaluer les molécules trouvées. Des techniques telles que la recherche de similitudes, la recherche de sous-structures, la modélisation pharmacophore, le docking protéine-ligand, la modélisation QSAR et les simulations de dynamique moléculaire sont des exemples de méthodes de conception de médicaments assistée par ordinateur utilisées dans cette thèse de doctorat pour examiner les modes de liaison et les affinités d'une bibliothèque de composés virtuels. Ces instruments de calcul sont essentiels pour identifier et hiérarchiser les candidats moléculaires potentiels, réduire la durée et les dépenses associées au développement de petites molécules et améliorer le taux de réussite des "hits"[3].

La recherche de similitude[4] est une méthode employée pour identifier les molécules ayant des propriétés similaires à celles de composés actifs connus. Elle repose sur le principe de la similarité moléculaire, selon lequel des molécules structurellement similaires sont statistiquement

susceptibles d'avoir des propriétés similaires. La recherche de similitude consiste à comparer les caractéristiques structurelles et chimiques de différentes molécules, le plus souvent à l'aide de fingerprints moléculaire, qui sont des représentations sous forme de chaînes de bits de la structure et des propriétés moléculaires. Chaque bit de l'empreinte digitale représente la présence ou l'absence d'une caractéristique structurelle particulière au sein de la molécule. Il est également possible d'utiliser d'autres descripteurs moléculaires. En classant les composés sur la base de leurs scores de similarité, la recherche par similarité permet de cribler efficacement de grandes bibliothèques de composés afin de trouver des médicaments candidats potentiels dès les premiers stades de la découverte de médicaments. Les mesures de similarité couramment utilisées comprennent le coefficient de Tanimoto, le coefficient de Dice, la distance euclidienne ou la distance de Manhattan. La combinaison idéale de descripteurs et de fonctions métriques est celle qui garantit la meilleure "conformité au comportement de voisinage"[5]. Il s'agit de minimiser les situations où des paires de composés semblent très similaires malgré des valeurs de propriétés différentes, connues sous le nom de "falaises de propriétés"[5].

La recherche de sous-structures[6] est une opération fondamentale qui implique l'identification de fragments ou de motifs moléculaires spécifiques au sein de structures moléculaires plus larges. Les principes clés qui sous-tendent le concept de recherche de sous-structures sont ancrés dans la théorie des graphes. Les molécules et leurs sous-structures peuvent être représentées comme des graphes, où les atomes sont des nœuds et les liaisons des arêtes. Le problème de la recherche de sous-structures est alors transformé en un problème d'isomorphisme de sous-graphes, qui consiste à trouver une correspondance biunivoque entre les nœuds et les arêtes du graphe de sous-structures et un sous-ensemble de nœuds et d'arêtes du graphe de molécules. La recherche de sous-structures est utilisée dans le cadre du criblage virtuel, car elle permet de filtrer et de sélectionner efficacement des composés à partir de grandes bases de données en fonction de la présence de groupes fonctionnels ou de sous-structures spécifiques. Cette méthode est particulièrement utile si l'on sait que certaines sous-structures sont liées à des propriétés souhaitables, telles que l'affinité de liaison avec une protéine cible. En identifiant les composés qui contiennent ces sous-structures, les chercheurs peuvent donner la priorité à certains composés pour des tests et des analyses plus poussés[7].

La modélisation des relations quantitatives structure-activité (QSAR)[8] est une méthode qui établit une relation fonctionnelle entre un ensemble de descripteurs moléculaires et une propriété quantifiable d'une molécule. Ainsi, un modèle QSAR est une fonction $\mathcal{F}$ qui produit une estimation raisonnable d'une propriété cible $\mathcal{Y}$ à partir d'un ensemble de descripteurs moléculaires $\mathcal{D}$, $\mathcal{Y} = \mathcal{F}(\mathcal{D})$. Les descripteurs peuvent être classés en descripteurs 1D, 2D et 3D, chacun capturant des aspects différents de la structure de la molécule. Les descripteurs 1D sont simples et

comprennent des propriétés telles que le poids moléculaire, tandis que les descripteurs 2D capturent des informations topologiques ou de connectivité, et les descripteurs 3D reflètent l'arrangement spatial des atomes dans une molécule. Le processus de recherche de cette dépendance fonctionnelle s'appelle l'ajustement du modèle ou l'entraînement.

Pour effectuer une modélisation QSAR, il faut disposer d'un ensemble de données comprenant les structures moléculaires et les propriétés expérimentales correspondantes[8]. Les données doivent être soigneusement contrôlées avant la modélisation, en veillant à supprimer les doublons, à normaliser la structure (transformation des formes tautomériques et de résonance en une seule forme, neutralisation des charges et élimination des petits fragments des sels), à vérifier l'exactitude des données et à transformer les données biologiques en une forme adaptée à la modélisation mathématique.

En appliquant un modèle QSAR entraîné aux vecteurs de descripteurs de composés inconnus, il est possible de prédire leurs propriétés, ce qui s'avère utile aux premiers stades de la conception et de la découverte de médicaments pour identifier de nouveaux composés potentiels[8].

Une autre méthode importante utilisée dans ce travail est la modélisation du pharmacophore[9]. Un modèle de pharmacophore est un ensemble de caractéristiques stériques et électroniques essentielles d'un ligand qui assurent des interactions supramoléculaires optimales avec une cible biologique spécifique. Il existe plusieurs types de caractéristiques pharmacophores communes, notamment les accepteurs et les donneurs de liaisons hydrogène, les groupes chargés ou ionisables, les résidus hydrophobes et les anneaux aromatiques. Ces caractéristiques reflètent le concept de bioisostérisme, reconnaissant que différents groupes fonctionnels peuvent présenter des propriétés physicochimiques similaires[9].

Dans un modèle pharmacophore tridimensionnel[10], les éléments ont des relations spatiales spécifiques les uns avec les autres, sous forme de distances ou de plages de distances entre les éléments. Les coordonnées spatiales des éléments sont généralement complétées par une région de tolérance sphérique pour tenir compte de la variabilité de la distance.

La source des données pour générer un modèle pharmacophore peut varier. Il existe deux méthodes courantes : la modélisation basée sur la structure, qui se repose sur la structure tridimensionnelle d'un complexe ligand-protéine, et la modélisation basée sur le ligand, qui dépend uniquement des informations structurelles des composés actifs[10].

Avant le criblage[10], chaque molécule d'une bibliothèque de composés est représentée par un ensemble de conformères, qui incluent potentiellement la géométrie bioactive supposée lors de l'interaction avec la protéine cible. Les correspondances entre le modèle pharmacophore et les conformères sont compilées dans une liste de résultats. Un système de scoring est ensuite utilisé pour classer les molécules de la liste de résultats. Ce score quantifie la qualité de la correspondance

entre chaque molécule et le modèle pharmacophore, fournissant une mesure de l'aptitude potentielle de chaque molécule en tant que candidat-médicament.

Une autre méthode computationnelle qui a été déterminante pour ce travail est le docking protéine-ligand[11]. Le docking protéine-ligand est une méthode largement utilisée pour estimer la manière dont un ligand interagit avec un site de liaison protéique spécifique. Ce processus est essentiel pour comprendre les interactions récepteur-ligand et les mécanismes d'action des médicaments, car il permet de prédire la pose de liaison du ligand et d'estimer grossièrement son affinité de liaison. Un programme de docking comprend généralement deux éléments : l'algorithme d'échantillonnage conformationnel et la fonction de notation. L'algorithme d'échantillonnage est chargé de générer un grand nombre d'orientations et de conformations potentielles du ligand dans le site de liaison de la protéine. L'objectif d'une fonction de notation est alors de prédire l'affinité de liaison de chaque conformation du ligand à l'aide d'une fonction d'énergie empirique. Les méthodes protéine-ligand utilisées dans ce travail considéraient que le squelette de la protéine était rigide et ne tenaient pas compte de la flexibilité des chaînes latérales[11].

Le présent travail de doctorat a également étudié l'application de l'apprentissage profond aux tâches de la chemoinformatique. L'apprentissage profond s'est récemment imposé comme un outil puissant dans la découverte de médicaments, jouant un rôle central dans le processus[8]. Il est utilisé dans différentes tâches chimiques, par exemple la prédiction des interactions entre médicaments et cibles, la conception de médicaments *de novo* et la modélisation des relations quantitatives structure-activité (QSAR). L'importance de l'apprentissage profond dans le contexte de la découverte de médicaments réside, en particulier, dans sa capacité avancée à prédire les propriétés et les fonctions moléculaires, et à générer de manière automatisée des entités chimiques innovantes dotées des propriétés souhaitées.

L'apprentissage par transfert, un concept de l'apprentissage automatique, a été appliqué à la découverte de médicaments pour relever le défi de l'identification de descripteurs appropriés pour les tâches de modélisation en aval. Par essence, l'apprentissage par transfert est une méthode qui consiste à adapter un modèle pré-entraîné à une tâche nouvelle, mais connexe. Il permet d'appliquer les connaissances acquises lors de la résolution d'un problème à un problème différent mais connexe. Cette méthode est particulièrement utile dans les situations où les données relatives à la tâche concernée sont rares ou lorsque la tâche est trop complexe pour être apprise à partir de zéro.

Le processus d'apprentissage par transfert dans la découverte de médicaments comporte deux étapes principales : la pré-entraînement et le réglage fin. Le pré-entraînement est la phase initiale au cours de laquelle un modèle est entraîné sur un grand ensemble de données afin d'apprendre une représentation générale des données. Dans le contexte de la découverte de

médicaments, la phase de pré-entraînement consiste à entraîner un modèle sur un grand ensemble de données afin d'apprendre des représentations générales des molécules. Cet objectif est atteint grâce à l'apprentissage auto-supervisé, où le modèle apprend à prédire certains aspects des données à partir d'autres parties des mêmes données. Pour ce faire, on utilise souvent l'apprentissage auto-supervisé, où le modèle apprend à prédire des parties des données d'entrée à partir d'autres parties, ce qui lui permet d'apprendre des représentations utiles des données. Ce processus permet au modèle d'apprendre des représentations utiles des molécules sans avoir besoin de données étiquetées pour la tâche spécifique à accomplir.

Après le pré-entraînement, le modèle subit un processus de réglage fin. Le réglage fin consiste à ajuster le modèle préapprenti pour le rendre plus adapté à la tâche spécifique à accomplir. Pour ce faire, la formation du modèle se poursuit sur les données de la tâche spécifique, ce qui permet au modèle d'adapter les représentations apprises aux caractéristiques spécifiques de la nouvelle tâche. Dans le contexte de la découverte de médicaments, le réglage fin peut contribuer à accroître les performances prédictives d'un modèle QSAR sur une tâche spécifique en adaptant les représentations moléculaires générales apprises lors du pré-entraînement aux propriétés spécifiques pertinentes pour la tâche QSAR en question.

Pour mettre en œuvre avec succès l'apprentissage par transfert dans la découverte de médicaments, certaines conditions doivent être remplies. Tout d'abord, il faut disposer d'un vaste ensemble de données pour le préapprentissage afin d'apprendre les représentations moléculaires générales. Cet ensemble de données devrait idéalement couvrir un large éventail d'espaces chimiques. Deuxièmement, des données spécifiques à une tâche sont nécessaires pour affiner le modèle. Ces données doivent être pertinentes pour la tâche spécifique à accomplir et doivent idéalement contenir des exemples de propriétés ou d'activités spécifiques que le modèle doit prédire. Enfin, un modèle d'apprentissage profond approprié, capable d'apprendre à partir des données de préformation et de s'adapter aux données de mise au point, est nécessaire. Ce modèle doit être capable d'apprendre des modèles et des relations complexes dans les données, et doit être suffisamment flexible pour adapter les représentations apprises à la tâche spécifique.

L'apprentissage par transfert offre une approche prometteuse pour relever le défi de l'identification de descripteurs appropriés pour les tâches de modélisation en aval dans la découverte de médicaments. En tirant parti de la puissance de l'apprentissage profond et du concept d'apprentissage par transfert, il est possible d'apprendre des représentations moléculaires utiles à partir de grands ensembles de données et d'adapter ces représentations à des tâches spécifiques, améliorant ainsi la performance prédictive des modèles dans la découverte de médicaments.

Chapitre 2. Découverte de possibles agents déstabilisateurs de microtubules ciblant le site de la maytansine

La polymérisation de la tubuline, un processus essentiel à la fonction cellulaire, implique l'alignement des unités α- et β-tubuline. Une étape critique de ce processus est l'interaction d'une boucle spécifique de la sous-unité α-tubuline avec une cavité unique de la sous-unité β-tubuline. La perturbation de cette interaction, par exemple par la liaison de molécules comme la maytansine et ses analogues à la cavité de la β-tubuline, inhibe la polymérisation de la tubuline[1]. Malgré leur activité inhibitrice connue, l'application pratique des ligands ciblant le site de la maytansine est limitée en raison de leur synthèse complexe, de leur coût élevé et de leur extrême cytotoxicité[12]. Notre étude visait à utiliser des techniques de chimie computationnelle pour identifier des molécules plus accessibles et moins cytotoxiques qui pourraient cibler le site de liaison de la maytansine (Figure R-3).

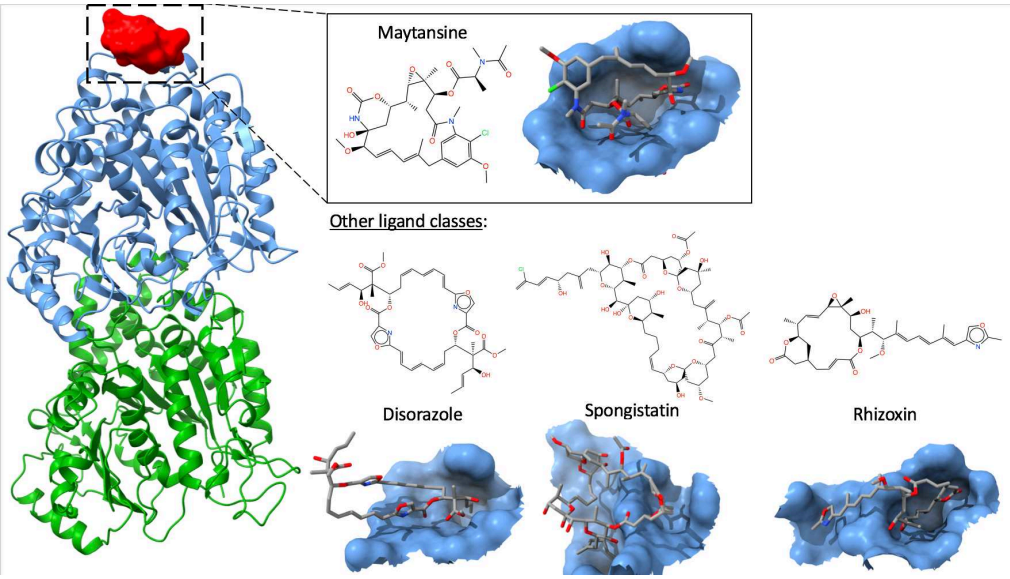


Figure R-3. Le site de liaison de la maytansine et certains ligands qui le ciblent

Notre projet a commencé par une évaluation des données disponibles. Nous avons constaté que la base de données ChEMBL[13], qui contient des composés médicamenteux et leurs résultats d'essais biologiques, ne contenait pas d'essais biologiques spécifiques sur les molécules ciblant le site de la maytansine. Par conséquent, nos données de lancement provenaient de plusieurs structures cristallines PDB obtenues à partir de la base de données PDB du RCSB[14].

Afin d'établir une structure de base pour la modélisation pharmacophore, nous avons réalisé une expérience de re-docking, un processus qui consiste à re-docker un ligand connu dans son récepteur natif. Pour ce faire, nous avons extrait des ligands natifs de toutes les structures PDB pertinentes, généré des conformations aléatoires pour ces ligands et les avons redocké dans le site

de liaison. Le ligand maytansinol de la structure 7E4Z a été redocké avec l'écart quadratique moyen le plus faible, ce qui indique que notre logiciel de docking pouvait reproduire correctement son mode de liaison. Par conséquent, nous l'avons choisie comme structure de base pour la construction automatisée d'un modèle pharmacophore avec le logiciel LigandScout[10]. Le modèle obtenu contenait cinq caractéristiques et classait correctement les autres dérivés du maytansinol co-cristallisés avec la tubuline comme actifs.

Nous avons criblé deux bibliothèques : ChEMBL (version 24, avec 1,5 million de molécules) et la Enamine High-Throughput Collection (contenant 2,7 millions de composés à l'époque). Les bibliothèques ont été standardisées à l'aide d'un pipeline interne basé sur ChemAxon, et 200 et 25 conformations ont été calculées pour chaque composé dans les bibliothèques ChEMBL et Enamine, respectivement. Le calcul des conformations a assuré la diversité entre les poses calculées dans une fourchette de 1.5 Å de RMSD.

Après le criblage pharmacophore, les molécules sélectionnées ont fait l'objet d'un docking rigide avec le logiciel PLANTS[15], le site de liaison étant dérivé de la structure PDB 7E4Z. Le ligand et la protéine ont été préparés avec SPORES[15], et les meilleures poses ont été réévaluées à l'aide du modèle pharmacophore.

Le criblage de la bibliothèque ChEMBL a mis en évidence la glycibridine, un produit naturel qui correspondait au modèle pharmacophore du maytansinol et qui présentait un score favorable dans le site de liaison de la maytansine. Après avoir généré tous les dérivés possibles et comparé leurs scores d'amarrage et leurs modes de liaison potentiels, nous avons déterminé que la structure originale était l'option la plus appropriée. Cependant, la synthèse de cette molécule a échoué.

Le criblage de la bibliothèque d'énamines a donné 11 molécules potentielles. Parmi celles-ci, deux composés ont montré une légère diminution de la polymérisation de la tubuline dans l'essai de polymérisation des microtubules *in vitro*, indiquant des interactions potentielles avec la tubuline ou les microtubules (Figure R-4). Cependant, la cristallographie aux rayons X n'a détecté aucune liaison entre ces molécules et la tubuline.

La complexité des molécules macrocycliques qui se lient au site de la maytansine provient de leur origine en tant que produits naturels, ce qui entraîne des processus de synthèse compliqués, qui entravent l'exploration et l'exploitation efficaces de ce site de liaison. Nous avons donc voulu étudier l'application des techniques de conception moléculaire *de novo* afin d'utiliser leur potentiel pour générer de nouvelles molécules[3]. Cependant, un inconvénient commun à ces méthodes est que la génération de molécules chimiquement valables néglige souvent les aspects pratiques et financiers de leur synthèse. Cette limitation souligne la nécessité d'un mécanisme de contrôle de la faisabilité chimique dans les méthodes de conception *de novo*[3].

Pour relever ce défi, il faut introduire un score d'accessibilité synthétique dans le pipeline de conception *de novo*. Une façon d'y parvenir est d'utiliser une approche de synthèse en amont. La tâche de prédiction de la synthèse en amont (également connue sous le nom de tâche de la prédiction de la réaction) consiste à trouver une chaîne de réactions chimiques synthétiquement valide appliquée à un nombre limité de blocs de construction chimiques facilement disponibles qui produisent la molécule cible requise, générant ainsi son arbre de synthèse. Cela contraste avec une approche rétrosynthétique plus couramment utilisée, dans laquelle la molécule cible est séquentiellement décomposée en petits fragments achetables. Les techniques de forward et de rétrosynthèse peuvent être mises en œuvre pour faciliter la génération de molécules *de novo*. Dans ce travail, nous avons mis en œuvre un pipeline de génération de molécules *de novo* basé sur la structure, qui utilise un outil interne capable de résoudre une tâche de prédiction de réaction en amont.

Pour s'assurer que les molécules sont spécifiquement conçues pour le site de liaison, nous avons mis au point une implémentation de cette approche basée sur un algorithme génétique pour le docking protéine-ligand, en utilisant l'approche de synthèse en amont comme mesure de l'accessibilité chimique des molécules générées. Au lieu de générer la molécule cible, l'outil de synthèse directe l'utilise comme référence. L'objectif de la construction de l'arbre synthétique est d'optimiser à la fois le score d'ancrage des produits dans chaque nœud et la similarité avec le ligand de référence. Cette dernière exigence limite l'outil à l'exploration de l'espace des composés semblables à des médicaments. La sélection des fragments et des transformations chimiques à chaque étape est guidée par un algorithme génétique. Cela conduit à la génération de molécules qui sont (1) similaires au ligand connu jusqu'à une valeur seuil de similarité spécifiée par l'utilisateur, garantissant ainsi la similarité des molécules générées avec les médicaments ; et (2) qui produisent un bon score de docking dans le site de liaison. L'orientation du processus en fonction du score de docking permet de produire des molécules dont l'affinité prévue pour le site est améliorée. L'examen des transformations chimiques qui génèrent les molécules ayant le meilleur score d'amarrage nous permet de reconstruire leur voie de synthèse. Par conséquent, cet outil sert à la fois d'instrument de conception *de novo* basée sur la structure et d'instrument de synthèse à terme.

Cette génération *de novo* de molécules adaptées au site de liaison de la maytansine, basée sur la structure, a donné à trois petites molécules. Grâce à la caractéristique unique de notre logiciel, nous avons pu extraire la séquence précise des éléments constitutifs et des transformations chimiques qui ont produit les molécules ayant le score de docking le plus élevé dans le site de liaison. L'évaluation des meilleures poses dockées de ces molécules ont révélé des caractéristiques

clés qui se chevauchent, ce qui est prometteur pour la synthèse et le développement futurs de ces molécules.

Les travaux futurs se concentreront sur l'utilisation d'un logiciel du docking plus avancé qui peut prendre en compte le site superficiel exposé au solvant, le développement d'un modèle pharmacophore plus complet qui incorpore des caractéristiques d'autres dérivés du maytansinol disponibles dans la PDB, et la préparation de bibliothèques de criblage par le calcul de conformations supplémentaires. En outre, les efforts visant à synthétiser le produit naturel glycibridin B en vue de le tester se poursuivront.

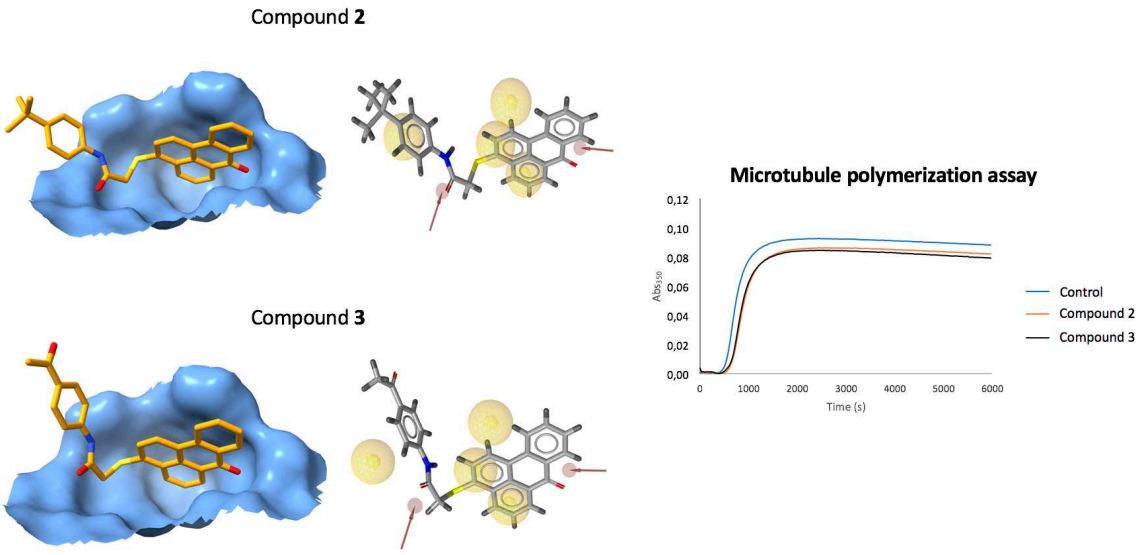


Figure R-4. Deux "hit molecules" qui montrent une certaine activité d'inhibition de la polymérisation des microtubules et leur mode d'action prédit

Chapitre 3. Criblage virtuel de nouveaux inhibiteurs de la polymérisation de la tubuline ciblant le site de la pironétine

Il est clair que perturber la polymérisation des microtubules en manipulant l'interaction entre la boucle spécifique d'une sous-unité d'α-tubuline et une cavité au sommet d'une sous-unité de β-tubuline constitue une approche stratégique pour la modulation de la dynamique des microtubules. Une opportunité intéressante se présente lorsque l'on considère la possibilité d'induire des changements de conformation dans la boucle elle-même, empêchant son placement dans la cavité. La pironétine (Figure R-5), un produit naturel isolé de *Streptomyces sp.* est connue pour provoquer ce déplacement en se liant à un site de liaison unique sur la sous-unité α-tubuline[1]. La pironétine se lie de manière covalente à une poche enfouie sur la sous-unité α-tubuline par une réaction d'addition de Michael avec un résidu cystéine à l'intérieur du site, près de la boucle

essentielle pour la liaison intra-tubuline[16]. Cependant, la difficulté de sa synthèse, sa cytotoxicité élevée et le manque d'options d'optimisation abordables limitent ses applications cliniques[17]. Par conséquent, dans ce projet, nous avons cherché à utiliser des techniques de chimie computationnelle pour identifier des analogues de la pironétine abordables et facilement accessibles qui pourraient réguler la polymérisation de la tubuline en suivant un mécanisme de liaison similaire.

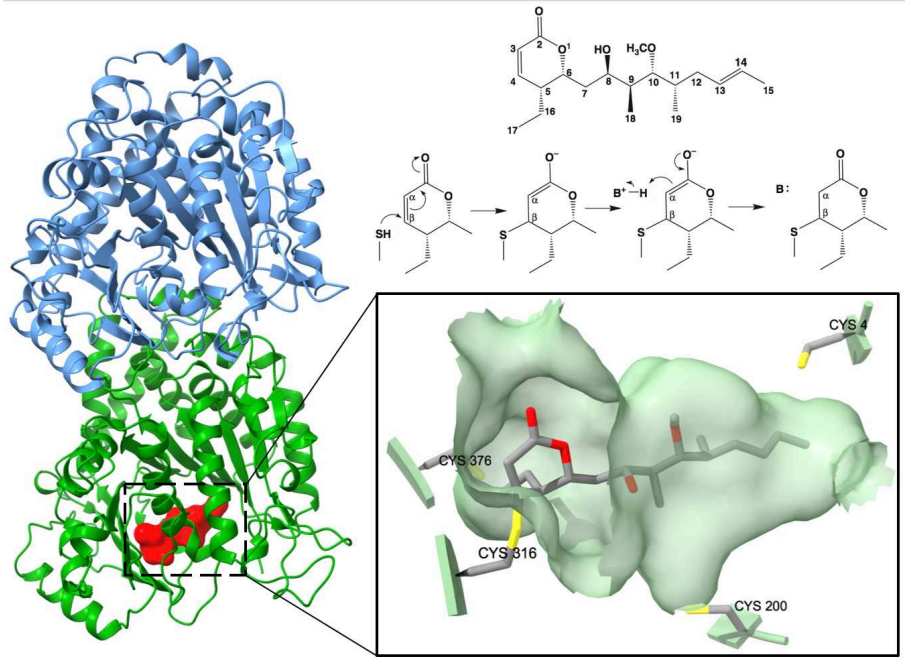


Figure R-5. Le site de liaison de la pironétine

Pour cette étude, nous avons employé des techniques de modélisation et de criblage pharmacophore, ainsi que des techniques d'amarrage protéine-ligand, afin d'identifier les molécules qui pourraient éventuellement se lier au site de la pironétine. Nous avons également mis en œuvre un pipeline de criblage virtuel efficace piloté par une prédiction itérative des scores de docking basée sur l'apprentissage automatique. Ce type de criblage itératif[18] utilise des modèles de relations quantitatives structure-activité (QSAR) qui sont entraîné sur les scores de docking de sous-ensembles d'une chimiothèque (Figure R-6). Ces modèles se rapprochent du résultat de docking pour les entrées non traitées et éliminent les molécules défavorables de manière itérative. La mise en œuvre du criblage itératif dans ce projet a impliqué l'utilisation d'une méthode de machine à vecteur de support pour apprendre les scores de docking produits par le logiciel PLANTS sur la base de la structure 2D d'un composé représentée par les descripteurs ISIDA[19].

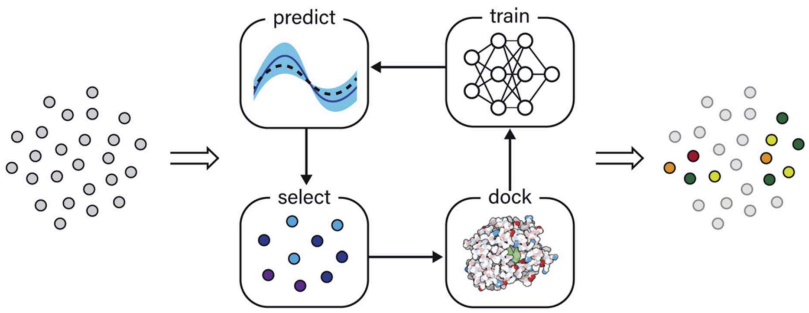


Figure R-6. Aperçu des approches de criblage itératif basées sur l'apprentissage automatique

Nos données initiales étaient constituées de deux structures PDB de la pironétine et de trois structures cristallines de petits fragments qui se sont liés à une cavité proche du site de liaison de la pironétine, publiées par nos collaborateurs[20]. Nous avons choisi la structure PDB 5FNV en raison de la meilleure résolution autour du ligand. Les modèles pharmacophores ont été générés automatiquement à partir des structures expérimentales par le logiciel LigandScout. Nous avons utilisé le Enamine High-Throughput Screening collection pour le criblage itératif et le criblage pharmacophore, ainsi que huit librairies Enamine plus petites et plus spécifiques. L'étape de préparation des chimiothèques a reflété notre approche dans le cadre du projet sur le site de la maytansine. Les hits virtuels identifiés ont été dockés dans le site de la pironétine et leurs scores de docking ont été comparés à ceux de la pironétine. Les 47 hits virtuels ont également été dockés dans le site de la colchicine à des fins d'analyse comparative.

Bien qu'aucun des composés sélectionnés par le criblage virtuel n'ait été détecté dans le site de la pironétine par cristallographie aux rayons X, trois petits fragments trouvés ont démontré une activité significative de déstabilisation des microtubules lors d'essais de polymérisation des microtubules. En outre, deux molécules trouvés se sont liées au site de la colchicine. Des tests supplémentaires ont révélé que l'une de ces molécules présentait une préférence pour l'isotype βIII de la tubuline, un biomarqueur surexprimé chez les patients cancéreux, en particulier ceux qui sont résistants à d'autres types de médicaments anticancéreux ciblant la tubuline (Figure R-7).

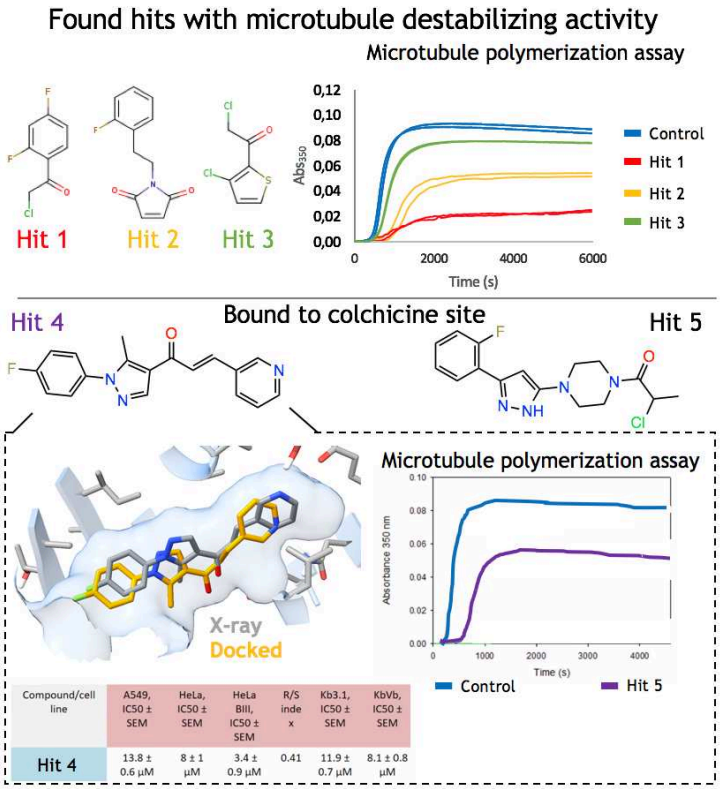


Figure R-7. Le criblage virtuel a révélé trois petits fragments puissants et deux molécules qui se lient au site de liaison de la colchicine, dont l'une présente une certaine spécificité pour l'isotype βIII-tubuline, important dans le cancer.

Chapitre 4. Découverte et conception d'agents ciblant le site todalam

Récemment, Muhlethaler et al. ont rapporté la conception rationnelle d'un nouvel inhibiteur de tubuline appelé "Todalam", identifié par criblage de fragments[21]. La caractéristique innovante du todalam est son site de liaison unique situé à l'interface de deux unités de tubuline (Figure R-8), à proximité du site de liaison de la pironétine et de la boucle de la sous-unité α-tubuline. Nous avons cherché à utiliser des outils de chimie computationnelle pour identifier de nouvelles molécules et de nouveaux échafaudages qui se lient à ce site, élargissant ainsi notre connaissance de la chimie de ce site de liaison et de la tubuline en général.

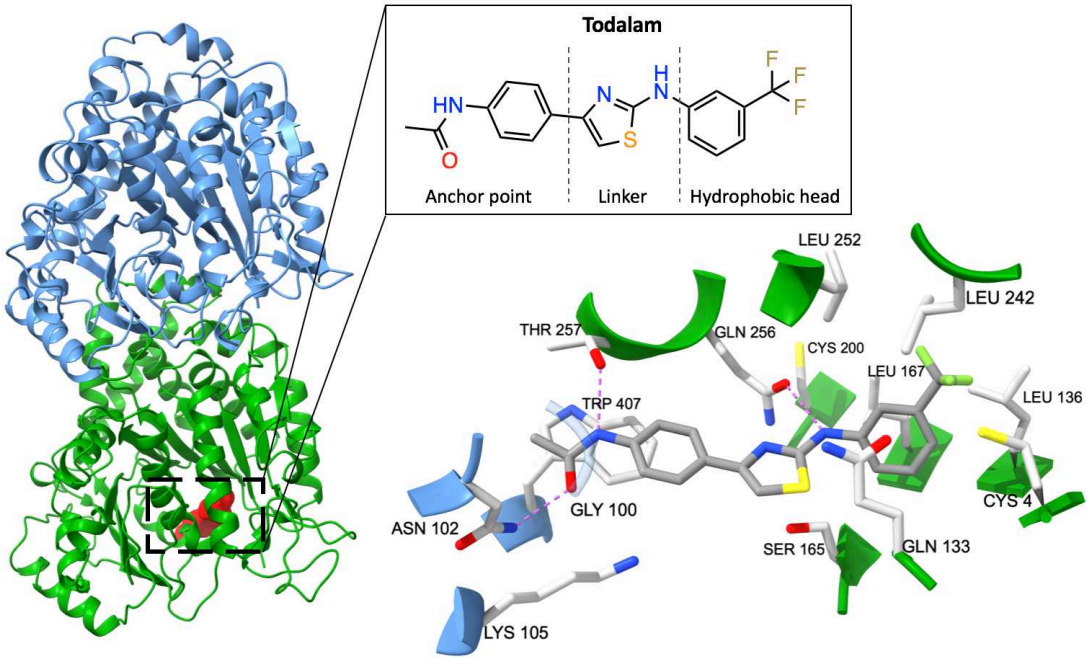


Figure R-8. Aperçu du site de liaison du todalam et de la structure du ligand du todalam

Ce projet a fait appel à plusieurs techniques chémoinformatiques, notamment la modélisation et le criblage pharmacophore, le docking protéine-ligand, la recherche de similarité et de sous-structure, la recherche de similarité de site de liaison et le docking covalent. PLANTS et AutoDock 4 a été utilisé pour les expériences de docking covalent et rigide, ainsi qu'AutoDock GPU[22].

Les données initiales se limitaient à une seule structure cristalline, PDB code 5SB7. La modélisation automatisée du pharmacophore a permis d'obtenir un modèle comportant huit caractéristiques. Le criblage pharmacophore a été effectué sur la bibliothèque HTS Enamine, la bibliothèque HTS Ambinter et une bibliothèque interne de composés triazoles. Les molécules identifiées ont été obtenues et soumises à la cristallographie aux rayons X et à des essais de déplacement de microtubules. Sur les 57 molécules trouvées par le criblage virtuel, 21 ont été confirmées comme se liant au site todalam par cristallographie aux rayons X, et trois d'entre elles ont également montré des effets dépolymérisants significatifs sur les microtubules.

En raison de la proximité d'un résidu cystéine à la position de liaison du ligand, nous avons cherché à concevoir un liant covalent ciblant ce résidu. La création réussie d'un ligand réagissant de manière covalente avec le résidu cystéine dans ce site constituerait un composé inédit, qui pourrait servir de sonde moléculaire utile pour étudier la dynamique des microtubules *in vivo*.

La réactivité du résidu cystéine a été évaluée à l'aide de diverses méthodes. Bien que toutes les méthodes indiquent une faible réactivité, nous avons émis l'hypothèse qu'un fragment organique suffisamment réactif, s'il est maintenu à proximité du résidu par un ligand bien lié,

pourrait potentiellement réagir avec la cystéine. Une analyse exhaustive de la littérature a permis de dresser une liste de 32 fragments réactifs possibles.

Nous avons identifié une liste d'échafaudages uniques dont il a été prouvé expérimentalement qu'ils se liaient bien au site todalam. En particulier, les molécules contenant l'échafaudage triazole ont démontré de bonnes propriétés de liaison. Profitant de la synthèse facile et rapide des triazoles par chimie click, nous avons recherché des fragments contenant un fragment réactif d'un côté et une triple liaison de l'autre pour synthétiser des triazoles intéressants.

Nous avons découvert quelques fragments achetables convenant à la chimie click et avons dénombré des molécules de triazole avec ces fragments. Ces molécules ont été dockées à l'aide d'AutoDock et réévaluées à l'aide du modèle pharmacophore complet contenant huit caractéristiques pharmacophoriques. Un docking covalent a également été réalisé pour simuler la façon dont la pose changerait après une éventuelle réaction covalente. Parallèlement, nous avons également recherché des échafaudages de liaison connus contenant un fragment réactif convenablement situé à l'aide de l'outil SciFinder. Les molécules proposées ont été achetées, synthétisées et testées. Quelques molécules conçues se sont effectivement liées au site. Cependant, aucune réaction covalente n'a encore été observée.

Grâce à ce projet, nous avons identifié 21 molécules cibles qui se lient au site de liaison de todalam (Figure R-9). Malgré les tentatives infructueuses de liaison covalente, ces résultats contribuent de manière significative à notre compréhension de l'inhibition de la polymérisation de la tubuline. Les efforts futurs pourraient impliquer une optimisation plus poussée des molécules les plus prometteuses afin d'améliorer les propriétés de déstabilisation des microtubules et d'obtenir une liaison covalente avec le résidu cystéine du site de todalam.

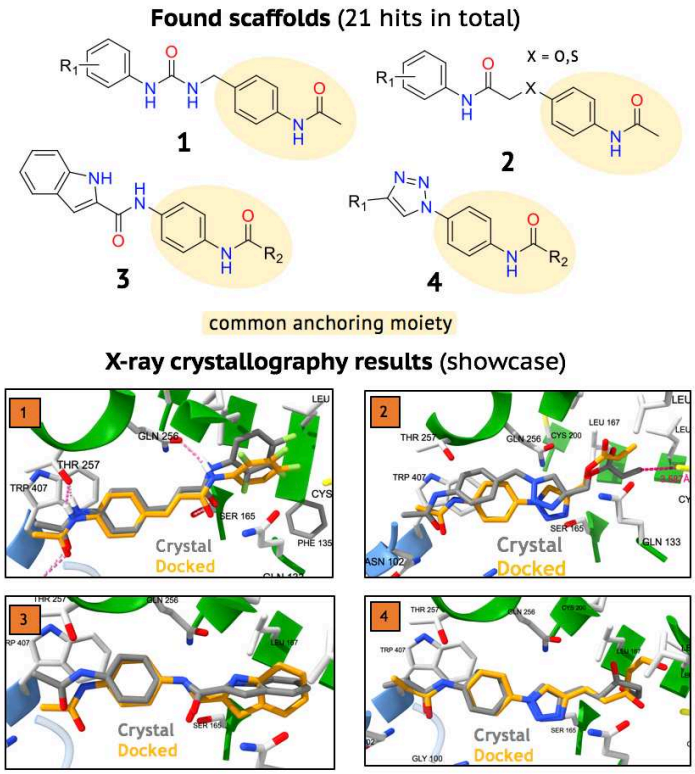


Figure R-9. Aperçu des molécules obtenues

Chapitre 5. Conception de novo d'agents ciblant le site de la colchicine en utilisant l'approche QSAR inverse

Le site de liaison de la colchicine, une poche profonde située à l'interface entre les sous-unités α et β de l'hétérodimère tubuline, joue un rôle important dans les changements de conformation qui se produisent pendant l'assemblage des microtubules (Figure R-10). La liaison d'un ligand à ce site inhibant ainsi la formation des microtubules. La plupart des agents connus se liant au site de la colchicine ont des profils pharmacologiques peu satisfaisants et sont des dérivés d'un nombre limité d'échafaudages, ce qui limite la diversité[23]. L'objectif de ce projet était de concevoir de nouveaux agents de liaison au site de liaison de la colchicine en mettant l'accent sur la diversification des échafaudages connus en utilisant la méthodologie QSAR inverse[24].

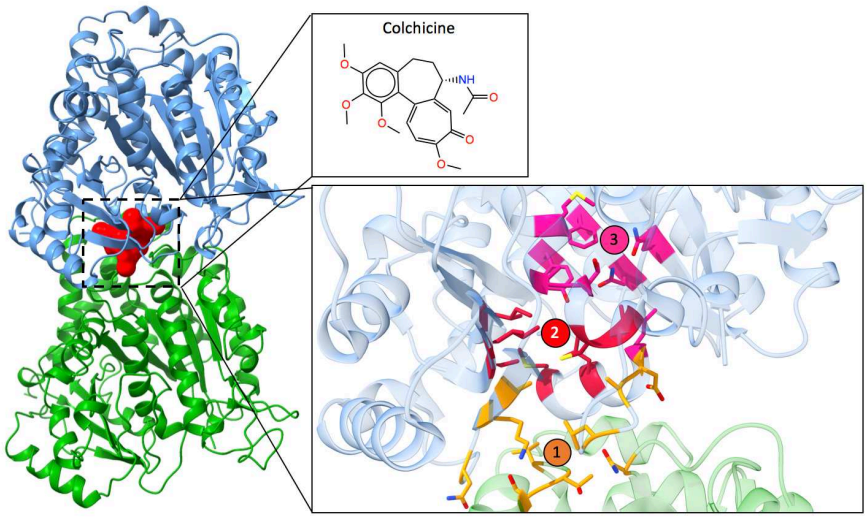


Figure R-10. Le site de liaison de la colchicine

Au lieu de prédire la propriété d'une molécule donnée sur la base des valeurs des descripteurs moléculaires, la QSAR inverse concerne la conception de nouvelles molécules ayant des propriétés spécifiques souhaitées en s'appuyant sur les relations entre les structures moléculaires et leurs activités établies par les modèles QSAR (Figure R-11). Dans cette étude, nous avons utilisé un auto-encodeur conditionnel basé sur l'attention de Bort et al[24]. Ce type de modèle apprend la distribution des composés dans l'espace chimique défini par les descripteurs moléculaires et permet l'échantillonnage de molécules à partir de graines spécifiées dans cet espace latent. Le modèle autoencodeur a été entraîné sur l'ensemble de la base de données ChEMBL (v 26, 1,8 million de composés) pour apprendre à reconstruire des structures moléculaires valides à partir de descripteurs moléculaires.

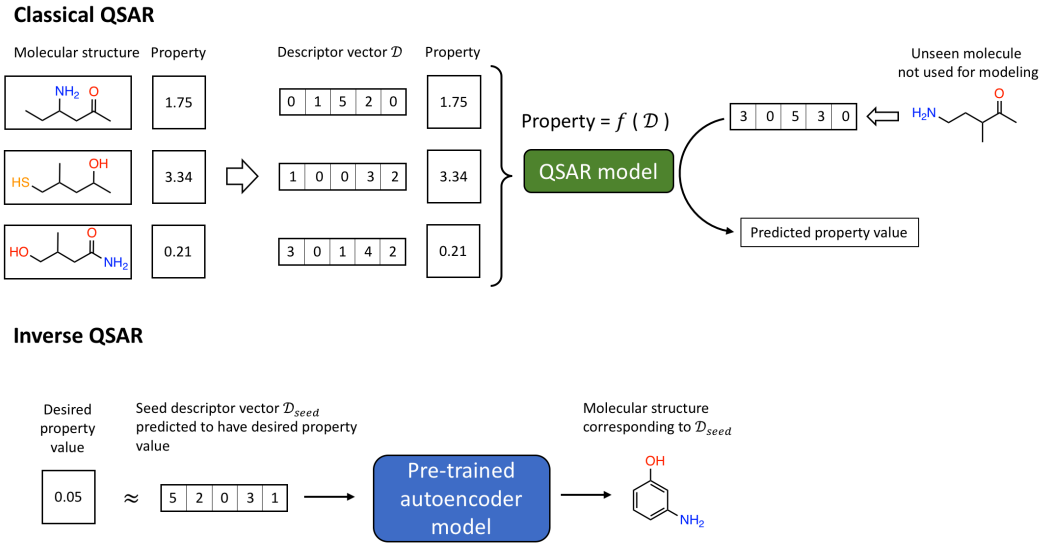


Figure R-11. Aperçu des différences entre les approches QSAR classiques et les approches QSAR inverses

Pour déterminer les valeurs des descripteurs ISIDA qui sont en corrélation avec l'activité souhaitée, un modèle QSAR de régression a été entraîné sur les valeurs IC50 de 379 composés ciblant le site de la colchicine contre les cellules HeLa[25]. L'algorithme de la forêt d'arbres décisionnels a été employé pour l'entraînement, et les données ont été standardisées comme décrit précédemment. Un modèle à validation croisée performant avec R^2 de 0.68 a été utilisé pour cribler la collection de composés de la bibliothèque phénotypique d'Enamine. Le domaine d'applicabilité du modèle a été évalué par la méthode de la boîte englobante, ce qui a permis de réduire le nombre de composés à 421 sur les 5760 initiaux. Ces composés ont été standardisés et codés de la même manière que les données d'apprentissage, et leurs valeurs IC50 prédites contre les cellules HeLa ont été déterminées. Les 15 molécules présentant les meilleures valeurs prédites de IC50 ont été sélectionnées pour un examen plus approfondi.

Les 15 composés les plus actifs prédits ont servi de molécules de graines pour la génération *de novo*. L'introduction d'un petit bruit aléatoire dans la représentation des descripteurs a permis la variabilité, ce qui a abouti à la génération de 782 composés uniques, présentant tous des caractéristiques structurelles compatibles avec la liaison au site de la colchicine et des valeurs IC50 prédites élevées contre les cellules HeLa.

Pour classer par ordre de priorité certains des composés générés, on a réalisé un docking protéine-ligand contre le site de la colchicine de la structure PDB de la tubuline 1SA0. Les molécules qui ont obtenu des scores de docking supérieurs à ceux du ligand natif ont été sélectionnées, ce qui a permis d'obtenir 50 molécules ayant obtenu les meilleurs scores.

Étant donné qu'aucun de ces composés ne pouvait être obtenu directement auprès d'un fournisseur et que la synthèse organique était irréalisable, une recherche de similarité a été effectuée pour trouver des molécules analogues dans les bibliothèques de composés achetables, en utilisant un score de Tanimoto de 0.8 ou plus comme critère de similarité (Figure R-12). Malheureusement, les composés achetés, qui sont des analogues structurels proches des molécules générées *de novo*, ne présentaient pas de liaison au site de la colchicine ni d'activité de dépolymérisation des microtubules, testés respectivement par cristallographie aux rayons X et par essai de dépolymérisation des microtubules *in vitro*.

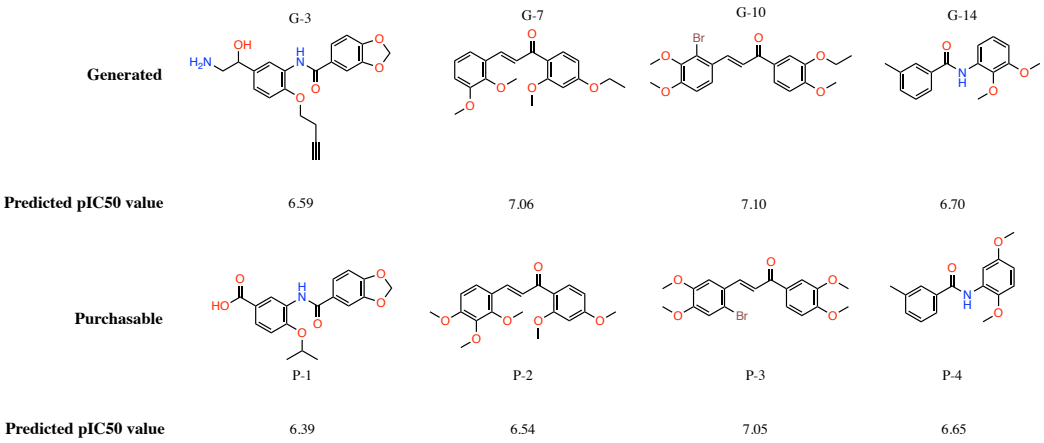


Figure R-12. Comparaison de la structure moléculaire et des valeurs pIC50 prédites contre les cellules HeLa des molécules générées de novo et de leurs analogues disponibles à l'achat

Chapitre 6. Evaluation de l'applicabilité du concept d'apprentissage par transfert pour la modélisation QSAR

Les modèles QSAR permettent d'établir une corrélation entre les structures chimiques et leurs propriétés. Une étape critique de ce processus consiste à représenter les molécules sous forme de vecteurs numériques à l'aide de divers descripteurs moléculaires. La sélection des descripteurs est essentielle pour développer des modèles prédictifs performants, mais l'ensemble optimal de descripteurs n'est généralement pas connu a priori. Les choix sont souvent dictés par l'intuition d'un expert ou par des pipelines complexes de sélection d'ensembles de descripteurs. Malgré cela, l'optimalité du jeu de descripteurs final n'est pas garantie. L'objectif de ce projet était de vérifier si l'apprentissage par transfert, un concept qui gagne du terrain dans le domaine de l'apprentissage profond, pouvait être exploité pour apprendre des représentations moléculaires significatives pour une tâche à accomplir.

L'apprentissage par transfert implique généralement des étapes de pré-entraînement, peaufinage et de prédiction (Figure R-13). Le pré-entraînement consiste à apprendre des représentations moléculaires utiles à partir d'un vaste ensemble de données moléculaires de manière autosupervisée. Un plus petit ensemble de données est ensuite utilisé pour le peaufinage, où les représentations apprises sont ajustées de bout en bout pour faire correspondre la structure moléculaire d'entrée à la propriété cible. Ces représentations affinées peuvent potentiellement améliorer les performances prédictives lorsqu'elles sont extraites de modèles affinés en vue d'une utilisation ultérieure. Pour cette étude, nous avons utilisé GROVER de Rong et al.[26], un outil qui met en œuvre toutes les étapes du pipeline d'apprentissage par transfert avec une interface pratique.

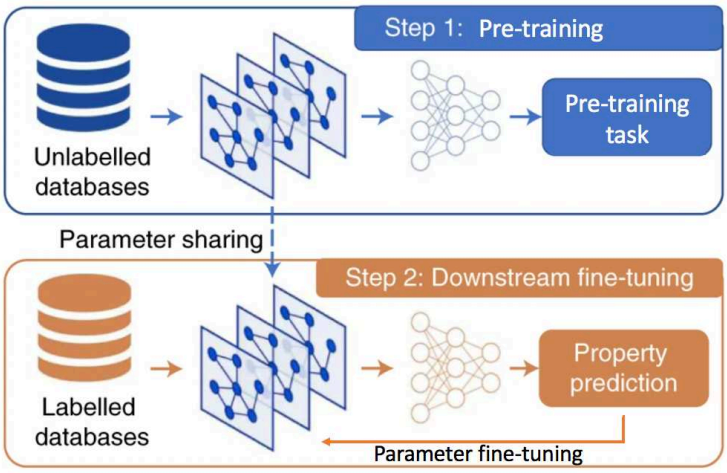


Figure R-13. Aperçu du concept d'apprentissage par transfert

Nous avons trouvé les paramètres optimaux de la méthode et nous avons les appliqué à une étude de cas axée sur la classification de l'activité des composés ciblant le site de la colchicine. Nous avons utilisé un ensemble de données comprenant 766 points de données de structures moléculaires avec des étiquettes d'activité correspondantes pour une tâche de classification binaire[25].

La méthodologie d'apprentissage par transfert de GROVER démontre des performances comparables à celles de la méthode de pointe de la machine à vecteur de support sur les descripteurs ISIDA (Figure R-14). Elle surpasse la référence de base établie par une forêt d'arbres decisionelle sur les descripteurs physico-chimiques. Par conséquent, l'apprentissage par transfert présente des performances compétitives par rapport aux principales méthodologies actuelles et peut être encore optimisé en combinant les représentations apprises avec d'autres descripteurs. Les travaux futurs pourraient explorer un schéma de validation croisée plus robuste afin d'améliorer encore cette approche.

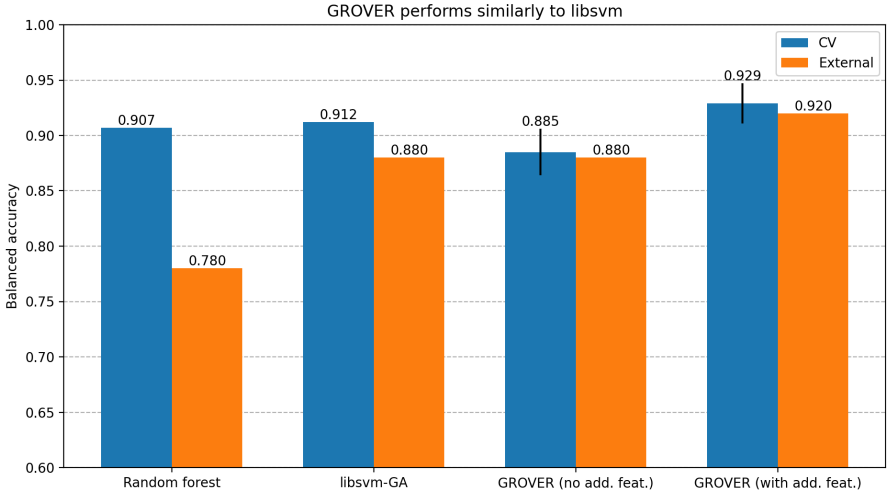


Figure R-14. Performance de classification de quatre modèles différents sur des données de structure-activité associées au site de la colchicine

Chapitre 7. Exploration de poches de liaison cryptiques dans la tubuline à l'aide de simulations de dynamique moléculaire accélérées par la gaussienne

Les protéines sont des entités dynamiques qui subissent continuellement des changements de conformation. Ce dynamisme est étroitement lié à la fonction biologique des protéines. La nature dynamique des protéines donne lieu à des poches de liaison cryptiques. Les poches de liaison cryptiques sont des sites de liaison dans les protéines qui ne sont pas toujours présents dans la structure de la protéine, mais qui émergent pendant une courte période en raison de changements de conformation intrinsèques à la protéine ou induits par un ligand (Figure R-15). Elles offrent des possibilités uniques pour le développement de médicaments[27].

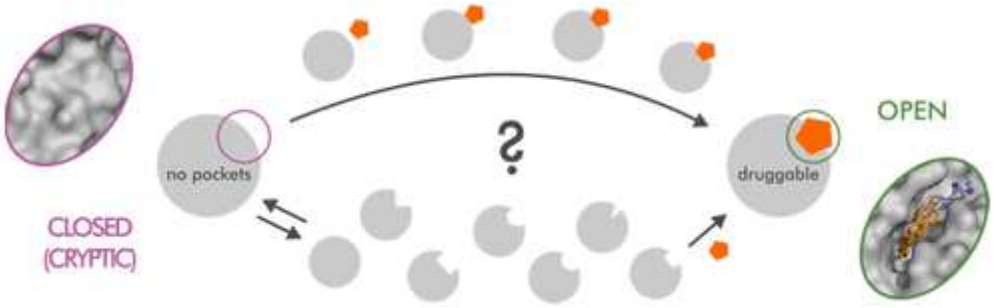


Figure R-15. Représentation schématique de deux mécanismes possibles de formation de poches cryptiques (adapté de Kuzmanic et al.[27])

Les simulations computationnelles jouent un rôle crucial dans l'identification de ces poches cryptiques. Les méthodes computationnelles telles que les simulations de dynamique moléculaire peuvent capturer les changements de conformation de la protéine au fil du temps, révélant ainsi les poches transitoires. Cependant, les approches conventionnelles des simulations de dynamique moléculaire ne peuvent pas échantillonner efficacement de nombreux changements de conformation biologiquement pertinents lorsqu'il s'agit de systèmes qui ont des barrières énergétiques élevées sur leurs surfaces d'énergie potentielle, car le système peut rester piégé dans des minima locaux pendant de longues périodes[27].

Une façon de surmonter cette limitation est d'utiliser une technique d'échantillonnage améliorée, telle que la dynamique moléculaire accélérée par la gaussienne (GaMD)[28]. La GaMD introduit un potentiel d'accélération harmonique sur la surface d'énergie potentielle du système, ce qui lisse efficacement le paysage énergétique et facilite les transitions conformationnelles[28]. L'objectif de ce projet était donc d'utiliser la GaMD pour simuler la dynamique conformationnelle de la protéine tubuline et d'identifier les poches cryptiques potentielles à sa surface.

À cette fin, nous avons préparé le système de simulation en récupérant la structure tridimensionnelle de l'hétérodimère α,β-tubuline dans le dossier 7E4Z de la Protein Data Bank et en supprimant toutes les molécules de solvant et les petites molécules organiques, à l'exception du GTP. Nous avons conservé les ions manganèse. Nous avons supprimé toutes les chaînes à l'exception des chaînes C et D, correspondant aux parties α-tubuline et β-tubuline de l'hétérodimère. La protéine préparée a été immergée dans une boîte cubique remplie de molécules d'eau TIP3P équilibrées. Pour neutraliser le système, certaines molécules d'eau ont été remplacées par des ions Na^+ ou Cl^-. Le module LEaP du logiciel AMBER18 a été utilisé pour corriger les problèmes de protonation et les atomes manquants.

Après la préparation du système, la minimisation de l'énergie a été effectuée par étapes. Après la minimisation, le système a été équilibré. Ensuite, nous avons effectué trois cycles de production de 1 μs. Tous les calculs ont été effectués à l'aide du code PMEMD (Particle Mesh Ewald Molecular Dynamics) du logiciel AMBER18 dans sa version CUDA, en utilisant le champ de force AMBER ff14SB.

Après la simulation, nous avons calculé les valeurs RMSD et RMSF. Cette analyse a montré que le système a atteint une stabilité structurelle après une brève période de stabilisation.

Nous avons ensuite utilisé l'analyse en composantes principales pour déterminer les principales variations structurelles du système étudié. Nous avons ensuite projeté chaque instantané de trajectoire MD sur les deux composantes principales présentant la variance expliquée la plus élevée. Nous avons montré que chacune des trois courses GaMD explorait des sections distinctes de l'espace conformationnel.

Ensuite, nous avons identifié des conformations distinctes de la tubuline en classant les structures similaires de l'ensemble de la trajectoire de simulation en 15 groupes différents. Des structures représentatives de chaque groupe ont ensuite été utilisées pour identifier les points chauds à l'aide du serveur web FTMap[29]. Notre étude a révélé quatre nouvelles poches cryptiques, qui n'ont pas été identifiées auparavant par cristallographie aux rayons X ou par des simulations MD classiques, et qui ne sont pas connues pour héberger des ligands connus (Figure R-16).

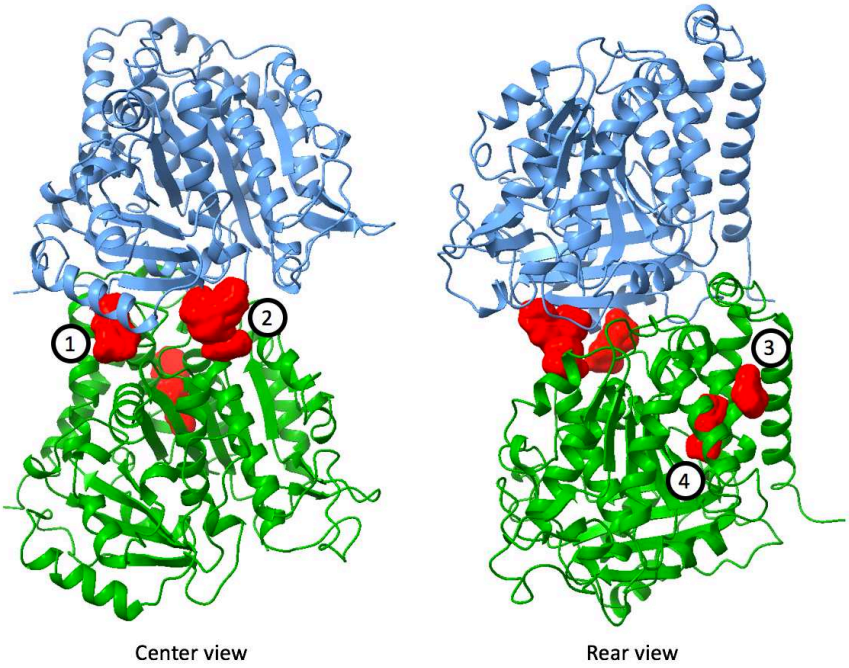


Figure R-16. Localisation des poches cryptiques identifiées par FTMap

Ces poches nouvellement identifiées peuvent constituer des cibles prometteuses pour les criblages de composés et les études d'amarrage. Toutefois, des recherches supplémentaires sont nécessaires pour comprendre pleinement ces sites de liaison. La pharmacocinétique des poches doit être évaluée pour toutes les poches identifiées. À l'avenir, nous avons l'intention d'explorer davantage ces nouvelles poches, en nous concentrant particulièrement sur leur potentiel en tant que cibles médicamenteuses.

Chapitre 8. Développement d'une application graphique pour l'analyse automatique des images de diffraction des fibres de microtubules.

L'interaction de petites molécules organiques avec la tubuline peut avoir un impact significatif sur les caractéristiques physiques des microtubules, notamment le rayon des microtubules, le nombre de protofilaments constitutifs et la longueur moyenne des monomères de tubuline dans la structure tubulaire. La diffraction des fibres de microtubules aux rayons X constitue une approche permettant d'étudier quantitativement ces changements[30].

La diffraction des fibres, une technique largement utilisée en biologie structurale, permet d'explorer la structure des filaments biologiques, en particulier des microtubules, dans des conditions physiologiques sans qu'il soit nécessaire de les fixer, de les cristalliser ou de les congeler[30]. Les échantillons étudiés sont souvent disposés, naturellement ou artificiellement, en une ligne de structures filamenteuses présentant un certain degré de régularité, de périodicité ou de structure hélicoïdale. La diffraction des fibres offre une compréhension structurelle plus

complète que les autres techniques basées sur la diffraction des rayons X, en fournissant des informations détaillées sur la périodicité longitudinale et l'espacement latéral des molécules à l'intérieur d'un filament arrangé[30].

Dans une expérience typique de diffraction de fibres de microtubules, les images de diffraction de fibres de rayons X sont capturées dans des lignes de faisceaux de rayonnement synchrotron[31] (Figure R-17). Les rayons X diffractés sont recueillis par un détecteur, ce qui permet d'obtenir une seule image de diffraction par exposition au faisceau. En général, 16 à 24 images de diffraction sont collectées à partir de 4 à 6 échantillons indépendants pour une expérience donnée. En outre, des images de fond sont obtenues dans les mêmes conditions à l'aide d'une solution tampon. L'étape finale est l'étalonnage spatial, réalisé à l'aide de la diffraction de poudre d'Ag-Behenate, qui prend en compte une diffusion élastique et fournit les distances au centre du faisceau des intensités de diffraction des vecteurs de diffusion[31].

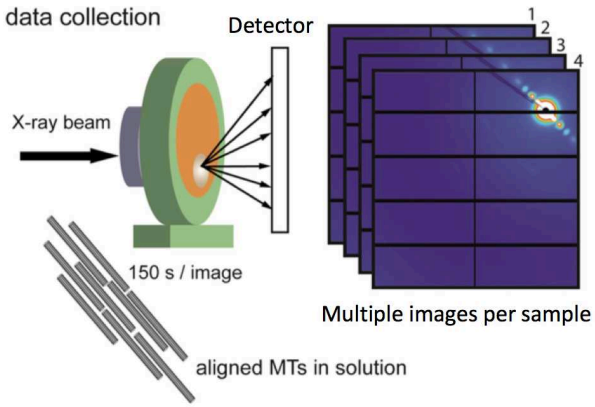


Figure R-17. Montage expérimental d'une expérience de diffraction de fibres de microtubules

Ainsi, les résultats d'une expérience typique de diffraction sur fibre de microtubules comprennent un fichier avec les paramètres du détecteur, un fichier d'étalonnage de l'expérience de diffraction sur poudre d'Ag-Behenate, les images de la solution tampon et, surtout, les diagrammes de diffraction des rayons X obtenus à partir de microtubules alignés en présence d'agents testés ciblant les microtubules[31].

L'analyse des images produites par les expériences de diffraction des rayons X sur les fibres de microtubules par flux de cisaillement implique le traitement des images, l'intégration numérique des données visuelles et l'approximation fonctionnelle des résultats de l'intégration (Figure R-18). En règle générale, cette analyse est manuelle et exige beaucoup de travail, ce qui nécessite l'utilisation de plusieurs applications spécialisées de traitement d'images et de données statistiques. La nécessité d'une analyse fastidieuse limitait le nombre d'expériences que les chercheurs pouvaient réaliser pendant le temps d'expérimentation très limité et coûteux dont ils disposaient sur les installations synchrotron.

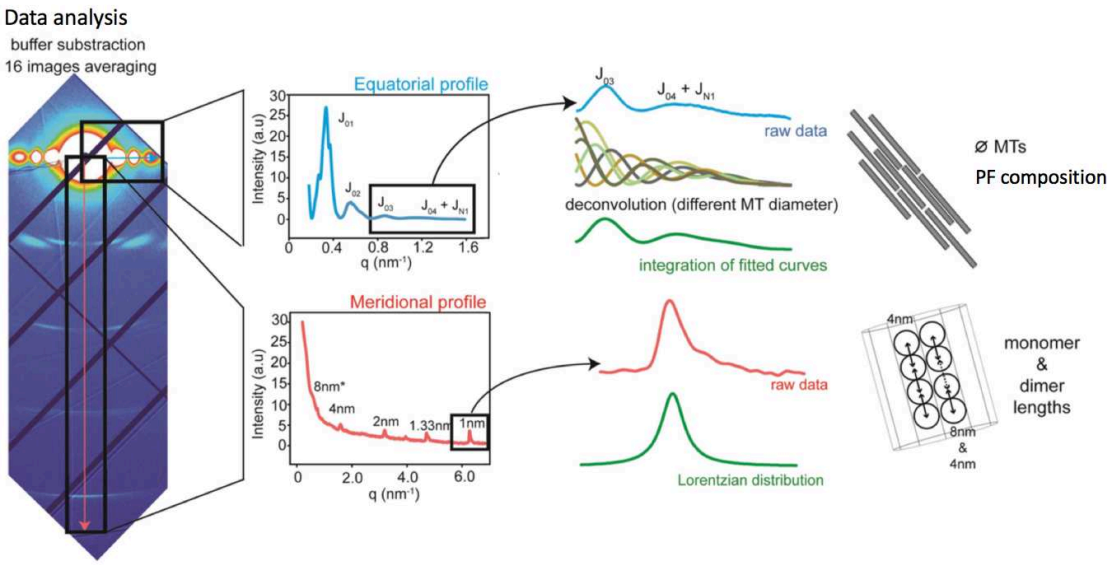


Figure R-18. Résumé schématique de l'analyse des diagrammes de diffraction et des informations qui peuvent en être déduites

Par conséquent, l'objectif de ce projet était de développer un programme d'analyse automatisée des images de diffraction des rayons X des fibres de microtubules avec une interface graphique simple d'utilisation qui augmenterait la vitesse de l'analyse et, par conséquent, permettrait de réaliser plus d'expériences plus rapidement, augmentant ainsi le rendement des expériences de diffraction des fibres de microtubules par les rayons X de l'écoulement cisaillé[30,31].

Nos efforts ont abouti à la création de FiDAT (Fiber Diffraction Analysis for microTubules), une application autonome qui rationalise les trois étapes de l'analyse des données expérimentales. Cette application est conçue pour être simple d'utilisation, ne nécessite pas d'expertise préalable et accélère de manière significative le débit de ces expériences. FiDAT est équipé pour gérer toutes les étapes de l'analyse des résultats des expériences de diffraction sur fibre, ce qui en fait une solution complète pour les chercheurs dans ce domaine. Le logiciel est fourni avec des paramètres par défaut préconfigurés, garantissant un traitement des données et un ajustement du modèle fiableы immédiatement. Cependant, nous avons également répondu aux besoins des utilisateurs experts en leur offrant la possibilité de personnaliser les paramètres pour chaque étape de l'analyse.

L'efficacité de FiDAT a été validée par nos collègues du consortium TubInTrain. L'efficacité du processus d'analyse automatisé de FiDAT a non seulement facilité l'exécution d'expériences supplémentaires, mais a également permis aux chercheurs d'intégrer rapidement des idées nouvelles dans leur conception expérimentale. Cette adaptabilité en temps réel a permis une prise de décision sur place et une vérification immédiate des hypothèses, éliminant ainsi la nécessité d'une analyse manuelle longue et laborieuse.

À l'avenir, nous prévoyons d'améliorer encore FiDAT en résolvant des bogues mineurs liés au traçage d'images et en incorporant des fonctions d'enregistrement supplémentaires, comme l'ont suggéré nos collègues. Après cela, nous prévoyons également de publier FiDAT en tant que logiciel libre, rendant ainsi cet outil puissant accessible à la communauté scientifique dans son ensemble. En outre, la direction du synchrotron Alba à Barcelone, en Espagne, a manifesté un intérêt considérable pour l'intégration de FiDAT dans son ordinateur central. Une fois les améliorations prévues mises en œuvre, FiDAT sera mis à la disposition de tous les chercheurs du synchrotron en tant qu'outil par défaut, ce qui renforcera encore son rôle en tant qu'atout indispensable dans les expériences de diffraction des fibres de microtubules.

Conclusion

Cette thèse de recherche a abouti à la découverte d'une série de nouveaux composés qui ciblent différents sites de liaison sur la protéine tubuline, ouvrant la voie à des approches innovantes pour moduler la polymérisation de la tubuline. La première découverte importante comprend l'identification de deux molécules qui présentent une affinité possible pour le site de liaison de la maytansine. En outre, trois fragments actifs ont été découverts, démontrant un puissant effet inhibiteur sur la polymérisation des microtubules. Deux nouvelles molécules rentables ont été trouvées pour se lier au site de liaison de la colchicine, l'une d'entre elles montrant une spécificité pour l'isotype βIII-tubuline. Une découverte importante est l'identification de 21 molécules qui ciblent le site todalam, jusqu'à présent peu étudié, ce qui enrichit considérablement notre compréhension de la diversité chimique de ce site de liaison. Parallèlement à ces découvertes, de nouvelles méthodologies d'apprentissage profond ont été développées et peuvent être utiles pour les recherches futures dans ce domaine. Par ailleurs, une analyse conformationnelle du système hétérodimère de la tubuline a été réalisée à l'aide d'une technique de dynamique moléculaire à échantillonnage amélioré, mettant en évidence quatre poches de liaison cryptiques précédemment inconnues à la surface de la tubuline. En outre, une interface utilisateur graphique (GUI) a été développée pour automatiser et rationaliser le processus d'analyse de la diffraction des microtubules. Les futurs efforts de recherche devraient se concentrer sur l'optimisation de ces composés prometteurs et sur l'affinement des méthodologies de modélisation utilisées dans cette étude.

Références

1. Steinmetz, M. O. & Prota, A. E. Microtubule-Targeting Agents: Strategies To Hijack the Cytoskeleton. Trends Cell Biol. 28, 776–792 (2018).

2. Čermák, V. et al. Microtubule-targeting agents and their impact on cancer treatment. Eur. J. Cell Biol. 99, 151075 (2020).

3. Macalino, S. J. Y., Gosu, V., Hong, S. & Choi, S. Role of computer-aided drug design in modern drug discovery. Arch. Pharm. Res. 38, 1686–1701 (2015).

4. Stumpfe, D. & Bajorath, J. Similarity searching. WIREs Comput. Mol. Sci. 1, 260–282 (2011).

5. Horvath, D., Koch, C., Schneider, G., Marcou, G. & Varnek, A. Local neighborhood behavior in a combinatorial library context. J. Comput. Aided. Mol. Des. 25, 237–252 (2011).

6. Ehrlich, H.-C. & Rarey, M. Systematic benchmark of substructure search in molecular graphs - From Ullmann to VF2. J. Cheminform. 4, 13 (2012).

7. Kochev, N., Monev, V. & Bangov, I. Searching Chemical Structures. in Chemoinformatics 291–318 (Wiley-VCH Verlag GmbH & Co. KGaA). doi:10.1002/3527601643.ch6.

8. Cherkasov, A. et al. QSAR Modeling: Where Have You Been? Where Are You Going To? J. Med. Chem. 57, 4977–5010 (2014).

9. Seidel, T., Wolber, G. & Murgueitio, M. S. Pharmacophore Perception and Applications. in Applied Chemoinformatics 259–282 (Wiley-VCH Verlag GmbH & Co. KGaA, 2018). doi:10.1002/9783527806539.ch6f.

10. Wolber, G. & Langer, T. LigandScout: 3-D Pharmacophores Derived from Protein-Bound Ligands and Their Use as Virtual Screening Filters. J. Chem. Inf. Model. 45, 160–169 (2005).

11. Kolodzik, A., Schneider, N. & Rarey, M. Structure-Based Virtual Screening. in Applied Chemoinformatics 313–331 (Wiley-VCH Verlag GmbH & Co. KGaA, 2018). doi:10.1002/9783527806539.ch6h.

12. Li, W. et al. C3 ester side chain plays a pivotal role in the antitumor activity of Maytansinoids. Biochem. Biophys. Res. Commun. 566, 197–203 (2021).

13. Mendez, D. et al. ChEMBL: towards direct deposition of bioassay data. Nucleic Acids Res. 47, D930–D940 (2019).

14. Berman, H. M. The Protein Data Bank. Nucleic Acids Res. 28, 235–242 (2000).

15. Korb, O., Stuztle, T. & Exner, T. PLANTS: Application of Ant Colony Optimization to Structure-Based Drug Design. in Ant Colony Optimization and Swarm Intelligence (eds. Dorigo, M., Gambardella, L. M. & Birattari, M.) 247–259 (Springer, 2006). doi:https://doi.org/10.1007/11839088_22.

16. Yang, J. et al. Pironetin reacts covalently with cysteine-316 of α-tubulin to destabilize microtubule. Nat. Commun. 7, 12103 (2016).

17. Coulup, S. K. & Georg, G. I. Revisiting microtubule targeting agents: α-Tubulin and the pironetin binding site as unexplored targets for cancer therapeutics. Bioorganic Med. Chem. Lett. 29, 1865–1873 (2019).
18. Gentile, F. et al. Deep Docking: A Deep Learning Platform for Augmentation of Structure Based Drug Discovery. ACS Cent. Sci. 6, 939–949 (2020).
19. Ruggiu, F., Marcou, G., Varnek, A. & Horvath, D. ISIDA Property-Labelled Fragment Descriptors. Mol. Inform. 29, 855–868 (2010).
20. Mühlethaler, T. et al. Comprehensive Analysis of Binding Sites in Tubulin. Angew. Chemie 133, 13443–13454 (2021).
21. Mühlethaler, T. et al. Rational Design of a Novel Tubulin Inhibitor with a Unique Mechanism of Action. Angew. Chemie Int. Ed. 61, (2022).
22. Santos-Martins, D. et al. Accelerating AutoDock 4 with GPUs and Gradient-Based Local Search. J. Chem. Theory Comput. 17, 1060–1073 (2021).
23. Du, T. et al. A novel orally active microtubule destabilizing agent S-40 targets the colchicine-binding site and shows potent antitumor activity. Cancer Lett. 495, 22–32 (2020).
24. Bort, W. et al. Inverse QSAR: Reversing Descriptor-Driven Prediction Pipeline Using Attention-Based Conditional Variational Autoencoder. J. Chem. Inf. Model. 62, 5471–5484 (2022).
25. López-López, E., Cerda-García-Rojas, C. M. & Medina-Franco, J. L. Tubulin Inhibitors: A Chemoinformatic Analysis Using Cell-Based Data. Molecules 26, 2483 (2021).
26. Rong, Y. et al. GROVER: Self-supervised Message Passing Transformer on Large-scale Molecular Data. Adv. Neural Inf. Process. Syst. 1–13 (2020).
27. Kuzmanic, A., Bowman, G. R., Juarez-Jimenez, J., Michel, J. & Gervasio, F. L. Investigating Cryptic Binding Sites by Molecular Dynamics Simulations. ACS Appl. Mater. Interfaces (2020) doi:10.1021/acs.accounts.9b00613.
28. Miao, Y., Feher, V. A. & McCammon, J. A. Gaussian Accelerated Molecular Dynamics: Unconstrained Enhanced Sampling and Free Energy Calculation. J. Chem. Theory Comput. 11, 3584–3595 (2015).
29. Kozakov, D. et al. The FTMap family of web servers for determining and characterizing ligand-binding hot spots of proteins. Nat. Protoc. 10, 733–755 (2015).
30. Kamimura, S., Fujita, Y., Wada, Y., Yagi, T. & Iwamoto, H. X-ray fiber diffraction analysis shows dynamic changes in axial tubulin repeats in native microtubules depending on paclitaxel content, temperature and GTP-hydrolysis. Cytoskeleton 73, 131–144 (2016).
31. Oliva, M. Á., Gago, F., Kamimura, S. & Díaz, J. F. Alternative Approaches to Understand Microtubule Cap Morphology and Function. ACS Omega (2022) doi:10.1021/acsomega.2c06926.

Chapter 1. Bibliographic overview

Humanity is facing several healthcare challenges, including cancer and neurodegenerative diseases. Cancer is a leading cause of death worldwide, accounting for nearly 10 million deaths in 2020[1]. The burden of cancer is projected to increase in the future, with a predicted 28.4 million new cancer cases and 13 million cancer-related deaths annually by 2040[1]. Similarly, neurodegenerative diseases, in which neurons in the brain can no longer function properly causing loss of cognitive function (dementia), also cause significant morbidity and mortality. For instance, the World Health Organization estimated that in 2019, 55.2 million people worldwide were living with dementia, predicting a one and a half times rise in this number by 2030[2]. Neurodegenerative diseases were estimated to cause 1.6 million deaths worldwide in 2019[2]. Currently, there are no definitive cures for either of these conditions.

The tubulin protein is a promising biological target for the development of therapies aimed at addressing the pathogenesis of both cancer and neurodegenerative diseases. Tubulin is present in all eukaryotic cells and plays a crucial role in regulating cell division and intracellular transport. Recent research efforts have focused on modulating the polymerization of tubulin as a way to influence the cellular processes involved in cancer and neurodegenerative diseases.

Currently, despite promising research potential, tubulin-targeting agents have only been approved by FDA as a treatment for cancer, and not neurodegenerative diseases. Despite their effectiveness against malignant tumours, these agents are associated with high cytotoxicity and severe side effects, including peripheral neuropathy, hair loss, nausea, and vomiting, which further limit their clinical use[3]. Furthermore, most of the developed agents were designed to target only a small number of binding sites on tubulin, while several other binding sites remain largely unexplored.

Given the limitations of existing tubulin-targeting agents, it is promising to develop novel and potent agents to overcome these challenges. The eventual aim is to develop more effective drugs for cancer and to explore their potential use in developing treatments for neurodegenerative diseases. Additionally, there is promise in designing new molecular probes to study tubulin behavior both *in vitro* and *in vivo*.

1.1. The structural dynamics of tubulin polymerization

1.1.1. Structure of an individual α,β-tubulin unit

The tubulin protein is a heterodimeric globular complex comprising two subunits, α-tubulin and β-tubulin. Each of these subunits weighs approximately 50 kDa, resulting in a heterodimer with a total molecular weight of approximately 100 kDa[4]. This heterodimer serves as the fundamental building block of microtubules, which are integral to the cytoskeleton in all eukaryotic cells[4]. The α-tubulin and β-tubulin subunits are structurally similar and connect in a head-to-tail manner through non-covalent, longitudinal interactions[5]. Each subunit encases a core of two β sheets surrounded by α helices, and they both consist of 445-450 amino acids[4,6]. Figure 1 shows the amino acid sequence of both tubulin subunits, labeling fragments of the sequence by the structural element they form. Throughout this thesis, we will refer to this designation when referring to specific structural elements of the tubulin protein.

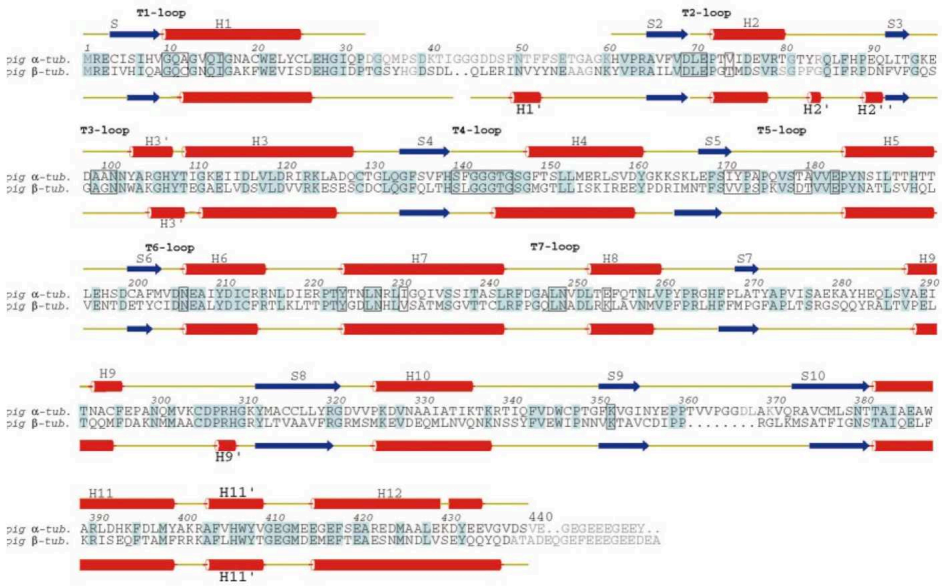


Figure 1. The amino acid sequence of both tubulin subunits (adapted from Löwe et al. [7])

Both α-tubulin and β-tubulin subunits have a binding site for guanosine triphosphate (GTP), a molecule with a crucial role in energy transfer, as it stores and releases energy through the breakdown (hydrolysis) of its high-energy phosphate bonds[5]. In the α-tubulin subunit, the bound GTP is non-exchangeable and non-hydrolysable, making it an intrinsic part of the heterodimeric structure[5]. However, β-tubulin can bind either GTP or GDP and is the site of GTP hydrolysis, which is important for the protein's biological function[5,8]. Figure 2a shows the quaternary structure of the tubulin protein with the nucleotide binding site highlighted, while Figure 2b has the structural elements colored by the secondary structure as designated in Figure 1.

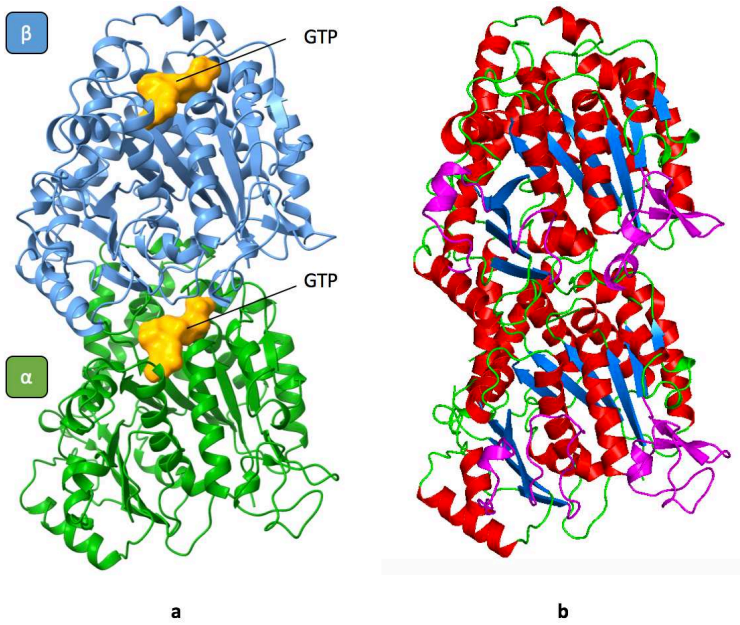


Figure 2. (a) The structure of a single α,β-tubulin heterodimer; (b) The heterodimer's structural elements highlighted according to designation in Figure 1. Red – α-helices, blue – β-sheets, green – loops, purple – structurally important loops.

Multiple isoforms of the tubulin heterodimers exist, as the human genome contains several α- and β-tubulin genes[5]. The amino acid sequence of tubulin is highly conserved throughout evolution, indicating that this sequence has remained largely unchanged over time, underscoring its biological importance and functionality[4]. At least six different types of tubulin isotypes are found in mammals, each performing subtly different functions in cells and tissues. Most differences between isotypes are localized within the last 15 residues of the sequences[4,9].

The α-β tubulin heterodimer undergoes post-translational modifications executed by a set of enzymes, which include polyglutamylation, polyglycylation, acetylation, and detyrosination[9]. Post-translational modifications occur after protein biosynthesis and often take place in the endoplasmic reticulum and the Golgi apparatus[9]. They are key mechanisms to increase proteomic diversity and play a fundamental role in functional proteomics by regulating protein activity, localization, and interaction with other cellular molecules, such as other proteins, nucleic acids, lipids, and cofactors[6,9].

Mutations in both α- and β-tubulin genes have implications in both cancer and neurodegeneration research[9–11]. As such, cancer cells are well-documented to develop resistance to tubulin-targeting chemotherapy treatment by upregulating specific tubulin isotypes, particularly βIII-tubulin[12]. At the same time, tubulin gene mutations disrupt normal neurodevelopment and are associated with a range of neurodevelopment disorders[11].

1.1.2. Proto-filament formation mechanism

Tubulin naturally aggregates into long, hollow, cylindrical biopolymers known as microtubules. Microtubules are composed of 13 parallel protofilaments, each of which is a longitudinal assembly of αβ-tubulin heterodimers arranged in a head-to-tail manner, ultimately folded into a tubular structure and held together by longitudinal and lateral contacts[5,6]. They are a crucial component of a cell's cytoskeleton, playing an indispensable role in such biological processes as cell division and long-distance intracellular transport (e.g., along axons and dendrites in neurons). In the former, they are integral in forming the mitotic spindle, which aids in chromosome segregation during mitosis. In the latter, they provide the highways along which vesicles, organelles, and other cellular components move to facilitate normal cell functioning.

The formation of a microtubule starts with the longitudinal assembly of tubulin dimers, forming extended chains[6]. These chains, referred to as protofilaments, act as the building blocks of the microtubule structure.

The polymerization of tubulin into protofilaments is a GTP-dependent process. Only tubulin bound to GTP, rather than GDP, can undergo such polymerization[5]. This process starts with nucleation, a rate-limiting step wherein tubulin dimers randomly form proto-filaments through Brownian motion. Proto-filaments are elongated chains of tightly linked dimers, arrayed in a head-to-tail manner, meaning that the α-tubulin of one dimer is non-covalently bound to the β-tubulin of the preceding one (Figure 3)[6]. Consequently, a proto-filament always exposes α-tubulin at one side (designated the minus (-) end) and β-tubulin at the opposite side (the plus (+) end). An interesting aspect of this polymerization is that the addition of new dimers typically occurs more rapidly at the plus end[6].

During tubulin polymerization, the GTP bound to β-tubulin undergoes hydrolysis as the polymer extends. This hydrolysis can occur concurrently with the addition of a new dimer or slightly afterward. To facilitate the addition of new dimers, it is critical for the growing polymer to maintain a 'GTP cap' – a layer of tubulin-GTP at the developing tip. Only tubulin-GTP can form stable structures, hence its vital role in tubulin polymerization. When the hydrolysis of GTP catches up with the polymer's growth, polymerization halts, and the structure begins to de-polymerize, leading to separate tubulin units. This feature of tubulin-GDP also causes microtubules to slowly disassemble from the α-side *in vitro*, albeit at a slower rate than polymerization[5,6].

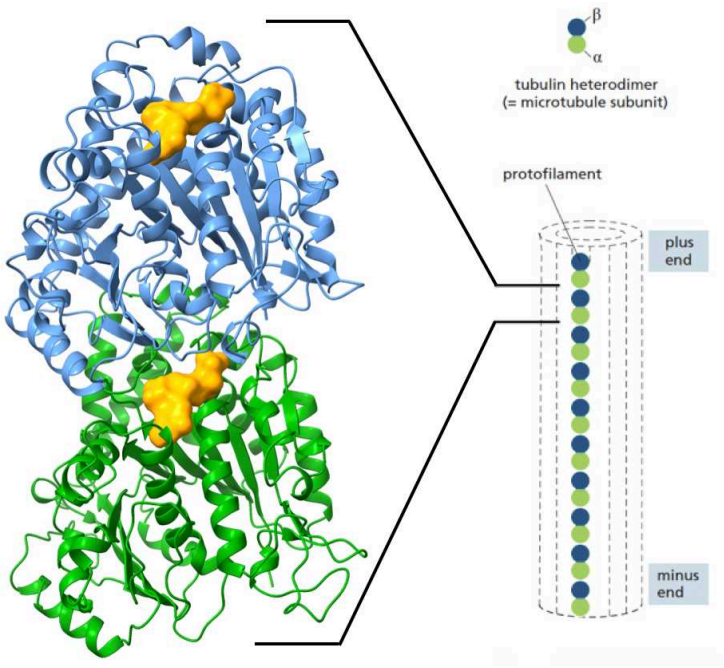


Figure 3. A protofilament is made of individual tubulin heterodimers bound together in a head-to-tail fashion.

1.1.3. Microtubule formation mechanism

When sufficient number of protofilaments is available, they stochastically come together and fold to form a microtubule. In the biological context, a microtubule is typically composed of 13 protofilament chains, thereby shaping a nanotube with a diameter of approximately 25 nm. In laboratory conditions, the count of protofilaments in a microtubule can range from 11 to 15. Microtubules are the most rigid and straightest structural elements in most eukaryotic cells [6,13].

The arrangement of protofilaments within the microtubule follows a helical pattern, forming a helix with a skew of 1.5 tubulin dimer units between the first and last chains at the seam line. This configuration allows the helix to complete one turn across three protofilament subunits[5] (Figure 4).

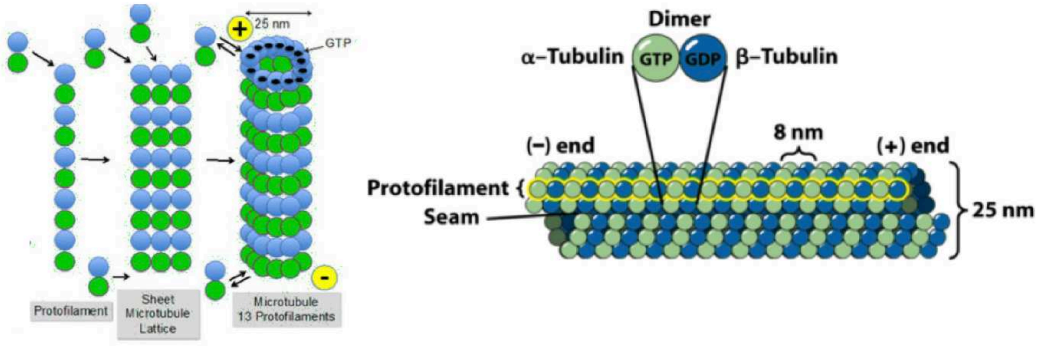


Figure 4. Microtubule structure

Two types of bonds contribute to the structural integrity of the microtubule: longitudinal bonds that span each protofilament chain and lateral bonds that connect neighboring tubulin subunits (Figure 5). The longitudinal interface, similar to the one that links individual αβ-tubulin units, possesses high binding energy. Lateral bonds are formed between identical subunits, either α-tubulin to α-tubulin or β-tubulin to β-tubulin, with the exception of the seam where α-tubulin is adjacent to β-tubulin [5,8].

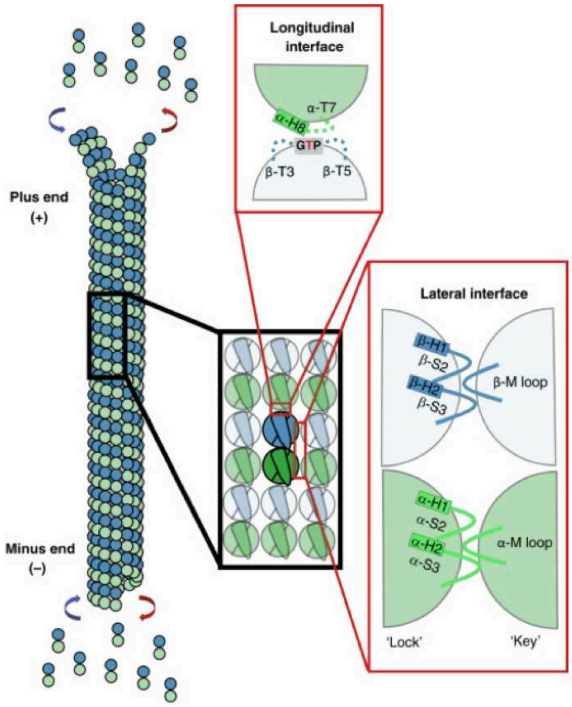


Figure 5. Lateral and longitudinal contacts contribute to the microtubule stability

Microtubules, being polar structures, expose α-tubulin at one end and β-tubulin at the other. Given the tight bonding within the microtubules, the addition or removal of tubulin dimers is typically restricted to the ends of the microtubule. This dynamic process, regulated by GTP hydrolysis and taking place predominantly at the microtubule's plus end, is referred to as *dynamic instability*[4,5].

1.1.4. Dynamic instability of microtubules

Microtubule dynamic instability is a hallmark of eukaryotic cellular function, underpinned by intricate processes involving tubulin-GTP and tubulin-GDP dimers, and their respective roles in microtubule polymerization and depolymerization. This interplay of conformational changes and GTP hydrolysis drives the characteristic behavior of microtubules and shapes their functional role in the cell [5,14].

Initial polymerization of microtubules is facilitated by a structure known as the GTP cap, composed of tubulin-GTP dimers. This cap stabilizes the microtubule, facilitating the addition of

further tubulin-GTP dimers. The cap results from the fact that the protofilaments that made up the microtubule had tubulin-GTP at their plus ends when folding into a microtubule. Tubulin-GTP dimers, which initially exhibit a curved conformation, are added to growing microtubule ends. Upon incorporation into microtubule lattices, these curved dimers undergo a gradual conformational transition towards a straighter structure, in line with the overall microtubule architecture. This is known as a "curved-to-straight" conformation change (Figure 6). The delay in GTP hydrolysis and phosphate release relative to microtubule growth allows the ends to maintain a GTP-tubulin cap, thus stabilizing the structure. Subsequent hydrolysis of the GTP in β-tubulin units is a crucial step, notably slower in a free tubulin dimer compared to when the dimer is part of a microtubule[5,8].

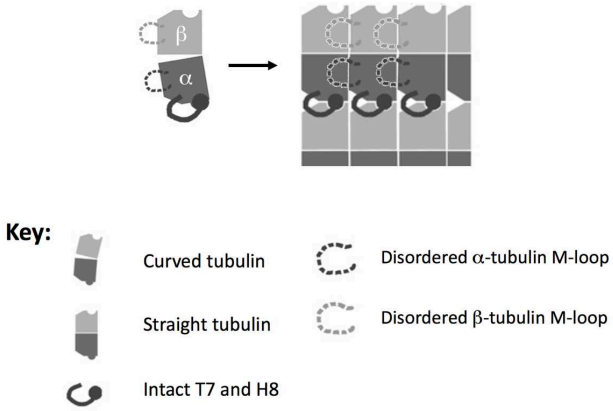


Figure 6. Schematic representation of curved-to-straight conformational change[5].

The main shaft of the microtubule consists of tubulin-GDP dimers, and a random loss of the GTP cap induced by an imbalance between the hydrolysis rate and the addition of new tubulin-GTP dimers, can cause rapid depolymerization, a process known as "catastrophe" (Figure 7). This is characterized by protofilaments dissociating from the microtubule shaft into tubulin-GDP dimers and small curved oligomers. In this event, the tubulin-GDP units assume a curved conformation that weakens the microtubule structure, leading to protofilament dissociation, and potential total microtubule disassembly. However, mechanisms exist for microtubule "rescue", whereby the structure can either stabilize itself or undergo reconstruction [5,15].

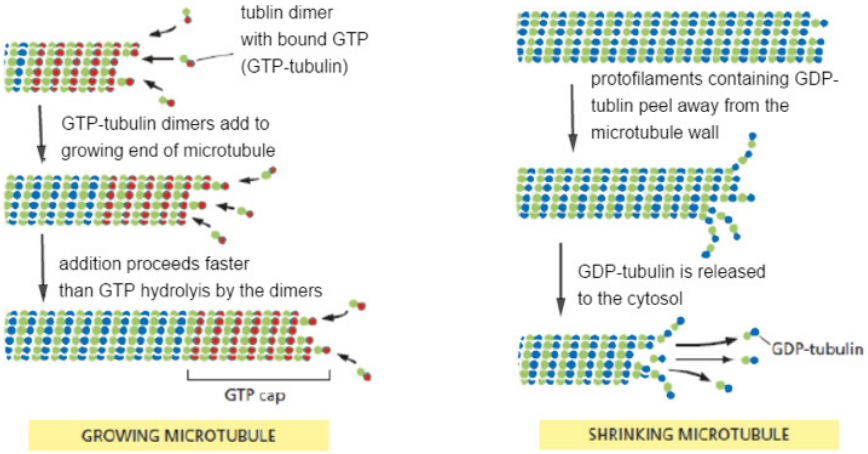


Figure 7. Schematic representation of microtubule growth and shrinkage

There are two main hypotheses for the rescue process. The first postulates the existence of "GTP islands" within the microtubule structure, where β-tubulin remains unhydrolyzed during polymerization. In the event of a catastrophe, these GTP-rich regions provide the foundation for microtubule rebuilding. The second hypothesis suggests the intervention of specific proteins which can influence the GDP-tubulin within a disassembling microtubule, encouraging it to abandon its curved conformation and adopt a structure more akin to that of GTP-tubulin, thereby stabilizing the overall microtubule structure and facilitating the binding of GTP-tubulin to the microtubule tip[16,17].

Although the exact mechanisms of the rescue process remain under debate, it is universally recognized that the cyclical processes of catastrophe and rescue are integral to microtubule function. They enable spatial and temporal regulation of microtubule assembly and disassembly within the cell. The balance between these processes, including phases of growth, shortening, and pausing, defines the dynamic instability of microtubules. These processes enable microtubules to be constructed and deconstructed at different sites of the cell whenever necessary, regulated by complex cellular biochemical machinery[5].

1.2. Microtubules and intracellular transport

1.2.1. Intracellular transport as part of normal cell functioning

Microtubules are integral to the intricate network of intracellular transport, performing crucial roles in the delivery of vesicles and organelles to specific locations within the cell, thereby ensuring their proper function. These microtubule-based transport mechanisms also mediate the transfer of signaling proteins between different organelles. Cellular components or cargo molecules bind to designated motor proteins that traverse the length of the microtubules, directed to precise intracellular sites by regulatory proteins[18].

The inherent polarity of microtubules facilitates bidirectional movement along their length. Special motor proteins can journey from the minus end to the plus end, a movement known as anterograde transport, which is oriented from the inner cellular space towards the cell periphery. Alternatively, they can move from the plus end towards the minus end, a process referred to as retrograde transport, which proceeds from the cell exterior towards its interior. This bidirectional transport is executed by two families of motor proteins: kinesins and dyneins, discussed in more detail below[6,19].

