# THÈSE présentée par :

## Kossiwa Ikafui Merveille EGUIDA

soutenue le : **23 Septembre 2022**

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

Discipline / Spécialité : Chimie

## Comparaison de cavités protéiques par traitement numérique de nuages de points : principes et applications en drug design

**THÈSE dirigée par :**
  M. ROGNAN Didier                    Directeur de recherche, CNRS

**RAPPORTEURS :**
  Mme DOUGUET Dominique         Chargée de recherche, INSERM
  M. LEWIS Richard                      Directeur CADD & Data Science, Novartis

**AUTRES MEMBRES DU JURY :**
  M. HIBERT Marcel                      Professeur émérite, Université de Strasbourg
  Mme KELLENBERGER Esther     Professeure, Université de Strasbourg
  Mme KRIER Mireille                    Chercheure, OpenEye Scientific Software

**MEMBRE INVITÉ :**
  M. MARCOU Gilles                     Maître de conférences, Université de Strasbourg

*ÉCOLE DOCTORALE DES SCIENCES CHIMIQUES*

Laboratoire d'Innovation Thérapeutique (UMR 7200)
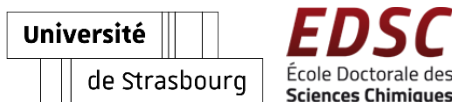
# DOCTORAL THESIS

## Comparison of protein cavities by point cloud processing: principles and applications in drug design

presented by

## Kossiwa Ikafui Merveille EGUIDA

to the University of Strasbourg

for the degree of Doctor of Philosophy
Chemistry / Cheminformatics

**Université** de Strasbourg

**EDSC**
École Doctorale des
Sciences Chimiques

September 23$^{rd}$, 2022

PhD Advisor
  **Dr Didier ROGNAN**          Research Director, CNRS

External Examiners
  **Dr Dominique DOUGUET**   Research Investigator, INSERM
  **Dr Richard LEWIS**          Director of CADD & Data Science, Novartis

Examiners
  **Pr Marcel HIBERT**          Professor, University of Strasbourg
  **Pr Esther KELLENBERGER**   Professor, University of Strasbourg
  **Dr Mireille KRIER**         Research Investigator, OpenEye Scientific Software

Guest Examiner
  **Dr Gilles MARCOU**          Associate Professor, University of Strasbourg

To my loved ones.

In the words of an Ewe proverb, "Nunya adidoe, asi mesunɛ o"

(Knowledge is like a baobab tree, no one can embrace it all, nor alone).

# Abstract

Protein cavities are the heart of molecular interactions that trigger and regulate biological processes in living organisms. Supported by the constant augmentation of characterized pockets in three-dimensional protein structures, methods to assess the similarity between protein cavities have multiple applications in drug design but face many challenges. This thesis proposes new algorithms based on three-dimensional (3D) image processing to compare global and subtle patterns in different protein (sub-) pockets represented by point clouds. Through prospective applications validated by *in vitro* biological experiments, we showed how these methods can predict a secondary target at the proteome scale and design a target-focused library for faster small molecule hit identification. In the next stages, better characterization of the cavities for pharmacophore elaboration and the development of virtual screening methods were investigated.

Keywords: protein subpocket comparison, point cloud, 3D alignment, secondary target prediction, focused library, virtual screening, pharmacophore, graph matching, machine learning, drug design, structure-based, Cheminformatics.

# Résumé (Abstract in French)

Les cavités de protéines sont au cœur d'interactions moléculaires nécessaires aux fonctions biologiques du vivant. Grâce à l'augmentation incessante des données structurales, les méthodes de comparaison de cavités protéiques offrent diverses applications en conception de molécules bioactives mais doivent relever plusieurs défis. Cette thèse propose de nouveaux algorithmes basés sur le traitement d'images tridimensionnelles pour comparer les motifs globaux et locaux de (sous-) cavités protéiques, représentées en nuages de points. Leurs applications concrètes, validées par des essais biologiques *in vitro*, illustrent leurs utilisations pour prédire des cibles secondaires à l'échelle du protéome structural et pour générer des chimiothèques focalisées permettant d'augmenter le taux de touches en criblage virtuel. A partir de la caractérisation des cavités, l'élaboration de pharmacophores et le développement de méthodes de criblage virtuel ont été investigués.

Mots-Clés : comparaison de sites de protéines, nuage de points, alignement 3D, prédiction de cible secondaire, chimiothèque focalisée, criblage virtuel, pharmacophore, alignement de graphe, intelligence artificielle, conception de molécules bioactives, structure, Chémoinformatique.

# Table of contents

# Acknowledgements

**UNIVERSITE DE STRASBOURG**

**ECOLE DOCTORALE DES SCIENCES CHIMIQUES**

**RESUME DE LA THESE DE DOCTORAT**

**(THESIS SUMMARY IN FRENCH)**

Discipline : Chimie

Spécialité (facultative) : Chimie-informatique et théorique

Présentée par : Eguida Kossiwa Ikafui Merveille
(*Nom Prénom du candidat*)

**Titre : comparaison de cavités protéiques par traitement numérique de
nuages de points : principes et applications en drug design**

Unité de Recherche :   UMR 7200 Laboratoire d'Innovation Thérapeutique
(*N° et Nom de l'Unité*)

Directeur de Thèse :   Dr. Rognan Didier, Directeur de Recherche DRCE, HDR
(*Nom Prénom – Grade*)

Localisation : Faculté de Pharmacie,
               64 route du Rhin, 64700 Illkirch-Graffenstaden

Thèse confidentielle :     ■ NON      □ OUI

# 1. Introduction

Un des problèmes fondamentaux de la conception de candidat-médicaments reste l'identification de molécules bioactives ayant de bonnes propriétés pharmacologiques, ou du moins optimisables aux mêmes fins. Expérimentalement, des banques de molécules de masse molaire allant de 200 à 800 g.mol-1 (chimiothèques) sont évaluées dans des essais biologiques à haut-débit afin d'identifier des touches. Cette approche requiert des infrastructures particulières, en plus de la mise en place des essais biologiques, et est par conséquence coûteuse. Au contraire, la conception assistée par ordinateur (CAO) offre l'avantage d'être rapide et beaucoup moins onéreuse, mais s'applique lorsque certaines données sont connues : par exemple, la structure tri-dimensionnelle (3D) de la cible, les structures chimiques d'inhibiteurs, etc. Une approche populaire de la CAO est l'arrimage moléculaire ou « docking »[1] dont le principe est de prédire l'affinité de molécules à la cible, par proposition de potentiels modes de liaison et évaluation des contributions énergétiques à des fins de classement, avant de tester expérimentalement les meilleures propositions. Classiquement, un programme de docking commence par le choix de la chimiothèque à cribler, étape cruciale car les chercheurs partent d'un ensemble fini de molécules et espèrent y trouver, sans garantie, des touches pour une protéine particulière. Même si les chances d'identifier des molécules bioactives augmentent avec la taille de la chimiothèque,[2] il reste la question de la priorisation des touches. Le criblage de chimiothèques focalisées, conçues pour être enrichies en touches pour une cible donnée, s'avère avantageux.[3] Il existe donc un besoin de méthodes alternatives au docking classique, comme la comparaison de poches protéiques, en tirant profit de l'augmentation incessante des données structurales publiques de cavités de complexes protéine-ligand.[4]

Les petites molécules interagissent avec une protéine en se liant à des cavités compatibles avec leurs formes et propriétés physicochimiques. La comparaison de cavités de protéines a pour but d'estimer la similarité entre des sites de liaison de différentes protéines. Cette approche est utilisée en CAO à plusieurs fins selon le principe de similarité : générer des hypothèses de touches et identifier des cibles secondaires. Quelques applications réussies de prédictions, d'explications d'observations expérimentales ou de confirmation ont été rapportées dans la littérature.[5] Depuis la création de la banque de données structurales Protein Data Bank ou PDB, permettant la caractérisation des sites de liaison protéiques, plusieurs méthodes de comparaison de cavités ont vu le jour. Cependant, elles se différencient par la combinaison des quatre principales étapes d'une comparaison : la détection de la cavité, la sélection et représentation de motifs pertinents du site, l'algorithme de comparaison (alignement de graphe ou de motifs géométriques, comparaison d'empreintes ou d'histogrammes de distances, apprentissage automatique) et l'estimation du degré de ressemblance (scoring). La comparaison de site, reste une tâche difficile, non directement mesurable expérimentalement mais sensible à la précision de chacune des étapes énumérées ci-dessus. La délimitation du site peut être suggérée par un ligand en complexe avec la protéine, lorsqu'il est présent. Toutefois les algorithmes

opérant par détection de cavité *de novo* offre l'avantage de s'appliquer à de nouvelles cavités pour lesquelles aucune information n'est connue. Aussi, observons-nous que la majorité des méthodes existantes effectuent des comparaisons globales des sites, alors qu'une comparaison locale pourrait mettre en évidence des similarités cachées expliquant la liaison du ligand à une cible secondaire.[6]

Notre laboratoire a préalablement développé une représentation en nuage de points des sites de protéines (IChem VolSite, **Figure 1**).[7] L'objectif de cette thèse est de développer des méthodes basées sur la vision par ordinateur pour traiter et comparer les nuages de points de cavités protéiques puis d'évaluer leurs usages dans la conception de molécules bioactives.
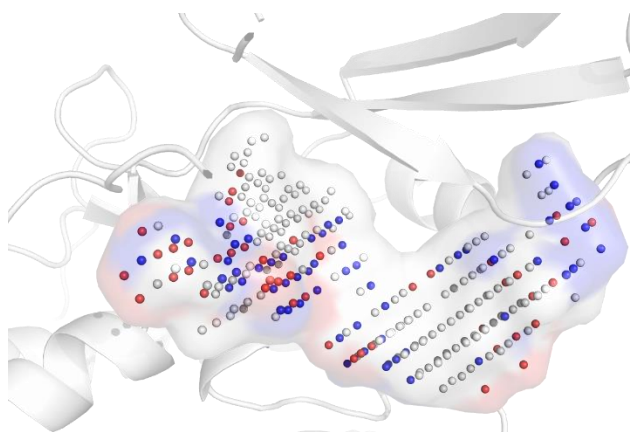


**Figure 1.** Example de nuage de points de cavité de protéine calculé par VolSite. Chaque point, est associé à une propriété pharmacophorique complémentaire à celle de l'atome protéique le plus proche : en bleu, donneur de liaison hydrogène, positivement ionisable, en rouge accepteur ou accepteur/donneur de liaison hydrogène, négativement ionisable, en blanc hydrophobe, aromatique et nul. La surface transparente du nuage est déterminée par Pymol 2.1 (Schrödinger, New York, USA), code PDB: 5HBH.

## 2. Traitements et comparaisons de cavités protéiques

La comparaison de cavités protéiques repose sur une représentation des propriétés importantes du site. Généralement, il s'agit d'encoder les relations spatiales et pharmacophoriques des atomes du site protéique, mais celle-ci peut prendre la forme d'une surface continue, d'un graphe, d'une empreinte ou de nuages de points. Mes travaux se basent sur cette dernière représentation car elle offre plusieurs avantages : les points occupent l'espace discrétisé 3D du ligand, encodent les courbures et les propriétés pharmacophoriques du site. Cependant cette discrétisation a pour inconvénient d'introduire du bruit dans la représentation, un défi pour les algorithmes de comparaison. En vision par ordinateur et robotique, des procédures particulières d'alignement de nuages de points sont utilisées pour superposer des images 3D bruitées[8] mais elles n'avaient jamais été adaptées pour aligner des cavités protéiques. Le principe

d'un de ces l'algorithmes le rend intéressant pour notre problème car il permettrait une comparaison locale tout en étant robuste aux bruits. A partir des données de la sc-PDB, une base de données de complexes de protéines-ligands non-redondants, plusieurs stratégies CAO et leurs applications concrètes ont été élaborées (**Figure 2**).
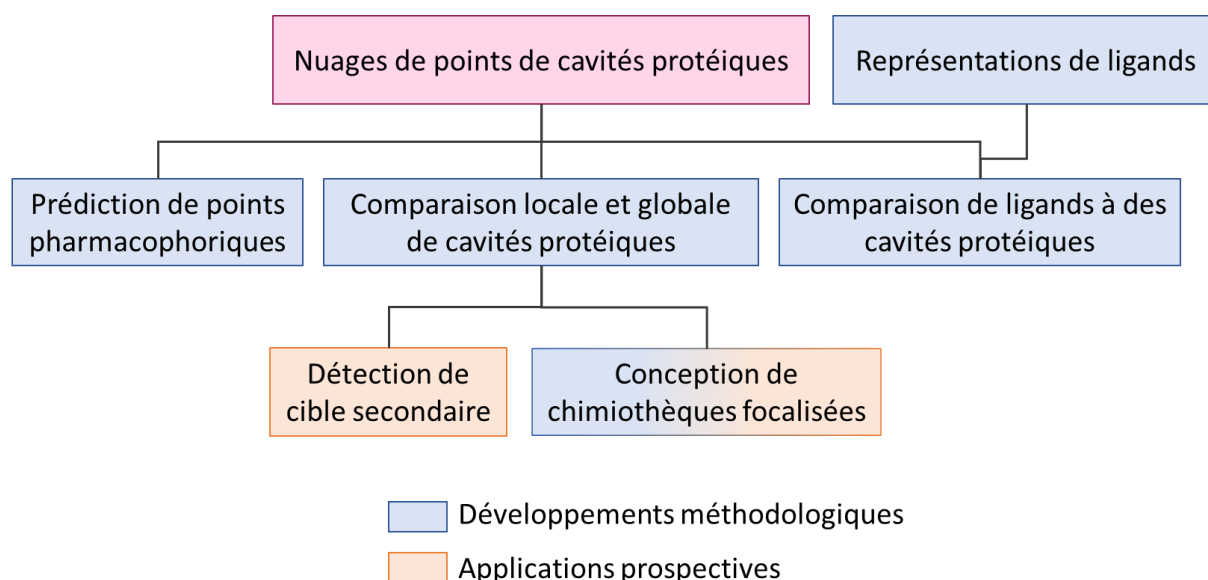


**Figure 2.** Stratégies CAO par traitement de nuages de points élaborées dans cette thèse.

## 2.1. ProCare : développement d'une nouvelle méthode de comparaison locale de cavités protéiques

ProCare est une méthode codée en C++ et en Python permettant de comparer deux nuages de points de cavités protéiques.[9] Elle est basée sur la librairie de traitement d'image Open3D,[10] adaptée et optimisée pour traiter nos représentations des cavités protéiques. La comparaison de deux cavités se déroulent en cinq étapes : (1) calcul des descripteurs de chaque point, (2) échantillonnage aléatoire d'au moins trois points de la première cavité et associations avec des points de la deuxième cavité les plus similaires dans l'espace des descripteurs et par leur topologie commune, (3) alignement grossier à partir des points associés, (4) raffinement de l'alignement par la méthode itérative du point le plus proche (« iterative closest point ») qui associe naïvement les points les plus proches dans l'espace Euclidien et enfin (5) quantification de la similarité.

Du fait que Open3D ait été développé originellement pour une autre application, nous avons dans un premier lieu optimisé les paramètres géométriques en évaluant 157 465 conditions d'alignement couvrant 15 paramètres. Ensuite, le descripteur représentant la forme locale autour de chaque point a été

modifié en y introduisant l'information pharmacophorique, ce qui a amélioré les comparaisons. Enfin, plusieurs fonctions de score ont été développées, implémentées, optimisées et finalement, un score symétrique comptant les points ayant un équivalent de même propriété dans l'autre cavité a été défini comme score principal.

Afin d'évaluer les performances de la méthode, nous avons assemblé 8 jeux de données, de taille allant de dix paires à deux millions de paires d'entrées, représentant différents scénarios de similarité de cavités (classification fonctionnelle, reconnaissance de mêmes ligands, comparaison de sous-poches de fragments avec des cavités entières de protéines différentes, sensibilités aux variations de coordonnées) et permettant la détermination statistique d'un seuil de similarité.

ProCare a montré une performance de similarité globale équivalente aux méthodes de l'état de l'art et supérieure en ce qui concerne la détection de similarité locale. Elle est sensible aux déformations globales du squelette de la cavité d'environ 2.5 Å et indique une similarité significative à partir d'un score de 0.47, la zone grise étant estimée à 0.39. Tout en reconnaissant que ces valeurs peuvent être biaisées par la composition des jeux de données, elles forment néanmoins une base de comparaison à haut-débit. Le principe de comparaison locale a été appliqué pour comparer des sous poches de protéines à des cavités entières de protéines dont les structures venaient d'être nouvellement résolues. L'alignement ainsi obtenu a été appliqué aux fragments issus de ces sous-poches afin de suggérer des blocs de construction de ligands (**Figures 3**). Ce protocole sera exploité dans les parties **2.2** et **2.3**.
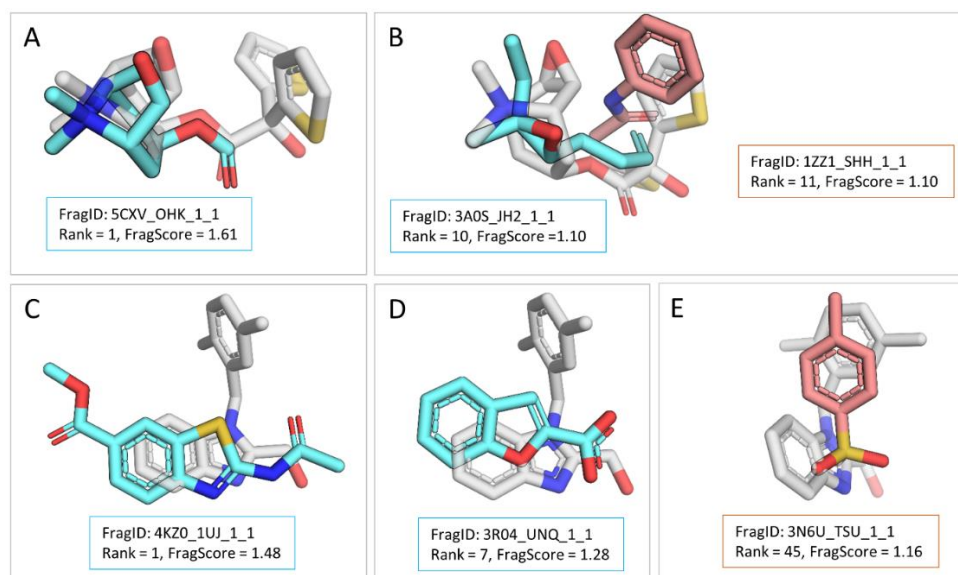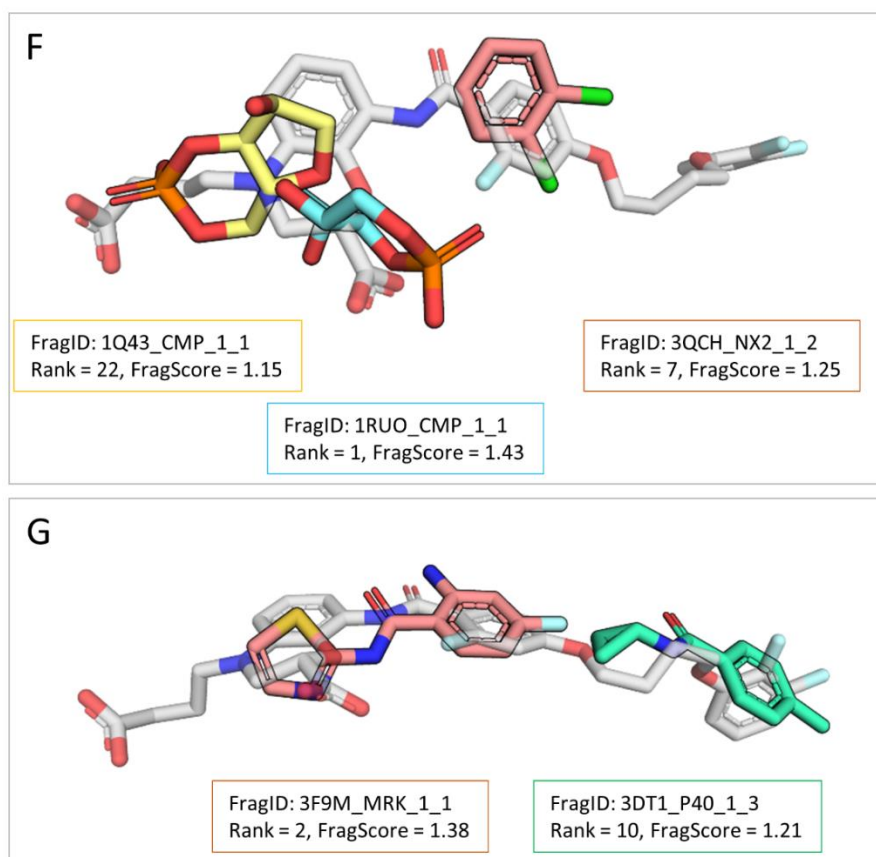
**Figure 3.** Positionnement de fragments de la sc-PDB dans de nouvelles cavités protéiques par alignement de sous-poches avec ProCare. Code couleur des atomes (azote : bleu ; oxygène : rouge; souffre : jaune ; carbone du fragment: cyan / jaune vif / rose orangé / vert ; carbone du ligand, blanc). Les codes PDB, HET, le site sc-PDB et le numéro du fragment sont indiqués. Cibles : A-B) récepteur muscarinique M5 (PDB : 6OL9), C-E) facteur de nécrose tumorale alpha (PDB : 6OOY), F-G) Récepteur des cystéinyl-leucotriènes 2 (PDB : 6RZ8).

À la suite de ces évaluations rétrospectives concluantes, nous avons évalué ProCare dans les applications prospectives en drug design.