Neurons, the functional units of the brain responsible for processing and transmitting information via electrical and chemical signals, are particularly reliant on intracellular transport. Neurons comprise axons and dendrites, both of which are packed with microtubule assemblies. The polarity of these microtubules, specifically the orientation of their plus and minus ends, is of paramount importance. Within axons, all microtubules are aligned such that their minus ends point towards the cell body and their plus ends extend towards the axonal terminals. This arrangement serves as a microtubule highway, guiding the transport of specific proteins and vesicle-bound biological cargo to the dendrites[20] (Figure 8).

For instance, neurotransmitters, which are synthesized in the cell body near the nucleus, must traverse extensive distances to reach the axonal termini where they participate in synaptic transmission. It is along the microtubule routes that these vital molecules are transported from their point of origin to their site of action[6].

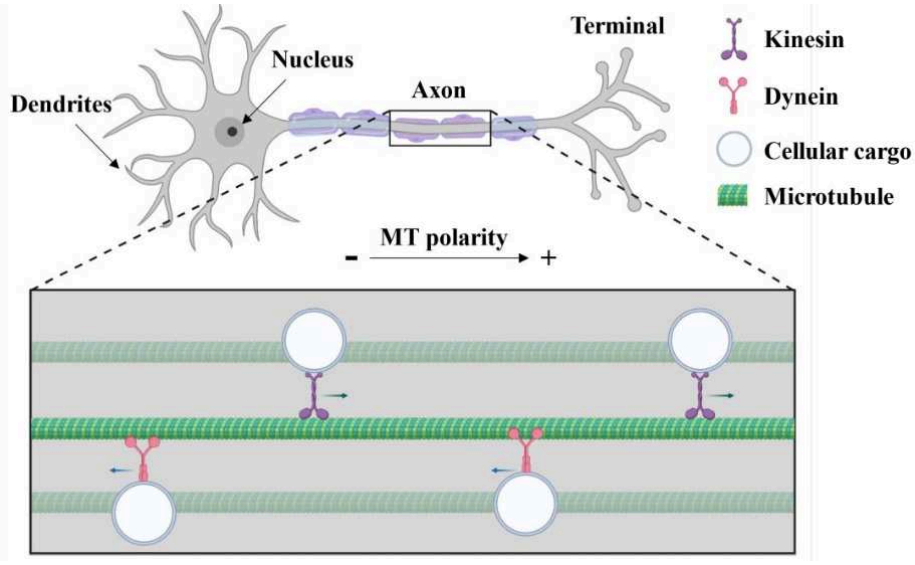


Figure 8. Microtubules are involved in intracellular transport

Microtubules are not only indispensable for axonal transport but are also essential in other eukaryotic cells. They play a crucial role in intracellular transport by providing a structural framework that guides and facilitates the movement of cargo. This includes the transport of organelles, such as mitochondria and lysosomes, as well as vesicles carrying proteins and lipids,

to their required destinations within the cell. This ensures the proper localization of cellular components and allows for efficient cellular functioning[18].

1.2.2. Microtubule-associated proteins (MAPs)

The intricate dynamics of microtubules is a characteristic that cells frequently exploit for their own functionality. This is made possible by an array of biochemical machinery within cells that governs and guides the dynamic instability of microtubules, a task primarily executed by microtubule-associated proteins (MAPs). As a diverse set of proteins, MAPs interact with microtubules to regulate their dynamics, stability, and organization. Their critical role extends to numerous cellular processes. The cellular composition typically includes more than 100 different microtubule-binding proteins[6].

The structure of the tubulin protein includes a short C-terminal (~20 amino acids) enriched with glutamic and aspartic acids. Therefore, when tubulin assembles into microtubules, the surface of the latter carries a net negative charge. As a result, many MAPs, which are positively charged, bind to microtubules via electrostatic interactions[21].

MAPs can be broadly categorized into two groups: those that interact with individual tubulin heterodimers, and those that engage with fully formed microtubules.

A key representative of MAPs that interact with individual tubulin units is the γ-tubulin ring complex (γ-TuRC). Microtubule formation necessitates the interaction between numerous tubulin dimers, which, to occur spontaneously, demands an extraordinarily high concentration of free tubulin in one place - a condition challenging to meet *in vivo*. Therefore, additional factors are required to accelerate and induce microtubule nucleation. g-Tubulin, an isotype of tubulin, serves precisely this function: it acts as a nucleation site to foster the growth of microtubules. However, g-tubulin is usually found as part of a larger structure, the γ-TuRC, in all cells[22].

The γ-TuRC, a protein complex, initiates the formation of microtubules by providing a template for the polymerization of 13 protofilament microtubules (Figure 9). It consists of g-tubulin molecules bound to various members of a g-tubulin complex protein family. The complex has a cone-like structure, with g-tubulin molecules arranged in a single-turn helix, which facilitates the addition of αβ-tubulin dimers from the cytosol. It does this by forming robust bonds with α-tubulin and promoting lateral contacts between forming tubulin proto-filaments. When these proto-filaments reach a critical size, they rapidly polymerize into a microtubule filament, held together by their bonds with γ-TuRC. As γ-TuRC only binds to the α-tubulin subunit, the resulting microtubules retain their intrinsic polarity, with the unstable α-side (the minus pole) attached to and stabilized by the organizing center, and the reactive β-side (the plus pole) exposed towards the

cell. Furthermore, γ-TuRCs can bind to microtubules that have formed stochastically within the cell from the α-side, thereby anchoring them[6,22].

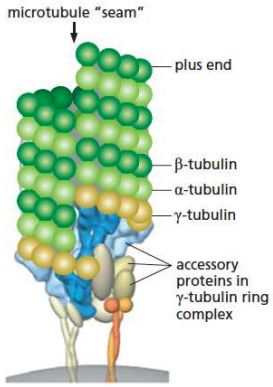


Figure 9. γ-TuRC is a nucleation center that promotes formation of microtubules with 13 protofilaments. Image adapted from Alberts et al.[6]

Stathmin is another example of a microtubule-associated protein modulating microtubule polymerization. Stathmin is a highly conserved 17 kDa protein that plays a crucial role in the regulation of the cell cytoskeleton. Unlike the γ-TuRC, stathmin binds to small oligomers of tubulin, inhibiting their further polymerization[23].

MAPs' interactions with formed microtubules can be categorized into three broad classes: (1) those that promote microtubule polymerization, (2) those that induce microtubule disassembly, and (3) those that cross-link microtubules to form complex arrays (Figure 10).

A representative of the first class is XMAP215. This protein belongs to a highly conserved group of MAPs, distinct due to their primary interaction with the growing-end (plus-end) of microtubules. This interaction places XMAP215 within the family of plus-end tracking proteins (+TIPs)[24,25].

In contrast, kinesin-13 serves as an example from the second class. Kinesin-13s are microtubule depolymerizing enzymes, thus promoting microtubule disassembly and playing a significant role in the catastrophe process[6].

The tau protein is an important member of the third class, known for its role in microtubule stabilization and cross-linking in human brain neurons. Tau helps stabilize neuronal microtubules, promoting axonal outgrowth and ensuring long-distance cargo trafficking in neurons for correct neuroactivity[6].

In summary, cells harbor a plethora of mechanisms to control microtubule dynamics, involving multiple signaling pathways and a multitude of proteins with specific roles. These diverse mechanisms highlight the cell's ability to fine-tune microtubule dynamics, facilitating the intricate processes of intracellular transport and cargo distribution.

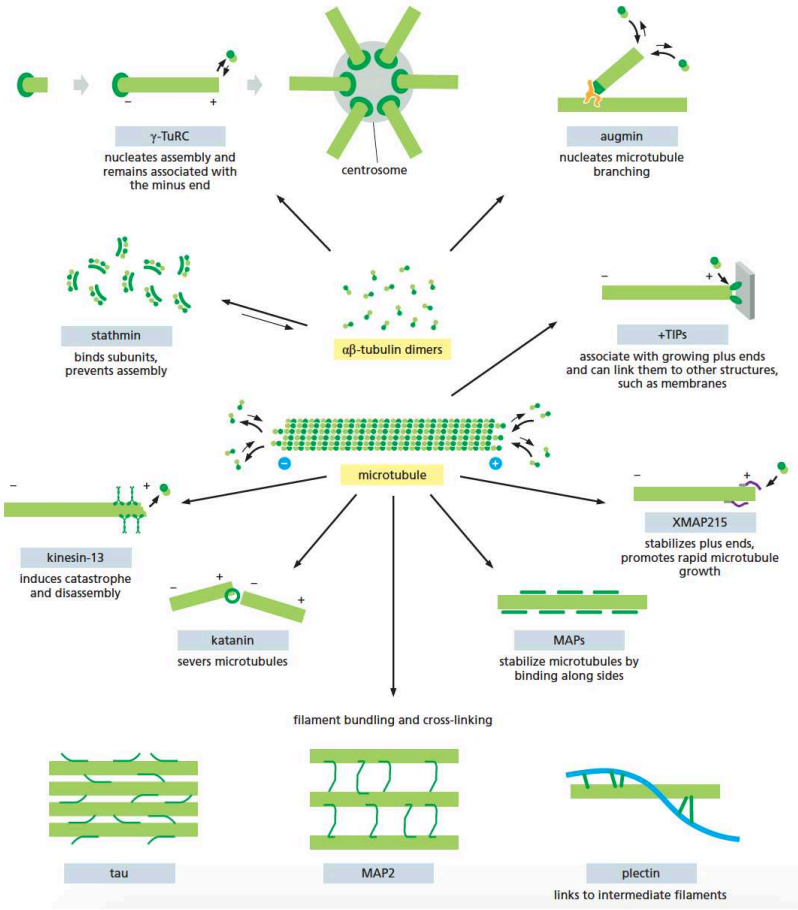


Figure 10. Schematic overview of different types of microtubule-associated proteins[6]

1.2.2. Motor proteins and their interaction with microtubules

Motor proteins constitute a critical class of MAPs, with their primary function being the transportation of various cellular components along microtubules and contributing to the positioning of microtubules relative to each other. The fundamental architecture of motor proteins includes three core elements: the motor domain, the cargo-binding domain, and the linker chain between them. Despite the existence of many members within a given motor protein family, such as the 45 identified members of the kinesin family in humans, the motor domain remains conserved across them. In contrast, the cargo-binding domain is specialized to bind to specific cargo targets [26].

The intracellular transport of molecules and organelles facilitated by microtubules is a crucial process for many cellular functions, including cognitive processes. Here, motor proteins play a pivotal role, carrying out two key functions: (1) the transportation of cargo like organelles and macromolecules over long distances within the cell; (2) facilitating the sliding of microtubules relative to one another, thus enabling the specific arrangements of microtubules[6].

Motor proteins are primarily categorized into two types: kinesins and dyneins (Figure 11). These proteins traverse microtubules in different directions and possess distinct structural features. Kinesins have two motor units at the head of the structure and traverse microtubules from the

minus end to the plus end. In contrast, dyneins, being eight times larger than kinesins, comprise two motor units attached to each other and the cargo, and traverse microtubules from the plus end to the minus end [26].

The mechanism of kinesin interaction with microtubules begins with the binding of the kinesin to the cargo it needs to transport. The motors then bind to a β-tubulin subunit in the microtubule, triggering a series of conformational changes that propel the motor protein forward along the microtubule. This process continues until the kinesin interacts with regulatory proteins that detach the cargo. Specifically, kinesins are involved in fast anterograde axonal transport, moving from minus to plus ends of a microtubule, carrying mitochondria, secretary vesicle precursors, and various synapse components to distant nerve terminals[27].

Conversely, dyneins move in the opposite direction to kinesins, from the plus end to the minus end of a microtubule. The structure of dynein motor proteins differs from kinesins, with each dynein protein consisting of two motor units attached to each other and the cargo. The precise mode of attachment remains unknown. Dynein transport operates through a process of stochastic binding and rebinding, which results in a worm-like movement pattern. Dynein family proteins are involved in retrograde axonal transport, moving from plus to minus ends of a microtubule[27].

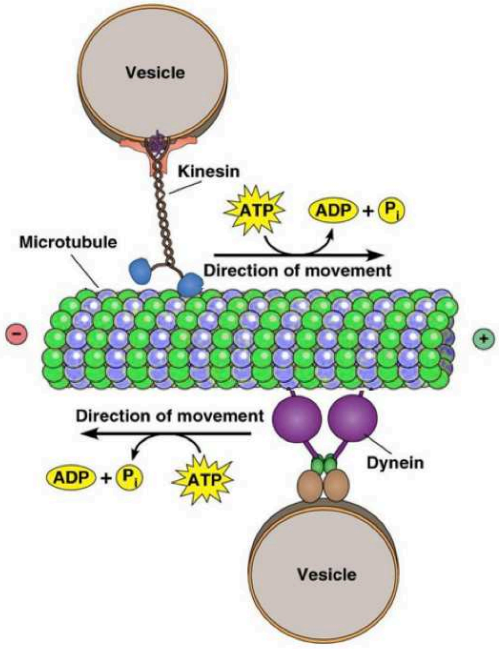


Figure 11. Kinesin and dynein families of motor proteins have distinct mechanisms of traversing microtubules

The significance of microtubules for the proper functioning of cells cannot be understated. As the primary conduits for intracellular transport, microtubules form an essential network within the cell. Their dynamic nature - characterized by constant growth and shrinkage - allows for the effective and timely delivery of various cellular components, from organelles to macromolecules,

to where they are needed most. This is achieved through the strategic regulation of their dynamics by the cell, using microtubule-associated proteins, including motor proteins primarily responsible for physical cargo transportation down the microtubules [6,26,27].

Importantly, recent research has implicated microtubule dysfunction in neurodegenerative diseases in relation to the dysregulation of microtubule dynamics in unhealthy neurons, highlighting the importance of their further study. Thus, investigating microtubules' dynamic behavior offers a promising avenue for gaining insights into not only the fundamental mechanisms of intracellular transport but also the pathogenesis of neurodegenerative diseases [10,11].

1.3. Microtubules in cell division

1.3.1. The cell cycle

In eukaryotic organisms, the continuous process of cell division is central to both the survival and reproduction of the organism. This intricate process involves a carefully orchestrated series of events, governed by a combination of external stimuli and internal checkpoints. The ultimate aim of the cell cycle is to duplicate the DNA of each chromosome of the parent cell and distribute these copies to two daughter cells, ensuring that each newly formed cell possesses an identical genome to its predecessor[28].

Fundamentally, the cell division process in eukaryotes comprises two main phases: the Synthesis (S) phase and the Mitosis (M) phase. During the S phase, the cell duplicates its DNA, preparing for the subsequent division. This is a time-consuming process, often taking tens of hours to complete. Following the S phase is the M phase, which is characterized by the physical segregation of chromosomes into daughter nuclei, culminating in the formation of two genetically identical cells. Compared to the S phase, the M phase is relatively swift, often completed in less than an hour[6].

Intersecting the S and M phases are periods of pause, the Gap 1 (G1) and Gap 2 (G2) phases. These intervals allow the cell to assess internal conditions and respond to external signals, ensuring it is ready to proceed to the next stage of the cycle. Therefore, the eukaryotic cell cycle is traditionally divided into four sequential phases: G1, S, G2, and M. Collectively, the G1, S, and G2 phases are often referred to as the interphase[6] (Figure 12).

Several additional steps occur within the M phase, each reliant on the key driving force of microtubules. These steps include prophase, prometaphase, metaphase, and anaphase, all of which occur within the broader mitotic stage of the M phase. The active participation of microtubules during these stages highlights their critical role in cell division. The following sections will discuss these stages, shedding light on the fundamental importance of microtubules in the process of cell division[6,28].

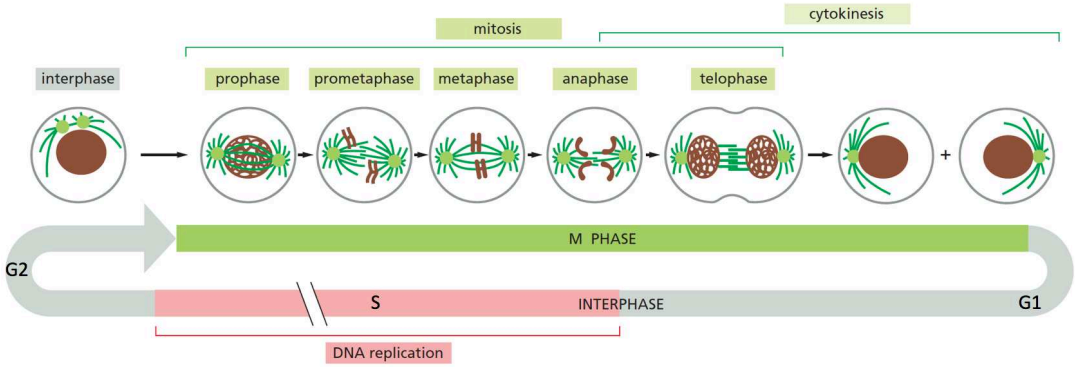


Figure 12. General scheme of cell division[6]

1.3.2. Prophase

In the initial stages of mitosis, the cell receives a signal to start replication, triggering a series of complex processes within the nucleus. At the crux of this process is the duplication of the DNA. During this phase, a chromosome splits into individual chromatids, and each chromatid is copied to form an identical pair. These pairs, referred to as sister chromatids, are bound together at a region known as the centromere. Each sister chromatid thus carries two identical copies of a chromosome, marking a successful completion of the DNA replication process during the Synthesis (S) phase of mitosis[6].

Simultaneously, the cell begins the process of centrosome duplication. Located adjacent to the nucleus, the centrosome serves as the primary microtubule-organizing center (MTOC) in most cells during interphase, nucleating the majority of the cell's cytoplasmic microtubules. As the cell enters the cell cycle, the centrosome duplicates, ensuring that two centrosomes are present by the onset of mitosis. This duplication is initiated in tandem with the cell's entry into the S phase, under the influence of proteins that trigger cell-cycle entry[29].

The centrosome, an organelle specialized for the nucleation of microtubules, is a distinct location within the cell from which microtubules emerge and spread (Figure 13). Comprising two tubular structures known as centrioles and a protein-rich pericentriolar material (PCM), the centrosome serves as the main MTOC in eukaryotic cells. The PCM plays a crucial role in efficient microtubule nucleation, recruiting multiprotein γ-tubulin ring complexes (γ-TuRCs) at its surface. These γ-TuRCs facilitate the nucleation of microtubules, with their minus ends anchored at the centrosome and their plus ends extending outwards, continually growing and shrinking to explore the cell's three-dimensional volume. It is worth noting that although centrosomes are the most extensively studied MTOCs, any organelle can serve as an MTOC provided it recruits γ-TuRCs at its surface[6,22,29].

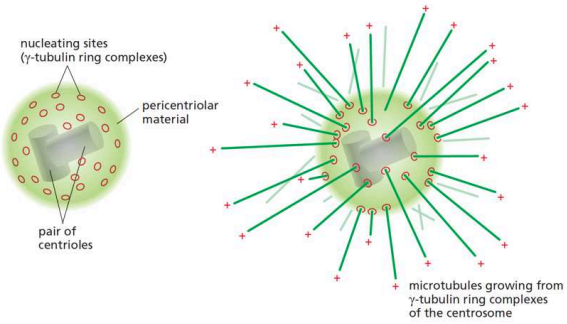


Figure 13. The centrosome

Following the initiation of microtubule formation by the centrosomes, the two closely located centrosomes require motor proteins from the cytoplasm for their distribution to opposite ends of the cell (Figure 14). This separation of the centrosomes forms the two poles of a complex protein structure known as the mitotic spindle[6].

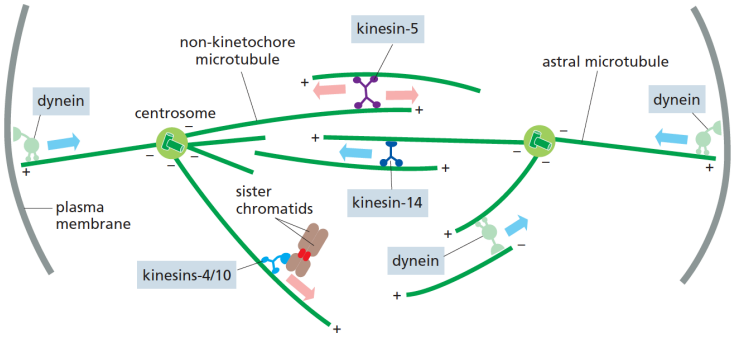


Figure 14. Distribution of centrosomes

The mitotic spindle is a bipolar network of highly dynamic microtubules, characterized by their minus ends oriented towards the poles and their plus ends directed outwards (Figure 15). This network of microtubules undergoes a significant transformation at the onset of mitosis, becoming far more dynamic than those present during interphase. This increased instability culminates in a remarkably dense and dynamic array of spindle microtubules, setting the stage for the subsequent steps of mitosis[8].

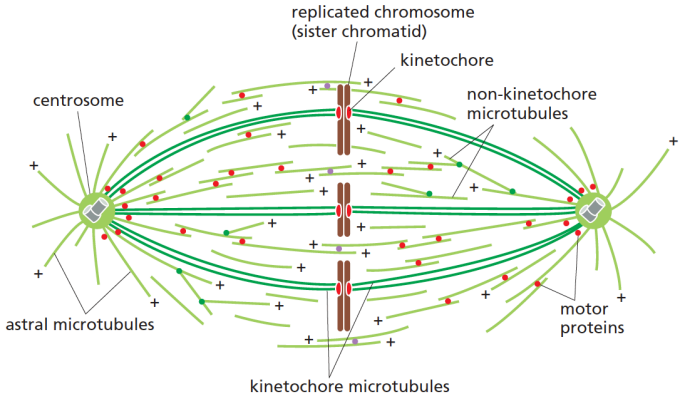


Figure 15. The mitotic spindle

The mitotic spindle is made of three primary types of microtubules: kinetochore, astral, and polar microtubules. Each type has distinct roles and characteristics, contributing to the orchestration of successful cell division[6,8].

Kinetochore microtubules play a pivotal role in chromosome segregation during the metaphase. They connect to kinetochores, large protein structures located at the centromere of each sister chromatid. This connection is established at the plus ends of the kinetochore microtubules during metaphase, effectively linking sister-chromatid pairs to the mitotic spindle. Each kinetochore forms a robust attachment with numerous microtubules, which are cross-linked to create substantial microtubule bundles. Overall, the kinetochore microtubules function to connect the chromosomes to the spindle via kinetochores, with each kinetochore accommodating the attachment of approximately 20-30 microtubules during metaphase[8].

Polar microtubules, also known as non-kinetochore microtubules, originate from opposing spindle poles and interact with each other through motor proteins. This interaction facilitates the separation and stabilization of the mitotic spindle. Despite their short and unstable nature, polar microtubules contribute to the structural integrity of the spindle. They are cross-linked by various proteins to form a dynamic scaffolding network, capable of adapting to ensure the stability of the spindle[6].

Finally, astral microtubules extend from the spindle poles towards the cell cortex. This outward radiation allows them to interact with the cell cortex, playing a vital role in spindle positioning within the cell. By maintaining contact with the cell cortex, astral microtubules assist in correctly situating the spindle, ensuring a successful execution of mitosis[6].

1.3.3. Prometaphase

Prometaphase represents a pivotal shift in the process of mitosis, marked by the rupture of the nuclear envelope (Figure 16). This rupture liberates the sister chromatids, enabling their interaction with microtubules, and marking the start of their distribution within the cell. Once released, sister chromatids situated at the cell's center are readily seized by microtubules emanating from the mitotic spindle[6].

With the nuclear envelope now dissolved, a complex series of interactions is involved to position the sister chromatids precisely at the midpoint of the cell. This process relies heavily on a swarm of microtubules, which rapidly polymerize in the vicinity of the newly freed sister chromatids, as the latter create a suitable environment for nucleating microtubule polymerization[6].

Motor proteins play a vital role in this process as well, interacting with both the newly formed microtubules and released sister chromatids to facilitate their accurate positioning. A sister chromatid’s precise placement at the equatorial plate is achieved through the balanced

polymerization and depolymerization of microtubules. This intricate system involves regulatory microtubule-associated proteins to ensure the perfect alignment of sister chromatids at the cell's center[18,21].

To ensure such configuration, cells employ a mechanism known as the spindle assembly checkpoint. This surveillance system detects and corrects errors, promoting the accurate positioning of chromosomes. A multitude of proteins and mechanisms coordinate to secure the precise placement of chromosomes at the cell's midpoint[30].

As a result, prometaphase orchestrates the meticulous positioning of sister chromatids precisely at the center of the dividing cell, setting the stage for the subsequent steps of mitosis.

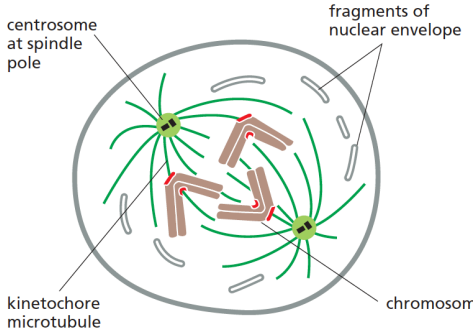


Figure 16. Prometaphase

1.3.4. Metaphase

At the metaphase stage of mitosis, the sister chromatids, carrying crucial genetic material, have been meticulously positioned at the cell's center. This central placement facilitates their attachment to the microtubules, which marks the onset of one of the most critical steps in the cell division process[6].

Following the successful assembly of a bipolar microtubule array, the next major step involves this array's attachment to the sister chromatids. The spindle microtubules find their attachment points at each chromatid's kinetochore, a multilayered protein structure located at the chromatid's centromeric region. During metaphase, the plus ends of these kinetochore microtubules embed head-on into specialized attachment sites within the kinetochore's outer region (the Ndc80 protein complex, not reviewed here), the area most distant from the DNA[31]. This strategic attachment of the kinetochore protein complex to the microtubule creates an important linkage between the microtubule and the kinetochore and, by extension, the chromosome. Crucially, this attachment process does not impede the addition or removal of free tubulin subunits at the microtubule's plus end[32].

During metaphase, the chromosomes align at the cell's equatorial plane, forming the metaphase plate. This is achieved through the balanced pulling forces generated by the kinetochore

microtubules from opposite spindle poles. This tug-of-war-like tension ensures that each chromosome is properly aligned for the next stage of mitosis [8] (Figure 17).

Moreover, metaphase hosts one of the cell cycle checkpoints, ensuring that all chromosomes are correctly attached to microtubules and properly aligned before the cell proceeds to anaphase. This stage in mitosis is not only significant for its pivotal role in the orderly segregation of genetic material but also for its role in maintaining the integrity of cell division through stringent regulatory checkpoints[33].

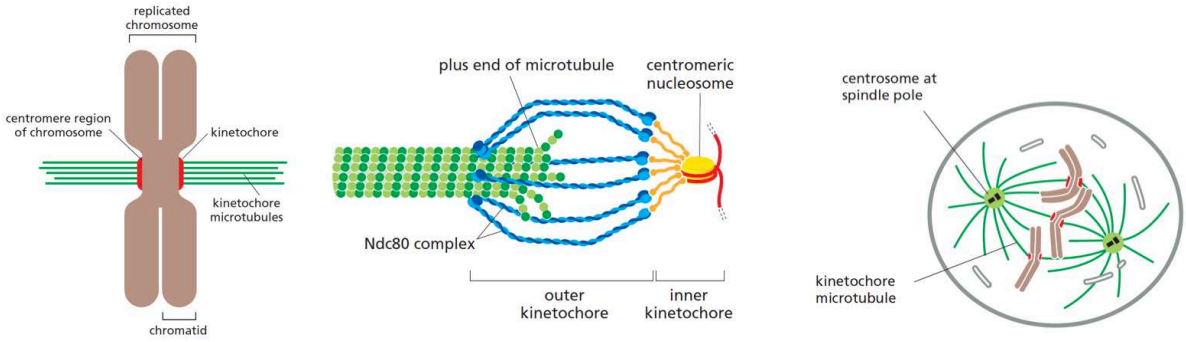


Figure 17. The metaphase

1.3.5. Anaphase

Anaphase represents the pinnacle of mitotic activity wherein the sister chromatids are not only separated but also relocated to opposite poles of the cell (Figure 18). The dissolution of sister-chromatid cohesion enables the separation of the sister chromatids, thus permitting the forces of the mitotic spindle to pull the sisters towards the cell's opposite poles, effectuating chromosome segregation. This stage is the culmination of the processes that have preceded it and relies on two primary forces to facilitate the requisite movement[6].

The first force is generated through the depolymerization of the microtubule's plus end, to which the chromosome's kinetochore is attached. This force propels the kinetochore and its associated chromatid along the kinetochore microtubule towards the spindle pole. As the kinetochore moves further along the microtubule, the microtubule itself collapses behind it, effectively pulling the kinetochore towards the spindle pole. Even as the microtubule depolymerizes, the kinetochore remains attached due to the multiple low-affinity attachments formed along the side of the microtubule. These attachments are continuously breaking and re-forming at new sites, ensuring the kinetochore's continual movement. The dynamic behavior of tubulin, particularly its depolymerization, is integral to the successful execution of this process[6,8].

A second force, referred to as microtubule flux, contributes to the poleward movement in certain cell types. This process involves the microtubules being pulled towards the spindle poles and dismantled at their minus ends. While the exact mechanism underlying this movement remains unclear, it likely involves forces generated by motor proteins and minus-end depolymerization at

the spindle pole. The addition of new tubulin at the plus end compensates for the loss at the minus end, maintaining a constant microtubule length despite the movement towards the spindle pole. It is the stored energy within the microtubule, primarily derived from the hydrolysis of GTP after a tubulin subunit has been added to the microtubule plus end, that drives this movement. A kinetochore attached to a microtubule undergoing such flux experiences a poleward force. This force, in conjunction with the kinetochore-based forces, contributes to moving the sister chromatids after their separation in anaphase[6].

Finally, the subsequent stages of telophase and cytokinesis essentially complete the segregation of chromosomes by enclosing the divided genetic material within new membranes. These stages, however, do not involve active participation of tubulin or microtubules, and are thus not discussed here in detail.

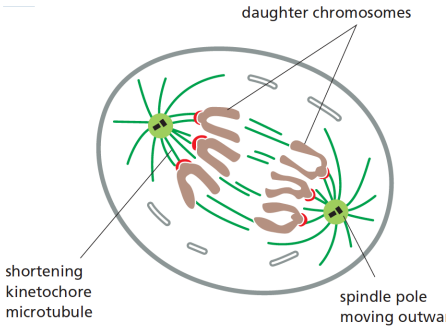


Figure 18. Results of the anaphase

1.4. Small molecule modulators of tubulin polymerization

Microtubules are essential for the correct functioning of all eukaryotic cells. Specifically, the intrinsic property of tubulin heterodimers to spontaneously polymerize into microtubules, and then dynamically switch between the growth and shrinkage phases both stochastically and under the influence of external factors, is something that drives two of the most important cellular processes: intracellular transport and cell division.

Of particular importance is the role of tubulin in intracellular transport, which has especially profound implications in neurodegeneration. Neurons heavily rely on the continuous transport of biochemicals between the neuron cell body and axon terminals. Therefore, any disruption to tubulin polymerization or microtubule stability inflicts considerable damage upon neurons, potentially leading to their malfunction or even death. Understanding the role of microtubules in neuronal function and disease thus paves the way for developing effective treatments for neurological disorders[10,11].

Similarly, tubulin is involved in a key stage of mitosis, what makes it a promising target for anti-cancer research. By interfering with microtubule dynamic instability during the cell cycle, it becomes possible to halt the cycle and induce the death of cancer cells[5,34].

Thus, tubulin's multifaceted role makes it a significant biological target for both cancer- and neurodegeneration-related research. In particular, the design and discovery of small molecules capable of modulating tubulin polymerization present promising ways to regulate cancer cell proliferation and investigate neuronal behavior. By understanding the processes that regulate tubulin dynamics, we may be able to manipulate them in a targeted manner. Such manipulations could lead to the development of new therapeutics.

As such, microtubule-targeting agents (MTAs) have emerged as powerful tools in the fight against various diseases, including cancer, due to their ability to interfere with microtubule dynamics. The specific binding of MTAs to microtubules (MTs) allows them to interfere with the functional performance of these structures, making them invaluable tools in studying MT function and its role in diverse cellular processes. By using MTAs as molecular probes, it becomes possible to track the interactions of MTs with other cellular components and perturb MT functions in cells, thereby studying the effects of MT disruption on cellular processes. MTAs can be labeled with fluorescent dyes or other markers, such as radioisotopes, broadening their application in research[5,35].

They have also been employed as therapeutic agents in the treatment of diseases associated with dysregulation of MT dynamics, most notably – cancer. MTAs predominantly induce cell death in dividing cells, given the crucial role of microtubule dynamics in maintaining the functional integrity of the mitotic spindle. However, these drugs also exhibit toxicity towards healthy, rapidly dividing cells, such as those present in bone marrow, intestine, and hair follicles[36].

A broad range of chemical classes of MTAs have been identified, the majority of which are natural products or their synthetic derivatives extracted from diverse natural sources such as marine sponges, plants, or bacteria. First developments of MTAs in the context of anti-cancer research started in the 1950s, culminating in the development of several FDA-approved anti-cancer MTAs including vinblastine (Velban®), vincristine (Vincrex®), paclitaxel (Taxol®), epothilone (Ixempra®), eribulin (Heleven®), auristatin (Adcetris®), and Trastuzumab emtansine (Kadcyla®, an antibody-drug conjugate). While many of these tubulin binders display promising *in vitro* profiles, they also present significant off-target effects when tested in patients, necessitating the continued search for safer and more efficient MTAs. The potential of MTAs in treating injuries and diseases of the nervous system is underscored by their ability to affect intracellular trafficking of vital molecules and organelles. Moreover, the application of MTAs in cancer treatment has been limited by factors such as high cytotoxicity, complex non-scalable

synthesis, adverse properties, and the development of resistance. With the increasing availability of tubulin structural data, computer-aided design techniques can be instrumental in focusing on the relevant chemical space and guiding the design process of new MTAs[5].

As of this writing, eight MTA binding sites have been identified and characterized (Figure 19), with the most recent discoveries made in the last two years, characterized extensively via X-ray crystallography. MTAs can be broadly classified as microtubule stabilizing (MSA) and destabilizing (MDA) agents based on their action on microtubules. While the former promote microtubule assembly and enhance the stability of the formed microtubules, the latter provoke microtubule disassembly into separate tubulin dimers or small oligomers. The exact action exerted by a MTA on tubulin polymerization depends on the binding site that the molecule binds to, as the binding interferes with conformational changes required for normal functioning of the protein[5].

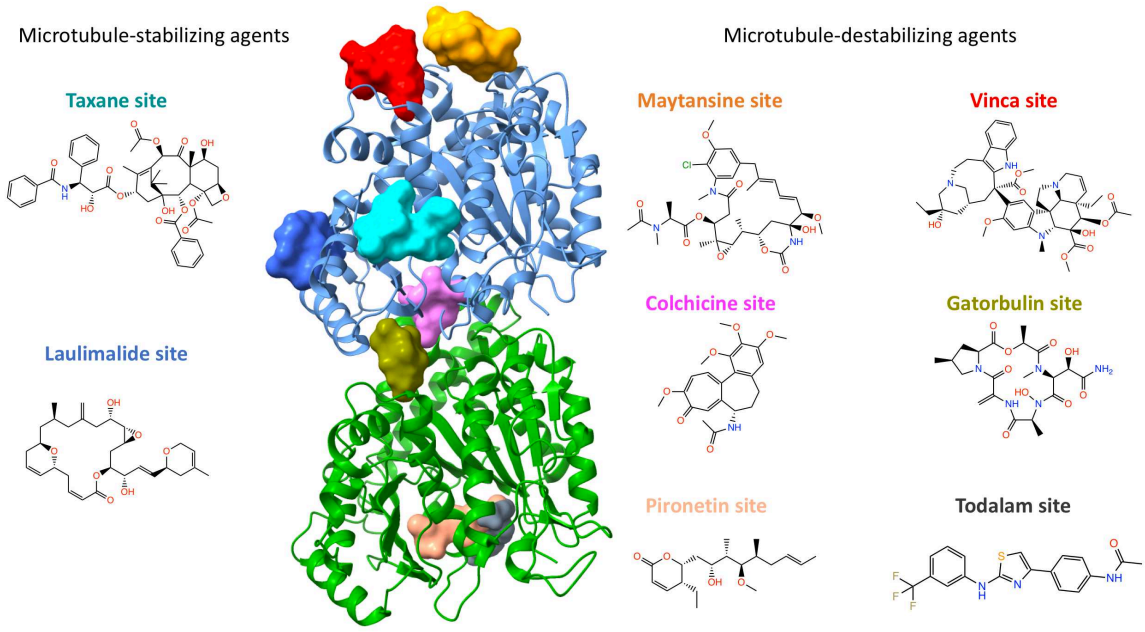


Figure 19. Microtubule-targeting agents have different action on microtubule polymerization

1.4.1. Microtubule-stabilizing agents

1.4.1.1. Taxane site

Located on the luminal side of microtubules, the taxane binding site is a pocket within β-tubulin predominantly formed by hydrophobic residues of H7, S7, loops H6-H7, S7-H9 (the M-loop), and S9-S10. Ligands of this site establish both hydrophobic and polar contacts with several of these secondary structural elements. Some ligands bind to this site by forming covalent bonds with residues within it[5].

As the site is close to the unordered M-loop, certain ligands (zampanolide, epothilone A) structure it into a short helix. Such induced structuring greatly contributes to the stability of

microtubules because the M-loop plays a significant role in establishing lateral tubulin contacts within microtubules. On the other hand, other ligands (paclitaxel, discodermolide) stabilize microtubules by enhancing longitudinal tubulin contacts via an allosteric mechanism (Figure 20)[5,37].

These diverse mechanisms suggest that although different taxane-site ligands bind to the same pocket on β-tubulin, they may achieve their microtubule-stabilizing effects through varying molecular pathways. Importantly, several ligands that target the taxane site (e.g., paclitaxel (Taxol®), epothilone (Ixempra®), eribulin (Heleven®)) have gained FDA approval for use as cancer therapeutics[5,37].

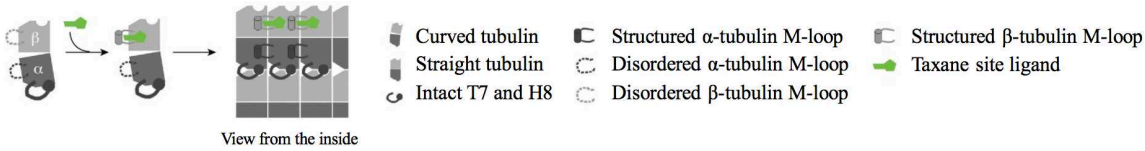


Figure 20. Schematic representation of taxane site-targeting molecules' action

1.4.1.2. Laulimalide/Peloruside site

Laulimalide and peloruside A are two microtubule-stabilizing agents that demonstrate notable cytotoxic activity against a diverse range of cancer cell lines bind to a site different from the taxane site[5,37].

The binding pocket common to both of these ligands is located on β-tubulin on the microtubule's exterior. This site is composed of hydrophobic and polar residues of helices H9 and H10, along with the loops H9–H90 and H10–S9 of β-tubulin. Due to its location on the microtubule, this site allows the two ligands to establish lateral contacts with adjacent protofilaments, thereby inhibiting microtubule disassembly[5] (Figure 21).

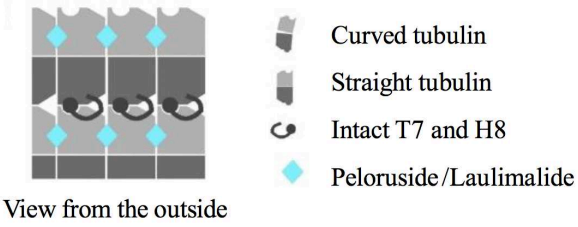


Figure 21. Schematic representation of lauliamalide site-targeting agents action

1.4.2. Microtubule-destabilizing agents

1.4.2.1. Vinca site

The vinca alkaloids are the earliest known microtubule-targeting agents. A variety of compounds, both natural and synthetic, target the vinca site and span several chemical classes [38]. These ligands bind at the inter-dimer interface between two longitudinally aligned tubulin dimers.

The structural elements that form the core zone of the vinca binding site include the C-terminal turn of helix H6, loops T5 and H6– H7 of β-tubulin, and helix H10, strand S9, and loop T7 of α-tubulin of another heterodimer[5] (Figure 22).

The destabilization of microtubules by vinca-site ligands is achieved in one of two ways. Some ligands obstructing the curved-to-straight transition of tubulin on the microtubule tip. Alternatively, other ligands only allow tubulin dimers to form ring-like oligomers that are incompatible with the straight protofilament structure in microtubules [38]. Notably, some vinca site-targeting agents have received FDA approval for use as anti-cancer drugs (e.g., vinblastine (Velban®), vincristine (Vincrex®))[5].

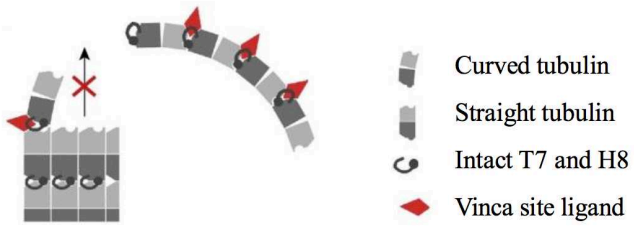


Figure 22. Schematic representation of the vinca site-targeting agents action

1.4.2.2. Colchicine site

The colchicine binding site is located in the intra-dimer interface between the α- and β-tubulin subunits. The ligands targeting the colchicine site display a remarkable structural variety around a limited number of structural frameworks[39]. Despite being a well-studied target for microtubule-targeting agents, none of the ligands directed at this site have yet advanced to the commercial phase[5].

Colchicine and its related ligands primarily inhibit microtubule formation by blocking the "curved-to-straight" conformational change in tubulin[39] (Figure 23). Within the bound state, the core secondary structural elements of the colchicine site interact predominantly through hydrophobic, and to a lesser extent, polar contacts with the ligand[5]. Chapter 5 of this thesis describes work on *de novo* design of novel colchicine site-targeting agents.

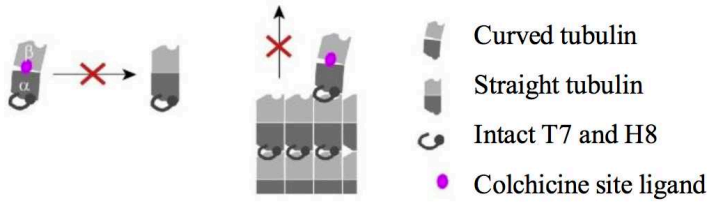


Figure 23. Schematic representation of the mechanism of action of colchicine site-targeting compounds

1.4.2.3. Gatorbulin site

A previously unknown binding site on the tubulin dimer has been identified recently[40]. This discovery involves a cydodepsipeptide named gatorbulin, which is derived from marine cyanobacteria. Gatorbulin has been observed to bind to a specific region within the tubulin dimer, located next to the colchicine binding site. However, the exact mechanism by which gatorbulin influences tubulin function is not yet fully understood. Further research is required to elucidate the precise mechanism of gatorbulin action on the tubulin dimer and microtubules [40].

1.4.2.4. Maytansine site

The maytansine binding site is an exposed pocket on the β-tubulin, located near the guanosine nucleotide binding site. It is shaped by a combination of hydrophobic and polar residues from helices H3', H11, and H11', and certain loops like S3-H3' (T3-loop), S5-H5 (T5-loop), and H11-H11' [5].

Binding of ligands to this site interferes with microtubule formation either by preventing tubulin dimers from joining the growing ends of microtubules by blocking the formation of new longitudinal interactions, or by creating tubulin-ligand that cannot participate in assembly, especially at high concentrations of the ligands (Figure 24). As a result, maytansine site-ligands exhibit strong anti-tumor activity *in vitro* and *in vivo*[41].

A derivative of maytansine, a natural compound after which the site was named for, is a part of the FDA-approved antibody-drug conjugate trastuzumab emtansine (Kadcyla®), used in the treatment of metastatic breast cancer[5].

Chapter 2 of this thesis discusses the work performed on discovery of novel maytansine site-targeting ligands.

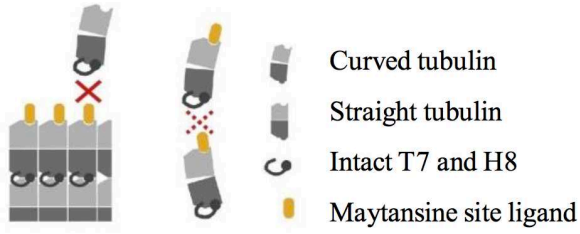


Figure 24. Schematic representation of the mechanism of action of maytansine site ligands.

1.4.2.5. Pironetin site

Until recently, the only known binding site on the α-tubulin subunit was the pironetin binding site[42]. Pironetin is a natural compound known for its promising anti-cancer properties. Upon binding, it forms a covalent bond with a cysteine residue within the site[43].

Pironetin binding induces significant changes in tubulin structure. Specifically, it disrupts the T7 loop and causes a conformational change in the N-terminal region of helix H8. Additionally, the pyrone ring of pironetin interacts with certain amino acid residues found in strands S8 and S10, as well as helix H7. The side chain of pironetin is buried within a pocket formed by amino acid residues of helix H7 and strands S4, S5, and S6[5].

Structural changes caused by the binding of pironetin have important implications for the dynamics of microtubules. They hinder the formation of microtubules by either creating complexes between tubulin and pironetin that are unable to assemble when the concentration of the ligand is high or by preventing the addition of more tubulin dimers to the minus ends of microtubules, where α-tubulin subunits are exposed (Figure 25). As a result, pironetin binding impairs the assembly of tubulin into microtubules, highlighting its potential as an effective agent for cancer chemotherapy[5].

Chapter 3 of this thesis described the work performed on attempted discovery of novel pironetin site-targeting ligands.

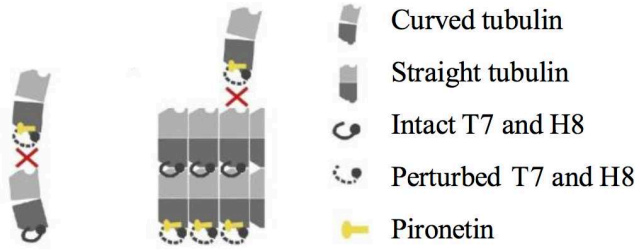


Figure 25. Schematic representation of the mechanism of action of pironetin site ligands

1.4.2.6. Todalam site

In a recent study by Mühlethaler et al., a crystallographic fragment screening of the tubulin protein[44] discovered a hitherto unknown binding site, now named the todalam binding site after the first ligand rationally designed by combining the small fragments that demonstrated affinity for the pocket and subsequently optimizing the resultant structure [45].

The todalam binding site is located between two longitudinally arranged αβ-tubulin heterodimers, in an interface between the maytansine site on β-tubulin and the end of the pironetin pocket on α-tubulin. Todalam is thought to hinder the formation of microtubules by creating a wedge in the tubulin-oligomer structure.[45] At the time of this writing, no other compounds have been reported to engage this particular binding site. Chapter 4 of this thesis discusses the work performed on discovering novel molecules that target this binding site.

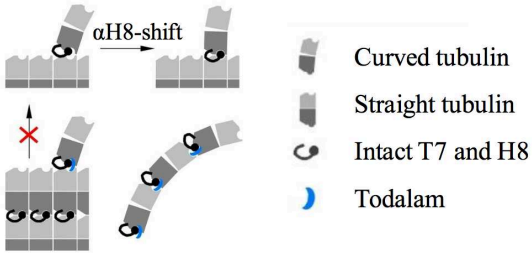


Figure 26. Schematic representation of the mechanism of action of todalam site ligands

1.5. Computer-aided drug design (CADD) methodologies used in this work

Computer-aided drug design (CADD) is a comprehensive field that employs a wide array of theoretical and computational methodologies, forming an integral part of modern drug discovery. These methods have significantly contributed to the development of numerous drugs that are either in clinical use or undergoing clinical trials, demonstrating their practical relevance in the pharmaceutical industry.

CADD methods can be broadly categorized into ligand-based and structure-based approaches based on the type of data required to utilize a given method. Computational methods are used for modeling small molecules and macromolecules on various levels, as well as for mining chemical data, analyzing and predicting protein-ligand interactions. This predictive capability of CADD is particularly valuable in the early stages of drug discovery, where it can guide the design of new molecules with desired properties, or optimization of found hit molecules[46].

Computer-aided methods are fast, efficient, and straightforward tools for identifying new molecules that bind to a specific protein site and for optimizing this process by elucidating its mechanism. By predicting and optimizing small molecules' interactions with macromolecular biological targets, they accelerate the drug discovery process, reduce associated costs, and enhance the success rate of identifying effective molecules[47].

1.5.1. Ligand-based virtual screening

1.5.1.1. Molecular structure representation in chemoinformatics

Chemoinformatics manages, interprets, and extracts knowledge from chemical data, requiring digital representation of chemical structures. This digital representation is critical to the success of chemoinformatics methodologies, and it is typically achieved through several levels of representation.

At the primary level, molecules can be represented as molecular graphs[46]. In this representation, nodes represent atoms, and edges correspond to chemical bonds. However, a graph

is inherently a topological concept, which, while easily understood by humans, is not well-suited for direct algebraic operations. This limitation arises from the graph representation's dependence on an arbitrary atom numbering scheme. Consequently, a secondary level of chemical structure representation, known as molecular descriptors, is employed. In this approach, structural information is extracted from the molecular graph and encoded in a numerical format, typically a vector of numbers, $\mathcal{D}_i$, where each component i represents a specific structural feature. These descriptor vectors are particularly effective for computational analysis. Success of any chemical structure-dependent modeling depends on the inclusion of property-relevant information in the descriptor vector $\mathcal{D}$.

Molecular descriptors can be categorized into three main types[48]. 1D descriptors, derived from the molecular formula, capture bulk properties and physicochemical parameters such as molecular weight and atom count. 2D descriptors, calculated from the 2D molecular graph, capture information on atoms connectivity. Examples include topological indices, fragment descriptors, and topological fingerprints. Topological indices capture the structural information related to the connectivity and arrangement of atoms and bonds within a molecule, providing a numerical representation of its overall topology. Fragment descriptors encode the structure of a molecule into a vector, with each index corresponding to a predefined structural feature. The presence or absence of a feature is indicated either by a total count of times this fragment is present in the molecule, or by a binary value of 1 or 0, respectively. Topological fingerprints, on the other hand, do not require a predefined fragment library. They are generated by enumerating all possible fragments within a molecule that are not larger than a certain size and converting these fragments into numeric values. 3D descriptors are derived from the 3D structure of the molecule and capture information on geometric, electronic, and thermodynamic properties, thereby providing a comprehensive representation of its spatial configuration and physicochemical characteristics[49].

In this study, we specifically used ISIDA fragment descriptors, a type of 2D fragment descriptors[50]. These descriptors encode a compound's structure by counting the occurrence of different substructural fragments, which could be linear sequences, augmented atoms (central atoms with their environment), or triplets that encode the compound's atoms and/or bond types. Additionally, these fragments can be colored to provide extra information, such as pharmacophore types of atoms, formal charges, and force-field atom types.

Given that the same chemical compound can be represented in multiple ways due to the presence of tautomers, isomers, and varying charge states, standardization is crucial to ensure that identical or similar structures are recognized as such, regardless of their initial representation. Chemical structure standardization is thus an integral part of any modelling pipeline[51]. It refers to the process of transforming and normalizing chemical structures into a consistent format. This

process may involve several steps, such as the removal of salts, the normalization of specific functional groups, the correction of bond types, and the generation of a canonical tautomeric form. This consistency is vital for many chemoinformatics operations. By standardizing chemical structures, we can ensure that these operations are performed on a consistent and meaningful representation of the chemical space, thereby enhancing the reliability and interpretability of the results.

1.5.1.2. Similarity search

Similarity search is a computational method employed in drug discovery to identify molecules with similar properties to known active compounds. It is based on the molecular similarity principle, stating that structurally similar molecules are statistically likely to have similar properties[52]. In the context of drug discovery, if two molecules have similar structures, they are likely to interact with biological systems in similar ways, and thus, may have similar therapeutic profiles.

Similarity search involves comparing the structural and chemical characteristics of different molecules, most often using fingerprints, which are bit string representations of molecular structure and properties[53]. Each bit in the fingerprint represents the presence or absence of a particular structural feature within the molecule. It is possible to use other molecular descriptors, too. By ranking compounds based on their similarity scores, similarity search allows for the efficient screening of large compound libraries to find potential drug candidates in the early stages of drug discovery. Commonly used similarity metrics include the Tanimoto coefficient, Dice coefficient, Euclidean distance, or Manhattan distance [54]. The ideal combination of descriptors and metric function is the one that ensures the best "Neighborhood Behavior Compliance".[55] This means minimizing situations where pairs of compounds appear very similar despite having different property values, known as "property cliffs".

Applying similarity search to large compound libraries requires efficient algorithms and appropriate hardware due to computational considerations. The results of a similarity search can be utilized to predict the biological activity of a compound by comparing its fingerprint to those of known active compounds (e.g., by means of the *k* nearest neighbors approach). However, determining the appropriate similarity threshold values for different fingerprint types and compound classes remains a challenge, as similarity search calculations are influenced by compound class-dependence and database composition[53].

Despite its limitations, such as the inability to handle structure-activity relationship discontinuity, similarity search in virtual screening offers advantages by providing a holistic molecular view and allowing for screening in the absence of detailed knowledge about activity-

determining features. It accelerates the drug discovery process by reducing the number of compounds requiring experimental testing [54].

In this work, similarity search was part of the pipeline used to discover novel todalam binding site-targeting molecules (Chapter 4), as well as to find commercially-available compounds structurally similar to *de novo* generated colchicine site-targeting agents (Chapter 5).

1.5.1.3. Substructure search

Substructure search is a fundamental operation that involves the identification of specific molecular fragments or patterns within larger molecular structures[56]. Patterns for substructure search are encoded using the SMARTS language[57] that provides flexibility in the search criteria, as the substructures can be defined in various ways to capture different chemical features. Furthermore, it allows for the definition of complex substructural patterns, which can be useful in identifying compounds with specific structural characteristics [56,58].

The key principles that underpin the concept of substructure search are rooted in graph theory. Molecules and their substructures can be represented as graphs, where atoms are nodes and bonds are edges. The substructure search problem is then translated into a subgraph isomorphism problem, which involves finding a one-to-one correspondence between the nodes and edges of the substructure graph and a subset of the nodes and edges of the molecule graph[58].

Substructure search is used in virtual screening because it helps to efficiently filter and select compounds from large databases based on the presence of specific functional groups or substructures. This is particularly useful if certain substructures are known to be related to desirable properties, such as binding affinity to a target protein. By identifying compounds that contain these substructures, researchers can prioritize certain compounds for further testing and analysis.[58]

However, substructure search also has its limitations. The performance of the search can be affected by the size of the molecules and the complexity of the substructures. Large molecules and complex substructures can increase the computational cost of the search. Additionally, the search is sensitive to the way the substructures are defined. Different formulations of the same substructure can lead to different search results. Therefore, careful consideration must be given to the definition of the substructures to ensure that the search results are meaningful and relevant.[56,58]

In conclusion, substructure search is a powerful tool in chemoinformatics and virtual screening, providing a means to efficiently navigate large compound databases and identify potential drug candidates. Despite its limitations, its benefits in terms of speed, flexibility, and the ability to capture complex chemical features make it an indispensable part of the drug discovery process.

In this work, substructure search was performed as part of the virtual screening pipeline designed to discover novel inhibitors of the todalam binding site (Chapter 4).

1.5.1.4. Quantitative structure-activity relationship (QSAR) modeling

Quantitative structure-activity relationship (QSAR) modeling is a computational method establishes a mathematical relationship between chemical structure and some property of interest. This method helps rationalize underlying relationships between molecular structure and property within a series of molecules. It can also be used to screen virtual libraries of compounds by predicting their properties and using the predicted value as a filter [48].

To perform QSAR modeling, one needs a dataset comprising molecular structures and their corresponding experimental property data. The data should be carefully curated before modeling, ensuring the removal of duplicates, structure standardization (transformation of tautomeric and resonance forms into a single form, the neutralization of charges, and the removal of small fragments from salts), the verification of the accuracy of primary data, and the transformation of biological data into a form suitable for mathematical modeling[51]. In particular, it involves encoding the chemical information of the modelled compounds in the descriptor form.

Classification and regression modeling are two key approaches in QSAR modeling. Classification modeling is used when the target variable is categorical, such as an "active"/"inactive" label, while regression modeling is used when the target variable is continuous, for example – the IC_{50} value. The choice of machine learning methods for these modeling tasks varies and depends on the character of the target property[59].

The performance of a QSAR model can be assessed using various metrics. For classification models, the balanced accuracy metric (BA) is often used. For regression models, the coefficient of determination (R^2) metric is commonly used[48]. Both metrics are discussed in detail in Chapter 6.

K-fold cross-validation is a technique often used in QSAR modeling to estimate the predictive performance of a model. It involves partitioning the data into subsets, training the model on a portion of the data, and then testing it on the remaining data. This process is repeated multiple times with different partitions. Cross-validation is crucial in QSAR modeling as it provides a robust estimate of the model's predictive performance and helps prevent overfitting[51].

The applicability domain of a QSAR model refers to the chemical space within which the model can make reliable predictions. It can be assessed using various methods, such as the leverage approach or the distance-based approach. Assessing a model's applicability domain is crucial to ensure that predictions are only made within the domain for which the model is valid[60].

QSAR modeling in virtual screening has both advantages and disadvantages. On the one hand, it can significantly reduce the cost and time of drug discovery by predicting the target property of compounds before synthesis or testing. Additionally, it can be used in virtual screening to filter large compound sets by the predicted values of the target property. On the other hand, the accuracy of QSAR models depends on the quality of the input data and the choice of descriptors and modeling techniques. Furthermore, QSAR models are limited to their applicability domain and may not make accurate predictions for compounds outside this domain.

In this thesis, QSAR modeling was performed to model anti-proliferative activity of colchicine site-targeting compounds against HeLa cells in Chapter 5, as one of the steps in *de novo* inverse QSAR drug design pipeline. It was also used in Chapter 6 as a benchmark for the application of the transfer learning approach. The descriptor calculation steps were performed using the RDKit chemoinformatics toolkit, as well as the ISIDA Fragmentor tool. The machine learning and cross-validation steps were performed using the scikit-learn Python package.

1.5.1.5. Transfer learning

Transfer learning, a concept in machine learning, has been applied to drug discovery to address the challenge of identifying suitable descriptors for downstream modeling tasks. In essence, transfer learning is a method where a pre-trained model is adapted for a new, but related task. It allows the application of knowledge gained while solving one problem to a different but related problem. This is particularly useful in situations where the data for the task of interest is scarce or when the task is too complex to be learned from scratch [61,62].

The process of transfer learning in drug discovery involves two main steps: pre-training and fine-tuning. Pre-training is the initial phase where a model is trained on a large dataset to learn a general representation of the data. In the context of drug discovery, the pre-training phase involves training a model on a large dataset to learn general representations of molecules. This is achieved through self-supervised learning, where the model learns to predict some aspect of the data from other parts of the same data. This is often done using self-supervised learning, where the model learns to predict parts of the input data from other parts, thereby learning useful representations of the data. This process allows the model to learn useful representations of molecules without needing labeled data for the specific task at hand[63].

Following pre-training, the model undergoes a fine-tuning process. Fine-tuning is the adjustment of the pre-trained model to make it more suitable for the specific task at hand. This is done by continuing the training of the model on the specific task data, allowing the model to adapt its learned representations to the specific characteristics of the new task. In the context of drug discovery, fine-tuning can help increase the predictive performance of a QSAR model on a

downstream task by adapting the general molecular representations learned during pre-training to the specific properties relevant to the QSAR task at hand[64].

To successfully implement transfer learning in drug discovery, certain requirements need to be met. Firstly, a large dataset for pre-training is needed to learn general molecular representations. This dataset should ideally cover a wide range of chemical space. Secondly, task-specific data is required for fine-tuning the model. This data should be relevant to the specific task at hand and should ideally contain examples of the specific properties or activities that the model needs to predict. Lastly, a suitable deep learning model that can learn from the pre-training data and adapt to the fine-tuning data is required. This model should be capable of learning complex patterns and relationships in the data, and should be flexible enough to adapt its learned representations to the specific task.

In conclusion, transfer learning offers a promising approach to address the challenge of identifying suitable descriptors for downstream modeling tasks in drug discovery. By leveraging the power of deep learning and the concept of transfer learning, it is possible to learn useful molecular representations from large datasets and adapt these representations to specific tasks, thereby improving the predictive performance of models in drug discovery.

In this thesis, Chapter 6 discusses the application of transfer learning to downstream QSAR modeling tasks and its comparison to state-of-the-art approaches.

1.5.1.6. Inverse QSAR-based de novo ligand generation

De novo drug design is a strategy in drug discovery that involves the design of novel molecules with desirable properties from scratch, rather than relying on the modification of pre-existing molecules. This approach offers several advantages. Firstly, it allows for the exploration of a vast chemical space, potentially leading to the discovery of novel drug candidates that would not be identified through the modification of known molecules. Secondly, it can be guided by the target structure, allowing for the design of molecules that are specifically tailored to interact with the target in a desired manner. Lastly, it can be automated, making it a highly efficient method for drug discovery[65].

Inverse QSAR modeling, on the other hand, is a process that reverses the conventional QSAR modeling approach. Instead of mapping a set of descriptors to a target property, inverse QSAR modeling maps a target property to a required compound via a set of descriptor values. The purpose of this approach is to identify descriptor values that correspond to optimal properties, and then to generate molecules that possess these descriptor values. This is challenging due to the difficulty of reconstructing a molecule from descriptor values. However, the development of autoencoder neural network models has made this process feasible[66].

An autoencoder is a type of artificial neural network that is trained to encode input data into a lower-dimensional representation and then decode this representation back into the original data. In the context of chemoinformatics, an autoencoder can be trained to learn the correspondence between chemical structures and a latent chemical space. This is achieved by training the autoencoder on a large dataset of chemical structures, allowing it to learn how to encode these structures into points in the latent space and decode points in this space back into chemical structures[66].

The process of performing inverse QSAR modeling involves several steps. Firstly, a QSAR model is built to identify the relationship between the descriptors and the target property. Then, compounds whose descriptor vectors are predicted to correspond to high affinity values are found. A trained autoencoder model is then used to generate molecules that correspond to these descriptor values. This process involves the addition of random noise to the values of a point in the latent space, allowing for the sampling of multiple molecules from the region around the selected seed molecule[66].

In the context of *de novo* drug design, the application of inverse QSAR modeling offers a powerful approach for the generation of novel drug candidates. By first identifying descriptor values that correspond to optimal properties, and then using an autoencoder to generate molecules that possess these descriptor values, it is possible to design molecules that are specifically tailored to exhibit desirable properties. This approach, therefore, offers a highly efficient and targeted method for drug discovery, allowing for the exploration of a vast chemical space and the identification of novel drug candidates that may not be discovered through traditional methods[66].

In this thesis, Chapter 5 discusses the application of inverse QSAR methodology to tailored *de novo* drug design of colchicine site-targeting agents.

1.5.1.7. Forward- and retrosynthesis route prediction

Retrosynthesis, one of the core tasks in reaction informatics, is a method used to plan the synthesis of organic molecules by deconstructing them into commercially available precursors. The process begins with a target molecule and involves the iterative application of reaction rules to break down the molecule into simpler components. These reaction rules, also known as transformations, are derived from known chemical reactions and reflect which bond needs to form/break and what products are formed as a result of that. The process continues until commercially available starting building blocks are identified, or a limit of steps is reached[67].

The retrosynthesis process is guided by reaction mapping, which is a representation of how atoms rearrange in reactions. Such mapping is often used to automate the labeling of reactants and

products, and plays a crucial role in understanding the transformation of molecules during a reaction[67].

To navigate the retrosynthetic trees generated during this process, a technique known as Monte-Carlo tree search (MCTS) is employed. MCTS is a heuristic search algorithm used in decision-making processes. In the context of retrosynthesis, it is used to explore the space of possible synthetic routes, making decisions based on the expected outcome of the reactions[68].

On the other hand, the reaction prediction task is centered on predicting a possible product of a chemical reaction in one or more reactants[67]. In one of the implementations, it involves creating a virtual reactor where virtual building blocks are iteratively subjected to chemical transformation rules until a desired target molecule, or a close analog, is obtained. This process is guided by similarity metrics and is crucial for determining if a molecule that showed up in screening or *de novo* generation is actually synthesizable and how much it would cost to make it[68].

To predict if a reaction between two reagents will proceed, machine learning models are trained on experimental procedures extracted from patent data. These models expect the input reaction to specify all the species involved in the reaction, including solvents and catalysts. The models capture the functional dependence between the input parameters and the reaction outcomes, and they are highly specific for a single reaction or a family of reactions[67].

Knowing the synthesis steps required to get to a target molecule with desired properties is immensely useful. It not only aids in the discovery of new molecules and materials but also accelerates the R&D processes in academia and across chemical and pharmaceutical industries. This knowledge can lead to more efficient and cost-effective drug design, ultimately contributing to the advancement of medicine and healthcare.

In this project, both forward and retrosynthesis approaches were used in the project dedicated to design and discovery of novel maytansine site-targeting ligands, Chapter 2.

1.5.2. Structure-based virtual screening

1.5.2.1. Pharmacophore screening

A pharmacophore model represents a set of essential steric and electronic features that ensure optimal supramolecular interactions of a ligand with a specific biological target. These interactions can either activate or inhibit the biological response of the target[69].

There are several common pharmacophore feature types, which include hydrogen bond acceptors and donors, charged or ionizable groups, hydrophobic residues, and aromatic rings. These features reflect the concept of bioisosterism, acknowledging that different functional groups may exhibit similar physicochemical properties[69].

In a three-dimensional pharmacophore model, the features have specific spatial relationships with each other, captured as distances or distance ranges between the features. The spatial coordinates of the features are typically supplemented with a spherical tolerance region to account for distance variability[70].

The source of data for generating a pharmacophore model can vary. There are two common methods: structure-based modeling, which relies on the three-dimensional structure of a ligand-protein complex, and ligand-based modeling, which depends solely on the structural information of active compounds. This work extensively employed structure-based pharmacophore modeling.

Several tools have been developed for pharmacophore modeling and screening, with Catalyst, MOE, Phase, and LigandScout among the most commonly employed for lead discovery. Despite differences in their respective screening algorithms, these tools share a common approach: they utilize a pharmacophore model as a query to screen databases of small molecules' 3D structures[70].

Before screening, each molecule in the database is represented by a set of conformers, which potentially include the bioactive geometry assumed during interaction with the target protein. Thus, the quality and robustness of conformational sampling performed during database preparation has a strong influence on the quality of screening results. The resulting matches between the pharmacophore model and the conformers are compiled into a hit list. Depending on the selectivity and rigor of the model, a virtual screening of chemical databases containing millions of small molecules can yield from tens to thousands potential hits[71].

A scoring system is then employed to rank the molecules in the hit list. This score quantifies the quality of the match between each molecule and the pharmacophore model, providing a measure of the potential fitness of each molecule as a drug candidate[69].

In this work, pharmacophore screening was performed using the LigandScout software (v. 4.4.8). LigandScout, developed by Inte:Ligand GmbH, is a powerful tool that allows users to automatically generate a feature-based pharmacophore model from a ligand-target complex structure. This process can be performed using either a co-crystallized or docked complex[70,71].

The pharmacophore model creation begins with ligand perception, a two-step process that involves the interpretation and assignment of the ligand's molecular information. This information, which includes hybridization status and bond characteristics, is often not explicitly defined in the input data files, particularly those in the Protein Data Bank file format[71].

The next phase involves generating feature-based pharmacophore models by identifying interactions between the ligand and target atoms. The structure of both the ligand and the binding pocket undergoes a thorough analysis to identify atoms and groups capable of participating in a variety of interactions. These include hydrogen bonding, hydrophobic, aromatic, ionic, and metal

binding interactions. The decision to include a feature in the final pharmacophore model is guided by its location relative to corresponding features within the binding site. Once all complementary feature pairs within the complex have been identified, their corresponding ligand-side features are incorporated into the pharmacophore model. The final step in this process involves the addition of exclusion volume spheres to the model. These spheres mimic the shape of the binding pocket, adding another layer of spatial detail to the model and contributing to its ability to accurately represent the molecular interaction landscape[69,71].

After the initial model is created, it can be further refined using binding data to increase its predictive accuracy. Alternatively, multiple models can be merged to form a single, comprehensive feature pharmacophore model. The combined model offers the potential for a more complete and holistic representation of the pharmacophore[71].

In LigandScout, the quality of alignment is evaluated using four distinct scoring functions. The first of these, the pharmacophore fit score, is a straightforward geometric scoring function. This score prioritizes solutions with a high number of geometrically matched feature pairs, while solutions with higher root mean square deviations among these pairs receive penalties[71].

The second scoring function, the atom sphere overlap score, quantifies the overlap of atom van der Waals spheres. The Gaussian shape similarity score, the third function, measures the overlap of Gaussian function representations of molecular volume[71].

The final scoring function is a combination of the pharmacophore fit and atom overlap scores. In the present study, the pharmacophore fit score was used as the default scoring function (Equation 1).

$$S_{FCR} = c \cdot N_{MFP} + \left(9 - 3 \cdot \min(RMS_{FP}, 3)\right) \quad (1)$$

In Equation 1, S_{FCR} is the feature count/RMSD distance score, c is a weighting factor for the number of matched feature pairs, N_{MFP} is the number of geometrically matched feature pairs, RMS_{FP} is the root mean square deviation of the matched feature pair distances.

To validate the created pharmacophore models, a dataset of ligands with known activity levels can be used. This validation dataset is tested against the pharmacophore model, and statistical metrics are calculated to assess how effectively the model distinguishes between active and inactive compounds. Once validated, the model is ready to be applied in virtual screening processes[69].

The list of pharmacophore features available in LigandScout is shown in Figure 27. Each feature in the generated pharmacophore can be labeled optional, meaning it is not obligatory for a valid alignment. Therefore, during the screening process, even if an optional feature is not matched, the molecule can still be considered a valid hit.

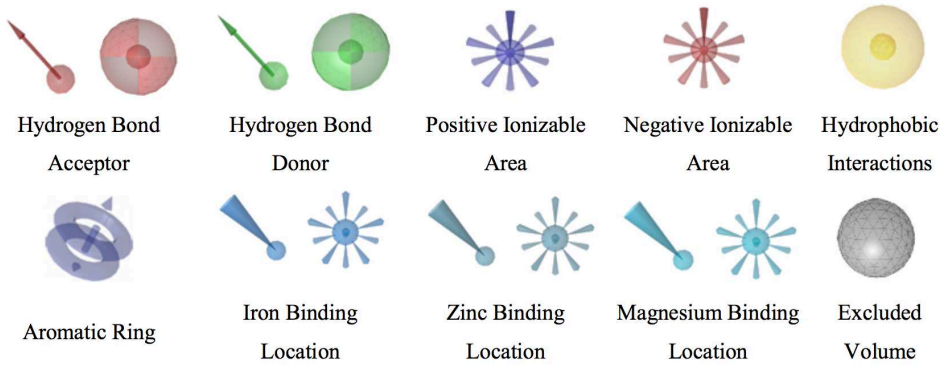


Figure 27. Pharmacophore features implemented in LigandScout

Pharmacophore modeling and screening played was an important method used throughout the whole thesis. Of particular importance is its application to discovery of novel maytansine site-targeting ligands (Chapter 2) and todalam site-targeting agents (Chapter 4).

1.5.2.2. Binding site similarity search

Application of binding site similarity search to drug design is based on the assumption that similar binding sites may accommodate similar ligands. The focus on binding sites is due to the increasing knowledge on ligand bioactivity data. Binding sites can be represented on a computer in two primary ways: sequence-based and structure-based[72].

Sequence-based representation involves deriving important residues lining the ligand-binding site from a set of aligned protein sequences from the same family. These residues are then mapped onto a reference 3D structure, and consensus binding site amino acids are concatenated into a gapless cavity sequence. A binding site-based phylogenetic tree is then derived from computed distances between cavity sequences using either sequence identity (homology) or physicochemical properties as a distance metric[72].

On the other hand, structure-based representation of binding sites involves defining a binding site from residues interacting with a particular compound. This definition can vary depending on whether global or local similarities are desirable. For instance, a cavity will encompass all protein residues potentially accessible to a ligand at the protein surface, whereas a specific binding site is only defined from residues interacting with a particular compound[72].

Comparing binding sites with each other involves the use of various algorithms[73]. Generally, they can be grouped in five groups. Clique-based methods, the first group, are primarily based on the concept of identifying maximal cliques in a graph, where a graph represents a protein's binding site. The nodes of the graph correspond to the atoms or residues in the binding pocket, and the edges represent the spatial relationships between them. The main idea behind these methods is to find the largest common subgraph between two protein binding sites, which

corresponds to the largest clique in the product graph. This approach uses the representation of Cα atoms or functional groups in the binding pocket.

The second group of methods solves the assignment problem. The main idea here is to assign residues or atoms in one binding pocket to those in another pocket, aiming to maximize the overall similarity or minimize the total cost. These methods often use the Hungarian algorithm or its variants to solve the assignment problem. The pocket representation used by these methods includes Cα atoms or Cα-Cβ vectors[73].

The third group of methods combines the clique detection and the assignment algorithm. The main idea is to use the strengths of both approaches to achieve better performance. They first detect cliques in the graph representation of binding pockets and then solve the assignment problem within these cliques. The pocket representation used by these methods includes chemical feature points or Cα atoms and Cα-Cβ vectors[73].

The fourth group of methods employs geometric hashing and sorting. The main idea behind these methods is to use geometric hashing to index the features of binding pockets and then use sorting to quickly find similar features. The pocket representation used by these methods includes N, Cα, C, O, Cβ and side-chain centroid atoms or microenvironments[73].

The fifth group of methods employs the rotational and translational search. The main idea is to rotate and translate one binding pocket in the three-dimensional space to find the best match with another pocket. This approach often involves a comprehensive search in the rotational and translational space, which can be computationally intensive. These methods represent the binding pockets as ensembles of non-hydrogen atoms[73].

The process of performing 3D binding site similarity search thus involves several steps. First, the binding site of interest is represented in a format suitable for the required alignment method. Then, pocket alignment and similarity search is run to identify similar binding sites. The ligands that bind to these similar binding sites are then analyzed, and this information is used to design new ligands that can target the original binding site of interest[72].

In this work, binding site similarity search was used as one of the first steps to the design of novel todalam site-targeting molecules, described in Chapter 4.

1.5.2.3. Protein-ligand docking

Protein-ligand docking is widely-used computational approach to estimate how a ligand interacts with a specific protein binding site. This process is central to understanding receptor-ligand interactions and the mechanisms of drug action, as it aids in predicting the ligand's binding pose and roughly estimating its binding affinity[74].

A docking program typically includes two components: the sampling algorithm and the scoring function. The sampling algorithm is tasked with generating a large number of potential orientations and conformations of the ligand within the protein's binding site. The goal of a scoring function then is to predict the binding affinity of each ligand orientation or conformation using either a physical or an empirical energy function. The pose with the lowest energy score is predicted to be the "best match", i.e. the most probable binding pose[75].

Ligand sampling involves the exploration of a multitude of conformational degrees of freedom, even for relatively simple organic molecules. Thus, the accuracy of this exploration is vital in identifying the conformation best suited to the receptor structure, and the process must be rapid enough to evaluate thousands of compounds efficiently. It is important to note that ligand binding often triggers changes in protein conformation as well. These can range from minor side-chain rearrangements to large domain motions. Given the extensive size and multiple degrees of freedom inherent in proteins, the modeling of protein flexibility represents a significant challenge in molecular docking[76]. Protein-ligand docking methods used in this work considered the protein backbone to be rigid and did not account for flexibility of the side chains.

Speed and accuracy of the chosen scoring function are also important characteristics of a protein-ligand docking pipeline. There are four main types of scoring functions: empirical, knowledge-based, force-field methods, and machine learning-based[77].

Empirical scoring functions are equations that include various terms representing physicochemical properties that are known to influence drug binding. These terms generally describe polar and apolar interactions, the loss of ligand flexibility, and desolvation effects. However, these functions require a training set to determine the weight factors of individual energy terms, which is a significant drawback[77].

Force-field methods, on the other hand, rely on the non-bonded terms of a classical molecular mechanics force field. Force fields are mathematical models that estimate the energy and forces acting on atoms within a molecule, thereby facilitating the simulation of the molecule's physical behavior and properties. The underlying idea of force fields is to approximate the potential energy surface of a molecular system using a series of empirical equations that describe the interactions between atoms, including bond stretching, angle bending, and non-bonded interactions, thereby providing a means to predict the molecule's geometric and energetic properties[77].

The main drawback of force-field calculations is the exclusion of the entropic component of the binding free energy[77].

Knowledge-based scoring functions use structural information gathered from resolved protein-ligand system coordinates to encode the free interaction energies of protein-ligand atom

pairs. The score is thus the sum of all interatomic interactions in the protein-ligand complex. Primary limitation of such scoring functions is that their derivation is largely reliant on information implicitly encoded in limited sets of protein-ligand complex structures[77].

Machine learning-based scoring functions present an emerging and promising approach in molecular docking. They mine the relevant physicochemical patterns from available protein-ligand complex data without explicit programming or derivation of rules, hence avoiding the bias in predicting a binding affinity value. However, these methods largely depend on the initially chosen representation of the training set protein-ligand complexes and are limited in their domain of applicability by the contents of the training set[77].

In addition to the challenges mentioned above, another issue is the preparation of protein binding sites for docking simulations. Issues may arise due to the low resolution of crystallographic structures of proteins and their complexes, the positioning of nitrogen and oxygen atoms in side chains of asparagine and glutamine residues, determining the correct tautomeric state of some amino acid residues, and the appropriate positioning and orientation of water molecules, which can participate in protein-ligand interactions. As such, these factors need to be taken into account to ensure the reliability and accuracy of docking simulations[74].

One of the projects performed as part of the thesis work also involved covalent protein-ligand docking. This approach is distinct from conventional, unconstrained protein-ligand docking, where the ligand is free to rotate within the binding site. In contrast, covalent docking involves a constrained rotation around a fixed point in three-dimensional space, specifically around the reacting residue and the warhead of the ligand. However, the underlying sampling and scoring techniques are the same. The warhead, an electrophilic group within the ligand, and the target residue, a nucleophilic component of the protein, are central to the process of covalent docking. Covalent docking is particularly useful in drug discovery due to its ability to model and predict the behavior of covalent binders. These molecules, which form a covalent bond with their target, can offer unique activity profiles compared to non-covalent ligands[78].

It is possible to assess the accuracy of docking via a benchmarking process that measures the root mean square deviation of atomic positions between the poses of the crystallized and docked ligand, obtained through re-docking and cross-docking procedures. Re-docking involves reintroducing the ligand, extracted from the X-ray structure of a protein-ligand complex, back into its original binding site. Cross-docking, on the other hand, is placing the native ligand of a protein-ligand complex into a slightly different shaped protein binding site, either from another complex or from the ligand-free protein. A satisfactory benchmark is typically indicated by RMSD values that are less than 2Å[74].

This work used two docking programs, namely PLANTS[79], AutoDock 4[80], and AutoDock GPU[81], a version of AutoDock that supports GPU acceleration for conformational sampling. In all three approaches the ligand is flexible and the protein is kept rigid. The software differs in the underlying sampling algorithms and scoring functions, as well as docking system preparation routines.

The PLANTS docking software represents a protein's binding site by including all protein atoms within a certain distance from the geometric center of a ligand. To normalize raw PDB structures of complexes and determine the tautomeric and protonation states of binding site amino acids, a complementary Structure Protonation and Recognition System (SPORES) software is utilized. It can also be used to prepare the ligand structure in a similar manner[79].

The docking algorithm employed by PLANTS is built on a stochastic optimization algorithm known as ant colony optimization (ACO). The principle underlying ACO is drawn from the natural behavior of ants in their quest for the shortest path between their nest and a food source. As ants move, they deposit pheromones to denote the paths they've already followed. When faced with a choice between multiple paths, they are more likely to select the ones marked with a higher concentration of pheromones. ACO algorithms emulate this behavior using virtual pheromones, which are represented as numerical values associated with each possible conformation. The algorithm then gradually converges on ligand poses with the most number of favorable interactions, thereby optimizing ligand conformation[79].

PLANTS has two empirical scoring functions: $PLANTS_{PLP}$ and $PLANTS_{CHEMPLP}$. Both functions are built on elements of previously published scoring functions and force fields, with the piecewise linear potential (PLP) scoring function utilized to model the steric complementarity between the protein and the ligand. $PLANTS_{CHEMPLP}$ (Equation 2) introduces angle-dependent terms for hydrogen bonding and metal binding, drawing from the terms of GOLD's Chemscore implementation. Additionally, the torsional potential from the Tripos force field, along with a heavy-atom clash term, is employed to account for intra-ligand interactions. These components collectively contribute to the robustness and efficacy of the PLANTS docking software in the realm of drug design. In this work, the $PLANTS_{CHEMPLP}$ scoring function was used[82].

$$PLANTS_{CHEMPLP} = f_{plp} + f_{hb} + f_{hb-ch} + f_{hb-CHO} + f_{met} + \\ + f_{met-coord} + f_{met-ch} + f_{met-coord-ch} + f_{clash} + f_{tors} + C_{site} \quad (2)$$

In Equation 2, f_{plp} is the piecewise linear potential; f_{hb} is the potential for the donor-acceptor pairs; f_{hb-ch} is the potential for the charged donor and charged acceptor pairs; f_{hb-CHO} is the potential for the hydrogen bonding pairs containing an oxygen-acceptor; f_{met} is the distance- and angle-dependent potential; $f_{met-coord}$ is the metal coordination potential; f_{met-ch} is the potential

for charged acceptor atom, involved in a metal interaction; $f_{met-coord-ch}$ is the potential for charged acceptor atom, involved in a metal interaction; f_{clash} is the empirical heavy-atom potential; f_{tors} is the torsional potential; C_{site} is the quadratic potential.

AutoDock 4 is a widely-used docking program that employs a free energy force field to evaluate conformations during docking simulations. The force field was parameterized using a large number of protein-inhibitor complexes for which both structure and inhibition constants are known. The force field evaluates binding in two steps. The ligand and protein start in an unbound conformation. In the first step, the intramolecular energies are estimated for the transition from these unbound states (based on user input) to the conformation of the ligand and protein in the bound state. The purpose of this initial assessment is to provide a baseline for comparing the energies of the bound states that are generated during the docking process. The second step then evaluates the intermolecular energies of combining the ligand and protein in their bound conformation. The force field scoring function of AutoDock 4 includes six pair-wise evaluations (Equation 3) and an estimate of the conformational entropy lost upon binding. Each of the pair-wise evaluations include energy terms for dispersion/repulsion, hydrogen bonding, electrostatics, and desolvation (Equation 4)[80].

$$\Delta G = (V_{bound}^{L-L} - V_{unbound}^{L-L}) + (V_{bound}^{P-P} - V_{unbound}^{P-P}) + (V_{bound}^{P-L} - V_{unbound}^{P-L} + \Delta S_{conf}) \quad (3)$$

L – ligand, P – protein.

$$V = W_{vdw}\sum_{i,j}\left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}\right) + W_{hbond}\sum_{i,j}E(t)\left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}}\right) + W_{elec}\sum_{i,j}\frac{q_i q_j}{e(r_{ij})r_{ij}} + W_{sol}\sum_{i,j}\left(S_i V_j + S_j V_i\right)e^{(-r_{ij}^2/2\sigma^2)} \quad (4)$$

In Equation 4, first term is a typical 6/12 potential for dispersion/repulsion interactions; second term is a directional H-bond term based on a 10/12 potential, with *C* and *D* parameters assigned to control energies for hydrogen bonds with oxygen, nitrogen, and sulfur; the *E(t)* function introduces directionality of a hydrogen bond based on the angle t from an ideal H-bonding geometry; the third term is a Coulomb electrostatic potential; finally, the last term is a desolvation potential based on the volume of atoms (V) that surround a given atom and shelter it from solvent, weighted by a solvation parameter (S) and an exponential term with distance-weighting factor σ = 3.5Å.

AutoDock 4 employs the concept of a grid map, which is a three-dimensional grid where the protein is embedded, and a probe atom is placed at each grid point. The energy of interaction of this single atom with the protein is assigned to the grid point. AutoGrid affinity grids are calculated for each type of atom in the ligand, typically carbon, oxygen, nitrogen, and hydrogen,

as well as grids of electrostatic and desolvation potentials. This grid map concept allows AutoDock 4 to make rapid energy estimations during the docking process. The energetics of a particular ligand configuration is evaluated using the values from the grids[80].

AutoDock 4 uses a genetic algorithm-based conformational sampling algorithm, specifically a Lamarckian genetic algorithm. A Lamarckian genetic algorithm integrates the principles of natural selection with the idea of inheritance of acquired traits. The algorithm begins with an initial population of potential solutions, each represented as a *chromosome* – a string of parameters that define the solution. This population undergoes a process of evolution over a series of iterations (*generations*). In each generation, a *fitness function* is used to evaluate the performance (*fitness*) of each individual in the population. The fittest individuals are then selected to create offspring for the next generation through operations that mimic biological processes: crossover (or recombination), where parts of two parent chromosomes are combined to create a new offspring chromosome, and mutation, where random changes are introduced to a chromosome. This iterative process of selection, crossover, mutation, and local search continues until a satisfactory solution is found or a stopping criterion is met. In the case of protein-ligand docking, chromosomes encode ligand conformations, and their selection is based on the values of the scoring function[80].

It is also possible to use AutoDock 4 for covalent protein-ligand docking experiments. The first approach involves specifying a restricting grid map around the warhead atom that would be covalently attached to a nucleophilic residue. This would penalize the movement of the atom away from the initially defined coordinates, essentially fixing it in place. The second approach considers a bound ligand as an extension of the binding site residue to which it is bound, and samples conformations of the new residue within the pocket. These two approaches provide flexibility in modeling covalent interactions, expanding the range of docking simulations that can be performed with AutoDock 4[78].

AutoDock GPU is recently released implementation of AutoDock designed to exploit both GPU and CPU parallel architectures. The conformational search is performed using either the original random optimizer Solis-Wets (SW) or the newly implemented ADADELTA gradient-based local search function. The key difference between AutoDock GPU and its predecessor, AutoDock 4, lies in their computational efficiency, as AutoDock GPU is designed to exploit the parallel nature of docking and the underlying algorithms, outperforming AutoDock 4 by a factor of 30 times. This significant speedup is achieved without compromising the accuracy of the docking results. AutoDock GPU uses the same free-energy force field scoring function as AutoDock 4. However, while AutoDock 4 uses precalculated and cached interaction energy maps

for ligand atoms, AutoDock GPU calculates exact values by evaluating the analytical form of the scoring function[81].

Protein-ligand docking was instrumental in most of the projects described throughout the thesis.

1.5.2.4. Gaussian-accelerated molecular dynamics simulations

Molecular dynamics (MD) simulation is a computational method that simulates the movement of atoms in protein systems. These simulations are crucial in understanding the behavior of biological systems in real life, where they are in constant motion. The fundamental idea behind MD simulations is to calculate the force exerted on each atom by all other atoms, using this information to predict the spatial position of each atom over time. Accuracy of such simulations relies on the used force fields – mathematical models that describe the interactions between atoms[83].

To prepare a system for MD simulations, a three dimensional protein structure is placed in a simulation box, and solvent molecules, typically water, are added, along with ions to neutralize the system. This is crucial to mimic the natural environment of the protein. The simulation process involves two main steps: equilibration and production. Equilibration is the initial phase where the system is allowed to stabilize, while the production phase is where the actual simulation is run[83].

When performed over suitable timescales, molecular dynamics simulations can yield a wealth of information. They can provide insights into protein function and interaction with small molecules, which helps in elucidating the mechanisms underlying the behaviour of targeted biological systems[83].

Conventional molecular dynamics simulations, while powerful, have a notable limitation. They often struggle with the issue of inefficient sampling due to the tendency of biological systems to get trapped in local minima of potential energy. This means that conventional MD simulations may not fully capture all possible dynamics of a studied biological system[84].

To overcome this challenge, the Gaussian-accelerated molecular dynamics (GaMD) approach was developed. GaMD is a computational enhanced sampling technique that works by adding a harmonic boost potential that follows a Gaussian distribution to the system's potential energy. The boost potential is determined based on a harmonic force constant and the difference between the reference energy and the system's potential energy. This addition is performed when the system's potential energy is lower than a reference energy. The use of a Gaussian distribution allows for a smooth and continuous modification of the potential energy surface. The modified potential of the system is then calculated as the sum of the system's potential and the harmonic

boost potential. This process effectively smoothens the potential energy surface and reduces system energy barriers, enabling the system to escape from local energy minima more easily. This approach allows GaMD to induce more conformational changes in the simulated systems, thereby enhancing the conformational sampling[84].

GaMD simulation proceeds in three stages: short conventional MD, GaMD equilibration, and GaMD production. During the first stage of short conventional MD, system potential statistics (including the minimum, maximum, average, and standard deviation) are collected to calculate the GaMD acceleration parameters. In the second stage of GaMD equilibration, the system potential statistics are updated to recalculate the GaMD acceleration parameters on the fly. In the third stage of GaMD production, the boost potential is applied to the system with GaMD acceleration parameters fixed. Simulation frames and the corresponding boost potential values are saved for analysis[84].

After the GaMD simulation, energy reweighting is done to analyze the boost potential distribution and calculate free energy profiles. Because the boost potential follows a Gaussian distribution, the original free energy profiles of studied biomolecules can be recovered through a process known as "Gaussian approximation" or cumulant expansion to the second order. By reweighting the obtained potential energy profiles, it is possible to understand how energetically favorable or unfavorable certain regions of conformational space are[84].

The application of GaMD is particularly beneficial in the study of highly dynamic biological systems. It allows for the exploration of a wider range of protein conformational dynamics, including the identification of rare conformational change episodes[84].

In this thesis, chapter 7 discusses the application of GaMD in modeling the dynamics of the tubulin protein.

1.6. Review of published works on computer-aided drug design techniques for discovering new modulators of tubulin polymerization

It's essential to appreciate the role that CADD techniques have played in the search for small molecules modulators of tubulin polymerization. CADD approaches have been instrumental in directing the discovery and development of novel tubulin modulators. In the following section, we review published works where these computational tools have been used to explore the vast chemical space in search of small molecule modulators of tubulin polymerization. We aim to highlight the inherent versatility and efficiency of CADD methodologies, showcasing how they have paved pathways to potential therapeutic agents and enriched our understanding of tubulin's complex biochemistry.

biomolecules

MDPI

Review

Computational Approaches to the Rational Design of Tubulin-Targeting Agents

Helena Pérez-Peña [1,2,†], Anne-Catherine Abel [1,3,†], Maxim Shevelev [2,4,†], Andrea E. Prota [3], Stefano Pieraccini [1] and Dragos Horvath [2,*]

[1] Department of Chemistry, Università degli Studi di Milano, Via Golgi 19, 20133 Milan, Italy
[2] Laboratory of Chemoinformatics, Faculty of Chemistry, University of Strasbourg, 4, Rue Blaise Pascal, 67081 Strasbourg, France
[3] Laboratory of Biomolecular Research, Paul Scherrer Institute, Forschungsstrasse 111, 5232 Villigen, Switzerland
[4] Department of Biochemistry and Molecular Biology, Universitat de Barcelona, Gran Via de les Corts Catalanes, 585, 08007 Barcelona, Spain
* Correspondence: dhorvath@unistra.fr
† These authors contributed equally to this work.

Abstract: Microtubules are highly dynamic polymers of α,β-tubulin dimers which play an essential role in numerous cellular processes such as cell proliferation and intracellular transport, making them an attractive target for cancer and neurodegeneration research. To date, a large number of known tubulin binders were derived from natural products, while only one was developed by rational structure-based drug design. Several of these tubulin binders show promising in vitro profiles while presenting unacceptable off-target effects when tested in patients. Therefore, there is a continuing demand for the discovery of safer and more efficient tubulin-targeting agents. Since tubulin structural data is readily available, the employment of computer-aided design techniques can be a key element to focus on the relevant chemical space and guide the design process. Due to the high diversity and quantity of structural data available, we compiled here a guide to the accessible tubulin-ligand structures. Furthermore, we review different ligand and structure-based methods recently used for the successful selection and design of new tubulin-targeting agents.

Keywords: computer-aided drug design; microtubules; microtubule targeting agents; virtual screening; molecular docking; molecular dynamics simulations; pharmacophore screening; QSAR

Citation: Pérez-Peña, H.; Abel, A.-C.; Shevelev, M.; Prota, A.E.; Pieraccini, S.; Horvath, D. Computational Approaches to the Rational Design of Tubulin-Targeting Agents. *Biomolecules* **2023**, *13*, 285. https://doi.org/10.3390/biom13020285

Academic Editor: Scott Thomas Forth

Received: 15 December 2022
Revised: 27 January 2023
Accepted: 31 January 2023
Published: 2 February 2023

1. Introduction

Microtubules (MTs) are an essential part of the eukaryotic cytoskeleton and are implicated in various diseases. They are highly dynamic polymers composed of α,β-tubulin dimers in which each monomer is able to bind GTP. GTP hydrolysis is limited to the β-monomer (E-site), providing energy for conformational changes required for MT formation. Within the α-monomer GTP is always retained (N-site). Together, these proteins form hollow, cylindrical structures, in cells mostly containing 13 protofilaments. Within the cell, they are involved in numerous cellular processes such as cell signaling, morphology, motility, growth, and long-distance trafficking regulation [1].

Naturally, any perturbation of the MT network severely affects cell survival, thus making MTs attractive targets for cancer therapy. Presently, several MT targeting agents (MTAs) such as vinca alkaloids and taxanes are used to treat different types of cancer. By altering the MT homeostasis, they promote apoptosis of cancer cells via several independent mechanisms [2]. Moreover, there is an increasing interest in MTs as a target for the treatment of diabetes [3]. Furthermore, abnormal dynamics of MTs in neuronal cells is implicated to play an important role in several neurodegenerative diseases (reviewed in [4]).

Almost 40 years after the first mechanism was proposed [5], the details of MT formation still remain an ongoing topic of discussion; the main steps as understood today are outlined below: Nucleation of MTs occurs in cells at MT organizing centers (MTOCs) such as the γ-TuRC complex (reviewed in [6–8]). Based on this template structure, MTs grow by addition of a dimer carrying GTP in both nucleotide binding sites in a head–to-tail fashion, always adding α-tubulin onto exposed β-tubulin. Thus, the MT is formed as a polar structure and exposes β-tubulin at the growing end (MT plus end). Incorporation of tubulin dimers into the MT lattice is accompanied by a conformational change of the dimer from a curved towards a more rigid, straight structure (curved-to-straight transition), which is then followed by GTP hydrolysis in the β-monomer [9]. Only at the plus end of the MT a so-called "GTP-cap" consisting of dimers that contain GTP in both sites is sustained, which is thought to stabilize the end against depolymerization [10].

Within cells, the MT cytoskeleton is maintained in what is termed the "dynamic equilibrium", alternating between phases of growth and shrinkage of individual MTs, which allows them to perform their various physiological activities (Figure 1). MT associated proteins, post-translational modifications, as well as small molecules MT targeting agents (MTAs), modulate the dynamics of the MT network. MTAs at high concentrations exert different mechanisms of actions, which are used to categorize them into two classes: MT stabilizing agents (MSAs) that lead to an increased stability of the present MT by promoting assembly or stabilization of the lattice structure, and MT destabilizing agents (MDAs) which prevent the assembly of dimers into MTs.

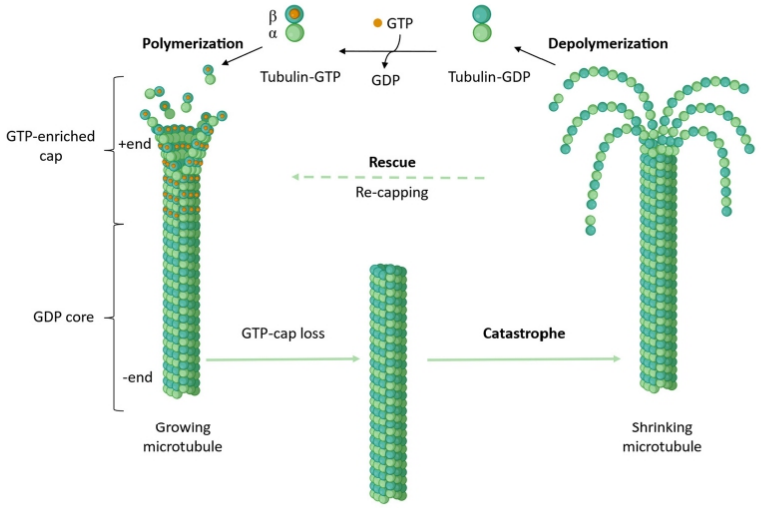


Figure 1. Microtubule dynamic equilibrium. MTs are constantly alternating between growth and shrinkage phases, while the −end of the MT is displaying some dynamics the overall stability is governed by quicker processes at the MT +end. Growth of an MT is facilitated by incorporation of two GTP containing tubulin dimers onto the +tip, followed by lattice incorporation, which leads to subsequent GTP hydrolysis. On the top of the growing MT a "GTP-cap" consisting of GTP-dimers stabilizes the structure. Exchange of this capping dimers against GDP tubulin leads to depolymerization. Adapted from "Microtubule (polymerizing and depolymerizing)" by BioRender.com (accessed on 15 December 2022).

MTAs have been widely studied and characterized due to their long-standing use as anti-cancer drugs. Routinely, MTAs are probed on their cytotoxicity and their ability to influence MT polymerization. Further, to understand their mode of action a lot of effort has been dedicated to solving high-resolution MT and ligand–tubulin complex

structures. Up to 2021, seven distinct binding sites for small molecules had been thoroughly characterized by X-ray crystallography. In 2021, a combination of crystallographic fragment-based screening and molecular dynamics (MD) simulations evidenced 10 binding sites occupied by 56 chemically diverse fragments, of which six sites were completely novel [11]. A selection of these fragments was subsequently used in a straight-forward fashion to develop a lead-like molecule from non-cytotoxic building blocks. It was named todalam and occupies the 8th binding site on tubulin located at the inter-dimer interface [12]. Together, the large amount of biochemical data and ever-growing amount of structural data available lay a solid foundation for the computer-aided development of novel tubulin-targeting agents.

Computer-aided molecular design methods, such as ligand-based and structure-based approaches, open new possibilities to further exploit current knowledge on MTs, tubulin and MTAs. These two in silico strategies have been considered essential for accelerating the research of MTAs assisting in the identification, design, and selection of new compounds. Both are used to discover molecules with desired biological activity, but differ in terms of the initial information exploited to generate their predictions. Ligand-based methods "learn" from previously discovered ligands of a target, and their measured affinities. They are agnostic in terms of ligand-target interaction mechanisms, but rely on interpolation and extrapolation of predicted affinity of a new candidate based on the nearest known examples of ligands. On the contrary, structure-based approaches base their predictions on explicit modeling of presumed interactions between ligands and given biological targets.

The aim of this review is to summarize recent applications of state-of-the-art methods of both computational ligand and structure-based approaches to successful design of new MTAs. Note, however, that using in silico methodology to "discover" putatively active compounds makes no sense unless those compounds are actually synthesized and tested. Publishing in silico predictions without further validation should, in our opinion, be strongly discouraged, because the likelihood of experimentalist readers embarking on the difficult task of synthesis and testing of someone else's predictions is very low (actually null, as far as we can tell). Therefore, this work will only cite computer-aided design work which is either (a) methodologically innovative, (b) reporting tool benchmarking studies or (c) backed up by experimental validation.

2. Ligand-Based Approaches

Ligand-based strategies may be employed if rich and balanced structure-activity information (at least ~100 known tested small molecules, including binders and non-binders to the target) is available. They are of course the only option if no structure of the target protein has been solved, but are irrespectively useful in the early stages of a virtual screening (VS) campaign, as they are typically much faster than structure-based algorithms. These methods algorithmically analyze molecules encoded by molecular descriptors or ensembles of calculated conformations and extract chemical knowledge to predict a given compound's property. Such screening usually highlights structural patterns deemed important for exhibiting a desired property.

Historically, these methods were the first to be applied to the problem of discovering novel modulators of tubulin polymerization. This was mostly due to the low quality of tubulin-related structural data at that time (reviewed in [13]). However, despite considerable progress in tubulin crystallography and prevalence of structure-based methods in modern tubulin research, ligand-based approaches are still useful and yield promising results. This section highlights recent examples of successful application of such computational methods in tubulin-related drug design.