## 2.2. Prédiction de cible secondaire par comparaison de sous-poches de protéines

La capacité de ProCare à effectuer des alignements locaux le rend prometteur pour détecter des similarités non-évidentes mais suffisantes pour favoriser la reconnaissance d'un même ligand/fragment. Nous avons comparé la poche à l'interface de la protéine homotrimérique du facteur de nécrose tumorale TNF-α[11] à une collection de 31 000 sous-poches, correspondant à diverses protéines. ProCare a prédit

une similarité locale avec des sous-poches du site non-nucléosidique de la transcriptase inverse du virus-1 de l'immunodéficience humain (HIV1-RT) de manière significative[12] : les scores sont élevés et statistiquement indépendants de la structure 3D utilisée, l'alignement des points de cavités résulte en un alignement pertinent des résidus protéiques délimitant les deux cavités, les alignements des fragments correspondent à des propositions de docking **(Figure 4)**.
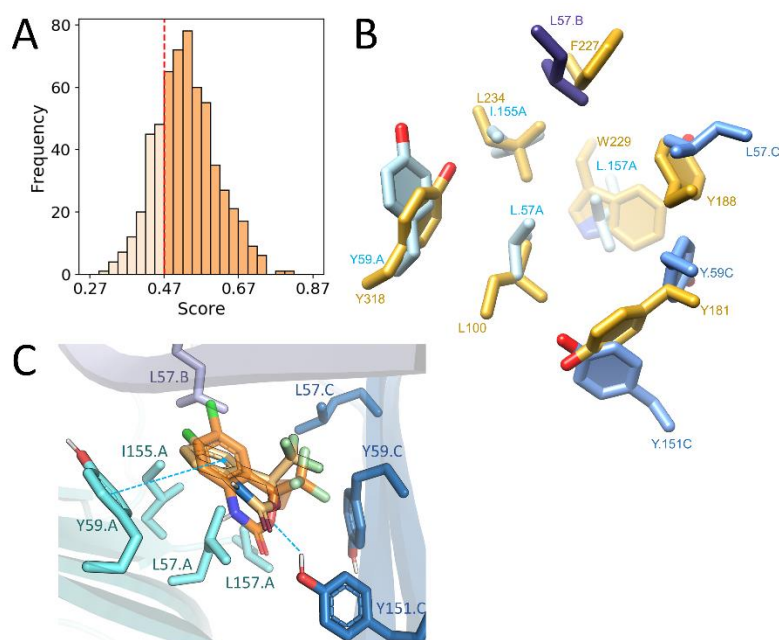


**Figure 4.** Comparaison des cavités de TNF-α et HIV1-RT avec ProCare. A) Distribution des scores de similarité. B) Résidus alignés de TNF-α (chaine A: cyan, chaine B: bleu, chaine C: bleu ciel; code PDB: 6OOZ) sur ceux de HIV1-RT (orange, code PDB: 1FKO) après rotation et translation résultant de l'alignement des cavités par ProCare. C) Alignement correspondant du fragment principal d'efavirenz (orange clair) dans la poche de TNF-α, superposé à une solution de docking (orange foncé transparent). L'interaction aromatique avec TYR59-TNF-α et la liaison hydrogène avec TYR151-TNF-α sont représentées par le trait en pointillé bleu.

Nous donc avons émis l'hypothèse que des ligands HIV1-RT peuvent se lier au TNF-α. Afin de vérifier ou de réfuter cette hypothèse, 3 inhibiteurs commercialisés (delavirdine, efavirenz et nevirapine) du site non-nucléosidique du HIV1-RT ont été testés *in vitro* pour leur capacité à se lier au TNF-α **(Figure 5)**. L'efavirenz et la delavirdine se lient au TNF-α avec une constante de dissociation à l'équilibre $K_D$ de 24±8 µM et 39±9 µM respectivement, de même ordre de grandeur que de celle du fragment co-cristallisé avec TNF-α (UCB-6876 $K_D$ = 22 µM).[11] Cette similarité non évidente entre des protéines fonctionnellement et structuralement différentes n'a pu être détectée par les méthodes existantes de comparaison de cavités protéiques, ou de similarités bi- et tri-dimensionnelles de ligands.

14
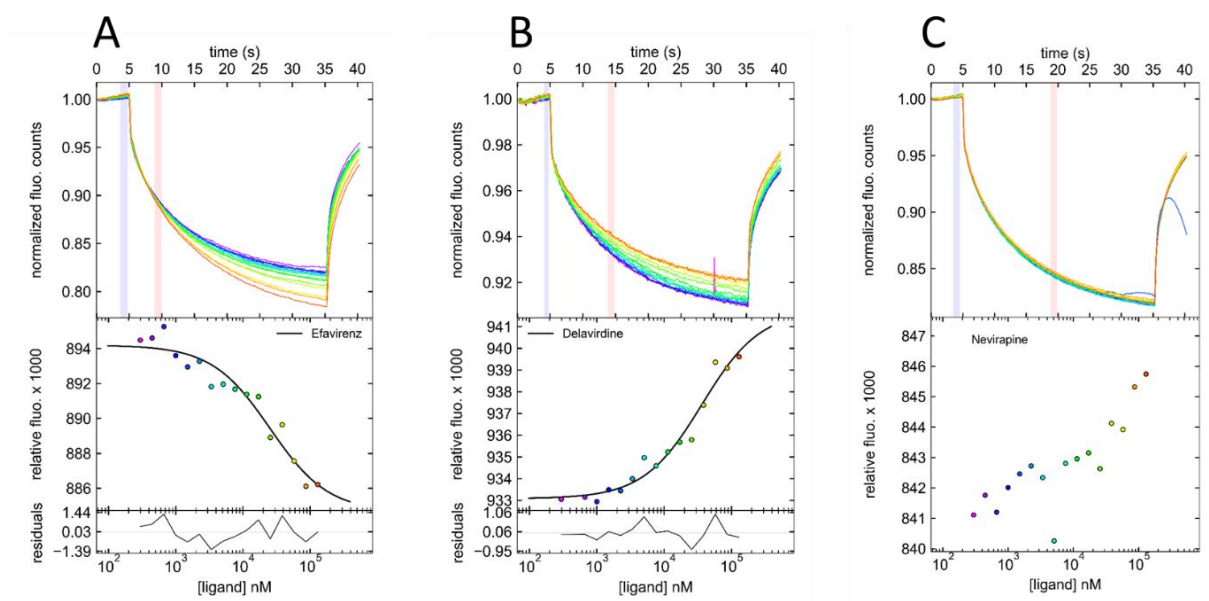
**Figure 5.** L'essai biophysique par thermophorèse (MST) démontre une liaison directe entre deux inhibiteurs non-nucléosidiques du HIV1-RT et le TNF-α. A) efavirenz ($K_D = 24 \pm 8$ µM); B) delavirdine ($K_D = 39 \pm 9$ µM); C) nevirapine (pas de liaison).

Nous avons ainsi validé l'usage de ProCare à déterminer des similarités non-évidentes et locales entre sous-poches de protéines de différentes familles.

## 2.3. Conception de chimiothèque focalisée

Une chimiothèque focalisée est une petite collection de molécules, enrichie en touches pour la cible choisie, permettant ainsi un criblage rapide et un taux de touches plus élevé.[3] De nombreuses approches publiées requièrent des ligands connus pour élaborer une chimiothèque focalisée, ce qui les rend inutilisables pour les cibles dont la seule information connue est structure protéique. Nous avons donc conçu une approche (POEM, Pocket-Oriented Elaboration of Molecule ou élaboration de molécules focalisés sur les caractéristiques de la cavité protéique, **Figure 6**) qui, à partir de la cavité de la cible, positionne des fragments obtenus de complexes protéine-ligand sur la base de la similarité de leurs microenvironnements protéiques avec la cavité cible. Les fragments sont filtrés, annotés selon zone de la cavité cible qu'ils occupent, puis liés par un algorithme d'apprentissage profond génératif[13] pour énumérer des molécules complètes. Les molécules sont ensuite vérifiées et filtrées selon leurs propriétés physico-chimiques et leur accessibilité synthétique.[14]

15
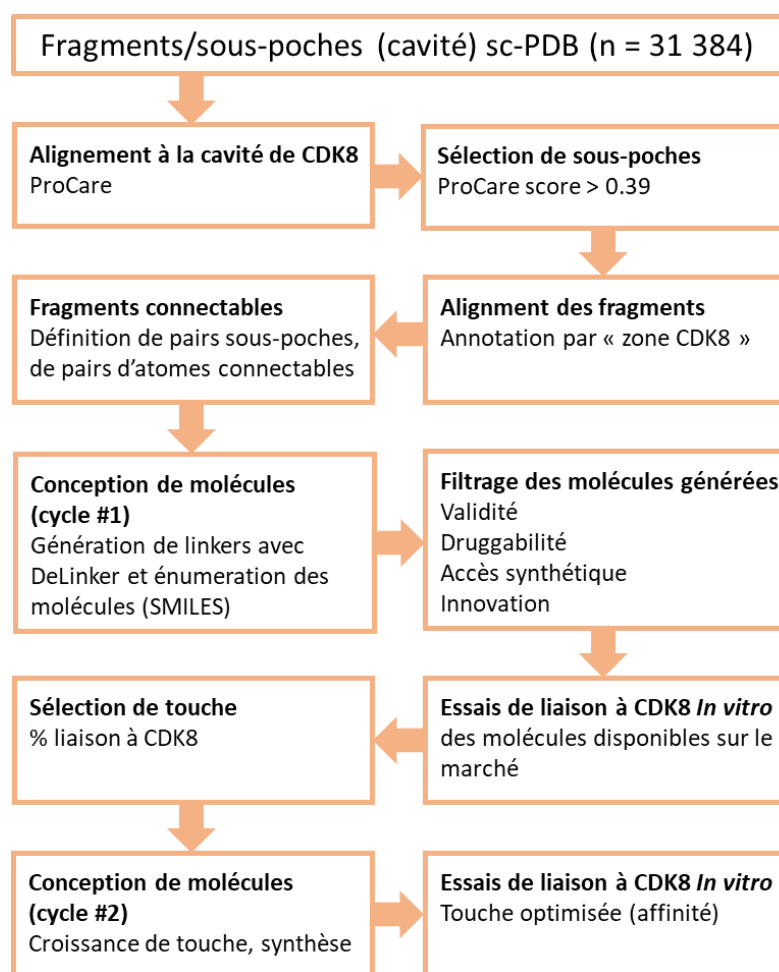
**Figure 6**. La méthode POEM (Pocket-Oriented Elaboration of Molecule) pour concevoir une chimiothèque focalisée. La preuve de concept a été appliquée à la protéine kinase dépendante des cyclines 8 (CDK8).

L'application de POEM à la protéine kinase dépendante des cyclines 8 (CDK8) a conduit à l'identification de molécules similaires à des inhibiteurs connus, mais surtout à de nouveaux inhibiteurs d'affinité micromolaire, voire nanomolaire pour les meilleurs d'entre eux (**Figure 7**), avec un taux de touches de 16%.
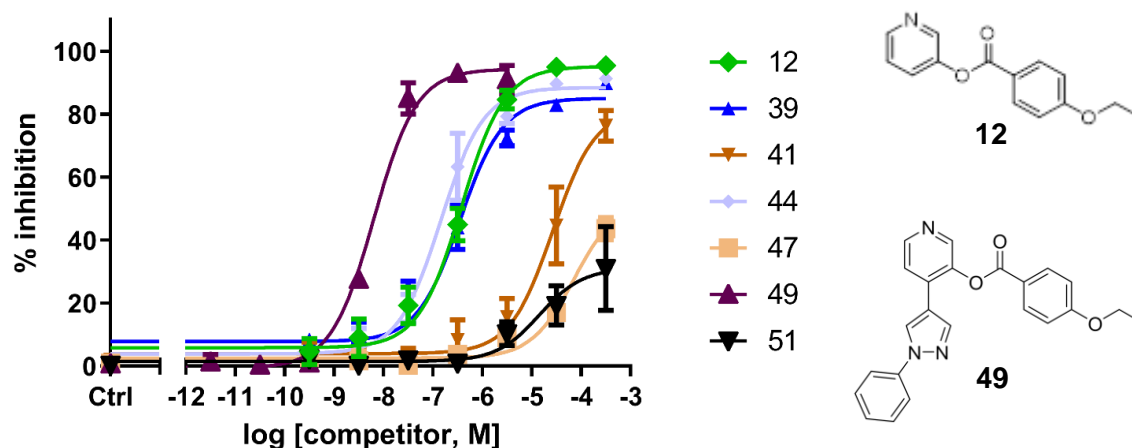
**Figure 7.** Inhibition de CDK8 par 7 molécules générées par POEM. Les courbes dose-réponse sont dérivées de trois expériences de compétition (TR-FRET, Fluorescence en temps résolu) indépendantes avec duplicatas par expérience. Les molécules **12** (issu du cycle #1, **Figure 6**) et **49** (cycle #2) ont respectivement une affinité (IC$_{50}$) de 376 nM et 6.4 nM.

Ces molécules ont été générées à partir de fragments aussi bien dérivés de complexes avec des protéines kinases que de complexes avec des protéines non-kinases, démontrant la capacité de la méthode à transposer des fragments pertinents en opérant dans tout le protéome structural connu. L'application à d'autres cibles thérapeutiques (quinolinate synthase NadA, domaine WD40 de la leucine rich-repeats kinase 2 LRRK2) a permis d'améliorer le protocole (positionnement et regroupement des fragments, atomes connectables) mais aussi d'identifier les limites de l'approche. Les résultats des essais biologiques de ces deux dernières applications sont attendus prochainement de nos collaborateurs.

## 2.4. Alignement de petites molécules à des cavités de protéines

La comparaison des nuages de points de cavités à des petites molécules, sur la base de règles pharmacophoriques et topologiques simples peut être une alternative intéressante au docking si elle génère des hypothèses orthogonales. Nous avons exploré et développé différentes approches pour superposer des petites molécules à des nuages points de cavités protéiques, puis les classer (scoring) par complémentarité décroissante : (1) implémentation d'un modèle pharmacophorique des molécules afin de les rendre comparables aux points de cavités, (2) développement de modèles de nuage de points des petites molécules pour une utilisation avec ProCare, (3) développement d'algorithmes d'alignement de graphes cavité-molécule, (4) développement d'une autre représentation de la cavité afin de contourner

les bruits des cavités VolSite, tout en respectant les contraintes de temps de calculs pour rester compétitif avec les méthodes existantes. Les résultats suggèrent que la recherche et l'estimation d'alignement rigide telle qu'implémentée ne sont pas efficaces pour résoudre ce problème, les performances restant inférieures à celles de méthodes de docking (**Figure 8**).[15] Cependant, ils montrent également que les jeux de représentations de cavités protéiques et de ligands contiennent parfois des informations riches, exploitables à des fins de classification.
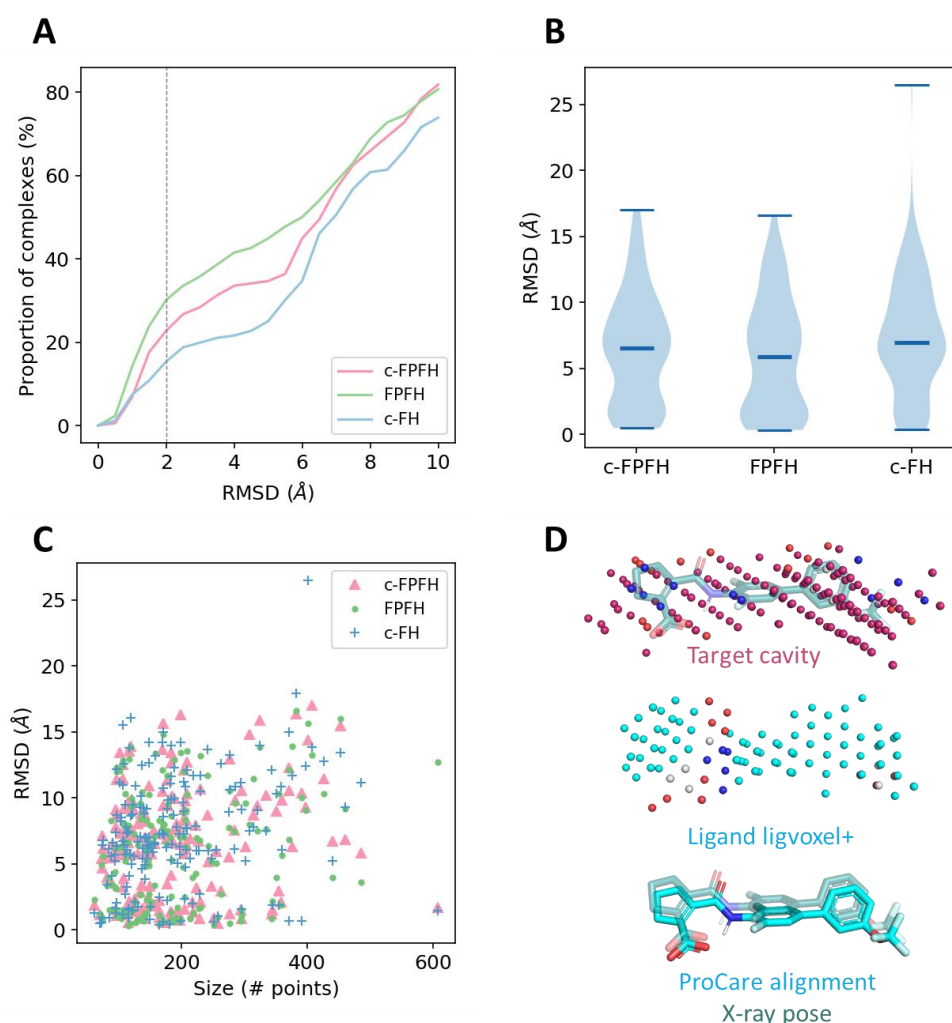


**Figure 8**. Alignement de 176 ligands de la sc-PDB sur leurs cavités correspondantes par comparaison de nuages de points. Trois descripteurs FPFH (forme), c-FH (forme et propriétés pharmacophoriques) et c-FPFH (hybride des deux précédents) sont utilisés. A) Pourcentage cumulatif de ligands alignés en deçà d'un certain seuil de déviation (RMSD) par rapport à la position du ligand déterminé par rayons X. B) Distribution en tracé de violon, montrant une RMSD médian d'environ 6 Å. C) RMSD des ligands en fonction du nombre de points dans la cavité protéique. D) Example d'alignement de l'entrée PDB 2FPT donnant une RMSD de 0.94 Å.

## 2.5. Apprentissage automatique des points de cavités pertinents

Identifier les points de cavités pertinents permettrait plusieurs applications directes en CAO : amélioration des comparaisons/alignements des petites molécules/cavités protéiques, priorisation de touches en criblage virtuel, interprétation de résultats d'activités. Nous avons conçu des modèles d'apprentissage pour discriminer les points pertinents des points non-pertinents, capable d'opérer sur de larges nuages de points de cavités, même en l'absence de ligands connus. Les descripteurs représentent la densité pharmacophorique dans des sphères concentriques, l'enfouissement et la distance au centroïde. Les points sont annotés en deux classes, selon leur distance et la compatibilité pharmacophorique avec les atomes du ligand qui interagissent avec la cible : les points *importants* (classe positive) sont situés à moins de 2 Å d'un atome du ligand de même propriété pharmacophorique, tout autre point est de classe négative. Les données sont ensuite équilibrées en jeux d'apprentissage (~450 000 points), d'évaluation externe (~150 000 points), puis d'application externe (1000 cavités). Les résultats préliminaires montrent que les modèles individuels pour chaque type pharmacophorique se généralisent mieux qu'un modèle global et permettent d'élaguer 60% des points négatifs tout en conservant les points positifs (**Figure 9**). Ces résultats sont encourageants pour des études plus approfondies.
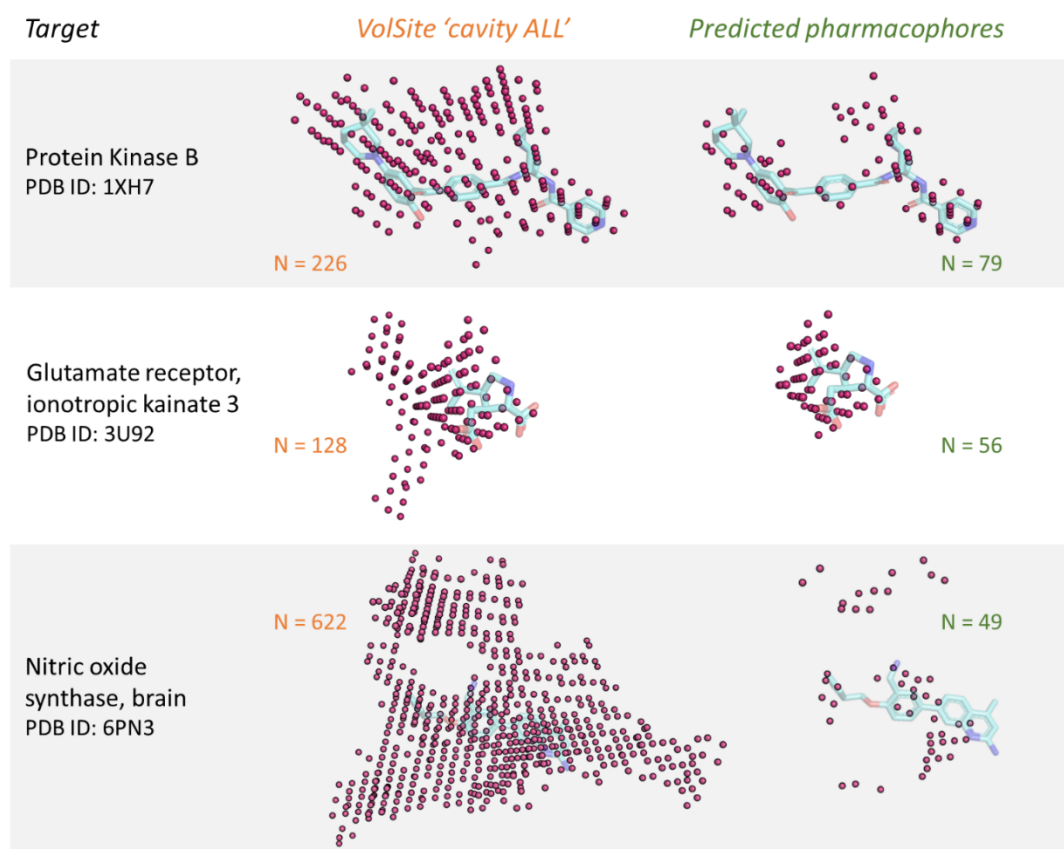


**Figure 9.** Prédiction des points importants des cavités protéiques. Les deux premiers exemples montrent une bonne délimitation des points autour des ligands, le dernier exemple une mauvaise délimitation.