2.1. Similarity Search

A similarity search is used to filter a set of molecules, in search of those that display similar features to a query molecule. This method assumes that similar molecules exhibit—statistically speaking—similar properties [14]. There is no absolute best way to

encode molecular similarity, typically rendered by the metric (distance) of the two points representing molecules in "descriptor space". Fragment-based fingerprints (monitoring the presence of specific substructures in each molecule) are common molecular descriptors for this task; however, other features such as descriptors of molecular shape, topological pharmacophores can be used. Any function that measures distance between two points in a metric space is applicable to characterize "molecular dissimilarity". The best combination of descriptors and metric function is the one that guarantees the best "Neighborhood Behavior Compliance", e.g., by minimizing the occurrence of "property cliffs"—pairs of compounds perceived as highly similar in spite of using widely different property values [15].

A similarity search is often used as a first step in VS. For example, Aoyub et al. [16] and Guo et al. [17] performed 2D similarity searches in large compound databases as initial phases of drug design cycles that resulted in development of novel MTAs binding to the taxane and colchicine site, respectively. Several novel colchicine-site targeting agents were also discovered by Mangiatordi et al., who based their design on a 3D shape similarity screening [18]. Another two colchicine-site targeting hits were found by Federico et al., who used not only 3D shape, but also electrostatic potential similarity in their VS campaign [19].

Coupling known active compound structures with information on their targets can make the similarity search useful for establishing targets of novel compounds. This was demonstrated by Lo et al., who developed chemical similarity networks based on two and three-dimensional compound similarity (CSNAP2D and CSNAP3D, respectively). By calculating similarities of molecules with cytotoxic action of unknown mechanism to molecules within the network, the authors correctly predicted tubulin as a target for 37 novel compounds targeting the colchicine and taxane binding sites [20,21].

In Table A1 (Appendix A) we have summarized the implementations of the technique used in mentioned references.

2.2. QSAR Modeling

Quantitative structure-activity relationship (QSAR) modeling finds a mathematical function that relates chemical structure to values of some desired property, e.g. biological activity. The process of fitting such a function is called model training. Typically, two- or three-dimensional molecular structures are digitally encoded by various descriptors, which are then input to machine learning algorithms along with corresponding target property values, available from biological assays. These values can be continuous (pIC_{50} values, binding affinity) or discrete (active/inactive classification), corresponding to either regression or classification problems. Afterwards, a trained model can be used to predict target values for new molecules, not included in the training set. The predictive power of a QSAR model depends on careful curation of input data, rigorous validation, and adequate assessment of its applicability domain. State-of-the-art approaches in these topics are described in more detail in [22–24].

This method is particularly useful for rational drug design as it provides insight into which molecular features correlate the most with changes of desired property values. For example, Gaikwad et al. used two-dimensional QSAR modeling to establish structural patterns that significantly correlate with cytotoxicity of colchicine site-targeting phenylindoles against cancer cells [25]. High utility of QSAR modeling in VS was demonstrated in works by Guo et al. [26] and Stefanski et al. [27], who used consensus QSAR modeling in VS campaigns that yielded a total of three novel colchicine site targeting tubulin polymerization inhibitors.

3D QSAR was shown to be a convenient way to rationalize ligand optimization in works by Quan et al. [28] and Pandit et al. [29]. Both works used CoMFA and CoMSIA methods to rationalize structure-activity data for limited datasets of similar scaffold-based compounds, suggesting possible structure optimization patterns, which, in case of the latter work, yielded a new class of cytotoxic in vitro tubulysin derivatives targeting the vinca binding site. A summary of the experimental conditions for the above-mentioned QSAR works is provided in Table A2.

It is worth noting that the use of machine learning in this field has been limited due to the scarcity of publicly available data. The lack of large, diverse tubulin-related structure-activity datasets makes it difficult to train adequate machine learning models that can be used in a large-scale virtual screening context. For example, querying the ChEMBL database (v.26) for "Tubulin" returns more than 8000 raw structure-activity records, but these are a heterogeneous collection of results from widely different assays at diverse experimental setups, using the MTs or tubulin of widely different species (from *Arabidopsis* to *Homo Sapiens*). Or, machine learning requires homogeneous, comparable experimental activity entries to serve for calibration of empirical functions trying to approximate them upon input of a molecular structure. Thus, only entries sourcing from a same experimental setup (listed under a same ChEMBL Assay ID) can be safely compared. Deceivingly, there is only one such assay (CHEMBL817769; Inhibition of tubulin polymerization interacting at the colchicine binding site of *Sus Scrofa*) featuring more than 100 entries (103, precisely)—a rule-of-thumb minimal threshold of training set size to start envisaging machine learning. Size is necessary, but far from sufficient—a balanced presence of active and inactive compounds is of paramount importance, whereas the chemical diversity of the compounds sets the limit for the applicability domain of the model. Machine learning is likely to play a more prominent role in this regard if more relevant data becomes publicly available.

2.3. Pharmacophore Screening

A pharmacophore is an abstract description of the set of local steric or electronic properties (hydrophobicity, H-bond acceptor/donor features, charged groups) that a molecule should contain in order to interact with a particular biological target at a specific site. A set of such properties, with defined positions in space relative to each other is called a pharmacophore model. For a given ligand, it is mostly related to fragments of chemical structure and is binding site-specific. It is assumed that molecules that follow the same pharmacophore pattern may have similar biological activity (even though they may differ in other, less relevant structural aspects). This makes pharmacophore-based VS useful for searching and designing new drugs, escaping the rather narrow domain accessible by strict similarity-driven searching.

In particular, experimental structure-activity data can be used to automatically construct ligand-based pharmacophore models. A detailed explanation of pharmacophore model generation steps is given by Giordano et al. [30]. In short, models are obtained by computing and aligning 3D conformations of selected molecules, with pharmacophore features assigned to overlapping structural fragments. Several models may be built for different alignments. A fitness function estimates how well the molecules fit into a given model, leading to selection of the best model.

Screening with such models can be used to filter compounds in a large library, leaving only those that match the required model in at least one of several conformations. Models always need to be validated before use in VS. A model is considered valid if it can discriminate known active molecules from decoys—structurally similar compounds not showing the desired activity [31].

Ligand-based pharmacophore screening is often used in combinations with other computational methods to lower the number of candidates that need to be tested by subsequent approaches. For example, Zhang et al. used a pharmacophore model based on taxane-site ligands to reduce the number of compounds processed by structure-based pharmacophore model and protein-ligand docking, eventually leading to a discovery of two novel tubulin-targeting cytotoxic agents targeting this site [32]. In a similar manner, a ligand-based pharmacophore model developed by Lone et al. was shown to be useful for vinca-site targeting agents design [33]. Moreover, Niu et al. successfully applied a ligand-based pharmacophore model to discover two novel colchicine-site targeting modulators of tubulin polymerization [34]. Stefanski et al. used a ligand-based pharmacophore model in a VS campaign that discovered two potent in vitro cytotoxic colchicine-site targeting agents [27].

As can be seen, despite ligand-based pharmacophore screening not being featured in many recent tubulin-related computational studies (structure-based pharmacophores or docking being preferable, as soon as experimental protein structures are available), it is still a viable method that is used to design and screen for novel modulators of tubulin polymerization. Table A3 provides an overview of recent works that used this approach.

3. Structure-Based Approaches

Contrarily to ligand-based methods, structure-based approaches exploit the 3D structure of a macromolecular biological target to estimate a given molecule's affinity to a targeted binding site. The main sources of information for these methods are either experimental data generated by X-ray crystallography, NMR spectroscopy, cryo-electron microscopy or computationally predicted data. Analyzing bound ligand poses helps to determine the key residues defining the binding site, as well as pinpoint to the key fragments of molecular structure that contribute to interaction with the target protein. Success in high-resolution determination of biological macromolecule structures drove the usage of these structure-based techniques in modern drug discovery pipelines, and tubulin-related research is no exception. In this section, we review recent examples of structure-based methods application in search and design for novel modulators of tubulin polymerization [35,36].

3.1. Structural Data on Tubulin

3.1.1. Tools to Study Tubulin 3D Structures

Possibly, the most important decision in carrying out a structure-based drug design project on tubulin is the selection of the correct tubulin model. While the sheer abundance of accessible information is a huge benefit for any of such projects, the numbers and diversity of available structures can be overwhelming. In order to select the best possible model for one's purpose, it is important to consider the method and system in which the structure was obtained. Therefore, we will give a brief overview of the available structures and setups that were used to determine them, as well as highlight a few key points to consider when selecting the structure.

By comparing the different structures obtained of tubulin and MTs, it was observed that tubulin dimers are able to adopt two prominent tubulin conformations that are related to its assembly state: a "straight" conformation is present in assembled MTs and a "curved" conformation is observed in soluble tubulin. The conformational transition from curved-to-straight is needed to establish lateral tubulin contacts between protofilaments in MTs. This curved-to-straight transition requires rearrangements of the tubulin monomers, in which the intermediate domain of the tubulin monomer moves with respect to a larger ensemble comprising both the N- and C-terminal domains. Due to this repositioning within the straight MT lattice, the α monomers are almost perfectly aligned with the β monomers, thus it is possible to superpose α onto β simply by translation (Figure 2A). Whereas, within the soluble dimer there is an intrinsic curvature of one monomer against the other, thus translation alone is not sufficient to superpose one monomer onto another (Figure 2B). The degree of this curvature varies; it can range from 9–18 degrees depending on the binding partners present [37].

This conformational state is one of the main differences observed between all available crystal structures and the CryoEM data on MTs: All crystal structures depict the soluble and "curved" conformation of tubulin and all MT structures show the "straight" conformation. Thus, it is important to consider on which "state" of the tubulin structure is used, as basis for the computational work. Despite these major differences, the crystal structures are remarkably well suited for the design and optimization of drugs. Up to now, five different systems have been described for the crystallization of tubulin. All rely on proteins stabilizing the tubulin in its dimeric or tetrameric form, as the uncoordinated, soluble tubulin is polymerizing rather than forming nicely diffracting crystals. This is highlighted

by the fact that the first high-resolution crystal structure has only been reported after the tubulin–stathmin interaction had been discovered and exploited [38,39].

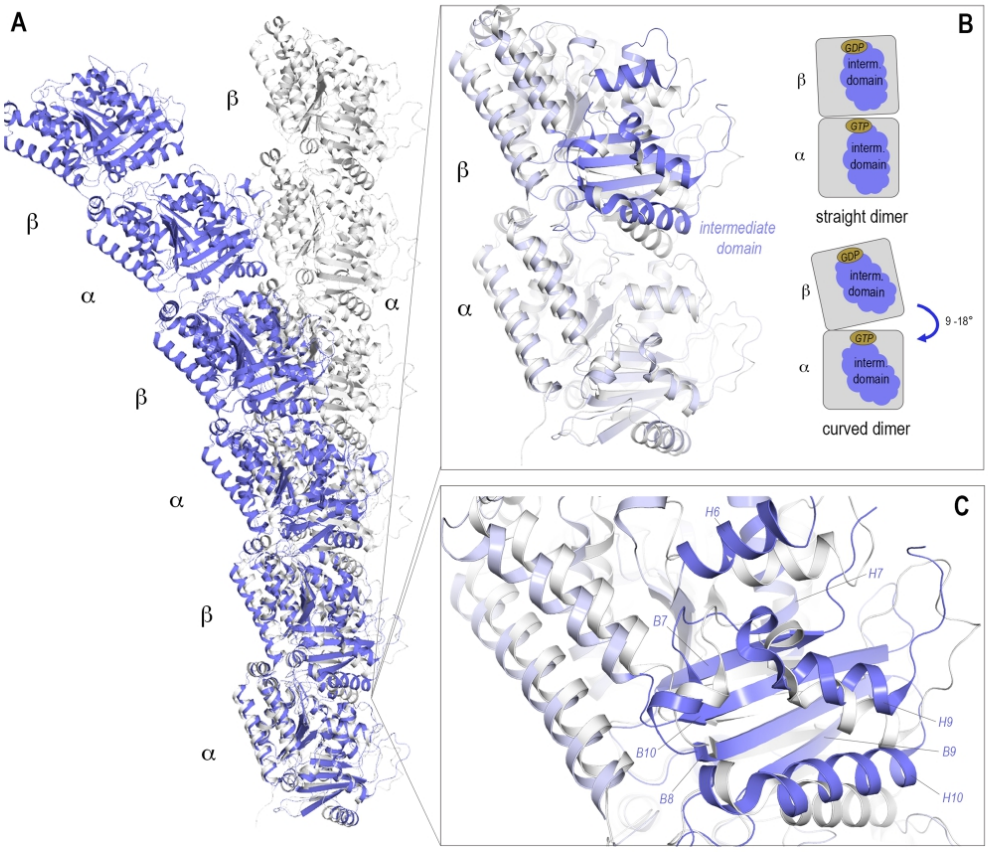


Figure 2. The "curved" and "straight" tubulin conformations. (**A**) A straight protofilament, as present in the MT lattice, is shown in ribbon presentation in light gray (PDB ID 7SJ7). A protofilament constituted of tubulin in a curved conformation is shown in blue (from PDB ID 5LXT). (**B**) The intrinsic curvature and structural differences on a single dimer are shown: A heterodimer in the straight conformation is depicted in light gray and the curved conformation in light blue. The main differences in the structures are within the intermediate domain (residues 206–384), highlighted in darker blue, which upon curved-to-straight transition moves relative to the other domains. This is also indicated in the schematic drawing of both straight and curved dimers. The angle corresponds to the relative curvature of one monomer to the other. (**C**) The structural elements of the intermediate domain are shown in more detail, the changes necessary for "straightening" are mainly translation of the shown H7 as well as rotation of the neighboring structural elements H6-10 and B7-10.

The very first structural information on tubulin was obtained in 1998 by Nogales et al. using electron crystallography on taxol stabilized zinc-induced protofilaments. This allowed the determination of a first model of the structure of tubulins, the assignment of domains and identified the taxol binding site on β-tubulin [40]. However, the arrangement of the protofilaments in this crystal system is antiparallel and does not reflect the

protofilament-assembly found in MTs. Accordingly, this system was not further used for X-ray crystallographic studies.

Soon afterwards, the tubulin stathmin-like domain SLD (T_2R) system was the beginning of tubulin complex crystallization with the first crystal structure in 2000 [41], followed by the first tubulin-small molecule complex in 2004 [42], which revealed the position of the colchicine site. Later, it was noted that cleavage of the C-terminal tubulin tails increases the resolution of the T_2R system significantly. Furthermore, this system evolved to be the most commonly used T_2R-tubulin tyrosine ligase setup (T_2R-TTL, Figure 3A) [43,44], which was used to solve most tubulin-small molecule structures. In both complexes, two tubulin dimers are coordinated by a stathmin-like protein RB3 that prevents tubulin polymerization by its N-terminal β-hairpin cap bound to α1 tubulin. In the T_2R-TTL system, the TTL protein is bound at the same end of the tetramer on α1 tubulin. The overall tubulin structure does not differ significantly between the two setups.

Since the SLDs and TTL used in these crystallization systems may prevent binding of proteins to tubulin, alternatives have been developed. The tubulin Designed Ankryin Repeat Protein DARPin crystallization system (Figure 3B) [45] is the second most frequently used one. This system allows to achieve even higher resolution compared to the T_2R-TTL one, with the best resolved structure ranking at 1.5 Å resolution (PDB ID 6S8K, [46]). In this system, only one tubulin dimer is coordinated by the selected DARPin, resulting in a much more densely packed and smaller unit cell.

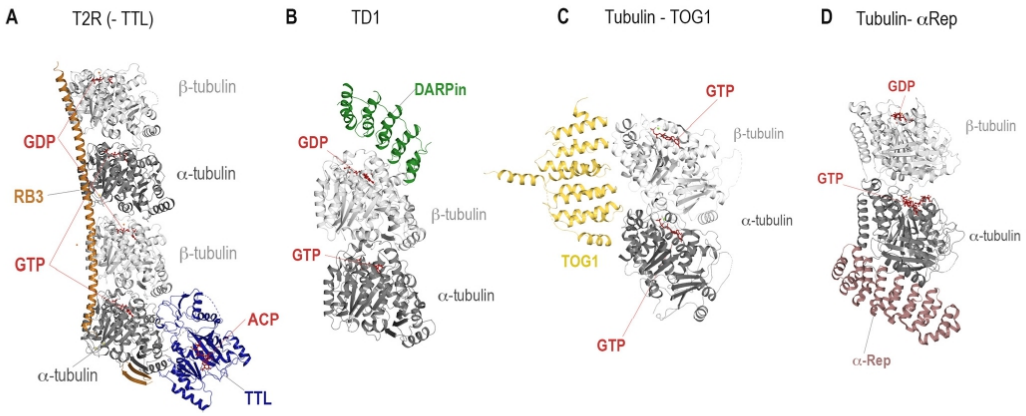


Figure 3. The crystallization systems (**A**) T_2R(-TTL) (PDIB ID: 4I55, [44]), (**B**) TD1 (PDB ID 4DRX, [45], (**C**) Tubulin-TOG1 (PDB ID: 4FFB, [47]) and (**D**) Tubulin-αRep (PDB ID: 6GWC, [37]) are depicted. The proteins are shown in ribbon representation, α- and β-tubulin are colored dark and light grey, respectively. The SLD/RB3 protein is colored orange, the TTL in blue, DARPin in green, TOG1 in yellow and alpha-Rep in brownish color. Nucleotides are shown in sticks representation and colored red. The structure of the SLD tubulin complex, T_2R crystallization system corresponds to the T_2R-TTL structure without the bound TTL and thus was not shown separately.

Up to now, the described systems T_2R, T_2R-TTL and TD1 are the only ones that have been used to elucidate the structures of tubulin-small molecule complexes. Nevertheless, the following two crystallization systems for the study of protein-protein interactions have been included to provide a complete overview of tubulin crystallization systems.

In order to investigate the interaction of the cellular MT growth factor, Stu2p, Ayaz et al. co-crystallized its tumor overexpressed gene domain TOG1 with tubulin [47]. Surprisingly, it was found that TOG1 was establishing interactions with both α- and β-tubulin and preferentially bound to the curved state of soluble tubulin dimers (Figure 3C).

More recently, a fifth crystallization system, targeting MT binding proteins, has been introduced. Therein, one artificially designed α-Rep protein is used to prevent tubulin polymerization and to enable crystallization of the complex (Figure 3D). α-Rep was specifically designed to bind to tubulin sites involved in longitudinal protofilament interactions in order to expose the surface of tubulin, which would be on the exterior site of the MT [37]. So far, the system has been used to elucidate the structural details of centrosomal P4.1-associated protein CPAP [48], allowing a more throughout investigation compared to the previously published CPAP –tubulin DARPin structures [49,50].

3.1.2. Binding Sites on Tubulin

As mentioned in the introduction, extensive work has been done on determining the binding mode of tubulin-targeting agents. Here, we would like to give a brief overview of the eight established binding sites (Figure 4) and their mode of action on modulating MT dynamics (in more detail reviewed in [51]). The most prominent member of MTAs is paclitaxel, sold as a blockbuster drug under the name Taxol®, which is an MSA that binds to an exposed pocket on β-tubulin. **Taxane-site** ligands are able to enhance MT stability, either by promoting the curved-to-straight transition, e.g., paclitaxel [52,53] or by direct structural stabilization of the βS7-βH9 loop (M-loop), a key structural element forming inter-dimer contacts in MTs [54], e.g., epothilone A or zampanolide [44]. **Laulimalide-/Peloruside-site** agents strengthen the interactions of tubulin dimers across neighboring protofilaments in MTs by binding to a pocket near the lateral protofilament interface. Moreover, these agents have been described to allosterically stabilize the M loop to some extent [55,56].

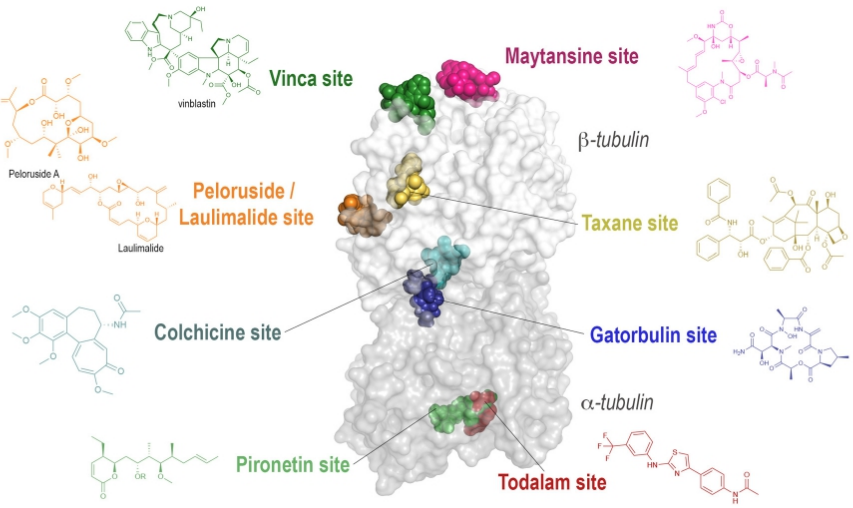


Figure 4. The eight distinct binding sites are highlighted on one tubulin dimer with all their representative ligands in colored sphere representation. The protein is shown in a transparent surface representation α- and β-tubulin chains are colored dark and light grey, respectively. The chemical structures of the ligands after which the binding sites were named are indicated next to the labels and colored following the color code of their sphere model.

In the group of MDAs, **colchicine-site** ligands are present with a great variety and a high number of different scaffolds. They bind in a buried pocket at the intra-dimer interface of α and β tubulin, flipping the βT7 loop out of its native position. By occupying this binding site, they effectively prevent the curved-to-straight transition by blocking the compaction of the pocket formed by the strands βS8 and βS9, and by the helices βH8 and αH7 [42,57].

Another well-known group of MDAs are the vinca alkaloids, which bind at the longitudinal interface between tubulin dimers. **Vinca-site** ligands induce a 'wedge' [58] at the tip of the MT and thus prevent the straightening of the dimers. Additionally, they promote the assembly of small helical tubulin polymers, thereby effectively reducing the amount of assembly-competent tubulin. It has also been noted that vinca-site ligands interfere with the hydrolysis of GTP by blocking the proper alignment of the catalytic residues, thereby further hindering the polymerization process [59,60].

The group of **maytansine-site** ligands blocks the assembly of MTs by inhibiting the addition of new tubulin dimers to the growing end. This is achieved by binding to the exposed site of β-tubulin and then effectively blocking the site that should accommodate the αH8 and αT7 loop of the binding tubulin dimer [61]. Ligands bound at this site not only block further growth of MTs, but are also capable of fully blocking the formation of smaller tubulin oligomers, at high concentration, effectively keeping tubulin within the dimeric state.

So far, the only ligand known to exclusively bind to α-tubulin is **pironetin**, which binds to a buried pocket by covalent attachment to Cys316 [62,63]. Binding of pironetin perturbs the above-mentioned helix αH8 and the αT7 loop, thus similar to maytansine preventing the interaction of these elements with the neighboring tubulin and fixing tubulin in an assembly-incompetent state. Furthermore, pironetin also prevents the growth at the −end of the MT, which exposes the α-tubulin surface harboring both the helix αH8 and the αT7 loop and thus eventually promotes the disassembly of already formed MTs [62].

Recently, both the 7th and the 8th distinct binding sites on the tubulin dimer have been described. **Gatorbulin**, a cydodepsipeptide isolated from marine cyanobacteria, was found to bind to the intra-dimer interface adjacent to the well-known colchicine binding site [64]. **Todalam**, the first rationally designed tubulin binder, which emerged from a crystallographic fragment screen [11], binds at the inter-dimer interface at a site located between the maytansine site on β-tubulin and the end of the pironetin pocket on α-tubulin [12]. Both compounds are thought to hinder MT formation by a mechanism similar to that of the vinca-site ligands, by creating a wedge into the tubulin-oligomer structure. As observed for vinblastine, todalam as well was shown to promote the formation of ring-like tubulin oligomers, further decreasing the pool of tubulin available for polymerization.

The position of the binding sites has clear implications on the choice of the crystallization system: due to its size, the TD1 crystallization system is well suited for molecules bound internally within one dimer (e.g., colchicine, gatorbulin), however the binding sites at the inter-dimer interface such as for example the vinca-site can only be targeted by using the T_2R(-TTL) systems.

3.1.3. System Selection for Virtual Screening (VS) and MD Simulations

Not all out of the more than 300 crystal structures within the PDB database were equally often used in computational experiments, as we noticed in our analysis of the most recent MD simulation literature (overview in Table A6). Surprisingly, we found that even 20 years after the first description of the tubulin structure at near-atomic resolution [54], simulations of taxane-site ligands or apo tubulin are often based on some of the very first tubulin datasets obtained with electron diffraction in 1998 (PDB ID 1TUB, 3.7 Å, [40]) and 2001 (PDB ID 1JFF, 3.5 Å, [54]). There is a bit more of variety in the colchicine site structures that were selected for simulations, although only a fraction of the great number of available high-resolution tubulin colchicine site structures have been considered: PDB ID 1SA0 2004 3.6 Å [42], PDB ID 1Z2B 2005 4.1 Å [58], PDB ID 3E22 2008 3.8 Å [59], PDB ID 3HKC 2009 3.8 Å [57], PDB ID 4O2B 2014 2.3 Å [65], PDB ID 6Y6D 2020 2.2 Å [66]. For simulations of other ligands, since a lower number of structures is available, the choice of the starting model was obvious: vinca-site ligands PDB ID 3E22 2008 3.8 Å [59], PDB ID 4O4J 2014 2.2 Å [56], PDB ID 5JH7 2016 2.2 Å [67], and laulimalide site: PDB ID 4O4H 2014 2.1Å [56].

While this analysis reflects on only a fraction of the most recent literature, we see a trend that not always the most recent or high-resolution structures are selected. Due to the

importance of the selection of the starting model for virtual screening and MD simulations we provide in Table 1 an overview of the highest resolution structures available to support the selection process. Further, in Table 2 we have compiled a list of the CryoEM models for MT structures with highest resolution for tubulin-small molecule complexes, a field in which not many structures are available yet.

Table 1. List of high-resolution tubulin crystal structures by binding site.

Binding Site	PDB ID	Resolution (Å)	Crystallization System	Bound Ligand
Apo	5NQU [68]	1.8	TD1	-
	3RYC [69]	2.1	T_2R	-
	4I55 [44]	2.2	T_2R-TTL	-
Taxane site	4I4T [44]	1.8	T_2R-TTL	Zampanolide
	5LXT [70]	1.9	T_2R-TTL	Discodermolide
	6SES [71]	2.0	T_2R-TTL	B2
Laulimalide/Peloruside	4O4H [56]	2.1	T_2R-TTL	Laulimalide
	4O4J [56]	2.2	T_2R-TTL	Peroluside A
Maytansine	4TV9 [61]	2.0	T_2R-TTL	PM060184
	6FJM [72]	2.1	T_2R-TTL	Disorazole Z
	4TV8 [61]	2.1	T_2R-TTL	Maytansine
Colchicine	6S8K [46]	1.5	TD1	Plinabulin
	6ZWB [73]	1.7	TD1	Z-SBTub3 photoswitch
	7Z2P [74]	2.0	T_2R-TTL	Nocodazole
	5M7E [75]	2.0	T_2R-TTL	BKM120
	6TH4 [76]	2.1	T_2R	*exo*-methylene-nor-colchicine
Vinca	5IYZ [77]	1.8	T_2R-TTL	Monomethylauristatin E
	5J2T [77]	2.2	T_2R-TTL	Vinblastine
	5JH7 [67]	2.3	T_2R-TTL	Eribulin
Pironetin	5LA6 [62]	2.1	T_2R-TTL	Pironetin
	5FNV [63]	2.6	T_2R-TTL	Pironetin
Todalam	5SB3 [12]	2.2	T_2R-TTL	Todalam precursor 4
	5SB6 [12]	2.3	T_2R-TTL	Todalam derivative 10
Gatorbulin	7ALR [64]	1.9	TD1	Gatorbulin

Table 2. High-resolution CryoEM MT structures.

MT Structure	PDBID	Resolution (Å)
Taxol-stablized MTs	6WVR [78]	2.9
Peloruside stabilized MTs	5SYC [55]	3.5
Taxol/Peloruside MTs	5SYE [55]	3.5
Taxol MTs	5SYF [55]	3.5
Zampanolide MTs	5SYG [55]	3.5
Undecorated MTs recombinant tubulin	7SJ7 [79]	3.8

When choosing the VS system, one should also consider the target of the desired molecule. If one is aiming for an MT-binder, one might compare the binding pocket found in crystallization systems with the CryoEM MT structures to evaluate the differences and the impact of MT formation on the specific binding site. However, one needs to be careful because most of the structures have been obtained by stabilizing the MT with small molecules, most often paclitaxel, or using non-hydrolyzable nucleotides. Therefore, these structures could also be different from the MT structure in the absence of stabilizers or artificial nucleotides.

The next consideration on the selection of the system for MD simulation is the assembly of tubulin into protofilaments and MT structures. If the binding site studied is far from any

tubulin inter-dimer interface (e.g., colchicine site, gatorbulin) or is considered to completely prevent the interaction of two dimers (e.g., maytansine site, pironetin site), a dimer can serve as a model for tubulin binders. It can be extracted from either T_2R, T_2R-TTL or TD1 structures, however the presence of the stabilizing proteins could artificially modify the tubulin structure in the proximity of their binding site. Ideally, the site of VS should be far from crystal contacts established in the system and the binding sites of the stabilizing proteins DARPin, RB3 and TTL.

If the binding site is present at the longitudinal inter-dimer interface (e.g., vinca, todalam, gatorbulin) or the lateral axes (e.g., taxanes), a more complex system may need to be considered. To extract two dimers in the curved conformation either T_2R or T_2R-TTL structures can be used to generate longitudinally linked tetramers. In the case of both longitudinal and lateral axes as present only within the context of an MT, a CryoEM structure should be used as a basis. For example, scientists such as Castro-Álvarez et al. [80] opted to study a 'tetramer' model to investigate binders at the taxane site, since the M loop stabilized by some taxane-site ligands is establishing lateral interactions with the neighboring tubulin dimer. The choice of the system size is a trade-off between the accuracy of the site and the computational effort needed.

3.2. Tubulin-Related VS Strategies

3.2.1. Pharmacophore Screening

We already discussed ligand-based pharmacophore modeling and its application in VS, where models are generated from structures of active molecules relying on conformational space sampling and ligand alignment. In structure-based pharmacophore modeling, a ligand's bioactive conformation in the binding site along with knowledge of the receptor structure guides the pharmacophore features placement and often provides higher quality models than those deduced by the ligand-based approach [31].

It is common to start such modeling by choosing one or several protein structures with bound ligands. Then, possible interactions are estimated between ligand and binding site atoms. After that, pharmacophore features are automatically assigned to regions of binding site space based on estimated H-bond formation, charge, and hydrophobic contact. Such models can be combined by merging over common features or refined manually [81]. The same validation strategy is applied before usage in VS, as described for ligand-based models.

Structure-based pharmacophore screening has shown significant value in tubulin-related research. It has been mostly used as one of the steps in multi-step VS campaigns that yielded novel colchicine and taxane-site targeting modulators of tubulin polymerization. Interestingly, recent successful works used different approaches to model building and selection. As such, Nagarajan et al. [82] built six colchicine-site interaction models based on relevant crystal structures and merged them by common features to obtain a model later used in a VS. Mangiatordi et al. [18] built seven colchicine-site models based on manually selected relevant PDB structures, validated them with a set of actives and decoys, and used the model with the best discriminative performance for VS. On the contrary, Zhou et al. [83] built four pharmacophore models based on relevant well-resolved PDB structures containing colchicine-site ligands and refined them manually, putting emphasis on interactions with experimentally known key residues. Similarly, Zhang et al. [32] derived seven pharmacophore models of the taxane site interactions from a single PDB crystal structure and refined all of them to highlight only the most important features. However, Gallego-Yerga et al. [84] noted that defining a single pharmacophore model puts unnecessary constraints on the model. Instead, they used an ensemble of 118 pharmacophore models derived from all resolved structures of tubulin with different bound colchicine-site targeting ligands in an attempt to capture flexibility of the site and variating nature of ligands. By contrast, Elseginy et al. [85] was able to produce good results by using a single model automatically extracted from a relevant colchicine site structure without any additional refinement.

Table A4 provides an overview of pharmacophore screening implementations from each mentioned VS campaigns.

3.2.2. Protein-Ligand Docking

One of the most frequently used structure-based drug design methods is protein-ligand docking. It is used to estimate with a considerable degree of accuracy the most likely conformation of a ligand within a given binding site, and therefrom extrapolate—with, unfortunately, not very good accuracy—its binding affinity.

By computationally predicting the binding affinity of tubulin-targeting agents, researchers identify compounds that have a high binding affinity for tubulin and are therefore more likely to be effective binders. This information can be used to prioritize compounds for further experimental validation, such as performing in vitro or in vivo assays to confirm their binding activity and efficacy. It's worth noting that computational predictions of binding affinity are not always accurate, and experimental validation is needed to confirm the predictions. However, computational predictions can be very useful for rapidly and efficiently identifying potential binders and prioritizing them for further experimental validation. Then, the success rate can vary depending on several factors, such as the quality of the computational method, the quality of the input data, and the complexity of the system being studied.

Protein-ligand docking tools operate on 3D structures of proteins and ligands. Typical docking computations involve sampling of a ligand's conformational space, and ranking the computed poses by estimating the (free) energy of interaction between the ligand in a given pose and the binding site using specific scoring functions. These computations may consider the binding pocket's residues to be rigid or flexible. Rigid docking is computationally faster, but unable to account for ligand-specific adjustments of the protein site geometry.

Algorithms for conformation sampling modify torsional, translational, and rotational degrees of freedom of a given ligand in a site in either a systematic sequential or a stochastic randomized fashion. Detailed reviews of sampling methods were compiled previously for example by Sulimov et al. [86] or Halperin et al. [87].

Sampling algorithms visit many putative poses of a ligand within the site and the docking software ranks all of them according to a scoring function. These functions aim to estimate a ligand's affinity toward the binding site in each specific sampled pose, taking into account intermolecular interactions and other physicochemical effects. The calculations are based on either force fields, modeled contribution of empirically defined physicochemical parameters, or knowledge of different atom-type interactions statistically extracted from resolved co-crystallized protein-ligand structures.

Before use, protein structures are pre-processed by adding missing hydrogens, computing charges, removing solvent molecules, ligands, and other heteroatoms. It is considered good practice to validate the suitability of a chosen docking software to model a desired binding pocket, which is most often done by re-docking. It consists of removing a native ligand from the modeled system and placing it back using the docking method of choice. If the best pose output by the software matches the bioactive pose of the native ligand, it is assumed that both the conformation sampling algorithm and the scoring function adequately describe the modeled system and can be used to model interactions of novel ligands with the pocket [88,89].

With protein-ligand docking being an efficient and quick way to obtain significant intuition for drug design and optimization, it has been used in several contexts of tubulin-related drug design. For example, it is often included in VS campaigns as one of the last steps to prioritize a virtual hit for further investigation. As such, Mangiatordi et al. used protein-ligand docking to further filter the results of a prior pharmacophore screening and prioritize remaining compounds, the latter containing 31 novel colchicine-site targeting agents with in vitro anti-proliferative properties [18]. In a similar manner, Guo et al. reported protein-ligand docking as an essential step that allowed them to discover eight confirmed cytotoxic agents targeting the colchicine binding site [26]. Moreover, Zhou

et al. used protein-ligand docking to highlight five virtual hits found by pharmacophore screening as most promising ones, their cytotoxic action related to binding at colchicine site was later confirmed in vitro [83]. A work by Ayoub et al. showed how docking-based optimization of VS hits could benefit from pose rescoring using the MM/PBSA method [16].

A noteworthy work by Zhang et al. compared five docking programs by re-docking 10 complexes of tubulin co-crystallized with taxane-site targeting ligands and selecting the three best software programs for evaluation of virtual hits found by pharmacophore screening; among the prioritized molecules, two were established as cytotoxic agents, supposedly targeting the taxane binding site [32]. Protein-ligand docking was instrumental in highlighting 15 virtual hits found by pharmacophore screening in the work by Nagarajan et al., later experimentally confirmed to be cytotoxic in vitro due to targeting the colchicine site of the tubulin protein [82]. Similarly, Federico et al. used docking to evaluate potential affinity of found virtual hits toward tubulin's colchicine site, eventually discovering seven micromolar inhibitors of tubulin polymerization [19]. Consensus docking of pharmacophore screening virtual hits helped Elseginy et al. establish four novel compounds with significant antiproliferative activity against cancer cells due to targeting the colchicine site of the tubulin protein [85]. Interestingly, Mao et al. incorporated protein-ligand docking and interaction fingerprint similarity comparison to discover a novel taxane-site targeting promoter of tubulin polymerization [90]. Lastly, Stefanski et al. also combined docking and fingerprint similarity measure of protein-ligand interactions as a last step of a VS campaign that yielded two potent in vitro cytotoxic colchicine-site targeting agents [27].

Protein-ligand docking is a powerful VS tool that alone can produce high-quality results. For example, Zúñiga-Bustos et al. used only protein-ligand docking to screen a large compound library, with virtual hits being confirmed promotors of tubulin polymerization targeting the laulimalide binding site [91]. In another study, Liu et al. screened a large database with consecutive docking experiments with increasing rigor of conformational sampling, eventually yielding six hits with in vitro antitumor activity due to targeting the colchicine binding site [92]. In a similar manner, Liu et al. docked a large compound library and discovered two colchicine-site targeting in vitro inhibitors of tubulin polymerization among the highest ranked molecules [93].

Often, protein-ligand docking is used as a way to provide rationale for a tubulin-targeting agent's biological action. In such case, designed molecules are docked into one or several potentially targeted binding sites. Best estimated poses are then examined in terms of docking scores and physicochemical interactions within the site. Such analysis may also provide ideas for further compound optimization. For example, docking studies were used to assess possible binding modes and guide rational design of colchicine-site targeting compounds of different classes independently reported by Ameri et al. [94], Guo et al. [17], Riu et al. [95], Patel et al. [96], and Mustafa et al. [97]. In a similar manner, Tripathi et al. [98], Ayoub et al. [99], and Chávez-Estrada et al. [100] used protein-ligand docking to estimate putative binding modes of taxane-site targeting molecules. Interestingly, Forero et al. [101] predicted possible binding modes of the designed compounds by docking them into both colchicine and taxane site, eventually settling on colchicine site as the possible target of the designed compounds based on interaction analysis. Finally, Pandit et al. [29] used docking to evaluate binding regimes of vinca-site targeting peptides. Table A5 provides an overview of exact implementations of docking protocols used in mentioned works.

3.3. Molecular Dynamics (MD) Simulations to Study Tubulin-Ligand Complexes

3.3.1. Classical MD Simulations Used on Tubulin

Molecular dynamics (MD) is a computational simulation technique that allows exploration of the behavior of a molecular system over time by solving Newton's equations of motion. This is of great importance for research, as biomolecules are dynamic entities whose atoms are in constant motion. In this way, by using MD, time-dependent processes in molecular systems can be monitored to facilitate the analysis of their structural, dynamic, and thermodynamic properties.

MD simulations can provide valuable information that is not accessible from experiments, allowing the formulation of new hypotheses. In addition, technical progress, both in algorithm efficiency and computational power, allows the study of biological macromolecules of larger dimensions on longer timescales, and the predictions that are inferred from these simulations make MD simulations a very valuable computational approach in the drug design field.

MD is widely used as a computational technique to examine protein-ligand complexes, such as the binding of molecules to tubulin and MTs, to analyze the effects on the tubulin structure upon ligand binding.

In the study of MTAs in complex with tubulin using classical MD simulations, different settings need to be considered during system preparation. For instance, the choice of the force field that best suits the system under study is important, since the quality of the MD simulations results depends on the quality of the energy function used to treat the interactions among atoms in the system. Additionally, the simulation time and the MD engine used are important factors that also condition the accuracy of the simulations.

In this review, Table A6 summarizes the settings used by scientists to set up classical MD simulations to investigate tubulin-ligand complexes. Due to the number of articles related to this topic published since 2019, we have decided to dedicate the review of classical tubulin MD simulations to the articles which were published in the last three years and thus are the most up-to-date manuscripts.

By analyzing Table A6, we can observe that most often the tubulin-ligand complex systems are simulated under periodic boundary conditions, solvated in explicit water (TIP3P or SPC water model) in cubic or octahedral box at room temperature and atmospheric pressure. The typical simulation time is ~100 ns. While different force fields are explored, the most prevalent are Amber Force Fields FF99SB and the more recent one FF14SB.

3.3.2. Enhanced Sampling Methods

Enhanced sampling algorithms have appeared as a powerful tool for increasing the efficiency of classical MD simulations. During a certain simulation time, enhanced sampling methods allow for the sampling of larger areas of a complex system configuration space. The accuracy of the results is highly dependent on the selection of the simulation settings. Here, we outline three different enhanced sampling methods used to study tubulin-ligand binding mechanisms.

4. Umbrella Sampling (US)

Umbrella sampling (US) is an enhanced sampling computational technique applied to expand the sampling of a system in which ergodicity is hampered by the form of the energy landscape of the system. US is used to calculate the thermodynamic parameters for the binding of a ligand to a protein. In the tubulin field, US has been used to predict the strength of binding (binding energy) of a ligand to tubulin by slowly pulling away the ligand from the binding site. ΔG_{bind} derives from the potential of mean force (PMF), obtained from a series of US simulations. Several initial positions of the ligand with respect to the protein of interest are generated, each corresponding to a location where the ligand is harmonically restrained at increasing center of mass (COM) distance from other selected groups via an umbrella biasing potential. These restraints allow the ligand to sample the conformational space in a defined area along a single degree of freedom (reaction coordinate) [102].

US is subject to certain limitations, such as biases in sampling due to improper selection of reaction coordinates (RCs), challenges in identifying appropriate RCs for complex systems, the need for multiple RCs in systems with multiple reaction pathways, and the method being dependent on the choice of RC. Additionally, the method can be computationally expensive and limited to systems with multiple reaction pathways and high-dimensional systems.

Zhang et al. used US simulations to retrieve the free energy potential of αβ-tubulin separation upon binding to a certain ligand [103]. Also, Zhou et al. and Mane et al.

simulated the αβ-tubulin dissociation free energy under different system conditions using the US method [104,105].

5. Steered Molecular Dynamics Simulations (SMD)

Steered molecular dynamics (SMD) is another enhanced sampling method in which an additional external force is applied to one or more atoms in the studied system to maintain the constant speed of motion along a selected coordinate [106]. SMD emulates atomic force microscopy (AFM) experiments. It allows the study of molecular processes, such as the protein-ligand unbinding mechanism, by focusing on selected degrees of freedom. It is important to keep in mind that in SMD the force applied is not necessarily proportional to the binding free energy, as it aims to simulate the process of binding a molecule to another, rather than the equilibrium state of the bound complex.

Rai et al. performed SMD to study the bonding strength between eribulin and tubulin isotypes to which it presented the highest (aVIIIbIII) and lowest (aIbII) binding energies, which were previously calculated computationally. They kept the tubulin structures fixed by setting position restraints on their heavy atoms, whereas the eribulin structure was dynamic. They observed that a three-fold greater force was required to pull out eribulin from the active site of one tubulin isotype in comparison to that of another isotype [107].

6. Metadynamics (MetaD)

Metadynamics is an enhanced sampling technique that enables conformational sampling of the free energy landscape of a system through the use of collective variables that describe it. Castro-Álvarez et al. used MetaD to study the effect in the tubulin M loop on the binding of laulimalide and peloruside A to the taxane site [80].

Binding pose metadynamics (BPMD) allows for the assessment of the stability of the ligand in solution. This is because BPMD can differentiate between stable and unstable binding geometries. It is expected that the unstable ligand poses will rarely be occupied in the energy landscape under MetaD simulation bias. As a result, unstable ligand poses make a minimal contribution to binding affinity.

Boichuk et al. applied BPMD to evaluate the stability of a colchicine binder in complex with tubulin and to select its most stable conformation using as collective variables the RMSD values of the heavy atoms of the ligand [108]. Fusani et al. compared the binding mode of epothilone A in complex with tubulin of the first published 3D structure solved by Nettles et al. (PDB: 1TVK) and a later one solved by Prota et al. (PDB: 4I50) using BPMD. Fusani et al. wanted to differentiate between the correct and incorrect ligand binding poses by applying BPMD [109].

Moreover, Gaspari et al. used MetaD to induce the *cis*-to-*trans* isomerization of a colchicine binder in complex with tubulin. This allowed the authors to calculate the difference in binding free energy between the *cis* and *trans* isomers of the ligand via a thermodynamic cycle. Furthermore, Gaspari et al. also used MetaD to gain insight into the differences in the unbinding process of colchicine and another colchicine site binder studied in complex with tubulin [110].

When using MetaD as an enhanced sampling method, it is important to be aware of its limitations, particularly in relation to the selection of the collective variable (CV). These limitations include potential bias in sampling, challenges in identifying appropriate CV for complex systems, increased computational cost for high-dimensional systems, and limitations in exploring the free energy surface.

7. Applications of MD for Tubulin-Ligand Studies

7.1. Docking Validation and Refinement

MD is often used as a post-processing technique to validate and refine the binding modes of the protein-ligand complexes obtained from docking experiments. MD applied for docking validation has also been used in the tubulin research field.

For example, Hadizadeh et al. investigated the possible binding mode of an active tubulin binder (9IV-c) that showed high activity against human tumor cell lines. For this, they used computational methods such as docking and MD. First, they docked 9IV-c in the colchicine site, and the output was later submitted to MD simulations to evaluate and refine the docking results. The simulation of the complex was analyzed using root mean square deviation (RMSD), radius of gyration (Rg), and hydrogen bond stability values. In this way, they obtained a successful prediction of the way 9IV-c binds to tubulin, allowing them to conduct further computational studies to identify new potent tubulin inhibitors [111].

El-Mernissi et al. designed four new colchicine site binders using 3D-QSAR models and docking based on a series of 2-oxoquinoline arylaminothiazole derivatives that were identified as promising tubulin inhibitors. Among the four newly designed binders, MD simulations of the compound with the best docking score were performed to validate its docking binding pose using the RMSD, root mean square fluctuation (RMSF), Rg, and solvent accessible surface area (SASA) metrics. By performing MD simulations, they confirmed the conformational stability of the complex, thus validating their docking experiments [112].

Zhang et al. performed VS using a combination of molecular docking methods of 50 compounds in the taxane site to search for novel tubulin polymerization inhibitors. Subsequently, the best hits were submitted to IC50 experiments, from which the two compounds with the highest antiproliferative activity were selected for MD simulations along with the tubulin-paclitaxel complex. By performing MD simulations, they further studied the binding mode, stability, and molecular interaction pattern of the docking results. Apart from using RMSD, RMSF, and Rg as MD analysis metrics, they performed clustering analysis to extract information on how tubulin in complex with the three studied taxane-site binders is sampling the conformational space. They used 'BitClust' [113], which is a relatively new faster implementation of the Daura et al. clustering algorithm that performs rapid structural clustering of long trajectories [114]. In this way, using MD simulations, they validated the stability of tubulin in complex with the two compounds and probed the mechanism of their interactions, which aligned with the experimental results [115].

Elhemely et al. observed that a meta-substituted 3-arylisoquinolinone that had shown a high cytotoxic effect in several cancer cell lines mimicked the structure of colchicine. They hypothesized that its mode of action could be related to its binding to the colchicine site of tubulin. To test the suitability of the compound to bind to this site, the authors first performed docking experiments, which were later refined by MD. These computational studies suggested that the meta-substituted 3-arylisoquinolinone was able to bind well to the colchicine binding site [116].

7.2. Comparison of the Binding Free Energy of Different Ligands

The resulting trajectories from MD simulations are also used to compute the free energy of binding of different molecules binding to the same site to obtain a quantitative measure to compare and rank the best hits normally resulting from docking studies. There are different methods to estimate the free energy of binding of protein-ligand complexes such as Free Energy Perturbation (FEP), Molecular Mechanics Generalized-Born Surface Area (MM-GBSA), and Molecular Mechanics Poisson-Boltzmann Surface Area (MM-PBSA). Due to the numerous computational resources required for the performance of MD simulations, this approach can only be used to rank a low number of molecules, in the tens range.

Elhemely et al., in the article mentioned above, computed the free energy of binding applying the MM-GBSA method using the MD-based refined complexes of two 3-arylisoquinolinones bound to tubulin that only differed in the location of a substituent in their structure (meta versus para). The authors wanted to investigate how the change in the substituent position could alter the free energy of binding and compare the binding mode of the molecules in the tubulin sub-pocket. The computational results aligned with the experimental ones, concluding that the meta-substituted molecule was a better colchicine site binder than the para-substituted compound [116].

Stroylov et al. used FEP calculations based on MD simulations for predicting tubulin-ligand free binding energy differences of new tubulin polymerization inhibitors targeting the colchicine site [117].

Mao et al. with the goal of discovering new tubulin inhibitors capable of binding to the taxane site, performed a VS of ~1.6M molecules retrieved from the ChemDiv database. After applying different computational filters, 17 hit compounds were selected and submitted for experimental evaluation. The in vitro tubulin polymerization assay found P2 to be the most promising compound. Therefore, P2 was submitted to MD simulations not only to further investigate the interactions between P2 and tubulin based on the docking results but also to compare it with paclitaxel, an already known active taxane-site binder. They calculated the free energy of binding of both complexes using the MM-PBSA method obtaining—68.25 ± 12.98 kJ mol^{-1} for the tubulin-P2 complex and—146.05 ± 16.17 kJ mol^{-1} for the tubulin-paclitaxel complex. These results were in line with the experimental evidences, defining P2 as a lead compound that could be used for new tubulin inhibitors drug design campaigns [90].

7.3. Identification of Key Binding Site Residues

MD is also used to further investigate the mechanisms of interactions between tubulin and hits, as previously reported, and to find key amino acids in the protein that are especially important for binding to the studied ligand within a given tubulin binding site, also called 'hot spots'.

Neto et al. studied a series of chalcones predicted to bind to the taxane site using both experimental and computational approaches, including MD simulations. To identify the key binding site residues establishing the strongest interactions with the studied ligands, the authors performed Computational Alanine Scanning (CAS) of each tubulin-ligand interface. This allowed analysis of the free energy contribution of the amino acids located at the taxane site, bringing new insights into this tubulin site for further exploitation using chalcones [118].

Gamya et al. reported a noscapine derivative (VPN) discovered and validated using computational tools such as docking and MD simulations. VPN was able to be properly accommodated in the colchicine site according to the docking results, which were then submitted to MD studies for validation of its stability at the site by calculating the RMSD and RMSF values, and its binding free energy using the MM-GBSA and MM-PBSA methods. Furthermore, they performed a deeper analysis of the interactions established between the residues of the receptor with the ligand by calculating the energy contribution of each residue in the binding of VPN by performing Per Residue Energy Decomposition (PRED) analysis using the MM-GBSA method. In this way, they were able to identify the residues that have the greatest impact on the binding and stability of VPN, the 'hotspots' [119]. Other researchers have also applied PRED analysis to the search for 'hotspots' to investigate the details of tubulin-ligand interactions at the atomic level [90,120].

7.4. Analysis of Local and Global Effects upon Ligand Binding

Structure-based computational approaches have also been used to investigate the effect of different MTAs on the local geometry of tubulin. Moreover, since MTs are formed by allosteric proteins, the effect of binding of a ligand at one site can also cause non-local effects in MTs, and therefore, the study of global effects caused by ligand binding is also important.

For example, the M loop has been widely studied by X-ray crystallography and other structural techniques to understand the effect of taxane site binders on this loop [44,70]. This is due to the fact that the M loop is found at the β1/β2 interface and is involved in the stability of the interaction. However, the dynamics of M loops remains unclear, and other research groups approach these questions using SB computational techniques. Castro-Álvarez et al. performed MetaD simulations of laulimalide and peloruside A to analyze the changes produced in the M loop upon binding of these ligands [80]. MetaD

helped explain how laulimalide and peloruside A shift the M loop to an α-helix structure by bringing together different residues at the external site of β1.

Basu et al. studied the collective changes that the tubulin over-stabilizing agents paclitaxel and taxotere induce on the structure and dynamics of the α,β-tubulin dimer by performing MD simulations. To study the conformational effects of tubulin induced by the binding of the ligands, they also performed MD of the apo protein to compare the results of the simulations of apo tubulin with those of holo tubulin. They investigated the influence of ligand binding on the essential dynamics of tubulin using Principal Components Analysis (PCA). They observed that the apo tubulin samples a broader range of conformations than that of the holo tubulin. Therefore, the presence of the ligands biases the system toward a more stabilized conformation. Moreover, for a more local structural exploration, the authors performed a Define Secondary Structure of Proteins (DSSP) analysis to study the conformational changes of the M loop and its associated regions induced by the binding of the two ligands. More computational analyzes were performed to thoroughly investigate the effect of binding of both paclitaxel and taxotere on the dimeric structure, concluding that these ligands enhance the α,β-tubulin dimer to be more favorably accommodated into the MT superstructure [121].

7.5. Exploration of Ligand Binding to Different Tubulin Isotypes

The α and β tubulin in eukaryotes consist of isotypes that differ in their aminoacidic sequence. Therefore, in the field of tubulin, researchers study not only the binding of different ligands to the same binding site of a certain tubulin isotype, but also the binding of the same ligand to different tubulin isotypes [122]. In silico approaches have a great advantage in the study of tubulin isotypes, since they are rarely accessible to be investigated experimentally. In silico strategies allow for the analysis of the sensitivity of a certain ligand to bind to tubulin isotypes which would be highly demanding to do experimentally. Rai et al. performed MD simulations of the potent anticancer drug eribulin bound to different tubulin isotypes to report differential binding affinities. However, it remains to be explored how the residue composition at the binding site between tubulin isotypes translates into major changes in the tubulin conformation and the binding affinities with ligands [107].

7.6. MD Analysis Metrics

As previously described, MD simulations have multiple applications in the *in silico* study of tubulin-ligand complexes. To extract the information of interest from the output of MD simulations (trajectory), different analysis metrics are available. In Table 3 we present the techniques that have been used in the selected tubulin-related articles to analyze MD simulations of tubulin and its interactions with MTAs.

Table 3. A glossary of key parameters and procedures used to analyze observed conformational changes during MD trajectories.

MD Analysis Metrics	Definition	Examples of Application
RMSD	The root mean square deviation (RMSD) is a standard measure of the structural distance between coordinates: it measures the average distance between a group of atoms. RMSD values help to evaluate the global structural stability of the system studied in the simulation.	Dash 2022 [119], El-Mernissi 2022 [112], Zhang 2022 [115], Zhao 2022 [120], Radha 2022 [123]
RMSF	The root mean square fluctuation (RMSF) represents the quadratic deviation of the atoms in temporal averages. RMSF values help to evaluate the internal structural flexibility of the studied system in the simulation.	Dash 2022 [119], El-Mernissi 2022, Zhang 2022 [115] Radha 2022 [123], Talimarada 2022 [124]

Table 3. *Cont.*

MD Analysis Metrics	Definition	Examples of Application
Rg	The radius of gyration (Rg) is defined as the mass-weighted root mean square atomic distance from the center-of-mass and can be applied to measure the level of structural compactness of a protein at different time points during the trajectory.	Hadizadeh 2022 [111], El-Mernissi 2022 [112], Zhang 2022 [115], Radha 2022 [123]. Rai 2022 [107]
SASA	The solvent accessible surface area (SASA) permits assessment of the overall changes in the tertiary structure of a molecule and its solvent accessibility over the course of the simulation.	El-Mernissi 2022 [112] Rai 2022 [107]
2D interaction analysis	2D interactions established between the protein and the ligand along the course of the simulations help to identify the residues within the binding site that play an important role in the binding of the ligand to the receptor and to list the 'hot spots' between the ligand and the protein.	Basu 2022 [121], Mao 2022 [90], Zhao 2022 [120], Rai 2022 [107], Zhang 2022 [103], Majumdar 2022 [125], Mao 2022 [90], Hadizadeh 2022 [111], Zhang 2022 [115]
DSSP	The Define Secondary Structure of Proteins (DSSP) algorithm is the standard method for assigning a secondary structure to amino acids of a protein given the atomic resolution coordinates of the protein.	Mao 2022 [90], Basu 2022 [121]
Clustering	Clustering is a data mining technique that allows molecular configurations to be grouped into subsets based on the similarity of their conformations.	Zhang 2022 [115]
Binding free energy	The Gibbs free energy (G) provides valuable information about the structure and stability of biomolecules. It is possible to calculate the predicted binding energy (ΔG_{bind}) of a given tubulin-ligand complex using the MD simulation trajectory of this biomolecular association.	Zhao 2022 [120], Zhang 2022 [115], Elhemely 2022 [116], Rai 2022 [107], Radha 2022 [123], Majumdar 2019 [125]
PRED	The Per Residue Energy Decomposition (PRED) is a computational tool that is used to obtain the residue-wise contribution to the total binding free energy. It provides information on the key residues that contribute to protein-ligand association, the so-called 'hot spots'.	Dash 2022 [119], Mao 2022 [90], Zhao 2022 [120], Zhang 2022 [120]
CAS	Computational Alanine Scanning (CAS) is a technique that consists of the mutation of amino acids present on the interaction surface between the protein and the ligand to alanine, and the measurement of the difference in binding free energy between the ligand and the native protein and the ligand and the multiple mutated proteins to identify 'hot spots'.	Neto 2022 [118]
PCA	Principal Component Analysis (PCA) is a linear dimensionality reduction tool used in the MD field to map the coordinates of each frame of the trajectory to a linear combination of orthogonal vectors and to investigate the internal modes of motion of the system under study.	Basu 2022 [121]

8. Conclusions

In this review, we provide an overall picture of the different ligand and structure-based computational methods that have been used in recent years for the study of tubulin-targeting agents, and an overview on the available MT and tubulin structural data. We observed that computer-aided methods have had significant contribution to the field of tubulin-targeting drug design. VS of compounds, applying both ligand and structure-based approaches, provided many hits with in vitro bioactivity. An advantage of ligand-based methods is their computational efficiency and ability to work with big data. They are often

beneficial to the early stages of VS, where the goal is to filter out compounds irrelevant to the task at hand in a fast manner. These initial results are well suited for subsequent filtering by structure-based methods, which provide more intuition behind the physico-chemistry of potential interactions between a given virtual hit and the desired biomolecular target. Computational methods were also shown to guide in rational design and optimization of novel tubulin-targeting agents.

Moreover, despite the large number of available tubulin binding sites (8), our analysis shows that the colchicine and the taxane sites are the most studied ones in tubulin-related computational research while the rest are underrepresented. We also observed a tendency to mainly use structure-based methods to find tubulin-targeting agents such as molecular docking for VS and MD for the refinement of the resulting docking hits.

MD simulations have widely been used in the tubulin-directed drug discovery field. In the recent literature, there is a tendency to use MD as a computational docking post-processing method that allow the validation and refinement of the docking results, the analysis of the ligand–tubulin dynamics and the estimation of binding free energies.

We expect growth of interest in these computationally understudied sites in the near future since computational strategies are becoming essential in the first steps of the drug design campaigns.

Author Contributions: H.P.-P., A.-C.A. and M.S.—extensive literature review and original draft preparation, A.E.P., S.P. and D.H.—editing, review and supervision. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the H2020-MSCA-ITN-2019 (860070 TUBINTRAIN).

Institutional Review Board Statement: No bibliographic database was harmed during this study.

Informed Consent Statement: Not applicable.

Data Availability Statement: This is a review article, no new data was created.

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

Appendix A

Table A1. Summary of similarity search implementations.

Reference	Screened Dataset	Dataset Size	Software	Descriptors	Similarity Metric	Result
Ayoub 2013 [16]	PubChem	33×10^6	Built-in web search	881-bit PubChem subgraph fingerprints [URL]	Tanimoto, 80% similarity threshold	Virtual hits ranked by protein-ligand docking, one compound used as a reference for further successful design
Guo 2019 [17]	ChemDiv	1.7×10^6	Discovery Studio (Biovia)	ECFP-4	Tanimoto, 50% similarity threshold	Virtual hits were found to be cytotoxic, one was confirmed as a colchicine site binder
Lo 2015 [20]	ChEMBL, PubChem	35×10^6	CSNAP2D	OpenBabel FP2	Tanimoto, 85% similarity threshold	Correctly identified and validated tubulin as a target for 36 molecules that showed cytotoxicity in a HTS setting
Lo 2016 [21]			CSNAP3D	ShapeAlign	3D Tanimoto, 85% similarity threshold	A virtual hit was established to promote tubulin polymerization by binding at the taxane site
Magiatordi [18]	CoCoCo	3.7×10^6	Phase (Schrödin-ger)	Atom-type-based 3D shape	Atom-type volume scoring, 0.65 similarity threshold	31 virtual hits have been confirmed to decrease microtubule polymerization in vitro.
Federico 2020 [19]	ZINC *, Chembridge Diverset CL, Chembridge Diverset EXP, BindingDB FDA, MayBridge	164×717	ROCS (OpenEye)	Smooth 3D Gaussian functions for each atom	Tanimoto similarities of aligned overlap volumes (no threshold, top 5000 selected for ROCS, top 2000 for EON)	Two virtual hits established by shape and electrostatic similarity to a known active were shown to inhibit tubulin polymerization in vitro.
			EON (OpenEye)	Electrostatic potential maps of pre-aligned molecules		

* (Drug Database, Naturals).

Table A2. Summary of recent ligand-based QSAR modeling.

Reference	Modeled Data	Descriptor Type	Algorithm	Validation Strategy	Application and Result
Gaikwad 2018 [25]	IC_{50} of 102 phenylindoles cytotoxic against MCF7 cancer cell line	Fragment-based holograms implemented in SYBYL-X (Certara) Extended connectivity fingerprints, physicochemical descriptors	PLS Naïve Bayes (Discovery Studio 3.0, Accelrys)	Two sets were used: training (77) and test (25). Leave-one-out and five-fold cross-validation were used.	Analysis of literature data allowed the authors to highlight structural features important for cytotoxicity.
Guo 2020 [26]	1076 diverse colchicine-site targeting small molecules extracted from the ChEMBL database	Extended-connectivity fingerprints, path-based fingerprints	Naïve Bayes Single Tree Random Forest	Five-fold cross-validation.	A colchicine site-binding inhibitor of tubulin polymerization was established after a virtual screening campaign.
Stefanski 2018 [27]	IC_{50} of 83 thio-derivatives of combretastatin-A4 mined from literature	Extended connectivity fingerprints, physicochemical descriptors	Naïve Bayes Multiple Linear Regression	Leave-one-out, cross-validation, and external test set methods. The external validation test set was composed of 20 tubulin inhibitors and 800 decoys.	Two virtual hits selected by consensus QSAR modeling were later confirmed to be cytotoxic due to perturbing microtubule polymerization by binding at the colchicine site.
Quan 2018 [28]	IC_{50} values of 64 literature-mined derivatives of combretastatin A-4	CoMFA (steric and electrostatic fields) CoMSIA (steric, electrostatic, hydrophobic, hydrogen bond donor, and hydrogen bond acceptor fields)	PLS (SYBYL-X 2.0, Tripos)	Leave-one-out validation	A 3D QSAR study highlighted structural elements with pronounced relation to activity value, useful for further optimization.
Pandit 2021 [29]	IC_{50} values of 49 tubulysin derivatives reported in the literature	CoMFA (steric and electrostatic fields) CoMSIA (steric, electrostatic, hydrophobic, hydrogen bond donor, and hydrogen bond acceptor fields)	PLS (SYBYL-X 2.0, Tripos)	Cross-validation	3D QSAR investigation of structure-activity data on tubulysins lead to rational design and synthesis of a new class of cytotoxic in vitro tubulysin derivatives

Table A3. Summary of ligand-based pharmacophore screening campaigns.

Reference	Compound Library	Compound Set Used to Build the Model	Software Used to Build the Model	Model Generation and Validation Settings	Validation Set	Validation Metric and Score	Screening Result
Zhang 2021 [32]	BioDiversity, 30,000 molecules	Six agents targeting taxane site	HipHop algorithm from Discovery Studio 3.5 (Accelrys)	Five features were used (HBA, HBD, HP, HP-A, and R-A) [1], paclitaxel used as reference	467 inactive molecules from ZINC15 database, 33 known inhibitors	Gunner-Henry (GH) score of 0.62	Large database filtered to focus on a subset that eventually led to discovery of two taxane-site targeting cytotoxic agents
Lone 2017 [33]	IBScreen Natural Product Database, 84,215 molecules	Four C20 substituted vinblastine analogues extracted from literature	Phase (Schrödinger)	HBA, HBD, HP, PI, and R-A [1]	35 inactive and four active C20 substituted vinblastine analogues	The Survival-inactive score of 4.006.	Possibility of scaffold-hopping for vinca-site-targeting compounds design was shown
Niu 2014 [34]	Specs Screening Database, 202,919 molecules	26 compounds designed to target colchicine site with known cytotoxic action	HypoGen module from Discovery Studio 2.5 (Accelrys)	HBD, HBA, HP, and R-A [1]	66 colchicine site-targeting compounds with known cytotoxicity (26 actives, 40 inactives)	Cost difference	Two compounds with good fitness to the developed pharmaco-phore model were shown to be tubulin polymerization inhibitors in vitro.
Stefanski 2018 [27]	A custom-designed virtual combinatorial library of 1159 combretastatin A-4 analogs	21 active colchicine site-targeting molecules mined from literature	Discovery Studio 3.5 (Accelrys)	HBA, HBD, HP, and HP-A [1]	20 tubulin inhibitors and 800 decoys mined from ChEMBL	Area under receiver-operator curve (AUROC)	Two virtual hits were established as in vitro cytotoxic agents targeting colchicine binding site

[1] HBA = hydrogen bond acceptor, HBD = hydrogen bond donor, HP = hydrophobic, HP-A = hydrophobic-aromatic, R-A = ring-aromatic, PI = positive ionizable bond.

Table A4. An overview of structure-based pharmacophore screening implementations.

Reference	Data Used to Build the Model	Software	Validation Set	Screened Data	Result
Nagarajan 2015 [82]	1SA0, 1SA1, 3HKC, 3HKE, 3HKD, 3N2K, 3N2G; model derived from 1SA0 was manually removed of a hydrogen bond feature, shown best result in validation	Model building: LigandScout v3.1 (Inte:Ligand); Screening—Phase v3.4 (Schrödinger)	52 active colchicine site binders mined from literature, 1800 decoy molecules from the DUD database	The CoCoCo database, containing multiconformer data on 3.7 million purchasable compounds	31 novel colchicine site-targeting inhibitors of tubulin polymerization were established that match the derived pharmacophore model
Mangiatordi 2017 [18]	6F7C, 5EYP, 5YL2, 4O2B (common feature model)	MOE (Chemical Computing Group Inc.)	970 inactive molecules and 30 known inhibitors with experimental activity mined from literature	Specs database, 202,919 molecules	The screening established five virtual hits that are cytotoxic in vitro, one most potent hit confirmed to bind at the colchicine site
Zhou 2019 [83]	118 crystal structures of tubulin co-crystalized with colchicine site binding ligands	LigandScout v3.1 (Inte:Ligand), Phase v3.4 (Schrödinger), Pharmer	81 co-crystalized ligands and 3354 decoys randomly extracted from the DUD-E database	A subset of specifically selected 8918 purchasable compounds from the ZINC database	Ensemble of many pharmacophore models based on colchicine site-bound ligands structures was used in virtual screening which led to discovery of a potent tubulin-targeting cytotoxic agent
Zhang 2021 [32]	1JFF	Discovery Studio 3.5 (Accelrys)	467 inactive molecules from ZINC15 database and 33 known inhibitors with experimental activity	BioDiversity, 30,000 molecules	Large database filtered to focus on a subset that eventually led to discovery of two taxane-site targeting cytotoxic agents
Gallego-Yerga 2021 [84]	1SA0	Protein-ligand interaction fingerprints (PLIF) implemented in MOE (Chemical Computing Group Inc.)	No additional validation performed	A subset of 100, 000 compounds from ZINC15 database	Virtual screening campaign yielded a novel cytotoxic agent disrupting tubulin polymerization by binding at colchicine site
Elseginy 2022 [85]	3E22, 3HKD, 3HKE, 3HKC, 1Z2B, and 1SA1	Discovery Studio 2.5 (Accelrys)	40 literature-mined tubulin inhibitors targeting the colchicine site, 2000 decoy molecules randomly selected from ChemDiv library	ChemDiv library, 700,000 molecules	A virtual screening campaign discovered an in vitro potent cytotoxic hit targeting the colchicine binding site

Table A5. Protein-Ligand Docking.

Screening Setup				Results			Reference
Binding Site	**Binding Site Definition**	**Docking Software**	**Screened Set**	**Hit No**	**Hit Rate, %**	**Best Compound's Activity**	
Virtual screening in succession to other computational methods							
Colchicine	Extracted from 1SA0 as a 10 Å-wide cubic box around the center-of-mass of the native ligand	Glide SP	25,146 virtual hits established by pharmacophore screening of CoCoCo database	68	35%	Inhibition of tubulin polymerization at IC_{50} of 3 μM	Mangiatordi 2017 [18]
Colchicine	Extracted from 5H7O as 10 Å-wide cubic box around the center-of-mass of the native ligand	Glide SP, Glide XP	30,327 virtual hits of pharmacophore screening of SPECS library	8	20%	Anti-proliferative activity (IC_{50}) against different cancer cell lines in range 6.14–15.06 μM	Guo 2020 [26]
Colchicine	Extracted from 6F7C (exact settings not specified)	MOE	3135 virtual hits found by pharmacophore screening of SPECS library	5	100%	80% growth inhibition rate against five different cell lines	Zhou 2019 [83]
Taxane	Extracted from 1TVK as a grid box centered around the native ligand with each dimension a size of 5.8 Å	AutoDock 4.2	645 virtual hits yielded by similarity search in PubChem	1	20%	Established hit got satisfactory predicted physiochemical properties; later work saw an analog compound synthesized and tested	Ayoub 2013 [16]
Taxane	Extracted from 1JFF as a sphere containing the residues within 11.5 Å from the ligand	AutoDock Vina Gold CDOCKER	1309 virtual hits established by a pharmacophore screening of the BioDiversity database	11	22%	Anti-proliferative activity (IC_{50}) against four cancer cell lines ranging from 10.31 μM to 21.04 μM	Zhang 2021 [32]
Colchicine	Extracted from 1SA1 as all residues around the ligand at a 6.5 Å distance	SurFlex-Dock	1739 virtual hits found by pharmacophore screening of the ChemDiv library	1	1.78%	Tubulin polymerization inhibition IC_{50} value of 17.6 μM	Nagarajan 2015 [82]
Colchicine	Extracted from 4O2A as a sphere of 8 Å radius around the native ligand	GOLD	Around 3000 virtual hits procured by ligand-based virtual screening of six chemical libraries	3	43%	IC_{50} of 83.61 μM in hepatotoxicity model	Federico 2020 [19]
			Three databases: Chembridge Diverset EXP, Chembridge Diverset CL, and ZINC natural products	4	66%		

Table A5. *Cont.*

Screening Setup						Results	Reference
Binding Site	**Binding Site Definition**	**Docking Software**	**Screened Set**	**Hit No**	**Hit Rate, %**	**Best Compound's Activity**	
Colchicine	Defined as a 20 Å-wide grid box around the centroid of the native ligand from the 1SA0 structure	MOE, BUDE, AutoDock 4.2	2746 virtual hits from a pharmacophore screening of a subset of ZINC15 library	4	30%	Tubulin polymerization inhibition IC_{50} = 6.1 μM	Elseginy 2020 [85]
Taxane	Extracted from 1JFF as a 23 Å-wide box around the native ligand	AutoDock Vina	1,601,806 compounds from the ChemDiv library	1	5.8%	IC_{50} value against four cancer cells in range from 9.21 to 17.30 μM	Mao 2022 [90]
Colchicine	Extracted from 1SA0, 1SA1 (exact procedure not specified)	Glide SP	1159 compounds from an in-house library	6	35%	Tubulin polymerization inhibition at IC_{50} = 0.85 μM	Stefanski 2018 [27]
Virtual screening based on protein-ligand docking only							
Peloruside	Extracted from 4O4J as a cubic grid of 20 Å in size	AutoDock 4.2	2000 virtual hits established after docking a 6 million ZINC subset with AutoDock Vina	3	48%	Cell viability of HeLa cells decreased after 48 h by 60% at 100 μM	Zuniga-Bustos 2020 [91]
Colchicine	Extracted from 4O2B as all residues closer than 12 Å to the centroid of the native ligand	Glide SP, GOLD	40,000 virtual hits obtained by high-throughput docking with Glide HTVS of IBScreen library	2	13%	Tubulin polymerization inhibition IC_{50} = 23.5 μM	Liu 2022 [92]
Colchicine	Extracted from 4O2B as a cubic grid of 20 Å in size	AutoDock 4.2	212,449 compounds from the SPECS library	2	5.5%	Tubulin polymerization inhibition activity with IC_{50} value of 1.68 μM	Liu 2019 [93]
Binding mode assessment							
Colchicine	Extracted from 1SA0 as a 15 Å-wide cubic grid box centered on root point of native ligand	AutoDock 4.2	An in-house library of 48 Schiff bases	1	–	Tubulin polymerization inhibition activity with IC_{50} value of 0.16 μM	Ameri 2018 [94]
Colchicine	Extracted from 4O2B as a sphere of 12 Å in diameter center on the native ligand	CDOCKER	A virtual hit from a ligand-based screening of the ChemDiv library	1	–	IC_{50} of 2.99 μM against CNE2 cancer cell line	Guo 2019 [17]
Colchicine	Extracted from 4O2B as a 30 Å-wide cubic grid box centered on root point of native ligand	AutoDock Vina	A single compound from an in-house designed library of colchicine site targeting ligands	1	–	IC_{50} = 0.6 μM in an anti-proliferative assay against the HeLa cancer cell line	Riu 2022 [95]

Table A5. *Cont.*

Screening Setup						Results	Reference
Binding Site	**Binding Site Definition**	**Docking Software**	**Screened Set**	**Hit No**	**Hit Rate, %**	**Best Compound's Activity**	
Colchicine	Extracted from 6Y6D as a 12 Å-wide grid box around the native ligand	Glide XP	In-house library of 9-arylimino noscapinoids	3	–	Anti-proliferative activity with IC_{50} of 10.8 μM against MCF-17 cancer cell line	Patel 2021 [96]
Colchicine	Extracted from 1SA0 as a 25 Å-wide box around the native ligand	AutoDock Vina	An in-house library of combretastatin A4 derivatives	2	–	Anti-proliferative activity with IC_{50} = 0.62 μM against HepG2 cancer cell line	Mustafa 2017 [97]
Taxane	Extracted from 1JFF and 1TUB a 30 Å-wide grid box around the native ligand	AutoDock 4.2	Only a paclitaxel molecule was docked into tubulin mutants	1	–	Docking was used to provide rationale for paclitaxel resistance in mutant cancer cells	Tripathi 2016 [98]
Taxane	Extracted from 1TVK, 5MF4, 5LXT, and 3J6G as all residues within 6 Å distance from each native ligand	GOLD FRED	Only a lankacidin C molecule was docked into several conformations of taxane site	1	–	Ensemble docking was used to account for binding site flexibility and establish the binding mode of a recently discovered microtubules stabilizer targeting the taxane site	Ayoub 2019 [99]
Taxane	Extracted from 1JFF as a grid rectangle with a size of x = 30, y = 34, z = 26 centered on the native ligand	AutoDock 4	A single hit with the best in vitro microtubule stabilizing properties	1	–	Binding to taxane suggested as a mechanism of action, promotion of tubulin polymerization by 76% at 50 μM	Chavez-Estrada 2020 [100]
Taxane Colchicine	Extracted from 1JFF as a 21 Å-wide grid box centered on the native ligand Extracted from 1SA0 as a 21 Å-wide grid box centered on the native ligand	AutoDock 4	Three compounds with the best in vitro anti-proliferative properties from a library of 32 marine natural and semisynthetic diterpenes	3	–	Interactions fingerprint analysis after docking prioritized the taxane site as the probable binding site for designed molecules with IC_{50} < 1 μM against three cancer cell lines	Forero 2021 [101]
Vinca	Extracted from 4ZOL following an unspecified protocol	SurFlex-Dock	A known vinca-site ligand	1	–	Docking was used to guide the rational design of novel derivatives of tubulysin, which led to synthesis and validation of a hit with pronounced anti-proliferative properties attributed to binding at the vinca-site (IC_{50} = 9.4 nM against HeLa cell line)	Pandit 2021 [29]

Table A6. Details of implemented classical molecular dynamics protocols for the study of tubulin-ligand complexes.

Reference	PDB	Object of Study	MD Engine	Force field	Water Model	Time
Zhang 2019 [103]	1Z2B	Docking refinement of DVB-α,β-tubulin complex	GROMACS 4.5		SPC	100 ns
Majumdar 2019 [125]	3HKB 3HKC	Comparison of the apo α,β-tubulin dimer and α,β-tubulin dimer bound to E7010	NAMD 2.9	Tubulin: CHARMM36 Ligand: CGenFF	TIP3P	120 ns
Zhang 2021 [32]	1JFF	Docking validation of ligand–tubulin complex for	GROMACS 2019.1	Tubulin: Amber99sb-ildn Ligand: ACPYPE	SPC216	90 ns
Kumbhar 2021 [122]	4O4J	Docking validation of PLA in complex with α,β-tubulin isotypes	GROMACS 5.0	Tubulin: ff99SB-ildn Ligand: GAFF	TIP3P	100 ns
Elhemely 2022 [116]	4O2B	Docking of molecules at the colchicine site using an α,β-tubulin dimer. MD was used to study interactions and validate ligand persistence in binding site and SAR studies.	AMBER 19	Tubulin: ff14SB Ligand: antechamber GAFF2	TIP3P	50 ns
Dash 2022 [119]	1SA0	Docking of molecules in the αβ-tubulin interface using a tubulin dimer. MD was used to study interactions, validate ligand persistence at the binding site, and calculate binding free energies.	AMBER 16	Tubulin: ff14SB Ligand: GAFF	TIP3P	100 ns
Hadizadeh 2022 [111]	4O2B	Docking of molecules at the colchicine site. MD was used to study interactions and validate ligand persistence at the binding site.	NAMD 2.12	Tubulin: CHARMM27 Ligand: provided by SwissParam	TIP3	100 ns
Mao 2022 [90]	1JFF	Docking of molecules in the taxane site using a monomer of β-tubulin. MD was used to study interactions, validate ligand persistence at the binding site, and calculate binding free energies.	GROMACS 2019.1	Tubulin: Amber99sb-ildn Ligand: ACPYPE	TIP3P	80 ns
Neto 2022 [118]	4O2B	Docking of chalcones in the colchicine site. MD was used to study interactions, validate ligand persistence at the binding site, and calculate binding free energies.	Discovery Studio software		implicit	1000 ns

Table A6. *Cont.*

Reference	PDB	Object of Study	MD Engine	Force field	Water Model	Time
Pragyandipta 2022 [126]	6Y6D	Docking of molecules in the noscapinoids site. MD was used to study interactions, validate ligand persistence at the binding site, and calculate binding free energies.	GROMACS 2019.2	Tubulin: GROMOS96 Ligand: ACPYPE	TIP3P	100 ns
Yang 2022 [127]	1JFF 4O4H	Study of wangzaozin as a binder for the taxane and laulimalide sites.	GROMACS 2019.1	Tubulin: Amber99sb-ildn Ligand: ACPYPE	TIP3P	90 ns
Boichuk 2022 [108]	4O2B	Assess the position of the ligand at the colchicine binding site and determine key amino acid interactions using the EAPC-67-tubulin complex.	Desmond in Schrödinger suite 2021-2		SPC	100 ns
Basu 2022 [121]	1JFF 1TUB	Comparison of apo α,β-tubulin dimer, bound to taxol, and bound to Taxotere.	NAMD 2.11	Tubulin: CHARMM36 Ligand: CGenFF	TIP3P	200 ns
El-Mernissi 2022 [112]	3E22	3E22-colchicine in complex with tubulin and two selected tubulin compound complexes to examine protein-ligand interactions.	Desmond Dynamics	OPLS		50 ns
Zhang 2022 [115]	1JFF	Docking validation of hits bound to the taxane site	GROMACS 2019.1	Tubulin: Amber99sb-ildn	SPC216	90 ns
Zhao 2022 [115]	4O2B	Docking validation of styrylquinoline tubulin inhibitors	AMBER16	Tubulin: Amber ff99SB Ligand: GAFF	TIP3P	100 ns
Radha 2022 [123]	6Y6D	Docking validation of shikonin as a tubulin inhibitor	GROMACS 2019.2	Tubulin: Amber ff99SB Ligand: GAFF	TIP3P	100 ns
Rai 2022 [107]		MD used for the analysis of the Interactions between eribulin and different tubulin isotypes	AMBER 12	Tubulin: Amber ff99SB Ligand: Antechamber tool	implicit	60 ns

Note: the majority of these simulations were performed at a temperature of ~300 K, a pressure of 1 bar, in Periodic Boundary Conditions (PBC) at a constant temperature and pressure (NPT ensemble).

References

1. Pellegrini, L.; Wetzel, A.; Granno, S.; Heaton, G.; Harvey, K. Back to the tubule: Microtubule dynamics in Parkinson's disease. *Cell. Mol. Life Sci.* **2017**, *74*, 409–434. [CrossRef]
2. Čermák, V.; Dostál, V.; Jelínek, M.; Libusová, L.; Kovář, J.; Rösel, D.; Brábek, J. Microtubule-targeting agents and their impact on cancer treatment. *Eur. J. Cell Biol.* **2020**, *99*, 151075. [CrossRef] [PubMed]
3. Bracey, K.M.; Ho, K.H.; Yampolsky, D.; Gu, G.; Kaverina, I.; Holmes, W.R. Microtubules Regulate Localization and Availability of Insulin Granules in Pancreatic Beta Cells. *Biophys. J.* **2020**, *118*, 193–206. [CrossRef]
4. Dubey, J.; Ratnakaran, N.; Koushika, S.P. Neurodegeneration and microtubule dynamics: Death by a thousand cuts. *Front. Cell. Neurosci.* **2015**, *9*, 343. [CrossRef]
5. Mitchison, T.; Kirschner, M. Dynamic instability of microtubule growth. *Nature* **1984**, *312*, 237–242. [CrossRef]
6. Kollman, J.M.; Merdes, A.; Mourey, L.; Agard, D.A. Microtubule nucleation by gamma-tubulin complexes. *Nat. Rev. Mol. Cell Biol.* **2011**, *12*, 709–721. [CrossRef] [PubMed]
7. Teixido-Travesa, N.; Roig, J.; Luders, J. The where, when and how of microtubule nucleation-one ring to rule them all. *J. Cell Sci.* **2012**, *125*, 4445–4456. [CrossRef] [PubMed]
8. Liu, P.; Wurtz, M.; Zupa, E.; Pfeffer, S.; Schiebel, E. Microtubule nucleation: The waltz between gamma-tubulin ring complex and associated proteins. *Curr. Opin. Cell Biol.* **2021**, *68*, 124–131. [CrossRef]
9. Brouhard, G.J.; Rice, L.M. Microtubule dynamics: An interplay of biochemistry and mechanics. *Nat. Rev. Mol. Cell Biol.* **2018**, *19*, 451–463. [CrossRef]
10. Roostalu, J.; Thomas, C.; Cade, N.I.; Kunzelmann, S.; Taylor, I.A.; Surrey, T. The speed of GTP hydrolysis determines GTP cap size and controls microtubule stability. *Elife* **2020**, *9*, e51992. [CrossRef]
11. Mühlethaler, T.; Gioia, D.; Prota, A.E.; Sharpe, M.E.; Cavalli, A.; Steinmetz, M.O. Comprehensive Analysis of Binding Sites in Tubulin. *Angew. Chem. Int. Ed. Engl.* **2021**, *60*, 13331–13342. [CrossRef] [PubMed]
12. Muhlethaler, T.; Milanos, L.; Ortega, J.A.; Blum, T.B.; Gioia, D.; Roy, B.; Prota, A.E.; Cavalli, A.; Steinmetz, M.O. Rational Design of a Novel Tubulin Inhibitor with a Unique Mechanism of Action. *Angew. Chem. Int. Ed. Engl.* **2022**, *61*, e202204052. [CrossRef] [PubMed]
13. Marzaro, G.; Chilin, A. QSAR and 3D-QSAR models in the field of tubulin inhibitors as anticancer agents. *Curr. Top. Med. Chem.* **2014**, *14*, 2253–2262. [CrossRef]
14. Johnson, M.A.; Maggiora, G.M. *Concepts and Applications of Molecular Similarity*; Wiley: Hoboken, NJ, USA, 1990.
15. Horvath, D.; Koch, C.; Schneider, G.; Marcou, G.; Varnek, A. Local neighborhood behavior in a combinatorial library context. *J. Comput. Aided. Mol. Des.* **2011**, *25*, 237–252. [CrossRef]
16. Ayoub, A.T.; Klobukowski, M.; Tuszynski, J. Similarity-based virtual screening for microtubule stabilizers reveals novel antimitotic scaffold. *J. Mol. Graph. Model.* **2013**, *44*, 188–196. [CrossRef] [PubMed]
17. Guo, Q.; Luo, Y.; Zhai, S.; Jiang, Z.; Zhao, C.; Xu, J.; Wang, L. Discovery, biological evaluation, structure-activity relationships and mechanism of action of pyrazolo[3,4-b]pyridin-6-one derivatives as a new class of anticancer agents. *Org. Biomol. Chem.* **2019**, *17*, 6201–6214. [CrossRef] [PubMed]
18. Mangiatordi, G.F.; Trisciuzzi, D.; Alberga, D.; Denora, N.; Iacobazzi, R.M.; Gadaleta, D.; Catto, M.; Nicolotti, O. Novel chemotypes targeting tubulin at the colchicine binding site and unbiasing P-glycoprotein. *Eur. J. Med. Chem.* **2017**, *139*, 792–803. [CrossRef] [PubMed]
19. Federico, L.B.; Silva, G.M.; de Fraga Dias, A.; Figueiro, F.; Battastini, A.M.O.; Dos Santos, C.B.R.; Costa, L.T.; Rosa, J.M.C.; de Paula da Silva, C.H.T. Identification of novel alphabeta-tubulin modulators with antiproliferative activity directed to cancer therapy using ligand and structure-based virtual screening. *Int. J. Biol. Macromol.* **2020**, *165*, 3040–3050. [CrossRef]
20. Lo, Y.C.; Senese, S.; Li, C.M.; Hu, Q.; Huang, Y.; Damoiseaux, R.; Torres, J.Z. Large-scale chemical similarity networks for target profiling of compounds identified in cell-based chemical screens. *PLoS Comput. Biol.* **2015**, *11*, e1004153. [CrossRef]
21. Lo, Y.C.; Senese, S.; Damoiseaux, R.; Torres, J.Z. 3D Chemical Similarity Networks for Structure-Based Target Prediction and Scaffold Hopping. *ACS Chem. Biol.* **2016**, *11*, 2244–2253. [CrossRef]
22. Muratov, E.N.; Bajorath, J.; Sheridan, R.P.; Tetko, I.V.; Filimonov, D.; Poroikov, V.; Oprea, T.I.; Baskin, I.I.; Varnek, A.; Roitberg, A.; et al. QSAR without borders. *Chem. Soc. Rev.* **2020**, *49*, 3525–3564. [CrossRef]
23. Tropsha, A.; Gramatica, P.; Gombar, V.K. The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *Qsar Comb. Sci.* **2003**, *22*, 69–77. [CrossRef]
24. Golbraikh, A.; Tropsha, A. Beware of q2! *J. Mol. Graph. Model.* **2002**, *20*, 269–276. [CrossRef] [PubMed]
25. Gaikwad, R.; Amin, S.A.; Adhikari, N.; Ghorai, S.; Jha, T.; Gayen, S. Identification of molecular fingerprints of phenylindole derivatives as cytotoxic agents: A multi-QSAR approach. *Struct. Chem.* **2018**, *29*, 1095–1107. [CrossRef]
26. Guo, Q.; Zhang, H.; Deng, Y.; Zhai, S.; Jiang, Z.; Zhu, D.; Wang, L. Ligand- and structural-based discovery of potential small molecules that target the colchicine site of tubulin for cancer treatment. *Eur. J. Med. Chem.* **2020**, *196*, 112328. [CrossRef]
27. Stefanski, T.; Mikstacka, R.; Kurczab, R.; Dutkiewicz, Z.; Kucinska, M.; Murias, M.; Zielinska-Przyjemska, M.; Cichocki, M.; Teubert, A.; Kaczmarek, M.; et al. Design, synthesis, and biological evaluation of novel combretastatin A-4 thio derivatives as microtubule targeting agents. *Eur. J. Med. Chem.* **2018**, *144*, 797–816. [CrossRef] [PubMed]
28. Quan, Y.P.; Cheng, L.P.; Wang, T.C.; Pang, W.; Wu, F.H.; Huang, J.W. Molecular modeling study, synthesis and biological evaluation of combretastatin A-4 analogues as anticancer agents and tubulin inhibitors. *Medchemcomm* **2018**, *9*, 316–327. [CrossRef] [PubMed]

29. Pandit, A.; Yadav, K.; Reddy, R.B.; Sengupta, S.; Sharma, R.; Chelvam, V. Structure activity relationships (SAR) study to design and synthesize new tubulin inhibitors with enhanced anti-tubulin activity: In silico and in vitro analysis. *J. Mol. Struct.* **2021**, *1223*, 129204. [CrossRef]
30. Giordano, D.; Biancaniello, C.; Argenio, M.A.; Facchiano, A. Drug Design by Pharmacophore and Virtual Screening Approach. *Pharmaceuticals* **2022**, *15*, 646. [CrossRef]
31. Seidel, T.; Wieder, O.; Garon, A.; Langer, T. Applications of the Pharmacophore Concept in Natural Product inspired Drug Design. *Mol. Inform.* **2020**, *39*, e2000059. [CrossRef]
32. Zhang, H.; Mao, J.; Yang, Y.L.; Liu, C.T.; Shen, C.; Zhang, H.R.; Xie, H.Z.; Ding, L. Discovery of novel tubulin inhibitors targeting taxanes site by virtual screening, molecular dynamic simulation, and biological evaluation. *J. Cell. Biochem.* **2021**, *122*, 1609–1624. [CrossRef] [PubMed]
33. Lone, M.Y.; Athar, M.; Manhas, A.; Jha, P.C.; Bhatt, S.; Shah, A. In Silico Exploration of Vinca Domain Tubulin Inhibitors: A Combination of 3D-QSAR-Based Pharmacophore Modeling, Docking and Molecular Dynamics Simulations. *ChemistrySelect* **2017**, *2*, 10848–10853. [CrossRef]
34. Niu, M.M.; Qin, J.Y.; Tian, C.P.; Yan, X.F.; Dong, F.G.; Cheng, Z.Q.; Fida, G.; Yang, M.; Chen, H.Y.; Gu, Y.Q. Tubulin inhibitors: Pharmacophore modeling, virtual screening and molecular docking. *Acta Pharmacol. Sin.* **2014**, *35*, 967–979. [CrossRef] [PubMed]
35. Lionta, E.; Spyrou, G.; Vassilatis, D.K.; Cournia, Z. Structure-based virtual screening for drug discovery: Principles, applications and recent advances. *Curr. Top. Med. Chem.* **2014**, *14*, 1923–1938. [CrossRef] [PubMed]
36. Ferreira, L.G.; Dos Santos, R.N.; Oliva, G.; Andricopulo, A.D. Molecular docking and structure-based drug design strategies. *Molecules* **2015**, *20*, 13384–13421. [CrossRef] [PubMed]
37. Campanacci, V.; Urvoas, A.; Consolati, T.; Cantos-Fernandes, S.; Aumont-Nicaise, M.; Valerio-Lepiniec, M.; Surrey, T.; Minard, P.; Gigant, B. Selection and Characterization of Artificial Proteins Targeting the Tubulin alpha Subunit. *Structure* **2019**, *27*, 497–506.e494. [CrossRef]
38. Curmi, P.A.; Andersen, S.S.; Lachkar, S.; Gavet, O.; Karsenti, E.; Knossow, M.; Sobel, A. The stathmin/tubulin interaction in vitro. *J. Biol. Chem.* **1997**, *272*, 25029–25036. [CrossRef]
39. Steinmetz, M.O.; Kammerer, R.A.; Jahnke, W.; Goldie, K.N.; Lustig, A.; van Oostrum, J. Op18/stathmin caps a kinked protofilament-like tubulin tetramer. *EMBO J.* **2000**, *19*, 572–580. [CrossRef]
40. Nogales, E.; Wolf, S.G.; Downing, K.H. Structure of the alpha beta tubulin dimer by electron crystallography. *Nature* **1998**, *391*, 199–203. [CrossRef]
41. Gigant, B.; Curmi, P.A.; Martin-Barbey, C.; Charbaut, E.; Lachkar, S.; Lebeau, L.; Siavoshian, S.; Sobel, A.; Knossow, M. The 4 angstrom X-ray structure of a tubulin: Stathmin-like domain complex. *Cell* **2000**, *102*, 809–816. [CrossRef] [PubMed]
42. Ravelli, R.B.G.; Gigant, B.; Curmi, P.A.; Jourdain, I.; Lachkar, S.; Sobel, A.; Knossow, M. Insight into tubulin regulation from a complex with colchicine and a stathmin-like domain. *Nature* **2004**, *428*, 198–202. [CrossRef]
43. Prota, A.E.; Magiera, M.M.; Kuijpers, M.; Bargsten, K.; Frey, D.; Wieser, M.; Jaussi, R.; Hoogenraad, C.C.; Kammerer, R.A.; Janke, C.; et al. Structural basis of tubulin tyrosination by tubulin tyrosine ligase. *J. Cell Biol.* **2013**, *200*, 259–270. [CrossRef]
44. Prota, A.E.; Bargsten, K.; Zurwerra, D.; Field, J.J.; Díaz, J.F.; Altmann, K.-H.; Steinmetz, M.O. Molecular Mechanism of Action of Microtubule-Stabilizing Anticancer Agents. *Science* **2013**, *339*, 587–590. [CrossRef]
45. Pecqueur, L.; Duellberg, C.; Dreier, B.; Jiang, Q.; Wang, C.; Pluckthun, A.; Surrey, T.; Gigant, B.; Knossow, M. A designed ankyrin repeat protein selected to bind to tubulin caps the microtubule plus end. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 12011–12016. [CrossRef]
46. La Sala, G.; Olieric, N.; Sharma, A.; Viti, F.; Perez, F.D.B.; Huang, L.; Tonra, J.R.; Lloyd, G.K.; Decherchi, S.; Diaz, J.F.; et al. Structure, Thermodynamics, and Kinetics of Plinabulin Binding to Two Tubulin Isotypes. *Chem* **2019**, *5*, 2969–2986. [CrossRef]
47. Ayaz, P.; Ye, X.; Huddleston, P.; Brautigam, C.A.; Rice, L.M. A TOG:alphabeta-tubulin complex structure reveals conformation-based mechanisms for a microtubule polymerase. *Science* **2012**, *337*, 857–860. [CrossRef]
48. Campanacci, V.; Urvoas, A.; Ammar Khodja, L.; Aumont-Nicaise, M.; Noiray, M.; Lachkar, S.; Curmi, P.A.; Minard, P.; Gigant, B. Structural convergence for tubulin binding of CPAP and vinca domain microtubule inhibitors. *Proc. Natl. Acad. Sci. USA* **2022**, *119*, e2120098119. [CrossRef]
49. Sharma, A.; Aher, A.; Dynes, N.J.; Frey, D.; Katrukha, E.A.; Jaussi, R.; Grigoriev, I.; Croisier, M.; Kammerer, R.A.; Akhmanova, A.; et al. Centriolar CPAP/SAS-4 Imparts Slow Processive Microtubule Growth. *Dev. Cell* **2016**, *37*, 362–376. [CrossRef]
50. Zheng, X.; Ramani, A.; Soni, K.; Gottardo, M.; Zheng, S.; Ming Gooi, L.; Li, W.; Feng, S.; Mariappan, A.; Wason, A.; et al. Molecular basis for CPAP-tubulin interaction in controlling centriolar and ciliary length. *Nat. Commun.* **2016**, *7*, 11874. [CrossRef]
51. Steinmetz, M.O.; Prota, A.E. Microtubule-Targeting Agents: Strategies To Hijack the Cytoskeleton. *Trends Cell Biol.* **2018**, *28*, 776–792. [CrossRef]
52. Elie-Caille, C.; Severin, F.; Helenius, J.; Howard, J.; Muller, D.J.; Hyman, A.A. Straight GDP-tubulin protofilaments form in the presence of taxol. *Curr. Biol.* **2007**, *17*, 1765–1770. [CrossRef] [PubMed]
53. Alushin, G.M.; Lander, G.C.; Kellogg, E.H.; Zhang, R.; Baker, D.; Nogales, E. High-resolution microtubule structures reveal the structural transitions in alphabeta-tubulin upon GTP hydrolysis. *Cell* **2014**, *157*, 1117–1129. [CrossRef]
54. Lowe, J.; Li, H.; Downing, K.H.; Nogales, E. Refined structure of alpha beta-tubulin at 3.5 A resolution. *J. Mol. Biol.* **2001**, *313*, 1045–1057. [CrossRef]