# 3. Conclusion générale

A travers les travaux présentés dans cette thèse, nous avons proposé de nouvelles approches computationnelles pour la conception de molécules bioactives, en exploitant les cavités protéiques disponibles et représentées sous forme de nuage de points. Les projets ont été progressivement construits pour résoudre plusieurs problèmes : (1) estimation de la similarité des cavités protéiques à l'échelle du protéome structural et leurs applications prospectives à (2) la prédiction de cibles secondaires et (3) la conception de chimiothèques focalisées, (4) la comparaison de ligands aux cavités protéiques, (5) la prédiction des points de cavité en interaction (**Figure 2**).

La revue des méthodes existantes a révélé les difficultés de la comparaison des cavités protéiques et le besoin de méthodes permettant la comparaison de micro-environnements protéiques. En développant ProCare à cette fin, nous avons montré que traitement de nuages de points basé sur l'échantillonnage, appliqué à l'origine à d'autres tâches de la vision par ordinateur, peut identifier des motifs communs entre des sous-poches de protéines non apparentées. A partir des premières validations rétrospectives, nous avons procédé à l'évaluation de notre méthode en confrontant les prédictions computationnelles aux validations expérimentales. Ainsi, nous avons pu identifier une similarité locale entre les sites de liaison de deux protéines fonctionnellement et structurellement différentes, la cytokine facteur de nécrose tumorale alpha (TNF-α) et la transcriptase inverse (RT) du VIH-1. La mesure directe de la liaison *in vitro* a montré que deux inhibiteurs non nucléosidiques du RT-VIH-1 interagissent avec le trimère TNF-α avec une affinité comparable à un résultat de criblage à haut débit. De plus, nous avons développé une méthode, POEM, pour concevoir une chimiothèque focalisée de petites molécules, basée sur la prédiction de similarité de sous-poches. En appliquant POEM à la kinase dépendante des cyclines 8 (CDK8), nous avons réussi à concevoir un nouveau ligand nanomolaire en seulement deux étapes. Enfin, l'évaluation de POEM sur des cibles orphelines (quinolinate synthase, domaine WD40 de la leucine-rich repeat kinase 2), pour lesquelles aucun ligand pharmacologique n'est connu à ce jour, permet d'améliorer le workflow tout en proposant un défi à l'aveugle et en permettant d'identifier les limites de l'approche.

La représentation des cavités protéiques sous forme de nuages de points occupant tout l'espace des ligands offre l'avantage de développer des méthodes informatiques pour le criblage de petites molécules. Dans cette lancée, nous avons étudié l'alignement des nuages de points et de graphes des ligands aux cavités protéiques. Les informations contenues dans les nuages de la cavité se sont avérées riches pour être comparées à de petites molécules mais insuffisante pour générer de bons alignements, c'est pourquoi des modèles d'apprentissage automatique ont été développés pour prédire les points importants correspondant aux pharmacophores des ligands. Ces résultats sont encourageants et ont suggéré d'autres analyses pour approfondir ces études. Enfin, nous sommes intrigués par l'application de ces concepts à d'autres classes cibles.

Pour conclure, nous espérons que les nouvelles contributions de cette thèse par rapport à l'état de l'art ont fourni des informations utiles dans le cadre général de la conception de molécules assistée par ordinateur. Les diverses évaluations entreprises dans ces travaux de recherche nous ont suggéré des pistes d'améliorations, qui feront l'objet de travaux futurs.

## 4. Références

1. Lengauer, T.; Rarey, M. Computational Methods for Biomolecular Docking. *Curr. Opin. Struct. Biol.* **1996**, *6*, 402–406.

2. Lyu, J.; Wang, S.; Balius, T. E.; Singh, I.; Levit, A.; Moroz, Y. S.; O'Meara, M. J.; Che, T.; Algaa, E.; Tolmachova, K.; Tolmachev, A. A.; Shoichet, B. K.; Roth, B. L.; Irwin, J. J. Ultra-Large Library Docking for Discovering New Chemotypes. *Nature* **2019**, *566*, 224–229.

3. John Harris, C.; D. Hill, R.; W. Sheppard, D.; J. Slater, M.; F.W. Stouten, P. The Design and Application of Target-Focused Compound Libraries. *Comb. Chem. High Throughput Screen.* **2011**, *14*, 521–531.

4. Bhagavat, R.; Sankar, S.; Srinivasan, N.; Chandra, N. An Augmented Pocketome: Detection and Analysis of Small-Molecule Binding Pockets in Proteins of Known 3D Structure. *Structure* **2018**, *26*, 499-512.e2.

5. Ehrt, C.; Brinkjost, T.; Koch, O. Impact of Binding Site Comparisons on Medicinal Chemistry and Rational Molecular Design. *J. Med. Chem.* **2016**, *59*, 4121–4151.

6. Kalliokoski, T.; Olsson, T. S. G.; Vulpetti, A. Subpocket Analysis Method for Fragment-Based Drug Discovery. *J. Chem. Inf. Model.* **2013**, *53*, 131–141.

7. Desaphy, J.; Azdimousa, K.; Kellenberger, E.; Rognan, D. Comparison and Druggability Prediction of Protein–Ligand Binding Sites from Pharmacophore-Annotated Cavity Shapes. *J. Chem. Inf. Model.* **2012**, *52*, 2287–2299.

8. Rusu, R. B.; Cousins, S. 3D Is Here: Point Cloud Library (PCL). In *2011 IEEE International Conference on Robotics and Automation*; IEEE, 2011; Vol. I, pp 1–4.

9. Eguida, M.; Rognan, D. A Computer Vision Approach to Align and Compare Protein Cavities: Application to Fragment-Based Drug Design. *J. Med. Chem.* **2020**, *63*, 7127–7142.

10. Zhou, Q.-Y.; Park, J.; Koltun, V. Open3D: A Modern Library for 3D Data Processing. *arXiv:1801.09847* **2018**.

11. O'Connell, J.; Porter, J.; Kroeplien, B.; Norman, T.; Rapecki, S.; Davis, R.; McMillan, D.; Arakaki, T.; Burgin, A.; Fox III, D.; Ceska, T.; Lecomte, F.; Maloney, A.; Vugler, A.; Carrington, B.; Cossins, B. P.; Bourne, T.; Lawson, A. Small Molecules That Inhibit TNF Signalling by Stabilising an Asymmetric Form of the Trimer. *Nat. Commun.* **2019**, *10*, 5795.

12. Eguida, M.; Rognan, D. Unexpected Similarity between HIV-1 Reverse Transcriptase and Tumor Necrosis Factor Binding Sites Revealed by Computer Vision. *J. Cheminform.* **2021**, *13*,

1–13.

13. Imrie, F.; Bradley, A. R.; van der Schaar, M.; Deane, C. M. Deep Generative Models for 3D Linker Design. *J. Chem. Inf. Model.* **2020**, *60*, 1983–1995.

14. Ertl, P.; Schuffenhauer, A. Estimation of Synthetic Accessibility Score of Drug-like Molecules Based on Molecular Complexity and Fragment Contributions. *J. Cheminform.* **2009**, *1*, 1–11.

15. Cross, J. B.; Thompson, D. C.; Rai, B. K.; Baber, J. C.; Fan, K. Y.; Hu, Y.; Humblet, C. Comparison of Several Molecular Docking Programs: Pose Prediction and Virtual Screening Accuracy. *J. Chem. Inf. Model.* **2009**, *49*, 1455–1474.

# Publications and communications

## Journal articles

1. Eguida, M.; Valencia, C.; Hibert, M.; Villa, P. and Rognan, D. Target-focused library design by pocket-applied computer vision and fragment deep generative linking.
*Journal of Medicinal Chemistry.* **2022**, 65, 13771-13783.
https://doi.org/10.1021/acs.jmedchem.2c00931

2. Eguida, M. and Rognan, D. Estimating the similarity between protein pockets. (Review)
*International Journal of Molecular Sciences*. **2022**, 23, 12462.
https://doi.org/10.3390/ijms232012462

3. Eguida, M. and Rognan, D. Unexpected Similarity between HIV-1 Reverse Transcriptase and Tumor Necrosis Factor Binding Sites Revealed by Computer Vision.
*Journal of Cheminformatics*. **2021**, 13, 1–13.
https://doi.org/10.1186/s13321-021-00567-3

4. Eguida, M. and Rognan, D. A Computer Vision Approach to Align and Compare Protein Cavities: Application to Fragment-Based Drug Design.
*Journal of Medicinal Chemistry.* **2020**, 63, 7127–7142.
https://doi.org/10.1021/acs.jmedchem.0c00422

## Conference talks

1. Merveille Eguida and Didier Rognan. Subpocket similarity-based assembly of fragments: Applications to focused library design for CDK8 Inhibition.
*American Chemical Society Spring meetings*, Virtual, March 20 – 24th, 2022.

2. Merveille Eguida* and Marcel Hibert. Rational design of kinase inhibitors.
*French Association of Diamond-Blackfan Anemia*, Toulouse, October 29th – November 1st, 2021. (* co-presenter)

3. Merveille Eguida and Didier Rognan. Focused library design via fragment-bound subpocket alignment and deep generative linking: a proof-of-concept for CDK8 inhibitors.
*10th conference of the French Society of Cheminformatics & 22nd Graphism Group and Molecular Modeling*, Lille, September 29th – October 1st, 2021.

4. Merveille Eguida and Didier Rognan. Unexpected similarity between TNF-α and HIV-1 reverse transcriptase binding sites revealed by a novel 3D computer vision-inspired method.
*Illkirch Campus science Day JCI*, virtual, April 19th, 2021.

5. Merveille Eguida and Didier Rognan. ProCare – A computer vision approach to align protein cavities. Application to fragment-based drug design.
*Doctoral School of Chemical Sciences Day*, virtual, November 12th, 2020.

6. <u>Merveille Eguida</u> and Didier Rognan. ProCare – A computer vision approach to align protein cavities. Application to fragment-based drug design.
   16th *German Conference on Cheminformatics and EuroSAMPL*, virtual, November 2 – 4th, 2020.

## Conference posters

1. <u>Merveille Eguida</u>, and Didier Rognan. Unexpected similarity between HIV-1 reverse transcriptase and TNF-α binding sites revealed by protein subpocket cloud comparison.
   *23rd European Symposium on Quantitative Structure-Activity Relationship*, Heidelberg, Germany, September 26 – 30th, 2022. **Poster award.**

2. <u>Merveille Eguida</u>, Christel Schmitt-Valencia, Marcel Hibert, Pascal Villa and Didier Rognan. Protein-applied computer vision and deep generative linking generate potent kinase inhibitors.
   *Summer School of Cheminformatics,* Strasbourg, June 27th – July 1st, 2022. **Poster award.**

3. <u>Merveille Eguida</u>, Christel Schmitt-Valencia, Marcel Hibert, Pascal Villa and Didier Rognan. Focused library design via fragment-bound subpocket alignment and deep generative linking: a proof-of-concept for CDK8 inhibitors.
   *17th German Conference on Cheminformatics,* Garmisch-Partenkirchen, Germany, Mai 8 – 11th, 2022. **Poster award.**

4. <u>Merveille Eguida</u> and Didier Rognan. Protein subpocket cloud comparison revealed similarity between HIV-1 reverse transcriptase and tumor necrosis factor binding sites.
   *17th German Conference on Cheminformatics,* Garmisch-Partenkirchen, Germany, Mai 8 – 11th, 2022.