55. Kellogg, E.H.; Hejab, N.M.A.; Howes, S.; Northcote, P.; Miller, J.H.; Diaz, J.F.; Downing, K.H.; Nogales, E. Insights into the Distinct Mechanisms of Action of Taxane and Non-Taxane Microtubule Stabilizers from Cryo-EM Structures. *J. Mol. Biol.* **2017**, *429*, 633–646. [CrossRef]
56. Prota, A.E.; Bargsten, K.; Northcote, P.T.; Marsh, M.; Altmann, K.H.; Miller, J.H.; Diaz, J.F.; Steinmetz, M.O. Structural basis of microtubule stabilization by laulimalide and peloruside A. *Angew. Chem. Int. Ed. Engl.* **2014**, *53*, 1621–1625. [CrossRef]
57. Dorleans, A.; Gigant, B.; Ravelli, R.B.G.; Mailliet, P.; Mikol, V.; Knossow, M. Variations in the colchicine-binding domain provide insight into the structural switch of tubulin. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 13775–13779. [CrossRef] [PubMed]
58. Gigant, B.; Wang, C.; Ravelli, R.B.; Roussi, F.; Steinmetz, M.O.; Curmi, P.A.; Sobel, A.; Knossow, M. Structural basis for the regulation of tubulin by vinblastine. *Nature* **2005**, *435*, 519–522. [CrossRef]
59. Cormier, A.; Marchand, M.; Ravelli, R.B.; Knossow, M.; Gigant, B. Structural insight into the inhibition of tubulin by vinca domain peptide ligands. *EMBO Rep.* **2008**, *9*, 1101–1106. [CrossRef]
60. Maderna, A.; Doroski, M.; Subramanyam, C.; Porte, A.; Leverett, C.A.; Vetelino, B.C.; Chen, Z.; Risley, H.; Parris, K.; Pandit, J.; et al. Discovery of cytotoxic dolastatin 10 analogues with N-terminal modifications. *J. Med. Chem.* **2014**, *57*, 10527–10543. [CrossRef] [PubMed]
61. Prota, A.E.; Bargsten, K.; Diaz, J.F.; Marsh, M.; Cuevas, C.; Liniger, M.; Neuhaus, C.; Andreu, J.M.; Altmann, K.H.; Steinmetz, M.O. A new tubulin-binding site and pharmacophore for microtubule-destabilizing anticancer drugs. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 13817–13821. [CrossRef]
62. Prota, A.E.; Setter, J.; Waight, A.B.; Bargsten, K.; Murga, J.; Diaz, J.F.; Steinmetz, M.O. Pironetin Binds Covalently to alphaCys316 and Perturbs a Major Loop and Helix of alpha-Tubulin to Inhibit Microtubule Formation. *J. Mol. Biol.* **2016**, *428*, 2981–2988. [CrossRef]
63. Yang, J.; Wang, Y.; Wang, T.; Jiang, J.; Botting, C.H.; Liu, H.; Chen, Q.; Yang, J.; Naismith, J.H.; Zhu, X.; et al. Pironetin reacts covalently with cysteine-316 of alpha-tubulin to destabilize microtubule. *Nat. Commun.* **2016**, *7*, 12103. [CrossRef]
64. Matthew, S.; Chen, Q.Y.; Ratnayake, R.; Fermaintt, C.S.; Lucena-Agell, D.; Bonato, F.; Prota, A.E.; Lim, S.T.; Wang, X.; Diaz, J.F.; et al. Gatorbulin-1, a distinct cyclodepsipeptide chemotype, targets a seventh tubulin pharmacological site. *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2021847118. [CrossRef]
65. Prota, A.E.; Danel, F.; Bachmann, F.; Bargsten, K.; Buey, R.M.; Pohlmann, J.; Reinelt, S.; Lane, H.; Steinmetz, M.O. The novel microtubule-destabilizing drug BAL27862 binds to the colchicine site of tubulin with distinct effects on microtubule organization. *J. Mol. Biol.* **2014**, *426*, 1848–1860. [CrossRef] [PubMed]
66. Oliva, M.A.; Prota, A.E.; Rodríguez-Salarichs, J.; Bennani, Y.L.; Jiménez-Barbero, J.; Bargsten, K.; Canales, Á.; Steinmetz, M.O.; Díaz, J.F. Structural Basis of Noscapine Activation for Tubulin Binding. *J. Med. Chem.* **2020**, *63*, 8495–8501. [CrossRef] [PubMed]
67. Doodhi, H.; Prota, A.E.; Rodriguez-Garcia, R.; Xiao, H.; Custar, D.W.; Bargsten, K.; Katrukha, E.A.; Hilbert, M.; Hua, S.; Jiang, K.; et al. Termination of Protofilament Elongation by Eribulin Induces Lattice Defects that Promote Microtubule Catastrophes. *Curr. Biol.* **2016**, *26*, 1713–1721. [CrossRef]
68. Weinert, T.; Olieric, N.; Cheng, R.; Brunle, S.; James, D.; Ozerov, D.; Gashi, D.; Vera, L.; Marsh, M.; Jaeger, K.; et al. Serial millisecond crystallography for routine room-temperature structure determination at synchrotrons. *Nat. Commun.* **2017**, *8*, 542. [CrossRef] [PubMed]
69. Nawrotek, A.; Knossow, M.; Gigant, B. The determinants that govern microtubule assembly from the atomic structure of GTP-tubulin. *J. Mol. Biol.* **2011**, *412*, 35–42. [CrossRef] [PubMed]
70. Prota, A.E.; Bargsten, K.; Redondo-Horcajo, M.; Smith, A.B., III; Yang, C.H.; McDaid, H.M.; Paterson, I.; Horwitz, S.B.; Fernando Diaz, J.; Steinmetz, M.O. Structural Basis of Microtubule Stabilization by Discodermolide. *Chembiochem* **2017**, *18*, 905–909. [CrossRef]
71. Guo, B.; Rodriguez-Gabin, A.; Prota, A.E.; Muhlethaler, T.; Zhang, N.; Ye, K.; Steinmetz, M.O.; Horwitz, S.B.; Smith, A.B., III; McDaid, H.M. Structural Refinement of the Tubulin Ligand (+)-Discodermolide to Attenuate Chemotherapy-Mediated Senescence. *Mol. Pharmacol.* **2020**, *98*, 156–167. [CrossRef] [PubMed]
72. Menchon, G.; Prota, A.E.; Lucena-Agell, D.; Bucher, P.; Jansen, R.; Irschik, H.; Muller, R.; Paterson, I.; Diaz, J.F.; Altmann, K.H.; et al. A fluorescence anisotropy assay to discover and characterize ligands targeting the maytansine site of tubulin. *Nat. Commun.* **2018**, *9*, 2106. [CrossRef]
73. Gao, L.; Meiring, J.C.M.; Kraus, Y.; Wranik, M.; Weinert, T.; Pritzl, S.D.; Bingham, R.; Ntouliou, E.; Jansen, K.I.; Olieric, N.; et al. A Robust, GFP-Orthogonal Photoswitchable Inhibitor Scaffold Extends Optical Control over the Microtubule Cytoskeleton. *Cell Chem. Biol.* **2021**, *28*, 228–241.e226. [CrossRef] [PubMed]
74. De la Roche, N.M.; Muhlethaler, T.; Di Martino, R.M.C.; Ortega, J.A.; Gioia, D.; Roy, B.; Prota, A.E.; Steinmetz, M.O.; Cavalli, A. Novel fragment-derived colchicine-site binders as microtubule-destabilizing agents. *Eur. J. Med. Chem.* **2022**, *241*, 114614. [CrossRef] [PubMed]
75. Bohnacker, T.; Prota, A.E.; Beaufils, F.; Burke, J.E.; Melone, A.; Inglis, A.J.; Rageot, D.; Sele, A.M.; Cmiljanovic, V.; Cmiljanovic, N.; et al. Deconvolution of Buparlisib's mechanism of action defines specific PI3K and tubulin inhibitors for therapeutic intervention. *Nat. Commun.* **2017**, *8*, 14683. [CrossRef] [PubMed]
76. Stein, A.; Hilken Nee Thomopoulou, P.; Frias, C.; Hopff, S.M.; Varela, P.; Wilke, N.; Mariappan, A.; Neudorfl, J.M.; Fedorov, A.Y.; Gopalakrishnan, J.; et al. B-nor-methylene Colchicinoid PT-100 Selectively Induces Apoptosis in Multidrug-Resistant Human Cancer Cells via an Intrinsic Pathway in a Caspase-Independent Manner. *ACS Omega* **2022**, *7*, 2591–2603. [CrossRef] [PubMed]

77. Waight, A.B.; Bargsten, K.; Doronina, S.; Steinmetz, M.O.; Sussman, D.; Prota, A.E. Structural Basis of Microtubule Destabilization by Potent Auristatin Anti-Mitotics. *PLoS ONE* **2016**, *11*, e0160890. [CrossRef]
78. Debs, G.E.; Cha, M.; Liu, X.; Huehn, A.R.; Sindelar, C.V. Dynamic and asymmetric fluctuations in the microtubule wall captured by high-resolution cryoelectron microscopy. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 16976–16984. [CrossRef]
79. LaFrance, B.J.; Roostalu, J.; Henkin, G.; Greber, B.J.; Zhang, R.; Normanno, D.; McCollum, C.O.; Surrey, T.; Nogales, E. Structural transitions in the GTP cap visualized by cryo-electron microscopy of catalytically inactive microtubules. *Proc. Natl. Acad. Sci. USA* **2022**, *119*, e2114994119. [CrossRef]
80. Castro-Alvarez, A.; Pineda, O.; Vilarrasa, J. Further Insight into the Interactions of the Cytotoxic Macrolides Laulimalide and Peloruside A with Their Common Binding Site. *ACS Omega* **2018**, *3*, 1770–1782. [CrossRef]
81. Gaurav, A.; Gautam, V. Structure-based three-dimensional pharmacophores as an alternative to traditional methodologies. *J. Recept. Ligand Channel Res.* **2014**, *2014*, 27–38. [CrossRef]
82. Nagarajan, S.; Choi, M.J.; Cho, Y.S.; Min, S.J.; Keum, G.; Kim, S.J.; Lee, C.S.; Pae, A.N. Tubulin inhibitor identification by bioactive conformation alignment pharmacophore-guided virtual screening. *Chem. Biol. Drug Des.* **2015**, *86*, 998–1016. [CrossRef] [PubMed]
83. Zhou, Y.; Di, B.; Niu, M.M. Structure-Based Pharmacophore Design and Virtual Screening for Novel Tubulin Inhibitors with Potential Anticancer Activity. *Molecules* **2019**, *24*, 3181. [CrossRef] [PubMed]
84. Gallego-Yerga, L.; Ochoa, R.; Lans, I.; Pena-Varas, C.; Alegria-Arcos, M.; Cossio, P.; Ramirez, D.; Pelaez, R. Application of ensemble pharmacophore-based virtual screening to the discovery of novel antimitotic tubulin inhibitors. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 4360–4372. [CrossRef]
85. Elseginy, S.A.; Oliveira, A.S.F.; Shoemark, D.K.; Sessions, R.B. Identification and validation of novel microtubule suppressors with an imidazopyridine scaffold through structure-based virtual screening and docking. *RSC Med. Chem.* **2022**, *13*, 929–943. [CrossRef]
86. Sulimov, V.B.; Kutov, D.C.; Sulimov, A.V. Advances in Docking. *Curr. Med. Chem.* **2019**, *26*, 7555–7580. [CrossRef]
87. Halperin, I.; Ma, B.; Wolfson, H.; Nussinov, R. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins* **2002**, *47*, 409–443. [CrossRef]
88. Sabe, V.T.; Ntombela, T.; Jhamba, L.A.; Maguire, G.E.M.; Govender, T.; Naicker, T.; Kruger, H.G. Current trends in computer aided drug design and a highlight of drugs discovered via computational techniques: A review. *Eur. J. Med. Chem.* **2021**, *224*, 113705. [CrossRef]
89. Maia, E.H.B.; Assis, L.C.; de Oliveira, T.A.; da Silva, A.M.; Taranto, A.G. Structure-Based Virtual Screening: From Classical to Artificial Intelligence. *Front. Chem.* **2020**, *8*, 343. [CrossRef]
90. Mao, J.; Luo, Q.Q.; Zhang, H.R.; Zheng, X.H.; Shen, C.; Qi, H.Z.; Hu, M.L.; Zhang, H. Discovery of microtubule stabilizers with novel scaffold structures based on virtual screening, biological evaluation, and molecular dynamics simulation. *Chem. Biol. Interact.* **2022**, *352*, 109784. [CrossRef]
91. Zuniga-Bustos, M.; Vasquez, P.A.; Jana, G.A.; Guzman, J.L.; Alderete, J.B.; Jimenez, V.A. Mechanism-Based Rational Discovery and In Vitro Evaluation of Novel Microtubule Stabilizing Agents with Non-Taxol-Competitive Activity. *J. Chem. Inf. Model.* **2020**, *60*, 3204–3213. [CrossRef]
92. Liu, W.; Jia, H.; Guan, M.; Cui, M.; Lan, Z.; He, Y.; Guo, Z.; Jiang, R.; Dong, G.; Wang, S. Discovery of novel tubulin inhibitors targeting the colchicine binding site via virtual screening, structural optimization and antitumor evaluation. *Bioorganic Chem.* **2022**, *118*, 105486. [CrossRef]
93. Liu, G.; Jiao, Y.; Huang, C.; Chang, P. Identification of novel and potent small-molecule inhibitors of tubulin with antitumor activities by virtual screening and biological evaluations. *J. Comput. Aided. Mol. Des.* **2019**, *33*, 659–664. [CrossRef]
94. Ameri, A.; Khodarahmi, G.; Forootanfar, H.; Hassanzadeh, F.; Hakimelahi, G.H. Hybrid Pharmacophore Design, Molecular Docking, Synthesis, and Biological Evaluation of Novel Aldimine-Type Schiff Base Derivatives as Tubulin Polymerization Inhibitor. *Chem. Biodivers.* **2018**, *15*, e1700518. [CrossRef]
95. Riu, F.; Ibba, R.; Zoroddu, S.; Sestito, S.; Lai, M.; Piras, S.; Sanna, L.; Bordoni, V.; Bagella, L.; Carta, A. Design, synthesis, and biological screening of a series of 4'-fluoro-benzotriazole-acrylonitrile derivatives as microtubule-destabilising agents (MDAs). *J. Enzyme. Inhib. Med. Chem.* **2022**, *37*, 2223–2240. [CrossRef]
96. Patel, A.K.; Meher, R.K.; Nagireddy, P.K.; Pragyandipta, P.; Pedapati, R.K.; Kantevari, S.; Naik, P.K. 9-Arylimino noscapinoids as potent tubulin binding anticancer agent: Chemical synthesis and cellular evaluation against breast tumour cells. *SAR QSAR Environ. Res.* **2021**, *32*, 269–291. [CrossRef]
97. Mustafa, M.; Abdelhamid, D.; Abdelhafez, E.M.N.; Ibrahim, M.A.A.; Gamal-Eldeen, A.M.; Aly, O.M. Synthesis, antiproliferative, anti-tubulin activity, and docking study of new 1,2,4-triazoles as potential combretastatin analogues. *Eur. J. Med. Chem.* **2017**, *141*, 293–305. [CrossRef]
98. Tripathi, S.; Srivastava, G.; Sharma, A. Molecular dynamics simulation and free energy landscape methods in probing L215H, L217R and L225M betaI-tubulin mutations causing paclitaxel resistance in cancer cells. *Biochem. Biophys. Res. Commun.* **2016**, *476*, 273–279. [CrossRef]
99. Ayoub, A.T.; Elrefaiy, M.A.; Arakawa, K. Computational Prediction of the Mode of Binding of Antitumor Lankacidin C to Tubulin. *ACS Omega* **2019**, *4*, 4461–4471. [CrossRef]

100. Chávez-Estrada, E.J.; Cerda-García-Rojas, C.M.; Román-Marín, L.U.; Hernández-Hernández, J.D.; Joseph-Nathan, P. Synthesis, molecular docking, and saturation-transfer difference NMR spectroscopy of longipinane derivatives as novel microtubule stabilizers. *J. Mol. Struct.* **2020**, *1218*, 128519. [CrossRef]
101. Forero, A.M.; Castellanos, L.; Sandoval-Hernandez, A.G.; Magalhaes, A.; Tinoco, L.W.; Lopez-Vallejo, F.; Ramos, F.A. Integration of NMR studies, computational predictions, and in vitro assays in the search of marine diterpenes with antitumor activity. *Chem. Biol. Drug Des.* **2021**, *98*, 507–521. [CrossRef]
102. Ngo, S.T.; Vu, K.B.; Bui, L.M.; Vu, V.V. Effective Estimation of Ligand-Binding Affinity Using Biased Sampling Method. *ACS Omega* **2019**, *4*, 3887–3893. [CrossRef]
103. Zhang, Z.; Lu, C.; Wang, P.; Li, A.; Zhang, H.; Xu, S. Structural Basis and Mechanism for Vindoline Dimers Interacting with α,β-Tubulin. *ACS Omega* **2019**, *4*, 11938–11948. [CrossRef]
104. Zhou, X.; Xu, Z.; Li, A.; Zhang, Z.; Xu, S. Double-sides sticking mechanism of vinblastine interacting with α,β-tubulin to get activity against cancer cells. *J. Biomol. Struct. Dyn.* **2018**, *37*, 4080–4091. [CrossRef]
105. Mane, J.Y.; Semenchenko, V.; Perez-Pineiro, R.; Winter, P.; Wishart, D.; Tuszynski, J.A. Experimental and computational study of the interaction of novel colchicinoids with a recombinant human alphaI/betaI-tubulin heterodimer. *Chem. Biol. Drug Des.* **2013**, *82*, 60–70. [CrossRef]
106. Izrailev, S.; Stepaniants, S.; Balsera, M.; Oono, Y.; Schulten, K. Molecular dynamics study of unbinding of the avidin-biotin complex. *Biophys. J.* **1997**, *72*, 1568–1581. [CrossRef]
107. Rai, K.; Kumbhar, B.V.; Panda, D.; Kunwar, A. Computational study of interactions of anti-cancer drug eribulin with human tubulin isotypes. *Phys. Chem. Chem. Phys.* **2022**, *24*, 16694–16700. [CrossRef]
108. Boichuk, S.; Syuzov, K.; Bikinieva, F.; Galembikova, A.; Zykova, S.; Gankova, K.; Igidov, S.; Igidov, N. Computational-Based Discovery of the Anti-Cancer Activities of Pyrrole-Based Compounds Targeting the Colchicine-Binding Site of Tubulin. *Molecules* **2022**, *27*, 2873. [CrossRef]
109. Fusani, L.; Palmer, D.S.; Somers, D.O.; Wall, I.D. Exploring Ligand Stability in Protein Crystal Structures Using Binding Pose Metadynamics. *J. Chem. Inf. Model.* **2020**, *60*, 1528–1539. [CrossRef]
110. Gaspari, R.; Prota, A.E.; Bargsten, K.; Cavalli, A.; Steinmetz, M.O. Structural Basis of cis- and trans-Combretastatin Binding to Tubulin. *Chem* **2017**, *2*, 102–113. [CrossRef]
111. Hadizadeh, F.; Ghodsi, R.; Mirzaei, S.; Sahebkar, A. In Silico Exploration of Novel Tubulin Inhibitors: A Combination of Docking and Molecular Dynamics Simulations, Pharmacophore Modeling, and Virtual Screening. *Comput. Math. Methods Med.* **2022**, *2022*, 4004068. [CrossRef]
112. El-Mernissi, R.; El Khatabi, K.; Khaldan, A.; ElMchichi, L.; Shahinozzaman, M.; Ajana, M.A.; Lakhlifi, T.; Bouachrine, M. 2-Oxoquinoline Arylaminothiazole Derivatives in Identifying Novel Potential Anticancer Agents by Applying 3D-QSAR, Docking, and Molecular Dynamics Simulation Studies. *J. Mex. Chem. Soc.* **2021**, *66*. [CrossRef]
113. Gonzalez-Aleman, R.; Hernandez-Castillo, D.; Rodriguez-Serradet, A.; Caballero, J.; Hernandez-Rodriguez, E.W.; Montero-Cabrera, L. BitClust: Fast Geometrical Clustering of Long Molecular Dynamics Simulations. *J. Chem. Inf. Model.* **2020**, *60*, 444–448. [CrossRef]
114. Daura, X.; Gademann, K.; Jaun, B.; Seebach, D.; van Gunsteren, W.F.; Mark, A.E. Peptide folding: When simulation meets experiment. *Angew. Chem. Int. Ed.* **1999**, *38*, 236–240. [CrossRef]
115. Zhang, H.; Qi, H.Z.; Mao, J.; Zhang, H.R.; Luo, Q.Q.; Hu, M.L.; Shen, C.; Ding, L. Discovery of novel microtubule stabilizers targeting taxane binding site by applying molecular docking, molecular dynamics simulation, and anticancer activity testing. *Bioorganic Chem.* **2022**, *122*, 105722. [CrossRef]
116. Elhemely, M.A.; Belgath, A.A.; El-Sayed, S.; Burusco, K.K.; Kadirvel, M.; Tirella, A.; Finegan, K.; Bryce, R.A.; Stratford, I.J.; Freeman, S. SAR of Novel 3-Arylisoquinolinones: Meta-Substitution on the Aryl Ring Dramatically Enhances Antiproliferative Activity through Binding to Microtubules. *J. Med. Chem.* **2022**, *65*, 4783–4797. [CrossRef]
117. Stroylov, V.S.; Svitanko, I.V.; Maksimenko, A.S.; Kislyi, V.P.; Semenova, M.N.; Semenov, V.V. Computational modeling and target synthesis of monomethoxy-substituted o-diphenylisoxazoles with unexpectedly high antimitotic microtubule destabilizing activity. *Bioorganic Med. Chem. Let.t* **2020**, *30*, 127608. [CrossRef]
118. Neto, R.A.M.; Santos, C.B.R.; Henriques, S.V.C.; Machado, L.O.; Cruz, J.N.; da Silva, C.; Federico, L.B.; Oliveira, E.H.C.; de Souza, M.P.C.; da Silva, P.N.B.; et al. Novel chalcones derivatives with potential antineoplastic activity investigated by docking and molecular dynamics simulations. *J. Biomol. Struct. Dyn.* **2022**, *40*, 2204–2216. [CrossRef]
119. Dash, S.G.; Naik, P.K. 10. 9-VINYL PHENYL NOSCAPINE AS POTENTIAL TUBULIN BINDING ANTICANCER AGENT. *Biotechnology* **2022**, *102*, 102.
120. Zhao, X.; Zhang, R.; Yu, X.; Yu, N.; Shi, Y.; Shu, M.; Shen, Y. Discovery of novel tubulin polymerization inhibitors by utilizing 3D-QSAR, molecular docking and molecular dynamics simulation. *New J. Chem.* **2022**, *46*, 16426–16435. [CrossRef]
121. Basu, D.; Majumdar, S.; Mandal, N.; Dastidar, S.G. Mechanisms of influence of the microtubule over-stabilizing ligands on the structure and intrinsic dynamics of α,β-Tubulin. *Comput. Biol. Chem.* **2022**, *96*, 107617. [CrossRef]
122. Kumbhar, B.V.; Bhandare, V.V. Exploring the interaction of Peloruside-A with drug resistant alphabetaII and alphabetaIII tubulin isotypes in human ovarian carcinoma using a molecular modeling approach. *J. Biomol. Struct. Dyn.* **2021**, *39*, 1990–2002. [CrossRef] [PubMed]

123. Radha, G.; Naik, P.K.; Lopus, M. In vitro characterization and molecular dynamic simulation of shikonin as a tubulin-targeted anticancer agent. *Comput. Biol. Med.* **2022**, *147*, 105789. [CrossRef] [PubMed]
124. Talimarada, D.; Sharma, A.; Holla, H. Identification of dual binding mode of Orthodiffenes towards human topoisomerase-I and alpha-tubulin: Exploring the potential role in anti-cancer activity via in silico study. *J. Biomol. Struct. Dyn.* **2022**, 1–15. [CrossRef]
125. Majumdar, S.; Basu, D.; Ghosh Dastidar, S. Conformational States of E7010 Is Complemented by Microclusters of Water Inside the α,β-Tubulin Core. *J. Chem. Inf. Model.* **2019**, *59*, 2274–2286. [CrossRef]
126. Pragyandipta, P.; Meher, R.K.; Reddy, P.K.; Pedaparti, R.; Kantevari, S.; Naik, P.K. Structure Based Design of Tubulin Binding 9-Arylimino Noscapinoids: Chemical Synthesis and Experimental Validation Against Breast Cancer Cell Lines. *Anal. Chem. Lett.* **2022**, *12*, 29–43. [CrossRef]
127. Yang, M.-H.; Mao, J.; Zhu, J.-H.; Zhang, H.; Ding, L. Wangzaozin A, a potent novel microtubule stabilizer, targets both the taxane and laulimalide sites on β-tubulin through molecular dynamics simulations. *Life Sci.* **2022**, *301*, 120583. [CrossRef]

Chapter 2. Discovery of possible maytansine site-targeting microtubule destabilizing agents

2.1. Introduction

The polymerization of tubulin heterodimers into microtubules is a crucial physiological process in cell division and is fundamentally mediated by longitudinal interactions between dimers. This process propels the integration of new dimers into the microtubule lattice, thus facilitating the curved-to-straight conformational change[5,41]. Specifically, these interactions occur between helix H8 of α-tubulin and a pocket shaped by loops S3-H3, S5-H5, and H11-H11' of β-tubulin[41]. The conformational transition is further characterized by the movement of the intermediate domain of both α- and β-tubulin subunits, during which strands S8 and S9 move closer to helix H8[5]. Any interference with this site could potentially inhibit tubulin polymerization and block the formation of longitudinal tubulin contacts[5].

One such molecule is maytansine, an ansamacrolide isolated from the African shrub *Maytenus ovatus* in 1972, which has been noted for its potent cytotoxicity due to its inhibitory impact on microtubule assembly[85–87]. Alongside maytansine, a small number of ligands, classified into four structural types, are known to bind to this site - maytansine derivatives, disorazole analogues, rhizoxin and spongistatin[86] (Figure 28).

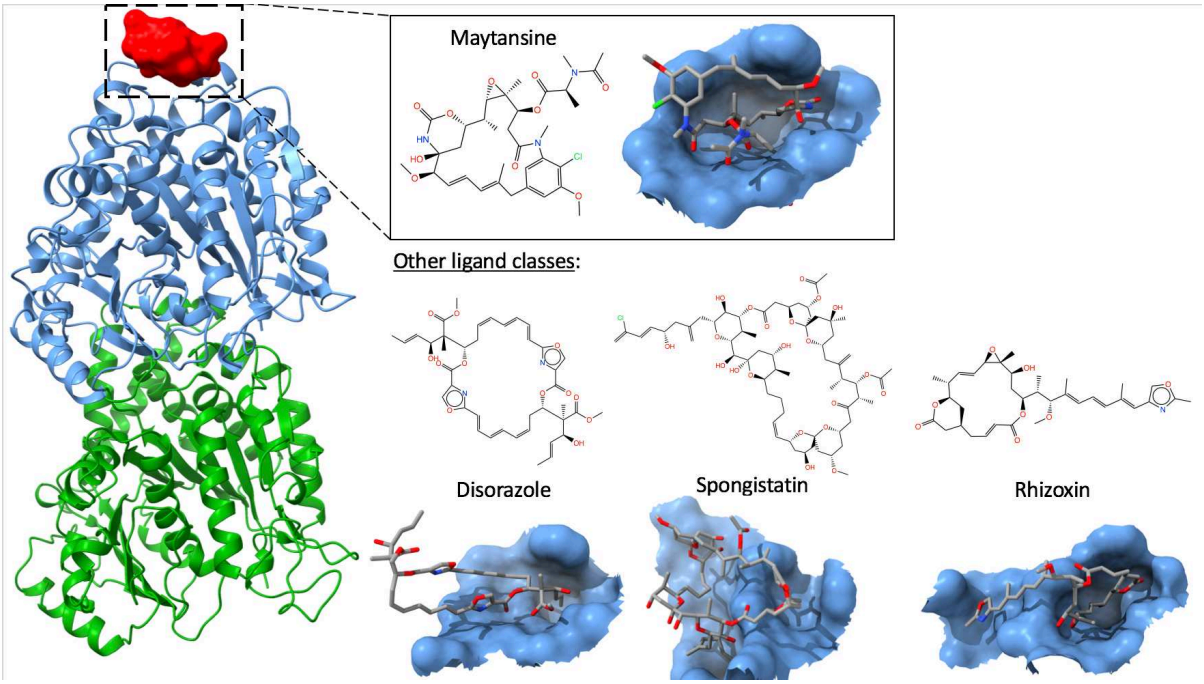


Figure 28. Maytansine binding site and some ligands that target it

However, the precise binding site of these ligands remained a matter of scientific debate until recent advancements in X-ray crystallography. This method facilitated a deeper understanding of the binding dynamics, revealing that these ligands, known to interfere with vinblastine binding, actually bind to a distinct site[5]. This site, found on an exposed pocket of β-tubulin, is located next to the guanosine nucleotide and is defined by the hydrophobic and polar residues of b-tubulin

helices H3', H11, and H11', along with the loops S3-H3' (T3-loop), S5-H5 (T5-loop), and H11-H11'[5]. A set of key β-tubulin residues were identified as being critical for the ligand binding. In essence, all compounds engage in hydrogen bonding with the main chain nitrogen atom of Val181, as well as the side chains of Lys105 and Asn102. Furthermore, a subset of these compounds create a hydrogen bond with the main chain carbonyl group of Gly100[5,41,87,88]. The conformation of the maytansine binding site remains fairly stable despite the curved-to-straight transition. This observation indicates that ligands targeting the maytansine site have the potential to bind with both curved and straight states of tubulin[5,41].

The cytotoxic potential of these ligands, especially at low concentrations, has resulted in their clinical application as part of antibody-drug conjugates (ADCs), with maytansine derivatives most commonly employed as cytotoxic warheads[5,86,89]. Despite this promising avenue, these ligands present significant drawbacks due to their complexity being extracted from natural sources, which pose challenges for synthesis, separation and purification[41,86]. Furthermore, the high costs associated with ADCs and their systemic toxicity observed in clinical trials further complicate their application[86–88]. Additionally, the solvent-exposed pocket is difficult to target[86].

Given these complexities, the maytansine binding site remains under-explored yet fascinating for research. Preliminary computational work has provided pharmacophore models, although these have not been extensively used for screening[41,86,90]. A notable work by Li et al. saw an application of a manually developed pharmacophore model to screening two libraries of commercially available macrocycle molecules, establishing 15 virtual hits that have not been further validated experimentally[86]. This research landscape underscores the need for the discovery of small, easily synthesizable molecules that can inhibit tubulin polymerization by binding to the maytansine site, offering new pathways for pharmacological profile modification. In light of this, the primary aim of this chapter is to attempt to discover such molecules, thus contributing to the broader understanding of tubulin dynamics and its role in cell division.

2.2. Virtual screening of the ChEMBL library

2.2.1. Survey of available data

Building upon previous work involving the creation and application of pharmacophore models[41,86,90], we decided to perform pharmacophore-based virtual screening of ChEMBL, a database containing known drugs and drug-like compounds and their respective bioassay results[91]. Our aim was to identify molecules that conformed to a pharmacophore model of maytansine site binding compounds and exhibited cytotoxic properties with an unknown mechanism of action, which could then potentially be linked to the maytansine binding site of the tubulin protein.

A suitable pharmacophore model could be derived automatically from a well-resolved crystal structure of tubulin co-crystallized with a maytansine site bound ligand. A search in the Protein Data Bank (PDB)[92], a primary source of protein-ligand crystal data, yielded six records where tubulin was cocrystalized with maytansine site-bound ligands (6FJM, 6FII, 6FJF, 4TV8, 4TUY, 4TV9). We proposed to employ pharmacophore screening followed by docking validation of the discovered virtual hits.

2.2.2. Re-docking

To ensure the effectiveness of our study, we decided to employ a pharmacophore model that corresponds to the binding pattern of a compound which our docking software can accurately reproduce. The rationale is that if the pharmacophore model is based on a molecule whose docking pose our software can accurately recreate, then the potential hits conforming to this pharmacophore model will likely also dock in a similar manner. This would potentially result in a more biologically relevant and plausible pose generated by the docking software within the binding site, leading to prioritization of more relevant compounds for further study.

To determine the most suitable protein-ligand complex, we conducted a re-docking experiment. This process involved extracting the native ligand from a protein-ligand complex, generating its random three-dimensional conformation, and using our docking software, PLANTS, to reintroduce the ligand into the binding site. We utilized the root mean square deviation of atomic positions between native and docked poses of the ligand as a metric to evaluate how well the pose was reproduced. Acceptable docking software performance is typically indicated by an RMSD value of less than 2.0 Å.

We started by extracting the ligand from each PDB structure. We then generated a random pose for each ligand using the conformation sampling tool provided by ChemAxon. Subsequently, all solvent molecules, ions, and minor organic molecules were eliminated from the protein structure to focus on the interaction of the ligand and the protein.

We used the SPORES software to prepare both the ligand and the ligand-free protein structure for further analysis. We defined the maytansine binding site as all the atoms within a distance of no more than 12 Å from the center of mass of each ligand. For scoring the docking simulations, we utilized the *chemplp* scoring function and set the software to generate 10 different docking poses for each ligand.

In defining the binding site, we chose not to include any water molecules. This decision was based on the fact that none of the studies we reviewed which discussed the structure of the binding site indicated any significant role of water molecules in ligand binding. This step helped

us streamline our docking simulation by focusing only on the crucial components of the binding site.

As a result, disorazole, a natural product known for its high cytotoxicity[88], demonstrated the best re-docking performance with an RMSD of 3.05 Å. However, it should be noted that a major setback in the RMSD value is caused by a solvent-exposed part of the molecule, while most of the main "body" of the molecule, especially parts important for the protein-ligand interaction, were re-docked correctly (Figure 29). Disorazole was therefore selected for pharmacophore model development.

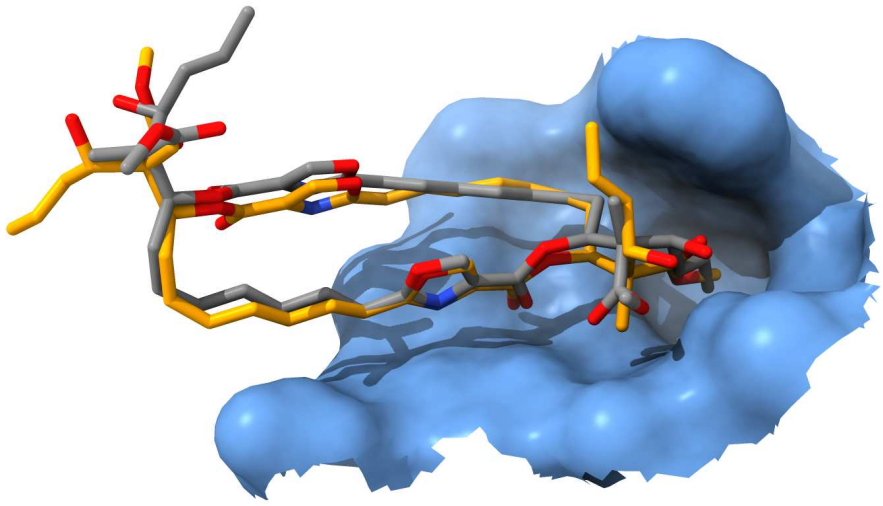


Figure 29. Re-docking of disorazole. Gray – experimental pose, orange – re-docked pose.

Failure to dock these complex compounds is clearly related to the sampling of macrocycle conformers. Alternative computational studies on maytansine (performed by Dr. Helena Perez-Peña within the TubInTrain consortium), showed that redocking of the PDB conformation of maytansine is successful with AutoDock[80], while software not using the correct initial ligand geometry would typically fail. Finally, the in-house program S4MPLE[93], developed as a general approach to "difficult" flexible and/or multiligand conformational sampling and docking programs was challenged to redock maytansine. S4MPLE is a Lamarckian evolutionary algorithm, always starting from a randomized set of conformers: even if the PDB file of maytansine is provided at input, that particular geometry will be ignored. It uses the AMBER/GAFF force field coupled to an empirical continuum desolvation model to score the stability of poses, and employs contact fingerprints to manage the non-redundancy of the evolving populations of conformers. Nonetheless, the initial S4MPLE run consisting of 5-fold repeated evolutionary simulations of 700 generations each also failed, although the number of generations was increased to 700 from the

default 500. Eventually, S4MPLE was deployed on 48 CPUs, and out of the 48 simulations, two were found to converge towards the native maytansine pose, ranked as the most stable of all other sampled conformations (Figure 30). However, at a cost of ~50 CPU hours/molecule, this approach is not applicable to high throughput virtual screening and was not pursued.

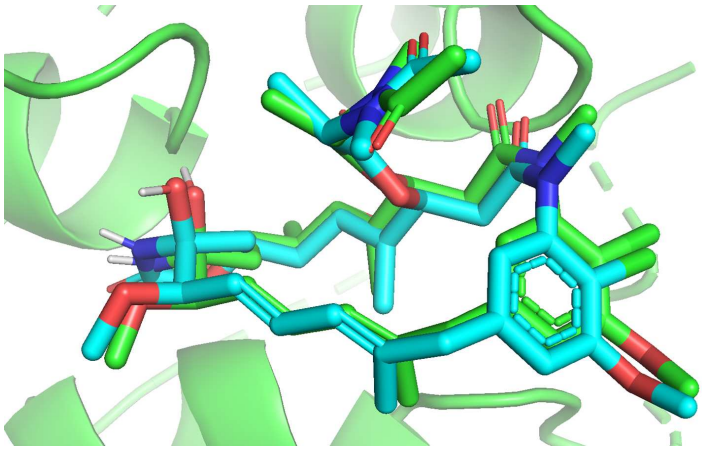


Figure 30. S4MPLE-generated native pose of maytansine, after an aggressive parallel deployment on 48 CPUs.

2.2.3. Pharmacophore modelling

We used the 6FJM protein-ligand crystal structure in conjunction with LigandScout software to automatically generate a pharmacophore model. The model comprised 10 features (Figure 32): 6 hydrophobic spheres corresponding to hydrophobic interactions of the conjugated polyalkene side chain with βTRP407, βPHE404, βVAL182, βTYR408, βVAL181, 3 hydrogen bond acceptors targeted at βVAL181, βASN101, βVAL182 residues, and 1 hydrogen bond donor with βGLY100. The model also included exclusion zones mimicking the binding site shape, not shown for brevity. The resulting model included interactions with key residues highlighted earlier by Prota et al., Porter et al., and Li et al., which affirmed the validity of our model.

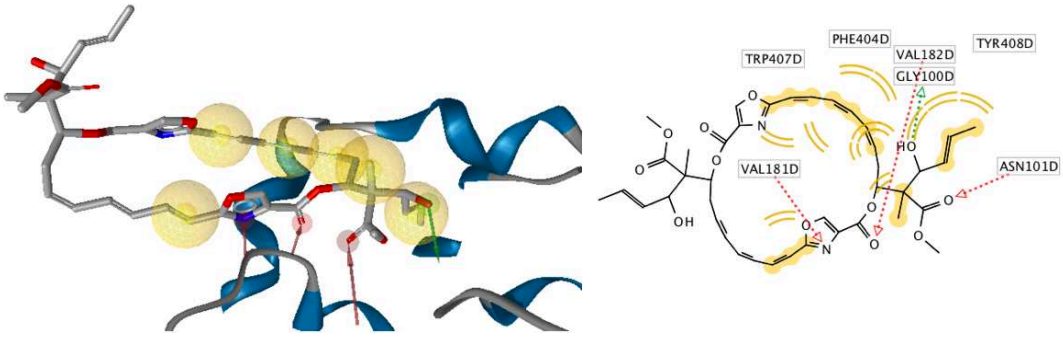


Figure 31. Overview of the derived pharmacophore model

Furthermore, using ChimeraX software, we superimposed the 4TV8 protein structure onto the 6FJM protein structure to observe the spatial relationship between them. Specifically, we utilized the “align” command with the “matchAtomNames” option set as a boolean “True” to

ensure optimal alignment. After this, we extracted the pose of the maytansine ligand, which is native to the 4TV8 PDB record, to evaluate how it overlapped with the pharmacophore model we derived from the 6FJM disorazole. Upon visual inspection, we were satisfied with the results (Figure 33). The overlap was particularly striking for features corresponding to interactions with the βASN101 and βVAL181 residues. This similarity further substantiated the accuracy of our model, granting us confidence to proceed with its application in screening.

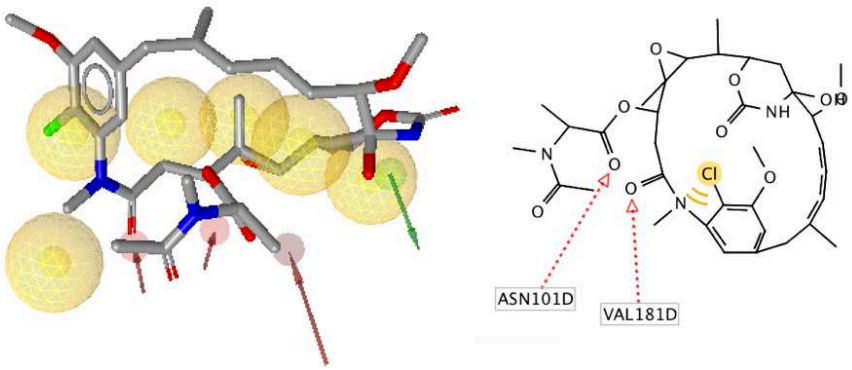


Figure 32. Maytansine aligns well with disorazole pharmacophore model

2.2.4. Screening library preparation

For the virtual screening, we selected the ChEMBL database (v. 26), containing 1,771,509 molecules. To ensure consistency and accuracy during the screening process, all the molecules in our study were standardized. This standardization process was conducted based on the protocol used on the Virtual Screening Web Server at the Laboratory of Chemoinformatics at the University of Strasbourg. We employed the ChemAxon Standardizer tool for this task.

The standardization process involved several steps. First, we performed dearomatization followed by final aromatization; however, heterocyclic compounds like pyridone were left unaromatized. The next step was dealkalization, where alkali ions were removed. Then, the molecular structures were converted into canonical Simplified Molecular Input Line Entry System (SMILES) format for easier processing and comparison. We also removed any salts and mixtures from the compounds to isolate the active molecules. This was followed by the neutralization of all species, except for nitrogen (IV). Finally, we generated the most stable tautomer of each compound, according to the protocol defined by ChemAxon. All these steps ensured that the molecules were in an optimal and standardized form for our screening and subsequent analyses.

Pharmacophore screening is based on the fundamental principle of matching a large collection of compounds, represented by various three-dimensional (3D) conformations, to a 3D pharmacophore model. The higher number of conformations is used, the higher is the chance to capture a biologically significant conformation for each molecule in the conformation set.

In this study, we used LigandScout's in-built conformational sampling tool, iCon, to prepare conformations for all 1,771,509 molecules in the ChEMBL database (v. 26). Specifically, we used the "iCon-best" option, meaning for each molecule we generated a maximum of 200 unique conformations, distinguished by a RMSD value of at least 0.7 Å between different conformations. This approach facilitated a comprehensive conformational analysis for each molecule, thereby enhancing the effectiveness of our pharmacophore screening.

2.2.5. Pharmacophore screening

After constructing the database, we initiated the pharmacophore screening process using LigandScout. We chose to employ the "Pharmacophore fit" scoring function. This score provides a reflection of the number of matched features, in addition to the RMSD of their positions relative to the feature sphere's center.

For the screening process, we opted for the "Match at least 3 query features" screening mode. We also set the retrieval mode to "Stop after first matching conformation" to speed up the screening process. To further refine the screening, we activated the check for excluded volume clashes.

Through these settings, we were able to identify 1,035 potential hits from the ChEMBL database that had a Pharmacophore-Fit score exceeding 64.

2.2.6. Protein-ligand docking

To further refine our results, we docked the 1,035 virtual hits into the maytansine binding site extracted from the 6FJM PDB structure alongside the native ligand, disorazole. This was accomplished using the PLANTS docking software. The binding site was defined as all atoms within 12 Å radius from the center of mass of the native ligand. We used the *chemplp* scoring function. For each compound we calculated 10 docking poses within the site. The pose with the lowest docking score was considered to be the most probable one. A total of 104 molecules exhibited a better docking score than disorazole and aligned with at least three features of the query pharmacophore model.

Upon cross-referencing these 104 molecules with the ChEMBL database, we identified 6 molecules with documented cytotoxic action in bioassays (Figure 34). These molecules are of particular interest due to their conformance to the pharmacophore model, superior docking score compared to the native ligand, and associated cytotoxic properties.

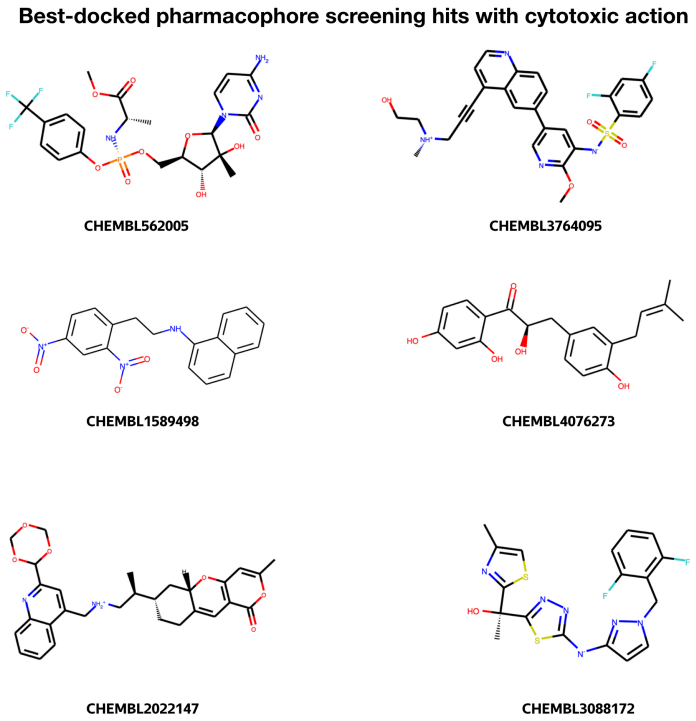


Figure 33. Six virtual hits with known cytotoxicity values

From the 6 virtual hits established by pharmacophore screening, compound ChEMBL4076273 attracted our attention in particular, because it is a natural product called Glycibridin B isolated from *Glycyrrhiza glabra* (commonly known as licorice)[94]. It had a docking score better than the original ligand, disorazole (-93.08 vs -91.05) and overlapped three pharmacophore features, showing potential interactions with key residues of the site (Figure 35). One work published its selective moderate cytotoxic activity on the MCF7 cancer cell line[94]. So we decided to pursue this molecule further.

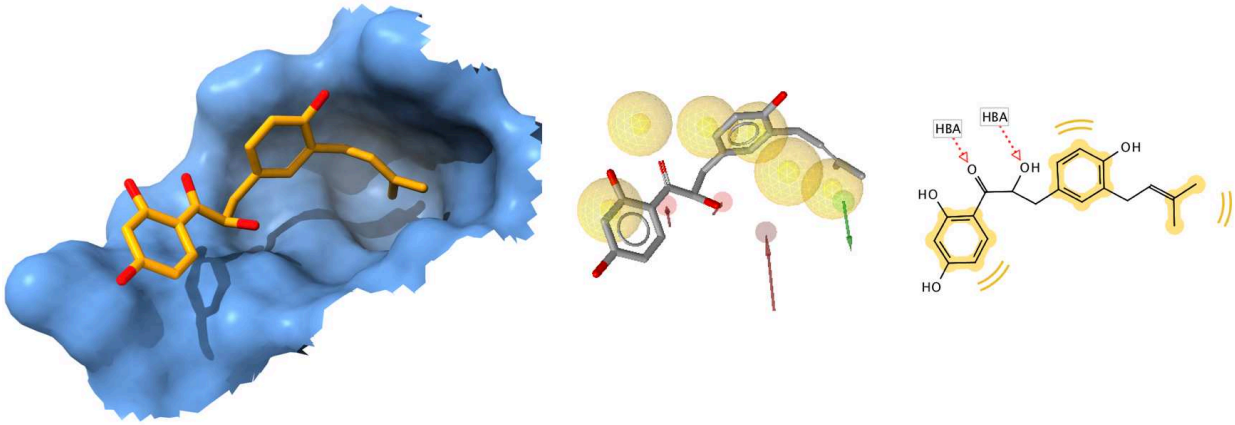


Figure 34. Virtual hit called glycibridin B has shown good docking score and overlaps five pharmacophore features of disorazole

The synthesis of this molecule was attempted by our collaborators within the TubInTrain network – the group of Prof. Daniele Passarella, University of Milano, Italy. Before starting to plan the synthesis, our colleagues asked us to see if certain structural modifications can simplify the molecule yet preserve its beneficial contacts with the key residues within the site.

2.2.7. Virtual hit optimization

The modifications discussed herein concern the scaffold of the molecule and its various substituents, located at the aromatic rings. When it comes to the scaffold, we were interested to check whether the chain connecting the two aromatic rings should contain a hydroxyketone, a diol, or just be made of carbon atoms, and whether the stereochemistry of the hydroxyl group(-s) in the chain matters (Figure 36).

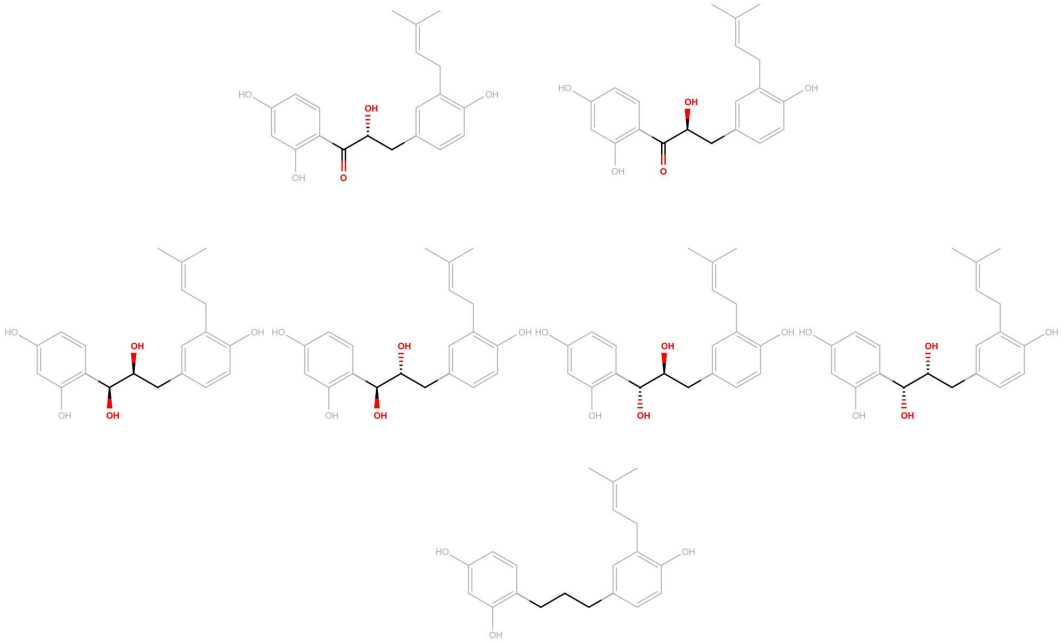


Figure 35. Examined modifications of the intermediate section of the virtual hit

Another question regarding the scaffold chain was whether it can be formed not by 3, but by 2 atoms (Figure 37).

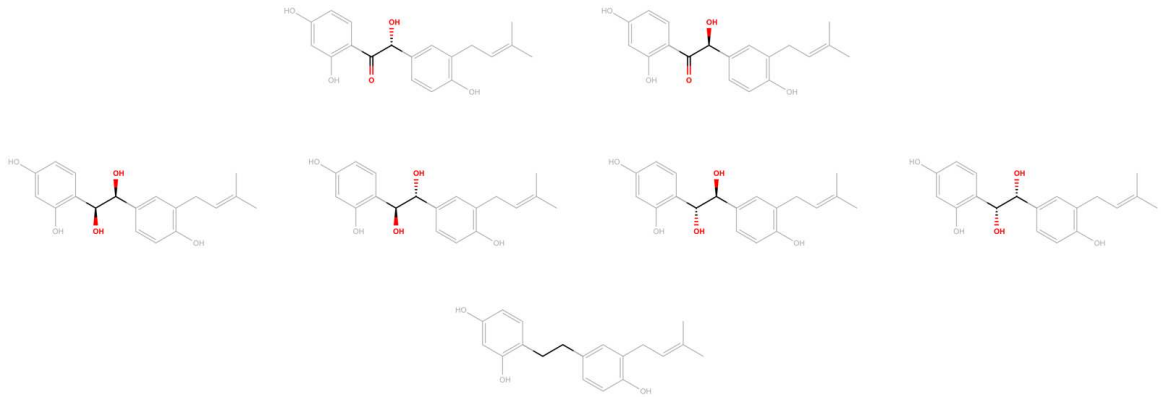


Figure 36. Tested scaffolds that are one carbon atom shorter in the intermediate section

In total, we have checked 6 scaffolds (Figure 38).

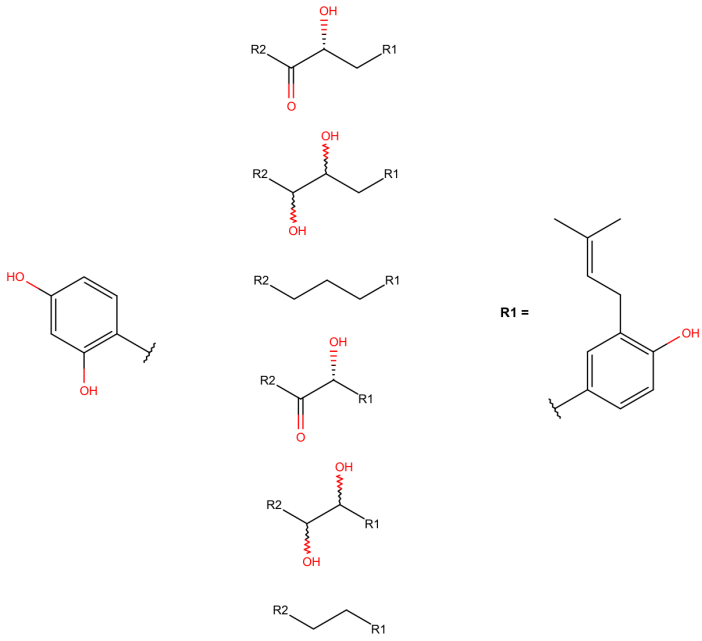


Figure 37. Overview of enumerated variants of the middle section

For each scaffold, we analyzed the importance of several factors for binding, such as (1) the presence of the dimethylallyl substituent (Figure 39a), (2) the variation of the hydroxyl groups positions in the aromatic rings (Figure 39b); (3) the replacement of the hydroxyl groups at their initial positions by a hydrogen atom (Figure 39c); (4) the replacement of the hydroxyl groups by another substituent from the list of suggested substituents: –F, –Cl, $–OCH_3$, $–NH_2$, $–NO_2$ (Figure 39d). In total, 1444 modifications of CHEMBL4076273 were examined.

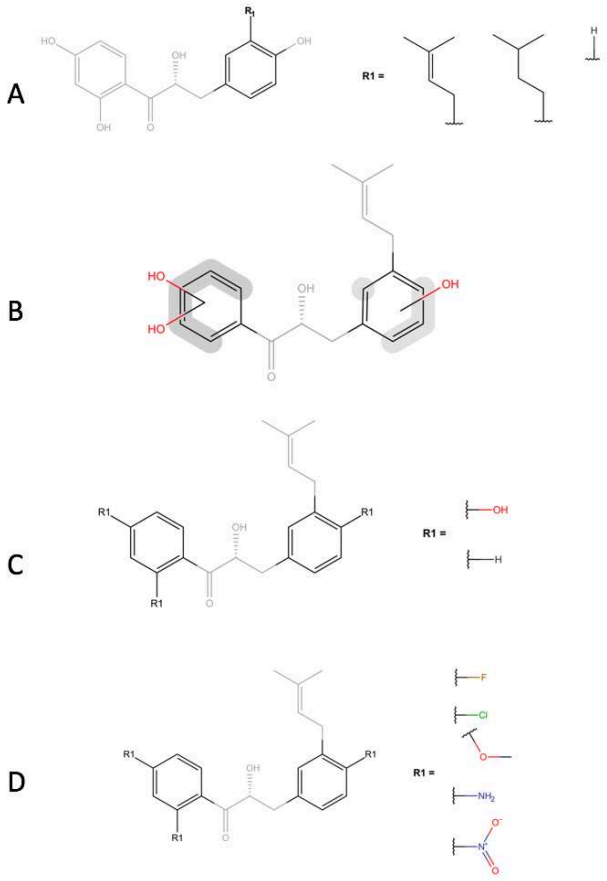


Figure 38. Summary of enumerated side chain modifications

Each of the generated modifications has been then docked into a region of the protein consisting of the maytansine site and two closely-located cavities using the PLANTS docking software. As such, it would be possible to distinguish between ligands that could potentially bind to the maytansine site, and those that won't. As a result of that, we have obtained 10 docking poses for each compound, each with an associated docking score. The lower the docking score value, the better the binding. Thus, by comparing the best-scoring docking poses of all ligands, it is possible to find out which modifications can potentially demonstrate better binding to tubulin.

According to our calculations, none of the modifications or simplifications of the original ligand's structure caused an increase in affinity towards the maytansine binding site, hence we recommended to continue with the original structure.

2.2.8. Virtual hit retrosynthesis route analysis

Additionally, we have generated a possible retrosynthetic pathway to the virtual hit compound using the freely available AiZynthFinder automatic retrosynthesis tool[95].

Retrosynthesis is a task of reconstructing a synthetic route given only the final structure of the target molecule and a database of known reaction rules. One of the available tools for this task

is the open-source retrosynthesis planning tool called AiZynthFinder. It uses the Monte Carlo tree search algorithm that recursively breaks down a molecule into smaller fragments until purchasable precursors are generated. The way a molecule is broken down is guided by a neural network that suggests possible precursors, being trained on a library of known reaction templates. In this particular case, we set the tool up in such a way to break down a molecule until it can generate precursor compounds that can be found in a subset of the ZINC database containing 17,422,831 purchasable compounds, which is a stock file made from the ZINC database on 17th of April, 2020, and comes together with the AiZynthFinder installation.

The tool can generate several possible synthetic pathways to a given molecule. Each of the generated pathways is described by a single number called “Score”. The “score” reflects the fraction of purchasable precursors in the generated route and the number of reactions required to synthesize the target compound. The score for a “solved” compound (i.e., a compound that is demonstrated to be synthesizable from purchasable building blocks) is close to 1.0, whereas the score for an “un-solved” compound is typically less than 0.8. However, it should be noted that the score was designed to support the tree search and is rather indiscriminate with regard to the quality of the route (i.e. if it’s a good route or not) and should be interpreted with care.

If a compound can indeed be “solved”, the results clearly display which precursors to procure in order to synthesize the target compound. The predicted route is drawn with precursors in stock in a green rectangle, and the precursors that are not in stock highlighted by an orange one.

Among the features of this tool are its high operation speed (generation of several synthetic routes for one molecule usually takes no more than 2-3 minutes) and the ability to run retrosynthetic jobs in batch mode via a dedicated command-line mode, while a more user-friendly GUI is available via an interactive Jupyter notebook interface.

When working with AiZynthFinder, the stereochemistry features of the target compound had to be omitted due to the limitations of the used software. The obtained path is shown in Figure 40.

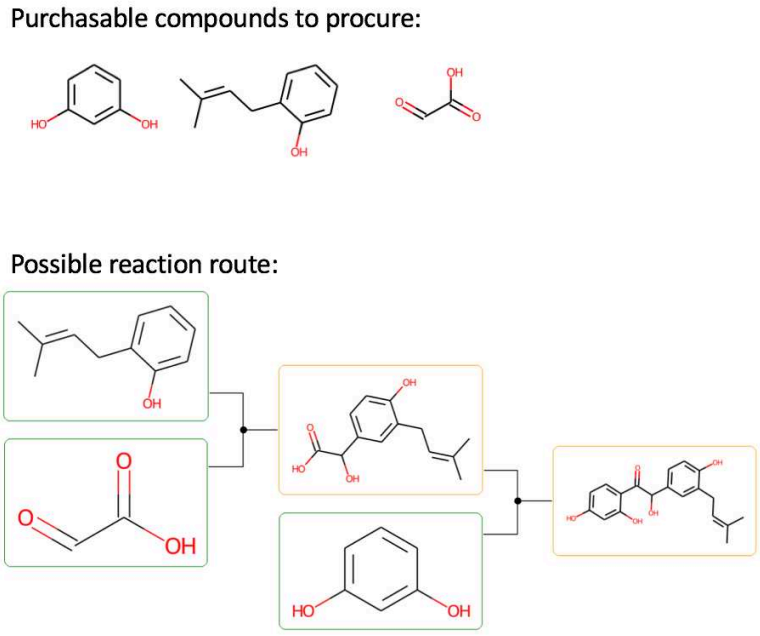


Figure 39. Retrosynthetic route suggested by AiZynthFinder for the virtual hit molecule

2.2.9. Virtual hit analogue selection

Our colleagues from the organic chemistry lab deemed our initially proposed synthetic route unrealistic, due to unaddressed regioselectivity and stereoselectivity issues. In response, they proposed an alternative, more intricate retrosynthetic pathway (Figure 41).

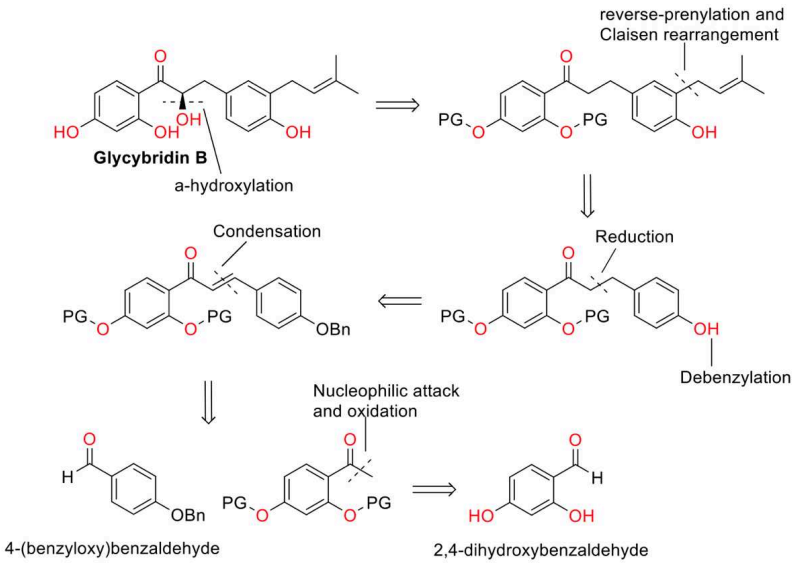


Figure 40. Alternative synthesis route devised by our colleagues

However, this path proved to be synthetically difficult, rendering the synthesis of Glycibridin B within a reasonable timeframe implausible. Consequently, we identified 14 potential molecules of interest that could be considered instead of the originally targeted natural product:

seven were intermediates already produced in the attempted synthetic route, while the remaining seven were prospective products easily accessible from these intermediates (Figure 42).

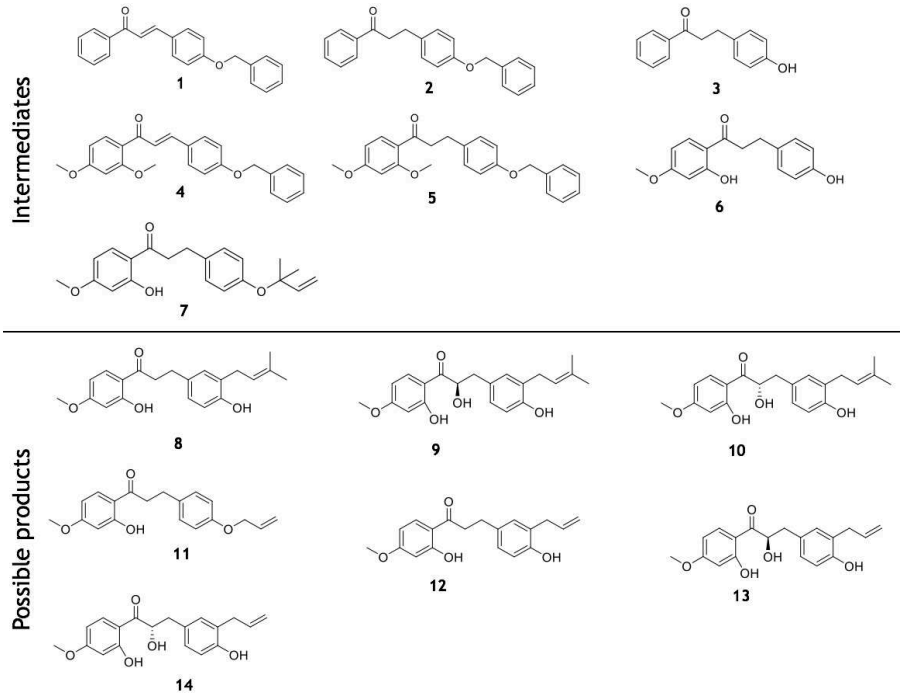


Figure 41. Intermediate and possible future product molecules that could be considered instead of the initially proposed glycibridin B molecule

Our objective thus was to see whether any of these 14 molecules could serve as effective binders to the maytansine site. To facilitate this, we executed a two-sided computational approach. Initially, we docked the 14 ligands, alongside disorazole (the native ligand), into the maytansine site using PLANTS software. We adhered to the previously established parameters for this process, which included using the same definition of the maytansine binding site, which included all atoms within 12Å from the center of mass of the native ligand of the 6fjm PDB structure. The pose with the lowest docking score was selected as the most probable one. However, docking results saw all 14 molecules get a better docking score than the native ligand, so it could not be used to differ or rank the ligands.

To accommodate for that, we also conducted a blind docking experiment. Blind docking explores whether a ligand in question forms more intra-molecular interactions with alternative protein sites, as opposed to the specifically assigned site. This approach stems from the characteristic nature of the docking process, which theoretically allows any ligand to be accommodated into a binding site. Consequently, blind docking serves as a surrogate method: it's rapid but may lack precision. Nonetheless, it provides a broad overview, indicating if there exist other pockets on the protein to which the docked ligand demonstrates a stronger affinity.

The blind docking was also done using the PLANTS software. We still used the α,β-tubulin heterodimer structure extracted from 6FJM PDB structure, pre-processed by removing everything but the C and D chains of the protein, and processed with SPORES before docking. However, this time, the "binding site" was specified as all atoms within 60 Å from the center of mass between four randomly selected residues on the interdimer surface. This ensured that the binding site definition covered all atoms in the tubulin protein. To ensure sufficient conformational sampling, we configured PLANTS to compute 40 conformations for each ligand. This extensive sampling provided a broader perspective on potential binding scenarios. Blind docking narrowed our focus to ligands 9 and 10, which displayed a better docking score in the maytansine site in the majority of their sampled poses.

Upon identifying ligands 9 and 10 as our most promising candidates, we further analyzed their overlap with the disorazole pharmacophore model. Both ligands' best docking poses overlapped with five features of the pharmacophore model and fit well within the binding site (Figure 43). These two molecules, yet to be synthesized, emerged as promising alternatives.

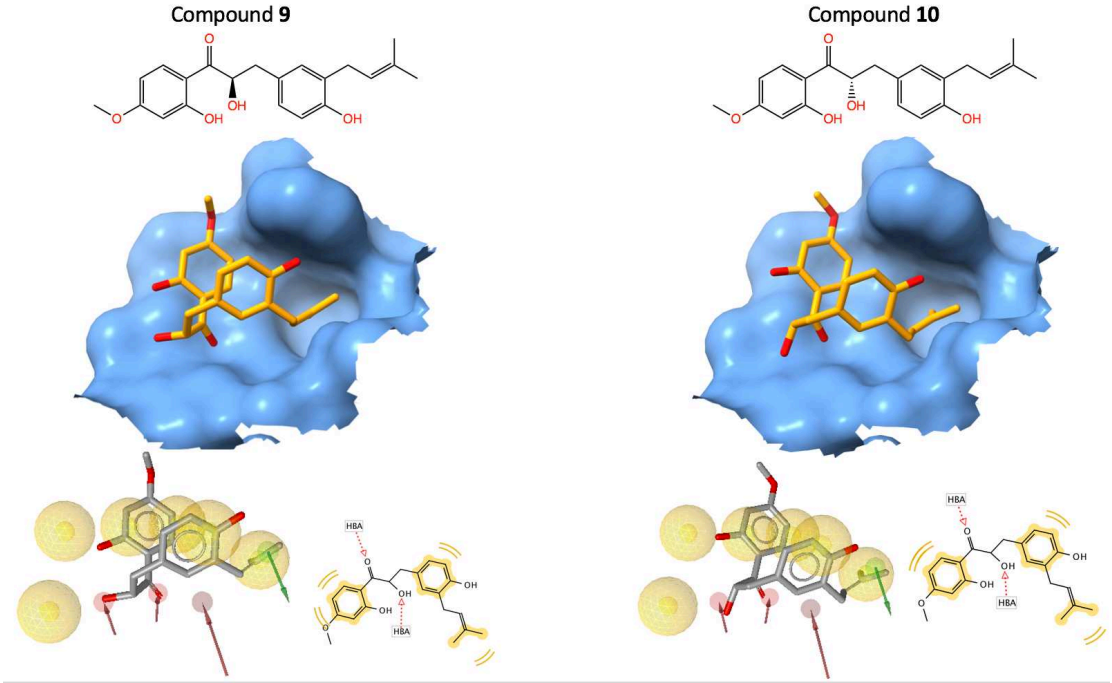


Figure 42. Two possible future products of glycibridin synthesis with promising fitness to the binding site and pharmacophore model

2.2.10. Results and discussion

Through the execution of this project, we developed a virtual screening pipeline that yielded 104 virtual hits, six of which had previously been identified in cytotoxic assays. We initially selected one molecule for synthesis and validation, but complexities in its synthesis necessitated a

search for alternatives. Computational analysis enabled us to propose two synthetically accessible molecules as potential candidates. These molecules are yet to be synthesized and experimentally tested. With the recent resolution of new tubulin crystal structures with bound maytansine site ligands, there is scope for re-exploring this project using these new structures. Additionally, while we didn't conduct cross-docking in this study due to similarities in the ligands' binding modes, the availability of new data opens the possibility for its inclusion in future experiments. It's also worth noting that the shallow character of the binding site was not accounted for by the docking software used in our study.

2.3. Virtual screening of the Enamine library

As a result of the virtual screening of the ChEMBL database, we discovered a virtual hit that we subjected to further study. Unfortunately, the complex process of synthesizing this molecule proved to be an obstacle, hindering our progress in the project. Consequently, we made a decision to switch our approach and screen a large library of commercially-available compounds instead. We specifically opted for the Enamine High Throughput Screening (HTS) library, which, at the time, consisted of 2,688,748 million compounds. However, new crystal structures of tubulin co-crystallized with maytansine site-bound ligands made us reconsider the choice of a pharmacophore model for the task.

2.3.1. Survey of available data

Sometime after we finished the screening of the ChEMBL library described in section 2.2, three new crystal structures of tubulin co-crystallized with maytansine site-binding ligands were published in the RCSB PDB. Their codes are **7E4R**, **7E4Q** and **7E4Z**. Ligands referenced in these structures come from a work by Li et al.[86], which studied the importance of an ester side chain in maytansinoids. The 2D structures of these molecules are shown in Figure 44.

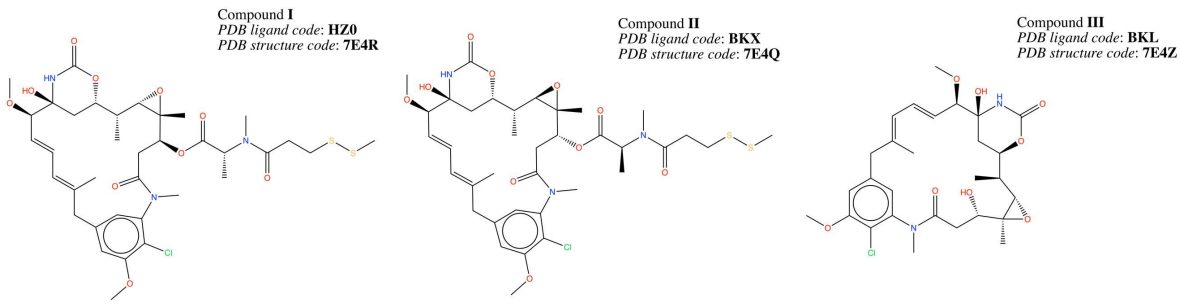


Figure 43. Three maytansine derivatives

Compared to the structure of maytansine itself, compounds **I** and **II** have a longer side chain with a disulfide group, and compound **III** doesn't have any sidechain at all.

We also compared these new structures' bound conformations to that of maytansine in 3D. For this, we aligned the three new protein-ligand complexes to the 4TV8 structure using ChimeraX software and the align command with the "matchAtomNames" flag turned on.

By analyzing the overlapped poses, we saw that the new structures keep exactly the same binding mode, establishing similar interactions to the key residues described in the literature (Figure 45).

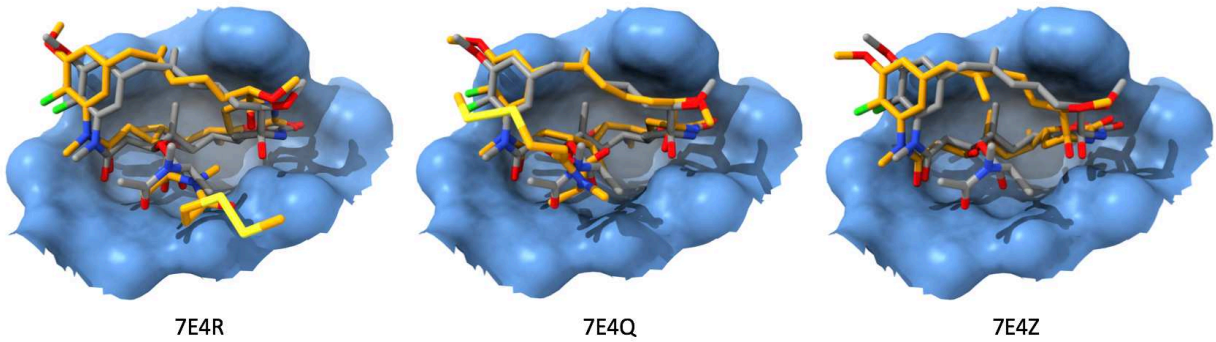


Figure 44. Overlap of novel maytansine derivatives with maytansine pose in the binding site. Gray – maytansine, orange – derivative molecule.

As such, all three compounds follow maytansine's binding mode. Thus, we decided to automatically build the pharmacophore model of maytansinol (compound III) using LigandScout, validate it on other maytansinoids with resolved crystal structures, and use this model to screen the Enamine HTS collection. We specifically chose this compound collection because of its high diversity and relatively low cost of compounds per one gram.

2.3.2. Pharmacophore modelling

We have then proceeded to construct the pharmacophore model of compound **III** (Figure 47).

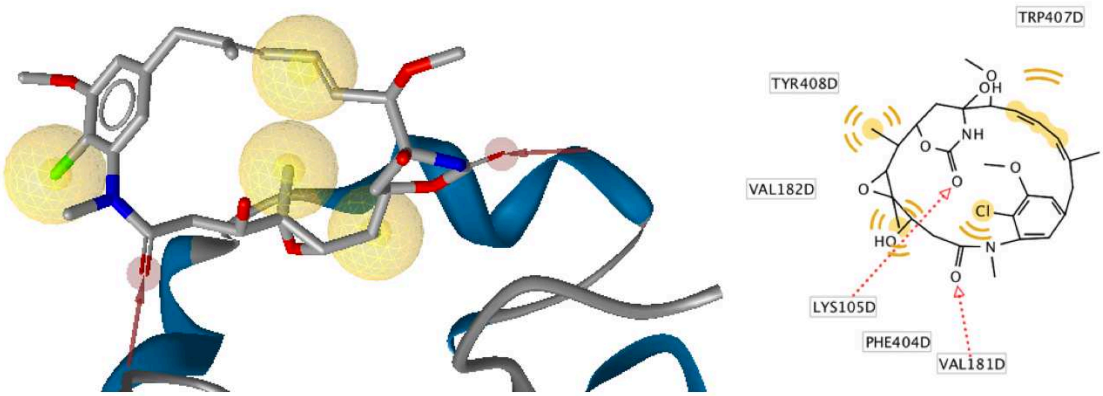


Figure 45. Compound III's pharmacophore model

The resulting model is similar to the one we used for pharmacophore screening of the ChEMBL database in the previous project. The difference between the models lies in a different arrangement of hydrophobic spheres, and a lower number of pharmacophore features in general (6 versus 10). The exact positions of some exclusion spheres (not shown in figures) were different, although the general shape of the site defined by the exclusion volume spheres remained largely the same.

2.3.3. Model validation

We validated the model by overlapping other maytansinoids with the 7E4Z system, visually inspecting the quality of feature overlap. Namely, we used 4TV8, 6FJF, 7E4Q, 7E4R for model validation. As a result, we saw that the maytansinoid compounds had a good alignment with this model, which let us use it in performing the virtual screening of the Enamine library.

2.3.4. Screening library preparation

As mentioned, we chose the Enamine HTS library, containing 2,688,748 molecules. To be consist in our screening process, we applied a standardization procedure to all molecules included in this screening campaign. The applied standardization process mirrored the approach employed for screening the ChEMBL library, as outlined in section 2.2.4.

In brief, the standardization involved several sequential steps. First, the standardization pipeline ensured proper aromatization. Second, it included dealkalization. Thirdly, compounds were converted into canonical SMILES strings. Additionally, salts and mixtures were removed, with active molecules isolated. Then, all species were neutralized, except for nitrogen (IV). Finally, the most stable tautomer was generated by ChemAxon tools for each compound.

In this study, we employed LigandScout's integrated conformational sampling tool, iCon, to prepare conformations for all 2,688,748 molecules sourced from the Enamine HTS database. Specifically, we utilized the "iCon-fast" option, which generated up to 25 unique conformations

for each molecule. These conformations were distinguished by a Root Mean Square Deviation (RMSD) value of at least 0.7 Å between different conformations. This approach offered a suitable balance between screening speed and quality, enabling us to efficiently prepare the database and conduct the screening process within a reasonable timeframe.

2.3.5. Pharmacophore screening

After preparing the database, we started the pharmacophore screening using the LigandScout software. We made the decision to utilize the "Pharmacophore fit" scoring function, which evaluates the degree of feature matching and the RMSD of their positions relative to the center of the feature sphere

For the screening process, we selected the "Match at least 3 query features" screening mode. To make the screening process more efficient, we configured the retrieval mode to "Stop after first matching conformation." Additionally, we enabled the check for excluded volume clashes to enhance the screening precision.

By employing these settings, we successfully identified 151 potential hits from the Enamine HTS collection. These hits exhibited a Pharmacophore-Fit score surpassing 65, indicating their compatibility with the pharmacophore model. A random sample of three best-matching virtual hits is shown in Figure 48.

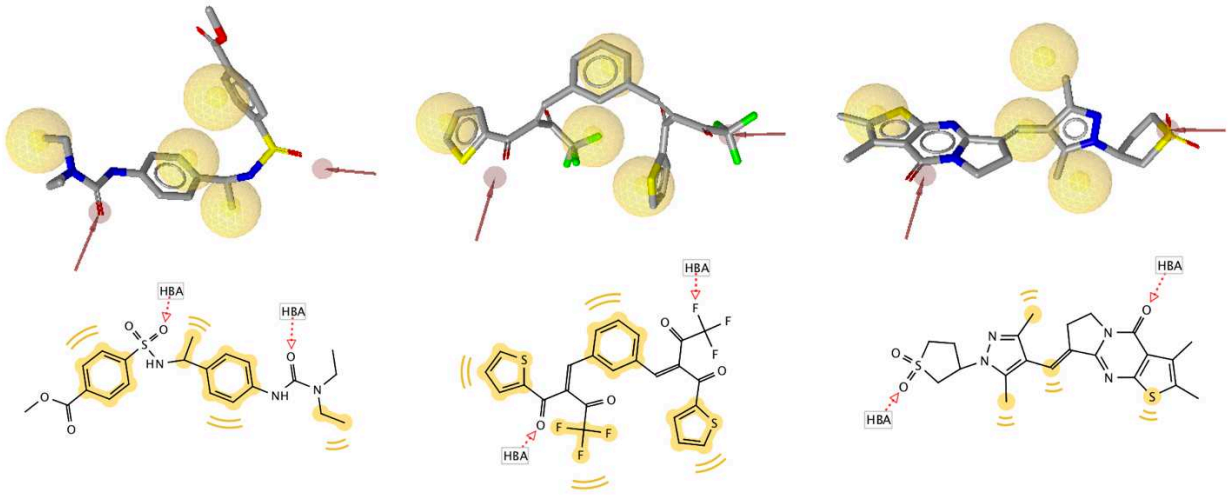


Figure 46. A random sample of three virtual hits from the Enamine HTS library that match the maytansinol pharmacophore model

2.3.6. Protein-ligand docking

We have then proceeded to dock these 151 molecules along compound **III** into its corresponding protein structure (PDB code: **7E4Z**). Docking was done using the PLANTS software. After docking, each compound had ten calculated poses within the binding site. The

best-scoring pose was used to characterize each compound by a *chemplp* PLANTS docking score and four ligand efficiency scores derived from it. Assuming *DS* is docking score, *n* is total number of ligand atoms, *b* is total number of rotatable bonds, the ligand efficiency scores were defined by Equations 5-8.

$$LE_1 = {}^{DS}/_{n} \quad \textbf{(5)}$$

$$LE_2 = {}^{DS}/_{\sqrt{n}} \quad \textbf{(6)}$$

$$LE_3 = {}^{DS}/_{(b+1)} \quad \textbf{(7)}$$

$$LE_4 = {}^{DS}\Big/_{\sqrt{(n \times b) + n}} \quad \textbf{(8)}$$

We reasoned that if a compound gets a docking score value and the ligand efficiency scores values better than the original ligand (compound **III**, in this case), it may demonstrate a similar or better binding affinity towards tubulin's maytansine binding site. The purpose of the ligand efficiency scores is to highlight the relevance of chemical interactions of a virtual hit with the binding site, reducing the influence of the sheer number of atoms and rotatable bonds on the final docking score.

We then chose most promising molecules by performing a Pareto front optimization of the list of 151 virtual hits based on the combination of docking score values and ligand efficiency scores.

Pareto front optimization, also known as multi-objective optimization or simply Pareto optimization, is a technique used to solve problems that involve multiple conflicting objectives. Pareto front optimization seeks to find a set of solutions that represents the best trade-offs between the objectives. These solutions are known as Pareto optimal solutions. A Pareto optimal solution is one that cannot be improved in any one objective without sacrificing performance in another objective. The Pareto front refers to the set of all Pareto optimal solutions, which forms a curve or surface in the objective space. It represents the trade-off relationship between the objectives, showing the best achievable performance for each objective combination. The process of Pareto front optimization involves exploring the solution space and evaluating the objective functions to identify and refine Pareto optimal solutions. The purpose of this exercise was to find molecules that are good binders not because they just have many atoms and as a consequence a higher docking score, but because they form meaningful interactions with the target binding pocket. Essentially, this is done to lower the influence of ligand atom count on the final docking score value, instead promoting the interactions themselves as the main factor contributing to the docking score value.

Following this logic, we established 11 molecules from the Enamine screening collection that fit compound **III**'s pharmacophore model and received a better combination of the docking and ligand efficiency scores than the original compound **III**. They are shown in Figure 49.

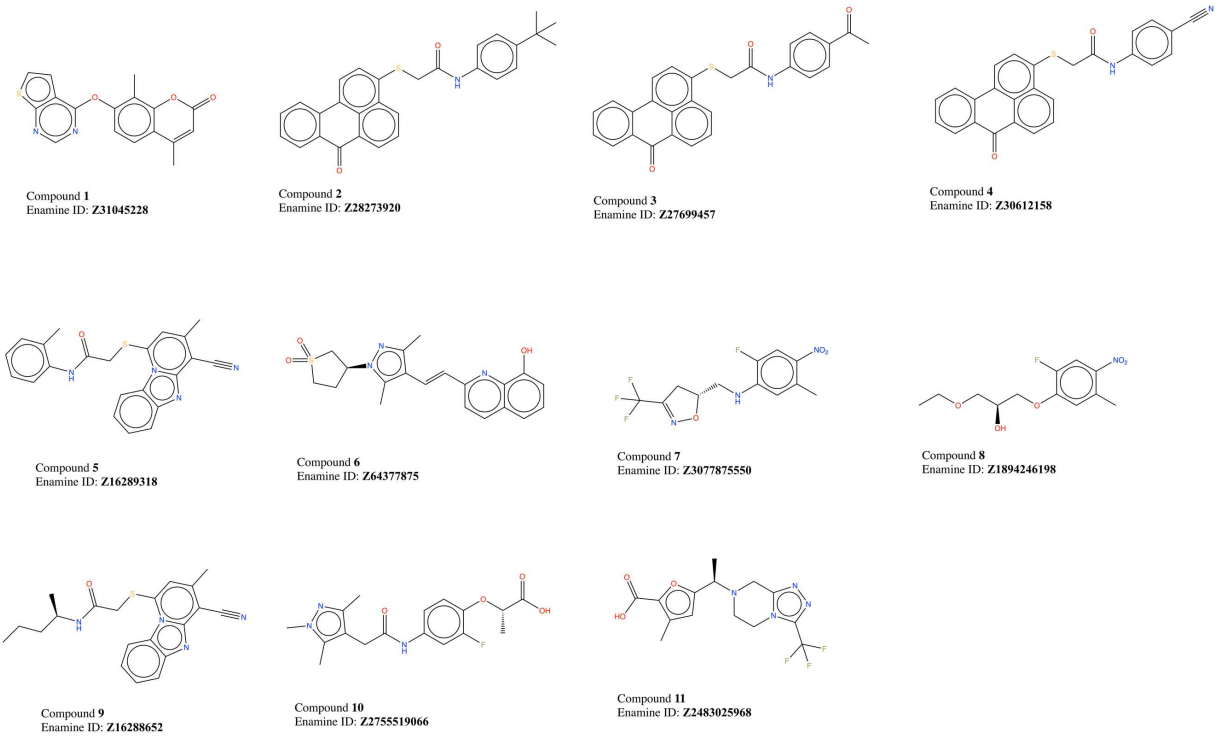


Figure 47. Eleven virtual hits found after docking the pharmacophore screening hits

2.3.7. Experimental validation of virtual hits

Two kinds of experiments were carried out with the 11 purchased virtual hits. The first involved X-ray crystallography, specifically the co-crystallization of these molecules with tubulin. This work was done by our collaborators from Dr. Andrea Prota's group at the Paul Scherrer Institut in Villigen, Switzerland. They performed soaking experiments using the T2R-TTL tubulin system, which consists of a protein complex comprising two bovine brain α,β-tubulin heterodimers, the rat stathmin-like protein RB3, and the chicken tubulin tyrosine ligase (TTL). The goal of this experiment was to see if any of the virtual hit molecules bound to the maytansine binding site of the tubulin protein.

The second test involved biochemical tubulin polymerization assays carried out by our colleagues at Dr. Fernando Díaz's group at the Centro de Investigaciones Biológicas Margarita Salas (CIB-CSIC) in Madrid, Spain. The goal of these assays was to examine the effect of these ligands on tubulin polymerization dynamics *in vitro*. Dr. Ahmed Soliman conducted the experiments by incubating the 11 ligands with tubulin purified from calf brains at 37°C. The formation of microtubules or other aggregates in solution was monitored by measuring the light absorption at 350 nm. The polymerization dynamics were tracked by measuring the absorbance at

350 nm for a duration of at least 4600 seconds. The tests included controls using tubulin with dimethyl sulfoxide (DMSO), the solvent used as a vehicle. The primary tubulin isotypes present in the tubulin preparation were tubulin βII (58%) and tubulin βIII (25%), with the remaining 17% consisting of other β-tubulin isotypes. In a typical bioassay of this kind, microtubule-stabilizing agents (e.g., paclitaxel) exhibit an increase in the amount of polymerized tubulin, reflected in an increased maximum slope and plateau in the absorption curve. Additionally, a higher number of nucleation events result in a reduced lag time, often too short to be detected in these experiments. Conversely, tubulin polymerization inhibitors (e.g., podophyllotoxin) show no increase in absorbance, indicating the complete prevention of microtubule formation.

Regarding the eleven virtual hits discovered through virtual screening of the Enamine HTS library, X-ray crystallography experiments did not demonstrate binding of any ligands to the maytansine binding site. However, two of the molecules (compound **2** and **3**) exhibited some inhibitory effects on microtubule polymerization. These molecules have favorable docking poses and alignment with the pharmacophore model (Figure 50). We hypothesize that these molecules may bind to the maytansine binding site, as predicted by computational methods, due to their binding mode aligning well with the pharmacophore features. Soaking these compounds into pre-formed crystals where the maytansine site is already occupied by the complementary tubulin monomer renders binding unachievable. For the binding to occur, it needs to transpire prior to the coupling of tubulin monomers, and the compound would, in essence, obstruct the formation of microtubules. However, when "mature" microtubule crystals are already formed, they cannot be disrupted by the compound. Additionally, the contacts formed by these two ligands as shown by the computational model are predominantly hydrophobic, which may contribute to their low affinity for the site, thus not being detected in the X-ray crystallography experiments.

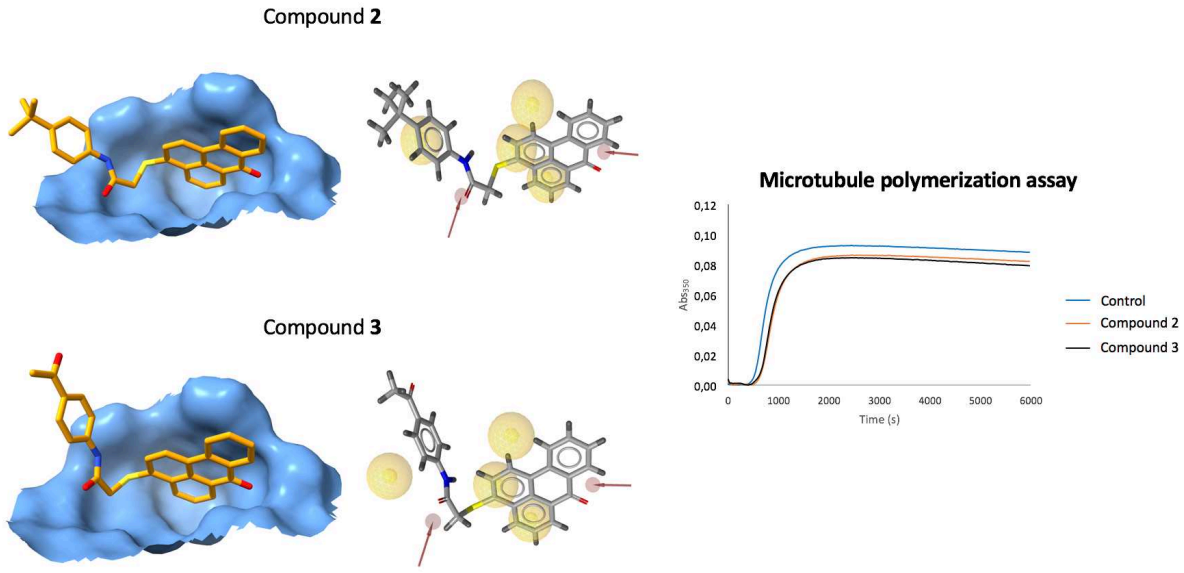


Figure 48. Two hits that show some microtubule polymerization inhibition activity

2.3.8. Results and discussion

In conclusion, the objective of this project was to identify inhibitors of tubulin polymerization that bind specifically to the maytansine binding site. We were particularly interested in discovering low-cost commercially available compounds, as these could be potentially modified and utilized for drug or molecular probe development. Through the screening of the Enamine HTS collection of 2.7 million compounds using a pharmacophore model derived from the maytansinol structure (7E4Z), we successfully identified 11 virtual hits. These hits underwent further experimental evaluation, including X-ray crystallography through co-crystallization with tubulin, and a standard tubulin polymerization assay.

The screening campaign yielded two molecules that exhibited inhibitory activity on tubulin polymerization, despite not being observed in X-ray structures. The observed effect on polymerization inhibition, although notable, was relatively modest. This could potentially be attributed to the sub-optimal affinity of these ligands for the binding site. Our computational analysis suggests that further optimization of these molecules may enhance their affinity to the site, potentially leading to the development of more potent binders. Exploring these possibilities in future research could pave the way for the design and synthesis of improved compounds for effective modulation of tubulin polymerization.

2.4. Structure-based de novo design

2.4.1. Docking-enabled forward synthesis-based de novo design

The complexity of current macrocyclic molecules that bind to the maytansine site stems from their origin as natural products, resulting in complicated synthetic processes, which hinders efficient exploration and exploitation of this binding site. Thus, we were interested to investigate the application of *de novo* design techniques in order to use their potential to generate novel molecules. Yet, a common drawback of such methods is that the generation of chemically valid molecules often overlooks the practical and financial aspects of their synthesis. This limitation underscores the need for a mechanism to control chemical feasibility in *de novo* design methods.

Addressing this challenge involves the introduction of a synthetic accessibility score into the *de novo* design pipeline. One way to do it is by using a forward synthesis approach. Forward synthesis prediction task (also known as reaction prediction task) is concerned with finding a synthetically valid chain of chemical reactions applied to a limited number of readily-available chemical building blocks that produce the required target molecule, thereby generating its synthetic tree. This contrasts with a more commonly used retrosynthetic approach, where the target molecule is sequentially broken down into small purchasable fragments. Both forward and retrosynthesis techniques can be implemented to assist in *de novo* molecule generation. In this

work, we attempted to implement a structure-based *de novo* molecule generation pipeline, which made use of an in-house tool capable of solving a “forward” reaction prediction task, central to which is the concept of a synthetic tree.

In this digital construct, nodes represent individual molecules, and edges are chemical transformations that link them. When the user inputs a target molecule (for which they would like to produce a synthetic route), the software performs a similarity search in the provided building block database. It then samples a user-defined number of building blocks as potential starting points for the synthetic tree. For each possible starting block, a pre-trained neural network predicts the chemical transformation and another building block that can be applied to yield a product structurally similar to the target molecule. The Tanimoto metric is employed to define such similarity. This process is conducted for all building blocks, assigning a special "reward" value to each node. This value measures the suitability of a given node to successfully grow towards the target molecule. At each step, the tool selects the node with the highest "reward" value to progress the route, thereby growing the synthetic tree. The tool's goal is to reach the target compound by maximizing the similarity of the products produced by the nodes. Alternatively, the process can be halted upon reaching a user-defined number of synthetic steps (i.e., number of applied chemical transformations). This approach offers the advantage of producing either the required molecule exactly, along with a complete synthetic pathway from the provided starting material and allowed reactions, or molecules highly similar to the target molecule that may still possess the required properties.

To ensure the molecules are specifically designed for the binding site, we developed a genetic algorithm-based protein-ligand docking implementation of this approach, re-configuring the forward synthesis approach (Figure 51). Instead of generating the target molecule, the forward synthesis tool uses it as a reference. The objective of the synthetic tree building is to optimize both the docking score of the products in each node and similarity to the reference ligand. The latter requirement limits the tool to explore the drug-like compound space. The selection of fragments and chemical transformations on each step is guided by a genetic algorithm. This leads to the generation of molecules that are (1) similar to the known ligand up to a user-specified threshold similarity value, thereby ensuring the drug-likeness of the generated molecules; and (2) yield a good docking score in the binding site. Guiding the process by docking score produces molecules with enhanced predicted affinity to the site. Reviewing the chemical transformations that generate the best-docking score molecules allows us to reconstruct their synthesis pathway. Consequently, this tool serves both as a structure-based *de novo* design and forward synthesis instrument.

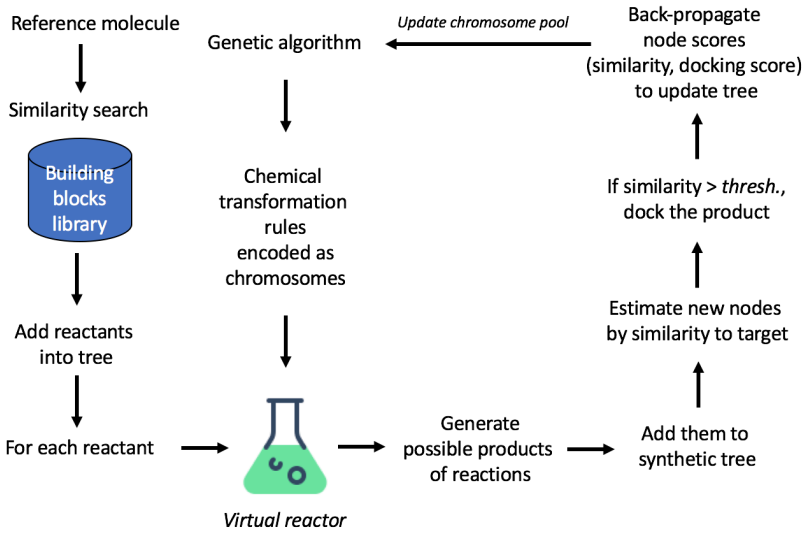


Figure 49. Overview of suggested approach to structure-based forward synthesis-guided de novo molecular design

2.4.2. Computational setup

The tool setup for this project required several components: (1) a dataset of chemical transformations; (2) a dataset of building blocks; (3) a reference ligand structure; and (4) a protein binding site. The dataset of chemical transformations came from our colleagues from the Kazan Federal University, Russia (Dr. Timur Madzhidov's group), who mined the dataset of chemical transformations from the USPTO dataset[96], which included 550 one-component rules and 1700 two-component rules. The building block dataset comprised of 390,000 building blocks from the ZINC database (v. 12), which were standardized using a procedure outlined in previous sections. We used disorazole as the reference ligand to guide the chemical design towards molecules that are likely to bind effectively to the targeted binding site. The binding site was defined as all atoms within 12 Å from the center of mass of disorazole from the 6FJM PDB structure, with the protein prepared by the SPORES software. All water molecules, ions, and other small ligands were removed. The genetic algorithm parameters included a starting population size of at least 50, number of generations set at 50,000, crossover probability of 0.7, an elitism fraction of 0.1, and the number of tolerated generations without fitness score (i.e., docking score) improvement as 50,000. Compounds encoded by individual chromosomes had to have at least 0.4 Tanimoto similarity score to the reference ligand to be docked and considered in the optimization process.

2.4.3. Results of the generation

The *de novo* design pipeline was run thrice to account for the variability in the docking scores and genetic algorithm that predicts the necessary building blocks/chemical transformations. The structure-based *de novo* generation of molecules tailored for the maytansine binding site

resulted in three small molecules. Owing to the unique feature of our software, we could extract the precise sequence of building blocks and chemical transformations that produced the molecules with the highest docking score in the binding site. The generated compounds and potential synthetic routes are presented in Figure 52, with compounds highlighted in green color being commercially-available starting building blocks as seen in the ZINC v. 12 building blocks subset. Evaluation of the best docked poses of these molecules and alignment with the pharmacophore model of disorazole revealed overlapping key features (Figure 53).

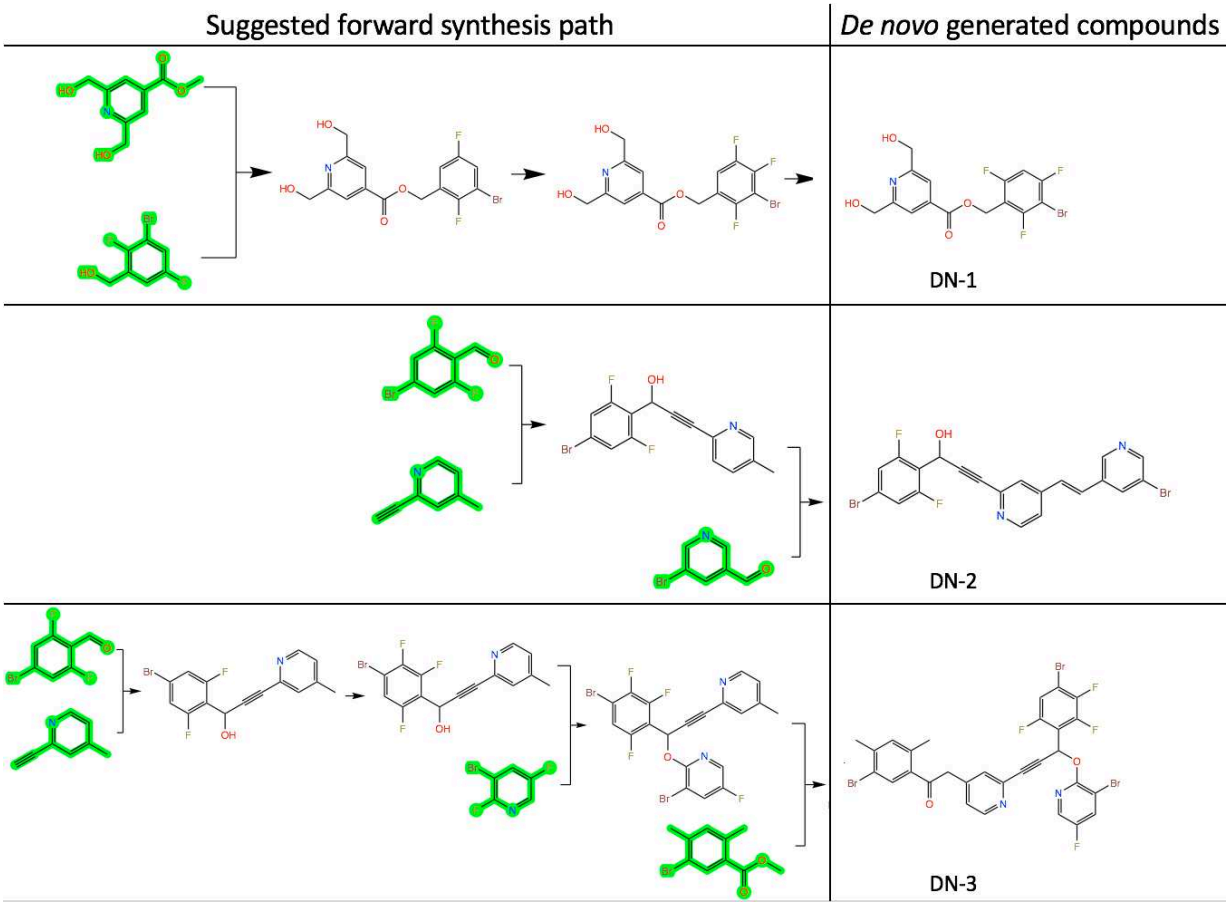


Figure 50. Three de novo generated compounds to target the maytansine binding site with synthetic pathways simultaneously suggested by the forward synthesis tool

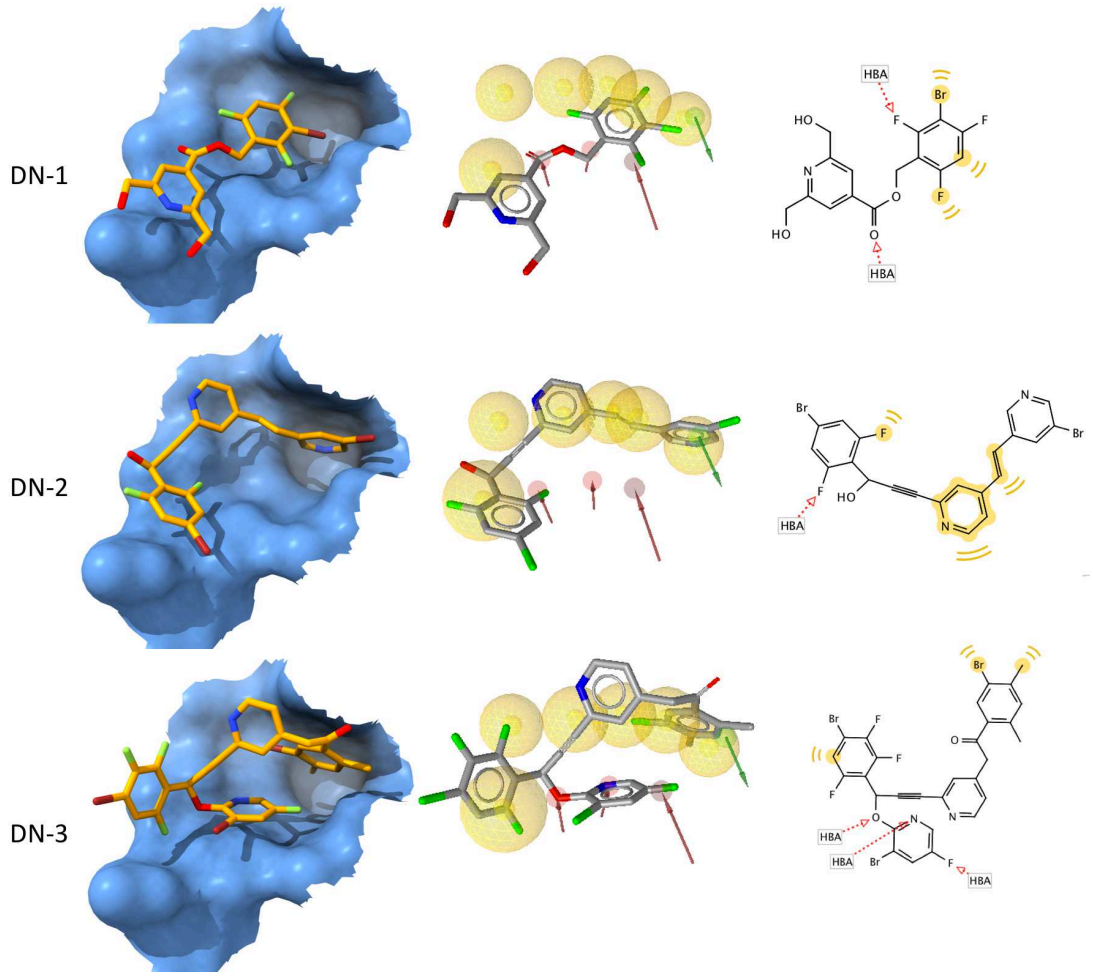


Figure 51. Three de novo generated ligands and their predicted overlap with pharmacophore features according to the docked pose with the lowest docking score value

2.4.4. Comparison to retrosynthesis tools

We aimed to determine if retrosynthesis tools could replicate the routes produced by the forward synthesis tool. For this comparison, we utilized AiZynthFinder, SciFinder (a proprietary web-based tool), and Spaya (a proprietary web-based tool with a time-limited free access option).