5. <u>Merveille Eguida</u> and Didier Rognan. Unexpected similarity between TNF-α and HIV-1 reverse transcriptase binding sites revealed by a 3D computer vision-inspired method.
   *American Chemical Society Fall meetings*, virtual, August 22 – 26th, 2021.

6. <u>Merveille Eguida</u> and Didier Rognan. Focused library design via fragment-bound subpocket alignment and deep generative linking: a proof-of-concept for CDK8 inhibitors.
   *Interfacing Chemical Biology and Drug Discovery RICT*, virtual, July 7 – 9th, 2021.

7. <u>Merveille Eguida</u> and Didier Rognan. Unexpected similarity between unrelated protein binding sites revealed by a novel 3D computer vision-inspired method.
   *European Chemical Biology Symposium*, virtual, Mai 26 – 28th, 2021. **Poster award.**

# GENERAL INTRODUCTION

# General introduction

In our contemporary era, designing a drug molecule to treat a particular disease is a long and costly process from the earlier generation of hypotheses to the distribution on the market. It takes on average 20 years, two billion US dollars,[1] thousands of scientists, operators, and participants, many failures[2] and one success to safely bring solutions to patients. In the early stages of the pharmaceutical industry, drugs were extracted from natural sources according to prior observations to treat symptoms or have been discovered accidentally.[3] The technological progress together with the accumulation of knowledge have enabled to adopt various strategies to characterize targets and find starting bioactive molecules on a rational basis while controlling the safety and costs. Many of these targets are proteins, one of the major building blocks that compose living organisms.[4] Proteins regulate biological processes by interacting with other molecules at specific areas on their surfaces.[5] Thus, it was discovered that inhibiting or activating key proteins involved in biological pathways relevant to a particular disease could restore a heathier function.[6] For more than a century, this was largely achieved by small molecular weight molecules. In 2021, 72% of FDA-approved drugs were new chemical entities.[7]

Before they ever reach clinical trials, drug candidates go through tedious "design-make-test-analyze" (DMTA) cycles to meet desired pharmacological and non-toxicity profiles, but the very beginning of this process is the identification of hit molecules that sufficiently interact with the target.[8] By accessing models of proteins three-dimensional structures thanks to advances in genomics and structural biology, it was shown that small molecules preferentially bind to buried cavities.[9] From then on, computational methods to model protein-small molecule interactions have flourished. The most popular, docking,[10] supports the screening of millions of molecules from well-thought virtual libraries to propose a few that have higher chances to bind in experimental assays.[11] Alternatively, methods which focus on assessing the resemblance of protein interaction sites quickly emerged and gain popularity in the first decade of this century.[12] This strategy is notably relevant now as the structural data on diverse proteins and the binding information on several molecules are constantly increasing.[13] Pure protein cavities comparison operates in the target space only, therefore is thought to provide at least a different perspective, at best an advantage against the combinatorial complexity of protein-ligand information and scoring problems known to docking.[10] When cavities of different targets are found similar, binding knowledge are hypothetically transferred to identify secondary targets, to design ligands or focused libraries for virtual screening.[14]

My host laboratory has contributed to the state-of-the-art binding site detection and comparison methods in the past two decades.[15–17] One of these methods (VolSite)[17] detects pockets in proteins irrespective of prior bound ligand coordinates and represents them as a cloud of points featuring a negative image of the cavity. Thus, it enables to reach previously non-characterized pockets, or those which prove to be difficult for classical approaches (small or large cavities). Then, another tool (Shaper) is used to compare

these clouds to estimate the similarity between two protein cavities.[17] Shaper is based on a commercial and proprietary toolkit from OpenEye Scientific Software (Santa Fe, USA), which performs global shape and property matching of two cavity clouds. Shaper have achieved good performance in evaluations, which validated the information carried by VolSite cavities. However, two aspirations have led to my dissertation:

- the access to a non-proprietary method to estimate the similarity of VolSite cavities,
- the exploration of pattern recognition methods used in image processing.

In **Chapter 1**, a review of previously published methods showed a diversity in how protein cavities are represented, compared and the similarity scored. Yet, the majority perform global searches for resemblance which might hinder the detection of subtle but relevant similarities at times. Therefore, the first part of my work consisted in identifying and implementing suitable algorithms to compare VolSite clouds, while striving for the following specifications:

- the possibility to estimate both global and local similarities,
- a computing time compatible with screening large databases on a daily basis,
- the interpretability of the results.

This led to the development and retrospective evaluation of a novel tool (ProCare), presented in **Chapter 2.** During the evaluation of ProCare on the tumor necrosis factor-alpha (TNF-α) protein, I observed a common pattern between the TNF-α trimer interface and the cavity of reverse transcriptase non-nucleoside inhibitors. The resulting similarity hypothesis was investigated in **Chapter 3**. In the same pursuit of providing a realistic assessment to the ProCare method, I designed a workflow for generating target-focused libraries using fragment moieties bound to subpockets that were locally estimated similar to the target cavity (**Chapter 4**). Finally, as a continuation of my laboratory goal to find alternative screening methods, I have explored the search of common patterns between VolSite cavities and small molecules in **Chapter 5**.

# References

1.  Wouters, O. J.; McKee, M.; Luyten, J. Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018. *JAMA* **2020**, *323*, 844.

2.  Kola, I.; Landis, J. Can the Pharmaceutical Industry Reduce Attrition Rates? *Nat. Rev. Drug Discov.* **2004**, *3*, 711–716.

3.  Pina, A. S.; Hussain, A.; Roque, A. C. A. An Historical Overview of Drug Discovery. In *Methods in Molecular Biology*; **2010**; Vol. 572, pp 3–12.

4.  Santos, R.; Ursu, O.; Gaulton, A.; Bento, A. P.; Donadi, R. S.; Bologa, C. G.; Karlsson, A.; Al-Lazikani, B.; Hersey, A.; Oprea, T. I.; Overington, J. P. A Comprehensive Map of Molecular Drug Targets. *Nat. Rev. Drug Discov.* **2017**, *16*, 19–34.

5.  Alberts, B.; Johnson, A.; Lewis, J.; Walter, P.; Raff, M.; Roberts, K. *Molecular Biology of the Cell 4th Edition*; Routledge, **2002**.

6.  Drews, J. Drug Discovery: A Historical Perspective. *Science* **2000**, *287*, 1960–1964.

7.  de la Torre, B. G.; Albericio, F. The Pharmaceutical Industry in 2021. An Analysis of FDA Drug Approvals from the Perspective of Molecules. *Molecules* **2022**, *27*, 1075.

8.  Hughes, J.; Rees, S.; Kalindjian, S.; Philpott, K. Principles of Early Drug Discovery. *Br. J. Pharmacol.* **2011**, *162*, 1239–1249.

9.  Liang, J.; Woodward, C.; Edelsbrunner, H. Anatomy of Protein Pockets and Cavities: Measurement of Binding Site Geometry and Implications for Ligand Design. *Protein Sci.* **1998**, *7*, 1884–1897.

10. Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and Scoring in Virtual Screening for Drug Discovery: Methods and Applications. *Nat. Rev. Drug Discov.* **2004**, *3*, 935–949.

11. Rognan, D. The Impact of in Silico Screening in the Discovery of Novel and Safer Drug Candidates. *Pharmacol. Ther.* **2017**, *175*, 47–66.

12. Kellenberger, E.; Schalon, C.; Rognan, D. How to Measure the Similarity Between Protein Ligand-Binding Sites? *Curr. Comput. Aided-Drug Des.* **2008**, *4*, 209–220.

13. Bhagavat, R.; Sankar, S.; Srinivasan, N.; Chandra, N. An Augmented Pocketome: Detection and Analysis of Small-Molecule Binding Pockets in Proteins of Known 3D Structure. *Structure* **2018**, *26*, 499-512.e2.

14. Ehrt, C.; Brinkjost, T.; Koch, O. Impact of Binding Site Comparisons on Medicinal Chemistry and Rational Molecular Design. *J. Med. Chem.* **2016**, *59*, 4121–4151.

15. Schalon, C.; Surgand, J. S.; Kellenberger, E.; Rognan, D. A Simple and Fuzzy Method to Align and Compare Druggable Ligand-Binding Sites. *Proteins Struct. Funct. Genet.* **2008**, *71*, 1755–1778.

16. Weill, N.; Rognan, D. Alignment-Free Ultra-High-Throughput Comparison of Druggable Protein-Ligand Binding Sites. *J. Chem. Inf. Model.* **2010**, *50*, 123–135.

17. Desaphy, J.; Azdimousa, K.; Kellenberger, E.; Rognan, D. Comparison and Druggability Prediction of Protein–Ligand Binding Sites from Pharmacophore-Annotated Cavity Shapes. *J.*

*Chem. Inf. Model.* **2012**, *52*, 2287–2299.

# CHAPTER 1

On the quest for estimating the similarity
between protein pockets

This Chapter was adapted and published in:

Merveille Eguida and Didier Rognan. *Int. J. Mol. Sci.* 2022, 23, 12462.

## 1.1. Introduction

In living organisms, biological processes are regulated through specific molecular recognition at local surfaces. Proteins, one of the major biomolecules composing our cells, interact with different partners: other proteins, peptides, nucleic acids, small molecules, transition metals. Proteins are made of amino acids chains, which spatially fold into particular shapes. To explore the proteome, sequence-based studies benefit from the boom of genomics since the early 2000, but their scope are quickly limited by the conservation of structure in proteins sharing less than 30% sequence homology.[1] Progress in molecular and structural biology have enabled to solve the three-dimensional (3D) structure of proteins, either by X-ray diffraction,[2–4] nuclear magnetic resonance (NMR)[5] or more recently cryo-electron microscopy (cryo-EM) at atomic scale.[6–9] Characterizing the binding cavities for small molecules have bolstered the rise of structure-based drug design.[10–12]

With the exponential increase of publicly-available protein structures,[13,14] coupled to the development of methods able to detect cavities,[15,16] the comparison of protein binding sites emerged naturally as a scientific topic to explain observations or generate hypothesis for ligand design or target fishing in drug design.[11] Possible applications span biological function prediction in bioinformatics to polypharmacology in medicinal chemistry.[17,18] Supported by the outlooks and successful case studies, many methods have been developed in the last three decades. The bottleneck of protein cavity comparison is common to all similarity estimation problems—similarity is a relative quantity which depends on the aspects considered. Therefore, generalizing a similarity quantification on different pairs of entries, without prior knowledge of the key points to compare is delicate.

Similarity is not directly measurable experimentally. Instead, derived hypotheses (e.g. function, ligand binding) are further evaluated. This presents many challenges for benchmarking methods and highlights the importance of carefully designing datasets in retrospective studies. For users as well as developers, knowing where we start from and what has been done in the field would enable realistic expectations and spot limitations to be addressed by future developments.