The ChemPlanner tool, now part of the proprietary SciFinder^n suite, operates in a similar manner to the AiZynthFinder (by breaking a molecule down into purchasable pieces according to certain chemical rules), but uses a different database of chemical transformation rules and purchasable compounds. Its features are the two options to (1) specify the first bond to break during the first step of the retro-synthesis procedure, or (2) protect some bonds from breaking (thus letting the user define a general direction for further retrosynthesis). It also can utilize three different databases of chemical transformations: Common (with commonly-used reactions), Uncommon and Rare (both including rare reactions with little examples in the literature, which may potentially lead to a more creative synthetic pathway being suggested). Additionally, when the retrosynthesis is finished, the user can examine the literature references that affected the selection, thus making the tool interpretable. Finally, if a user doesn't like a certain part of the

route, he can select an alternative one from a list of compatible suggestions. It is also worth noting that the tool can be used straight from the web browser and does not require any software installation or configuration. Moreover, this tool is capable of generating a detailed report for future use, and calculates an approximate cost of the synthesis.

Spaya is another proprietary retrosynthesis tool, developed by IKTOS. Unlike SciFinder^n, Spaya offers a free 1-month trial for individuals. Under the hood, it uses a neural network approach to molecule breakdown, much like the AiZynthFinder tool. The database of chemical transformations it uses is called Pistachio and comes from NextMove Softwares. Mcule provides its library of commercial starting materials and updated data related to their commercial availability and pricing. Among the features of this tool, one can mention the possibility of specifying specific intermediate molecules' SMILES prior to starting the retrosynthesis job, which may help the experienced user direct the retrosynthesis in a particular direction. When the job is started, the tool suggests several groups of possible routes, different in the bonds of the target molecule that they break on their first step. Each route is assigned a special score. The higher the score the better. As is the case with the SciFinder^n tool, if a user doesn't like one part of the proposed retrosynthesis route, he can selectively modify this part by selecting alternative routes from a list of suggestions. For each reaction in the route, it is possible to see similar reactions described in the literature – thus making the tool interpretable. It is also worth mentioning that all interaction with this tool happens via a browser-based GUI. Another interesting feature of Spaya is the availability of the API access, which means that the retrosynthetic accessibility of compounds of interest can be assessed in batch mode and without any installation, using the IKTOS servers to do the job. Moreover, this tool is also capable of generating a PDF with a detailed report of the route.

Below are the retrosynthesis paths generated by the respective tools for compounds DN-1 (Figure 54), DN-2 (Figure 55), and DN-3 (Figure 56). Molecules highlighted in green in these images are commercially-available starting building blocks.

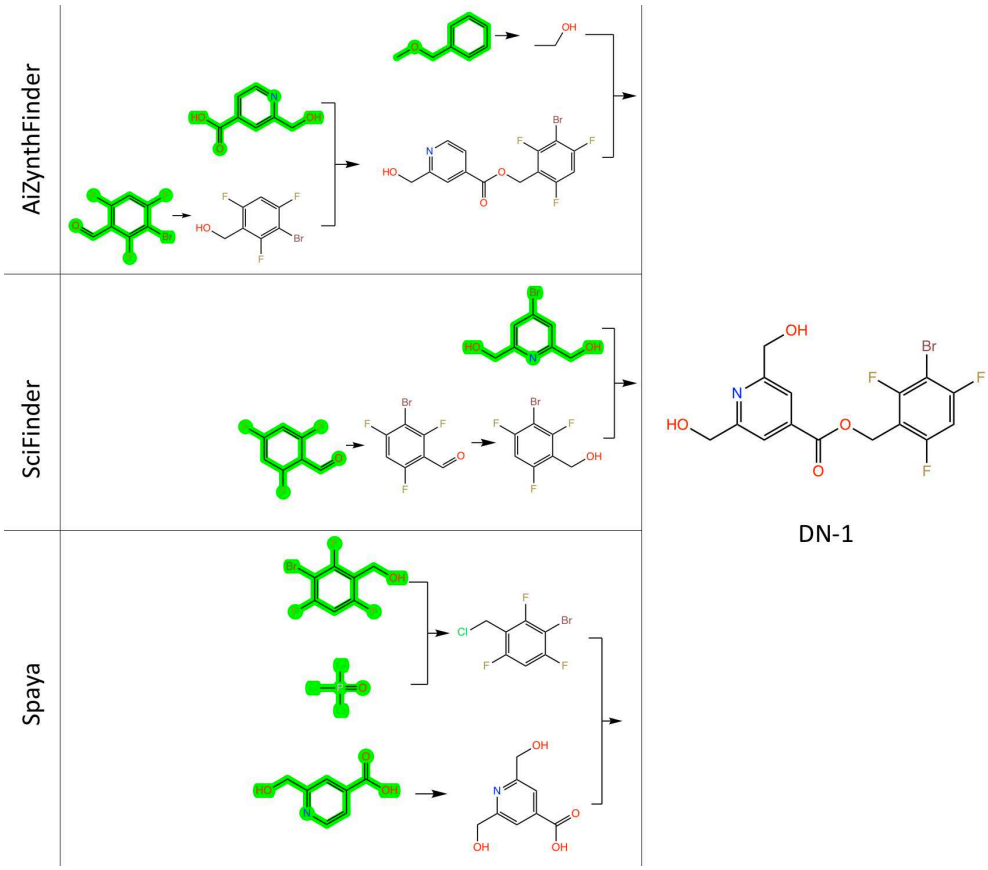


Figure 52. Retrosynthetic paths generated for compounds DN-1

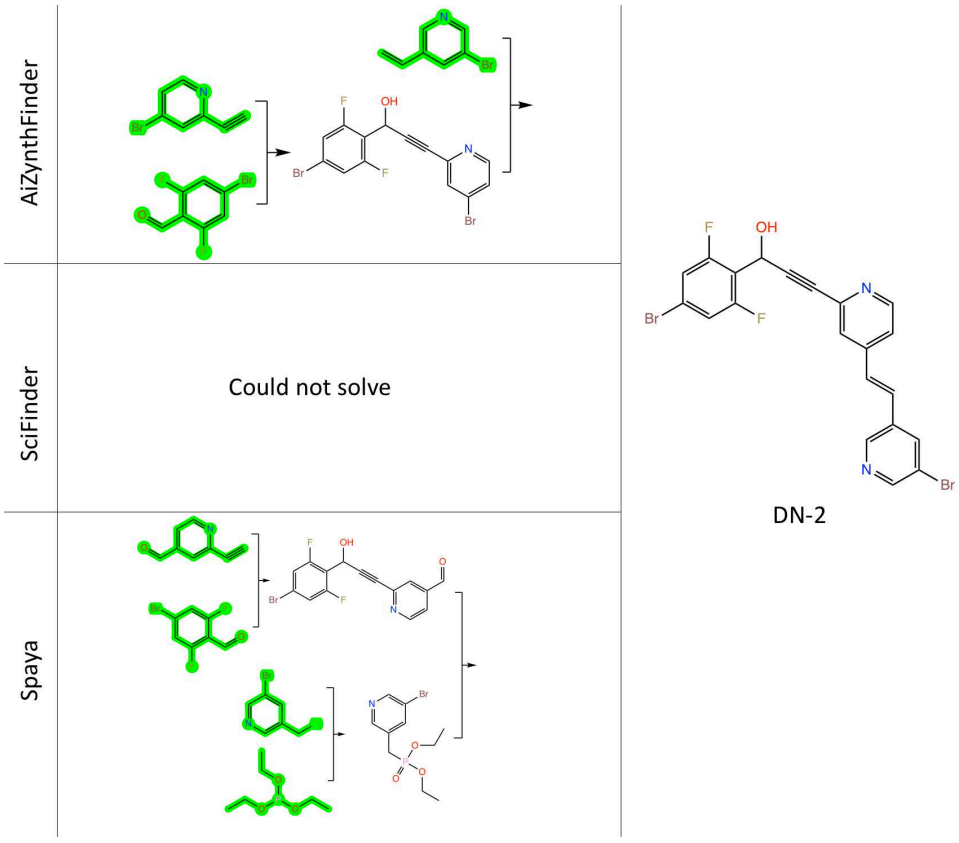


Figure 53. Retrosynthetic paths generated for compound DN-2

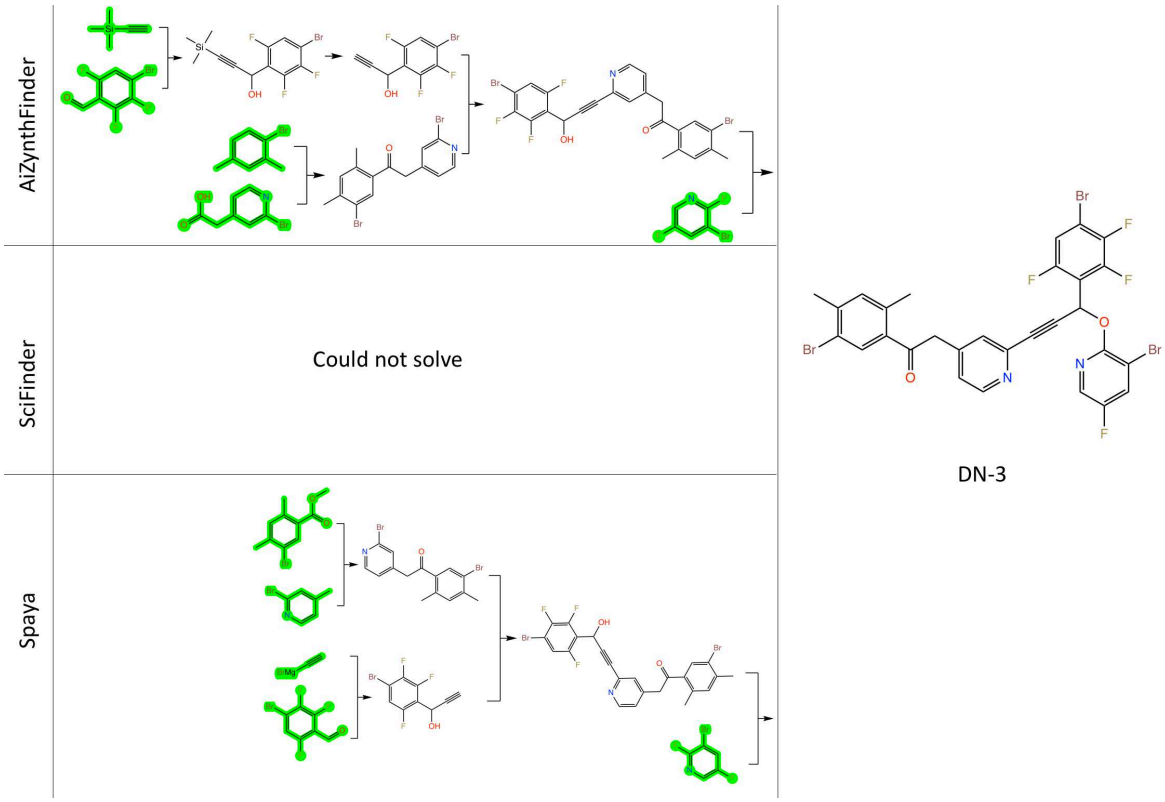


Figure 54. Retrosynthetic paths generated for compound DN-3

2.4.5. Results and discussion

In conclusion, in this project we successfully developed a structure-based *de novo* design pipeline, utilizing a forward synthesis prediction tool as a guiding limitation of synthetic accessibility. This novel implementation ensures the creation of molecules tailored for the maytansine binding site, with interactions with key residues, while also ensuring accessible synthesis. Notably, we can discern the exact reactions required for their synthesis. Upon comparison with routes generated by retrosynthesis tools for the *de novo* created molecules, we could draw insightful conclusions.

As such, the forward synthesis tool generally produces shorter synthetic paths than all of the retro-synthesis ones. However, sometimes the forward synthesis tool suggests sub-optimal routes. For example, in the case of compound **1**, the forward synthesis tool suggests using a reagent which then requires to perform complex rearrangements of fluorine and bromine atoms in the aromatic ring. Instead, it could have advised to use a bit different starting reagent, which would have been a better (easier) way. However, in the context of *de novo* design, the precise substituent pattern should not be taken as an absolute must – unless those substituents are providing (predicted) key interactions with the target, unlike in the case of the herein discussed halogens on the phenyl ring. All those analogues are roughly equally similar to the target product and dock comparably well – picking the synthetically easiest would have been the rational way to follow.

Unfortunately, that was not an option in the context of TubInTrain, because of insufficient resources. This computational experiment was not validated – showing again that in theory the exploration of the whole chemical space is appealing, but in practice exploring the chemical space of compounds for which the partner chemists have the (cheap) starting materials in stock and perfectly master the synthesis protocols is much likely to conclude with experimental validation.

Interestingly enough, we have observed that the AiZynthFinder tool in its default configuration tends to break already purchasable intermediate compounds further down into much smaller starting molecules. It also tends to break larger molecules down to small molecules like ethanol and acetylene. This leads to artificially more complicated synthetic paths. One of the reasons for that may be due to the contents of the default database of building blocks.

Generally, all of the retrosynthetic tools tend to separate larger molecules into 2 or more smaller parts, and then find ways to (1) separately synthesize them in a parallel fashion; (2) unite these parts together. On the contrary, the forward synthesis tool works by consequentially growing a starting fragment by enlarging its carbon chain, and everything is done in a sequential way.

Neither software recommends using protective groups, assuming the suggested reactions happen at required atoms, and not at other, un-protected ones. In some cases, all tools proposed similar reactions (of the same type) using slight modifications of the same starting reagents, once again highlighting the importance of the used building block database (e.g., compound 1, 2).

Thus, it's important to note that the effectiveness of our approach largely depends on the quality and extent of the databases of building blocks and chemical transformations available. Furthermore, the performance of the docking software is also crucial.

Despite these dependencies, our method has shown considerable utility in designing molecules targeting specific sites. Its uniqueness lies in its guidance by a forward synthesis tool, rather than a retrosynthesis tool, thus offering a distinctive perspective in the field of *de novo* drug design.

2.5. Conclusions and perspectives

In conclusion, we have performed two virtual screening campaigns to discover novel maytansine site-targeting agents - a combined approach only previously employed by one other study. The first campaign involved screening ChEMBL to identify a cytotoxic molecule with an unknown mechanism of action that could potentially be attributed to binding at the maytansine site. We did identify a natural product with cytotoxic action, although challenges in synthesizing this product led us to identify possible alternatives that would be easier to synthesize. Their synthesis and experimental validation remain to be done.

The second screening campaign, conducted with an Enamine HTS library, yielded 11 virtual hits, two of which showed moderate microtubule inhibiting action. While X-ray crystallization did not provide clear results, we believe the observed inhibition can be attributed to binding at the maytansine site due to their structural features and size. Although the exact binding dynamics remain unclear, these molecules present an exciting opportunity for further research, including potential modifications to develop new probes or inhibitors of tubulin heterodimer interactions.

Our third project marked the successful development of a *de novo* molecule design pipeline, allowing for the creation of chemically viable molecules specifically tailored for the maytansine binding site. By coupling a forward synthesis tool with a protein-ligand docking tool, we ensured that the generated molecules would be chemically valid and accessible via computed synthetic routes. The synthetic routes produced by our software were comparable to those generated by freely and commercially available retrosynthesis tools.

Despite these advances, our work also highlighted potential areas for improvement. Further research is necessary to fully capitalize on the successes of this study and refine our approach. While there is more work to be done, the promising results from our virtual screening campaigns and the development of a novel *de novo* design pipeline lay a solid foundation for future investigations.

Chapter 3. Virtual screening for novel pironetin site-targeting inhibitors of tubulin polymerization

3.1. Introduction

Disruption of longitudinal interactions between tubulin heterodimers is an effective strategy to inhibit the polymerization of microtubules, as seen in the mode of action of maytansine site-targeting agents and vinca alkaloids, who act as a "wedge," physically impeding the T7 loop and H8 helix of α-tubulin from locking themselves in a special cavity on the β-tubulin subunit[5]. However, a similar destabilizing effect may be achieved if a small molecule interferes with a-tubulin's T7 loop and H8 helix directly.

Indeed, such a molecule exists. It is pironetin, a natural product isolated in 1994 from the fermentation broths of *Streptomyces prunicolor* PA-48153 and *Streptomyces sp.* NK 10958[5,97] (Figure 55). Initially discovered as a plant growth regulator, it was later found to exhibit potent antiproliferative activity, thanks to its capability to impede microtubule dynamics. This ability gives it powerful *in vitro* activity against cell lines both sensitive and resistant to first-line therapeutics, and even those resistant to other microtubule-targeting drugs[42,43,98]. Pironetin is a dihydropyrone derivative and several studies have shown that its α,β-unsaturated lactone core fragment is essential for its microtubule inhibitory activity. The alkyl chain and the hydroxyl group at the 7-position are also important for the inhibition of tubulin polymerization[99].

Binding of pironetin disturbs the H8 helix and T7 loop of α-tubulin, disrupting longitudinal tubulin-tubulin interactions[5]. The molecule binds to an extended hydrophobic pocket on α-tubulin, interacting with strands S8, S10, and helix H7 (Figure 55). The sidechain of pironetin burrows further into a pocket shaped by helix H7 and strands S4, S5, and S6. This interaction disorganizes the T7 loop and provokes a conformational change in the N-terminal section of helix H8 of α-tubulin, with some residues shifting more than 10 Å[5,100]. The binding pocket is not present in the *apo* structure, hinting that pironetin may bind through an induced fit mechanism[42]. The inhibition was discovered to be essentially irreversible under physiological conditions[43].

A unique feature is that pironetin forms a covalent bond with a cysteine residue upon binding to the pocket. The reaction follows a Michael addition mechanism, made possible by the nucleophilic cysteine residue (Cys316) located close to ligand's position the binding site and the presence of an α,β-unsaturated carbonyl fragment (δ-lactone) in pironetin[43]. This covalent-binding action was confirmed by structure-activity relationship studies and verified by two independent X-ray crystallography studies, which identified Cys316 of α-tubulin as the reactive residue[5,42]. It is worth noting that the binding site contains three more cysteine residues in the vicinity of the ligand's binding pose (Figure 55).

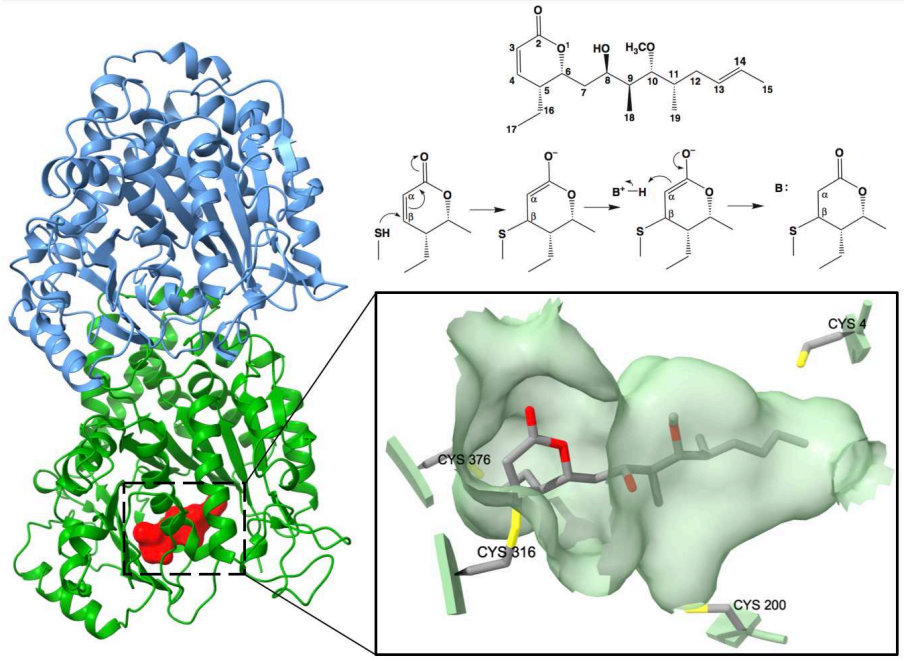


Figure 55. Pironetin binding site

Remarkably, until recently, pironetin was the only ligand characterized crystallographically to bind solely to α-tubulin[42]. This has prompted numerous groups to pursue total syntheses of this compound[101]. To date, 12 unique total syntheses of pironetin have been published, underlining both the synthetic challenges and the biological intrigue associated with this natural product[42].

The complex nature of pironetin, with its six stereo enters, results in lengthy and impractical synthesis processes for large-scale production[101]. Even the synthesis of its simplified analogues remains complex, with analogues showing a significant reduction in cytotoxicity compared to pironetin[42]. Moreover, *in vivo* studies of pironetin revealed a poor efficacy and potentially toxic side effects in mice, possibly related to potential off-target binding, making it unsuitable as a direct drug candidate[98]. Additionally, the structure-activity relationships of pironetin analogues are not easily rationalized, complicating the design of new generations of these analogues[43]. Little to no computational studies on the pironetin binding site-targeting molecules have been published so far. One computational study by Banuelos-Hernandez et al. used density functional theory computations, protein-ligand docking and molecular dynamics simulation to provide a rationale for the pironetin inhibitory activity[97]. Vergoten et al. established a list of potent cytotoxic agents that were structurally similar to pironetin and used protein-ligand docking to speculate that their mode of action is similar to that of pironetin[99]. No virtual screening attempts for this site have been published.

Nonetheless, we were interested to explore the pironetin binding site due to it being relatively understudied in the realm of cancer therapeutics and molecular probe design[42,99]. The binding site, nestled in the α-tubulin, represents a potentially fruitful area of drug design, especially due to β-

tubulin-targeting agent resistant mutations in tubulin in cancer isotypes[42]. Currently, the consequences and benefits of targeting α-tubulin instead of, or in conjunction with, β-tubulin remain largely unexplored. The lack of pironetin-resistant cell lines suggests that pironetin or α-tubulin binders could serve as useful probes to understand β-tubulin resistant cancers and show utility against tumors that have become resistant to first-line tubulin-binding chemotherapeutics[42]. Additionally, a molecular probe that would covalently bind to a cysteine residue within the pironetin binding site may be useful in the design of high-throughput surface-based tubulin binding assays, using such methods as surface plasmon resonance or wave-guided interferometry assay[102]. A key challenge for these surface-based assays lies in the immobilization of tubulin onto a matrix in a manner that facilitates small molecules interactions[103]. Current approaches are sub-optimal, often tampering with the useful signal[103]. However, combining the high sensitivity of surface-based assays with a properly immobilized, binding-competent tubulin through an easily accessible small molecule that binds to the pironetin binding site could potentially allow for the high-throughput determination of binding affinities of a wide array of small molecules. This would greatly accelerate the discovery and development of novel microtubule-targeting agents with high specificity and potency.

Thus, the goal of this work was to perform virtual screening of large compound collections to find more easily synthetically accessible molecules that could target the pironetin binding site, and, potentially, create a covalent bond with one of the cysteine residues in the binding site.

3.2. Virtual screening of the ChEMBL library

This project was aimed at discovering new small molecules that could bind to the pironetin site. We intended to expand the variety of molecules that could target this site, as it is presently known to be targeted by just one molecule, pironetin itself. With the moderate success of the pharmacophore-based virtual screening strategy as described in Chapter 2, we decided to adopt a similar pharmacophore screening approach to start this project.

3.2.1. Overview of available data

The crucial question was to select a starting protein-ligand complex to develop a pharmacophore model. The RCSB PDB database listed only two crystal structures of tubulin co-crystallized with pironetin, namely 5FNV and 5LA6. Both of these structures showed pironetin bound to Cys316 of the α-tubulin subunit in an identical conformation. Nevertheless, 5LA6 was not well-resolved around the site, as a significant portion of the T7 loop was missing. As a result, we selected the 5FNV PDB structure as our source of protein-ligand interaction data.

3.2.2. Pharmacophore modelling

An automatically-derived pharmacophore model was developed using the LigandScout software, based on the 5FNV structure (Figure 56). The model comprised 5 features: four hydrophobic spheres aligned linearly and 1 residue bonding point feature, which detects reactive groups essential for forming a covalent bond with a nucleophilic residue. Such groups include ketones, nitriles, or Michael acceptors, which were of particular interest to us. Given that only one ligand has been confirmed to bind at the pironetin site, we could not validate the model.

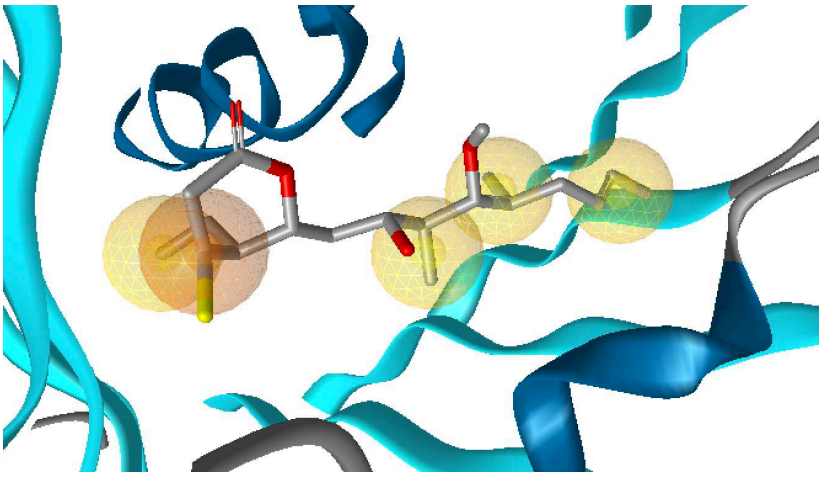


Figure 56. Pironetin's pharmacophore model derived from the 5FNV PDB structure

We redocked pironetin to verify if our software could replicate its pose within the site, anticipating that this could imply that other compounds adhering to this pharmacophore model would be docked accurately. Our redocking process involved breaking the covalent bond between pironetin and the cysteine residue and docking the non-bound form of the ligand. The docking procedure was executed using the PLANTS software. The ligand's binding site was derived from the 5FNV structure, which was prepared by removing all solvent molecules, ions, and other small organic molecules. The ChemAxon conformation sampling software within the MolConvert program calculated a random initial ligand's conformation from a stereoisomeric SMILES string representation of pironetin. Both protein and ligand structures were pre-processed by the SPORES software. The outcome was satisfactory as the molecule redocked correctly with an RMSD of 1.81 Å (Figure 57) with respect to the covalently bound form, expected to be offset from the (experimentally undetectable) non-covalent pose. The RMSD value was calculated using the CalcLigRMSD script, contributed to the RDKit chemoinformatics toolkit by Velázquez-Libera et al.[104]

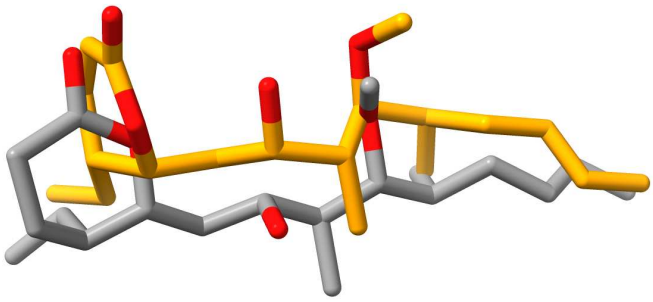


Figure 57. Grey – pironetin's native pose, orange – re-docked pose

3.2.3. Screening library preparation

We first screened the ChEMBL library (v. 26), as our initial preference was to work with molecules that already had established bioactivity data. This could aid in identifying potential cytotoxic agents without the need for complex experiments. The screening library preparation procedure followed the same steps as described in Chapter 2, section 2.2.4. Briefly, all molecules were first standardized using an in-house standardization routine (see section 2.2.4). Secondly, we calculated 200 conformations for all molecules in the library using the iCon-best option in the iCon built-in conformation sampler of LigandScout, with a 0.7 Å RMSD window between individual conformations.

3.2.4. Pharmacophore screening

We then applied the pironetin pharmacophore model for screening the compounds. We used the "Pharmacophore fit" scoring function to select compounds. The "Match at least 3 query features" screening mode was used. The retrieval mode was set to "Stop after first matching conformation" to speed up the screening process. Excluded volume clashed were accounted for during screening.

The result yielded 9229 virtual hits with pharmacophore fit score values higher than 54.69 due to the relative simplicity of the model. However, to proceed with the selection of virtual hits, we had to devise additional filtering methods to enhance the quality of the hits and improve their suitability for the binding site.

3.2.5. Protein-ligand docking

To do so, we docked all 9229 virtual hits into the pironetin site derived from the 5FNV structure using the PLANTS software. The binding site was defined as all atoms within 12 Å radius from the center of mass of the native ligand of the 5FNV structure, pironetin. Each compound was evaluated by a *chemplp* docking score and 4 ligand efficiency scores, which reduced the influence of the ligand size (in terms of atom numbers) on the docking score. The ligand efficiency scores

are derivatives of the docking score, defined in section 2.3.7. For each compound, we generated 10 docked poses, followed by Pareto optimization over the docking score of the best pose (i.e., the one with the lowest docking score) and ligand efficiency scores. This was to identify ligands that could yield highest docking scores with a lesser number of atoms. Post-optimization, 190 molecules were shortlisted. For all these molecules, we computed the synthetic accessibility score (SAscore) and ranked them accordingly. The SAscore, proposed by Ertl et al., is a combination of fragment contributions and a complexity penalty[105]. Fragment contributions were computed by examining a vast collection of representative molecules sourced from the PubChem database. The score accounts for molecule size, the presence of unconventional structural elements, including sizable rings, non-standard ring fusions, and stereocomplexity. It ranges from 1 to 10, where 1 signifies ease of synthesis, while 10 indicates a high level of complexity in synthesis. In this work, we used the SAscore implementation included in the RDKit chemoinformatics toolkit for the Python programming language.

As a result, we selected 10 molecules that had a docking score better than pironetin, largely adhered to the pharmacophore model, and had a favorable synthetic accessibility score of less than 2. They are shown in Figure 58. Calculated pose and pharmacophore feature overlap of a virtual hit with the lowest (i.e., best) value of the docking score is shown in Figure 59. Noticeably, the outcomes of this virtual screening campaign presented several molecules with warheads placed in molecular contexts where their reactivity is dubious, hinting that the structural match is a more important criterion for the used software than chemical reactivity. Moreover, it appeared as though the software prioritized overlap with the hydrophobic feature adjacent to the reactive group feature. For certain virtual hits, the reactive group was situated at a greater distance from the targeted cysteine residue than anticipated, suggesting an unexpected preference pattern within the screening process.

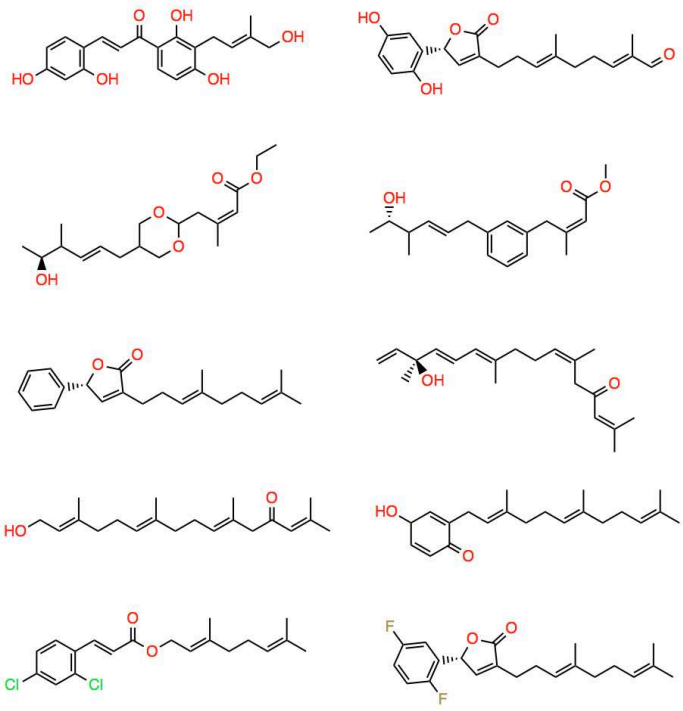


Figure 58. Ten virtual hits found by pharmacophore screening of the ChEMBL library

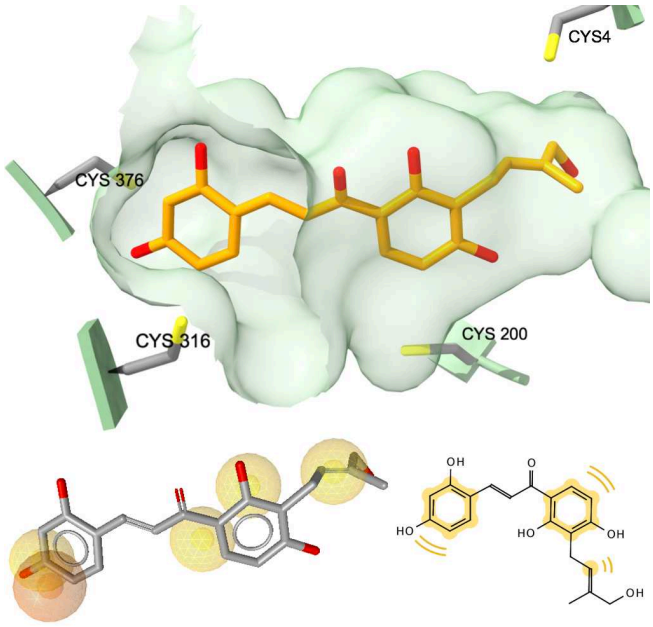


Figure 59. The calculated docked pose and pharmacophore feature overlap of the virtual hit with the best docking score

3.2.6. Results and discussion

Upon checking these 10 molecules in the ChEMBL database, it was found that none of them, despite being part of cytotoxic assays, had any significant cytotoxic effect. Consequently, we couldn't leverage these results to progress further and had to abandon the initiative. Possible future avenues could include refining the pharmacophore model to make it more complex or exploring different docking approaches. One potential alternative could be ensemble docking,

which takes into account the flexibility of the binding site and the induced-fit binding mechanism suggested by certain publications. It could also be that the database does not contain any cytotoxic molecules that cause the antiproliferative action via binding at the pironetin site.

3.3. Virtual screening of the Enamine libraries

When the screening of the ChEMBL library yielded no significant outcomes, we pivoted our approach to focus on the Enamine libraries of purchasable compounds.

3.3.1. Pharmacophore modelling

Our strategy for this project continued to use the pharmacophore model previously employed during the ChEMBL database screening. Moreover, we leveraged new findings from a fragment screening campaign conducted by our collaborators[106]. This campaign uncovered three small fragments with a propensity to bind to the pironetin binding site. From these protein-fragment complexes, we derived additional pharmacophore models. While two of the fragments expectedly produced straightforward models with a single feature, one fragment (2-chloro-N-methylbenzene-1-sulfonamide, PDB code: 5S5M) presented a more complex model containing three features (Figure 60).

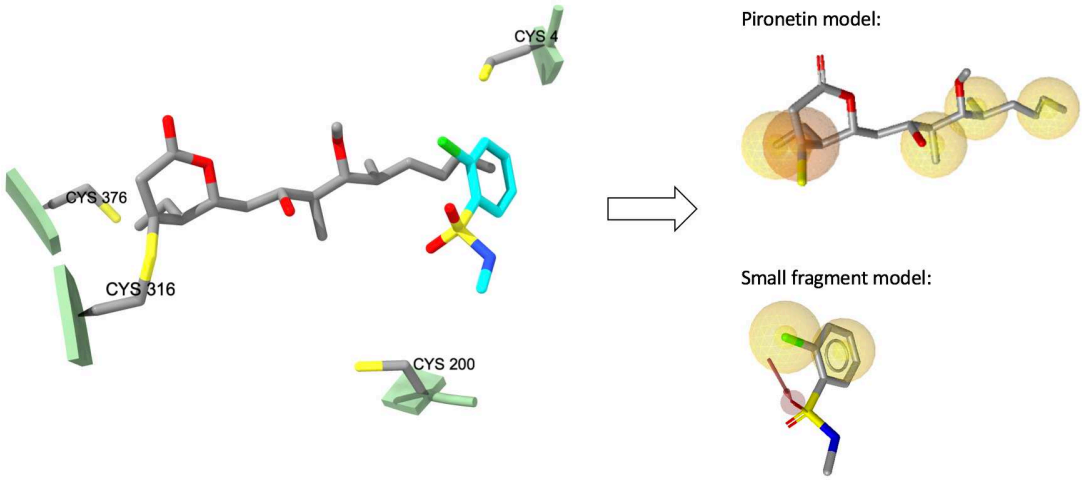


Figure 60. Gray – native pose of pironetin (5FNV). Cyan – crystal pose of the small fragment (5S5M). Both were used to derive a pharmacophore model.

Consequently, the larger pironetin model was used to screen libraries of larger purchasable molecules, approximating a molecular weight of 500 Da, due to its features residing in an elongated 3D shape. In contrast, the small fragment model, with closely situated features, was utilized to screen libraries of small fragments, with an approximate molecular weight of 200 Da.

3.3.2. Screening libraries preparation

The Enamine HTS compound library (2,688,748 compounds), covalent probes library (960 small fragments), tubulin-targeted library (3,452 compounds), cysteine-targeted library (3,200 compounds), phenotypic screening library (5,760 compounds), covalent screening library (11,760 compounds), 3D shape diverse library (1,200 small fragments), NP-like fragments (4,160 small fragments), covalent fragments (7,360 small fragments), DSI-poised library (860 small fragments), high fidelity fragment library (1,920 small fragments), and the essential fragment library (320 small fragments) were all used in this study. These libraries were all standardized using a previously described method. For the HTS library we calculated 25 conformations using the iCon-fast option of the built-in conformational sampler iCon, as part of the LigandScout software. For all other libraries, we computed 200 conformations using the iCon-best option of the same software.

3.3.3. Pharmacophore screening

We conducted the screening process using the LigandScout software, employing the "Pharmacophore fit" scoring function. This score reflects the number of matched features and the RMSD of their positions relative to the feature sphere's center. For the screening process, we used the "Match at least 3 query features" screening mode and set the retrieval mode to "Stop after first matching conformation" to make the process efficient. To enhance the screening accuracy, we activated the check for excluded volume clashes. The libraries screened with the pironetin model and the small fragment model produced 2340 and 4932 virtual hits, respectively (Table 1). These hits were then filtered to retain only those ligands best suited for targeting the pironetin binding site.

Table 1. Libraries screened by pironetin and small fragment's pharmacophore models

Library	**Number of records**	**Model**	**Pharmacophore screening hits found**
Enamine HTS compound library	2688748	Pironetin	2176
Covalent screening library	11760	Pironetin	120
Covalent fragments	7360	Small fragment	40
Phenotypic screening library	5760	Pironetin	21

NP-like fragments	4160	Small fragment	1507
Tubulin-targeted library	3452	Pironetin	12
Cysteine-targeted library	3200	Pironetin	11
		Small fragment	1693
High fidelity fragment library	1920	Small fragment	645
3D shape diverse library	1200	Small fragment	644
Covalent probes library	960	Small fragment	0
DSI-poised library	860	Small fragment	312
Essential fragment library	320	Small fragment	91

3.3.4. Protein-ligand docking

To refine our results, we conducted protein-ligand docking of the virtual hits identified within the pironetin binding site. We maintained the same binding site definition and docking parameters as outlined in section 3.2.6. For molecules identified by the pironetin pharmacophore model, we evaluated the best poses in the sites and performed Pareto optimization over the docking score values and the ligand efficiency score values. Then, we focused on the distances between the reactive functional groups of the remaining virtual hits and the cysteine residues in the binding site. Then, we manually computed distances to the four cysteine residues using the measurement wizard in the ChimeraX software. The top molecules were then re-ranked based on the distance values to the cysteine residues in the site, with emphasis on the distance value rather than the specific residue targeted.

For the virtual hits identified by the small fragment model, we ranked them based on the docking score and through visual inspection of their preferred position within the site, prioritizing those situated closer to any cysteine residue.

Our concerted efforts resulted in the identification of 32 promising virtual hits. 27 of these were derived from the screening with the pironetin model (eleven from the Enamine HTS collection (Figure 61), three from the tubulin-targeted library (Figure 62), and thirteen from the cysteine-targeted library separately found by two models (Figure 63)), and the remaining 5 were generated from the screening with the small fragment model (the electrophile covalent probes library (Figure 64)). Once again, for some of the virtual hit molecules, it seemed that the software

displayed a tendency to prioritize the overlap with the adjacent hydrophobic features rather than the reactive group feature. In some instances, the placement of the reactive group was more distanced than what was initially expected, implying an unanticipated bias within the screening algorithm.

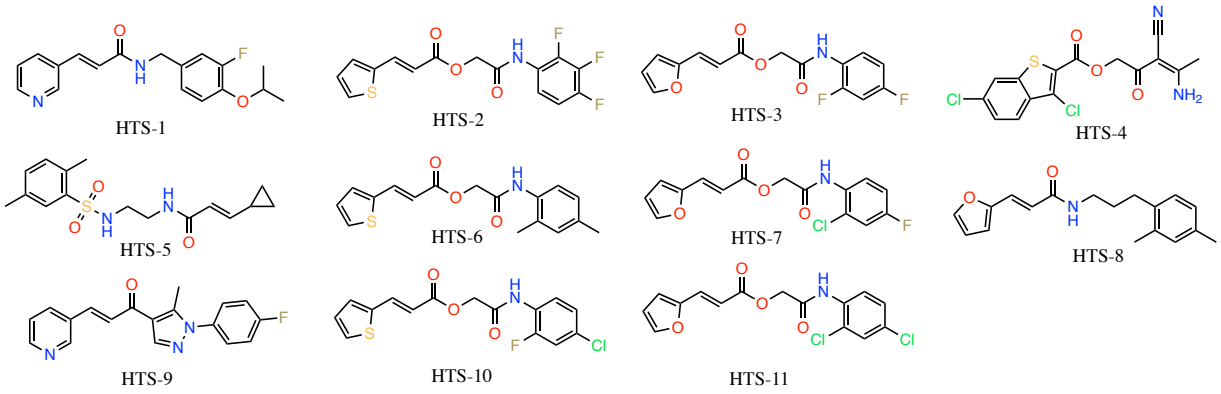


Figure 61. Virtual hits found by screening the Enamine HTS collection using pironetin's pharmacophore model

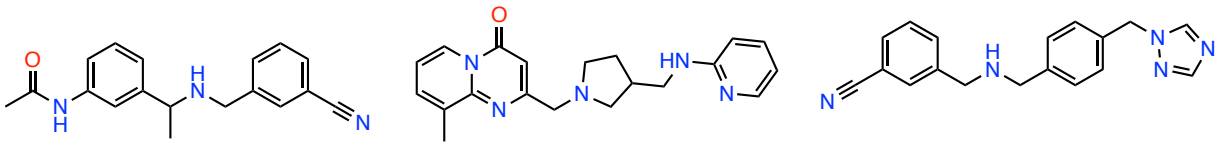


Figure 62. Virtual hits found by screening the tubulin-targeting library by pironetin's pharmacophore model

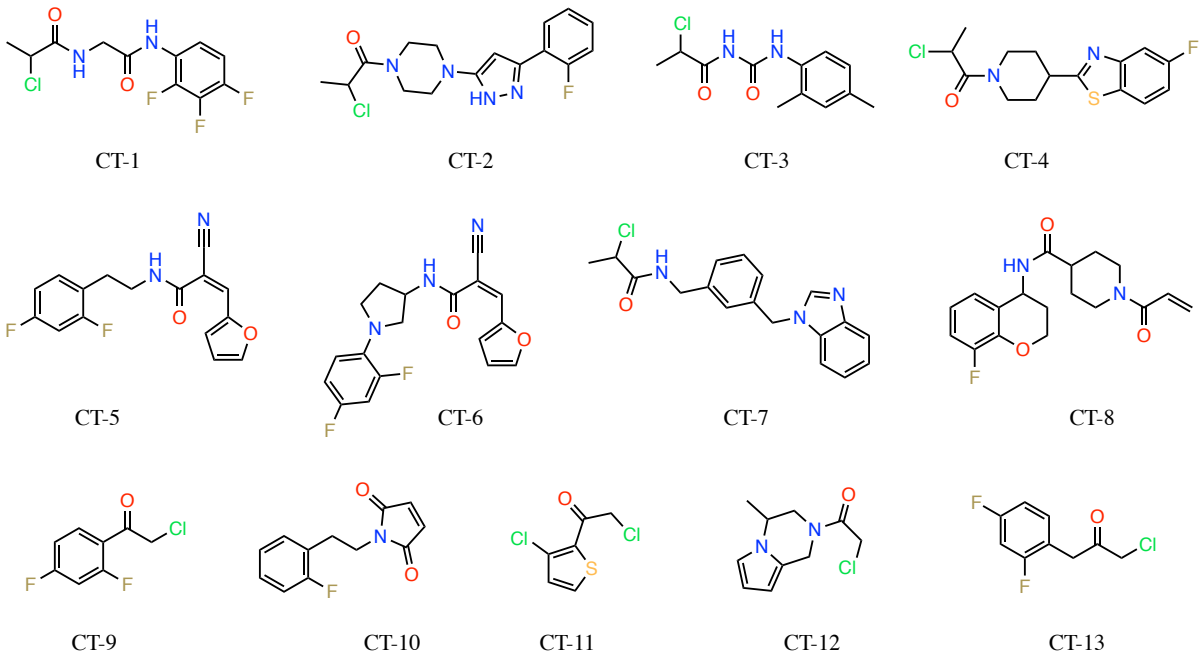


Figure 63. Virtual hits separately found by pironetin and small fragment's models from the cysteine-targeted Enamine library

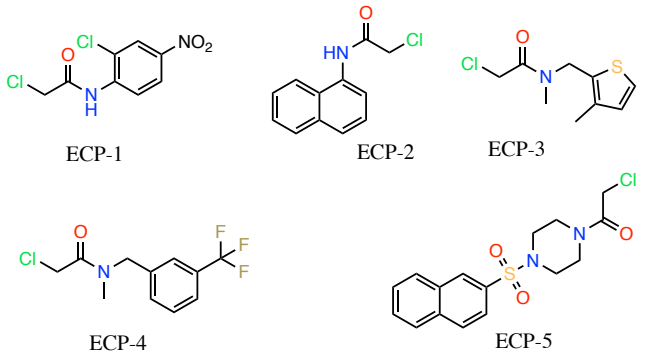


Figure 64. Virtual hits found by the small fragment model in the Enamine electrophile covalent probes library

The virtual hit molecules were purchased and subsequently underwent experimental validation, which is discussed in detail in section 3.3.7. Interestingly, a majority of the small fragment virtual hits and some larger virtual hits featured the α-chloroacetate reactive group, a group known for its exceptional reactivity due to the strong electrophilic character of the carbonyl carbon, which makes it readily susceptible to nucleophilic attack. This characteristic could make it a potent covalent modifier targeting the cysteine residues in the pironetin binding site. However, this same property also raises considerations about selectivity, as the high reactivity of the α-chloroacetate group could potentially lead to off-target interactions. Consequently, we anticipated that fragments containing this group may demonstrate non-specific tubulin polymerization modulator activity.

3.3.5. Machine learning-driven protein-ligand docking screening pipeline

Our pharmacophore model, being fairly rudimentary, may inadvertently exclude viable molecules due to an incomplete representation of all relevant protein-ligand interactions at the target site. Therefore, we sought to complement our existing approach with a more exploratory, albeit riskier strategy. We aimed to uncover entirely novel chemotypes targeting the pironetin site, which do not necessarily align with our existing pharmacophore model.

A docking-based virtual screening approach can be used to this end. It requires no pharmacophore hypothesis – any feasible interaction with the protein site can be exploited – but may fail in relative ranking of the impact of these interactions on binding affinity, and, foremost, is unable to tackle the covalent binding problem. It may discover compound conveniently fitting the pironetin binding site but without guarantees that these will feature any SH-binding warhead close to Cys316, and even less guarantees that if close, the warhead would actually react. Furthermore, neither docking nor pharmacophore model can guarantee that the selected ligands have a sufficient affinity to "power" (by whatever unknown mechanism) the induced-fit conformational change leading to the hypothesized binding site geometry to which they fit, according to the software. All in all, any state-of-art virtual screening approach is seriously challenged in this context, and the *a posteriori* analysis of hits by human experts is an absolute must.

Through the examination of docking scores, ligand poses, and the interactions a ligand forms within the site, we can gain insights into potential high-affinity molecules. However, large-scale application of this method is impeded by its significant computational demand. Literature provides examples of machine learning-driven iterative screening strategies developed to overcome this drawback (Figure 65) [107]. In these approaches, only a small subset of a large database is first selected for docking. A machine learning model is then trained on the results of the docking to predict the docking score based on a 2D representation of the ligands within this subset. This iterative process continues, with new subsets docked and their scores compared to the model's predictions. If the model's predictive accuracy, determined by a user-specified metric, is low, the model is re-fitted on the concatenated data from previous and current docking cycles. This process continues until the model achieves satisfactory predictive accuracy, after which it is used to predict the docking scores of all compounds in the larger database. Only compounds predicted to have a docking score higher than a user-defined threshold are docked, saving considerable time and computational resources without causing significant drops in the hit quality[108,109].

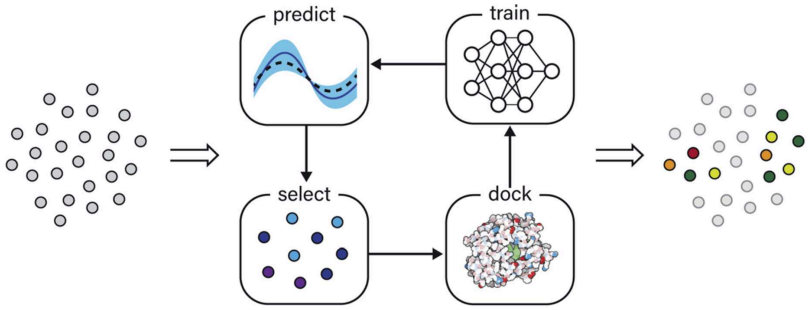


Figure 65. An overview of machine learning-based iterative screening approaches (adapted from Graff et al.[109])

In this study, we implemented such an iterative screening approach to screen the Enamine HTS collection using protein-ligand docking. Our setup used a support vector machine regression model and ISIDA descriptors as 2D molecular representations to learn the *chemplp* scoring function values from the protein-ligand complexes produced by the PLANTS software. The ISIDA property-labeled fragment descriptors encode molecular structures as counts of specific subgraphs' occurrences, with atoms represented as nodes, which can be labeled by element type or some local property/feature, and bonds represented as edges, with bond type information either present or omitted. Thus, many ISIDA fragmentation schemes can be produced for the same sets of molecules, different in the resolution of the chemical information extracted into the descriptors. The predictive performance of the model was measured by the regression coefficient (R^2) value between the predicted and actual docking scores, with a consistent R^2 value of 0.8 for three consecutive training cycles used as a threshold for model acceptance to screen the whole large database. A genetic algorithm, as described by Horvath et al.[110], was implemented to simultaneously optimize the SVM regression model's hyperparameters and the ISIDA descriptor set. Each model training step employed 3-fold cross-validation, repeated 12 times. The genetic algorithm parameters set the minimum initial population size at 500 points and a stopping criterion of 1000 generations without metric value progress. For each model fitting cycle, we randomly selected 3000 compounds from the large dataset. The training was configured to stop if the model did not require refitting after three optimization cycles.

We applied this setup to screen the Enamine HTS collection library. After five model refitting cycles, we obtained a predictive model with an R^2 metric value consistently exceeding 0.8, with the optimal fragmentation scheme being sequences fragmentation with atoms represented by atom symbol (no special coloration used) and the inclusion of the bond order information to produce fragments with the topological distance of at minimum 1 and at maximum 5. We then used this model and descriptor set to predict the docking scores of all compounds in the library. We chose to dock only those compounds predicted to have a docking score lower (indicating a better binding affinity) than -28.63. This cutoff was based on the 10% percentile of the best

docking scores from the model training data. Of the 2,670,898 predictions obtained, only 141,091 were docked using PLANTS, which is just 5% of the original library. Upon filtering for fragments with a double bond adjacent to an electron-withdrawing group (EWG) to promote cysteine binding, we were left with 1550 protein-ligand complexes. A Pareto front over the docking score and ligand efficiency scores was used to narrow down our selection to 52 compounds. These were then ranked by docking score, and only the top 15 molecules, which achieved a better docking score than pironetin and had their reactive group close to any of the four cysteine residues in the binding site, were retained for experimental validation (Figure 66).

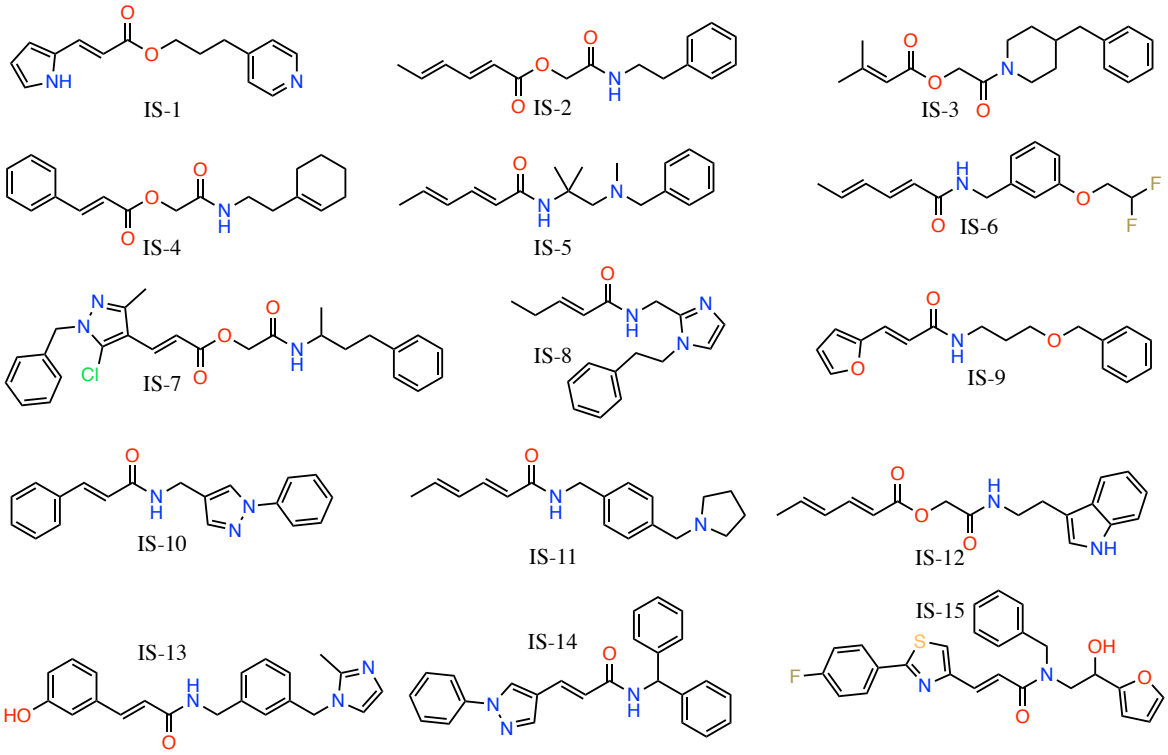


Figure 66. Fifteen virtual hits found by the iterative screening approach

3.3.6. Comparing the docked poses of virtual hits with those of known colchicine site agents

Our virtual screening and docking approaches yielded small molecules. In general, these molecules all feature an aromatic moiety and a significant number of hydrogen bind acceptors, which made us consider the possibility of alternative binding to the larger and more promiscuous colchicine site.

The colchicine site, primarily residing within the β-tubulin unit and described in more detail in Chapter 5, is a large cavity known to accommodate microtubule-destabilizing agents of diverse structures. Given its wide-open nature in contrast to the likely induced-fit opening of the pironetin site, it is plausible that our small, non-specific virtual screening hits could also bind there.

To assess the potential binding of our virtual hits to the colchicine site, we devised another pharmacophore modelling strategy. We conducted a literature review and mined the RCSB PDB database, producing a list of 104 PDB records of tubulin co-crystallized with colchicine site-targeting agents. From these records, we isolated the ligand-bound β-tubulin chains and aligned all of them to the ligand-bound β-tubulin from the high-resolution 6F7C PDB structure. After that, we automatically extracted pharmacophore models of all colchicine site-targeting compounds using LigandScout, merging models devised for structurally similar ligands that co-crystallized in the site following similar binding modes.

Then, we docked all of the 47 virtual screening hits into the colchicine binding site. The binding site was defined as all atoms within 8Å from the native ligand's (the colchicine ligand) center of mass in the 5EYP PDB structure, which was chosen due to being one of the most well-resolved crystal structures of tubulin bound to a colchicine site ligand (structure resolution 1.9 Å). Docking was done using the AutoDock GPU software to increase the sampling quality of the virtual hits position inside the binding site. It was setup to estimate 200 possible conformations for each ligand. The grid was centered around the native ligand's center of mass, stretching 62 points by the x-axis, 54 by the y-axis, and 78 by the z-axis, with a spacing of 0.375 Å between each point. All default atom types were used to calculate the grid maps. Default AutoDock scoring function was used.

Upon docking our virtual hits into the colchicine site, we investigated potential overlaps between the best docked poses (i.e., the ones with the lowest value of the docking score) and the pharmacophore models. In particular, we were only considering virtual hits that overlapped three or more pharmacophore features for any given model. This analysis identified two ligands that overlapped three pharmacophore features derived from the 3HKD PDB structure of a (3Z,5S)-5-benzyl-3-[1-(phenylamino)ethylidene]pyrrolidine-2,4-dione ligand (PDB code: N16) co-crystallized with tubulin in the colchicine binding site. These were HTS-9 and CT-2. Despite conforming to the pironetin pharmacophore model, these ligands, when docked in the colchicine site, occupied positions highly similar to that of N16, and arranged well with its pharmacophore model (Figure 67).

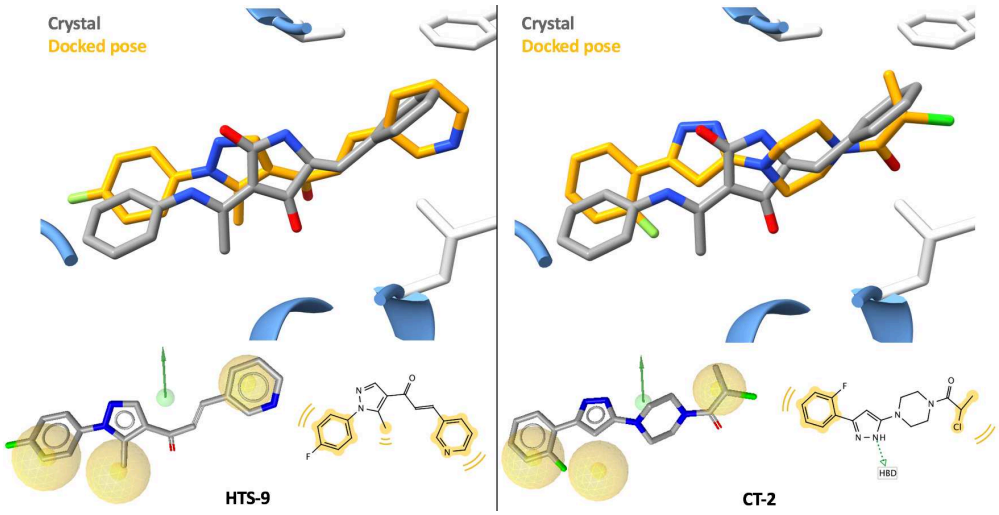


Figure 67. Docked poses of two virtual screening hits aligned well to a pharmacophore model of a known colchicine site inhibitor

Additionally, we superimposed the docked poses of virtual hits HTS-9 and CT-2 with all small fragments that were crystallographically shown to bind at the colchicine site in the recent fragment screening campaign mentioned above[106]. Remarkably, we saw good overlap (Figure 68) of a part of HTS-9 with cyclopropyl-[4-(4-fluorophenyl)piperazin-1-yl]methanone, a small fragment co-crystallized with tubulin in the colchicine site (PDB ligand ID: GX4, PDB structure code: 5S4U).

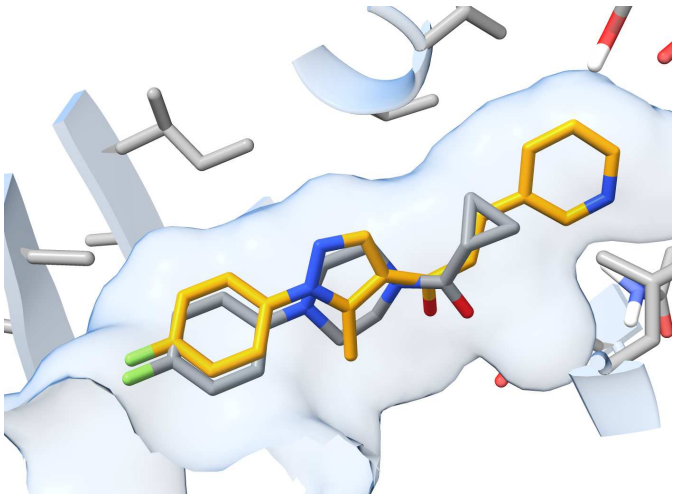


Figure 68. Overlap between docked pose of virtual hit HTS-9 and (orange) and crystal pose of small fragment GX4 (gray).

This led us to consider that HTS-9 and CT-2 might bind at the colchicine site, a consideration we bore in mind when proceeding to the experimental validation of all identified virtual hits.

3.3.7. Experimental validation of virtual hits

We procured all 47 virtual hits identified in our study for further evaluation, employing two primary assays: X-ray crystallography and tubulin polymerization inhibition, as detailed in Section 2.3.7. The X-ray crystallography experiments performed by our collaborators from the group of Dr. Andrea Prota in the Paul Scherrer Institut, Villigen, Switzerland did not detect any molecule or fragment bound to the prionetin site. However, three small fragments found by pharmacophore screening of the cysteine-targeted library with the small fragment model (**CT-9**, **CT-10**, **CT-11**) demonstrated a considerable inhibitory effect on tubulin polymerization (Figure 69), despite their absence in the X-ray crystal structures. We hypothesize that this could be attributed to non-selective acylation reactions, given the high reactivity of these fragments. While the exact site of acylation remains undetermined, it is most likely located on the tubulin protein surface. Given that the small fragments follow the pharmacophore model of the 2-chloro-N-methylbenzene-1-sulfonamide fragment that binds close to the entrance to the pironetin site, we speculate that the reactions may involve cysteine residues number 4 or 200 on the α-tubulin subunit. This warrants further investigation, as there is potential to develop these fragments into effective ligands.

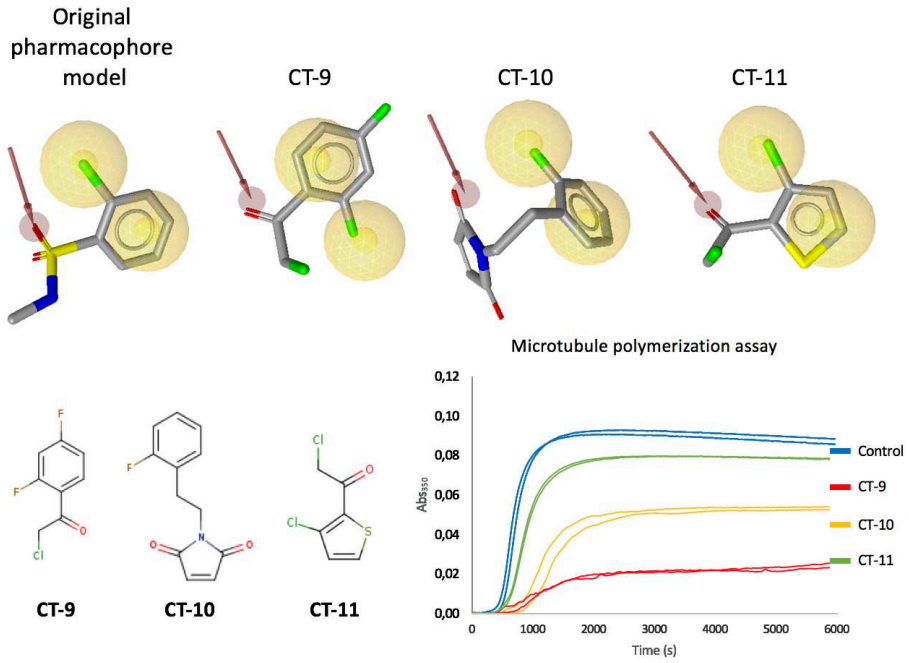


Figure 69. Microtubule polymerization assay results for three small fragment hits

Remarkably, two of the virtual hits (**HTS-9** and **CT-2**) were found to bind at the colchicine site, as predicted by our pharmacophore modelling and fragment overlap analysis. These molecules significantly inhibited tubulin polymerization, with compound HTS-9 showing a particularly pronounced effect. The X-ray structure of compound CT-2 was partially unresolved,

suggesting it may bind to the colchicine site, albeit not remain stable in the site. Conversely, the X-ray crystal structure of compound HTS-9 was well resolved, providing stronger evidence for its binding (Figure 70).

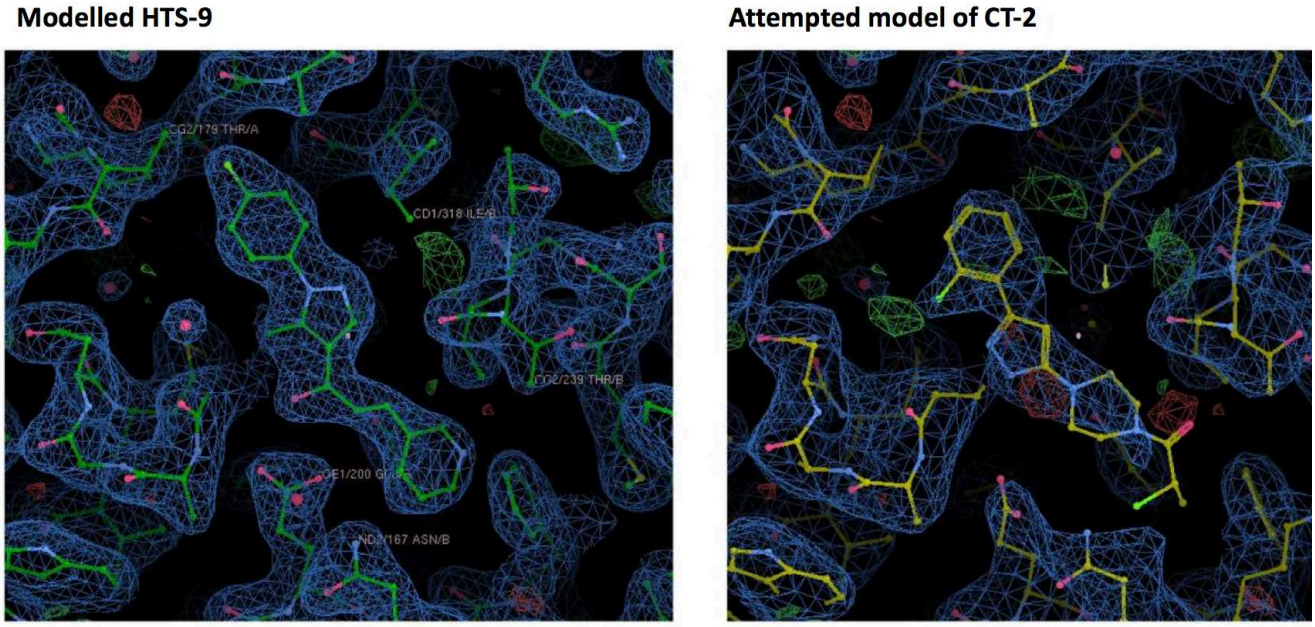


Figure 70. Experimental electron densities of hits HTS-9 and CT-2 in the colchicine site

We also validated the docking pose predicted by the AutoDock software for compound HTS-9. As a result, the docked pose that had the lowest (i.e., best) value of the docking score was 1.72 Å different from the pose that this compound takes in the colchicine binding site when co-crystallized with tubulin (Figure 71). The RMSD value was calculated using the CalcLigRMSD script, contributed to the RDKit chemoinformatics toolkit by Velázquez-Libera et al.[104]

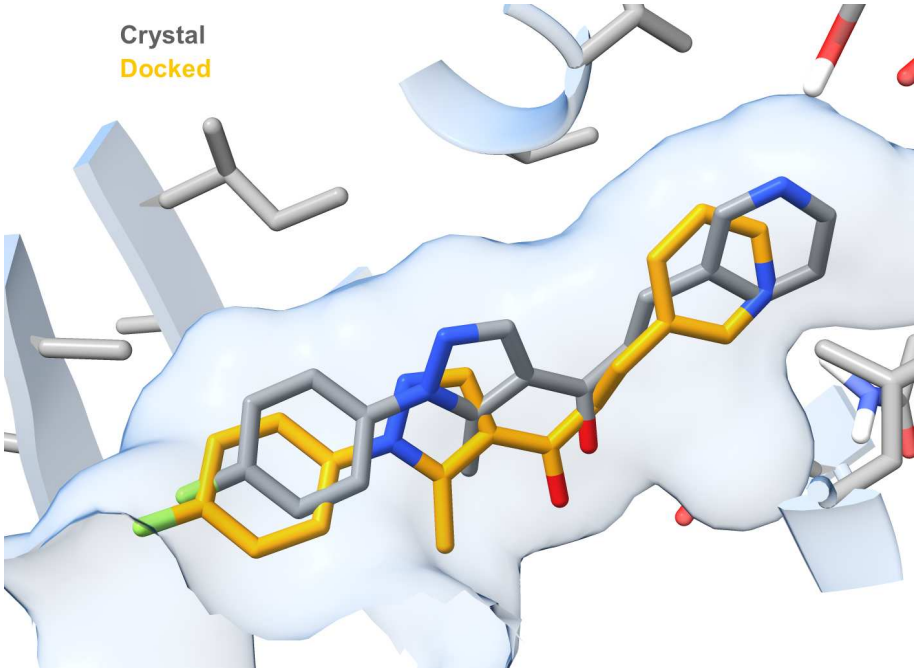
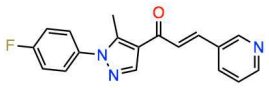


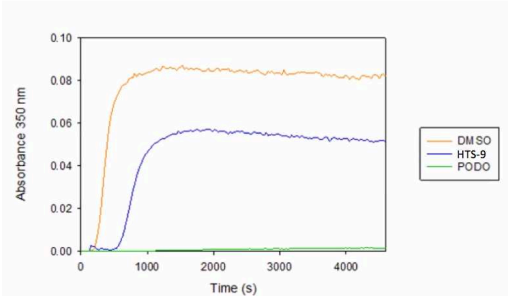
Figure 71. Crystal pose of hit HTS-9 (gray) vs. AutoDock generated best pose (orange)

Tubulin polymerization assay showed notable inhibition effect for compound HTS-9. Additionally, cell viability tests, performed by Francesca Bonato, a fellow TubInTrain PhD student from the group of Dr. Fernando Díaz at CIB-CSIC, Madrid, Spain revealed that compound HTS-9 was cytotoxic in the micromolar range (Figure 72). The goal of the *in vitro* assay was to measure the concentration of compound HTS-9 that inhibits proliferation of five different cancer cell lines by 50% (i.e., the IC50 value). Cytotoxicity of compound HTS-9 was measured in comparison with two standard nanomolar inhibitors of tubulin polymerization, podophyllotoxin and mebendazole, both binding to the colchicine site as well. Remarkably, HTS-9 exhibited pronounced cytotoxic action against all of the cell lines it was tested on. Interestingly, it exhibited some specificity towards the βIII-tubulin isotype expressed in the HeLa cells. This isotype is noteworthy as it is expressed by cancer cells resistant to other tubulin-targeting therapeutics. Thus, our discovery could pave the way for the development of future tubulin isotype-specific treatments.



HTS-9

Tubulin polymerization assay



Cytotoxicity assay

Compound/cell line	A549, IC50 ± SEM	HeLa, IC50 ± SEM	HeLa βIII, IC50 ± SEM	R/S index	Kb3.1, IC50 ± SEM	KbVb, IC50 ± SEM	R/S index
HTS-9	13.8 ± 0.6 µM	8 ± 1 µM	3.4 ± 0.9 µM	0.41	11.9 ± 0.7 µM	8.1 ± 0.8 µM	0.68
Podophyllotoxin	15.7 ± 0.7 nM	14 ± 1 nM	10 ± 1 nM	0.74	15 ± 1 nM	22 ± 2 nM	1.43
Mebendazol	422 ± 18 nM	200 ± 26 nM	318 ± 26 nM	1.58	592 ± 48 nM	486 ± 31 nM	0.82

Figure 72. Results of in vitro bioactivity tests performed for hit compound HTS-9.

3.4. Conclusion and perspectives

In summary, our research involved the screening of the ChEMBL library of drug-like molecules with known bioassay data, alongside multiple Enamine libraries of purchasable compounds and small fragments. Despite our efforts, the ChEMBL library failed to yield any significant hits, leading us to discontinue this line of study. The reasons for this lack of results may range from a potentially oversimplified pharmacophore model, unanticipated behaviour of the pharmacophore screening algorithm in the context of warhead matching, inaccuracies in docking filtering, or the possibility that the ChEMBL library does not contain ligands that induce cytotoxicity via pironetin binding.

Shifting our focus to the Enamine libraries, we applied two pharmacophore models and a novel machine learning-aided protein-ligand docking-based approach for screening the largest library, the HTS collection. This strategy proved more successful, yielding 47 virtual hits. Upon purchasing and evaluating these hits, we identified three small fragments with significant microtubule-depolymerization activity. Furthermore, we discovered two molecules that bind at the colchicine binding site and exert a notable inhibitory effect on microtubule polymerization. Importantly, one virtual hit demonstrated specificity towards the βIII-tubulin isotype, which is prevalent in drug-resistant cancer cells.

Future research should consider the further optimization of these small fragment hits to develop potent tubulin polymerization modulators. Of particular interest is the promising hit that binds at the colchicine site, which will require additional investigation and optimization. Our work paves the way for these future endeavors, contributing valuable insights to the development of targeted tubulin polymerization inhibitors.

Chapter 4. Discovery and design of todalam site-targeting agents

4.1. Introduction

In cells, tubulin is regulated by numerous proteins that modulate microtubule dynamics and organization, thus influencing fundamental physiological processes in all eukaryotes[106]. This binding capacity is not limited to proteins, but extends to a vast array of chemically diverse small molecules that interact with one of the seven distinct binding sites identified on the tubulin protein to date. Compounds that disrupt tubulin's function have demonstrated significant effectiveness in treating various human diseases. Given the diverse array of proteins and ligands that bind to tubulin, one can suggest that there may exist other, yet undiscovered binding sites on the tubulin protein, which could also be targeted to develop novel therapeutic agents or molecular probes[106]. One way to investigate this possibility is to perform crystallographic fragment screening, which is an experiment concerned with soaking a protein of interest with a large number of structurally diverse small chemical fragments, identifying their binding modes through X-ray crystallography experiments, and further developing the bound fragments into actual ligands with desired action[111].

In a recent study, Mühlethaler et al. employed such a crystallographic fragment screening campaign for tubulin, yielding significant results[106]. Using 708 different fragments, they identified 56 fragments that target ten unique, previously unidentified binding sites on the tubulin protein[106]. Notably, three fragments bound to a site at the inter-dimer interface between α- and β-tubulin, adjacent to the pironetin site, prompting further investigation due to its potential relevance in anti-cancer drug design and tubulin-targeting molecular probe studies. The authors then employed a fragment linking strategy to grow the fragments into a full ligand, named todalam[45].

Todalam binds to a unique binding site, formed by residues from βH3', βH11', and αH8 helices, αS4 strand, and various loops including βT3, βT5, αH3-S4, and αH4-αS5 when two tubulin heterodimers come together in a head-to-tail fashion[45]. The ligand's structure can be divided into three moieties (Figure 73): an anchor (the acetaminophenyl group), a central linker (an aminothiazole group), and a hydrophobic head (a hydrophobic trifluoromethylbenzene group). Todalam forms three hydrogen bonds with the βAsn102, αThr257, and αGln256 residues, and establishes parallel-displaced π-π stacking interactions with βTrp407. The binding site stretches into a region rich in hydrophobic amino acids (αLeu136, αLeu167, αLeu242, αLeu252), which supports the binding of todalam due to its own aromatic hydrophobic group[45].

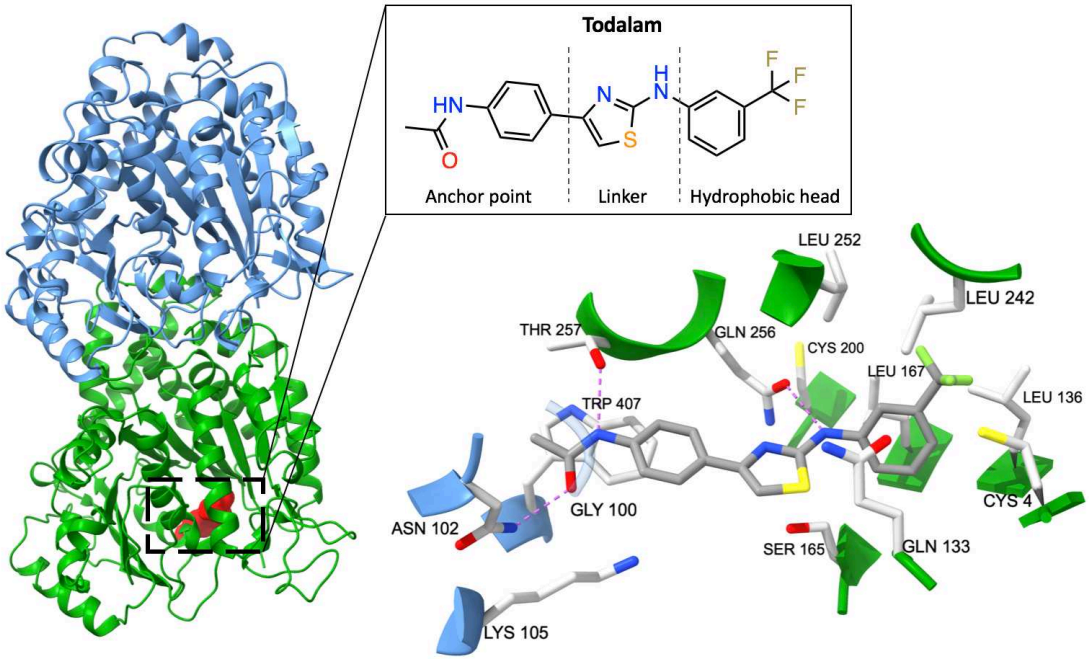


Figure 73. Overview of the todalam binding site and the todalam ligand's structure.

The transformation of tubulin from a curved to a straight shape when it integrates into the microtubule lattice is facilitated by the rotation of the intermediate domain of both the α- and β-tubulin monomers in relation to their N- and C-terminal domains[5]. A part of the todalam binding pocket on α-tubulin is formed by secondary structure elements from both the N-terminal (αS4 and αS5) and intermediate (αH8) domains of α-tubulin. When the α-tubulin monomer shifts from the curved to the straight form, the αH8 helix moves towards the αS4 and αS5 strands. However, in the presence of todalam, this motion results in a collision between the αH8 helix and the main part of the todalam compound. Therefore, todalam appears to function as a "molecular block", preventing the αH8 helix from moving closer to the αS4 and αS5 strands during this conformational shift[45]. Because todalam's binding site is positioned between two tubulin dimers aligned lengthwise, todalam is able to bind both the α- and β-tubulin monomers of two tubulin dimers simultaneously. This explains why the presence of todalam *in vitro* induces the formation of tubulin ring-like structures, effectively inhibiting microtubule polymerization and causing significant cytotoxicity[45].

What makes the todalam binding site particularly interesting is the presence of a targetable cysteine residue (αCys4) at the hydrophobic pocket in the a-tubulin subunit. If a molecule could bind to this site and form a covalent bond with the cysteine residue, it could greatly facilitate the development of assays for immobilizing tubulin and expedite high-throughput crystallographic studies on agents targeting tubulin by being developed into a molecular probe. The recent discoveries related to the todalam site open up new opportunities for the rational design of more accessible molecules, and further exploration and enhancement of new binders for this site.

In light of these findings, this project had two main objectives. Firstly, we aimed to identify diverse molecular scaffolds that can bind to the todalam binding site to explore structure-activity relationships in this so far uncharacterized binding site. Secondly, using this information, we sought to either find analogues of these compounds with reactive functional groups (warheads) capable of targeting the cysteine residue, or establish a limited number of easily accessible reactions that produce these well-binding scaffolds and identify purchasable small fragments that can functionalize them to create tailor-made covalent binders.

4.2. Discovery of novel chemical scaffolds that target the todalam site

4.2.1. Initial data analysis and library selection

We approached the project by first conducting a survey of available data on the todalam binding site and ligand that target it. At that time, several todalam site-related Protein Data Bank (PDB) structures were available, i.e. tubulin co-crystallized with todalam (PDB code: 5SB7) and its four simplified derivatives (PDB codes: 5SB3, 5SB4, 5SB5, 5SB6). A thorough analysis of these structures revealed the critical importance of the acetaminophenyl group and a hydrophobic group at the molecule's opposing end for effective binding, due to the interactions they formed with the binding site residues (Figure 74). Additionally, the central region of the todalam site contains two glutamine residues (αGln133 and αGln256) and a serine residue (αSer165), all of which provide opportunities for todalam site-targeting ligands to form hydrogen bonds with the site. To ensure the presence of these interactions in the ligands we aimed to discover, we opted to employ the pharmacophore modelling approach in this project.

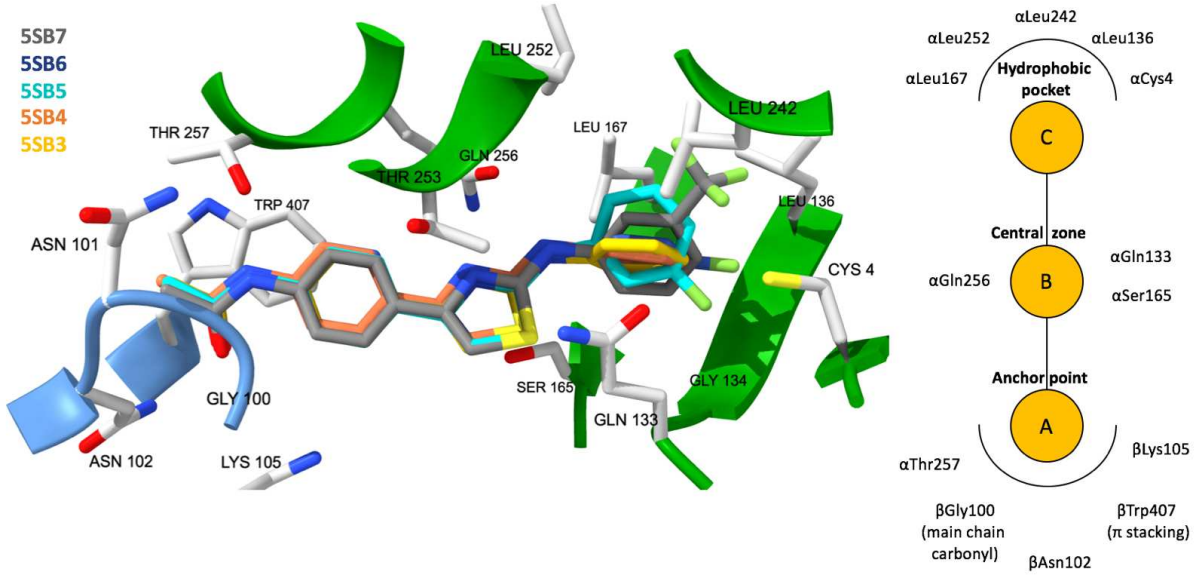


Figure 74. Overlap of five todalam derivatives in the binding site helped to come up with ligand moiety designations

Seven diverse Enamine libraries, including the HTS collection (2,688,748 compounds), the tubulin-targeted library (3,452 compounds), the protein-protein interactions inhibitors library (40,640 compounds), the covalent fragments library (7,360 compounds), the covalent screening library (11,200 compounds), the NP-like library (4,158 compounds), and the covalent compounds collection (88,259 compounds), were selected for pharmacophore screening. We anticipated that the chemical diversity of these libraries would ensure the diversity of the found virtual hits, while the ready availability of these compounds would help us advance the exploration of the binding site chemistry efficiently. Concurrently, the group of Prof. Daniele Passarella from the University of Milano, Italy, deigned a custom in-house library of 176 compounds, which they were capable of synthesizing in one or two steps using their expertise in amide formation, azide-alkyne Huisgen cycloaddition, and Suzuki coupling reactions. Virtual screening strategy for this small in-house set involved docking all molecules into the todalam site and re-scoring them based on their alignment with a chosen pharmacophore model. Altogether, the screening involved a total of 2,843,993 compounds.

4.2.2. Pharmacophore modelling

We used the crystal structure of the tubulin-bound todalam ligand (PDB code 5SB7) to automatically generate an initial pharmacophore model with LigandScout, resulting in an eight-feature model (Figure 75). The features included four hydrophobic spheres, two hydrogen bond donor and two hydrogen bond acceptor interactions between the ligand and the binding site residues.

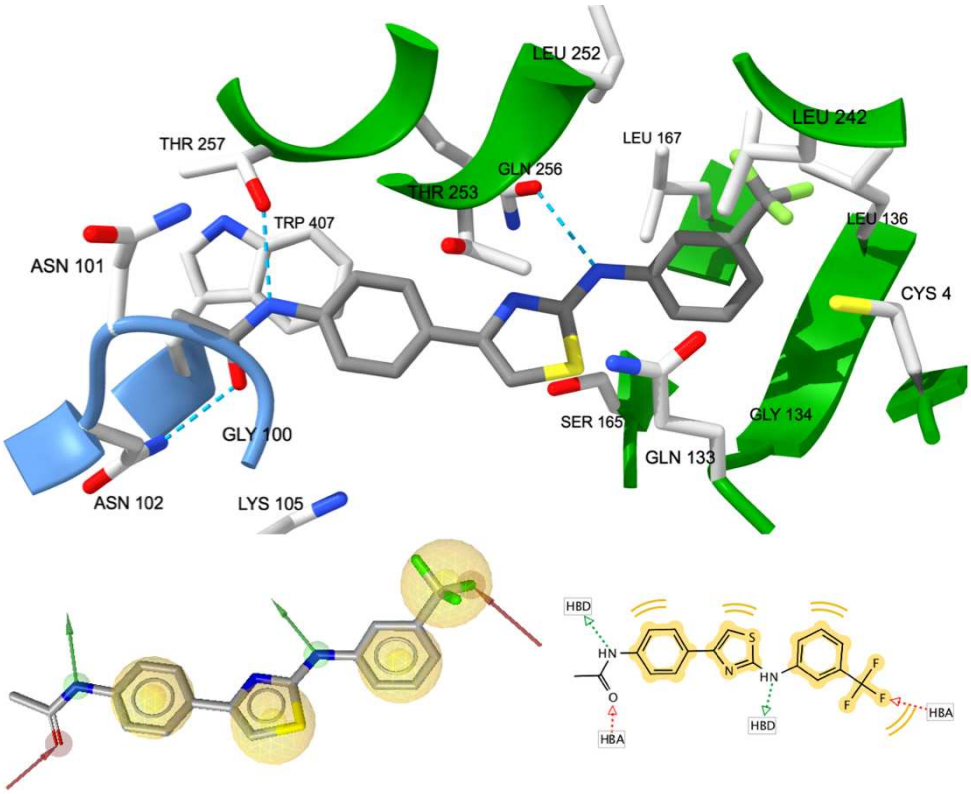


Figure 75. The initial eight-feature pharmacophore model of todalam; dashed blue line – possible hydrogen bonds between todalam and the binding site residues

The nature and specific arrangement of features of this model possibly made it restrictive for virtual screening. To address the potential lack of hit diversity during the virtual screening, five simpler models were created from the initial model by individually removing each of the five non-anchor features (Figure 76).

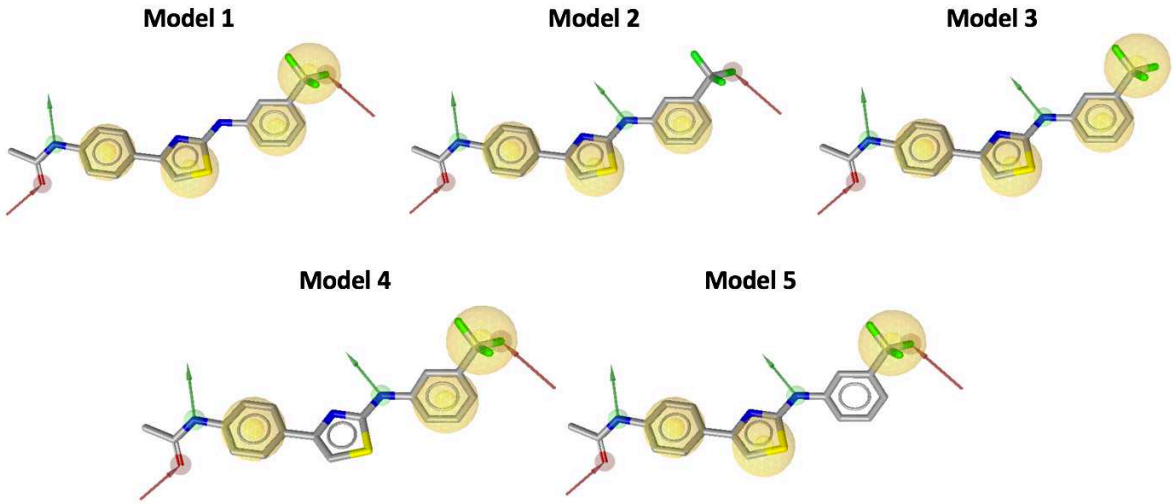


Figure 76. Five simplified analogues of the initial model, each missing one feature not related to the anchoring point

To validate the applicability of the PLANTS docking software for modelling our protein-ligand system, we performed a re-docking experiment of the todalam ligand. This re-docking was performed on a protein structure from the 5SB7 crystal structure, stripped of all solvent molecules, ions, and other small organic molecules. The ligand and protein were prepared using the SPORES software, with a random ligand conformation generated using ChemAxon's conformational sampling tool. We defined the binding site as all atoms of all residues from chains B and C (modelling b-tubulin and a-tubulin, respectively) of the 5SB7 PDB structure within 8 Å from the native ligand's pose in the site. The software was configured to produce ten docked poses, each evaluated by the *chemplp* scoring function. When compared to the native pose, the best-scoring docked pose had an RMSD value of 0.43 Å, confirming the software's reliability in accurately modeling the binding mode of todalam site-targeting ligands (Figure 77).

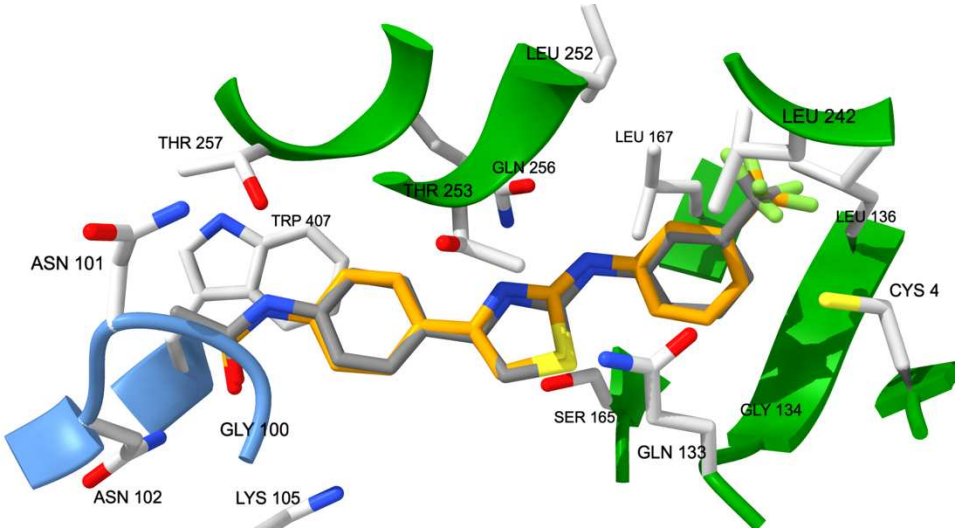


Figure 77. Native (gray) and re-docked (orange) poses of todalam

4.2.3. Screening libraries preparation

Following the selection of the Enamine libraries totalling 2,843,817 compounds, we standardized the molecules in these libraries by using a multi-step standardization process previously described in section 2.2.4. Briefly, it included proper aromatization, dealkalization, conversion into canonical SMILES strings, salt and mixture removal, species neutralization, and tautomer generation with ChemAxon tools. Each molecule then underwent conformational sampling using LigandScout's integrated iCon tool, generating up to 25 unique conformations with an RMSD value of at least 0.7 Å between conformations.

Similarly, the 176 compounds in the in-house library underwent the same standardization routine, although these were specifically prepared for docking rather than for pharmacophore screening, so conformational sampling was not performed for them.

4.2.4. Pharmacophore screening

As anticipated, the initial eight-feature model only identified one hit molecule, significantly similar to the native ligand, todalam, from the seven screened Enamine libraries. To ensure a diverse set of virtual hits and increase our design options, we repeated the screening using the five simplified models. Consequently, these yielded 175, 18, 53, 248, and 13 virtual hits from models 1, 2, 3, 4, and 5, respectively. This process, in total, produced 499 unique virtual hits from the Enamine libraries. Figure 78 shows examples of best-matching molecules found by each model. The next step was to dock these virtual hits into the todalam binding site and assess the alignment of their best scoring poses within the site with the full eight-feature pharmacophore model.

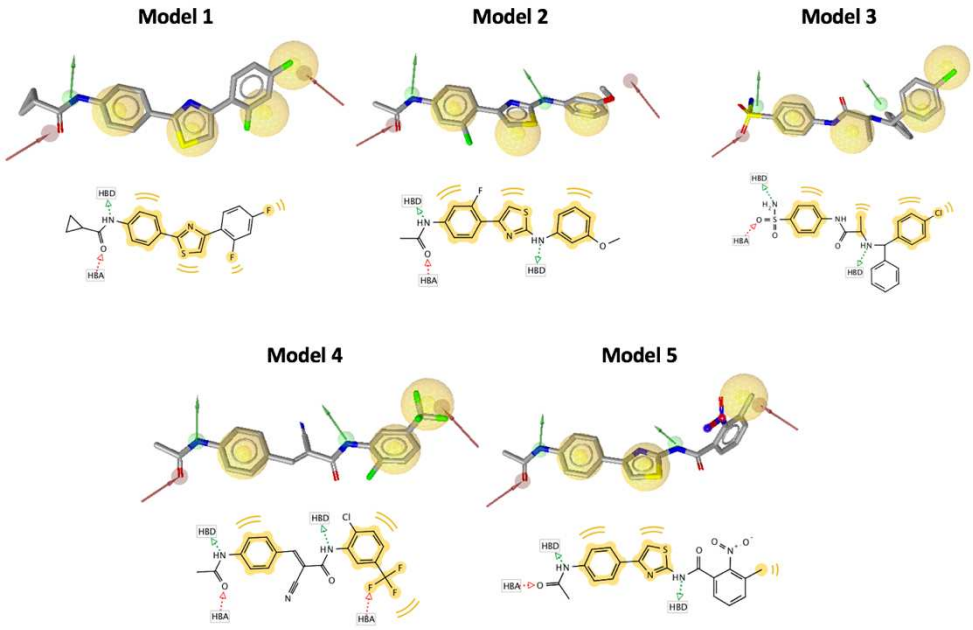


Figure 78. Example of well-fitting virtual hits found in the Enamine libraries after pharmacophore screening

4.2.5. Protein-ligand docking

All 499 unique virtual hits from the Enamine libraries were subsequently docked into the todalam binding site alongside the native ligand, utilizing the PLANTS docking software. The procedure and settings were consistent with those used during the re-docking experiment (section 4.2.2). Each of the ten poses calculated for each virtual hit were assigned a *chemplp* docking score. The molecules were then ranked based on the docking score values of the best scoring poses, and those with scores surpassing that of the native ligand, todalam, were selected (60 in total). These selected hits were then re-scored by overlapping each one's best-scoring docked pose in the todalam site with the full eight-feature pharmacophore model, which resulted in the final selection of 13 virtual hits from the Enamine libraries for experimental validation (Figure 79).

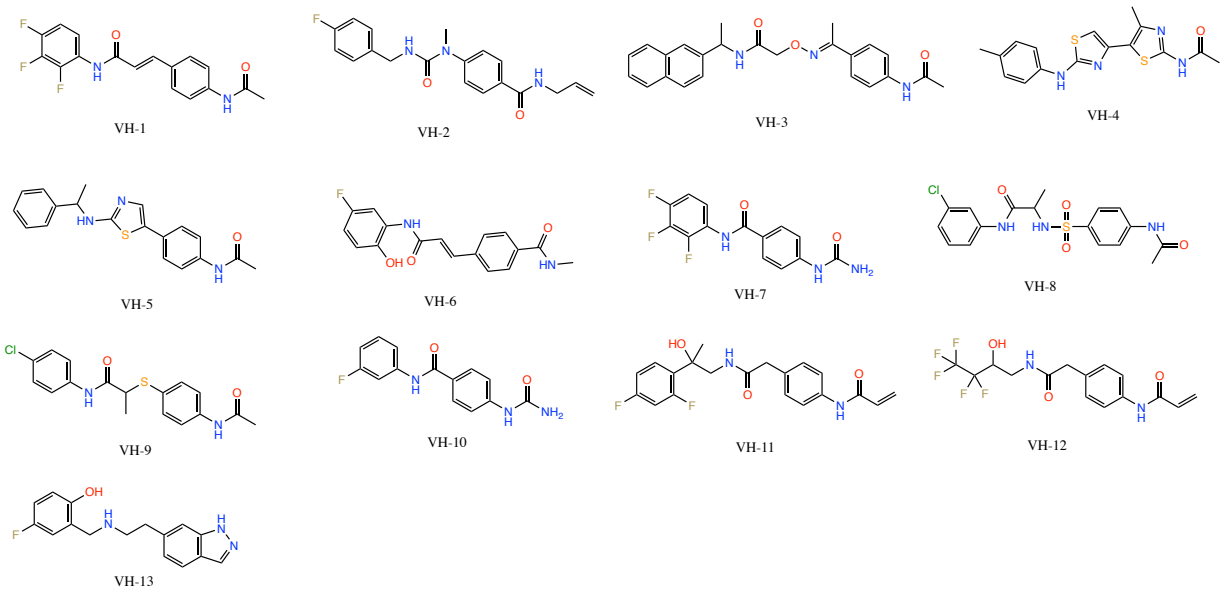


Figure 79. Virtual hits found after re-scoring pharmacophore screening hits by alignment of their best docked poses with the eight-feature pharmacophore model

The thirteen virtual hits were selected considering their potential interactions with the residues in the binding site. The calculated poses for all thirteen compounds are shown in Figure 80.

As such, VH-1 was chosen due to its excellent overlap with todalam's bound structure (in the image shown as transparent black sticks), facilitated by three hydrogen bonds and π-stacking interaction with identical residues to todalam. VH-2 was selected due to the possible dual hydrogen bonds and π-stacking with the b-tubulin anchor residues, and a possible hydrogen bond with serine. The selection of VH-3 was motivated by the desire to investigate the potential of a molecule to establish a hydrogen bond with αGln133 and provide an outward-facing functional group. VH-4 was chosen to examine the impact of substituting the phenyl ring in the anchor group, as docking revealed a favourable conformation. VH-5 was selected due to the possible hydrogen bonding with αSer165 and αGln256 via the sulphur atom of the thiazole fragment as suggested by the best-docked pose. This pose also displayed an advantageous placement of the phenyl ring within the hydrophobic pocket – a position we would have liked to leverage for future cysteine targeting.

VH-6 was chosen for its potential to form two hydrogen bonds with αGln256 to improve ligand stability in the site. VH-7 and VH-10, despite being shorter than todalam, were selected to explore an alternative anchor interaction point on the interdimer interface. Both molecules' anchors were calculated to form three hydrogen bonds with nearby residues. The choice of VH-8 was motivated by the desire to explore potential stabilizing effects resulting from hydrogen bonds with residues αGln133 and αGln256. We selected VH-9 with an interest in replacing the linker part in the middle of the molecule with a non-conjugated linear chain instead of a ring structure. Similarly, VH-11, with an easily accessible synthetic scaffold, was chosen for its potential binding

capabilities and the subsequent possibility of straightforward fragment modification through simple synthesis. VH-12 was selected due to the potential for replacing the hydrophobic aromatic ring with another hydrophobic functional group. This molecule followed todalam's binding mode closely and displayed the formation of an additional hydrogen bond with αSer165. Finally, VH-13 was chosen for its close alignment with the placement of todalam's hydrophobic groups within the pharmacophore model. We wanted to investigate whether prioritizing hydrophobic interactions over hydrogen bonding could aid this compound in reaching the site.

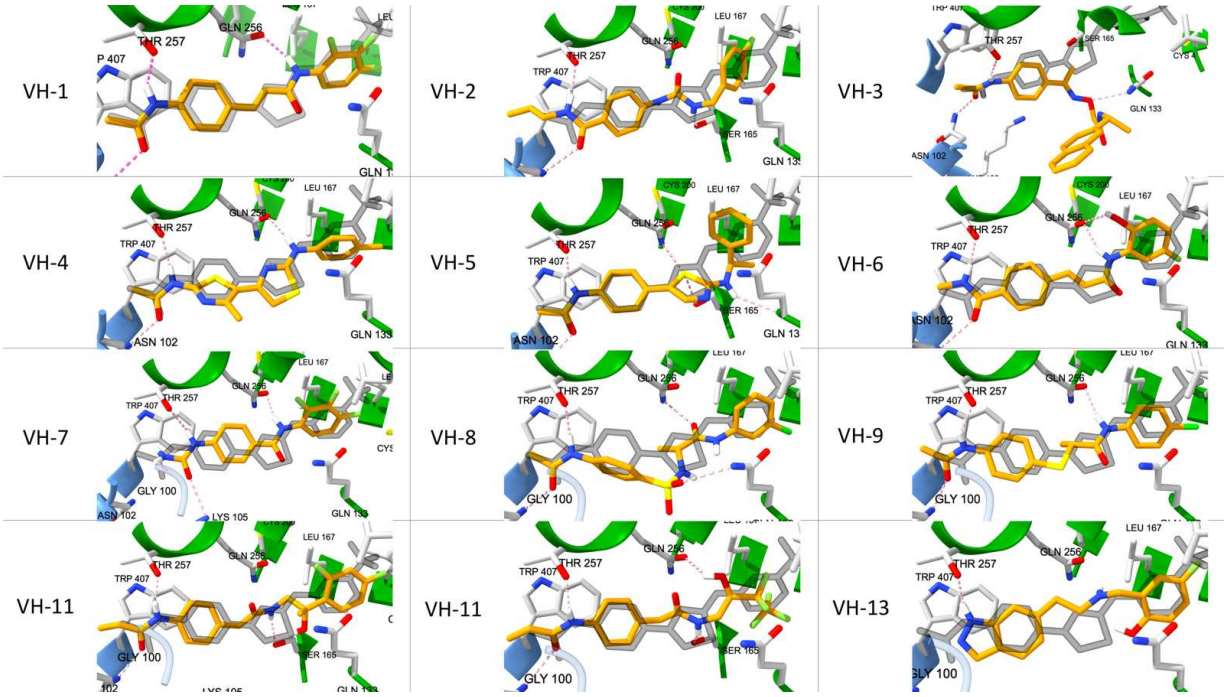


Figure 80. Predicted poses of virtual hits VH 1-13. Todalam's binding mode is shown as transparent gray sticks; best-scoring docked poses of the virtual hits – as orange sticks. Possible hydrogen bonds within the site are shown as pink dashed lines.

An identical procedure was applied to the molecules from the in-house library. Of the 176 designed molecules, 68 achieved a docking score superior to todalam, and five were prioritized for further investigation after being re-scored with the eight-feature model of todalam bound to the binding site (Figure 81 and 82).

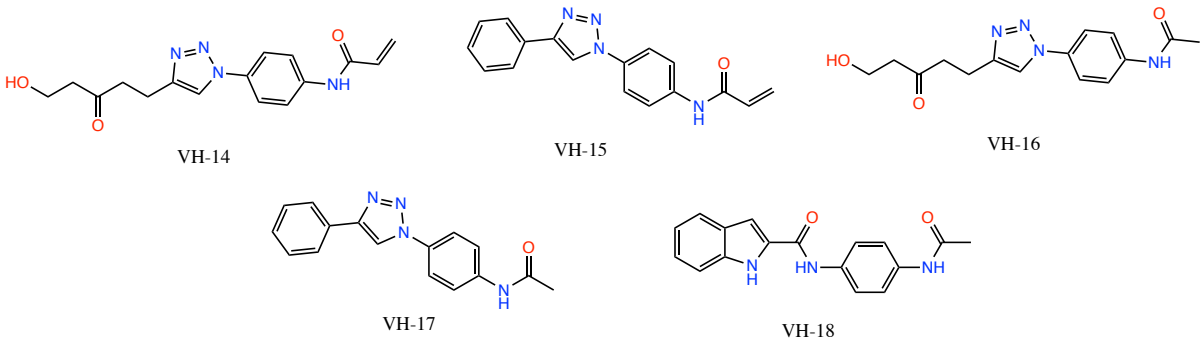


Figure 81. Five virtual hits selected from the in-house library

We selected in-house compound VH-14 due to its easy synthetic accessibility and a favorable docking pose, which aligns well with todalam's binding mode. Notably, the triazole ring of VH-14 shows a considerable overlap with todalam's thiazole group. However, the simplistic alkyl chain of the linker doesn't offer many opportunities for establishing stabilizing interactions within the site. Similarly, we opted for VH-16, as its docking pose exhibited good alignment with todalam's binding mode. Interestingly, the docked pose suggested a potential for a hydrogen bond between the triazole ring and the αSer165 residue. Compounds VH-15, VH-17, and VH-18 were selected because their calculated binding poses closely mirrored that of todalam. The easily accessible scaffold of these compounds offers the potential for future modifications. Additionally, we selected these compounds to understand the significance of having a phenylacetamide fragment in the anchor, as opposed to a phenylprop-2-enamide fragment, thereby contributing to our understanding of this binding site.

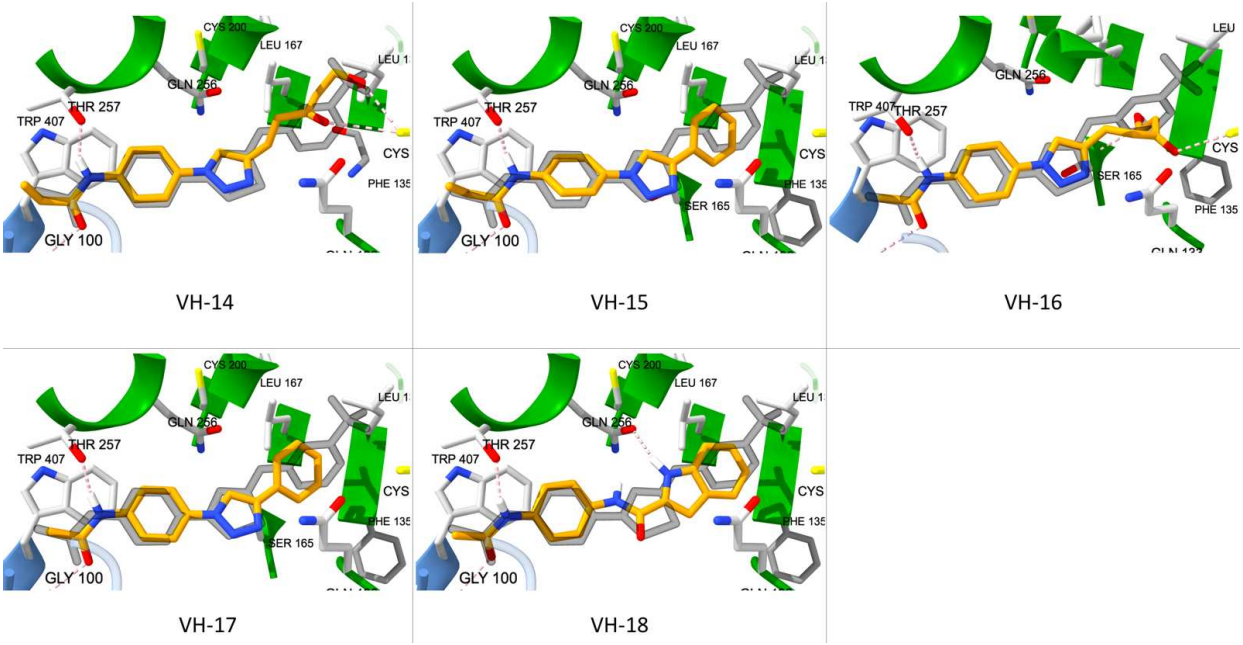


Figure 82. Best docked poses of in-house library virtual hits (orange) overlapped with todalam's native bound pose (transparent black)

4.2.6. Binding site similarity search

We also considered the possibility of searching for other binding sites that are structurally and chemically similar to the todalam site to potentially identify ligands that could target the todalam site due to similar binding environments. As such, we needed to select a set of protein-ligand complexes to search in, and a binding site similarity search algorithm. To this end, we used the scPDB database as the source of protein-ligand complexes[112]. It contained 16,034 entries: 6326 ligands bound to 4782 proteins. As a way to compare the sites, we used the ProBiS algorithm[113]. It conducts pocket matching by comparing the geometric and physicochemical characteristics of

protein binding pockets. This comparison is carried out at the level of amino acid functional groups. In this method, pockets are represented as graphs, where each vertex represents a functional group of surface residues that interact with other molecules. These vertices are assigned specific physicochemical properties, such as hydrogen bond acceptor or donor, mixed acceptor/donor, and aromatic or aliphatic characteristics. When comparing two proteins, a product graph is created, retaining only those edges whose lengths in the individual protein graphs differ by less than 2Å. The algorithm then identifies potential binding site similarities by applying the maximum clique algorithm. Here, the maximum clique represents the largest similarity between the two protein graphs, based on the number of vertices in the product graph. Each maximum clique corresponds to a single local structural alignment between the two proteins being compared. The final step involves scoring the constructed alignments using a function that takes into account surface vector angles, surface patch root-mean-square deviation, surface patch size, and expectation values[73].

Despite the meticulous search and the use of two separate representations of the todalam binding site (as a single pocket on the a-tubulin subunit, and as a pocket between b-tubulin and a-tubulin of two separate longitudinally aligned heterodimers), only the pironetin binding site was found to be similar. As this result did not offer substantial value for our research, we discontinued further binding site similarity searches.

4.2.7. Experimental validation of virtual screening hits

We proceeded to validate the 18 virtual hits (13 from the Enamine libraries, 5 from the in-house library) using X-ray crystallography and microtubule polymerization bioassays. Both experiments were performed by our collaborators from the TubInTrain consortium using the set up described in section 2.3.8.

From the in-house library compounds, three out of five were crystallographically confirmed to bind at the todalam site (Figure 83). Additionally, for the two other virtual hits that follow the same scaffold, electron density related to the anchoring moiety of the molecules was detected in the binding site, hinting that the molecules seem to bind at the site, but cannot adopt a stable conformation. Two of the bound molecules exhibited inhibitory influence on microtubule polymerization in the *in vitro* studies (Figure 84).

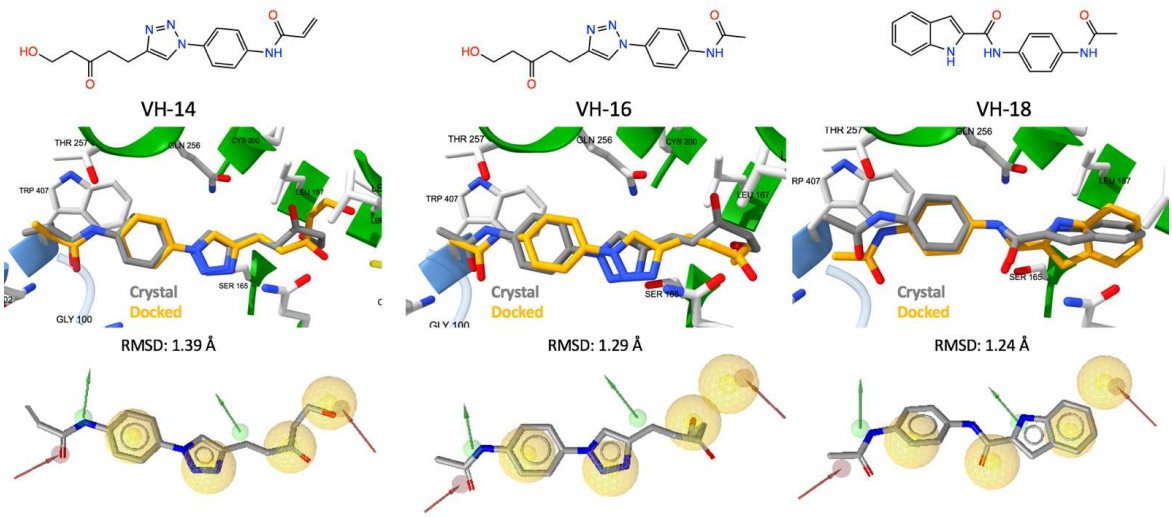


Figure 83. Three in-house hit molecules found by pharmacophore screening and protein-ligand docking

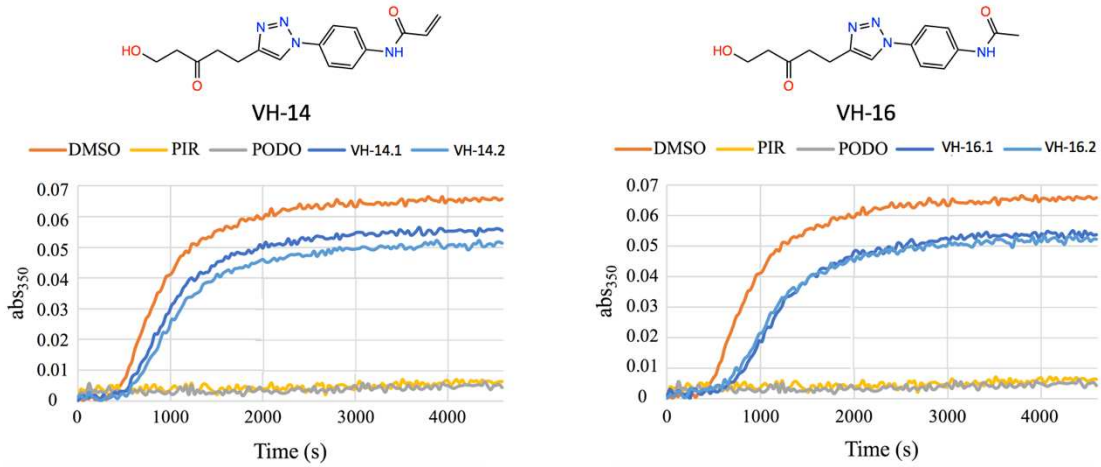


Figure 84. Microtubule polymerization inhibitory action shown by two in-house hit molecules

Two other hits from the in-house library (VH-15 and VH-17) that were poorly defined in the todalam site also caused notable inhibition of tubulin polymerization (Figure 85).

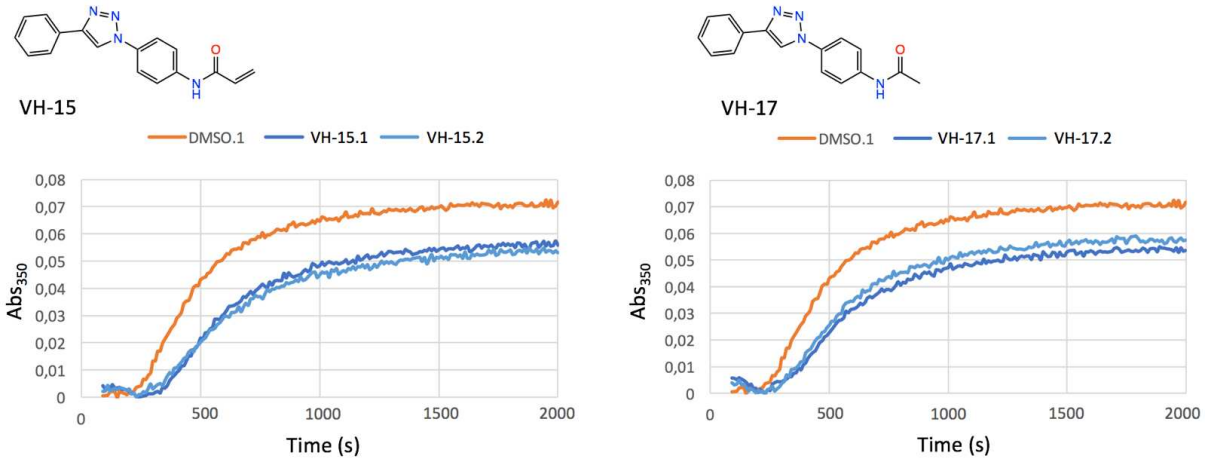


Figure 85. Microtubule polymerization inhibitory action displayed by hit molecules VH-15 and VH-17

Regarding the Enamine libraries hits, five out of thirteen compounds were crystallographically confirmed to bind at the todalam site. Crystal structure for one of the compounds (VH-12) could not be resolved due to poor resolution of the crystal structure. For the four other hits (VH-1, VH-5, VH-8, VH-11), their crystal poses are shown in Figure 86, overlapped with the best scoring docking poses calculated for them.

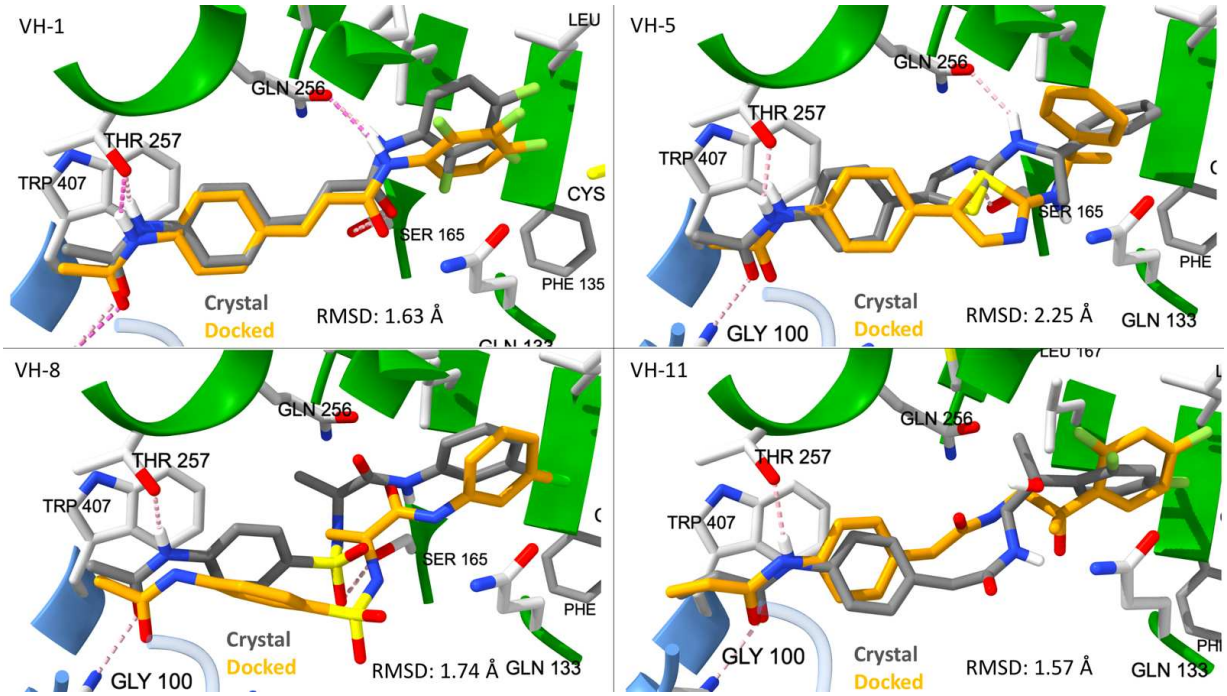


Figure 86. Overlap between the docked and experimentlally determined poses of four virtual hits in the todalam binding site

None of the 5 compounds had any effect on microtubule polymerization. Interestingly, one virtual hit from the Enamine libraries (VH-4) was not detected to bind at the todalam site, but

demonstrated a microtubule depolymerizing effect *in vitro*, comparable to that of todalam (Figure 87).

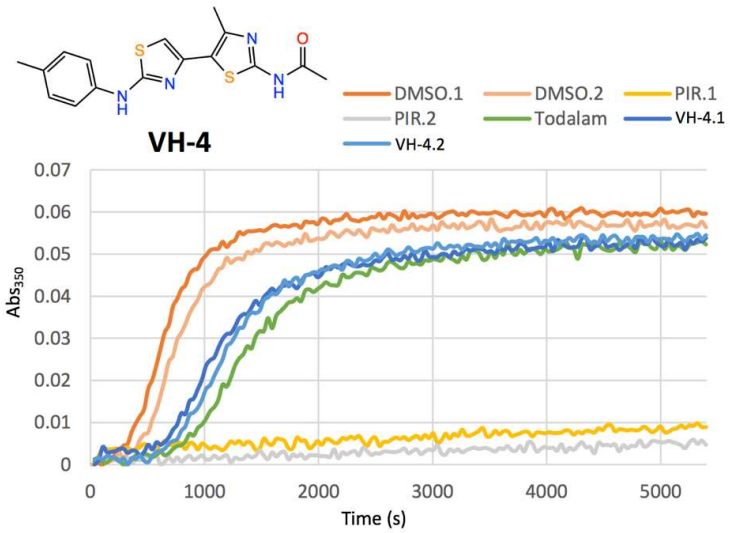


Figure 87. Virtual hit VH-4 demonstrated microtubule polymerization inhibition activity on a level comparable to todalam

In parallel, another virtual screening campaign, using a different computational approach, was performed by Dr. Helena Perez-Peña, in a collaboration between universities of Strasbourg and Milano under the TubInTrain consortium framework. Instead of pharmacophore screening, that work used substructure search to find purchasable molecules capable of targeting the todalam site in the desired binding mode. By combining the experimentally validated results of the virtual screening pipelines implemented in this thesis and the separate work of Dr. Perez-Peña, we were able to create a list of 7 distinct molecular scaffolds that bind to the todalam binding site in the binding mode that's similar to the native ligand's one (Figure 88).

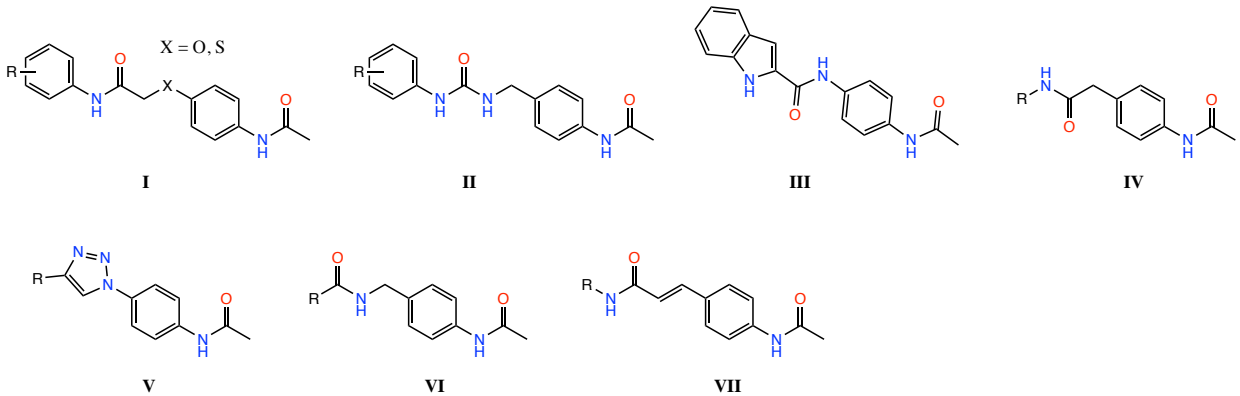


Figure 88. Six distinct scaffolds that target the todalam binding site

4.2.8. Results and discussion

Our virtual screening efforts of the in-house library have allowed us to select five molecules for synthesis and experimental evaluation. X-ray crystallography experiments have shown distinct binding of the three compounds to the todalam site. Two of the three bound in-house molecules caused notable microtubule polymerization inhibition action. Additionally, two other virtual hits from the in-house library, for whom only the anchoring moiety was determined in the site, were also shown to inhibit microtubule polymerization to some extent. In total, we obtained five hit molecules from this campaign.

As predicted by the docking computations, the binding mode of compounds VH-14 and VH-16 aligns well with todalam's binding mode. We observed that the triazole ring of these compounds is well-situated in the region of the binding site where todalam's thiazole fragment residues. Additionally, we established that both phenylacetamide and phenylprop-2-enamide fragments in the anchor moiety allowed for binding of the ligand at the site.

From the thirteen virtual hits coming from the pharmacophore screening of the Enamine libraries, X-ray crystallography experiments have outlined five molecules as binding to the todalam site. Bound poses of four hit molecules have been resolved. None of them had any effect on microtubule polymerization.

At the same time, another virtual hit from the Enamine libraries that was not detected in the binding site through crystallography was demonstrated to inhibit microtubule polymerization *in vitro*. Altogether, this screening campaign yielded six hit molecules.

In total, screening the in-house library and the Enamine libraries yielded eleven hit molecules. Experimental validation of the virtual hits allowed us to understand which molecular scaffolds facilitate binding with the todalam site. Additionally, we could estimate the length of the molecular fragments that is required to ensure a ligand's binding and a desired location of the ligand's outward-facing fragments in the hydrophobic pocket.

The reason for some molecules binding at the site without inhibiting tubulin polymerization, and others that follow the todalam's pharmacophore model but do not bind at the site while inhibiting polymerization, remains elusive. The possibilities may range from insufficient protein conformational changes upon ligand binding to the specifics of experimental conditions of the crystallography and *in vitro* tests. Further research is required to elucidate these findings.

The study broadened the structural diversity of molecules capable of binding to the todalam site. After it, we had a list of nine hit molecules, seven of which had a resolved structure in the todalam binding site after co-crystallization with tubulin, and the remaining two with notable microtubule polymerization inhibiting action. The diverse range of these molecules served as valuable starting points for the design of novel tubulin inhibitors discussed in the next section. The

unexpected effects of some molecules on tubulin polymerization also provided an intriguing insight into the complexities of protein-ligand interactions and the importance of experimental validation in conjunction with computational predictions.

4.3. Design of covalent todalam site binders

4.3.1. Overview of available data

Following the experimental testing of the virtual hits that were yielded by virtual screening of Enamine and our in-house libraries, we identified seven molecular scaffolds that bind to the todalam site. X-ray crystallography confirmed the *in silico* predicted orientations of these compounds within the binding site. Close to this hydrophobic domain in the a-tubulin lies the αCys4 cysteine residue. The aim of this project was to modify some of the found scaffolds by introducing a reactive functional group (a warhead) in such an orientation to promote a covalent interaction with this cysteine residue, leading to the creation of the first rationally designed covalent binder for the todalam site.

To achieve this, we first sought to evaluate the reactivity of the αCys4 residue within the binding site environment. Based on this data, we then planned to come up with a list of potential warheads for covalent bonding with the target cysteine. We would subsequently use the data on known molecular scaffolds that bind to the site, to either search for purchasable molecules with the required warheads, or generate a list of efficiently synthetically accessible scaffolds, seeking purchasable fragments that could be introduced into them to produce the desired covalent binders.

4.3.2. Estimating cysteine reactivity

While cysteine is one of the least abundant amino acids in many proteins, it plays a pivotal role in catalysis, signal transduction, and gene expression regulation[114]. With a pKa of ~8.5, its side chain thiol group can become deprotonated and thus nucleophilic under physiological pH conditions[114]. This behavior, which is unique among the natural amino acids, has triggered a surge of interest in cysteine-targeting warheads of diverse chemical compositions. Factors such as solvent exposure and dissociation of the thiol group into the thiolate anion influence cysteine reactivity[115]. Consequently, understanding the reactivity of the targeted cysteine is key for selecting an appropriate reactive group. Our objective was thus to examine the reactivity of the Cys4 residue of α-tubulin before attempting the design.

4.3.2.1. Literature review

Firstly, we surveyed the literature, looking for published results of experimental analysis of cysteine reactivity in tubulin. This led us to a study by Britto et al. which examined the reactivity of tubulin cysteines' SH groups for thioether formation[115]. They employed trypsin to cleave the tubulin protein downstream of every lysine or arginine, generating tryptic peptides. Post-separation by reverse-phase high-performance liquid chromatography, these peptides were analyzed using radio-labeled reactive probes and mass spectroscopy to determine cysteine residue interactions. This study identified the Cys4 residue of α-tubulin as exhibiting low reactivity.

4.3.2.2. Sequence-based machine learning model

Then, we attempted to use a recently published machine learning tool called sbPCR (sequence-based prediction of cysteine reactivity), to estimate the reactivity of the αCys4 residue[116]. The tool uses a "skip-gram"-like algorithm to generate motif features from local sequences containing cysteines, which are then passed to a pre-trained support vector machine (SVM) model to estimate reactivity, framed as a binary classification problem. This model also predicted the αCys4 residue as non-reactive when given a-tubulin's structure sequence from the 5SB7 PDB structure as input.

4.3.2.2. Structure-based pKa prediction

We also tried reformulating the problem as the task of pKa value estimation. The pKa value signifies the strength of a Bronsted acid, indicating how tightly it holds to a proton. A lower pKa indicates that a Bronsted acid can easily give up its proton, while a higher pKa suggests that the proton is more tightly held and is less likely to be released.

Determining pKa values using experimental methods, especially for complex biological systems, can be challenging. PROPKA is a widely used software for estimating pKa values based on protein three-dimensional structure[117]. It's favored because of its speed, accuracy, and ability to give insight into the structure behind the predicted pKa values. PROPKA calculates the pKa values of ionizable residues in a protein by considering how the protein's environment changes the modelled pKa value. Specifically regarding the reactivity of cysteine residues, PROPKA predicts the pKa based on features like hydrogen bonds, desolvation effects, and charge-charge interactions. These factors and their associated parameters are determined empirically. The method is designed to be computationally efficient and manageable, even for larger proteins or protein complexes, with the relationship between the perturbation and the structure explained by simple distance- and angle-dependent functions.

For our purposes, the pKa value predicted for the αCys4 residue was 12.15, based on an input of the todalam-hosting a-tubulin chain from the 5SB7 PDB structure, suggesting a low reactivity due to the difficulty in detaching a proton from the sulphur atom of this residue given the hydrophobic and solvent-inaccessible nature of the local environment.

4.3.2.3. Combined sequence- and structure-based approach

While most computational methods to evaluate cysteine reactivity are either sequence-based or structure-based, the Cpipe web server uses a combination of the two approaches[118]. The specifically developed HAL-Cy algorithm leverages both types of approaches to assess reactivity through parameters such as local hydrogen bond networks, solvent exposure, and resemblance to known nucleophilic cysteines. The different approached implemented in the algorithm act as an ensemble of weak predictors. Given an input three-dimensional structure, a majority consensus approach is then applied to their predictions: a cysteine is deemed reactive if predicted to be so by multiple methods, but if it was just one method, the prediction is putative.

We utilized Cpipe to predict the reactivity of all cysteine residues in the a-tubulin chain extracted from the 5SB7 PDB structure, seeking to compare these predictions with the experimental results reported by Wolff et al. The Cpipe tool's final predictions regarding all cysteine residues in a-tubulin were in good agreement with Wolff et al.'s findings. However, the Cpipe tool also predicted low reactivity for the αCys4 residue due to the nature of its surrounding binding site environment.

These findings suggest a challenge in targeting the αCys4 residue due to its possible low reactivity. Proceeding further by considering potential warheads to target the αCys4 residue within the todalam binding site, we accounted for these data. In a similar setting, Lu et al. have demonstrated that by employing a highly reactive warhead in combination with high-affinity binding molecular scaffold, it is possible to target non-reactive cysteine residues[119]. Moving on, we aspired that by employing a highly reactive warhead and optimizing the ligand structure for strong interaction with the binding pocket, we can maximize the exposure time and thus facilitate the formation of a covalent bond with the αCys4 residue.

4.3.3. Literature search for cysteine-targeting warheads

Our objective at this stage of the project was to design compounds capable of forming a covalent bond with the cysteine residue within the todalam binding site. To this end, we needed a diverse selection of reactive groups (also called "warheads") that could be incorporated into our molecules to ensure the covalent bond formation. The αCys4 residue's predicted non-reactivity, based on literature search and *in silico* modelling, required careful curation of such list.

To compile this list, we referred to three comprehensive databases: CovPDB[120], CovalentInDB[121], and CovBinderInPDB[122].

CovPDB, a public-access web database, is an exhaustive resource of high-resolution 3D structures of covalent protein-ligand complexes, gathered from the Protein Data Bank. Created to assist structure-based approaches in chemical biology and drug design, it helps in identifying warheads and reaction mechanisms that lead to covalent modification of the targetable residues in the binding sites. The information in this database is manually annotated by experts. The CovPDB database encompasses 2294 covalent complexes, 93 reactive warheads, 21 covalent mechanisms, and 14 targetable residues.

Likewise, the CovalentInDB (Covalent Inhibitor Database) is a vast web repository for covalent inhibitors and their corresponding targets. Its latest version has data on 8561 covalent inhibitors and 343 related protein targets, garnered from comprehensive literature research.

Finally, the CovBinderInPDB database encompasses 7375 covalent modifications mined from the PDB database, with 2189 unique covalent binders targeting nine types of amino acid residues (including cysteine) from 3555 protein-ligand complex structures.

These databases collectively offer extensive data on nearly all known protein complexes with covalently bound ligands. We leveraged all three to extract as much information on covalent warheads as possible.

From this research, we curated a list of 31 potential warheads, which we categorized into five groups based on their structure (Figure 89). Figure 90 shows the example possible mechanisms of cysteine reacting with each group of warheads. We intended for our potential covalent binders to the todalam site to incorporate one of these groups within molecular scaffolds known to bind at the todalam site.

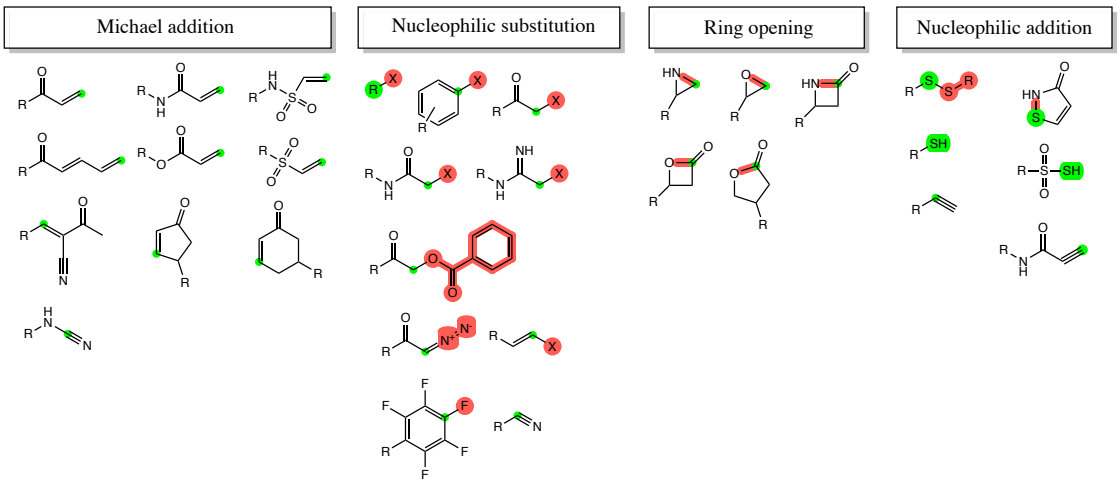


Figure 89. Widely-used reactive groups mined from the three databases of covalent protein-ligand complex data. Red color highlights leaving group's atoms or breaking bonds. Green

color highlights atoms that may get attacked by the nucleophilic sulfur atom in the thiolate anion of the cysteine residue.

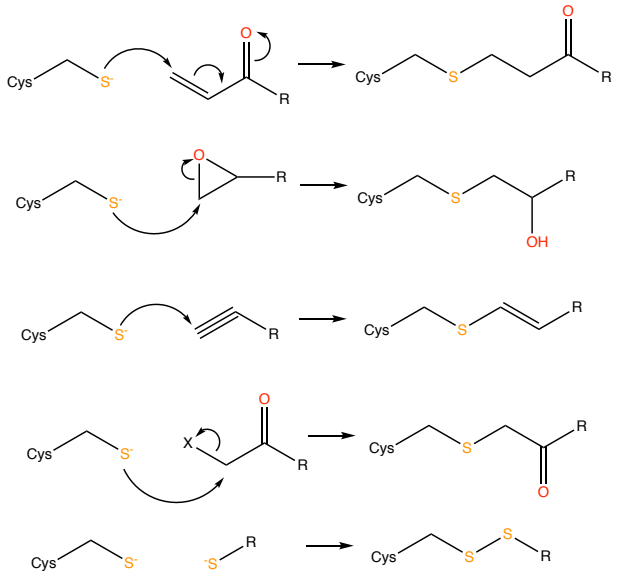
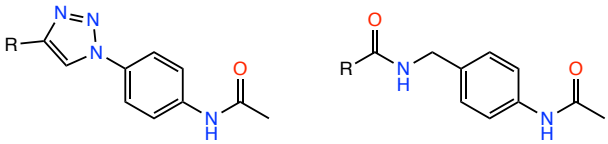


Figure 90. Possible reaction mechanisms between cysteine residue and several popular warhead types

Subsequently, our strategy involved identifying purchasable molecules that follow the well-binding molecular scaffolds and contain any of the identified warheads. Furthermore, out of the seven established molecular scaffolds suitable for binding, we chose scaffolds **V** and **VI** for additional modifications (Figure 91). The practical reason being, molecules from these scaffolds can be synthesized efficiently in one or two-step reactions from readily available starting components. Hence, one of our goals was to identify warhead-containing purchasable fragments that can be utilized in these one-step synthesis reactions to yield cysteine-targeting modifications of well-binding molecular scaffolds.



V **VI**

Figure 91. Scaffolds subjected to further modifications to obtain a covalent binder

4.3.4. Virtual screening for warhead-containing analogues of confirmed binders

Initially, we aimed to identify purchasable molecules that aligned with any of the seven established well-binding scaffolds and contained any of the 31 shortlisted warheads. For this purpose, we planned to conduct a substructure search using the generic structures of the seven scaffolds as queries.

To this end, we chose the ZINC library[123], comprising 727,549,993 purchasable compounds and small molecular fragments from various vendors, as our source for this search. Prior to computational work, the library was standardized using the standardization procedure, identical to the one outlined in section 4.2.3. Briefly, this involved structure aromatization, dealkylation, conversion to canonical SMILES strings, removal of salts and mixtures, species neutralization, and tautomer generation using ChemAxon tools.

Given the extensive size of the library, we were looking to develop an efficient approach for substructure search. For this, we used the substructure library class implemented the RDKit chemoinformatics toolchain. We first divided the large library into 73 segments, each consisting of 10,000,000 compounds (except for the 73rd segment which included the remaining 7,549,993 compounds). We then calculated special pattern fingerprints for each compound in each segment. This fingerprint type is unique to RDKit. Such fingerprints are used to detect molecular features through substructure searches using a limited number of very generic SMARTS patterns.

Pattern fingerprints act as pre-filters, indicating whether a substructure search is required for a given query. These fingerprints are calculated once for the whole dataset and stored locally as binary objects for quick reusability.

Hence, we built a substructure library correlating the SMILES strings of each compound in each segment with the locally stored hashed pattern fingerprints. This allowed for efficient querying of the library class by a required SMARTS pattern. The substructure library class object by default employs all available CPU threads to parallelize the substructure search, further accelerated by pre-filtering based on binary pattern fingerprints.

This method significantly expedited the screening process. As such, the screening of the entire ZINC library for a single query was reduced from a day with a straightforward looped RDKit substructure search to roughly an hour, indicating a substantial performance boost.

Once we have constructed the substructure library class object and populated it with SMILES and pattern fingerprints for the ZINC compounds, we queried it with the generic structures of the seven effective molecular scaffolds. The substructure search yielded 4853 potential hits. To refine this list, we conducted a simpler substructure search using RDKit's

SMARTS pattern matching capabilities, which resulted in 2188 compounds that contained any of the 31 warheads at any moiety within their structure.

To further refine this list, we conducted protein-ligand docking of these molecules into the todalam binding site using the PLANTS program, along with the native ligand itself. The binding site was defined as all atoms of all residues within an 8 Å radius of todalam's bound pose. This included both the α-tubulin and β-tubulin parts of the binding site. Solvent molecules, small organic molecules, and metal ions were removed from the protein. Both the protein and all ligands were prepared using the SPORES software. Ten poses were calculated for each compound, each described by the docking score value of the *chemplp* scoring function and four derivative ligand efficiency scores, as described in section 2.3.6.

Following this, molecules were ranked by the docking score value of the best-scoring pose, and those scoring worse than the native ligand were excluded. This left us with 1018 compounds. A Pareto front optimization was then applied over the docking and ligand efficiency scores, ensuring that compounds were selected based on their valuable interactions with the binding site, rather than the sheer number of atoms they contained. This further narrowed the list down to 65 potential hits.

Subsequent visual inspection of their best docked poses specifically assessed the location of the warhead within the site. Molecules unable to interact with αCys4 due to their length were discarded, leaving only compounds that placed the warhead within 3.7 Å to the αCys4 residue. The threshold distance of 3.7 Å was chosen arbitrarily as it was considered to allow for potential bond formation.

This resulted in a refined list of 9 potential hits (Figure 92). Only three of these could be procured at the time: FS-7, FS-8, and FS-9. These were purchased from the Ambinter chemical vendor company for experimental testing. The three procured molecules underwent the standard experimental tests outlined in this thesis: X-ray crystallography and microtubule polymerization bioassay. The exact details of the experimental setup as described in section 2.3.7.

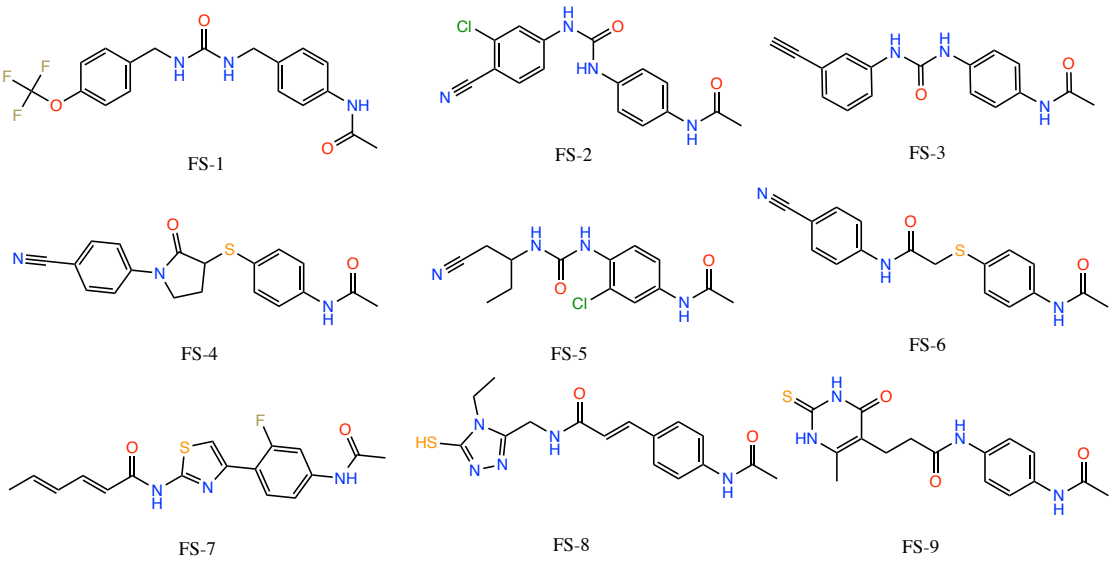


Figure 92. Virtual hits found after exhaustive substructure search in the ZINC library

Our results demonstrated that of the three procured virtual hits, only one molecule was definitively identified by X-ray crystallography to bind at the todalam site. Regrettably, no covalent bond was formed with αCys4 in this instance. The sulphur atom of the ligand that was expected to partake in a disulphide bond formation with the sulphur atom of αCys4 is located 4.42 Å away from the targeted cysteine's sulphur atom in the crystal structure, seemingly pushed away upon ligand binding. Figure 93 shows the overlap between the experimentally determined pose of hit FS-8 with the docking result, showing the change in αCys4 position between the rigid system used for docking (orange) and its experimentally determined position upon ligand binding (light gray).

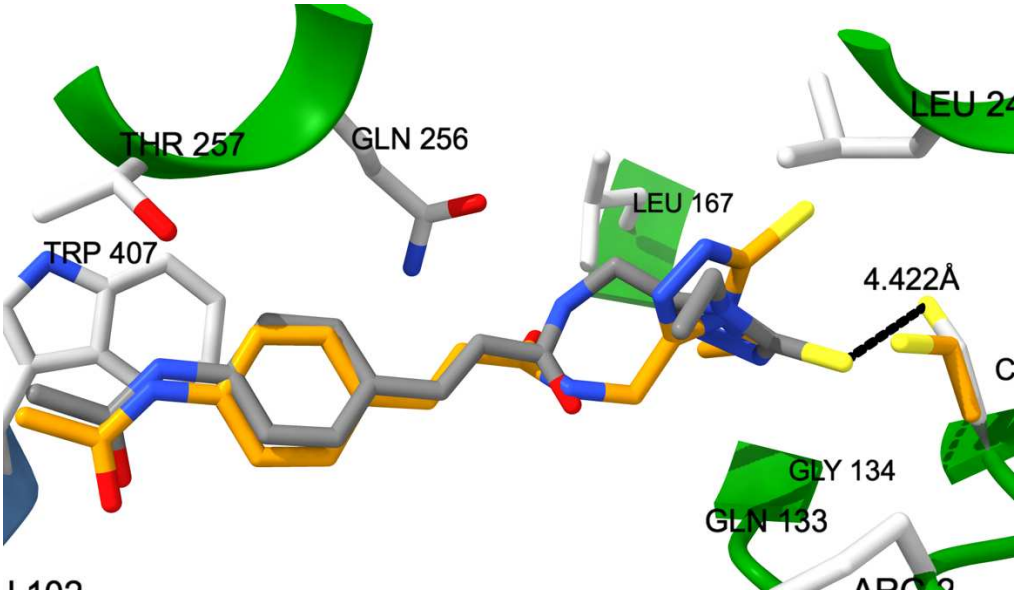


Figure 93. Docked (orange) and experimentally resolved (gray) poses of FS-8 in the todalam binding site. Notice the change in αCys4 position, orange - used in the rigid system for docking, light gray – experimental, after ligand binding.

However, another virtual hit molecule, despite not being detected as binding to the todalam site, has shown pronounced microtubule polymerization inhibition action *in vitro* (Figure 22).

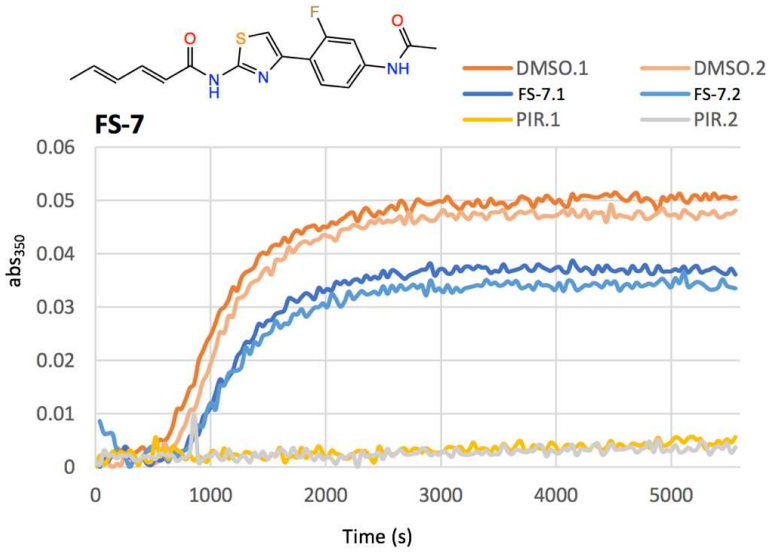


Figure 94. One of the virtual hits found in the ZINC library has a noticeable microtubule polymerization inhibitory action

Thus, further modification of these two molecules could potentially result in the desired formation of a covalent bond with αCys4, opening up new opportunities for drug development.

4.3.5. Optimization of scaffold V

Our TubInTrain consortium collaborators from the synthetic chemistry group of Prof. Daniele Passarella at the University of Milan, Italy, have profound expertise with the copper-catalyzed azide-alkyne cycloaddition reaction that facilitates the efficient and selective synthesis of 1,2,3-triazoles from alkyne and azide-containing compounds[124]. These triazole-containing molecules retain their structural stability in physiological media. Although our previous observations suggest that triazole-containing hits may not form as specific interactions at the todalam site compared to amide, acrylamide, or urea molecular scaffolds, the synthetic feasibility of triazole-containing molecules makes them a compelling focus. Our goal at this stage of the project was to design derivatives of this scaffold that could accommodate a reactive group proximate to αCys4 upon binding.

The particular reaction that results in molecules adhering to scaffold **V** involves a molecular fragment with a terminal alkyne bond and another fragment with an azide group. Therefore, we aimed to find purchasable "double-sided" small fragments, which needed to have any of the 31 shortlisted warheads on one side and a terminal triple bond on the other. We allowed up to four atoms (or small functional groups) in between the two, which should not be in a ring

structure and could include: a simple -CH_2- group; an oxygen atom; a nitrogen atom with 1 explicit hydrogen; a carbonyl group; a carbon atom connected with an alcohol functional group with 1 explicit hydrogen on the oxygen atom; a carbon atom connected with an amine functional group with 2 explicit hydrogens on the nitrogen atom; or an enamine group with 1 explicit hydrogen on the nitrogen atom (Figure 95). Our reasoning was that these specific types of atoms in between the two essential moieties would ensure the ligand's linear character to easily penetrate into the binding site, and at the same time establish useful interactions with the residues in the binding site.

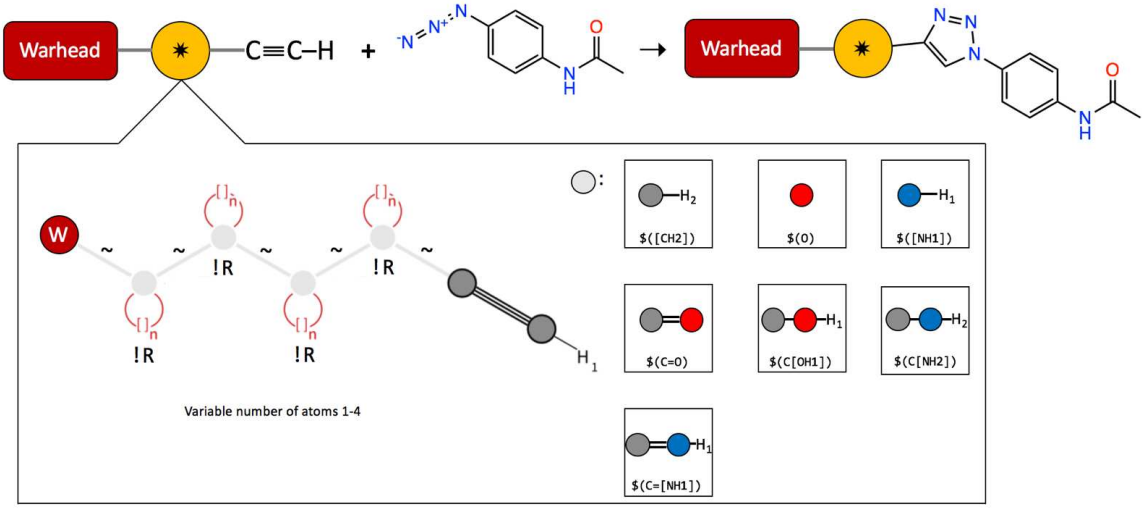


Figure 95. Overview of the azide-alkyne cycloaddition reaction and the SMARTS patterns used to find required fragments

Following this approach, we created 124 SMARTS pattern strings (with variable lengths of the linker chains (1-4) between the 31 warheads and the terminal alkyne). We then employed the previously described RDKit-based substructure screening routine to efficiently identify 439 purchasable fragments from the ZINC library that adhered to the rules outlined by the SMARTS patterns. We further enumerated the possible reaction products between all 439 fragments and the N-(4-azidophenyl)acetamide fragment, resulting in 439 warhead-containing derivative molecules of the scaffold **V**.

Next, we needed a method to assess how likely these molecules would be to bind and form a covalent bond with the αCys4 residue in the site. Therefore, we decided to perform two types of docking simulations for these compounds. The first type of simulation allowed the molecules to be conformationally sampled in an unconstrained manner in a rigid protein environment, similar to all docking campaigns described so far in this thesis. This approach provides a docked pose with the best docking score value, interpreted as the most likely placement of the molecule within the site, if we assume that it does bind.

The second type of simulation constrains the optimization space of a molecule's conformations by modelling the system as if a covalent reaction between the ligand and the

cysteine residue has occurred. For each ligand, we identified the targeted atom (and occasionally a leaving functional group). We then deleted the leaving group, if needed, and introduced a new bond between the ligand's reactive atom and a methanethiol fragment. By using this new methanethiol-bound structure to replace the cysteine residue in the binding site, we could model potential poses that a ligand could have in the site if the reaction proceeded by the mechanism we envisaged. Only the sulphur atom of the methanethiol group in the cysteine residue is allowed to move, as the backbone of the cysteine residue remains rigid. If the best pose following covalent docking closely resembled the best pose after unconstrained docking, we considered a reaction with the cysteine residue probable. If a ligand found another conformation not similar to the unconstrained docking pose, we considered the reaction less likely, but still possible. However, if the covalent docking approach could not identify any conformation of the bound ligand in the site, we considered a covalent reaction unlikely.

Applying this logic, we performed unconstrained docking using the PLANTS software on all 439 molecules, along with the native ligand, todalam, as a reference. The binding site was defined as all atoms within an 8Å radius from the center of mass of the native ligand. The protein was prepared by removing solvent molecules, ions, and other small organic molecules. Both the protein and ligands were pre-processed using the SPORES software before docking. We used the *chemplp* scoring function and generated 10 poses for each ligand, each pose characterized by a docking score value.

After performing unconstrained docking, the compounds were ranked by the value of their best pose's docking score, which reduced the number of considered compounds to 50. For these selected compounds, we performed covalent docking using the AutoDock 4 software. The binding site was similarly defined as all atoms within an 8Å radius from the center of mass of the native ligand, todalam. Manual structural editing was carried out for each ligand in the ChemAxon MarvinSketch program. Initial low-energy conformations of the methanethiol-bound ligands were computed using MarvinSketch's embedded conformational sampler tool utilizing the MMFF94 force field. Superimposition and docking preparations were conducted using scripts from the AutoDock suite. The grid parameter file and docking parameter file were created using default settings, including the default AutoDock scoring function. We then configured the software to calculate 10 poses for the covalently-bound ligands.

Once covalent docking was complete, we identified 16 molecules for which no covalently bound pose could be produced, excluding them from further investigation. This left us with 34 remaining virtual hits to consider. Among these, 13 had their covalently bound pose within the site, directed towards b-tubulin, and 19 had a covalent pose directed outside, not towards the pocket in the b-tubulin.

When considering which hits to synthesize, we also took into account the price of the identified fragments. The 34 fragments used to generate these compounds were sourced from vendors such as Ambinter, AKOS, Enamine, Fluorochem, Sigma Aldrich, Key Organics, BLD pharm, and TCI. The price per gram of these fragments ranged from 56 EUR to 2020 EUR. Following extensive consultations with our colleagues from the TubInTrain consortium, we shortlisted 15 fragments for purchase and subsequent synthesis of modified scaffold **V** compounds (Figure 96).

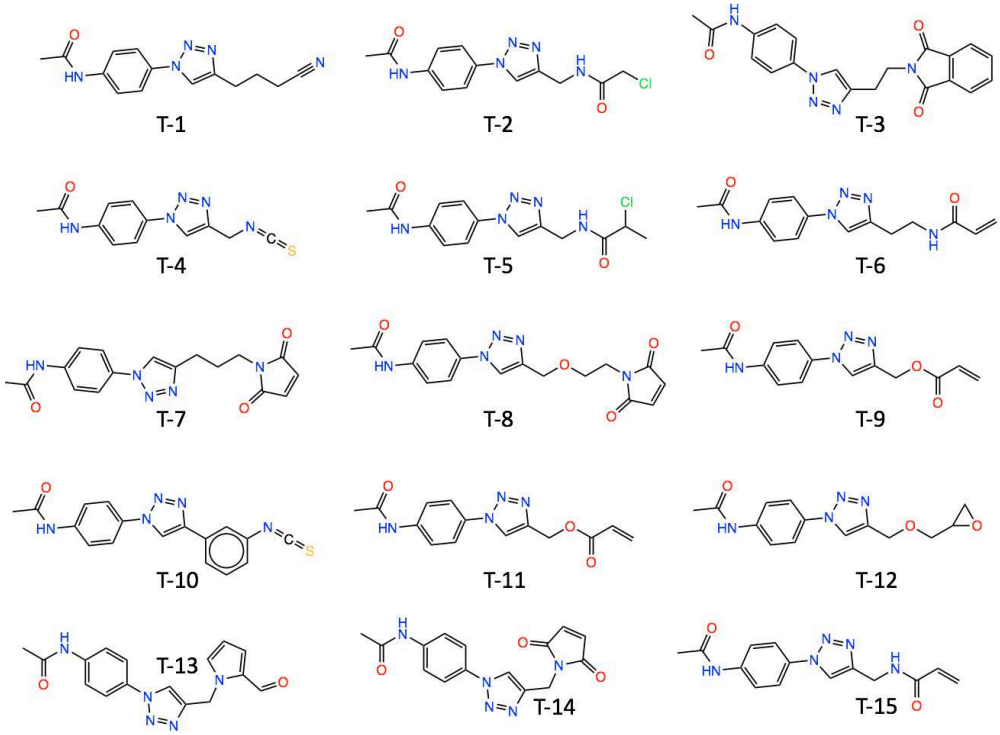


Figure 96. Fifteen virtual hits enumerated from small fragments found in the ZINC library

The synthesis of these compounds was performed by our TubInTrain collaborators, specifically Dr. Zlata Boiarska from the group of Prof. Daniele Passarella, University of Milano, Italy. Upon completion, the synthesized compounds were subjected to two experimental tests: X-ray crystallography and a microtubule polymerization assay, as defined in section 2.3.7. X-ray crystallography detected three of the synthesized compounds in the todalam site (**T-6**, **T-11**, **T-12**). Of these, compound **T-11** demonstrated strong binding affinity, particularly in the presence of vinblastine (Figure 97). Vinblastine is known to induce certain conformational changes in the todalam binding site, opening it up and providing more space for ligand binding. The difference between the docked and crystallographically resolved poses is 1.64 Å. We did not observe the formation of the covalent bond between the Michael acceptor prop-2-enoate fragment of **T-11** and the αCys4 residue. In the crystallographically resolved pose, the distance between the reactive atoms of ligand and binding site is 3.58 Å. The two other compounds (**T-6**, **T-12**), on the other hand, were not soaked with vinblastine, and only had the electron densities for their anchor parts

defined in the site, implying their intrinsic flexibility in the site. We also did not observe the formation of the desired covalent bond for any of them, neither did we observe any microtubule polymerization-related action from them.

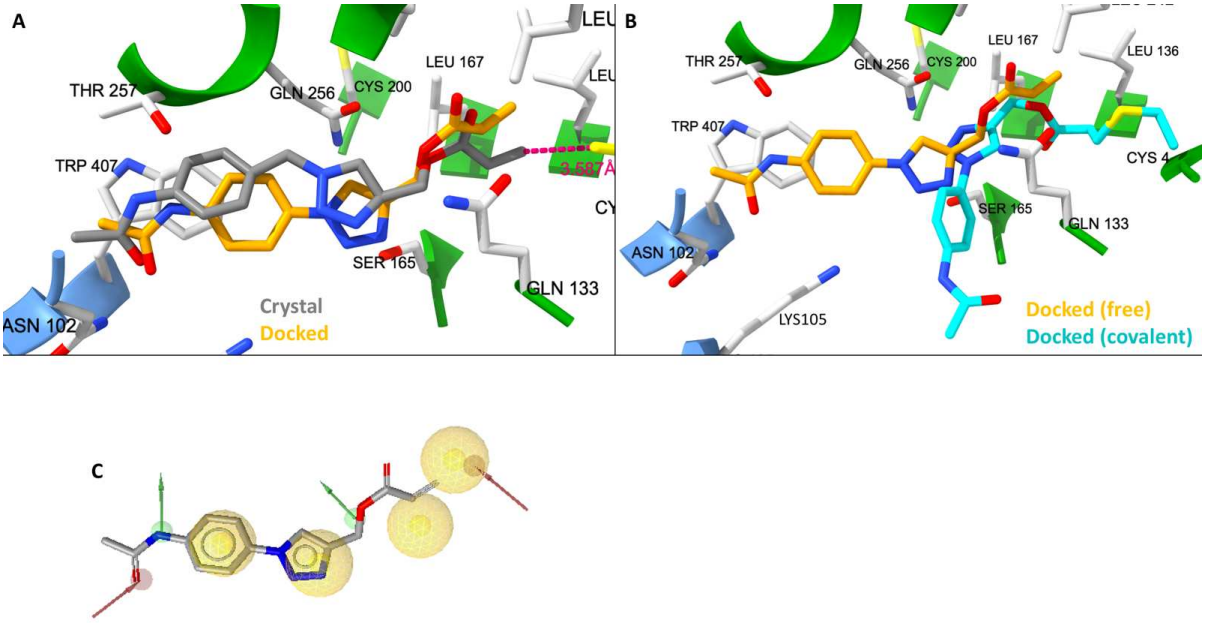


Figure 97. (A) Comparison of the best docked pose of T-11 (orange) and the experimentally determined pose (gray); (B) Overlap of best pose after free docking (orange) and best pose after covalent docking (teal); (C) overlap between T-11 best docked pose and todalam's pharmacophore model

However, an interesting observation was that one of the compounds, though not visible in the binding site, exhibited a substantial inhibitory effect on microtubule polymerization. This inhibition was even more pronounced than that caused by todalam (Figure 98).

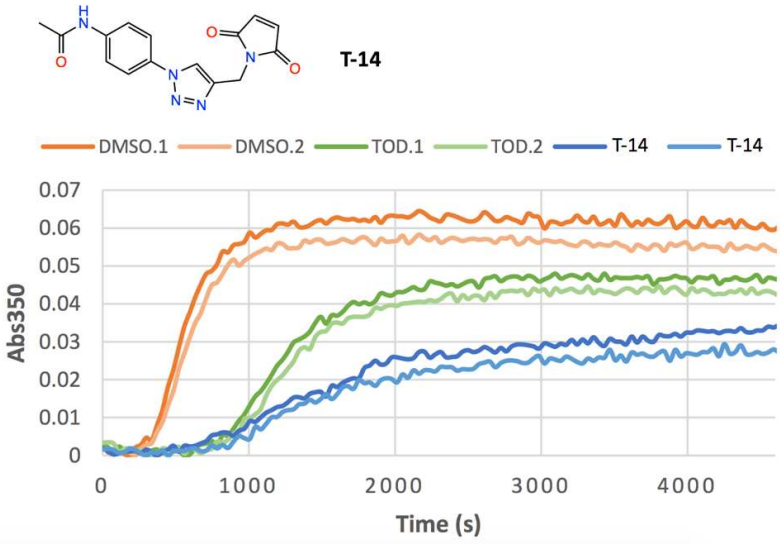


Figure 98. Hit molecule T-14 has a pronounced microtubule-depolymerizing effect

To summarize, the combination of substructure search, reaction product enumeration, and both free and constrained docking simulations resulted in the identification of four scaffold **V** derivatives. Three of these derivatives (**T-6**, **T-11**, **T-12**) exhibited some level of binding to the todalam site, and one (**T-14**) demonstrated a significant inhibitory effect on microtubule polymerization. The binding characteristics of the four hits were not ideal, but further optimization could potentially improve their affinity for the todalam site. The compound with a clear microtubule polymerization inhibitory effect contains a highly reactive maleimide warhead, and it's unclear whether its absence from the binding site is due to reactivity with other cysteine residues on the tubulin surface. This aspect warrants further investigation.

4.3.6. Optimization of scaffold VI

In collaboration with the TubInTrain consortium members, we also decided to explore covalent optimization of scaffold **VI**. Economic considerations and synthetic feasibility of this scaffold's derivatives were instrumental factors in this decision. The envisaged synthesis involved generating an amide bond between two molecular fragments: one with a reactive warhead and a carboxylic acid functional group, the other being an amine, specifically, N-[4-(aminomethyl)phenyl]acetamide (Figure 99).

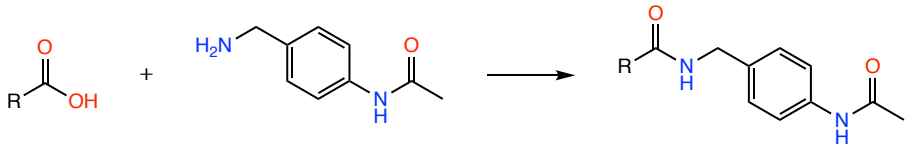


Figure 99. Suggested reaction scheme to obtain scaffold VI

Given the effective strategy employed in the modification of scaffold **V** (section 4.3.5), we decided to adopt a similar approach here. The aim was to find small, purchasable molecular fragments with dual-sided characteristics; having a warhead and a carboxylic acid functional group separated by up to four atoms or functional groups. The SMARTS patterns applied for this search mirrored the ones used to screen for fragments modifying scaffold **V** (section 4.3.5), the sole difference this time being the terminal group of the SMARTS patterns was a carboxylic acid rather than a terminal alkyne.

Similarly, we employed RDKit's efficient substructure screening algorithm to perform a substructure search in the ZINC library of purchasable compounds and small molecular fragments. This search produced 237 purchasable fragments containing the carboxylic acid functional group and at least one of the 31 desired warheads. However, some fragments featured multiple carboxylic acid groups or additional alcohol groups, which could potentially disrupt the regioselectivity of reactions. Upon eliminating these, 156 fragments of interest remained.

Following the strategy outlined in 4.3.5, we then enumerated potential products of reactions between the 156 purchasable fragments and the N-[4-(aminomethyl)phenyl]acetamide fragment. Next, docking simulations were executed for these 156 scaffold **VI** products. Firstly, the molecules were docked alongside todalam in an unconstrained fashion using the PLANTS software. We used the same binding site definitions and software parameters described in section 4.3.5. The molecules were then ranked based on the docking score values of their best-scoring poses. We excluded compounds with a worse docking score than todalam, leaving 42 compounds for the covalent docking step. Upon executing the covalent docking step (using the setup described in section 4.3.5), we narrowed the list down to 21 compounds.

The economic feasibility of synthesis remained a crucial factor. Most fragments identified in this screening were from the Enamine and Ambinter vendors, priced between 75 and 1125 EUR. Therefore, we filtered out the most expensive fragments, finalizing a list of six virtual hits suggested for synthesis and evaluation (Figure 100). Figure 101 shows the results of unconstrained and covalent docking simulations for these molecules. As can be seen, the calculated covalently bound poses mostly remain in the binding site and, in most cases, overlap the calculated unconstrained poses, which in turn retain the todalam's binding mode.

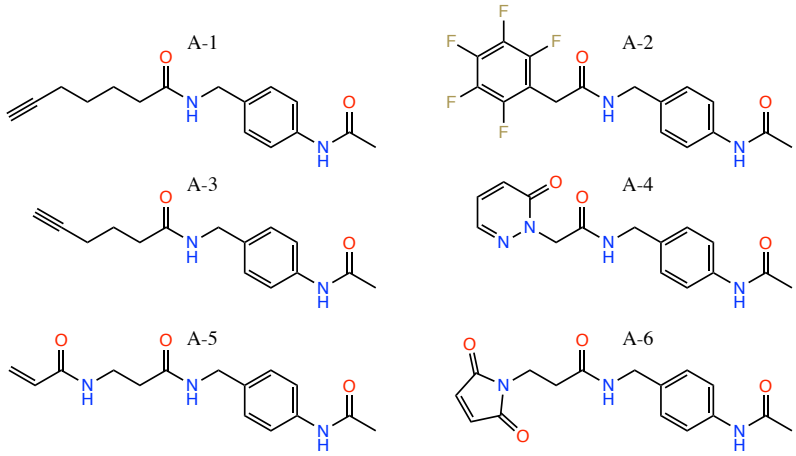


Figure 100. List of suggested scaffold VI derivatives for synthesis and evaluation

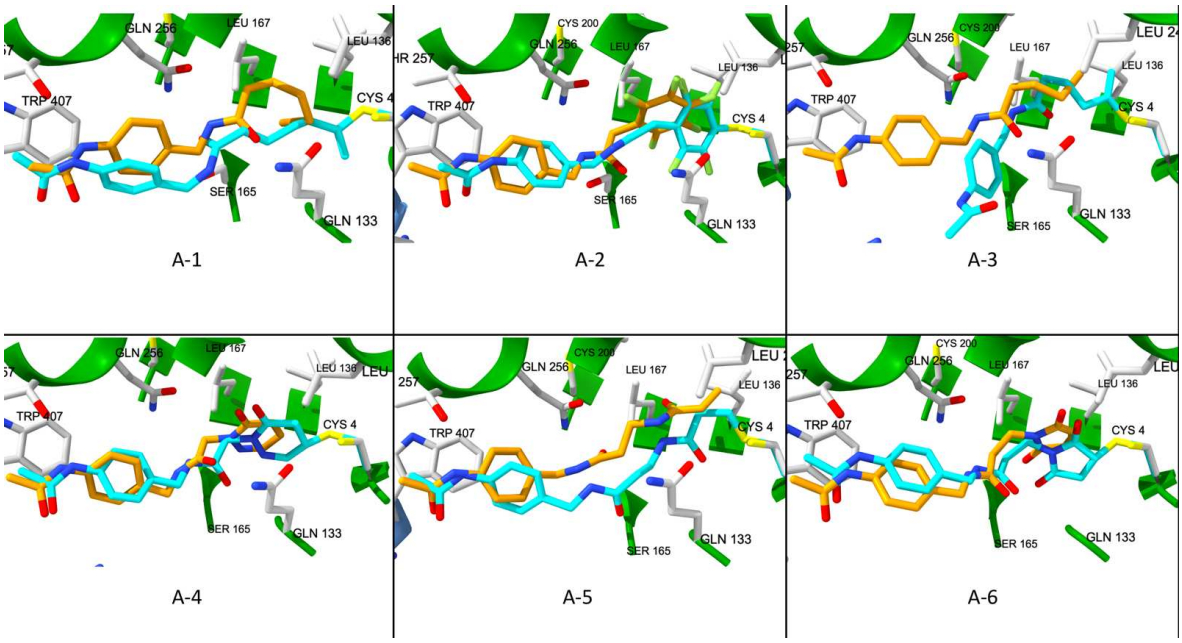


Figure 101. Results of unconstrained (orange) and covalent (cyan) docking for virtual hits A1-6

Upon synthesis, performed by Dr. Zlata Boiarska at the group of Prof. Daniele Passarella, University of Milano, Italy, the compounds underwent two experimental tests: X-ray crystallography and a microtubule polymerization bioassay, performed by our collaborators within the TubInTrain consortium as described in 2.3.7. X-ray crystallography confirmed the binding of four molecules, albeit not covalently. Bound conformations could be determined for two hits, **A-1** and **A-3** (Figure 102), while for two others – **A-2** and **A-5** – the electron density in the binding site is evident and is under ongoing resolution (Figure 103). Compound **A-1** is of particular notice as in the resolved structure, the distance between the warhead and the cysteine residue's sulfur atom is 2.579 Å. Compounds **A-4** and **A-6** were not seen in the todalam binding site in crystallographic experiments. Noticeably, despite binding to the site, neither of the compounds displays any action towards modulating tubulin polymerization *in vitro* in the standard microtubule polymerization bioassay.

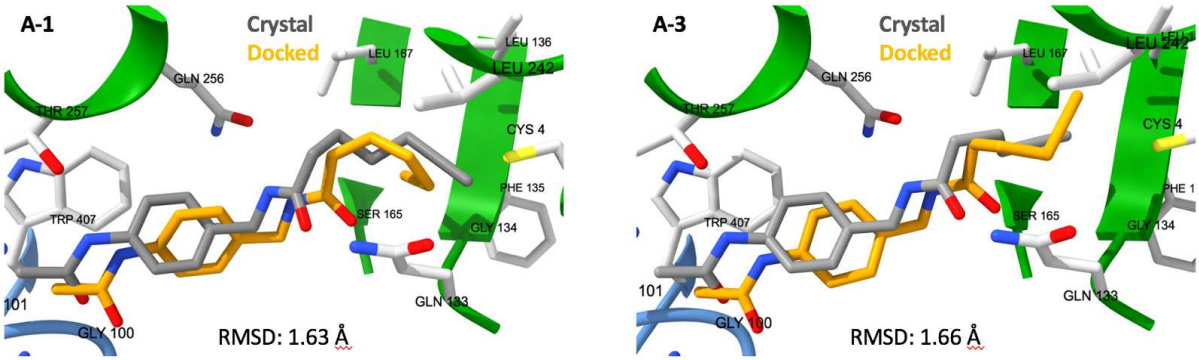


Figure 102. Comparison of the docked and experimentally determined poses shows good overlap

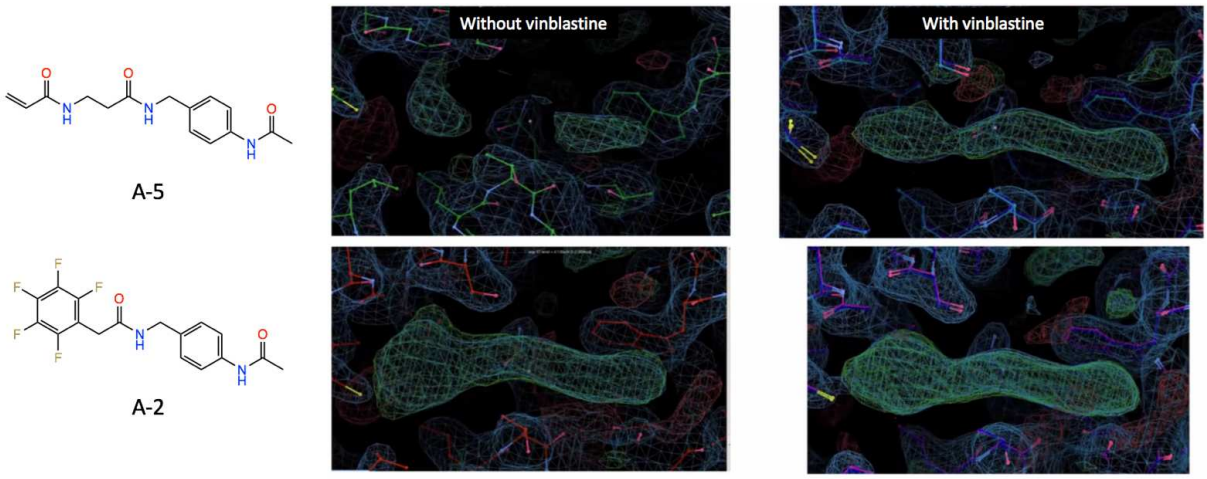


Figure 103. Experimentally determined electron densities for compounds A-5 and A-2 within the todalam binding site

However, one of the two virtual hits, undetected in the binding site (**A-6**), demonstrated a pronounced inhibitory effect on microtubule polymerization (Figure 104). Interestingly, the virtual hits with a significant inhibitory impact on microtubule polymerization, namely **T-14** and **A-6**, featured the maleimide reactive group, which is renowned for its high reactivity stemming from the electrophilic character of the carbon atom, which makes it susceptible to nucleophilic attack, especially from thiols such as the side chain of a cysteine residue. This allows it to form stable covalent bonds with proteins. However, this heightened reactivity could also lead to off-target interactions, implying potential non-specific action. As a result, fragments bearing this group could possibly react elsewhere on tubulin, perhaps binding somewhere on the protein's surface, and interfere with microtubule polymerization, but remain mobile, and hence, undetectable in the crystal structure.

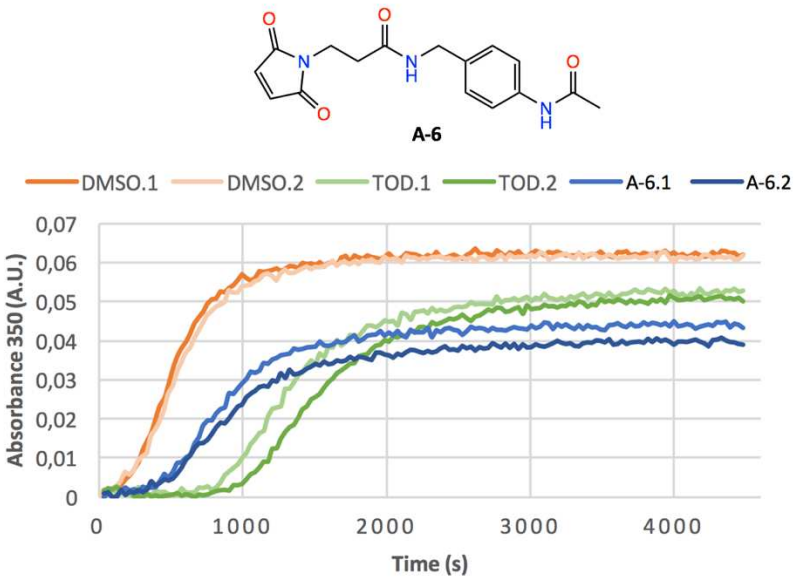


Figure 104. Hit molecule A-6 demonstrates considerable microtubule polymerization inhibitory action

Consequently, we deduced that scaffold **VI** provides a promising avenue for further molecular optimization. The warhead-bearing fragments required for synthesis are readily available. By selecting fragments with inherent linear structure, it is possible to have the synthesized derivatives bind to the todalam binding site. In this work, we used a substructure screening and combined free and constrained protein-ligand docking approaches to shortlist a set of six small fragments, that, upon synthesis into scaffold **VI** derivatives, produced four hit molecules that bound to the targeted site (**A-1**, **A-2**, **A-3**, **A-5**) and one hit molecule that did not bind to the site, but showed significant effect on microtubule polymerization (**A-6**).

4.4. Conclusion and perspectives

In conclusion, our research has substantially expanded our understanding of the small ligand chemistry that can be used to target the todalam binding site. We successfully devised several efficient and accurate virtual screening workflows, which included substructure search, pharmacophore screening, unconstrained rigid protein-ligand docking, and constrained (covalent) rigid protein-ligand docking.

Initially, we explored our in-house chemical library and the Enamine libraries of purchasable compounds, identifying eleven structurally diverse hit molecules. These served as a foundation for the strategic design of a molecule that could form a covalent bond with the site.

In the subsequent phase of the project, our focus shifted towards generating derivatives of established molecular scaffolds. The goal was to design molecules that would covalently interact

with the αCys4 residue within the binding site. We narrowed down a selection of warheads that could specifically target a low-reactive cysteine residue. At this stage of the project, we searched for purchasable analogues of known binding molecules that already included a warhead of interest. This effort led to the discovery of two additional hits that targeted the site and displayed inhibitory action on microtubule polymerization.

Our final phase focused on generating derivatives of two easily-accessible molecular scaffolds with the aim of covalently targeting the cysteine residue in the site. We searched for dual-sided small fragments that included any of the warheads on one end, and a functional group needed for the synthesis of scaffold derivatives on the other. This strategic design of known scaffold derivatives yielded eight promising hit molecules, some of which showcased strong microtubule polymerization inhibition along with good binding characteristics.

Ultimately, our research has significantly advanced the realm of rational small molecule design targeting the tubulin protein. In total, we have discovered 21 hit molecules with promising binding modes and, in some cases, potent inhibitory action on microtubule polymerization. Further exploration is warranted to unravel the intricate interplay between binding to the site and effective inhibition of microtubule polymerization.

Chapter 5. *De novo* design of colchicine site-targeting agents using the inverse QSAR approach

5.1. Introduction

Positioned at the interface between the α and β subunits of the tubulin heterodimer, the colchicine site is predominantly nestled within the β-subunit[39,125]. It is made by the residues of the T7 loop, H7 and H8 helices, and the S8 and S9 strands of β-tubulin, complemented by the T5 loop of α-tubulin[5] (Figure 105). The site features three hydrophobic pockets, which serve as key locations for ligand interaction, along with two hydrophilic regions capable of forming additional stabilizing hydrogen bonds with ligands[39]. This effectively divides the site into a central zone (Figure 105, zone 2, red) and two additional zones, one facing the α-tubulin subunit (Figure 105, zone 1, orange) and the other residing deeper within the β-tubulin subunit (Figure 105, zone 3, pink)[5,39]. Agents targeting the colchicine site usually occupy zones 1 and 2 or zones 2 and 3, but no known ligand simultaneously occupies all three zones[5].

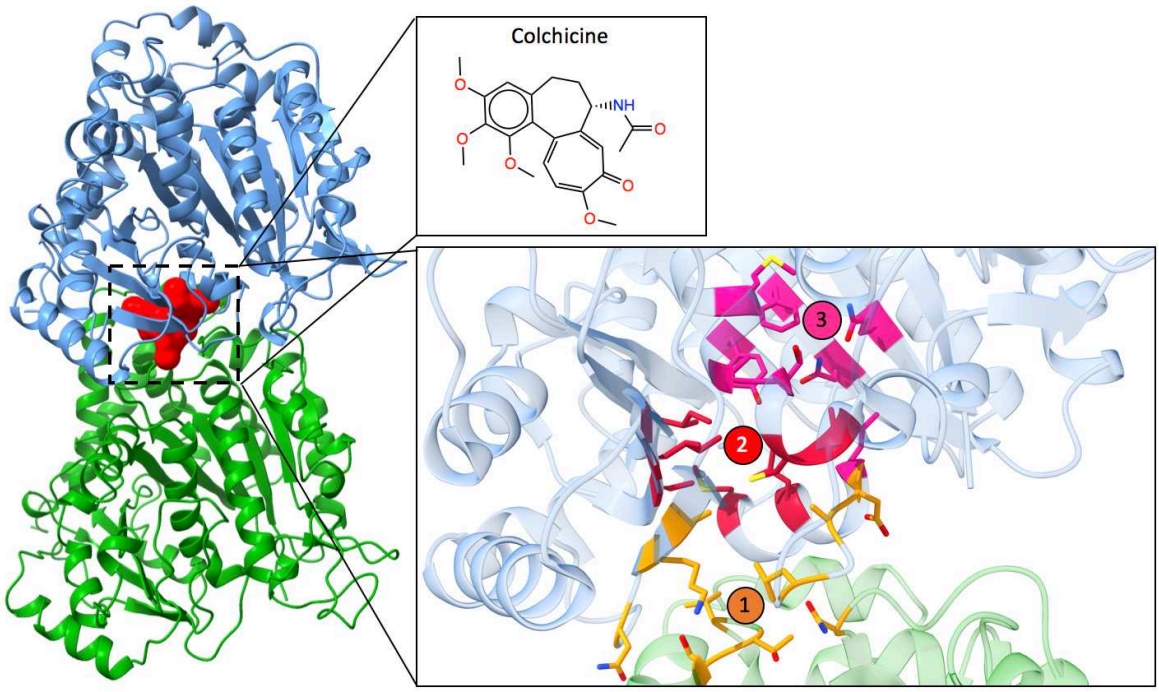


Figure 105. The colchicine binding site and its three zones

Ligand binding at the colchicine site inhibits microtubule formation by obstructing the "curved-to-straight" conformational shift in tubulin. This transition, key to microtubule assembly, involves movements of both α- and β-tubulin subunits' intermediate domains, bringing strands S8 and S9 closer to helix H8[5,100]. The presence of a colchicine-site binding ligand provokes a structural switch in the T7 loop, leading to a contraction of the colchicine site and thus preventing the required conformational change, consequently inhibiting microtubule formation[5,100]. Binding

of a ligand to the colchicine site predominantly occurs through hydrophobic interactions, supplemented by a minimal number of polar contacts [5,125].

Targeting the colchicine binding site has several benefits, including its distinctive structural and functional features compared to other binding sites of the tubulin protein, potentially leading to unique mechanisms of action of the ligands that target it, and its effectiveness against a wide array of cancer types, including drug-resistant variants[39,125].

However, colchicine site-targeted drugs also present challenges, such as their low selectivity for cancer cells which could induce off-target effects and toxicity, poor pharmacokinetic traits like low solubility, brief half-life, and suboptimal bioavailability. These factors collectively limit their clinical application. Additionally, resistance may develop over time due to tubulin mutations or efflux transporter upregulation, and the absence of compounds binding to all three zones of the colchicine site could limit the efficacy of these drugs[126].

Many inhibitors targeting the colchicine site have been developed from representative and commonly used scaffolds, limiting structural innovation and constraining exploration of the chemical space[127]. *De novo* drug design, particularly using inverse quantitative structure-activity relationship (*i*-QSAR) modeling, could help circumvent these issues.

Quantitative structure-activity relations (QSAR) are either regression or classification models capable of estimating a compound's property value given its molecular structure. This relationship can be expressed as activity = f (structure), where the function f requires tuning of internal parameters to produce accurate property value approximations for a given structure[48] (Figure 106). Typically, the "structure" argument in f (structure) is a molecular graph where vertices represent chemical elements and edges represent bond types. The molecular graph's information content is first translated into a numerical representation – a vector of N real numbers, also known as the molecular descriptor vector D. Then, in classical QSAR, a machine learning model is used to determine a relationship between a set of molecular descriptor vectors and the measured property values[48]. If such relationship between molecular structure and property holds, inverse mapping could be employed to retrieve the optimal chemical structure that would correspond to a specific property value[66]. Thus, the inverse QSAR problem can be formalized as two steps: firstly, identification of a "seed" descriptor vector D_{seed} that corresponds to the desired property value; secondly, finding valid molecular structures that correspond to D_{seed}[66]. The second task may be facilitated by an autoencoder neural network model that is pre-trained to map descriptor vectors to valid molecular structures[66].

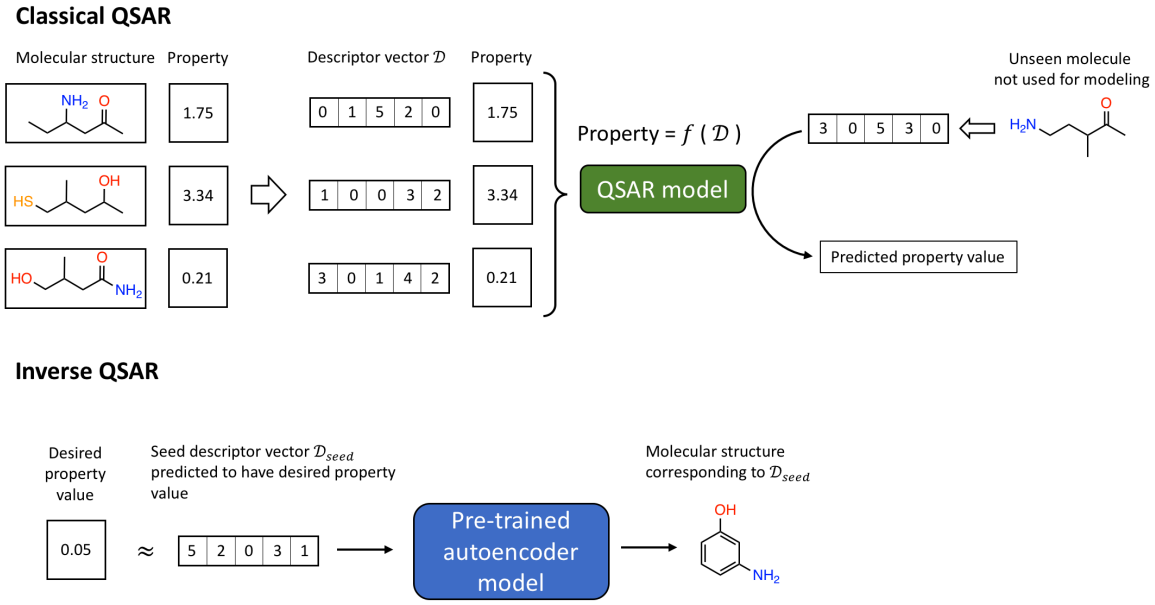


Figure 106. Overview of the difference between classical QSAR and inverse QSAR approaches

This methodology requires structure-activity data to build the initial QSAR model that will be used to select the seed descriptor vector to generate molecules from. The colchicine site, due to being extensively studied, provides enough small molecule structure-activity data for QSAR model training. Hence, the goal of this project was to design novel binders for the colchicine site using the inverse QSAR approach, thereby demonstrating the efficacy of this method.

5.2. Building a QSAR model for colchicine site binding propensity

This project aimed to generate novel molecules that inhibit tubulin polymerization by binding at the colchicine binding site. To identify seed descriptor vectors associated with structurally novel compounds that bind to the colchicine site, we opted first to train a QSAR model to map the relationship between the two-dimensional structure of colchicine site-targeting compounds and a property indicative of their binding efficiency. Our goal was to later use this model to perform virtual screening to find existing and synthetically accessible molecules with high predicted value of a property related to the affinity to the colchicine binding site. This way, we ensure that the chosen seed points actually correspond to chemical structures with pharmacology-compliant physicochemical properties. Thus, we initiated a survey of available structure-activity data for colchicine site-targeting compounds to understand what structure-activity data are available and what this property may be.

5.2.1. Survey of available structure-activity data

Since the discovery of anti-tubulin action of colchicine leading to significant anti-proliferative action on cells in the 1950s, the colchicine binding site has been well-studied, amassing substantial structure-activity data[128].

In a recent study, López-López et al. compiled a dataset of 851 unique compounds designed as tubulin inhibitors, including ones designed to target the colchicine binding site, with reported half maximal inhibitory concentration (IC50) values against different cancer cell lines[129]. The published dataset also included the SMILES strings, pIC50 values, and an activity label for all compounds. Thus, we decided that this experimentally measured value could serve as the target property that we would like to optimize in the potential *de novo* designed colchicine site-targeting compounds. To this end, we specifically selected a subset of 379 molecules that targeted the colchicine binding site and had an experimentally measured IC50 value against HeLa cells. We chose to focus on the measurements against HeLa cells because this subset had the highest number of related records in the data published by López-López et al. These IC50 values, coming from several individual bioassays, can be combined into a single set due to the standardized protocol followed during these tests and the consistent use of colchicine as the reference compound throughout.

5.2.2. Data preparation for QSAR modeling

From our chosen subset, 229 molecules were classified as active and 150 as inactive, with the "active" label denoting molecules with sub-micromolar range dose-response activity against HeLa cells, and "inactive" otherwise.

All compounds underwent standardization following ChemAxon tools' default protocol, including removal of large molecules, counter-ions, conversion to major microspecies of the most probable tautomeric form, and removal of stereochemical information (as the calculated molecular descriptors are stereochemistry-independent). We verified there were no conflicts in data annotations, such as differing activity labels or variances in their dose-dependent pIC50 values.

The selected compounds were initially represented by 95 sets of descriptors based on diverse ISIDA fragmentation schemes serving as the choice for descriptor representation[50]. These schemes included sequences, circular fragments, triplet counts, atom pairs color-coded by atom symbols, as well as pharmacophore features and force field types.

5.2.3. Model building and validation pipeline

We framed the problem as regression modeling, because our goal was to identify "seed" descriptor vectors correlating with high predicted activity values. We aimed to train a model that learns the relationship between the two-dimensional structure and pIC50 value, using the subset of 379 colchicine site-targeting molecules, to make predictions on a new set of compounds. The top-scoring compounds would be then selected as seed molecules for further generation.

Our modelling pipeline consisted of an evolutionary model-building procedure using the Random Forest Regressor estimator to optimize descriptor sets among the 95 proposed ISIDA fragmentation schemes. Models' hyperparameters were optimized and ranked based on a fitness score reflecting the mean coefficient of determination R^2 over a 12-times repeated 3-fold cross-validation scheme, implemented using the scikit-learn[130] and sklearn-genetic-opt python packages.

5.2.4. QSAR modeling results

The results of model building indicated that the optimal ISIDA fragmentation scheme was to count atom pair numbers at given topological distance, where atoms were rendered by their consistent-valence force field molecular mechanics force field types and topological distances (number of separating bonds) ranged from 1 to 5 (ISIDA notation: IA-FF-P-2-6). The top-performing random forest model built on this descriptor set achieved an R^2 value of 0.63 following the rigorous cross-validation procedure (Figure 107). Consequently, the model could be utilized to predict pIC50 values against HeLa cells for any compound within the model's applicability domain.

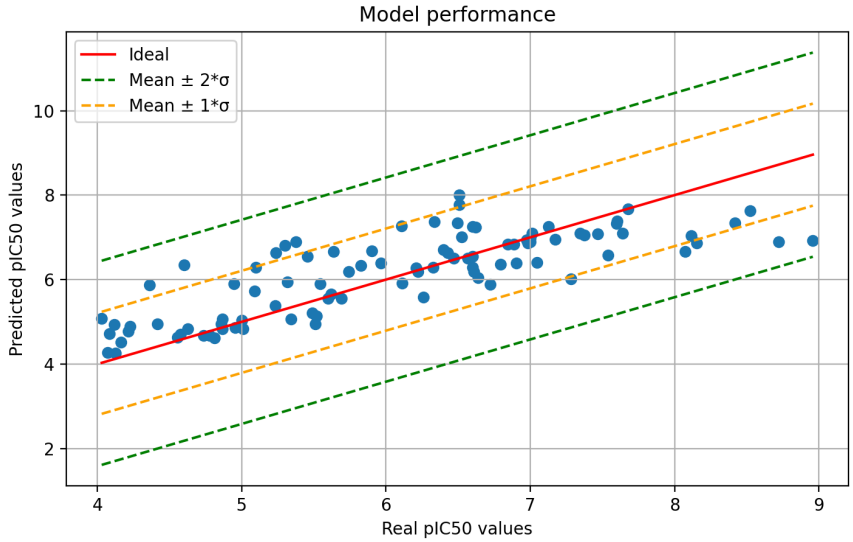


Figure 107. Model performance plotted as a real vs. predicted plot

5.3. Choosing seed descriptor vectors to generate from

Next, we needed to identify seed molecular descriptor vectors closely associated with colchicine site-targeting compounds and anticipated to possess high pIC50 values against HeLa cells, possibly indicative of potent binding affinity to the colchicine site. At this point, we could enumerate all possible artificial descriptor vectors and utilize our trained model until we identified some artificial descriptor vectors yielding high predicted pIC50 values. However, the issue was that the vector $\mathcal{D}$, which maximizes the predictive QSAR function pIC50 = $f(\mathcal{D})$, could be mathematically feasible, but may not correspond to any physical molecular structure. To counter this issue, we decided to perform a screening of a library of purchasable compounds. Molecules from such a library would not only possess pharmacologically compliant physicochemical properties but would be also commercially available for testing. The implementation of such an approach facilitated the exploration of intriguing regions of the chemical space while ensuring that the seed points that we select are not meaningless $\mathcal{D}$ vectors.

Thus, we decided to apply the trained QSAR model to an alternative set of compounds, selecting those with the highest predicted activity and generating molecules around these compounds. The colchicine site-targeting structural properties of the compound set would be ensured by initial filtering using the trained QSAR model's applicability domain.

5.3.1. Choosing data to filter by our predictive model

Thus, we needed to choose a library of compounds for filtering by the QSAR model. Our selection criteria included the purchase availability of the molecules, to ensure the generated compounds or their close analogs could be relatively easily synthesized or directly purchased. Additionally, the set should have been small yet diverse, ensuring a variety of scaffolds among top-ranking molecules predicted by pIC50 value.

Considering these aspects, we selected the Enamine phenotypic screening library comprising 5760 compounds. These compounds are structurally diverse and are known to have biological action on different targets *in vitro*. All compounds from the library are cell-permeable and possess pharmacology-compliant physicochemical properties.

5.3.2. Pre-processing the data

The data underwent standardization using the in-house ChemAxon-based procedure, including removal of large molecules, counter-ions, conversion to major microspecies of the most probable tautomeric form, and removal of stereochemical information for stereochemistry-independent descriptor calculation.

For these compounds, we calculated the IA-FF-P-2-6 ISIDA fragment descriptors, matching the ones used to train the QSAR model in section 5.2. The compounds from the chosen library needed to fall within the applicability domain of our trained QSAR model. We estimated it using the bounding box method, determining the minimum and maximum descriptor values for each dimension of the descriptor vector matrix used to train the QSAR model. The compounds from the Enamine phenotypic library that fell outside the established min-max range were removed, reducing the dataset from 5760 to 421 compliant compounds. Inherently, these compounds contained structural fragments characteristic of colchicine site-targeting compounds, suitable for prediction using our trained model.

5.3.3. Making predictions

Predictions were performed by inputting the 421 compliant compounds into the trained QSAR model, yielding predicted pIC50 values against HeLa cells. After ranking the compounds by predicted value in decreasing order, the IA-FF-P-2-6 descriptor vectors of the top 15 molecules by predicted pIC50 score were chosen as seed vectors for further compound generation (Figure 108). Some selected compounds included structural fragments that are present in the known binders (e.g., 1,3-diphenylprop-2-en-1-one or 1,2,3-trimethoxybenzene), while others have fragments not previously observed in crystallographically confirmed colchicine site binders (e.g., 2-(phenylamino)pyridine-3-carbaldehyde and phenylbenzamide).