Structure-based algorithms for protein site comparison emerged after the 1970s, a decade marked by the establishment of the Protein Data Bank (PDB) and the deposit of a few structures.[13,14,19] Initially, efforts were made to compare protein 3D structural motifs independently of sequence order and gaps. Computer vision approaches[20] were applied in structural biology for similar substructure identification even in the absence of sequence homology via rigid body alignments.[21–27] Protein functions could be predicted from a database of known 3D templates, by querying or inferring protein active sites.[28–32] Beyond functional annotations, cavity alignment and comparison quickly appeared promising for rational design of proteins

and ligands, since similar 3D arrangement of surface motifs may be similarly involved in molecular recognition.[31,33]

The path from the earlier to the current site comparison methods involved several implementations. It was common for the user to define researched features (e.g. set of atom/residues distances defining a motif: catalytic triads, similar ligands) from prior knowledge to initialize the search.[29,30,34,35] Subsequent advantages are a better control of the comparison, easier selection of relevant matches, and the reliability of the solutions. Progressively, methods enabling automatic identification of pockets[36–40] and of relevant patterns that are matching opened the doors to the analysis of the relationship between evolutionarily and structurally remote members of an entire database, without any *a priori* judgment.[41–45] Such predictions led to unexpected findings with implications for drug design.[18,46] Screening large databases require effective computing time. Together with the progress of computing technologies, fast methods were introduced but often at the cost of interpretability.[47–49]

The repertoire of possible comparison algorithms is tailored to the representation made of the pocket.[50] Pocket representation is a way to provide structured information to the algorithm, for exploration. Once delimited in the protein, a pocket can be modeled as list of residues, graphs, or unconnected pseudo atoms among other possibilities. Geometry constraints of alpha carbon tuples were extensively used to identify equivalenced areas.[51–53] Other cavity descriptors further encode the chemical properties of atoms or residues, hence reducing redundancy in the possible matches.[41,54,55] The intricacy of the representation lays in finding a good balance between fuzziness with a risk of false positive matches and preciseness with a risk of missing on remote similarities. In any case, similarity can only be properly reported with a fair scoring function. The scoring scheme aims at quantifying how two pockets resemble or differ. Often, a score threshold is applied in screening campaigns for decision making. How to assign the value of that threshold and assess the significance of that similarity is a genuine question raised by earlier studies.[47,56,57]

In practice, the variability of the pocketome (ensemble of all protein pockets) in terms of size, solvent accessibility, flexibility constitute obstacles to the performance of binding site comparison methods, as it is for other structure-based approaches.[11] It is perceived that comparing subpockets, instead of entire cavities might better handle the conformational variations, typically induced by ligand binding.[45,58–60] Noteworthy, the ability to detect local or global similarities is suitable for different purposes.

As the reader will notice, different parameters entail the success of protein cavity comparison, as discussed by previous articles.[18,61–64] In this review, we will provide a most recent and broad overview of all stages involved in pockets comparison, from the prediction of ligand binding sites, to the evaluation and prospective applications in drug design.

## 1.2. Pocket detection and druggability estimation

Identification of potential interaction sites is crucial to structure-based approaches and constitute the very first step of binding site comparison. Proteins can specifically bind to different classes of molecule (proteins, peptides, nucleic acids, small molecules, transition metals). Contact surfaces exhibit different geometric and physicochemical characteristics according to the nature of the binding partner. For examples, small molecule interaction sites are buried clefts while protein-protein interaction interfaces are rather flat and hydrophobic.[12,65–68] Although available methods for binding site detection covers the different applications above, they majorly concern small molecule pocket identification as a testimony of efforts to structure-based drug design of small chemical entities in the last decades. Accessibility to binding site identification is possible via standalone tools,[69] webservers,[70] or databases of precomputed sites.[71]

Methods can be classified at three levels: (i) the genomic or 3D structure nature of the input, (ii) the dependency to bound ligands and (iii) the class of the algorithm (**Figure 1.1**). Template or sequence-based methods such as ConSeq,[72] available from the ConSurf server[71,73] identifies functionally important residues in protein sequences by searching for evolutionary relations with other proteins.[74–77] 3DLigandSite is another approach which can take a protein sequence input, although it relies on homology models or *de novo* structure predictions.[78] Structure-based pocket identification uses only the 3D coordinates of structures as input and benefits from the augmentation of structural data[14].

Ligand-centric methods are restricted to protein-ligand complexes and is rather a site delimitation than prediction. Noticeably, the analysis of crystallization additives binding sites might suggest potential allosteric pockets.[79] Typically, a site is defined as all residues within a certain distance cutoff to the partner's heavy atoms, ca. 6 Å for protein-small molecule complexes. Alternatively, the set of residues can be restricted to those properly oriented and toward the ligand, with the particularity that the distance cutoff varies according to the interaction type. These approaches are available through integrated environments enabling to manipulate protein structure coordinates and interactions such as Molecular Operating Environment (Chemical Computing Group, Montreal, Canada), IChem[80], independent tools for parsing protein 3D structure data.
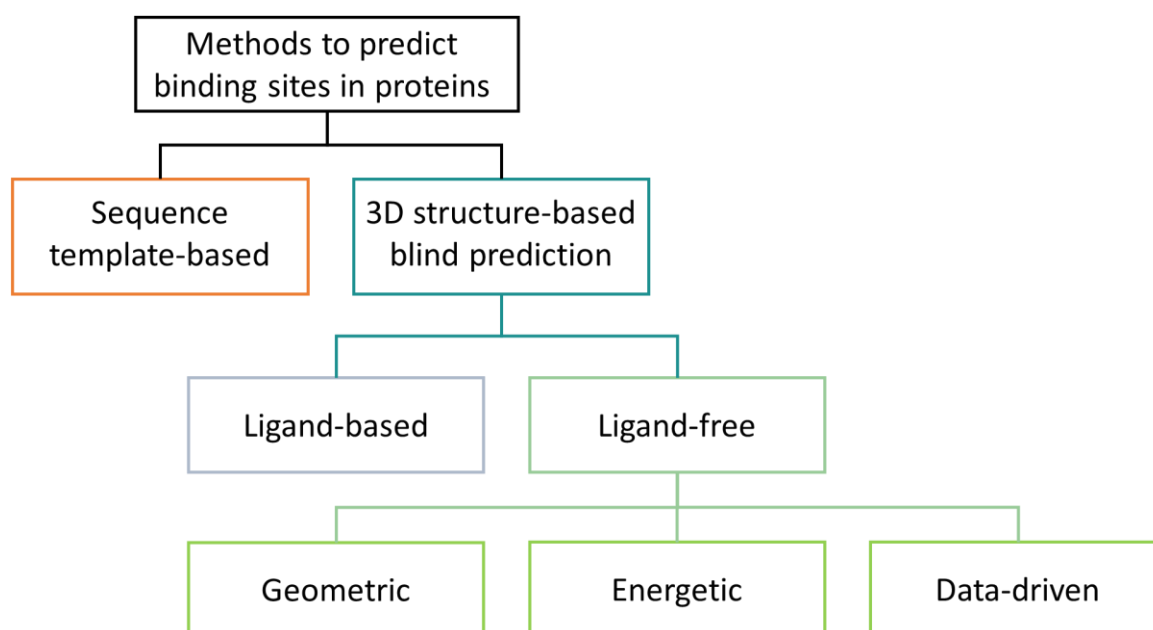
**Figure 1.1.** Classification of binding site detection methods.

Ligand-free approaches can operate on a larger range of structures, enabling the discovery of unprecedented sites. According to their search algorithm, they can be classified as geometric, energetic, or data-driven (**Table 1.1**). At first glance, all geometric methods aim at identifying sufficiently buried zones unoccupied by protein atoms, but differ in strategies to search for these areas. Grid-based methods place the protein into a cartesian grid and identify grid cells likely to be in a cleft by analyzing their neighborhood.[36,37,81–94] POCKET[37] and LIGSITE[89], two of the earliest methods, keep cells that correspond to a 'protein-solvent-protein' event by scanning respectively in three and seven directions. Such algorithms are sensitive to grid resolution and orientation but are powerful to detect cavities of different sizes and curvatures.

**Table 1.1.** Common structure-based methods to predict ligand binding pocket in proteins.

| Category | Search approach | Methods |
|---|---|---|
| Geometric | Grid | CAVIAR,[85] PROcket,[84] KVFinder,[83] VolSite[82], DoGSite,[81] McVol,[94] ghecom,[93] VICE,[92] PocketDepth,[91] PocketPicker,[90] LIGSITE[csc89], CAVER,[88] LIGSITE,[36] VOIDOO,[87] POCKET[37] |
| | Alpha-shape | Fpocket,[40] CASTp,[95,96] CAST,[10] APROPOS,[97] |
| | Spherical probes | DEPTH,[98] Roll,[99] HOLLOW,[100] PHECOM,[101] Xie and Bourne,[102] SURFNET-ConSurf,[103] PASS,[104] HOLE,[105] SURFNET[69] |
| | Other | MSPocket,[106] SplitPocket[107] |
| Energetic | Grid | FTSite,[108] SiteMap,[109] SITEHOUND,[110] AutoLigand,[111] Q-SiteFinder,[112] PocketFinder,[113] DrugSite,[114] *pocket-finder* (Surflex protomol),[115] GRID[116] |
| | Spherical probes | dPredGB,[117] Morita *et al.*[118] |
| | Other | Gaussian Network Model[119] |
| Data-driven | Classical machine learning | GRaSP,[120] P2Rank,[39] MCSVMBs,[121] PRANK,[122] SCREEN[123] |
| | Deep learning | PoinSite,[124] DeepPocket,[125] PUResNet,[126] DeepSurf,[127] BiteNet,[128] Jiang *et al.*,[129] DeepSite, ISMBLab-LIG[130] |

Contrarily, other methods process the protein coordinates directly and are not affected by the grid initialization phenomena. Based on the alpha-shape concept introduced by Edelsbrunner *et al.*,[131] they circumvent protein cavities by connecting adequate adjacent Delaunay triangles via the 'discrete flow' method,[10,95–97,107] or by clustering alpha spheres to satisfy pocket descriptors (e.g. Fpocket).[40] Alternative purely geometric approaches fill or coat the protein with spherical probes to delimit cavity void.[69,98–105] Finally, other concepts such as monitoring the direction of surface normal vectors were implemented.[106]

The second category of ligand-free methods estimate favorable surfaces for protein-ligand contacts by calculating the potential energy of probes at different positions. Generally, the Lennard-jones potentials[132,133] are used with hydrophobic probes. The nature and number of probes vary from a simple carbon probe in DrugSite[114] to 16 different in FTSite[108]. Potentials are either mapped to grid positions[108–116] or to probe coating the protein surface.[117,118] GRID, a very popular grid-based approach, has implemented an empirical force field to estimate van der Waals, electrostatic and hydrogen-bonding energies for 6 different probes with predefined parameters.[116] Obviously, the outputs of energy based methods are influenced by the force field, in addition to the initialization for grid-based ones.

The final class of methods use supervised models, trained on the features of well characterized ligand binding sites. Hence, they differ in the features representation, training models, set of parameters and datasets. P2RANK is one of the examples based on classical machine learning models. The protein solvent-exposed atoms are processed into a topological and physicochemical feature vector which serve as input to a Random Forest classifier.[39] Recently, many deep learning methods, majorly based on 3D-convolutional neural networks were introduced. PointSite is an example of point clouds segmentation using sparse convolution.[124] While these methods need to be challenged by prospective usages, recent advances on 3D point cloud deep learning[134] offers some long perspectives for this type of problem.