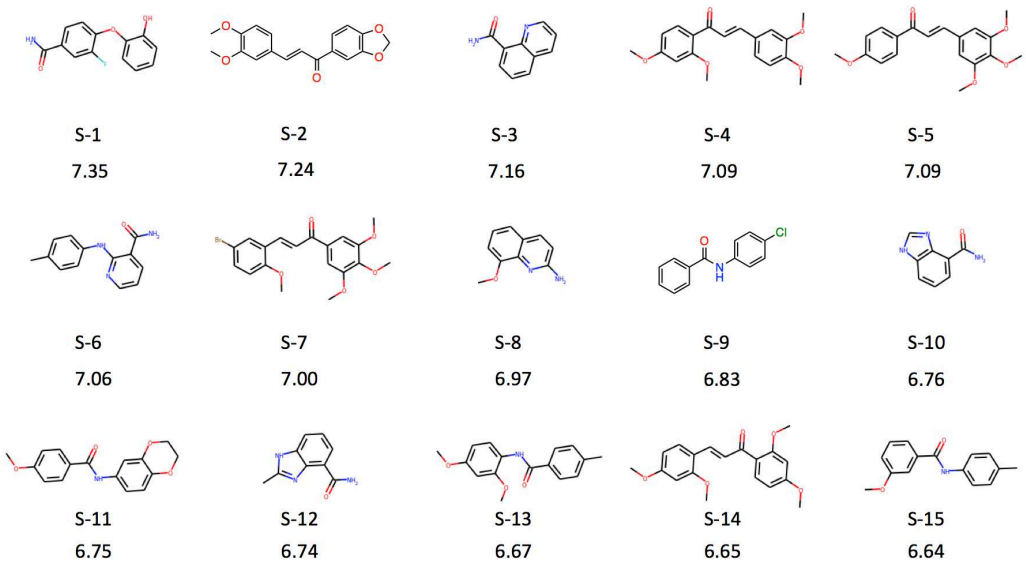


Figure 108. Fifteen seed molecules selected by predictions of a QSAR model. Numbers under molecular structures are predicted pIC50 values.

5.4. Training a variational autoencoder

Our approach considered the seed molecular descriptor vectors as points within a chemical space, where molecules can be sampled around this point. The sampled molecules are anticipated to possess the predicted property. The key challenge was to have a way of sampling chemically valid molecules from this chemical space around the seed points of interest. We addressed it by training an autoencoder neural network model.

5.4.1. What is a variational autoencoder

Autoencoders are specialized neural networks designed to reproduce their input as output. They achieve this by compressing the input into a lower-dimensional representation, or latent-space representation, and then reconstructing the output from this compressed form[131]. An autoencoder comprises three components: encoder, latent representation, and decoder. The encoder compresses the input to create the latent representation, and the decoder reconstructs the input using this latent representation. The learned latent representations form a latent space. In brief, autoencoders learn a function to map each input to a latent representation, and decoder learns the reverse mapping (Figure 109).

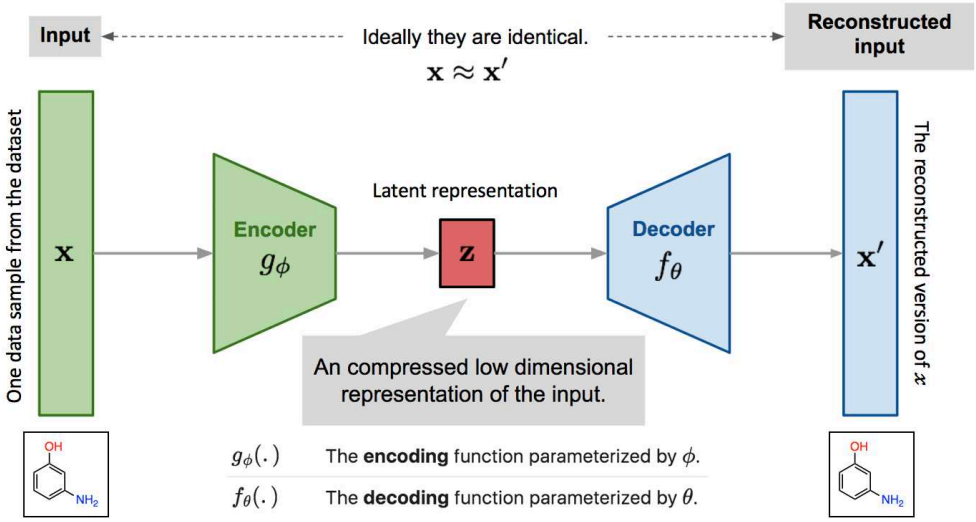


Figure 109. Schematic overview of an autoencoder neural network model (adapted from [132]).

Autoencoders are typically trained using raw input data, and are regarded as unsupervised learning techniques since they do not require explicit labels. More precisely, they can be considered self-supervised as they generate their own labels from the training data[131].

Standard autoencoders may not ensure continuity in the latent space formed by latent representations, which can complicate interpolation. Variational autoencoders (VAEs) address this

issue by modeling their latent representations as a probability distribution, thereby creating a continuous latent space that is easily sampled and interpolated[133] (Figure 110). VAEs are generative in nature, meaning these models are able to generate new instances that are similar to the original training dataset. Rather than mapping the input to a static vector, as simple autoencoder models, they map it to a specific distribution. However, while powerful, VAEs do not offer explicit control over the characteristics of the generated data[133].

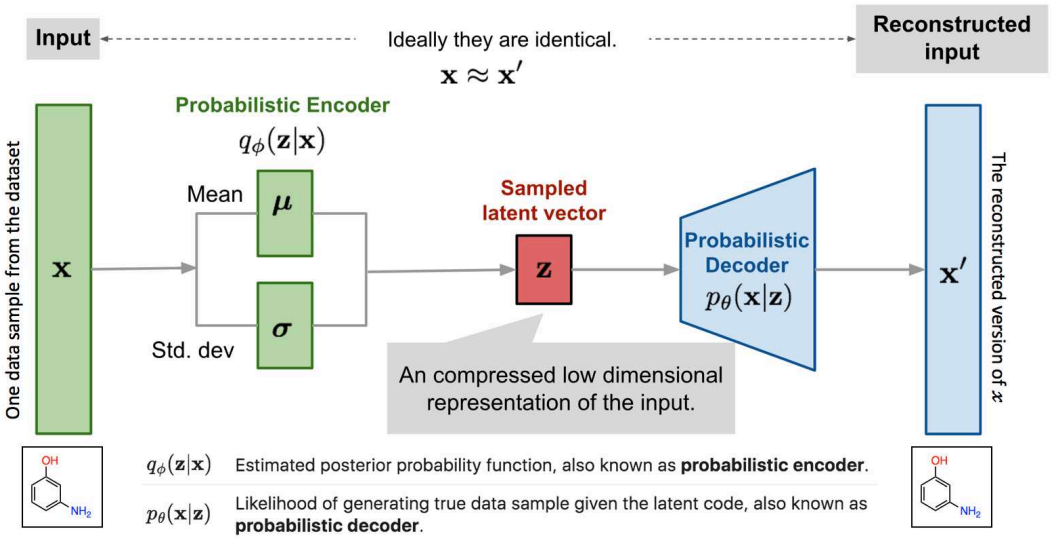


Figure 110. Schematic overview of a variational autoencoder (adapter from [132]).

Conditional variational autoencoders (CoVAEs) differ from VAEs by allowing for more controlled and versatile data generation[134] (Figure 111). By introducing conditional variables into the architecture of the encoder and decoder components, CVAEs can be instructed to generate data samples that not only resemble the original dataset but also satisfy a given set of conditions. Essentially, CVAEs learn the distribution of the input data, conditioned on specific attributes[134]. This conditional aspect can be any relevant feature or parameter of the data, which the model then leverages to guide the generation process.

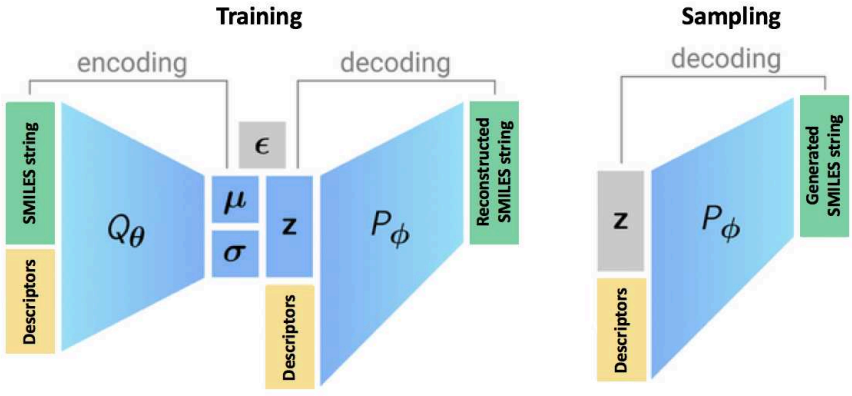


Figure 111. General scheme of a conditional variational autoencoder (adapted from [135])

In our work, we aimed to train a conditional autoencoder to generate syntactically valid SMILES strings based on the input molecular descriptors. Specifically, we used ACoVAE, an attention-based conditional variational autoencoder architecture proposed by Bort et al[66]. This model is trained on a set of SMILES strings and corresponding descriptor vectors. The training process involves a gated recurrent unit-based encoder parametrizing a random latent vector distribution, forming a (0, 1) hyperspherical distribution as the target latent vector distribution. During inference, the latent vector is sampled from the prior distribution, and a desired descriptor vector is used as a condition to generate the intended SMILES from the random and condition vector.

5.4.2. Preparing data for autoencoder training

To train the ACoVAE autoencoder, we required a large and diverse chemical library. For this purpose, we utilized the ChEMBL database (v. 26), which contains 1,721,154 molecules. The molecules were standardized through our typical in-house procedure involving the removal of large molecules, counter-ions, conversion to the major microspecies of the most probable tautomeric form, and removal of stereochemical information. We then computed the IA-FF-P-2-6 ISIDA fragment descriptors for all the molecules, resulting in a descriptor vector with 2901 fragment features for each molecule.

Next in our preprocessing pipeline, feature selection was conducted to reduce dimensionality of the resulting vector to constrain the amount of GPU RAM required to process the data. This was achieved by pruning features based on their standard deviation. We computed the standard deviation for each feature in the dataset and pruned those with a standard deviation of 0, meaning features with constant value throughout the dataset that provides no discriminatory information, as it doesn't vary across observations. This ensured that only features with substantial variability were retained, as they are more likely to contribute meaningful information to the machine learning model. Following this, the features were sorted by their standard deviations in descending order, and only the top 1207 features were selected for the subsequent autoencoder training.

5.4.3. Training a variational autoencoder

The ACoVAE implementation comprises of an encoder and a decoder. The encoder processes the input data and computes a latent representation, while the decoder generates data conditioned on the latent variables. It was specifically designed to handle SMILES strings and molecular descriptor vectors.

The construction of the model begins with an embedding layer that transforms the input SMILES strings into a higher-dimensional space by the process of SMILES tokenization. The encoder then processes the embedded input using an internal transformer model and returns two outputs: a mean and a log variance. A sample from the latent space is then drawn based on these outputs. This sample is generated using the reparameterization trick[136] to allow gradients to pass through the sample to the encoder. This sampled latent vector is passed to the decoder along with the input descriptor vector. The output of the model is the probability distribution of SMILES strings.

The model is trained with two components of the loss function: the reconstruction loss, computed as the sparse categorical cross-entropy between the input and the output, and the Kullback-Leibler divergence between the learned latent distribution and the prior distribution, which acts as a regularization term. The Kullback-Leibler divergence loss is scaled by a factor of 20 to control its influence relative to the reconstruction loss[137].

In this work, the AdaBelief optimizer[138] was used to optimize the model parameters. This optimizer has been shown to converge faster and generalize better than traditional optimizers such as Adam[138].

The model was trained for 200 epochs with a batch size of 512. The input SMILES strings were limited to a maximum length of 100 characters, and the latent space was a 64-dimensional hypersphere. The internal dimension of the transformer model was set to 256, with 4 layers and 8 heads in the multi-head attention mechanism.

During training, the model weights that achieved the best validation accuracy for any given epoch of training were saved for later use during inference, the exact one selected by the user. Moreover, the learning rate schedule was monitored using a custom callback, which allowed the model to adjust quickly early in the training process when the weights are randomly initialized, then slowly fine-tune as the training process progresses.

5.5. Generation of molecules around selected seed vectors

5.5.1. Description of the generation process

The ACoVAE transformer enables the sampling process by taking a descriptor vector as input to the trained decoder part of the model. Each descriptor vector acts as the "condition" part of the ACoVAE and is paired with a batch of random vectors drawn from a power spherical distribution. This distribution forms the latent space. Each pair of descriptor and random latent vector produces a generated SMILES sample. Through categorical sampling, which allows the exploration of different possibilities for the same input, a given descriptor vector can generate

multiple different SMILES. The ACoVAE software also incorporates a check for the validity of generated text strings, eliminating any incoherent or incorrect SMILES.

5.5.2. Generation results

Upon inputting the seed descriptor vectors into the trained ACoVAE model, we obtained 6623 generated molecules. Noting the presence of duplicates, we employed a standardization routine to eliminate these redundant molecules, resulting in a set of 782 unique molecules. A random sample of 10 generated compounds is shown in Figure 112. These molecules were derived from the 15 seed molecules, selected by the QSAR model due to their high predicted pIC50 values against HeLa cells.

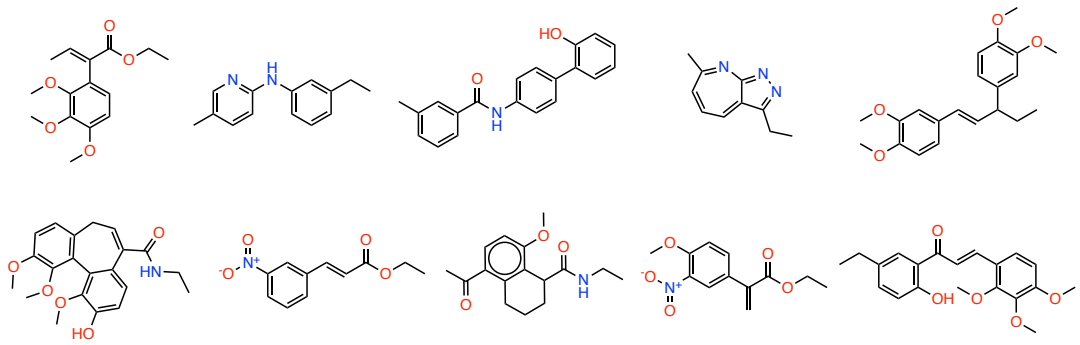


Figure 112. Examples of de novo generated molecules

5.5.3. Computational and experimental validation

For validation of the inverse QSAR approach, the 782 generated compounds were subjected to another round of QSAR model prediction. Our objective was to verify if the generated compounds retained high activity levels as predicted by the model, similar to the seed molecules used to generate them. We began this process by checking how many of the generated molecules were inside the model's applicability domain, which were determined using the bounding box method. This preliminary step revealed that only 163 out of the 782 generated molecules fell within the trained QSAR model's applicability domain.

We then proceeded to make activity predictions for these 163 molecules using the trained model. We present these results in Figure 113, which shows the distribution of predicted pIC50 values for these molecules. The pIC50 values offer a measure of the potency of the compounds, and higher values correspond to higher activity levels.

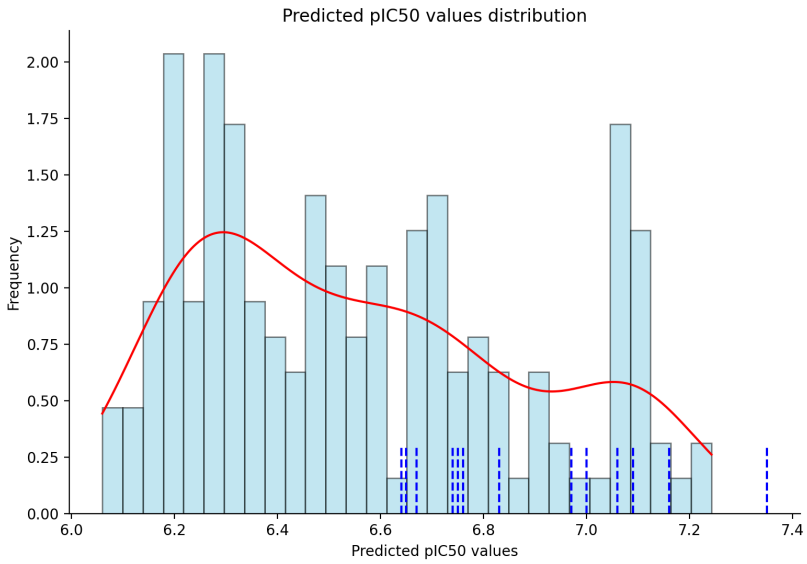


Figure 113. Distribution of the predicted pIC50 values against HeLa cells for 163 de novo generated compounds that are within the model's applicability domain

In our case, the predicted pIC50 values were found to be above 6, which meets the activity threshold set in this study. To provide a visual reference, we marked the predicted pIC50 values for the seed compounds on the histogram using blue dashed lines. This gives a sense of how the activity of the newly generated compounds compares to that of the original seeds. In addition, we included a kernel density estimate, represented by the red line. It serves to illustrate the probability density of the predicted pIC50 values, helping to indicate the most likely activity levels for compounds generated by our inverse QSAR approach. We see that all generated compounds were predicted active, but in general their predicted pIC50 values are lower than those predicted for the seed molecules.

Furthermore, docking simulations were conducted for these molecules. AutoDock GPU software was used for all the docking experiments in this section. Ligands were first represented using the standardized SMILES strings and then converted into SDF file format using ChemAxon's MolConvert software. This in turn allowed us to save them as PDBQT, which is the required input file format for AutoDock, by using Meeko, a python package that preserves connectivity information that is usually missing from PDBQT files by default. The resulting files included atomic coordinates, partial charges, and AutoDock atom types.

The protein structure was taken from the 4O2B PDB structure, retaining only the C and D chains (corresponding to α,β-subunits) while removing solvent and ions. We identified the binding site as all residues within 12 Angstroms from colchicine's center of mass to include all three zones of the binding site. Because the colchicine site is close to the nucleotide binding site on a-tubulin, we did not delete the GTP molecule bound to the a-tubulin. For the docking process, we selected the Lamarckian Genetic Algorithm as the search method due to its effectiveness in exploring the

ligand conformational space. Grid box parameters were defined to match the earlier described binding site. The size of the grid box was 32, determined based on the extent of the defined binding site, with a spacing of 0.375 Angstroms between the grid points. The grid box was centered on the colchicine binding site to ensure proper coverage. We performed 200 independent runs for each ligand to adequately sample the conformational space, with a maximum number of energy evaluations set to 2.5 million and an initial population size of 300.

Firstly, we checked if our software of choice could reproduce the binding mode of the native ligand, colchicine, to estimate the applicability of this software to modeling this system. Indeed, we were able to re-dock colchicine in the binding site with an RMSD value of 1.10 Å, indicating that our software can correctly model the native ligand's binding pose (Figure 114).

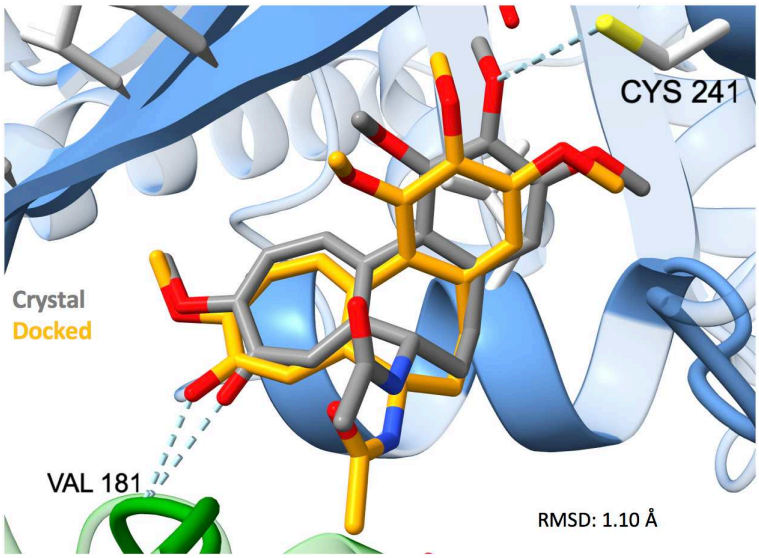


Figure 114. Re-docked pose of colchicine in the binding site (orange) vs crystallographically determined one (gray, 4O2B)

Then, we docked the 163 molecules, along with colchicine, into the colchicine binding site. Post docking, the molecules with docking scores superior to colchicine were ranked, and the top-20 were selected for further exploration (Figure 115).

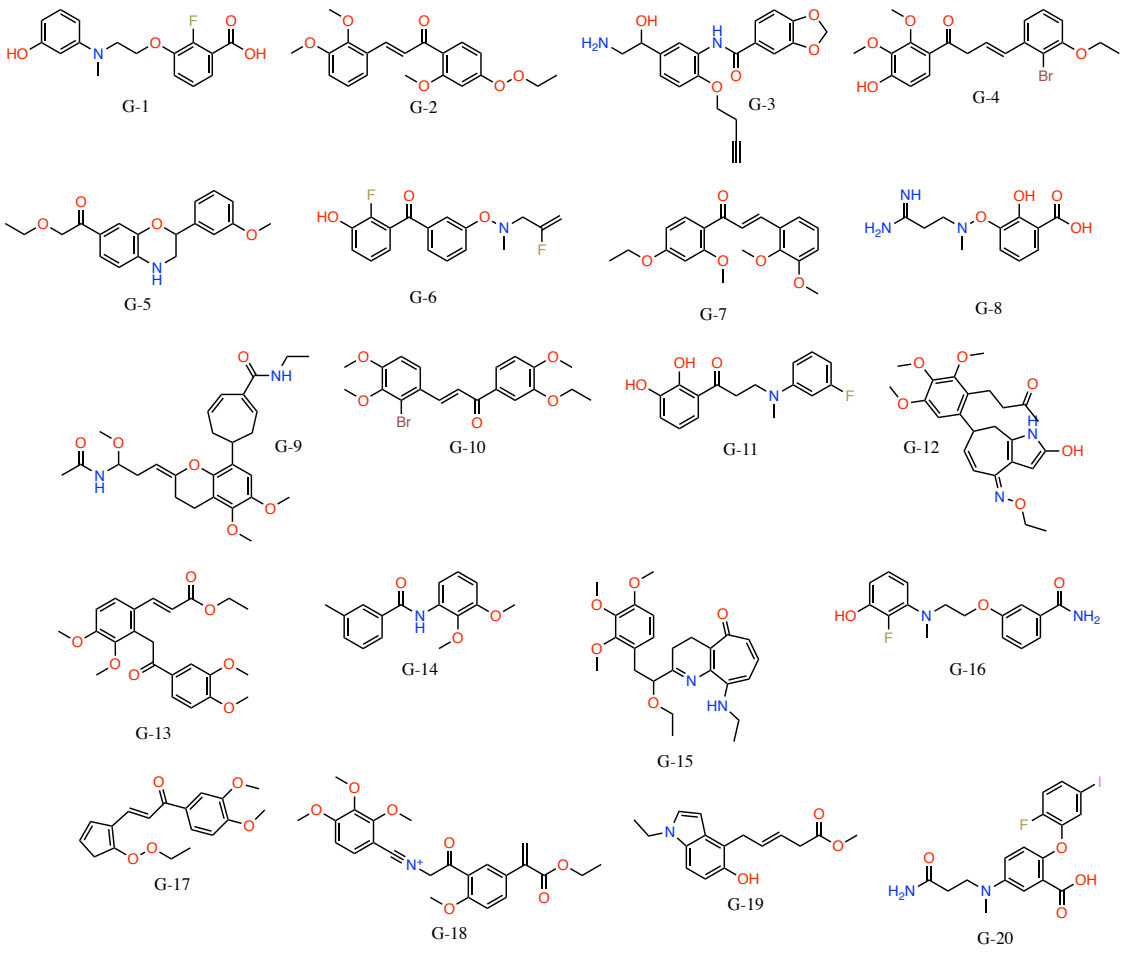


Figure 115. Top-20 de novo generated molecules by docking score

While none of the generated compounds were directly available for purchase from chemical vendors, close analogs for four of the *de novo* generated compounds (**G-3**, **G-7**, **G-10,** **G-14**) were found to be available in the Enamine store (Figure 116). The search was performed manually using the Enamine store web search engine. The similarity criterion was Tanimoto score higher than a threshold value of 0.7. Found analogs were inside the model's applicability domain, and so were subjected to another round of QSAR prediction and docking. Both QSAR prediction and docking results indicated high pIC50 values and a good fit within the colchicine binding site for all four compounds (Figure 117). Additionally, protein-ligand docking calculations have shown that the purchasable analogues of generated compounds still remain largely in the same pose within the binding site, indicating their good fit. Both generated compounds **G-3**, **G-7**, **G-10** and their purchasable analogues have structural fragments common to known colchicine site binders. Interestingly, generated compound **G-14** and its purchasable analogue has a structure that is not present in crystallographically confirmed colchicine site binders.

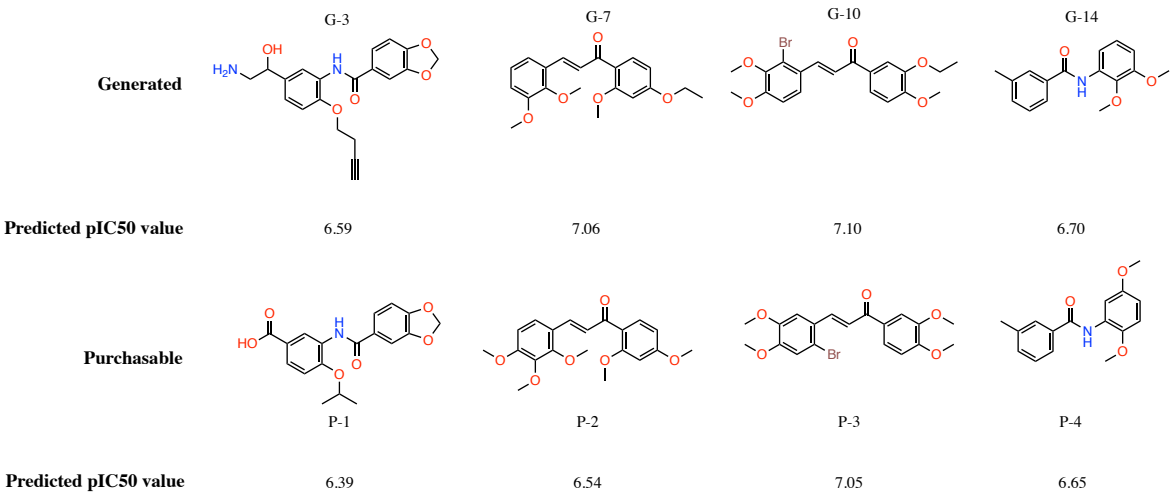


Figure 116. Comparison of the molecular structure and predicted pIC50 values against HeLa cells of de novo generated molecules and their purchasable analogues

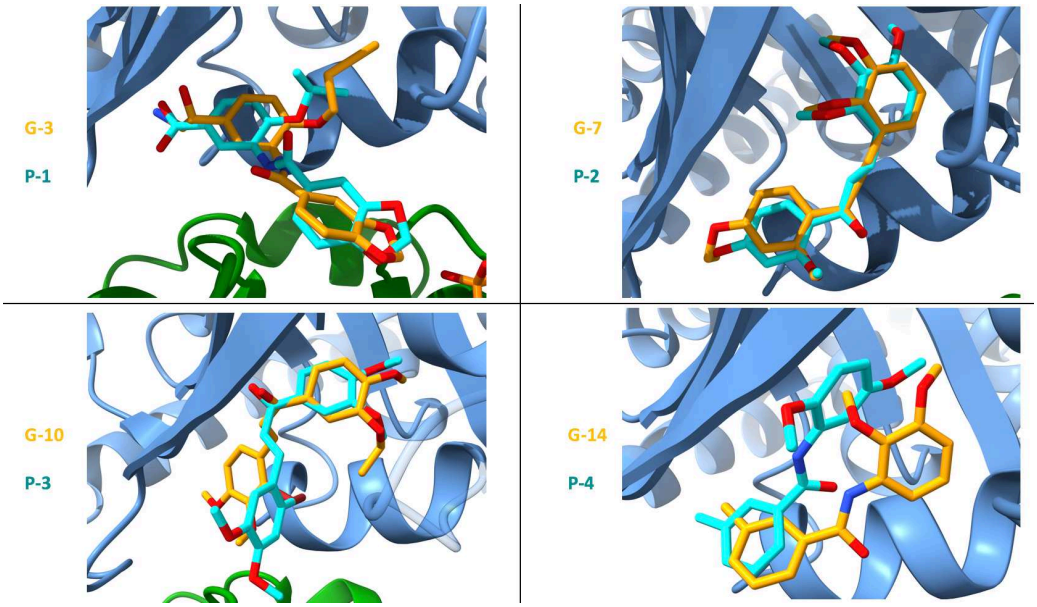


Figure 117. Comparison of the best scoring docked poses of generated compounds and their purchasable analogues

The four molecules **P-1**, **P-2**, **P-3**, and **P-4** were then tested experimentally using X-ray crystallography. However, the binding of these compounds at the colchicine site was not evident in the results. The effect of these molecules on tubulin polymerization *in vitro* remains to be determined and is currently under investigation.

Such lack of detection during X-ray crystallography experiments might be attributed to several factors inherent in the technique and the nature of the ligand-protein interaction. As such, ligand occupancy is a crucial aspect in determining the visibility of a ligand in an X-ray crystallography experiment. If the ligand binding is weak or the binding event is transient, the ligand may not be present in the binding site long enough or at a sufficient occupancy level for detection. Moreover, the crystallographic experiment is performed under conditions that may not

exactly replicate the physiological environment in which the ligand-protein interaction occurs. Variables such as temperature, pH, and crystal packing forces can impact the binding event and may lead to discrepancies between the experiment and the actual biological system.

Thus, further experimental studies of the potential inhibitory action of these molecules on tubulin polymerization are ongoing to elucidate this.

5.6. Conclusion and future perspectives

In this study, an inverse QSAR modeling method was applied to explore the chemical space of potential therapeutic agents targeting the colchicine binding site in tubulin. This method has two key aspects: identifying a "seed" descriptor vector that embodies the desired properties of a potential drug, and finding chemically valid molecular structures that align with the selected vector. This strategy may overcome some limitations in drug design, such as structural redundancy and narrow exploration of chemical space.

We used a set of 379 compounds known to target the colchicine binding site and exhibit activity against HeLa cells to train and validate a Random Forest Regressor model. This model, which demonstrated satisfactory predictive performance of $R^2 = 0.63$, was then applied to filter a diverse library of compounds from Enamine for high predicted pIC50 values against HeLa cells. Top-15 filtered compounds were used as seed vectors.

The resulting seed descriptor vectors were used as input for an autoencoder model trained using the ACoVAE transformer method on a large, diverse set of molecules from the ChEMBL database. This approach allowed us to generate 6623 molecules, 782 of which were unique, indicating the robustness and potential for discovery using this technique.

These *de novo* generated molecules were subsequently predicted to have high pIC50 values using the trained QSAR model, implying their potential for high activity against HeLa cells. To further validate their potential, the molecules were subjected to docking simulations, revealing that many had docking scores superior to colchicine, the native ligand. The top-20 generated molecules by docking score were selected for further investigation. Although the exact compounds generated were not commercially available, close analogs of four were found and were also predicted to have high pIC50 values and to fit well within the colchicine binding site.

While experimental validation of the generated compounds by X-ray crystallography has presented some challenges, they are being addressed, and further *in vitro* studies are ongoing. The comprehensive approach used in this study shows great promise for future research in drug discovery. This work exemplifies the potential of combining inverse QSAR modelling and experimental approaches to accelerate the discovery of novel therapeutic compounds.

Chapter 6. Application of transfer learning for QSAR modeling

6.1. Introduction

Quantitative structure-activity relationship modelling, established nearly half a century ago, has been instrumental in drug design and optimization. For a given set of molecules, the goal of QSAR modelling is to correlate specific target property values $\mathcal{P}$, typically measured experimentally, with structural features of the molecules[139,140]. The modeling process involves four key steps: (1) the computation of numerical representations of molecular structure (molecular descriptors $\mathcal{D}$); (2) descriptor selection, which involves identifying the subset of descriptors most relevant to modelling the desired property; (3) discovering an optimal, often nonlinear, relationship F between the descriptors and target property variable $\mathcal{P} = F(\mathcal{D})$; and (4) validating the model by assessing its predictive power, robustness, and applicability domain[140].

The success of QSAR modeling relies heavily on accurately numerically describing molecules in a manner that captures their relevant structural properties[140]. Various types of chemical descriptors, embodying different degrees of chemical structure representation, have been proposed, ranging from molecular formula (1D), two-dimensional structural formula (2D), three-dimensional, conformation-dependent (3D), to even higher levels considering mutual orientation and time-dependent molecular dynamics (4D and beyond)[48].

With many empirically appealing choices for the molecular descriptor set (DS) to be used as input $\mathcal{D}$, it's challenging to predict beforehand which descriptors will facilitate the most robust learning of $\mathcal{P} = F(\mathcal{D})$. This can be seen as a feature selection issue, where vectors of all potential DS candidates can be concatenated and filtered using a feature mask. With a large number of potential DS to be considered, each comprising large number of dimensions, selecting from the many possible descriptors represents a substantial computational challenge[110].

Various methods for descriptor set selection have been developed, including brute force enumeration, forward addition/backward elimination statistical techniques, Bayesian approaches such as automatic relevance determination, genetic algorithms, clustering methods, and self-organized maps. Sometimes the choice of a suitable descriptor set may be also guided by professional expertise and expert domain knowledge. However, none of these methods are optimal as they are computationally demanding or based on the assumption that the best descriptor set for a task is included in the optimized descriptor space, which might not be the case[110,139,140].

On the flip side, given sufficient data, deep learning methods can learn useful molecular representations from large molecular data sets that can significantly enhance predictive modelling performance. These methods, which do not require the meticulous design or selection of descriptors, have shown significant performance improvement over conventional methods. The

molecular representations learned by deep learning methods are task-specific and encode relevant information on par or even more effectively than traditional fingerprints or descriptors. However, these models require substantial amount of labelled data for training, and their performance decreases significantly when the training data is limited[141].

In this context, transfer learning emerges as a promising alternative. Transfer learning is an emerging concept within the field of deep learning, which has demonstrated promising results across various applications including computer vision and natural language processing. The idea behind transfer learning is to use the knowledge gained from learning one task, often with abundant available data, to improve the model's performance on another, related task, usually with less available data[142] (Figure 118). The typical routine for transfer learning at a minimum involves two core stages: pre-training and fine-tuning.

Within the context of chemoinformatics, pre-training involves learning representations of molecular structures from a substantial corpus of unlabelled molecular data in a self-supervised fashion. By exposing the model to a diverse set of molecular structures, it develops a broad understanding of various molecular features and relationships between them. The pre-training step is always followed by fine-tuning. This process involves adjusting the learned representations using a smaller structure-activity dataset specific to the task at hand. This fine-tuning process helps tailor the learned representations to model the specific task efficiently, thus optimizing the performance of the predictive model. This dual-stage approach in transfer learning not only enhances the model's predictive accuracy but also addresses data scarcity issues, often encountered in specialized tasks within the field of chemoinformatics[142,143].

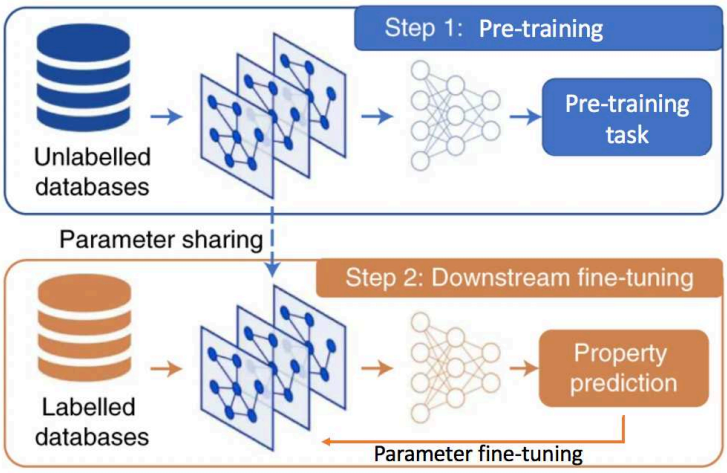


Figure 118. High level schematic overview of the transfer learning concept

Transfer learning has gained traction in chemoinformatics, proving particularly useful in *de novo* drug design. Several works have been published that have a model initially trained to understand SMILES grammar rules using a large pool of unlabelled data, and then generating valid SMILES strings under specific conditions. The potential of complex neural networks to learn

valuable features from unlabelled molecular data has also been demonstrated, indicating promising avenues for further research in this domain.[142,143]

The primary objective of this project was to investigate the effectiveness of transfer learning in QSAR modeling and compare it to a current state-of-the-art method that employs a support vector machine (SVM) in conjunction with a genetic algorithm-driven optimization process, involving descriptor selection and optimization, as the genetic algorithm simultaneously refines the hyperparameters of the SVM and the descriptor set used to train the models[110]. In contrast, transfer learning may allow to bypass the descriptor optimization step entirely. This approach learns useful molecular representations in an unsupervised end-to-end manner and then fine-tunes these representations to a specific downstream task[143].

Therefore, we aimed to assess whether these unsupervisedly learned and fine-tuned molecular representations lead to a predictive performance that at least matches, or even exceeds, that achieved by the genetic algorithm-optimized descriptor sets. By conducting this comparison, we sought to determine the potential of transfer learning as a more efficient and effective strategy for modeling in QSAR modelling.

6.2. Survey of open-source tools for transfer learning on molecular data

We aimed to test an existing methodology for learning representations from unstructured molecular data, rather than developing a new one. Thus, first we had to identify open-source tools that implement transfer learning on molecular data, following the pre-training, fine-tuning, and prediction workflow. Any given tool needed to meet several key criteria to be considered. Firstly, it should learn the general purpose representations in a self-supervised fashion from unlabeled molecular data with chemical intuition behind the design of the representation learning task. Secondly, the tool needed to offer models pre-trained on substantial data, while providing a convenient way to perform pre-training on user-defined datasets. Thirdly, it needed to implement fine-tuning of pre-trained models on any given downstream QSAR task, extraction of learned molecular representations for any given input molecule, and making predictions on new data.

Upon reviewing the literature, we identified several tools fitting these criteria: GeoSSL[144], TorchDrug[145], MolCLR[146], 3DInfoMax[147], Chemformer[148], and GROVER[149].

GeoSSL is a graph-based method for learning molecular representations which incorporates 3D conformation data into its learning process. The learning is framed under a 3D coordinate denoising pretraining framework, computing continuous motion of molecules in 3D Euclidean space to model an energy landscape. It simplifies the denoising task to denoising the pairwise atomic distances in a molecule in different conformations. This allows GeoSSL to capture a more

comprehensive view of the molecules, considering not only their topological structure but also their spatial conformation.

TorchDrug is a machine learning platform for drug discovery, employing self-supervised Graph Neural Network strategies during pre-training. Methods such as InfoGraph and Attribute Masking are utilized, leveraging structural information and node/edge attributes in molecules.

MolCLR (Molecular Contrastive Learning of Representations) is a graph-based method that uses a contrastive learning framework to embed molecules into graph-level representations with the task of distinguishing between similar and dissimilar data instances. The fundamental idea of this method is to have the model produce similar representations for similar (or related) molecules and dissimilar representations for dissimilar (or unrelated) molecules.

3DInfoMax is another graph-based method that incorporates both 2D and 3D molecular data in representation learning, while requiring only 2D data for fine-tuning and making predictions. This is achieved by training a graph neural network on molecular data with available 3D conformation data, inherently learning to generate implicit 3D data in latent representations.

Chemformer is a Transformer-based model which operates on SMILES strings by treating the task as a sequence-to-sequence problem. The process begins with tokenization of the SMILES strings into individual components representing atoms, bonds, or special characters. These tokens are then embedded into high-dimensional vectors, which are fed into the Transformer model. The Transformer, with its self-attention mechanisms, models the dependencies between tokens in the sequence, effectively capturing long-range interactions between atoms that are crucial in determining molecular properties.

Lastly, GROVER, Graph Representation frOm self-superVised mEssage passing tRansformer, employs a novel framework using self-supervised tasks in node-, edge- and graph-level for molecular representation learning. GROVER integrates message-passing networks into the Transformer-style architecture to devise more expressive molecular representations.

The selection of the most suitable tool for this study required careful consideration of the unique attributes and limitations of each open-source tool at our disposal. As such, GeoSSL was excluded due to its reliance on 3D representations of molecules, both for pre-training and downstream task fine-tuning. The computational burden associated with this requirement, combined with the tool's lack of a straightforward mechanism for extracting learned representations, rendered it unsuitable for our purposes.

TorchDrug, despite its potential, was also ruled out. The tool implements two graph-based pre-training methods, which proved highly time-consuming to train. Moreover, it did not provide any pre-trained models, and lacked a convenient interface for working with training/prediction data or straightforwardly extracting learned representations. Similarly, MolCLR, another graph-

based method, was eliminated from consideration due to its lack of an easy method for extracting learned representations or generating predictions for new data.

In the case of 3DInfoMax, despite its promising graph-based methodology, we encountered significant difficulties installing the tool and making it function properly. Chemformer, the sole SMILES-based method among the tools considered, was also deemed unsuitable due to installation difficulties arising from package version inconsistencies.

After extensive review and comparison, we chose to work with the GROVER tool. A graph-based method, GROVER provides convenient command-line implementations for model pre-training, model fine-tuning on any given downstream task, extraction of learned representations, and prediction using the fine-tuned representations and a multilayer perceptron model. The encoder model implemented in GROVER contains only graph convolution networks and operates on undirected graphs, and so, the latent representation is learned without any influence of atoms' order. Thus, GROVER was identified as the most appropriate tool for implementing transfer learning on molecular data in this study.

6.3. Transfer learning workflow with GROVER

The pre-training step with GROVER begins with the input of a list of SMILES strings, from which the tool generates a graph representation for each molecule. This step has two distinct pre-training tasks, both of which have user-definable hyperparameters within GROVER's interface. The first task concerns individual nodes and edges of each graph, which represent atoms and bonds, respectively. In contrast, the second task concerns the entire molecular graph.

For the atom- and bond-level pre-training task, each input graph is subdivided into smaller fragments, which are subsequently counted. Then, each graph is represented as a bit string, whose length corresponds to the total count of these generated fragments. Here, an active bit signifies the presence of a given fragment in the molecular graph, while a disabled bit indicates its absence.

Next, a portion of each graph corresponding to some fragment is masked, ensuring that the masked node and edge labels account for no more than 15% of the graph. Then, atom- and bond-level graph convolutions generate embeddings for the masked graph. The first self-supervised learning task is to use these embeddings of the masked graph to predict which fragment of the molecular graph has been masked (Figure 119). This task encapsulates the objective of learning molecular representations from the inherent structural features of the molecules.

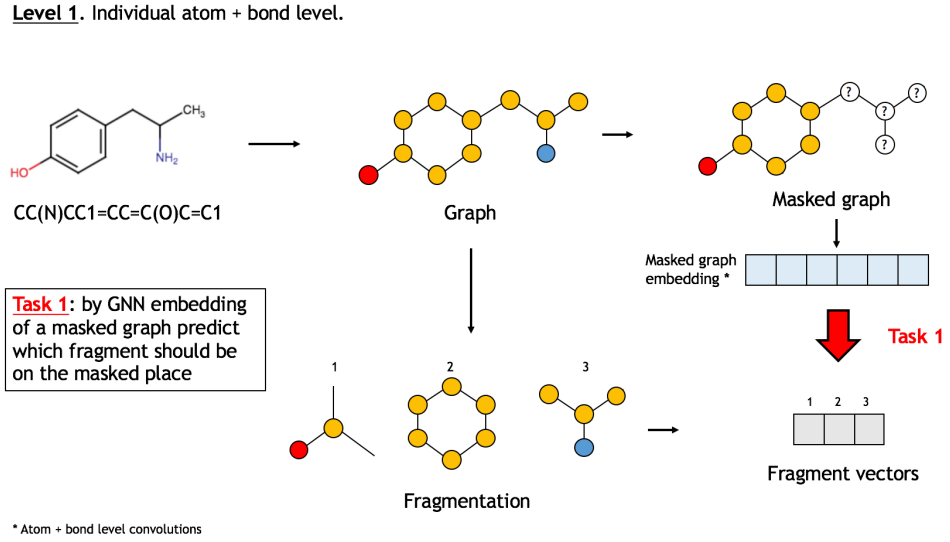


Figure 119. First self-supervised representation learning task.

For the second, graph-level pre-training task, each molecular graph is represented as a fixed-length fragment bit string, where each bit indicates the presence of one of 85 chemical fragments defined by a default fragmentation scheme in the RDKit Python package. Message-passing operations are applied to the entire unmasked graph, generating a graph-level embedding of the molecule. The second self-supervised learning task is to correctly identify which of the 85 fragments is present within the graph based on this graph-level embedding (Figure 120). This process further emphasizes the tool's capability to capture comprehensive structural information from molecular data.

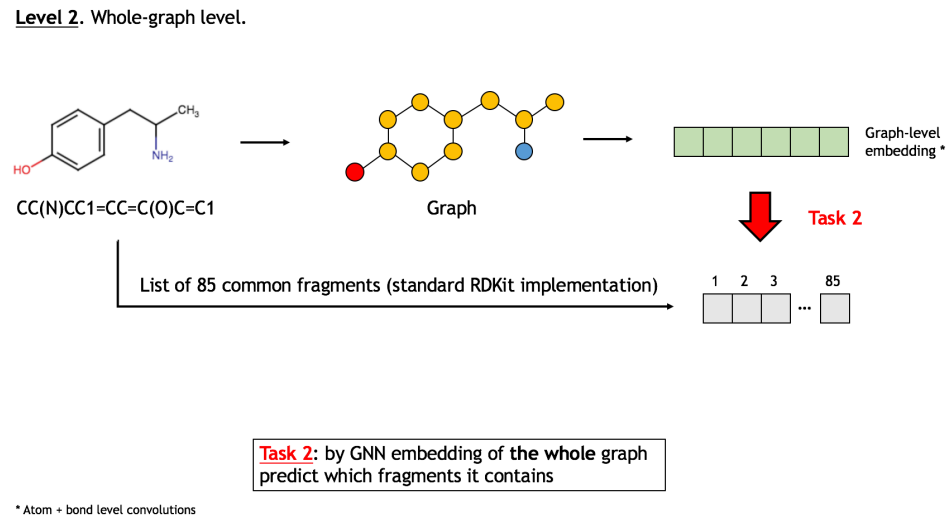


Figure 120. Second self-supervised representation learning task.

As a result, the final embedding obtained from both tasks constitutes atom- and bond-level embeddings for both tasks (Figure 121).

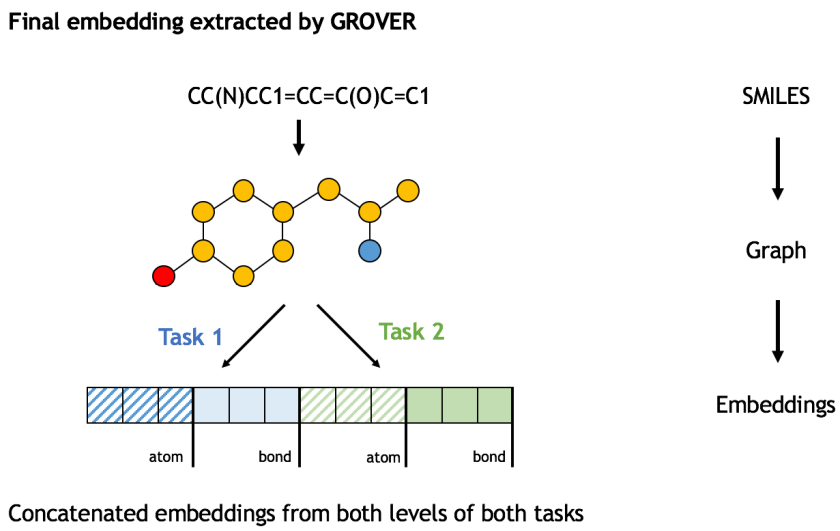


Figure 121. Final learned representation contains features from atom- and bond-level graph convolutions.

According to the authors, pre-training is an extremely time-and resource-consuming process, which required them to procure 250 top-budget NVIDIA graphics cards. Luckily, the authors provide two pre-trained models: GROVERlarge and GROVERbase. GROVERlarge was pre-trained on 11 million unlabelled molecules sampled from ZINC15 and ChEMBL datasets. The authors randomly split 10% of unlabelled molecules as the validation sets for model selection.

The lighter version, GROVERbase, was trained on the same data using the same architecture, but with smaller neural network layers, using a lesser number of hidden parameters. Specifically, GROVERbase contains ~48M parameters, and GROVERlarge contains ~100M parameters. The authors used 250 NVIDIA V100 GPUs to pre-train GROVERbase and GROVERlarge. Pre-training GROVERbase and GROVERlarge took 2.5 days and 4 days respectively.

Hence, for this work, we used the pre-trained GROVERlarge model to avoid having to pre-train it ourselves. Moreover, the original paper showed this model to have state-of-the-art performance on benchmark downstream QSAR tasks.

Fine-tuning of a pre-trained model with GROVER requires the input of a text file. Each line of this file should contain a SMILES string of a molecule and its corresponding target property value. For classification tasks, the target property should be a discrete integer value. In this project, we specifically focused on binary classification problems, leaving the tool's multi-class classification capabilities untested. For regression tasks, the target property should be denoted by a continuous number.

Only two metrics are supported to fine-tune regression or classification downstream tasks. For regression problems, the optimized metric is the coefficient of determination, R^2. The R^2

metric is a statistical measure that assesses the goodness of fit of a regression model to the observed data. It provides an indication of how well the model's predictions explain the variability in the dependent variable. R^2 is a real value between 0 and 1. It represents the proportion of the variance in the dependent variable that can be explained by the independent variables included in the model. In other words, it measures the proportion of the total variation in the dependent variable that is captured by the regression model. The R^2 value is calculated following equation 9, where SS_R (the residual sum of squares) is the sum of the squared differences between the predicted values and the actual values of the dependent variable, and SS_T (total sum of squares) is the sum of the squared differences between the actual values.

$$R^2 = 1 - (SS_R / SS_T) \quad (9)$$

To assess the predictive performance of classification models, we used the balanced accuracy (BA) metric, defined by equation 10, where *TP* stands for true positives (correctly labelled positive data points), *FN* stands for false negatives (incorrectly labelled positive data points), *TN* stands for true negatives (correctly labelled negative data points), *FP* stands for false positives (incorrectly labelled negative data points). In other words, it is calculated as the average of the proportion of correctly predicted observations in each class, namely, the sensitivity (true positive rate) and specificity (true negative rate). BA takes values from 1 (ideal case) to 0.5 (random predictions).

$$BA = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right) \quad (10)$$

Upon providing the input file, the tool randomly splits the dataset into training, testing, and validation sets in an 80-10-10 ratio (Figure 122). The user controls the number of different splits that will be generated from the input dataset. In this work, for all fine-tuning experiments, we fix this number to five data splits to ensure a robust and efficient sampling of the input data.

This sampling approach diverges from conventional K-fold cross-validation, as the partitioning into training, testing, and validation subsets occurs randomly. Therefore, it is not guaranteed that each molecule from the provided fine-tuning dataset is utilized for both training and testing.

During each optimization epoch, molecules from the training and validation sets are used to update the neural network's internal weights. After training concludes, the best weights are selected based on the best value of a task-defined metric on the validation set. These parameters are then utilized to make predictions for the compounds in the test set. Lastly, the metric values for these test set predictions are reported and used to compute the overall performance statistics.

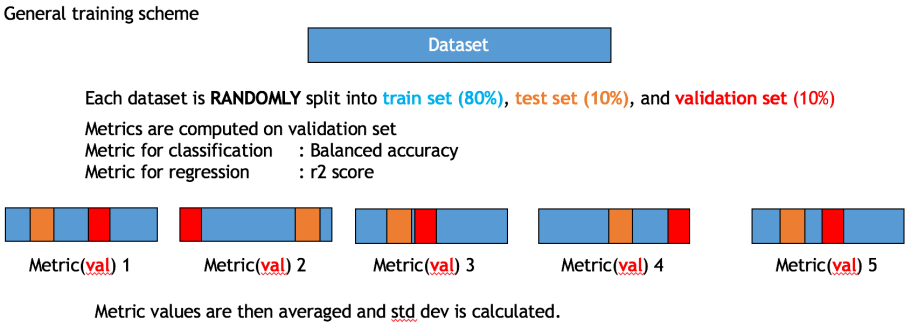


Figure 122. Schematic representation of used cross-validation strategy.

It is worth mentioning that the GROVER tool provides opportunity to concatenate other features to the learned representations. This often leads to better predictive performance, because selected additional features get concatenated with the learned representations.

After a pre-trained model was used to fine-tune representations for a downstream task, the prediction step is straightforward. The tool requires a fine-tuned model and a list of SMILES for prediction. If the user used additional features for fine-tuning, they also need to be pre-calculated for the prediction set before running a prediction job. The tool produces learned representations of the input molecules, concatenates them with pre-calculated additional features (if any are provided), and then uses a simple multilayer perceptron neural network model to make predictions of the target value. The output is a .csv file with each line containing a SMILES string and a predicted target value.

6.4. Optimization of hyperparameters

Our first task was to understand which hyperparameters of the GROVER tool lead to better predictive performance in the fine-tuning task when using a GROVERlarge pre-trained model. For benchmarking, we selected two datasets from MoleculeNet, which is a collection of structure-activity datasets widely accepted in machine learning community for benchmarking new contributions[150]. Specifically, we used datasets named freesolv and lipo. Freesolv is a benchmark dataset comprised of 642 small molecules along with their corresponding experimentally measured values of hydration free energy in water. Lipo is a benchmark dataset that contains experimentally measured values of octanol/water distribution coefficient (logD at pH 7.4) for 4200 small molecules. Both datasets concern the regression modelling task. However, lipo contains 4200 structure-property data pairs, while freesolv contains 642 such pairs. We deliberately chose these specific datasets because optimizing the hyperparameters using differently-sized benchmark

datasets allowed us to mitigate the effect of the dataset size on performance, focusing only on hyperparameter contributions.

The GROVER tool contains many configurable parameters that influence all parts of the transfer learning workflow. We arbitrarily selected 11 parameters that, in our opinion, had an influence on the quality of the resulting models. They are listed in Table 2 along with their conceptual meaning.

Table 2. Optimized parameters of the GROVER tool

Parameter	**Conceptual meaning**
features_generator	During fine-tuning, which additional features to concatenate with learned features
num_folds	How many different train, test, validation splits will be performed
split_type	How would molecules be selected into train, test, and validation sets for each fold
ensemble_size	How many GROVER models will be trained for each train, test, validation split
no_features_scaling	This is a flag that, if present, tells the tool **not** to scale the input features from 0 to 1
ffn_hidden_size	Size of the layer of a multi-layer perceptron network used in fine-tuning
ffn_num_layers	Number of hidden layers in a multi-layer perceptron network using in fine-tuning
self-attention	A flag that, if present, changes the way graph embeddings are calculated
bond_drop_rate	Probability of dropping random bonds from a graph (to prevent overfitting)
batch_size	During fine-tuning, how many molecules from the dataset will be used for a single iteration of internal weights change
epochs	During fine-tuning, how many times would the network iterate over the input dataset

The process of hyperparameter optimization consisted of performing fine-tuning on the freesolv and lipo datasets with different combinations of the parameters listed in Table 2. Each successfully finished fine-tuning run was described by an average value of the optimized metric over all the folds (specified by num_folds parameter), and a standard deviation of the R^2 metric value.

Thus, we were able to establish a set of the GROVER tool's hyperparameters that consistently lead to higher quality predictions than the default parameters. These parameters are listed in Table 3.

Table 3. Best hyperparameters of the GROVER tool found by iterative fine-tuning on freesolv and lipo datasets

Hyperparameter	**Best value after optimization**
features_generator (optional)	Normalized 2D descriptors derived from the topology of the molecule or counts of specific types of atoms or bonds; Counts-based Morgan fingerprint characterizing the local chemical environment around each atom in the molecule up to a specified radius, encoded as a fixed-length binary vector.
num_bits (optional)	2048
split_type	Random
ensemble_size	5
num_folds	5
no_features_scaling	(flag turned on)
ffn_hidden_size	400
ffn_num_layers	2
self_attention	(flag turned on)
bond_drop_rate	0.5
batch_size	32
epochs	100
init_lr	0.00015
select_by_loss	(flag turned on)

Then, we aimed to evaluate the predictive performance of the fine-tuned GROVER models against a state-of-the-art QSAR modeling approach. It relies on support vector machine models optimized through an evolutionary model building process that explores up to ninety-five ISIDA fragmentation schemes to select the best descriptor space. In this approach (henceforth referred to as libsvm-GA), models that perform optimally are identified through the evolutionary procedure and ranked based on their fitness score, which is calculated as the average BA across a 12-fold repetition of a leave-1/3-out cross-validation scheme.

For the purpose of benchmarking, five different datasets from the MoleculeNet benchmark set were used, namely bace, bbbp, esol, freesolv, and lipo. BACE is a dataset of 1513 compounds labelled with qualitative (binary label) binding results for human β-secretase 1 (BACE-1). BBBP is a dataset of 2039 compounds with binary labels of blood-brain barrier permeability. ESOL is regression modelling dataset with 1128 common organic small molecules and their water solubility data (log solubility in mols per litre).

Firstly, we used the state-of-the-art approach of using libsvm-GA on ISIDA descriptors. That means we computed 95 sets of ISIDA descriptors and used a genetic algorithm to find optimal descriptor set and hyperparameters of support vector machine models for all of the five datasets. Secondly, we used default GROVER parameters to model these datasets. Thirdly, we used the best GROVER hyperparameters, not using additional features. Finally, we used the best GROVER hyperparameters with additional features (two hundred normalized physicochemical 2D RDKit descriptors, and a count vector of Morgan fingerprints with radius of 2 and number of bits of 2048).

As a result of that, we saw that for the binary classification task involving the BACE and BBBP datasets, molecular representations learned by fine-tuning the GROVER model using best hyperparameters (Fir. 123, green) led to better performance that default GROVER hyperparameters (Figure 123, orange) and were on par with classification performance shown by the libsvm-GA approach (Figure 123, blue). Concatenation of additional features (normalized 2D RDKit and count of Morgan fingerprints (r=2, num_bits=2048)) to the learned representations did not necessarily improve the classification performance (Figure 123, red).

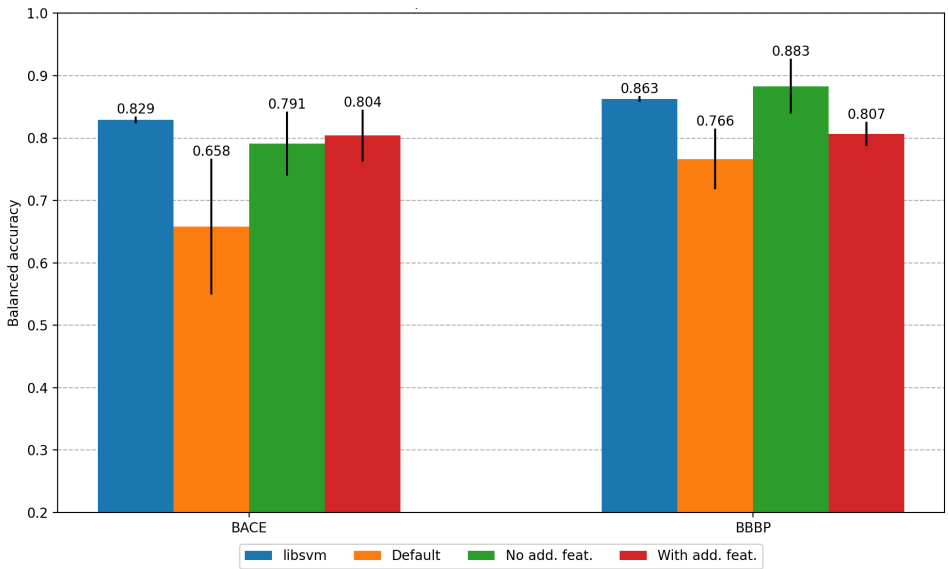


Figure 123. Reported are cross-validation scores achieved by four different methods.

Table 4. Cross-validation statistics of models' performance on binary classification datasets

Method	**Balanced accuracy score (BACE)**	**Balanced accuracy score (BBBP)**
Libsvm-GA	**0.829 ± 0.006**	0.863 ± 0.005
GROVER default	0.658 ± 0.109	0.766 ± 0.049
GROVER best (no added features)	0.791 ± 0.051	**0.883 ± 0.044**
GROVER best (with added features)	0.804 ± 0.042	0.807 ± 0.020

In a similar manner, regression modelling performance of the fine-tuned GROVER models with best found hyperparameters (Figure 124, green) was constantly higher than that achieved by fine-tuning with default parameters (Figure 124, orange). However, learned representations could not produce models scoring higher than the state-of-the-art approach (Figure 124, blue). At the same time, concatenation of the learned representations to the additional descriptors (normalized 2D RDKit and count of Morgan fingerprints (r=2, num_bits=2048) led to significant boost of performance (Figure 124, red).

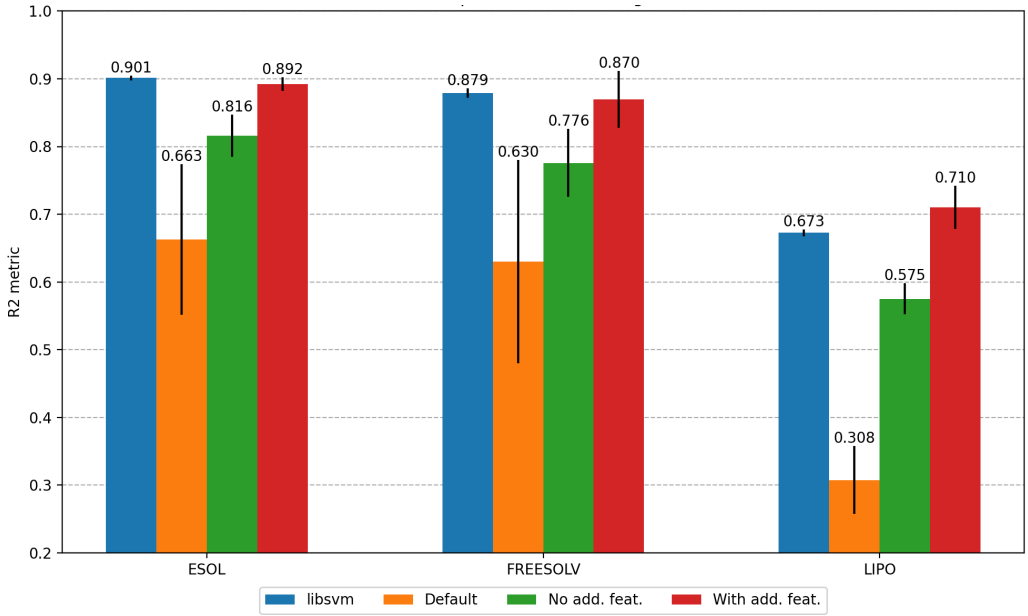


Figure 124. Reported are cross-validation scores achieved by four different methods.

Table 5. Cross-validation statistics of models' performance on regression datasets

Method	R^2 **score (ESOL)**	R^2 **score (FREESOLV)**	R^2 **score (LIPO)**
Libsvm-GA	**0.902 ± 0.004**	**0.879 ± 0.007**	0.673 ± 0.005
GROVER default	0.663 ± 0.111	0.630 ± 0.150	0.308 ± 0.051
GROVER best (no added features)	0.816 ± 0.031	0.776 ± 0.050	0.575 ± 0.023
GROVER best (with added features)	0.892 ± 0.011	0.870 ± 0.042	**0.710 ± 0.032**

Thus, we were able to establish the best hyperparameters of the GROVER models that constantly led to better predictive performance than the default parameters on various datasets and modelling tasks. We also show that combined use of the learned representations with other descriptors may further increase predictive performance of the models.

6.5. Downstream task fine-tuning performance

To assess the performance of the GROVER model applied via transfer learning on unseen data, we subjected it to the task of classifying compounds based on their cytotoxicity towards different cancer cell lines most likely caused by binding at the tubulin protein's colchicine binding site. For this purpose, we employed a dataset published by López-López et al. consisting of 766 structure-activity data points, each representing a molecular structure along with its corresponding activity label[129]. The dataset constitutes a binary classification task, with the activity label reflecting the compound's ability to inhibit the proliferation of various cancer cell lines. Molecules were labelled as active if they induced cytotoxic effects on any cell line at concentrations below micromolar.

This dataset was split into a training set for hyperparameters selection via cross-validation and an external validation set for assessing the predictive performance of the cross-validated models. The validation set contained 50 randomly chosen active and inactive molecules (25 of each class). We used the BA metric for cross-validation and performance assessment.

The classification performance benchmarking was carried out for the following models (Figure 125):

- Libsvm-GA using ISIDA descriptors (representing the state-of-the-art approach)
- A random forest model trained on 2D RDKit descriptors (representing a widely used baseline approach)
- The GROVER model with optimal parameters and without additional features
- The GROVER model with optimal parameters, incorporating additional features.

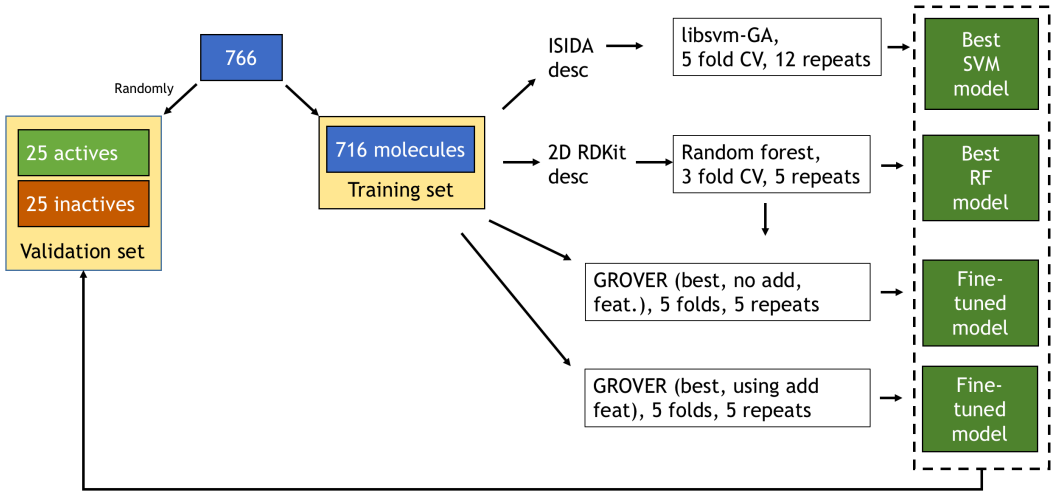


Figure 125. Scheme of model training and validation.

Figure 126 shows average balanced accuracy score of cross-validation attempts and prediction for the four tested models.

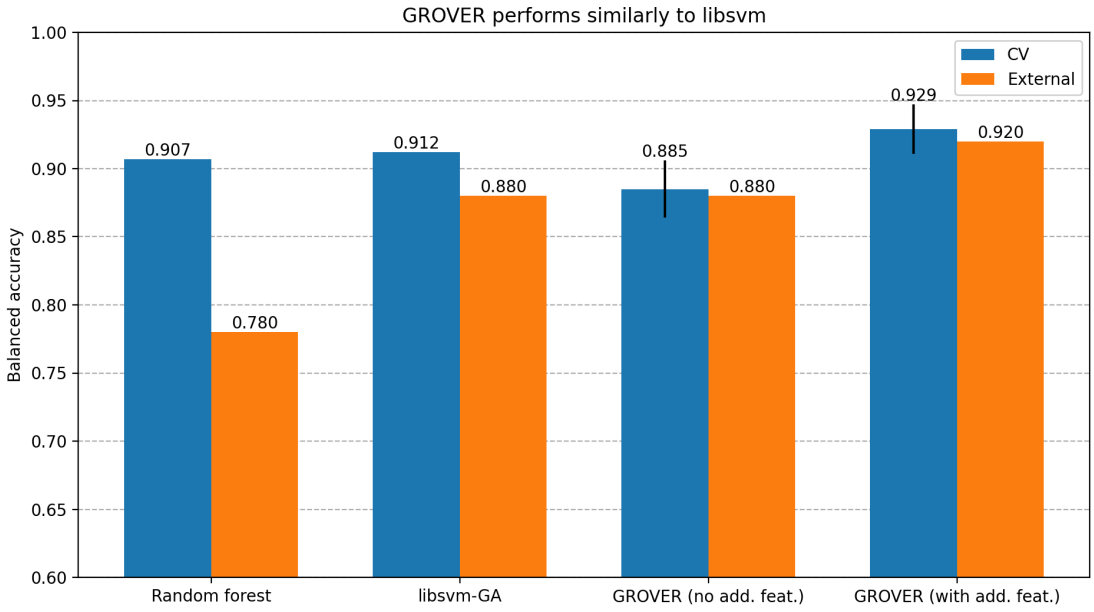


Figure 126. Performance of four different models on colchicine compounds classification.

As can be seen from the figure, transfer learning with GROVER is able to learn representations that allow for comparable performance with state-of-the-art libsvm-GA on ISIDA descriptors approach, outperforming the basic baseline of random forest on 2D RDKit descriptors. Additional features (normalized 2D RDKit descriptors and count vector of Morgan fingerprints (r=2, num_bits=2048)) concatenated to the learned representations further boost the performance (see *GROVER (with add. feat.)* on Figure 126).

Thus, transfer learning with GROVER shows comparable performance to the state-of-the-art approach, that can be further boosted by concatenating the learned representations to other descriptors, proving the utility of transfer learning for molecular representation learning in QSAR modeling.

6.6. Extraction of learned representations

We then wanted to understand whether learned representations from pre-trained and fine-tuned models could be successfully utilized with other machine learning methods beyond multilayer perceptron to yield high-performing models. For this, we isolated atom-level and bond-level representations from the GROVERlarge pre-trained model, provided by the authors of the original publication, for 716 molecules from the training set. These representations were then employed to train libsvm-GA models, during which only the hyperparameters of the SVM models were optimized. The numbers were compared to the performance results shown by libsvm-GA models after optimization over ISIDA fragment descriptor sets as performed in the previous section.

Building on this, we repeated the process but with representations extracted from a fine-tuned model, anticipating these would improve cross-validation and external validation predictive performance metrics. Although our predictions were correct and there was indeed an improvement, the increase was only modest and did not surpass the performance level of libsvm-GA trained on ISIDA descriptors. Figure 127 provides a visual representation of these findings.

This suggests that while fine-tuned representations may offer some advantages, pre-trained models are already quite proficient in capturing key molecular features, as reflected in their competitive performance. This conclusion underscores the robustness of the transfer learning methodology and its applicability in QSAR modelling, even if the performance gains from fine-tuning are not substantial. Our exploration serves as a valuable step forward in harnessing the full potential of transfer learning in chemoinformatics.

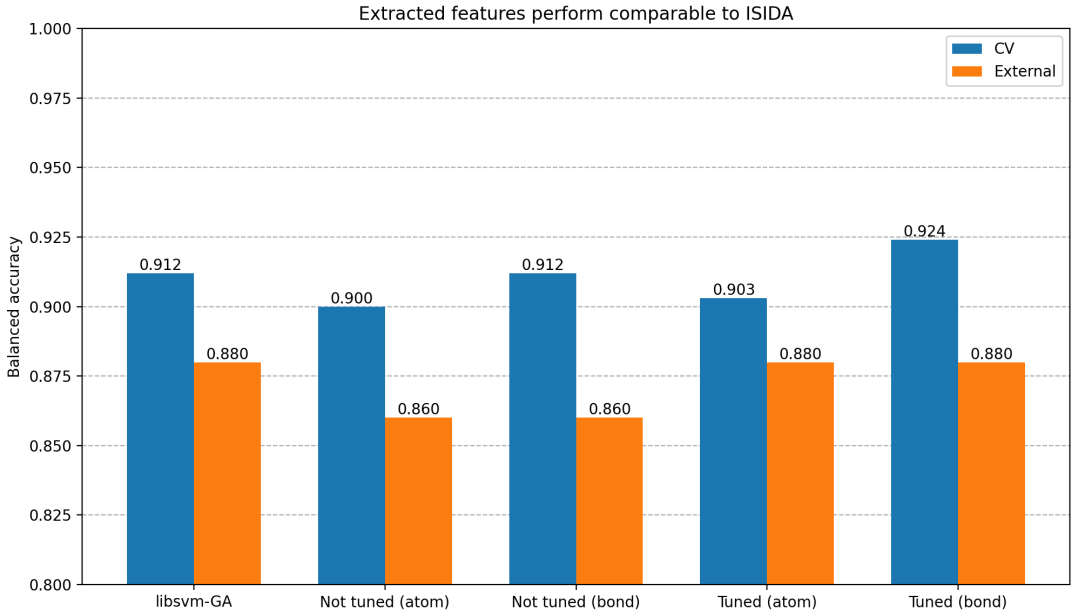


Figure 127. Pre-trained and fine-tuned features work well with other machine learning methods.

Table 6. Libsvm-GA performance on pre-trained and fine-tuned GROVER representations and ISIDA descriptors

	Representation level	**Cross-validation BA**	**External validation BA**
Not tuned GROVER representations	Atom	0.900	0.860
	Bond	0.912	0.860
Tuned GROVER representations	Atom	0.903	0.880
	Bond	0.942	0.880
ISIDA descriptors	–	0.912	0.880

6.7. Conclusion and perspectives

In summary, this work shows that the transfer learning approach is useful for QSAR modelling. It is capable of learning representations that are at least comparable in predictive performance to state-of-the-art evolutionary optimization of SVM models over the ISIDA fragment descriptor sets approach. As such, it indeed becomes possible to learn useful representations for a task at hand, rather than optimize descriptors sets. However, there are certain difficulties that need to be addressed if this approach is to be taken further.

The GROVER tool, selected for this project, implements a sub-optimal cross-validation strategy. Random sampling of the fine-tuned dataset repeatedly causes situations when

performance metric values considerably differ from fold to fold due to composition of train, validation, and test sets. We envisage that implementation of a more conventional *N*-fold cross-validation strategy to ensure that each data point has at least once been in train and validation sets would bring more robustness to the results.

The speed of fine-tuning highly depends on the batch size. Unfortunately, our GPUs did not allow us to use batch sized larger than 128. It means that for bigger datasets (where GROVER works best, e.g., for datasets containing more than 600 points), fine-tuning could take as much as several days. A workaround may be to utilize distributed GPU use.

For each fold of the fine-tuning dataset, it is possible to build not one, but several fine-tuned models. This is controlled by the ensemble_size parameter of the GROVER tool. Our study shows that this is a beneficial approach that improves performance overall. Most likely in a similar way to random forest, where an ensemble of weak predictors gives a good overall prediction. However, when it comes to extracting features, it is unclear which fine-tuned model should be used for feature extraction. In this work, we used first models from folds that reported the highest metric values. However, this doesn't feel optimal. Maybe a better approach would be to extract features from all models from the ensemble and take the average value of the features.

Although this was not the purpose of the project, we observed that concatenating learned representations with normalized 2D RDKit descriptors and a vector that counts unique Morgan fingerprint bits (r=2, num_bits=2048) adds considerable boost for modelling performance. Thus, when simply concerned with resulting model's quality, and not with learning representations *per se*, this may be a recommended strategy to apply.

Descriptor selection is a tedious task prone to finding sub-optimal solutions. Transfer learning is a concept that promises to learn useful molecular representations instead of selecting them from a pre-defined list of descriptors sets. Theoretical applicability of this approach for QSAR modelling has been demonstrated on classification and regression task. We have shown that, with right hyperparameters, transfer learning with the GROVER tool reaches the predictive performance level of the state-of-the-art evolutionary optimization of SVM models over the ISIDA fragment descriptor sets approach. However, challenges related to cross-validation strategy, ensemble modelling, and applicability domain of the learned representations remain open. They may be subjects of further research on this topic.

Chapter 7. Exploration of cryptic binding pockets in tubulin using Gaussian-accelerated molecular dynamics simulations

7.1. Introduction

Proteins are dynamic entities that continuously undergo conformational changes, which fundamentally define their biological structure[151]. This dynamism is intricately linked to the biological function of proteins. The conformational diversity that arises from protein dynamics is a crucial factor in understanding how a protein functions. A key aspect of protein structure and dynamics is the existence of binding sites, which are often open and ready for ligand interaction[152]. A binding site is a specific region on a protein where a ligand, such as a small molecule, can bind. Binding sites usually appear as pockets or grooves on the protein's surface and are typically present in the protein structure even in the absence of the drug. Ligand interaction with a binding pocket is typically specific, meaning that the binding site corresponds to the shape, size, and chemical characteristics of the ligand. The binding of ligands to these sites can induce alterations in the protein's structure and, hence, it's biological function[152].

However, the dynamic nature of proteins also gives rise to what are known as cryptic pockets. Cryptic binding pockets present a unique case where the pocket or groove only forms during the process of drug binding, and prior to this, the site lacks the usual geometric features of a typical binding site. As such, they are transiently occurring binding sites in proteins that are not present in the protein's static structure but emerge due to conformational changes intrinsic to the protein or induced by a given ligand (Figure 128). These pockets can be thought of as hidden opportunities for molecular interaction that only become apparent under certain conditions[153,154].

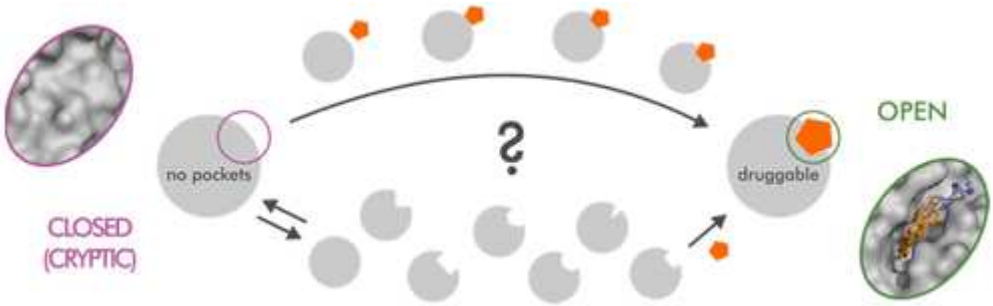


Figure 128. Schematic representation of two possible mechanisms of cryptic pockets formation (adapted from Kuzmanic et al.[155])

Cryptic pockets can sometimes be targetable by small molecules, and their study holds significant potential for drug discovery[152]. Such pockets are often found at the interfaces between proteins, suggesting that drugs designed to target these sites could potentially modulate specific protein-protein interactions, a strategy with considerable therapeutic promise[152]. They offer unique

and potentially more specific opportunities for drug binding, which can lead to the development of more effective and targeted therapeutics, particularly in cases where other known binding sites have proven challenging to target[154,155]. The identification and study of cryptic pockets thus provides a deeper understanding of protein function and dynamics, revealing potential new targets for drug discovery and contributing to the development of more effective and specific therapeutics. Despite this potential, the practical application of targeting cryptic sites is challenging, largely due to our limited ability to identify such cryptic pockets, which are often found by serendipity, and understanding of how small molecules interact with these cryptic sites compared to conventional well-defined binding sites[155].

Computational simulations play a crucial role in identifying such cryptic pockets[156]. Advanced computational methods, such as atomistic molecular dynamics simulations, can capture the protein's conformational changes over time, revealing the transient pockets that may not be visible in a static structure. Technological advancements in computer architectures specifically designed for MD simulations, coupled with the rise of distributed computing and the optimization of MD simulation packages for parallel processing and GPU utilization, have allowed scientists to simulate systems of unprecedented size and duration. It is now commonplace to conduct simulations spanning microseconds. Simultaneously, the accuracy of both protein and ligand force fields has improved to the point where they can reliably capture the key aspects of target dynamics and ligand binding mechanisms. These simulations can provide a dynamic view of the protein, allowing for the identification and characterization of cryptic pockets[156].

Conventional MD simulations are not without their limitations. Despite the substantial progress, conventional approaches in MD simulations cannot adequately sample many biologically and pharmaceutically relevant conformational changes. In classical molecular dynamics simulations, the system evolves over time according to Newton's laws of motion, with the potential energy surface (a multi-dimensional representation of the energy of a system as a function of its atomic positions) guiding the movements of the atoms. However, this approach can be inefficient when dealing with systems that have high energy barriers on their potential energy surfaces, as the system can get trapped in local minima for long periods, thereby hindering efficient conformational sampling. This is particularly problematic when studying cryptic pockets, which may only be transiently present during rare conformational changes[84,154,155,157].

To overcome these limitations, several enhanced sampling techniques have been developed for MD simulations. One such technique is Gaussian-accelerated molecular dynamics (GaMD). GaMD differs from conventional molecular dynamics in that it introduces a harmonic boost potential to the system's potential energy surface, effectively smoothing the energy landscape and facilitating easier conformational transitions Figure 129). This boost potential lowers the energy

barriers, allowing the system to escape from local minima more easily and explore the conformational space more efficiently. The boost potential is constructed using a Gaussian distribution, hence the name of the method. The parameters of the Gaussian distribution are adjusted dynamically during the simulation to ensure that the boost potential is always appropriate for the current state of the system[84,157].

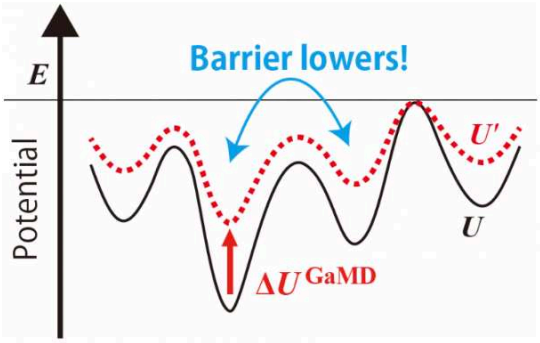


Figure 129. Scheme of GaMD boosting

GaMD has numerous advantages over classical MD. Firstly, it allows for more efficient sampling of the conformational space, which is crucial in studies of protein dynamics. Proteins can adopt a vast number of different conformations, and understanding this conformational variability is key to understanding protein function. By enabling the system to overcome high energy barriers more easily, GaMD allows for a more comprehensive exploration of the conformational space. Secondly, GaMD simulations can be performed without prior knowledge of the system's potential energy surface, making it a versatile tool for studying a wide range of systems[84,157].

In this study, we aimed to apply GaMD to the simulation of the tubulin protein, with the objective of identifying potential cryptic pockets on its surface. Previously, Muhlethaler et al. have performed classical MD simulation-based study of the tubulin protein's conformational dynamics and identified 26 distinct pockets on the tubulin surface, often related to known binding sites[106]. Hence, our goal in this project was to determine whether the analysis of the tubulin trajectory obtained by enhanced sampling with GaMD could not only reproduce the known binding sites but also discover new possible cryptic pockets. The discovery of such pockets could provide valuable insights for future research targeting tubulin. This work was done as part of my academic secondment to the Univeristy of Barcelona, Barcelona, Spain, under the guidance of Prof. J. Rubio and Prof. M. Cascante, in the framework of the TubInTrain curriculum.

7.2. Modelled system setup for simulations

We retrieved the three-dimensional structure of the α,β-tubulin heterodimer from the Protein Data Bank (PDB) record 7E4Z. The preparation of the protein structure involved the removal of all solvent molecules and small organic molecules, excluding GTP. We kept the manganese ions. We deleted all chains except for chains C and D, corresponding to α-tubulin and β-tubulin parts of the heterodimer. The residues of the gap in the loop between strand S7 and helix H8 of b-tubulin were sourced from the 4I4T PDB complex, where this loop is resolved. The missing atoms from these residues were completed using the LEaP module of the AMBER18 software[158].

We performed three Gaussian-accelerated molecular dynamics runs of 1 μs. All calculations were performed using the PMEMD (Particle Mesh Ewald Molecular Dynamics) code of the AMBER18 software in its CUDA version, employing the AMBER ff14SB force field[159].

The systems were prepared for molecular dynamics simulations following a common protocol. Initially, the prepared protein was immersed in a cubic box filled with equilibrated TIP3P water molecules. To neutralize the system, some water molecules were replaced with Na^+ or Cl^- ions, based on the electrostatic potential before solvation. The cubic periodic box was constructed to maintain a minimum distance of 16 Å between the protein and the box edges. We also deleted water molecules that were closer than 1 Å to the protein.

Subsequent to system preparation, energy minimization was carried out in a stepwise manner. The positions of the water molecules and ions were first optimized using the steepest descent (SD) algorithm up to 5000 cycles of minimization, while keeping the rest of the system fixed. The modeled residues were then relaxed in two stages, each consisting of 5000 cycles of SD, with the backbone positions of these modeled residues kept fixed and a decreasing force constant of 5.0 and 0.1 kcal/Å. Lastly, the minimization of the entire system was carried out without any restrictions using 10000 cycles of the SD method.

Following minimization, the system was heated in increments of 30 K every 20 ps, using a force constant of 1.0 kcal/mol·Å to maintain all backbone atoms constrained. The heating process was performed under the canonical (NVT) ensemble. After the heating process, a 200 ps trajectory at constant pressure (NPT ensemble) was performed for density equilibration using the Berendsen barostat to control and maintain the pressure at 1 atm.

The final structure served as the starting point for the production MD simulations in the NVT ensemble. Trajectories were calculated at 300 K using the Langevin thermostat to maintain a constant temperature with a collision frequency of 3.0 ps^{-1}. The SHAKE algorithm was used to fix all bonds involving hydrogen atoms, enabling us to use a time step of 2 fs for all the

simulations[160]. Nonbonded interactions were truncated using a cutoff of 9 Å, and long-range interactions were treated with the particle-mesh Ewald summation method with a grid spacing in the direct lattice of about 1 Å, a fourth-order B-spline interpolation for the gridded charge array, and a direct sum tolerance of 10^{-5}.

7.3. Root mean square deviation and fluctuation analysis

To evaluate the structural stability of the systems throughout the simulation, we calculated the root-mean square deviation using the cpptraj module from Amber18. The RMSD was computed relative to the initial structure obtained from LEaP. We reoriented each frame of all trajectories over all residues from both α-tubulin and β-tubulin subunits, utilizing the α carbons (Cα) of all the residues. In addition, we computed the root-mean square fluctuations (RMSF) for all tubulin residues using the cpptraj module of Amber18. This analysis provided insight into the local conformational flexibility of each residue during the MD simulations.

We calculated the RMSD for the entire trajectory to determine the structural stability of the Gaussian accelerated MD simulation. Figure 130 displays the average values from all three replicas of each system. As depicted, all systems achieved structural stability after a brief stabilization period. The RMSD of the Cα atoms of the α,β-tubulin heterodimer from the initial X-ray structure is plotted as a function of time, both with (red) and without (blue) the H1-S2, M, and S9-S10 loops of both tubulin monomers, following the same analysis done by Muhlethaler et al.[106] The lighter color represents the effective sampling of the RMSD during the simulation, while the darker lines represent a Bezier curve approximation of the data to minimize noise.

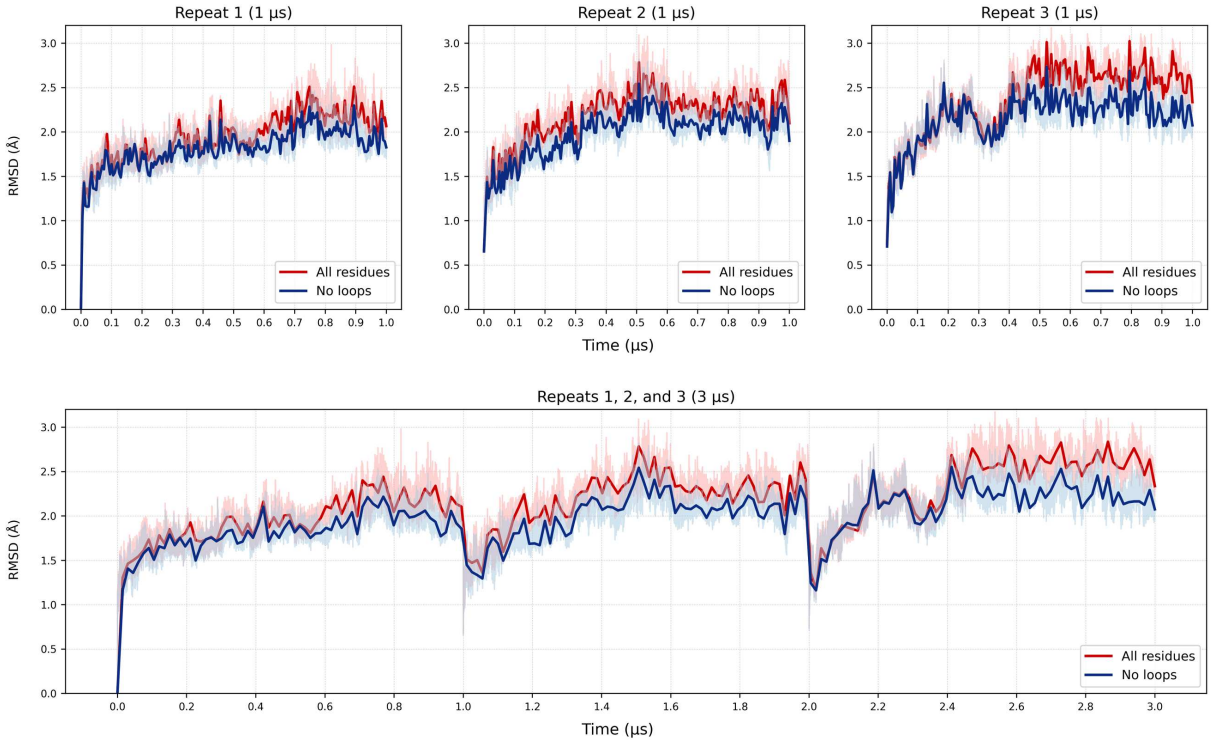


Figure 130. RMSD plots for the trajectory

We also computed the RMSF to evaluate the fluctuations of different regions of the tubulin structure (Figure 131). As anticipated, the structurally-important loops exhibited considerably higher fluctuations compared to the residues forming loops that do not participate in bonding with laterally or longitudinally located tubulin dimers to form a microtubule.

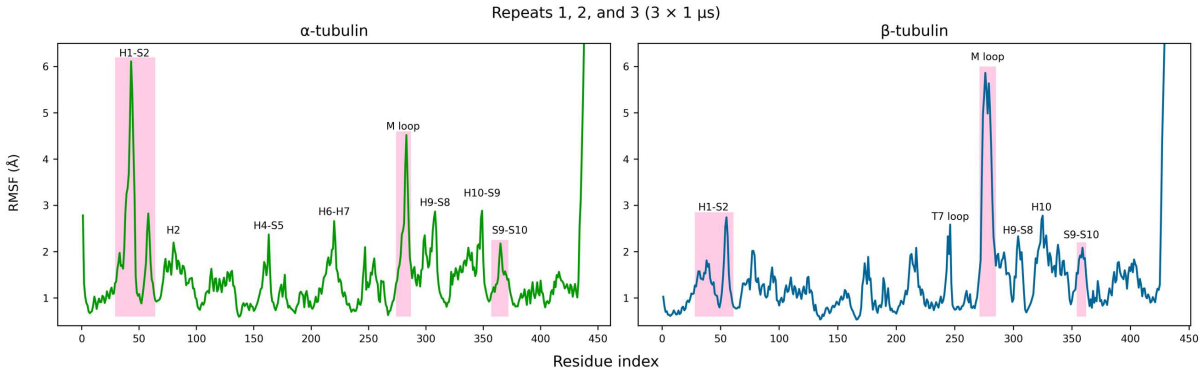


Figure 131. The degree of conformational flexibility across the residues in the tubulin protein during the molecular dynamics simulation

7.4. Principal component analysis

We used the principal component analysis (PCA), a multivariate statistical technique, to determine and analyze the primary structural variations of the studied system. PCA is particularly useful in capturing the most important features of protein dynamics while minimizing the

dimensions required for description. This dimensionality reduction is achieved through a decomposition process that ranks motions from the largest to the smallest spatial scales, thereby preserving as much variation in the data as possible.

The PCA methodology[161] involves the construction of a covariance matrix using the atomic coordinates of the alpha carbons (Cα) of each residue. This 3N × 3N symmetric matrix is subsequently diagonalized to yield a set of Principal Components or eigenvectors, along with their corresponding eigenvalues λ(*i*). The transformation of correlated variables into uncorrelated ones through PCA allows the first principal modes or eigenvectors to characterize large-scale protein motions. These first modes are sufficient to define the "essential" space or motions of the protein. The contribution of the *i*-th principal component PC(*i*) to the structural variance in the data set is given by equation 11, where the summation is performed over all 3N components, and N is the number of residues in the protein.

$$c\% = 100 \times \frac{\lambda^i}{\sum_{i=1}^{3N} \lambda^i} \qquad (11)$$

For our PCA analysis, we constructed the covariance matrix using the alpha carbons (Cα) of the α,β-tubulin residues, excluding both termini of each tubulin subunit. The first Principal Component (PC1) accounted for approximately 31% of protein fluctuations, and in conjunction with the second component, the two initial PCs described about 45% of protein fluctuations (Figure 132).

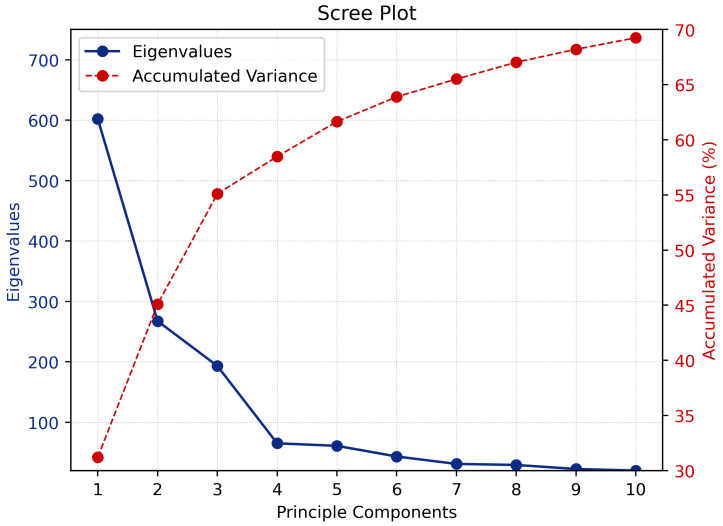


Figure 132. Scree plot illustrating the explained variance by each principal component in our PCA analysis of tubulin dynamics

The projection of each MD trajectory snapshot onto the respective two principal components is depicted in Figure 4. A comparison of these figures reveals that the sampling of the three replicas is not necessarily identical, reinforcing the notion that conducting multiple MD runs may be more efficient than generating a single trajectory from a single run. Each of the three GaMD runs explored distinct sections of the conformational space.

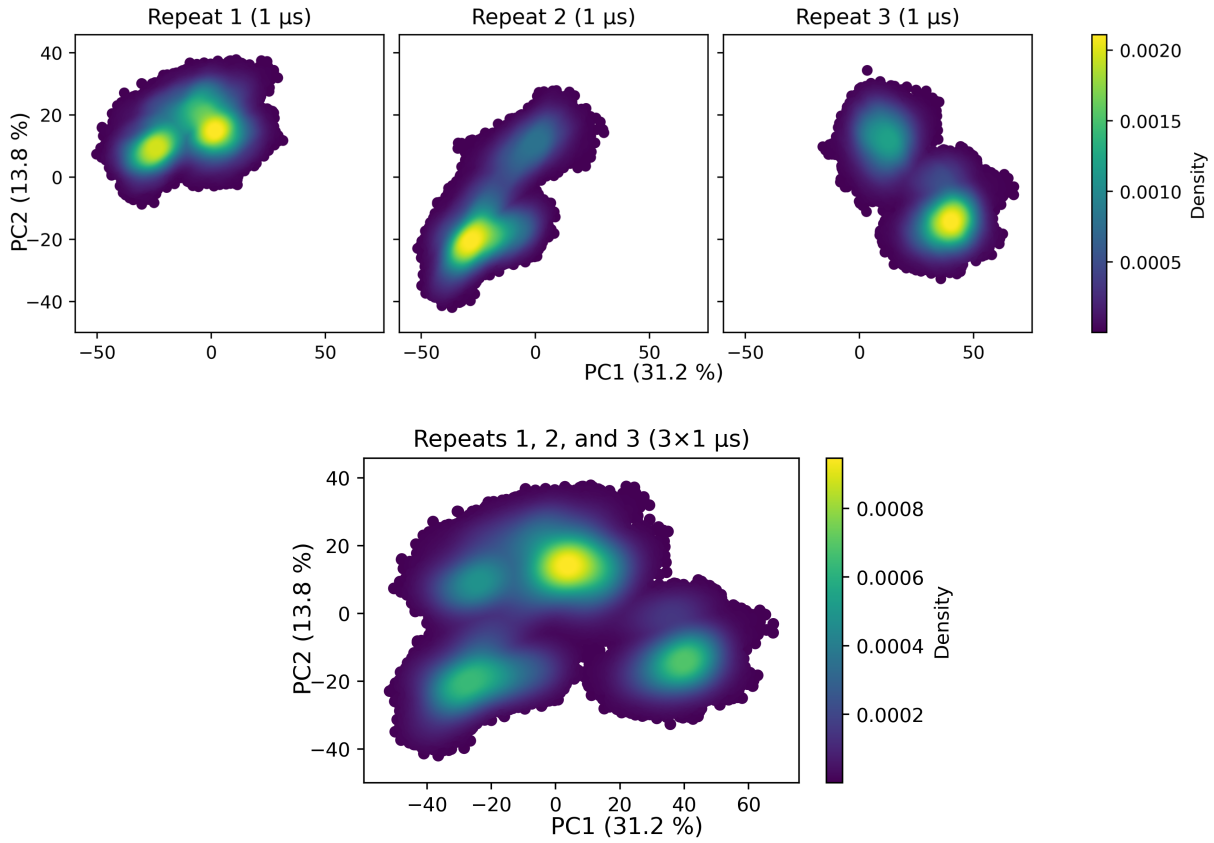


Figure 133. The conformational landscape of tubulin dynamics sampled by GaMD simulations

7.5. Cluster analysis

We then sought to identify the distinct structural features of tubulin by categorizing similar structures from the entire simulation trajectory into 15 different clusters. This was achieved using the average linkage algorithm, as implemented in the cpptraj module of AMBER18. The Root Mean Square Deviation (RMSD) of the Cα in residues αMet1-αVal440 and βArg1-βAsp441 served as the distance metric for this process. These residues were selected because they comprehensively represent the entire protein system.

The combined trajectory from the three runs was utilized for the clustering analysis to ensure the inclusion of all accessible states throughout the full length of the MD. This amounted to the clustering of 150,000 frames, equivalent to the extraction of one structure every 2

picoseconds. Furthermore, a sieve option of 4 was employed in the clustering process, which implies the use of only 37,500 frames, with the remaining frames added to the closest cluster in each instance. Consequently, the centroids of each of the 15 clusters were selected as representatives to shed light on the potential hotspots in the tubulin structure.

Figure 134 shows the coordinates of the representative structures mapped onto the PC1/PC2 plot. The representatives are numbered in accordance with the fraction of the total trajectory they describe, indicating that representatives with higher numbers cover a larger portion of the trajectory.

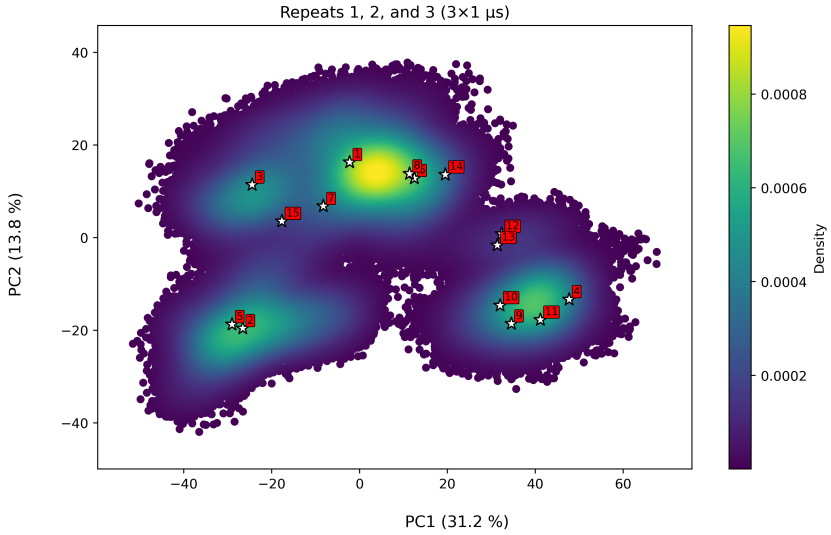


Figure 134. Representative structures selected after clustering projected on the sampled conformational landscape

7.6. Identification of cryptic pockets

We used the FTMap server to identify protein hotspots within our systems[162]. FTMap is a computational mapping server that is widely used to pinpoint binding hot spots in macromolecules. The algorithm operates by distributing a vast array of small organic molecules, or probes, of varying size, shape, and polarity. These probes are then scored based on a detailed ligand binding free energy calculation. The probes used in FTMap include acetaldehyde, acetamide, acetone, acetonitrile, benzaldehyde, benzene, cyclohexane, dimethyl ether, ethane, ethanol, isobutanol, isopropanol, methylamine, N,N-dimethylformamide, phenol, and urea. Binding hot spots are identified as regions where clusters of multiple probe types bind. FTMap distributes these probes on the protein surface, identifies the most favorable positions for each probe type, clusters the probes, and ranks the clusters based on their average energy. Regions that bind several different probe clusters are referred to as consensus sites and are predicted to be binding energy hot spots.

We utilized the FTMap server to identify cryptic pockets in tubulin and subsequently characterized these pockets for their potential as allosteric sites. This analysis was conducted on the representative structures of the 15 clusters. Our aim was to identify any binding pockets that did not correspond to the locations of known ligand binding, did not correspond to the locations where small fragments bound experimentally during fragment screening, and were not the pockets found by classical MD in a study published by Muhlethaler et al.[106]

Our analysis revealed that in all of the representative structures, we consistently found the gatorbulin site, the colchicine site, and the GTP on α-tubulin. Additionally, in some representative structures, we identified the taxane site, the vinblastine site, and the common entrance part of the todalam/pironetin sites. This not only validates the FTMap approach but also indicates that some binding sites are particularly susceptible to opening and closing due to the dynamics of the tubulin protein. Furthermore, all of the representative structures identified βIII, βV, and αII binding sites found experimentally in fragment screening (pocket names given following the notation used in [106]). The FTMap tool was also successful in identifying binding sites αI, aII, and bIV, which were found in a previous computational study by classical MD simulations (pocket names given following the notation used in [106]).

Intriguingly, our study unveiled four novel cryptic pockets (Figure 135), which have not been previously identified by either X-ray crystallography or classical MD simulations, and are not known to host any known ligands.

The first novel pocket (Pocket 1) is situated at the interface of α,β-tubulin. This pocket was identified through the binding of a variety of probes, including phenol, isopropanol, dimethyl ether, isobutanol, benzaldehyde, ethanol, ethane, benzene, N,N-dimethylformamide, acetonitrile, acetamide, acetone, and acetaldehyde. The residues that constitute this pocket include βAsp329, βGlu330, βMet332, βLeu333, βVal335, βGln336, αArg221, αTyr210, αGln176, and αArg214.

The second novel pocket (Pocket 2) is also located at the α,β-tubulin interface. It was discovered through the binding of benzaldehyde, cyclohexane, phenol, benzene, ethanol, N,N-dimethylformamide, isobutanol, acetone, and acetonitrile probes. This pocket is composed of residues βLeu248, βGln247, βPro245, αTyr224, αGln15, and αGln11.

The third novel pocket (Pocket 3) is located on α-tubulin. This pocket was identified through the binding of phenol, isopropanol, dimethyl ether, isobutanol, benzaldehyde, and cyclohexane probes. The residues that constitute this pocket include αLeu189, αThr193, and αHis192.

The fourth novel pocket (Pocket 4) is also situated on α-tubulin. It was discovered through the binding of urea, methylamine, acetaldehyde, ethanol, and ethane probes. This pocket is

composed of residues αSer198, αGlu155, αMet154, αGlu168, αThr194, αHis139, αThr150, αSer151, αThr190, and αTyr103.

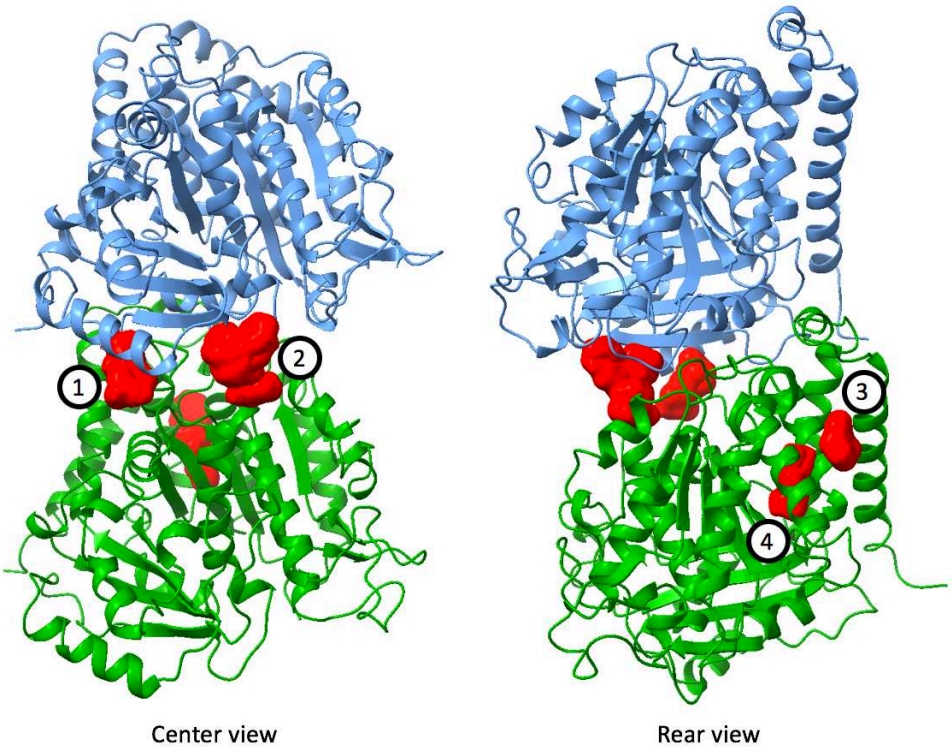


Figure 135. Location of the cryptic pockets identified by FTMap

7.7. Protein pocket dynamics analysis

To explore the dynamic properties of protein pockets within our molecular dynamics (MD) simulation trajectory, we utilized the D3Pockets web server[163]. This tool allows for the examination of pocket stability, continuity, and correlation, calculated using a set of trajectory points from all MD simulations. The process involves three fundamental steps: first, potential pockets in the various conformations within the trajectory are identified. Then, a grid of points is established for each detected pocket in each conformation. Lastly, the dynamic properties of the pocket are calculated by analyzing the frequency of a particular grid point's appearance throughout the MD trajectory.

The time-resolved stability of each pocket (P_S) is defined by the ratio of the number of conformations that include the i-th grid point (n) to the total number of conformations in the trajectory (N) for all points of the grid that make up this pocket (S_i), as shown by equation 12[161,163]. If m is the number of grid points in the pocket, the stability of the pocket (P_S) is defined as an array containing the S_i ratios of all the grid points defining the pocket (equation 13). D3Pockets color-codes the grid points that make up a pocket, with red points indicating the most frequently

occurring points during the MD, and blue points indicating the least frequent. Consequently, the red regions of a pocket are more stable than other regions.

$$S_i = \frac{n}{N} \tag{12}$$

$$P_S = \{S_1, S_2, \dots, S_i, \dots, S_m\} \tag{13}$$

Pocket correlation is determined by first clustering all potential binding pockets that appear in an MD trajectory, based on residues[161,163]. This generates sets of protein conformations corresponding to each cluster (C_i). The volume of each conformation belonging to cluster *i* (V_i) is then calculated. Finally, the coexistence and correlation matrices are calculated using Equations 14 and 15.

$$C_{i,j} = C_i \cap C_j \tag{14}$$

$$\rho_{i,j} = \frac{cov(V_i, V_j)}{\sigma_{V_i}\sigma_{V_j}} \tag{15}$$

In these equations, C_i and C_j represent the conformation sets of the protein corresponding to the *i*-th and *j*-th cluster pockets, respectively. V_i and V_j denote the volume sets of the *i*-th and *j*-th cluster pockets in the corresponding conformations. The term $cov(V_i, V_j)$ refers to the covariance of V_i and V_j, while σ represents the variance of V.

The correlation coefficient derived from these calculations provides insight into the relationship between the pockets. A positive correlation coefficient, which can reach a maximum of +1, indicates a positive correlation between pockets. This means that as the volume of pocket *i* increases, the volume of pocket *j* also increases. Conversely, a negative correlation coefficient, which can reach a minimum of -1, signifies a negative correlation between pockets. This implies that as the volume of pocket *i* increases, the volume of pocket *j* decreases. This correlation analysis, therefore, provides valuable insights into the interplay between different pockets during the dynamics of the protein.

We used the D3Pockets software to analyze the dynamics of protein pockets in our system. In this analysis, we focused solely on the four unique binding pockets identified by FTMap that were not identified in either X-ray screen or classical MD simulations reported by Muhlethaler et al.[106]

Upon comparing the stability of the predicted pockets in the tubulin system (Figure 136), it is evident that new pockets 1 and 3 are part of a larger pocket identified on the α,β-tubulin interface and remain stable throughout the MD simulation trajectory. New pocket 3 also exhibits stability throughout the trajectory, while new pocket 4 is not stable on the scale of the whole trajectory, and was not detected by D3Pockets.

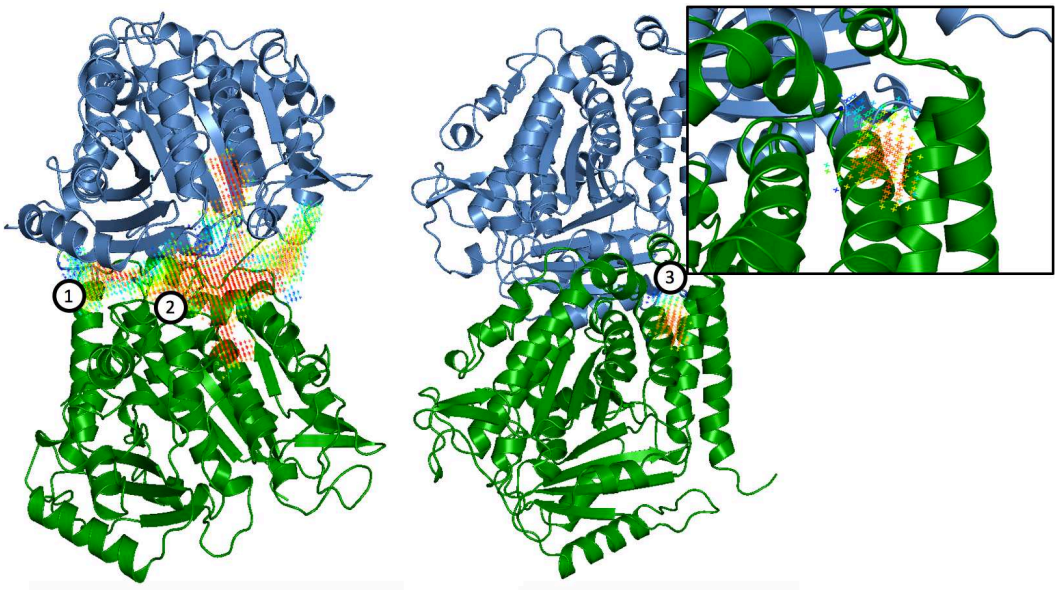


Figure 136. Analysis of protein pocket stability by D3Pockets

D3Pockets was also used to calculate a correlation between cryptic pockets appearing during the MD trajectory. D3Pockets could not distinguish new pockets 1 and 2 separately, instead, it perceived them as part of the large pocket on the α,β-tubulin interface. New pocket 4 was not detected due to, apparently, being rarely present in the trajectory in the open form. Therefore, we conducted this analysis solely for new pocket 3 (Figure 137). The results indicate that the size of new pocket 3 during tubulin dynamics has a direct correlation with the size of the taxane binding site (positive correlation coefficient of 0.81).

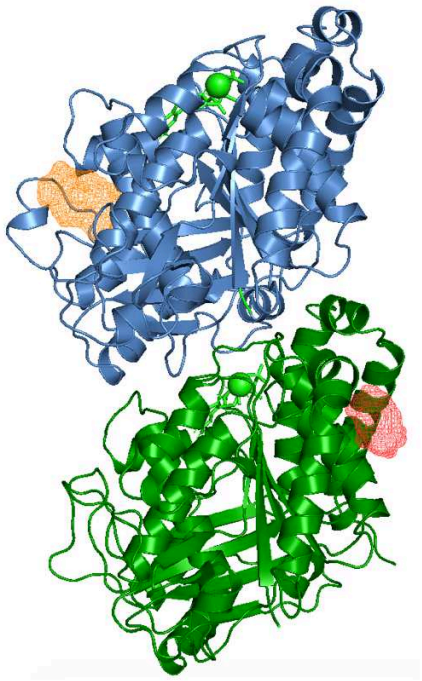


Figure 137. Pocket correlation analysis highlighted positive correlation between volumes of pocket 4 (red) and taxane binding site (orange)

7.8. Conclusion

In this study, we performed Gaussian-accelerated molecular dynamics simulations to explore the dynamics of the α,β-tubulin heterodimer in an aqueous environment. Our analysis of the simulation trajectory revealed that tubulin is a remarkably stable protein throughout the simulation period.

We utilized the FTMap computational server to identify potential binding pockets on the surface of tubulin. Our primary interest was in discovering pockets that have not been previously known to host ligands or small fragments and have not been highlighted in prior computational studies. This approach led to the identification of four such pockets.

Subsequently, we employed the D3Pockets analysis tool to evaluate the stability of these binding pockets over the course of the simulation. Our findings indicated that while one of the pockets was unstable, the remaining three were stable. Interestingly, two of these three stable pockets form part of a larger protein-protein interface. The third pocket is a novel finding and was demonstrated to have a direct correlation in size with the taxane binding site.

These newly identified pockets may present promising targets for compound screenings and docking studies. However, further investigation is required to fully understand these binding sites. Importantly, our findings underscore the value of integrating Gaussian-accelerated MD with classical MD and experimental work to gain a more comprehensive understanding of the system under study. This synergistic approach can provide deeper insights and open up new avenues for exploration in the field of computational chemistry and molecular dynamics simulations. Pocket drugability needs to be assessed for all identified pockets.

Looking forward, we aim to further explore these novel pockets, particularly focusing on their potential as drug targets. We plan to conduct docking studies with a range of compounds to assess their binding affinities. Additionally, we intend to refine our simulation parameters to better capture the dynamic behavior of these pockets. Ultimately, we hope that our work will contribute to the development of new therapeutic strategies targeting tubulin, and underscore the importance of integrating different computational approaches to fully understand complex biological systems.

Chapter 8. Development of a graphical application for automatic analysis of microtubule fiber diffraction pattern images

8.1. Introduction

Interaction of small organic molecules with tubulin can significantly impact the physical characteristics of microtubules, including the microtubule diameter, the number of constituent protofilaments, distance between the protofilaments in a microtubule, and the average length of the tubulin monomers in the tubular structure[164]. Experimental study of such changes is important to elucidate mechanisms of action of microtubule-targeting agents.

One approach to quantitatively study these changes involves shear flow X-ray microtubule fiber diffraction assay[164,165]. Fiber diffraction, a widely used technique in structural biology, is utilized to explore the structure of biological filaments, specifically microtubules, under physiological conditions without the necessity for fixation, crystallization, or freezing. The samples under investigation are often arranged, either naturally or artificially, in a line of filamentous structures exhibiting a degree of regularity, periodicity, or helical pattern. Fiber diffraction offers a more comprehensive structural understanding than alternative X-ray diffraction-based techniques, providing detailed information on the longitudinal periodicity and lateral spacing of molecules within an arranged filament[165].

The initial step of a microtubule fiber diffraction experiment is the alignment of the filaments in solution. Several methods exist for achieving this alignment, one of which is the technique of shear flow. This technique leverages the fluid-dynamic properties of the microtubule filaments in a medium stream with a certain gradient of flow velocity to a given shear. This method is not only cost-effective in terms of the materials required but also fast, enabling real-time experiments. It also allows for the simultaneous study of both physical and chemical parameters in real time as the shearing tool continuously mixes the specimen during ongoing data acquisition of diffraction[165,166].

In a typical microtubule fiber diffraction experiment, X-ray fiber diffraction images are captured in synchrotron radiation beamlines. The diffracted X-rays are collected by a detector, yielding a single diffraction image per beam exposure. Typically, 16-24 diffraction images are collected from 4-6 independent samples for a given experiment (Figure 138). Additionally, background images are obtained under the same conditions using a buffer solution. The final step is the spatial calibration, performed using Ag-Behenate powder diffraction, which considers an elastic scattering and provides the distances to the beam center of the scattering vectors' diffraction intensities[165,167].

Thus, the results of a typical microtubule fiber diffraction experiment include a file with the parameters of the detector, a calibration file of the Ag-Behenate powder diffraction experiment, the images of the buffer solution, and most importantly, the X-ray diffraction patterns obtained from aligned microtubules in the presence of tested microtubule-targeting agents. The first step in the typical analysis of the experimental results is the subtraction of buffer images to produce averaged experimental diffraction images[167].

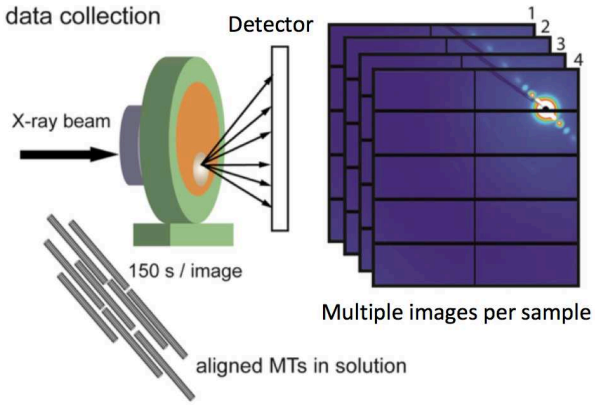


Figure 138. Experimental setup of a microtubule fiber diffraction assay

A typical X-ray microtubule diffraction pattern is shown in Figure 139. It has two regions of particular importance for the determination of microtubule structural parameters: the equatorial profile and the meridional profile. By integrating the diffraction intensities of the equatorial profile, it is possible to determine the average lateral microtubule structural parameters (radius, inter-protofilament distance, and protofilament number). The meridional intensity profile is used for the average axial monomer length determination[167,168].

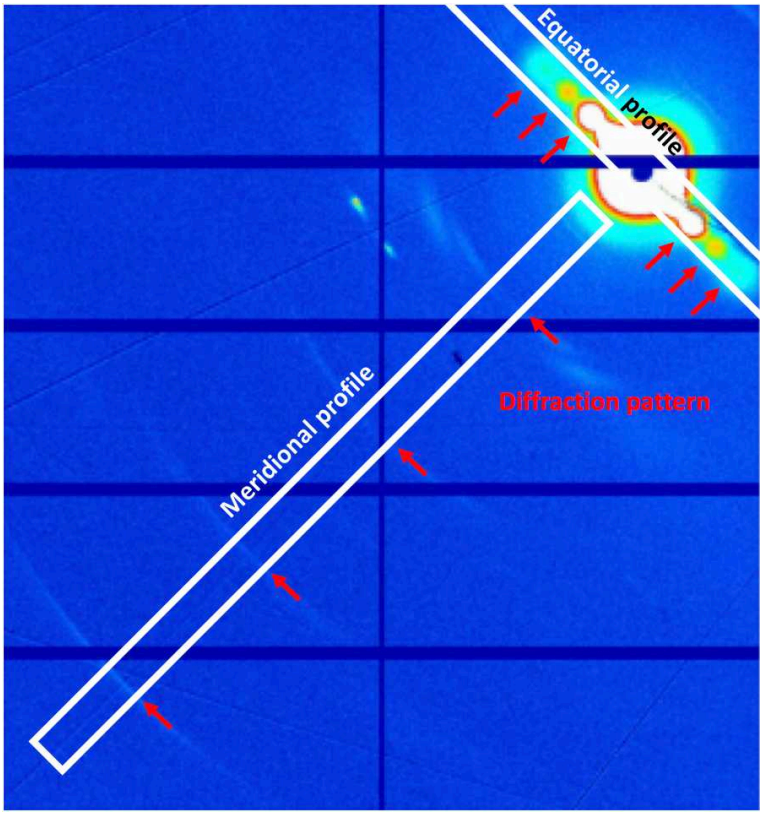


Figure 139. An example of a microtubule X-ray diffraction pattern

The equatorial profile is a scattering pattern that mirrors what is typically seen in molecules with a random orientation in solution. For microtubules, the primary scattering can be characterized by a zero-order Bessel function, which primarily consists of four peaks that are related to the microtubule diameter. The initial strong peak of scattering is masked and overlaid with background beam noise originating from the materials of the shear-flow device. However, the second and third peaks serve as useful tools for estimating the average diameter of the MTs. Peaks that occur at scattering angles beyond the fourth peak carry additional signals that reflect the number of protofilaments in MTs, which can be utilized to estimate the distribution of MTs with varying numbers of protofilaments[165,168].

The meridional signal profile aligns with a Fourier transform of the periodic organization of tubulin molecules. This is characterized by a 4 nm (x) peak, equivalent to the size of the tubulin monomer, and additional harmonic patterns that manifest as $N \times 1/x$, where N is an integer denoting the Nth-order diffraction. The variance in the intensity of each peak primarily depends on the helical configuration of tubulin dimers. The fourth-order signal (approximately 1 nm) typically exhibits the highest intensity on the meridional axis, and thus is employed to estimate the length of the tubulin molecules. During the analysis of the meridional profile, the meridional intensity profile of the 1 nm layer line peak is fitted to a Lorentzian distribution to accurately locate the intensity maximum. The intensities (height) of the signals offer insights into the density of the

structural regularity and the population of aligned molecules (Figure 140). This can be leveraged to monitor microtubule assembly and disassembly. Therefore, given the semi-crystalline structure of MTs, with tubulin units stacked in a regular pattern, X-ray diffraction serves as a potent tool for detecting minor yet significant structural changes in tubulin following the binding of specific ligands[164,165].

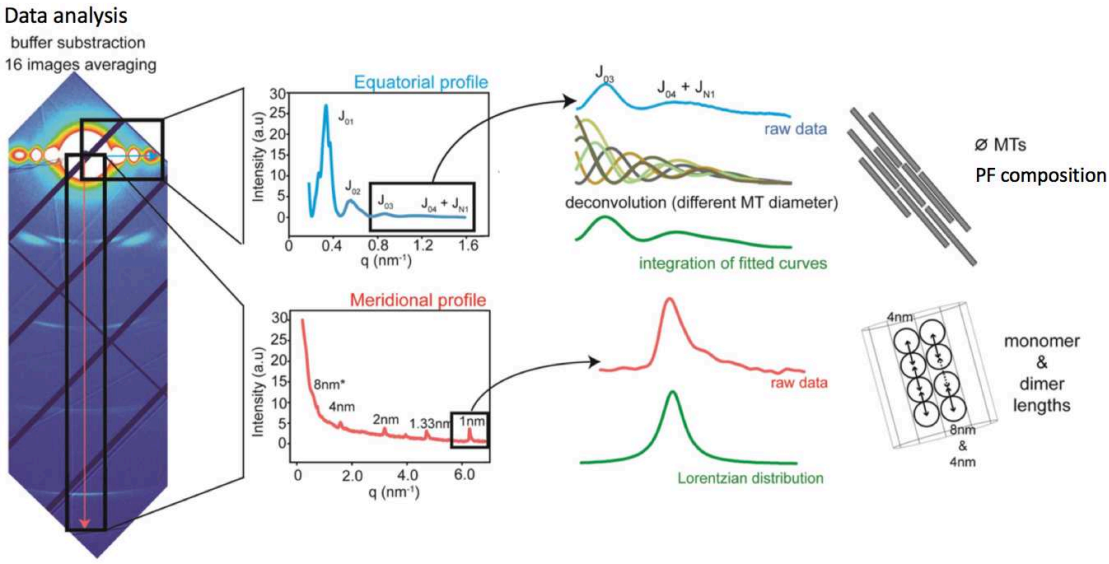


Figure 140. Schematic summary of the analysis of diffraction patterns and information that can be inferred from it. Adapted from [167].

The analysis of the images produced by shear flow X-ray microtubule fiber diffraction assay experiments involves image processing, numerical integration of visual data, and functional approximation of the integration results. Typically, this analysis has been manual and labor-intensive, necessitating the use of multiple specialized image processing and statistical data processing applications. The necessity of a time-consuming analysis limited the number of experiments researchers could perform during very limited and expensive experimental time they have on the synchrotron facilities. Therefore, the aim of this project was to develop a program for automated analysis of microtubule fiber X-ray diffraction patterns with a user-friendly graphical interface that would increase the speed of the analysis and, as such, allow for more experiments to be done quicker, increasing the throughput of shear flow X-ray microtubule fiber diffraction experiments. This work was performed in fruitful collaboration with Óscar Fernandez Blanco under the guidance of Dr. J. F. Díaz during my industrial secondment as part of the TubInTrain consortium curriculum at the AnkarPharma company, part of the Spanish National Research Council (CSIC), Madrid, Spain.

8.2. Overview of the developed software

Our efforts resulted in the creation of FiDAT (Fiber Diffraction Analysis for microTubules), a standalone graphical application that streamlines the three steps of experimental data analysis. It features a graphical interface, enhancing the throughput and adaptability of microtubule diffraction experiments. Developed using the Python 3.9 programming language, FiDAT's graphical interface is built with the Tkinter Python library, enhanced by the CustomTkInter Python package. Its image plotting functionality is based on matplotlib, while numerical integration and curve fitting were implemented using the pyFAI and scipy Python packages. Due to this, the application is fully cross-platform. It does not require any experience with programming to be installed or operated. It also doesn't require previous experience with analyzing fiber diffraction experiments results, since most of the steps have been completely automated. At the same time, advanced users can refine the results by changing the fully customizable settings.

8.3. Input images preprocessing

To start the analysis of experimental results with FiDAT, the user first needs to select a directory on their local machine where the results of the analysis will be stored (Figure 141). If a user wants to store the results of multiple analyses in the same directory, there is an option to provide a unique prefix that would be used to identify the files related to a particular analysis.

Next, the user can define the values of a normalization factor and buffer intensity. Both values are used to adjust the range of pixel intensity values in sample and buffer images, respectively. This is useful in cases where an image has poor contrast. A default value of 1.0 is usually enough for most applications.

Then, the user is required to provide a list of buffer images. These experimental images do not contain any diffraction pattern. They are used to subtract the buffer and background noise from the sample images in a subsequent step to enhance the signal-to-noise ratio. When loading a list of images, the user has an option to preview the images and choose which they would like to use for the analysis, and which they would like to discard (Figure 142). The same process is repeated for the sample images.

Finally, the user is asked to provide a point of normal incidence file, which contains a 6-parameter geometry definition from the synchrotron detector. This file is used to define the position of the beam relative to the image.

Once all files are loaded, the user can pre-process the images by pressing the "Prepare data" button (Figure 141). With this, the mean intensity value for each pixel across a given list of sample and buffer images is computed, thereby generating an averaged image representation that will be used for subsequent analysis.

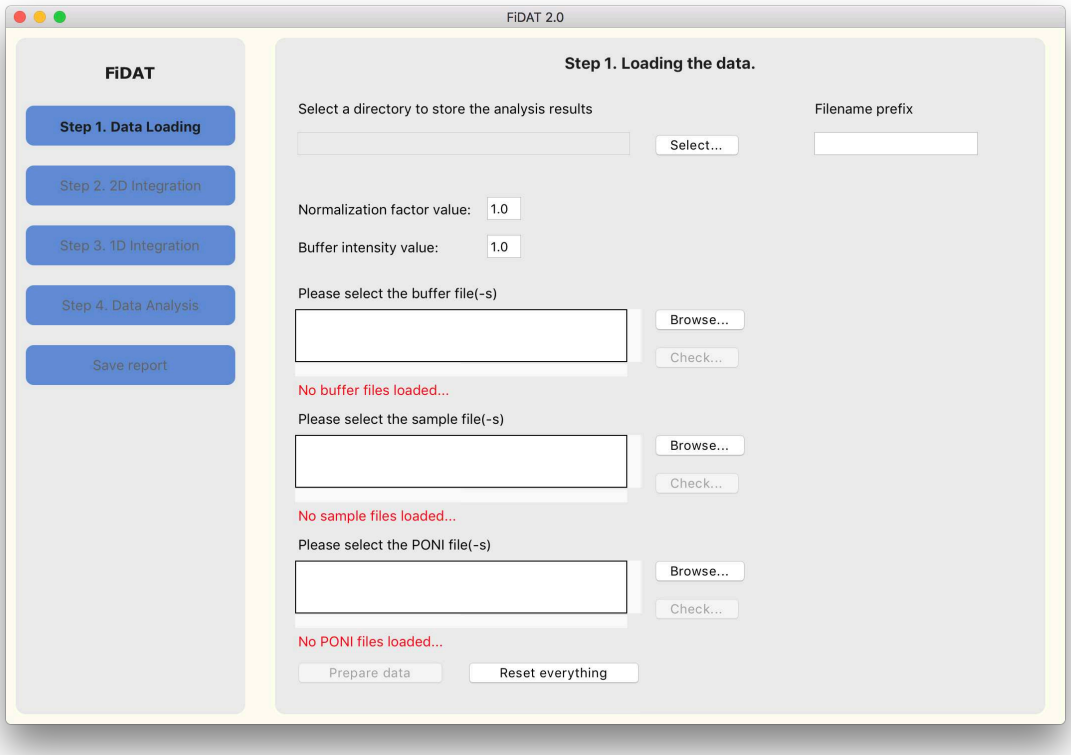


Figure 141. First screen of the FiDAT application

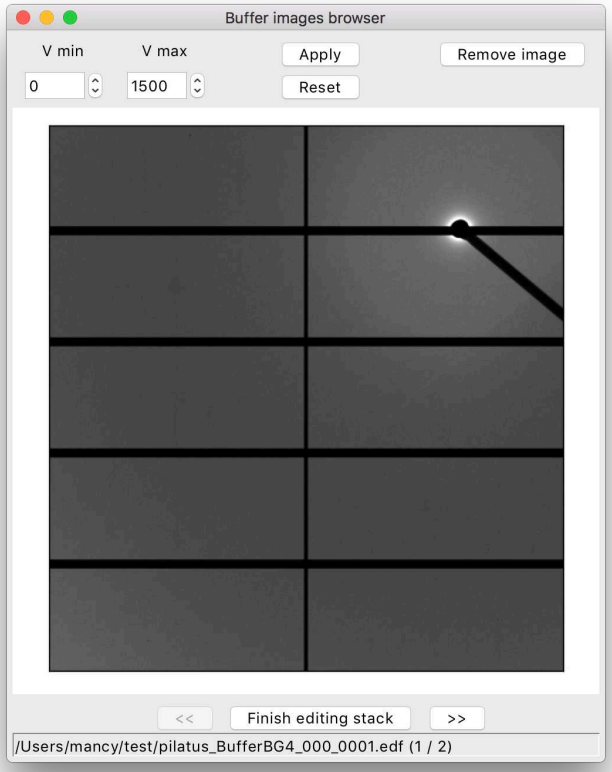


Figure 142. The user can pre-view selected buffer and sample images

8.4. Two-dimensional integration

Once an average image of the sample is produced, the user can perform a two-dimensional integration (also known as azimuthal regrouping) of this image. This is done to convert the 2D scattering pattern into a 2D image where the x-axis is the scattering angle, and the y-axis is the azimuthal angle. Each pixel in this image represents the integrated intensity over a small range of scattering angles and azimuthal angles. This is useful for visualizing the symmetry and orientation of the scattering pattern. The user can configure the parameters of the integration (Figure 143), or use default parameter values. In any case, after pressing the “Run 2D integration” button, the user obtains the required regrouped plot (Figure 144). The “Minimum” and “Maximum” input fields are used to control the contrast of the resulting image.

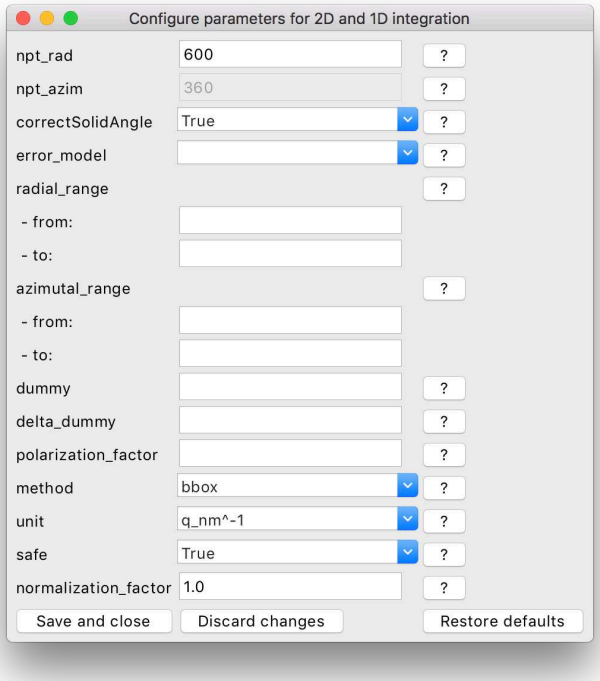


Figure 143. The user can configure the parameters of azimuthal regrouping

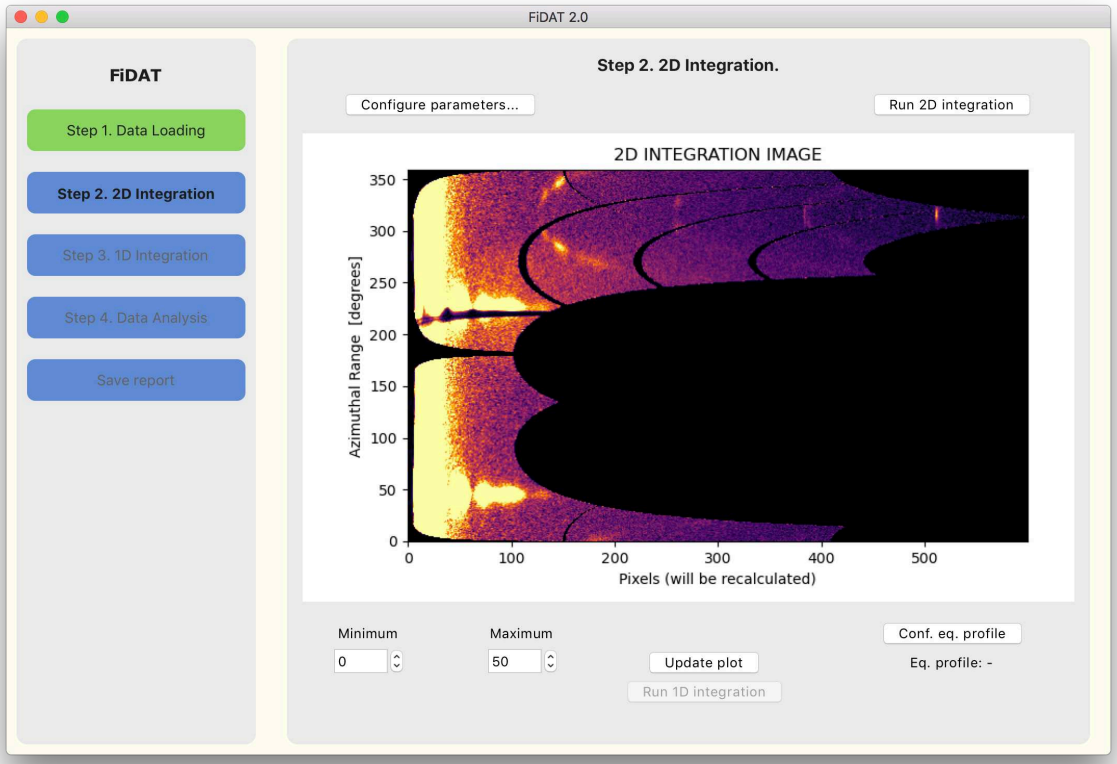


Figure 144. Result of the azimuthal regrouping

8.5. One-dimensional integration

The regrouped image is used in the second step, where the user selects an azimuthal angle range for one-dimensional integration using an interactive slider or an input field (Figure 145). This is done to transform the 2D diffraction pattern into a 1D plot of intensity versus scattering angle. This is achieved by averaging the intensity of pixels that are at the same distance from the center of the diffraction pattern, thus regrouping them into bins according to their radial distance, or azimuth, to simplify the analysis and interpretation of the diffraction data. The 2D diffraction pattern contains a wealth of information about the structure of the sample, but it can be challenging to interpret due to its complexity. By transforming it into a 1D plot, one can more easily identify and analyze the peaks of intensity, which correspond to specific structural features of the sample. The selection of the azimuthal angle that will be used for 1D integration thus defines the axes of the equatorial and meridional profiles (Figure 146). There are two possible equatorial profiles, EP1 and EP2, referring to the left and right halves of the equatorial profile with respect to the beam position (see Figure 139 for reference).

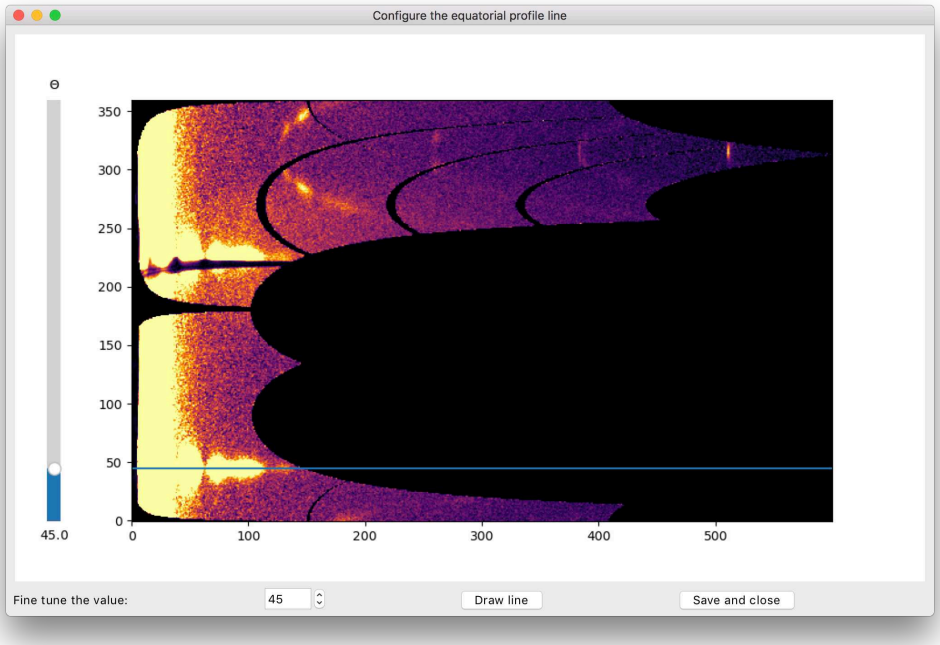


Figure 145. Selection of the azimuthal angle for 1D integration

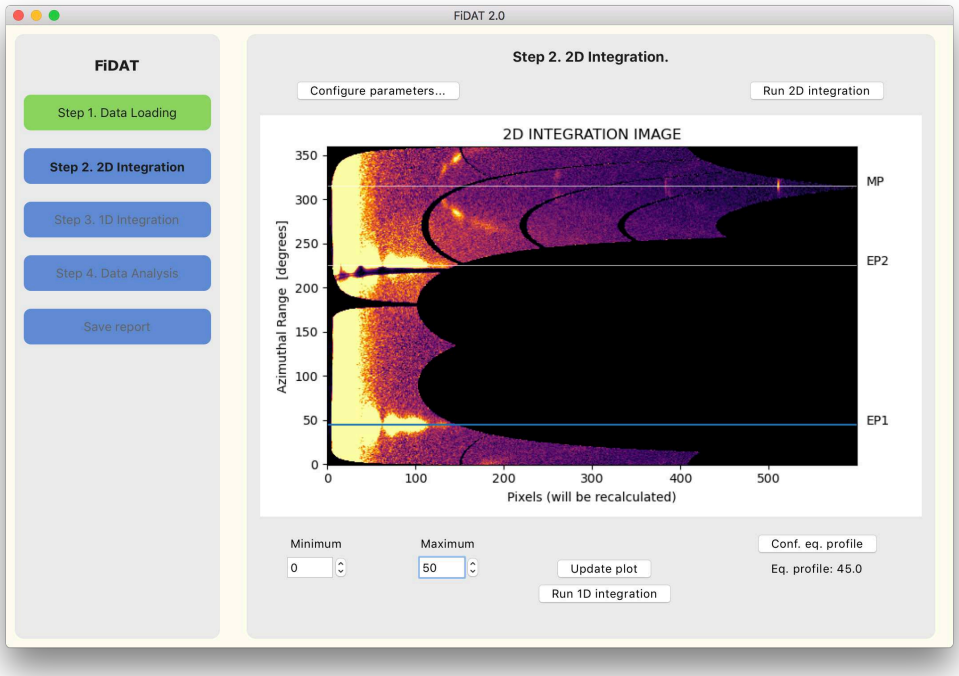


Figure 146. Selected azimuthal angle influences the selection of the axes for equatorial (EP1, EP2) and meridional (MP) profiles for subsequent analysis

Then, one dimensional integration is performed along the EP1, EP2, and MP axes specified on the previous step. What results from this are simple plots that are easy to analyze visually (Figure 147). It is useful to see whether the integration parameters used on the two previous steps were appropriate. However, to extract the structural parameters of microtubules from these data, a function fitting procedure is required.

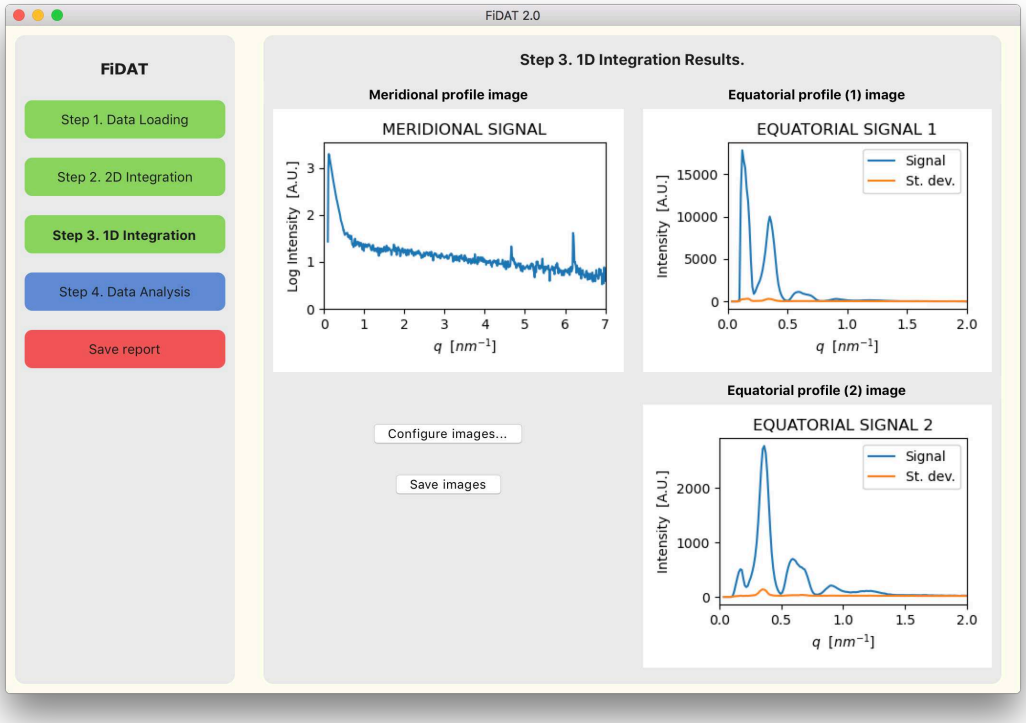


Figure 147. One-dimensional plots useful for the analysis of the integration results

The analysis of the meridional profile involves fitting a Lorentzian function to the meridional profile data (Figure 148). This is done to extract the value of the single tubulin monomer subunit length in Angstroms (see Figure 140). A Lorentzian function is defined by equation 16, where x is the input data, y_0 is the base level, a is the amplitude, x_0 is the center, and b is the width of the Lorentzian peak.

$$f(x) = y_0 + \frac{a}{1 + \left(\frac{x - x_0}{b}\right)^b} \qquad (16)$$

The curve fitting is done using the Levenberg-Marquardt algorithm implemented in the scipy library. The function finds the optimal parameters that minimize the residual sum of squares between the target outputs and the outputs predicted by the Lorentzian function. The R^2 coefficient of determination metric is used as a quality of fit. Finally, after the best parameters have been

found, the average tubulin monomer length is calculated using equation 17, where x_0 is the center of the Lorentzian peak. The factor of 4 comes from the fact that the scattering vector is 4 times the reciprocal of the monomer length[167].

$$AvgMonLength = \frac{2\pi}{x_0} \times 4 \qquad (17)$$

Figure 148. Results of the Lorentzian function fitting of a meridional profile plot data

The analysis of the equatorial profile involves fitting a Bessel function to the equatorial profile data (Figure 149). This is done to extract the value of luminal radius of the microtubule and investigate the protofilament composition of the microtubule (see Figure 140). The microtubule diffraction pattern comprises of several layer lines (l) each defined by a group of Bessel functions of order n. Their structural factor $F_{l,n}$ in the reciprocal space (R) is described by equation 18, where J_n is the Bessel function of the n-th order, r_m is the radius of a microtubule made of m protofilaments, and $f(R)$ is the structural factor defined by equation 19, where r_t is the radius of the tubulin monomer considered as a sphere, with a value of 2.48 nm. This expression is used to include the structural factor of the tubulin wall in the calculation.

$$F_{l,n}(R) = J_n(2\pi r_m R) f(R) \qquad (18)$$

$$f(R) = 4\pi r_t^3 \frac{\sin(2\pi r_m R)\cos(2\pi r_m R)}{(2\pi r_m R)^3} \qquad (19)$$

The Bessel functions are then weighted to model the scattering intensity from microtubules with different numbers of protofilaments. The weights (PN10, PN11, PN12, PN13, PN14, PN15) represent the relative proportions of microtubules with 10, 11, 12, 13, 14, and 15 protofilaments, respectively. By adjusting these weights, the model can better fit the experimental data. The weighting ensures that the model accurately reflects the physical properties of the system and can be defined as equation 20, where PN_i is the absolute ratio of *i* number of protofilaments.

$$f(x) = \sum_{i=10}^{15} \frac{PN_i}{\sum_{j=10}^{15} PN_j} \cdot f(x') \tag{20}$$

Consequently, $f(x')$ is defined by equation 21, where *Amp* is the amplitude of the function, *An* is the numerical aperture, r_m is the microtubule radius, and *n* is the order of the Bessel function.

$$f(x') = Amp(J_0(x \cdot r_m) \cdot F(U))^2 + An(J_n(x \cdot r_m) \cdot F(U))^2 \tag{21}$$

The curve fitting is done using the Trust Region Reflective algorithm implemented in the scipy package. The function finds the optimal parameters that minimize the residual sum of squares between the observed outputs in the dataset, and the outputs predicted by the mixture of weighted Bessel functions.

Finally, the interprotofilament distance (IPFD) is calculated as defined by equation 22, where *AvgN* is the average number of protofilaments and r_m is the microtubule radius.

$$IPFD = 2\sin\left(\frac{\pi}{AvgN}\right) \cdot r_m \tag{22}$$

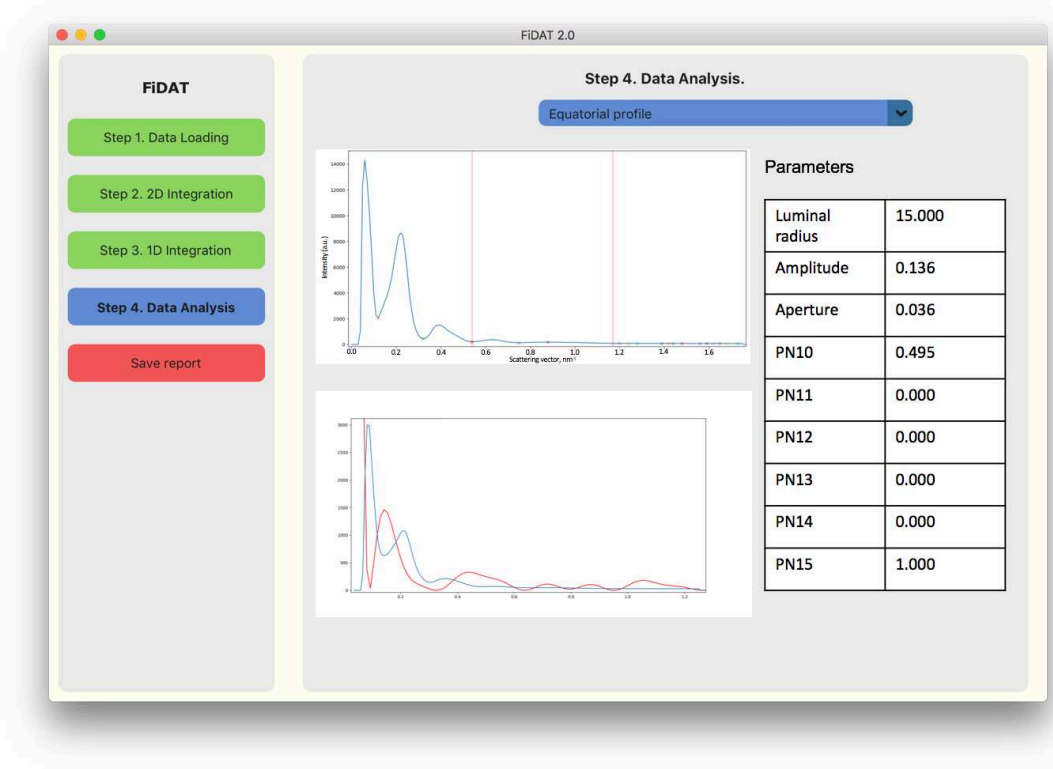


Figure 149. Results of a Bessel function fitting to the equatorial profile data

Having derived all the required microtubule parameters by fitting both Lorentzian and Bessel functions to the data obtained from one-dimensional integration of the azimuthal regrouped plot, the user can save a report as a Microsoft Office Word document with all the plots and parameters produced during data processing and function fitting. Additionally, raw data is saved for each plot, should a user decide to build the plots in other software.

8.7. Conclusion and perspectives

In conclusion, we have successfully developed FiDAT, a tool that facilitates the process of analysing the results of microtubule fiber diffraction experiments. This tool is designed to be user-friendly, requires no prior expertise, and significantly accelerates the throughput of these experiments. FiDAT is equipped to handle all stages of fiber diffraction experiment results analysis, making it a comprehensive solution for researchers in this field.

The software comes with pre-configured default parameters, ensuring reliable data processing and model fitting right out of the box. However, we have also catered to the needs of experienced users by providing the flexibility to customize the parameters for each step of the analysis.

The effectiveness of FiDAT has been validated by our colleagues from the TubInTrain consortium, led by Dr. J.F. Díaz at CSIC, Madrid, Spain. They successfully applied FiDAT to samples obtained during the research time at the Alba synchrotron in Barcelona, Spain, in June

2022. The efficiency of FiDAT's automated analysis process not only facilitated the execution of additional experiments but also empowered researchers to swiftly incorporate emerging ideas into their experimental design. This real-time adaptability has enabled on-the-spot decision-making and immediate verification of hypotheses, eliminating the need for time-consuming and labor-intensive manual analysis.

Looking ahead, we plan to further enhance FiDAT by resolving minor bugs related to image plotting and incorporating additional logging features, as suggested by our colleagues. These improvements aim to further streamline data reporting and enhance the user experience.

In our commitment to fostering open science, we also plan to release FiDAT as an open-source software, thereby making this powerful tool accessible to the wider scientific community.

Furthermore, the management of the Alba synchrotron in Barcelona, Spain, has shown considerable interest in integrating FiDAT into their mainframe. Once the planned enhancements are implemented, FiDAT will be made available as a default tool for all researchers at the synchrotron, further solidifying its role as an indispensable asset in microtubule fiber diffraction experiments.

General conclusion and perspectives

The present Ph.D. thesis titled "Computer-aided design of tubulin polymerization modulators" employed computational chemistry techniques to discover novel small molecule agents targeting underexplored binding sites on tubulin. By targeting these binding sites, we aspired to develop novel modulators of tubulin polymerization with diverse structure and mode of action, what has implications for cancer- and neurodegeneration-related research. We used virtual screening methodologies applied to a selection of drug-like, commercially available, and in-house developed chemical libraries. Through substructure and similarity search, pharmacophore screening, and different protein-ligand docking strategies, we were able to successfully identify small molecules of diverse chemical structure that bind to several sites on the tubulin protein. Additionally, we employed novel deep learning techniques to develop a computational pipeline for *de novo* design of small molecule tubulin polymerization modulators. Given the large size of the studied protein system, we performed accelerated molecular dynamics simulations to uncover potential cryptic binding pockets on the surface of the tubulin protein. Finally, we developed a software with a graphical user interface that facilitates the analysis of microtubule fiber diffraction experiments results.

In chapter 2 of this work, our goal was to discover novel microtubule-targeting agents that bind to the maytansine binding site of the tubulin protein. The motivation behind this was the complexity and high-cost associated with the synthesis of currently known macrocyclic binders for the maytansine site. To achieve this, we performed virtual screening of two resources: the ChEMBL database, which consists of drug-like compounds with known bioactivity properties, and a library of commercially available compounds provided by the Enamine company. This process led to the identification of six potential hits in the ChEMBL database, which were listed as cytotoxic compounds, although the origin of their cytotoxic effect remained unknown. We proposed a hypothesis that this cytotoxicity might be due to binding at the maytansine site, given the observed fit of these molecules with the pharmacophore model of a crystallographically confirmed binder. Notably, one of these hits is a natural product – a class of compounds with a rich history of acting as tubulin-targeting agents – thereby justifying its further synthesis and analysis. The screening of the commercially available compounds yielded a set of chemically diverse candidates that could potentially bind to the maytansine site. These were later subjected to experimental validation. Of the 11 molecules proposed, 2 were found to inhibit microtubule polymerization, as confirmed by the microtubule polymerization bioassay. These findings lay a robust groundwork for future development and computer-aided structural optimization of the identified scaffolds, driving the research toward potential maytansine site binders.

In chapter 3 of this work, our goal was to discover novel microtubule-targeting agents that bind to the pironetin binding site of the tubulin protein. The motivation behind this was the complexity and high-cost associated with the synthesis of the single currently known ligand that binds to the prionetin binding site. To this end, we performed virtual screening of several libraries of commercially available compounds provided by the Enamine company. We also implemented a machine learning-driven protein-ligand docking protocol, which allowed us to screen the largest Enamine library with protein-ligand docking. Combined, the screening of different libraries led to the identification of 47 virtual hits, which were subjected to experimental validation. Three small fragments were found to have significant inhibitory effect on microtubule polymerization, as confirmed by the microtubule polymerization bioassay. Additionally, two ligands were found to bind not at the pironetin, but at the colchicine binding site. One of the compounds had significant inhibitory effect on microtubule polymerization, with some experiments highlighting its specificity towards a βIII-tubulin isotype, especially relevant in cancer research. Thus, future work would see the further optimization of this hit molecule's structure to ensure even higher isotype specificity, as it would have a large impact on cancer research and could be developed into a therapeutic agent. The exploration of the pironetin binding site should continue as well by accounting for high flexibility of the binding site.

In chapter 4 of the present thesis, our goal was to develop novel molecules that target the recently discovered todalam binding site of the tubulin protein. We were motivated to explore this binding site because it was targeted by structurally simpler, rationally designed ligands, and contained a cysteine residue in proximity to the ligand binding pocket, which could be used to potentially design covalent ligands targeting this binding site, which has implications in molecular probe design. The virtual screening effort with subsequent experimental validation has been successful in identifying a novel set of chemically diverse compounds that bind to the todalam site on the tubulin protein. Out of 18 proposed molecules, 5 were confirmed to bind to the todalam site by X-ray crystallography. Interestingly, 3 other molecules were found to exert substantial inhibitory effects on microtubule polymerization, despite not being detected in the binding site. These results led to the discovery of five unique scaffolds capable of targeting the todalam site. These scaffolds were then developed into covalent binders, employing computer-aided design strategies. Our research efforts led to the design of 30 potential covalent binders for the todalam site. Of these, 8 were confirmed to bind to the site through X-ray crystallography experiments, albeit without forming a covalent bond with the targeted cysteine residue. Additionally, we observed that 3 molecules caused an inhibitory effect on microtubule polymerization, even though they were not detected in the binding site. The todalam site binders that we have discovered represent a significant advancement in our understanding of tubulin’s biochemistry and set the

stage for the development of more potent and selective binders. This breakthrough opens up exciting avenues for future research, with potential implications for drug discovery and the treatment of diseases related to microtubule function.

In chapter 5 of the present work, we aimed to apply the inverse QSAR methodology to the task of *de novo* design of novel, structurally diverse small molecules targeting the colchicine binding site. Many inhibitors targeting the colchicine site have been developed from representative and commonly used scaffolds, limiting structural innovation and constraining exploration of the chemical space. We aspired that *de novo* drug design using inverse QSAR could help circumvent these issues. To this end, we have first trained and validated an attention-based conditional variational autoencoder neural network model using unlabeled molecular data from the ChEMBL database. This model made it possible to sample SMILES strings of compounds corresponding to a user-defined point in a latent descriptor space. Next, we trained and validated a QSAR model on the HeLa cells cytotoxicity data collected for small molecules designed to target the colchicine binding site. Using this model, we screened a library of structurally diverse commercially available compounds from the Enamine company. Molecular descriptors of purchasable molecules with the highest predicted cytotoxic action were used as seed vectors for *de novo* generation using the trained autoencoder model. This was done to ensure that the chosen seed points actually correspond to chemical structures with pharmacology-compliant physicochemical properties. The generated molecules were then docked into the colchicine binding site, and top-20 molecules were selected by docking score. From these molecules, none were readily-available for purchase. However, for 4 of them, close structural analogues were found in catalogues of purchasable compounds. Additional cytotoxicity prediction and protein-ligand docking of these molecules have confirmed their high potential activity. These molecules were purchased and evaluated by X-ray crystallography. Unfortunately, none of the compounds were detected in the colchicine binding site. *In vitro* studies of these molecules' effect on microtubule polymerization are currently being performed. Despite the experimental validation of the generated compounds has presented some challenges, the comprehensive approach used in this study shows great promise for future research in drug discovery. This work exemplifies the potential of combining inverse QSAR modelling and experimental approaches to accelerate the discovery of novel therapeutic compounds. Future work would include different strategies of selecting the seed vectors for compound generation.

Chapter 6 of the present thesis is dedicated to exploring the utility of molecular representations learned in an unsupervised way in QSAR modeling. Specifically, we investigated the concept of transfer learning that has seen broad applications in the fields of computer vision and natural language processing, and sought to apply it to molecular structures. We conducted a comparative analysis of the predictive performance between a state-of-the-art support vector

machine model with evolutionarily optimized hyperparameters and descriptor sets, and a multi-layer perceptron model using learned molecular representations. The molecular representations were derived from a state-of-the-art graph neural network. Further, we compared the performance of evolutionarily optimized support vector machine models trained on extensively engineered descriptors, unsupervisedly learned molecular representations, and molecular representations fine-tuned for a specific downstream task. The results illustrated that unsupervisedly learned molecular representations yielded models with comparable performance to those trained on tailored descriptors. Fine-tuning these representations slightly improved the predictive performance, although the extent of this increase may be related to the specific downstream task. Thus, our findings highlight that molecular representations derived from large data corpus are highly effective as is. Future work will involve comparing different methods for learning these representations. In particular, we are interested in using a SMILES-based state-of-the-art neural network for learning these representations.

Chapter 7 of the present work is concerned with characterization and analysis of conformational dynamics of the tubulin protein in solution that may lead to the emergence of cryptic binding pockets. To efficiently sample tubulin's conformational dynamics, we employed the Gaussian-accelerated molecular dynamics simulation method. This method consists in adding a harmonic boost potential to smoothen the potential energy surface of the modelled system and decrease the energy barriers to accelerate the transitions between different low-energy states. The setup involved simulating an α,β-tubulin heterodimer in water. The simulation was separately performed three times starting from the same initial coordinates of the modelled system, exploring distinct sections of the conformational space. The resulting concatenated trajectory was then subjected to clustering analysis to include all the accessible states. The clustering has identified 15 distinct conformations of the α,β-tubulin heterodimer. The distinct conformations were then subjected to hotspot analysis by distributing small organic probes over the protein surface. The identified pockets were compared to the results of a previously published comprehensive analysis of the binding pockets of the tubulin protein, to highlight only those that have not been detected previously. As a result, we were able to discover four cryptic binding pockets. Additional analysis of the pocket dynamics has shown that from these four, only one is a distinct pocket that is stable during the simulation time and has a correlation with the taxane binding site. Future directions of work thus include a more detailed investigation of the size, shape, and electrostatic properties of the found pocket. Another promising direction involves conducting virtual screening campaigns to identify potential ligands for the newly discovered pocket. The selected hits can then be experimentally validated, potentially leading to the discovery of novel tubulin-targeting agents. Additionally, the application of machine learning techniques could prove beneficial in predicting

other cryptic pockets in the tubulin protein or in other related proteins. Given the success of machine learning in other areas of study, its application in the detection and analysis of cryptic pockets is certainly an exciting prospect.

Finally, chapter 8 of this thesis describes the development of FiDAT, a comprehensive and user-friendly software tool designed to streamline and expedite the process of analyzing the results of microtubule fiber diffraction experiments. FiDAT is capable of handling all stages of fiber diffraction experiment results analysis, with pre-configured default parameters for immediate use and customizable options for experienced users. The software's effectiveness has been validated by our colleagues from the TubInTrain consortium, who applied it to analyze experimental results obtained during their time at a synchrotron facility, demonstrating its ability to facilitate real-time decision-making and hypothesis testing. Future directions of work include enhancing FiDAT by resolving minor bugs and incorporating additional logging features to further streamline data reporting. We are also committed to releasing FiDAT as an open-source software, making this powerful tool accessible to the wider scientific community. Furthermore, we are in discussions with the management of the Alba synchrotron to integrate FiDAT into their mainframe, making it a default tool for all researchers at the synchrotron.

The computational ligand- and structure-based approaches used in the different studies for the exploration of new scaffolds and hit compounds that bind to the well-known and underexplored tubulin binding sites alike yielded numerous hits that were experimentally validated. In this work, we used state-of-the-art approaches such as efficient substructure and similarity search, automated pharmacophore modeling and screening, binding site similarity search, unconstrained protein-ligand docking, covalent protein-ligand docking, machine learning-driven protein-ligand docking, accelerated molecular dynamics simulations, deep learning-based *de novo* molecular design to identify potential tubulin-targeting agents. The study involved multiple iterations of design, testing, and optimization of various compounds, which ultimately provided novel insights into different binding sites on the α,β-tubulin heterodimer, offering promising avenues for further investigation.

The discovery of novel potential compounds for immobilizing tubulin and the design of molecular probes for studying microtubule dynamics are significant advancements with far-reaching therapeutic implications. These achievements could catalyze the development of innovative drugs that specifically target tubulin and modulate its polymerization, thus providing improved treatment strategies for diseases such as cancer and neurodegenerative disorders. Future research can focus on further optimization of compounds identified in this study to enhance their binding efficiency and specificity to different tubulin isotypes. Further *in vivo* testing could assess their efficacy and safety profiles, which are essential steps in the drug development process.

This Ph.D. thesis is a testament to the power of interdisciplinary collaboration, as evidenced by the successful integration of various areas of expertise within the TubInTrain team. The consortium's multifaceted approach encompassed the design of chemical scaffolds with target-oriented and customized biochemical properties. Computational chemistry served as a critical pillar in this endeavor, underpinning the design and optimization of chemical compounds. It guided the organic synthesis efforts and set the stage for X-ray crystallography, biochemical, and cellular experiments. TubInTrain's interdisciplinary approach has effectively established a new benchmark for future microtubule research. It underscores the value of cross-disciplinary cooperation, bringing together professionals in computational chemistry, organic chemistry, biochemistry, and structural biology. This thesis offers a compelling demonstration of how such collaboration can surmount challenges more efficiently and drive successful molecular design strategies. In essence, the combined efforts of these diverse fields enable a holistic, integrative approach to MT research, paving the way for future advancements.

List of abbreviations

ACO	Ant Colony Optimization
ACoVAE	Attention-Based Conditional Variational Autoencoder
ADC	Antibody-Drug Conjugate
BA	Balanced Accuracy
CADD	Computer-Aided Drug Design
ChemPLP	Chemical Piecewise Linear Potential
CIB	Center For Biological Research (Centro De Investigaciones Biológicas)
CPU	Central Processing Unit
CSIC	Spanish National Research Council (Consejo Superior De Investigaciones Científicas)
CUDA	Compute Unified Device Architecture
CVAE	Conditional Variational Autoencoder
DNA	Deoxyribonucleic Acid
DS	Descriptor Set
EWG	Electron-Withdrawing Group
FDA	Food And Drug Administration
γ-TuRC	γ-Tubulin Ring Complex
GDP	Guanosine Diphosphate
GPU	Graphics Processing Unit
GROVER	Graph Representation From Self-supervised Message Passing Transformer
GTP	Guanosine Triphosphate
GUI	Graphical User Interface
HTS	High-Throughput Screening
IC50	Half-Maximal Inhibitory Concentration
ISIDA	In Silico Design And Data Analysis
ITN	International Training Network
MAP	Microtubule-Associated Protein
MCTS	Monte-Carlo Tree Search
MD	Molecular Dynamics
MDA	Microtubule-Destabilizing Agents
MMFF	Merck Molecular Force Field
MSA	Microtubule-Stabilizing Agents
MT	Microtubule
MTA	Microtubule-Targeting Agents

MTOC	Microtubule-Organizing Center
NP	Natural Product
PCA	Principal Component Analysis
PCM	Pericentriolar Material
PDB	Protein Data Bank
PDBQT	Protein Data Bank, Partial Charge And Atom Type
PLANTS	Protein-Ligand Ant System
QSAR	Quantitative Structure-Activity Relationship
RAM	Random-Access Memory
RCSB	Research Collaboratory For Structural Bioinformatics
RMSD	Root Mean Square Deviation
RMSF	Root Mean Square Fluctuation
SDF	Structure-Data File
SMARTS	SMILES Arbitrary Target Specification
SMILES	Simplified Molecular Input Line Entry System
SPORES	Structure Protonation And Recognition System
SVM	Support Vector Machine
USPTO	United States Patent And Trademark Office
VAE	Variational Autoencoder
VS	Virtual Screening

References

1. Sung, H. *et al.* Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA. Cancer J. Clin.* **71**, 209–249 (2021).
2. *Global status report on the public health response to dementia. Geneva: World Health Organization; 2021*.
3. Mukhtar, E., Adhami, V. M. & Mukhtar, H. Targeting Microtubules by Natural Agents for Cancer Therapy. *Mol. Cancer Ther.* **13**, 275–284 (2014).
4. Nogales, E. Tubulin and Its Isoforms. in *Encyclopedia of Biological Chemistry* 450–453 (Elsevier, 2013). doi:10.1016/B978-0-12-378630-2.00479-5.
5. Steinmetz, M. O. & Prota, A. E. Microtubule-Targeting Agents: Strategies To Hijack the Cytoskeleton. *Trends Cell Biol.* **28**, 776–792 (2018).
6. Alberts, B. *et al. Molecular Biology of the Cell*. (W. W. Norton & Company, 2022).
7. Löwe, J., Li, H., Downing, K. . & Nogales, E. Refined structure of αβ-tubulin at 3.5 Å resolution. *J. Mol. Biol.* **313**, 1045–1057 (2001).
8. Gudimchuk, N. B. & McIntosh, J. R. Regulation of microtubule dynamics, mechanics and function through the growing tip. *Nat. Rev. Mol. Cell Biol.* **22**, 777–795 (2021).
9. Chakraborti, S., Natarajan, K., Curiel, J., Janke, C. & Liu, J. The emerging role of the tubulin code: From the tubulin molecule to neuronal function and disease. *Cytoskeleton* **73**, 521–550 (2016).
10. Boiarska, Z. & Passarella, D. Microtubule-targeting agents and neurodegeneration. *Drug Discov. Today* **26**, 604–615 (2021).
11. Hoff, K. J., Neumann, A. J. & Moore, J. K. The molecular biology of tubulinopathies: Understanding the impact of variants on tubulin structure and microtubule regulation. *Front. Cell. Neurosci.* **16**, (2022).
12. Kavallaris, M. Microtubules and resistance to tubulin-binding agents. *Nat. Rev. Cancer* **10**, 194–204 (2010).
13. Nogales, E. & Wang, H.-W. Structural mechanisms underlying nucleotide-dependent self-assembly of tubulin and its relatives. *Curr. Opin. Struct. Biol.* **16**, 221–229 (2006).
14. Manka, S. W. & Moores, C. A. Microtubule structure by cryo-EM: snapshots of dynamic instability. *Essays Biochem.* **62**, 737–751 (2018).
15. Brouhard, G. J. & Rice, L. M. Microtubule dynamics: an interplay of biochemistry and mechanics. *Nat. Rev. Mol. Cell Biol.* **19**, 451–463 (2018).
16. Bollinger, J. A., Imam, Z. I., Stevens, M. J. & Bachand, G. D. Tubulin islands containing slowly hydrolyzable GTP analogs regulate the mechanism and kinetics of microtubule

depolymerization. *Sci. Rep.* **10**, 13661 (2020).

17. Lawrence, E. & Zanic, M. Rescuing microtubules from the brink of catastrophe: CLASPs lead the way. *Curr. Opin. Cell Biol.* **56**, 94–101 (2019).
18. Burute, M. & Kapitein, L. C. Cellular Logistics: Unraveling the Interplay Between Microtubule Organization and Intracellular Transport. *Annu. Rev. Cell Dev. Biol.* **35**, 29–54 (2019).
19. Appert-Rolland, C., Ebbinghaus, M. & Santen, L. Intracellular transport driven by cytoskeletal motors: General mechanisms and defects. *Phys. Rep.* **593**, 1–59 (2015).
20. Murillo, B. & Mendes Sousa, M. Neuronal Intrinsic Regenerative Capacity: T he Impact of Microtubule Organization and Axonal Transport. *Dev. Neurobiol.* **78**, 952–959 (2018).
21. Peng, N. & Nakamura, F. Microtubule-associated proteins and enzymes modifying tubulin. *Cytoskeleton* **80**, 60–76 (2023).
22. Sulimenko, V., Dráberová, E. & Dráber, P. γ-Tubulin in microtubule nucleation and beyond. *Front. Cell Dev. Biol.* **10**, (2022).
23. Lu, Y. *et al.* Stathmin destabilizing microtubule dynamics promotes malignant potential in cancer cells by epithelial-mesenchymal transition. *Hepatobiliary Pancreat. Dis. Int.* **13**, 386–394 (2014).
24. Ferreira, J. G., Pereira, A. L. & Maiato, H. Microtubule Plus-End Tracking Proteins and Their Roles in Cell Division. in 59–140 (2014). doi:10.1016/B978-0-12-800255-1.00002-8.
25. van de Willige, D., Hoogenraad, C. C. & Akhmanova, A. Microtubule plus-end tracking proteins in neuronal development. *Cell. Mol. Life Sci.* **73**, 2053–2077 (2016).
26. Jain, K., Yadav, S. A. & Athale, C. A. Number Dependence of Microtubule Collective Transport by Kinesin and Dynein. *J. Indian Inst. Sci.* **101**, 19–30 (2021).
27. Lu, W. & Gelfand, V. I. Moonlighting Motors: Kinesin, Dynein, and Cell Polarity. *Trends Cell Biol.* **27**, 505–514 (2017).
28. Walczak, C. E. & Heald, R. Mechanisms of Mitotic Spindle Assembly and Function. in 111–158 (2008). doi:10.1016/S0074-7696(07)65003-7.
29. Jaiswal, S., Kasera, H., Jain, S., Khandelwal, S. & Singh, P. Centrosome: A Microtubule Nucleating Cellular Machinery. *J. Indian Inst. Sci.* **101**, 5–18 (2021).
30. Musacchio, A. Spindle assembly checkpoint: the third decade. *Philos. Trans. R. Soc. B Biol. Sci.* **366**, 3595–3604 (2011).
31. Oriola, D., Needleman, D. J. & Brugués, J. The Physics of the Metaphase Spindle. *Annu. Rev. Biophys.* **47**, 655–673 (2018).
32. Kapoor, T. Metaphase Spindle Assembly. *Biology (Basel).* **6**, 8 (2017).

33. Wang, W.-H. Meiotic spindle, spindle checkpoint and embryonic aneuploidy. *Front. Biosci.* **11**, 620 (2006).
34. Field, J. J., Kanakkanthara, A. & Miller, J. H. Microtubule-targeting agents are clinically successful due to both mitotic and interphase impairment of microtubule function. *Bioorg. Med. Chem.* **22**, 5050–5059 (2014).
35. Čermák, V. *et al.* Microtubule-targeting agents and their impact on cancer treatment. *Eur. J. Cell Biol.* **99**, 151075 (2020).
36. Visconti, R. & Grieco, D. Fighting tubulin-targeting anticancer drug toxicity and resistance. *Endocr. Relat. Cancer* **24**, T107–T117 (2017).
37. Cao, Y.-N. *et al.* Recent advances in microtubule-stabilizing agents. *Eur. J. Med. Chem.* **143**, 806–828 (2018).
38. Bates, D. & Eastman, A. Microtubule destabilising agents: far more than just antimitotic anticancer drugs. *Br. J. Clin. Pharmacol.* **83**, 255–268 (2017).
39. Wang, J., Miller, D. D. & Li, W. Molecular interactions at the colchicine binding site in tubulin: An X-ray crystallography perspective. *Drug Discov. Today* **27**, 759–776 (2022).
40. Matthew, S. *et al.* Gatorbulin-1, a distinct cyclodepsipeptide chemotype, targets a seventh tubulin pharmacological site. *Proc. Natl. Acad. Sci.* **118**, (2021).
41. Prota, A. E. *et al.* A new tubulin-binding site and pharmacophore for microtubule-destabilizing anticancer drugs. *Proc. Natl. Acad. Sci.* **111**, 13817–13821 (2014).
42. Coulup, S. K. & Georg, G. I. Revisiting microtubule targeting agents: α-Tubulin and the pironetin binding site as unexplored targets for cancer therapeutics. *Bioorg. Med. Chem. Lett.* **29**, 1865–1873 (2019).
43. Yang, J. *et al.* Pironetin reacts covalently with cysteine-316 of α-tubulin to destabilize microtubule. *Nat. Commun.* **7**, 12103 (2016).
44. Muehlethaler, T. *et al.* Comprehensive Analysis of Binding Sites in Tubulin. *Angew. CHEMIE-INTERNATIONAL Ed.* **60**, 13331–13342 (2021).
45. Mühlethaler, T. *et al.* Rational Design of a Novel Tubulin Inhibitor with a Unique Mechanism of Action. *Angew. Chemie Int. Ed.* **61**, (2022).
46. Macalino, S. J. Y., Gosu, V., Hong, S. & Choi, S. Role of computer-aided drug design in modern drug discovery. *Arch. Pharm. Res.* **38**, 1686–1701 (2015).
47. Sabe, V. T. *et al.* Current trends in computer aided drug design and a highlight of drugs discovered via computational techniques: A review. *Eur. J. Med. Chem.* **224**, 113705 (2021).
48. Cherkasov, A. *et al.* QSAR Modeling: Where Have You Been? Where Are You Going To? *J. Med. Chem.* **57**, 4977–5010 (2014).

49. Baskin, I. Fragment Descriptors in SAR/QSAR/QSPR Studies, Molecular Similarity Analysis and in Virtual Screening. in *Chemoinformatics Approaches to Virtual Screening* (ed. Varnek, A.) 1–43 (The Royal Society of Chemistry, 2008). doi:10.1039/9781847558879-00001.
50. Ruggiu, F., Marcou, G., Varnek, A. & Horvath, D. ISIDA Property-Labelled Fragment Descriptors. *Mol. Inform.* **29**, 855–868 (2010).
51. Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol. Inform.* **29**, 476–488 (2010).
52. Johnson, M. A. & Maggiora, G. M. *Concepts and applications of molecular similarity*. (Wiley, 1990).
53. Stumpfe, D. & Bajorath, J. Similarity searching. *WIREs Comput. Mol. Sci.* **1**, 260–282 (2011).
54. Muegge, I. & Mukherjee, P. An overview of molecular fingerprint similarity search in virtual screening. *Expert Opin. Drug Discov.* **11**, 137–148 (2016).
55. Horvath, D., Koch, C., Schneider, G., Marcou, G. & Varnek, A. Local neighborhood behavior in a combinatorial library context. *J. Comput. Aided. Mol. Des.* **25**, 237–252 (2011).
56. Ehrlich, H.-C. & Rarey, M. Systematic benchmark of substructure search in molecular graphs - From Ullmann to VF2. *J. Cheminform.* **4**, 13 (2012).
57. Daylight, I. SMARTS - A Language for Describing Molecular Patterns. *Official documentation* (1997).
58. Kochev, N., Monev, V. & Bangov, I. Searching Chemical Structures. in *Chemoinformatics* 291–318 (Wiley-VCH Verlag GmbH & Co. KGaA). doi:10.1002/3527601643.ch6.
59. Sippl, W. & Robaa, D. QSAR/QSPR. in *Applied Chemoinformatics* 9–52 (Wiley-VCH Verlag GmbH & Co. KGaA, 2018). doi:10.1002/9783527806539.ch2.
60. Sushko, I. *et al.* Applicability Domains for Classification Problems: Benchmarking of Distance to Models for Ames Mutagenicity Set. *J. Chem. Inf. Model.* **50**, 2094–2111 (2010).
61. Gawehn, E., Hiss, J. A. & Schneider, G. Deep Learning in Drug Discovery. *Mol. Inform.* **35**, 3–14 (2016).
62. Combs, K., Lu, H. & Bihl, T. J. Transfer Learning and Analogical Inference: A Critical Comparison of Algorithms, Methods, and Applications. *Algorithms* **16**, 146 (2023).
63. Osipenko, S., Botashev, K., Nikolaev, E. & Kostyukevich, Y. Transfer learning for small molecule retention predictions. *J. Chromatogr. A* **1644**, 462119 (2021).
64. Ju, R. *et al.* Deep Neural Network Pretrained by Weighted Autoencoders and Transfer Learning for Retention Time Prediction of Small Molecules. *Anal. Chem.* **93**, 15651–15658

(2021).

65. Schneider, G. & Clark, D. E. Automated De Novo Drug Design: Are We Nearly There Yet? *Angew. Chemie Int. Ed.* **58**, 10792–10803 (2019).
66. Bort, W. *et al.* Inverse QSAR: Reversing Descriptor-Driven Prediction Pipeline Using Attention-Based Conditional Variational Autoencoder. *J. Chem. Inf. Model.* **62**, 5471–5484 (2022).
67. Schwaller, P. *et al.* Machine intelligence for chemical reaction space. *WIREs Comput. Mol. Sci.* **12**, (2022).
68. Tu, Z., Stuyver, T. & Coley, C. W. Predictive chemistry: machine learning for reaction deployment, reaction development, and reaction discovery. *Chem. Sci.* **14**, 226–244 (2023).
69. Seidel, T., Wolber, G. & Murgueitio, M. S. Pharmacophore Perception and Applications. in *Applied Chemoinformatics* 259–282 (Wiley-VCH Verlag GmbH & Co. KGaA, 2018). doi:10.1002/9783527806539.ch6f.
70. Wolber, G. & Langer, T. LigandScout: 3-D Pharmacophores Derived from Protein-Bound Ligands and Their Use as Virtual Screening Filters. *J. Chem. Inf. Model.* **45**, 160–169 (2005).
71. Gaurav, A. & Gautam, V. Structure-based three-dimensional pharmacophores as an alternative to traditional methodologies. *J. Receptor. Ligand Channel Res.* 27 (2014) doi:10.2147/JRLCR.S46845.
72. Rognan, D. Binding Site Similarity Search to Identify Novel Target–Ligand Complexes. in *Computational Chemogenomics* (ed. Jacoby, E.) 183–206 (Jenny Stanford Publishing, 2013). doi:10.1201/b15631-8.
73. Naderi, M. *et al.* Binding site matching in rational drug design: algorithms and applications. *Brief. Bioinform.* **20**, 2167–2184 (2019).
74. Kolodzik, A., Schneider, N. & Rarey, M. Structure-Based Virtual Screening. in *Applied Chemoinformatics* 313–331 (Wiley-VCH Verlag GmbH & Co. KGaA, 2018). doi:10.1002/9783527806539.ch6h.
75. Halperin, I., Ma, B., Wolfson, H. & Nussinov, R. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins Struct. Funct. Genet.* **47**, 409–443 (2002).
76. Sulimov, V. B., Kutov, D. C. & Sulimov, A. V. Advances in Docking. *Curr. Med. Chem.* **26**, 7555–7580 (2020).
77. Li, J., Fu, A. & Zhang, L. An Overview of Scoring Functions Used for Protein–Ligand Interactions in Molecular Docking. *Interdiscip. Sci. Comput. Life Sci.* **11**, 320–328 (2019).
78. Scarpino, A., Ferenczy, G. G. & Keserű, G. M. Binding Mode Prediction and Virtual

Screening Applications by Covalent Docking. in 73–88 (2021). doi:10.1007/978-1-0716-1209-5_4.

79. Korb, O., Stützle, T. & Exner, T. E. PLANTS: Application of Ant Colony Optimization to Structure-Based Drug Design. in 247–258 (2006). doi:10.1007/11839088_22.
80. Morris, G. M. *et al.* AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J. Comput. Chem.* **30**, 2785–2791 (2009).
81. Santos-Martins, D. *et al.* Accelerating AutoDock 4 with GPUs and Gradient-Based Local Search. *J. Chem. Theory Comput.* **17**, 1060–1073 (2021).
82. Korb, O., Stützle, T. & Exner, T. E. Empirical Scoring Functions for Advanced Protein−Ligand Docking with PLANTS. *J. Chem. Inf. Model.* **49**, 84–96 (2009).
83. Hollingsworth, S. A. & Dror, R. O. Molecular Dynamics Simulation for All. *Neuron* **99**, 1129–1143 (2018).
84. Bhattarai, A. & Miao, Y. Gaussian accelerated molecular dynamics for elucidation of drug pathways. *Expert Opin. Drug Discov.* **13**, 1055–1065 (2018).
85. Kupchan, S. M. *et al.* Tumor inhibitors. LXXIII. Maytansine, a novel antileukemic ansa macrolide from Maytenus ovatus. *J. Am. Chem. Soc.* **94**, 1354–1356 (1972).
86. Li, W. *et al.* C3 ester side chain plays a pivotal role in the antitumor activity of Maytansinoids. *Biochem. Biophys. Res. Commun.* **566**, 197–203 (2021).
87. Marzullo, P. *et al.* Maytansinol Derivatives: Side Reactions as a Chance for New Tubulin Binders. *Chem. – A Eur. J.* **28**, (2022).
88. Menchon, G. *et al.* A fluorescence anisotropy assay to discover and characterize ligands targeting the maytansine site of tubulin. *Nat. Commun.* **9**, 2106 (2018).
89. Lambert, J. M. & Chari, R. V. J. Ado-trastuzumab Emtansine (T-DM1): An Antibody–Drug Conjugate (ADC) for HER2-Positive Breast Cancer. *J. Med. Chem.* **57**, 6949–6964 (2014).
90. Porter, J. *et al.* A highly potent maytansinoid analogue and its use as a cytotoxic therapeutic agent in gold nanoparticles for the treatment of hepatocellular carcinoma. *Bioorg. Med. Chem. Lett.* **30**, 127634 (2020).
91. Mendez, D. *et al.* ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.* **47**, D930–D940 (2019).
92. Burley, S. K. *et al.* RCSB Protein Data Bank (RCSB.org): delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning. *Nucleic Acids Res.* **51**, D488–D508 (2023).
93. Hoffer, L. & Horvath, D. S4MPLE – Sampler For Multiple Protein–Ligand Entities: Simultaneous Docking of Several Entities. *J. Chem. Inf. Model.* **53**, 88–102 (2013).
94. Li, K. *et al.* Glycybridins A–K, Bioactive Phenolic Compounds from Glycyrrhiza glabra. *J.*

Nat. Prod. **80**, 334–346 (2017).

95. Genheden, S. *et al.* AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning. *J. Cheminform.* **12**, 70 (2020).
96. Lowe, D. Chemical reactions from US patents (1976 - Sep 2016). *Figshare* https://figshare.com/articles/dataset/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873 (2017) doi:10.6084/m9.figshare.5104873.v1.
97. Bañuelos-Hernández, A. E., Mendoza-Espinoza, J. A., Pereda-Miranda, R. & Cerda-García-Rojas, C. M. Studies of (−)-Pironetin Binding to α-Tubulin: Conformation, Docking, and Molecular Dynamics. *J. Org. Chem.* **79**, 3752–3764 (2014).
98. Huang, D. S., Wong, H. L. & Georg, G. I. Synthesis and evaluation of C2 functionalized analogs of the α-tubulin-binding natural product pironetin. *Bioorg. Med. Chem. Lett.* **28**, 2789–2793 (2018).
99. Vergoten, G. & Bailly, C. Molecular Docking of Cryptoconcatones to α-Tubulin and Related Pironetin Analogues. *Plants* **12**, 296 (2023).
100. Alpízar-Pedraza, D., Veulens, A. de la N., Araujo, E. C., Piloto-Ferrer, J. & Sánchez-Lamar, Á. Microtubules destabilizing agents binding sites in tubulin. *J. Mol. Struct.* **1259**, 132723 (2022).
101. Marco, J. A. *et al.* Design and synthesis of pironetin analogues with simplified structure and study of their interactions with microtubules. *Eur. J. Med. Chem.* **46**, 1630–1637 (2011).
102. Gell, C. *et al.* Microtubule Dynamics Reconstituted In Vitro and Imaged by Single-Molecule Fluorescence Microscopy. in 221–245 (2010). doi:10.1016/S0091-679X(10)95013-9.
103. Donhauser, Z. J., Appadoo, V., Kliman, E. J., Jobs, W. B. & Sheffield, E. C. Structural Changes in Tubulin Sheets Caused by Immobilization on Solid Supports. *ACS Omega* **3**, 18196–18202 (2018).
104. Velázquez-Libera, J. L., Durán-Verdugo, F., Valdés-Jiménez, A., Núñez-Vivanco, G. & Caballero, J. LigRMSD: a web server for automatic structure matching and RMSD calculations among identical and similar compounds in protein-ligand docking. *Bioinformatics* **36**, 2912–2914 (2020).
105. Ertl, P. & Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminform.* **1**, 8 (2009).
106. Mühlethaler, T. *et al.* Comprehensive Analysis of Binding Sites in Tubulin. *Angew. Chemie - Int. Ed.* **60**, 13331–13342 (2021).

107. Dreiman, G. H. S., Bictash, M., Fish, P. V., Griffin, L. & Svensson, F. Changing the HTS Paradigm: AI-Driven Iterative Screening for Hit Finding. *SLAS Discov.* **26**, 257–262 (2021).
108. Gentile, F. *et al.* Deep Docking: A Deep Learning Platform for Augmentation of Structure Based Drug Discovery. *ACS Cent. Sci.* **6**, 939–949 (2020).
109. Graff, D. E., Shakhnovich, E. I. & Coley, C. W. Accelerating high-throughput virtual screening through molecular pool-based active learning. *Chem. Sci.* **12**, 7866–7881 (2021).
110. Horvath, D., Brown, J., Marcou, G. & Varnek, A. An Evolutionary Optimizer of libsvm Models. *Challenges* **5**, 450–472 (2014).
111. de Souza Neto, L. R. *et al.* In silico Strategies to Support Fragment-to-Lead Optimization in Drug Discovery. *Front. Chem.* **8**, 1–18 (2020).
112. Desaphy, J., Bret, G., Rognan, D. & Kellenberger, E. sc-PDB: a 3D-database of ligandable binding sites—10 years on. *Nucleic Acids Res.* **43**, D399–D404 (2015).
113. Konc, J. & Janežič, D. ProBiS algorithm for detection of structurally similar protein binding sites by local structural alignment. *Bioinformatics* **26**, 1160–1168 (2010).
114. Huang, F., Han, X., Xiao, X. & Zhou, J. Covalent Warheads Targeting Cysteine Residue: The Promising Approach in Drug Development. *Molecules* **27**, 7728 (2022).
115. Britto, P. J., Knipling, L. & Wolff, J. The Local Electrostatic Environment Determines Cysteine Reactivity of Tubulin. *J. Biol. Chem.* **277**, 29018–29027 (2002).
116. Wang, H. *et al.* Sequence-Based Prediction of Cysteine Reactivity Using Machine Learning. *Biochemistry* **57**, 451–460 (2018).
117. Bas, D. C., Rogers, D. M. & Jensen, J. H. Very fast prediction and rationalization of pKa values for protein-ligand complexes. *Proteins Struct. Funct. Bioinforma.* **73**, 765–783 (2008).
118. Soylu, I. & Marino, S. M. Cpipe: a comprehensive computational platform for sequence and structure-based analyses of Cysteine residues. *Bioinformatics* **33**, 2395–2396 (2017).
119. Lu, Y., Huang, F., Wang, J. & Xia, J. Affinity-Guided Covalent Conjugation Reactions Based on PDZ–Peptide and SH3–Peptide Interactions. *Bioconjug. Chem.* **25**, 989–999 (2014).
120. Gao, M., Moumbock, A. F. A., Qaseem, A., Xu, Q. & Günther, S. CovPDB: a high-resolution coverage of the covalent protein–ligand interactome. *Nucleic Acids Res.* **50**, D445–D450 (2022).
121. Du, H. *et al.* CovalentInDB: a comprehensive database facilitating the discovery of covalent inhibitors. *Nucleic Acids Res.* **49**, D1122–D1129 (2021).
122. Guo, X.-K. & Zhang, Y. CovBinderInPDB: A Structure-Based Covalent Binder Database.

J. Chem. Inf. Model. **62**, 6057–6068 (2022).

123. Irwin, J. J. *et al.* ZINC20—A Free Ultralarge-Scale Chemical Database for Ligand Discovery. *J. Chem. Inf. Model.* **60**, 6065–6073 (2020).
124. Hein, J. E. & Fokin, V. V. Copper-catalyzed azide–alkyne cycloaddition (CuAAC) and beyond: new reactivity of copper(i) acetylides. *Chem. Soc. Rev.* **39**, 1302 (2010).
125. Du, T. *et al.* A novel orally active microtubule destabilizing agent S-40 targets the colchicine-binding site and shows potent antitumor activity. *Cancer Lett.* **495**, 22–32 (2020).
126. Wang, J., Miller, D. D. & Li, W. Molecular interactions at the colchicine binding site in tubulin: An X-ray crystallography perspective. *Drug Discov. Today* **27**, 759–776 (2022).
127. López-López, E., Cerda-García-Rojas, C. M. & Medina-Franco, J. L. Consensus Virtual Screening Protocol Towards the Identification of Small Molecules Interacting with the Colchicine Binding Site of the Tubulin-microtubule System. *Mol. Inform.* **42**, 2200166 (2023).
128. Slobodnick, A., Shah, B., Pillinger, M. H. & Krasnokutsky, S. Colchicine: Old and New. *Am. J. Med.* **128**, 461–470 (2015).
129. López-López, E., Cerda-García-Rojas, C. M. & Medina-Franco, J. L. Tubulin Inhibitors: A Chemoinformatic Analysis Using Cell-Based Data. *Molecules* **26**, 2483 (2021).
130. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
131. Yang, Z., Xu, B., Luo, W. & Chen, F. Autoencoder-based representation learning and its application in intelligent fault diagnosis: A review. *Measurement* **189**, 110460 (2022).
132. Weng, L. From Autoencoder to Beta-VAE. *Personal blog* https://lilianweng.github.io/posts/2018-08-12-vae/ (2018).
133. Bian, Y. & Xie, X.-Q. Generative chemistry: drug discovery with deep learning generative models. *J. Mol. Model.* **27**, 71 (2021).
134. Lim, J., Ryu, S., Kim, J. W. & Kim, W. Y. Molecular generative model based on conditional variational autoencoder for de novo molecular design. *J. Cheminform.* **10**, 31 (2018).
135. Borysov, S. S. & Rich, J. Introducing synthetic pseudo panels: application to transport behaviour dynamics. *Transportation (Amst).* **48**, 2493–2520 (2021).
136. Kingma, D. P. & Welling, M. Auto-Encoding Variational Bayes. *arXiv* (2013).
137. Basseville, M. Divergence measures for statistical data processing—An annotated bibliography. *Signal Processing* **93**, 621–633 (2013).
138. Zhuang, J. *et al.* AdaBelief Optimizer: Adapting Stepsizes by the Belief in Observed Gradients. *arXiv* (2020) doi:arXiv:2010.07468.

139. Li, J., Luo, D., Wen, T., Liu, Q. & Mo, Z. Representative feature selection of molecular descriptors in QSAR modeling. *J. Mol. Struct.* **1244**, 131249 (2021).
140. Burden, F. R. & Winkler, D. A. Optimal Sparse Descriptor Selection for QSAR Using Bayesian Methods. *QSAR Comb. Sci.* **28**, 645–653 (2009).
141. Dey, V., Machiraju, R. & Ning, X. Improving Compound Activity Classification via Deep Transfer and Representation Learning. *ACS Omega* **7**, 9465–9483 (2022).
142. Yang, X., Wang, Y., Byrne, R., Schneider, G. & Yang, S. Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery. *Chem. Rev.* **119**, 10520–10594 (2019).
143. Cai, C. *et al.* Transfer Learning for Drug Discovery. *J. Med. Chem.* **63**, 8683–8694 (2020).
144. Liu, S., Guo, H. & Tang, J. Molecular Geometry Pretraining with {SE}(3)-Invariant Denoising Distance Matching. in *The Eleventh International Conference on Learning Representations* (2023).
145. Zhu, Z. *et al.* TorchDrug: A Powerful and Flexible Machine Learning Platform for Drug Discovery. *arXiv* (2022).
146. Wang, Y., Wang, J., Cao, Z. & Barati Farimani, A. Molecular contrastive learning of representations via graph neural networks. *Nat. Mach. Intell.* **4**, 279–287 (2022).
147. Stärk, H. *et al.* 3D Infomax improves GNNs for Molecular Property Prediction. in *Proceedings of the 39th International Conference on Machine Learning* (eds. Chaudhuri, K. et al.) vol. 162 20479–20502 (PMLR, 2022).
148. Irwin, R., Dimitriadis, S., He, J. & Bjerrum, E. J. Chemformer: a pre-trained transformer for computational chemistry. *Mach. Learn. Sci. Technol.* **3**, 015022 (2022).
149. Rong, Y. *et al.* GROVER: Self-supervised Message Passing Transformer on Large-scale Molecular Data. *Adv. Neural Inf. Process. Syst.* 1–13 (2020).
150. Wu, Z. *et al.* MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* **9**, 513–530 (2018).
151. Hollingsworth, S. A. & Dror, R. O. Molecular Dynamics Simulation for All. *Neuron* **99**, 1129–1143 (2018).
152. Shan, Y. *et al.* How does a small molecule bind at a cryptic binding site? *PLOS Comput. Biol.* **18**, e1009817 (2022).
153. Vajda, S., Beglov, D., Wakefield, A. E., Egbert, M. & Whitty, A. Cryptic binding sites on proteins: definition, detection, and druggability. *Curr. Opin. Chem. Biol.* **44**, 1–8 (2018).
154. Oleinikovas, V., Saladino, G., Cossins, B. P. & Gervasio, F. L. Understanding Cryptic Pocket Formation in Protein Targets by Enhanced Sampling Simulations. *J. Am. Chem. Soc.* **138**, 14257–14263 (2016).
155. Kuzmanic, A., Bowman, G. R., Juarez-Jimenez, J., Michel, J. & Gervasio, F. L.

Investigating Cryptic Binding Sites by Molecular Dynamics Simulations. *ACS Appl. Mater. Interfaces* (2020) doi:10.1021/acs.accounts.9b00613.

156. Knoverek, C. R., Amarasinghe, G. K. & Bowman, G. R. Advanced Methods for Accessing Protein Shape-Shifting Present New Therapeutic Opportunities. *Trends Biochem. Sci.* **44**, 351–364 (2019).
157. Miao, Y., Feher, V. A. & McCammon, J. A. Gaussian Accelerated Molecular Dynamics: Unconstrained Enhanced Sampling and Free Energy Calculation. *J. Chem. Theory Comput.* **11**, 3584–3595 (2015).
158. Case, D. A. *et al.* The Amber biomolecular simulation programs. *J. Comput. Chem.* **26**, 1668–1688 (2005).
159. Maier, J. A. *et al.* ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **11**, 3696–3713 (2015).
160. Ryckaert, J.-P., Ciccotti, G. & Berendsen, H. J. . Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **23**, 327–341 (1977).
161. Vila-Julià, G., Perez, J. J. & Rubio-Martinez, J. A Step Forward toward Selective Activation/Inhibition of Bak, a Pro-Apoptotic Member of the Bcl-2 Protein Family: Discovery of New Prospective Allosteric Sites Using Molecular Dynamics. *J. Chem. Inf. Model.* (2023) doi:10.1021/acs.jcim.3c00397.
162. Kozakov, D. *et al.* The FTMap family of web servers for determining and characterizing ligand-binding hot spots of proteins. *Nat. Protoc.* **10**, 733–755 (2015).
163. Chen, Z. *et al.* D3Pockets: A Method and Web Server for Systematic Analysis of Protein Pocket Dynamics. *J. Chem. Inf. Model.* **59**, 3353–3358 (2019).
164. Kamimura, S., Fujita, Y., Wada, Y., Yagi, T. & Iwamoto, H. X-ray fiber diffraction analysis shows dynamic changes in axial tubulin repeats in native microtubules depending on paclitaxel content, temperature and GTP-hydrolysis. *Cytoskeleton* **73**, 131–144 (2016).
165. Oliva, M. Á., Gago, F., Kamimura, S. & Díaz, J. F. Alternative Approaches to Understand Microtubule Cap Morphology and Function. *ACS Omega* (2022) doi:10.1021/acsomega.2c06926.
166. Sugiyama, T. *et al.* Quick shear-flow alignment of biological filaments for X-ray fiber diffraction facilitated by methylcellulose. *Biophys. J.* **97**, 3132–3138 (2009).
167. Estévez Gallego, J. Implications of the microtubule cap structure in the molecular mechanism of paclitaxel. (The Complutense University of Madrid, 2020).
168. Estévez-Gallego, J. *et al.* Structural model for differential cap maturation at growing microtubule ends. *Elife* **9**, (2020).

Maxim SHEVELEV

Conception assistée par ordinateur des modulateurs de polymérisation de la tubuline

Résumé

La protéine tubuline, cruciale pour la division cellulaire et le transport intracellulaire, est une cible clé dans la recherche sur le cancer et la neurodégénérescence. Les difficultés de synthèse et les propriétés pharmacologiques médiocres des agents existants ciblant la tubuline nécessitent de nouvelles découvertes. L'objectif de cette thèse était d'utiliser la conception de médicaments assistée par ordinateur pour identifier de nouvelles molécules qui ciblent des sites de liaison moins explorés et qui sont plus accessibles. La thèse a ciblé les sites peu étudiés de la maytansine, de la pironétine et du todalam avec des approches de criblage virtuel basées sur les ligands et la structure, et a conçu de nouvelles molécules pour le site de la colchicine en utilisant des technologies avancées d'apprentissage profond. La recherche a permis d'obtenir un total de 28 agents déstabilisateurs de microtubules nouveaux et structurellement diversifiés, ciblant les sites todalam, maytansine et colchicine. En outre, un logiciel d'analyse automatisée des images de microscope provenant d'expériences de diffraction de fibres de microtubules a été développé.

Mots clés : criblage virtuel, apprentissage profond, les agents antitubulines

Résumé en anglais

The tubulin protein, crucial for cell division and intracellular transport, is a key target in cancer and neurodegeneration research. Synthetic challenges and poor pharmacological properties of existing tubulin-targeting agents necessitate new discoveries. The goal of this thesis was to use computer-aided drug design to identify novel molecules that target less explored binding sites and are more synthetically accessible. The thesis targeted the understudied maytansine, pironetin, and todalam sites with ligand- and structure-based virtual screening approaches, and designed new molecules for the colchicine site using advanced deep learning technologies. The research yielded a total of twenty-eight structurally diverse and novel microtubule-destabilizing agents targeting the todalam, maytansine, and colchicine sites. Moreover, a software for automated analysis of microscope images from microtubule fiber diffraction experiments was developed.

Keywords: virtual screening, deep learning, microtubule-targeting agents