All in all, these methods have been evaluated on their performance to accurately predict binding pockets by comparing predictions on unbound proteins to true ligand locations in their corresponding bound structures. Not only the accuracy of the location, but also the delimitation or overlap with respect to the ligand are analyzed.[81] Indeed, all identified clefts do not forcibly correspond to the ability to accommodate a drug-like ligand (druggability). Detected pockets might be too large, or too small where a clustering is required. Thus, it might be convenient to post-process the results of other approaches.[135] Cleverly, meta-methods (e.g., MetaPocket) thrive to find consensus from different algorithms to increase the chances of correct predictions.[136,137] However, consensus might not always yield the right solution.

The concept of structural druggability[138–141] arose from observing the characteristics of pockets bound to pharmacological ligands: average volume between 200 to 800$Å^3$, a good balance of hydrophobic and polar atoms enabling some binding specificity, sufficient buriedness. A few methods were developed to predict target druggability.[38,82,142–146] Consistently, topological and physicochemical characteristics of the pockets sites are encoded into descriptors and trained on curated datasets to generate classification models (Support Vector Machines, linear regression).[38,82,144,145] Since pocket druggability does not guarantee that the bound ligand will also be druggable, the term may be replaced by ligandability[147] or bindability.[114] For more information, we refer the reader to a recent review.[141] Interestingly, some of the methods previously described have implemented a rule-based druggability prediction enabling to hit two targets with one bullet.[38,82,109] VolSite, the tool developed in my host laboratory, is one of them.

## Zoom on VolSite

In VolSite,[82] (**Figure 1.2**) grid points are sampled by projecting 120 rays of equally-spaced solid angle and 8 Å length. Positions that yield at least 80 rays overlapping with cells close to or occupied by a protein atom are further considered. Points having a protein atom within 4 Å are labeled with a pharmacophoric feature complementary to the physicochemical property of the closest protein atom (h-bond acceptor, h-bond donor, h-bond acceptor/donor, negative ionizable, positive ionizable, hydrophobic and aromatic), otherwise a dummy property. Isolated points, i.e., having less than three adjacent grid points are discarded. Later, VolSite was adapted so that at least three hydrophobic protein atoms are required in the neighborhood to assign that property to a grid point.[80] While hydrophobic and aromatic features happen to cluster in patches, in reality, the rarest features (e.g. negative ionizable) are diluted among other features.
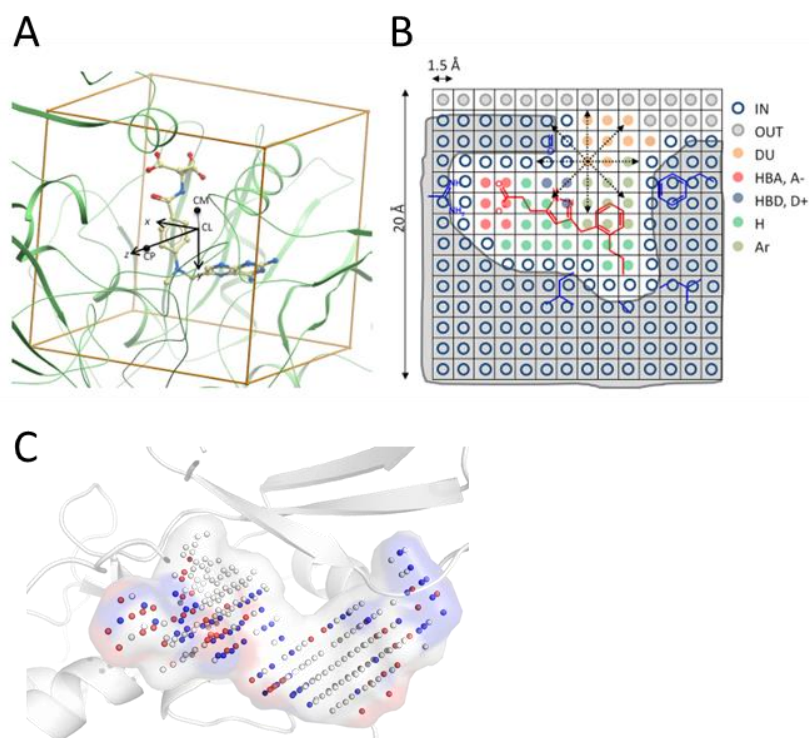


**Figure 1.2.** VolSite pocket detection. A) Grid initialization. B) Grid points can have one of the eight possible pharmacophoric points: h-bond acceptor HBA, h-bond donor HBD, h-bond acceptor and donor OG, negative ionizable A-, positive ionizable D+, hydrophobic H, aromatic Ar, dummy DU. C) Example of pockets detected in a kinase protein (PDB: 5HBH) by VolSite (molecular surface is depicted with PyMol 2.1, red points: HBA, A-, blue: HBD, D+, white: H, Ar, DU). A) and B) are adapted from Desaphy *et al.*[82]

VolSite has the particularity to output a cloud of points, occupying the volume of the cavity and not just the surface, therefore mimicking an ideal ligand (negative image of the cavity). It is therefore applicable to many structure-based scenarios ranging from ligand-binding site comparisons[82] (Chapter 2), secondary target identification[67] (Chapter 3), structure-based pharmacophore perception[148] (Chapter 5) and fragment-based library-design (Chapter 4).

In conclusion, we have seen in this section that methods to predict ligand pockets are diverse in the way they search and the features they consider. Predictions are subjected to uncertainties about the true delimitation of a ligand area and druggability, with implications for subsequent applications. In practice, some tools are specialized for predicting interaction sites with particular molecule classes: protein-protein interfaces,[67,149] nucleic acids,[150,151] peptides,[152] pores/channels,[153,154] phosphates.[155] In all cases, the output serves to delineate cavity-lining residues, and a few are directly processed by site comparison tools (e.g. DoGSite, LIGSITE, VolSite).

## 1.3. Steps for comparing cavities in proteins

Methods comparing protein cavities operate in three steps: the representation of the cavity characteristics, the comparison of these representations and finally a scoring or classification.[50,61,62] Hence, successful results reside in a coordinated performance of each of these tasks. Yet, cavity representation, which is the first step of the procedure is crucial as it influences the later steps. Principally, a poor representation where relevant characteristics are missing cannot be compensated by the most efficient algorithm. State-of-the art methods to compare protein cavities are summarized in **Table 1.2.** In the following sections, we will discuss these different algorithms to achieve this end.

**Table 1.2.** Methods to compare protein cavities.

| Year | Name | Detection[a] | Principle | Scoring | Evaluation datasets |
|---|---|---|---|---|---|
| 2002 | CavBase[41] | LIGSITE | Clique detection in graphs of pseudoatoms | Overlap of surface grid points, RMSD | Cofactor sites, kinases, serine proteases |
| 2002 | eF-site[156] | Ligand Databases | Clique detection in graph of surface normal vectors and electrostatic potentials | Normalized and weighed contributions of vectors angles, potentials, distances | Phosphate sites, antibodies, PROSITE classes |
| 2003 | SuMo[157] | Ligand | Incremental match of triplets of pseudocenters | Count of matches, RMSD, composite of euclidian and density distances | Protease catalytic sites, lectine sites |
| 2004 | SiteEngine[42] | Ligand | Match of triplets of points by hashing | Hierarchical scoring: count of matches, RMSD, overlap of patches, local shape | Cofactors, steroids, fatty acid sites, catalytic triad in proteases |
| 2004 | Brakoulias *et al.* (SiteBase)[158] | Ligand | Match of triplets of points | Count of matches, RMSD | Cofactors, phosphate sites |
| 2007 | Ramensky *et al.*[59] | Ligand | Clique detection in graph of atoms | Dice similarity of matches | Diverse |
| 2008 | Binkowski *et al.*[159] | CAST Ligand | Comparison of pairwise distance histograms | Kolmogorov-Smirnov divergence, overlap of volume, RMSD | Cofactor sites, HIV proteases |

[a] The site detection approaches used in the reference studies were reported. However, ligand-free methods might be employed depending on the input for the site comparison method.

**Table 1.2.** Methods to compare protein cavities (continued).

| Year | Name | Detection[a] | Principle | Scoring | Datasets |
|---|---|---|---|---|---|
| 2008 | PocketMatch[43] | Ligand | Comparison of sorted pairwise distances | Normalized count of matches | Diverse, SCOP[160] classes |
| 2008 | SiteAlign[44] | Ligand | Alignment of polyhedron fingerprints | Normalized distances of fingerprints | Functional groups, proteases, estrogen receptors, GPCRs |
| 2008 | SOIPPA[161] | Ligand | Clique detection in graphs of atoms | Composite weighted by frequencies, PSSM, distances | Cofactor sites, SCOP classes |
| 2009 | SMAP[56] | Ligand | Clique detection in graphs of atoms | Gaussian densities from distances, angles of normal vectors, BLOSSUM weights | Cofactor sites |
| 2010 | BSSF[48] | PASS Ligand | Comparison of fingerprints of binned distances and properties | Canberra distances of fingerprints | Diverse, synthetic data, SCOP classes |
| 2010 | Feldman *et al.*[53] | Ligand | Match of subsets of Cα atoms | Potential based on distances between matches | Diverses, kinases |
| 2010 | FuzCav[47] | Ligand | Fingerprints of triplets of atom features | Maximal proportion of matches | Diverse, functional groups, 8 difficult cases |
| 2010 | Milletti *et al.*[162] | Ligand | Comparison of 3 concentric spheres fingerprints encoding neighborhood for each point, solving linear assignment | Composite of fingerprint distances and RMSD | ATP sites, kinases |
| 2010 | P.A.R.I.S[163] (sup-CK) | Ligand | Initial alignment optimized by gradient ascent to maximize a Gaussian kernel | Gaussian kernel | Cofactor sites |

**Table 1.2.** Methods to compare protein cavities (continued).

| Year | Name | Detection[a] | Principle | Scoring | Datasets |
|---|---|---|---|---|---|
| 2010 | ProBiS[54] | Ligand | Maximum clique detection in graphs of surface atoms | Count of Matches, RMSD, angle between vectors | Cofactor/metal sites, protein-protein interfaces, protein-DNA complexes |
| 2011 | PocketAlign[164] | Ligand | Initial pairs from sorted lists of atom distances, then extend | Count of matches, RMSD | Cofactor sites, SCOP classes |
| 2011 | PocketFEATU-RE[165] | Ligand | Comparison of 7 concentric spheres fingerprints encoding neighborhood for each microenvironment | Normalized Tanimoto similarity of fingerprints | Kinases |
| 2012 | KRIPO[45] | Ligand | Fingerprints of triplets of pharmacophore | Modified Tanimoto of fingerprints | Diverse, search of bioisosteric substructures |
| 2012 | Patch-Surfer[166] | LIGSITE Ligand | Comparison of 3D Zernike of surface patches solving a weighted bipartite matching | Composite of surface match distances and size differences | Cofactor sites |
| 2012 | Shaper[82] | VolSite | Comparison of cloud of points by Gaussian shapes matching | Tanimoto, Tversky of matches | Diverse, GPCRs, proteases |
| 2012 | TIPSA[167] | Ligand | Match of quadruplets of points, iterative refinement by Hungarian algorithm | Tanimoto of matches, overlap of volume, normalized RMSD | Cofactor sites |

**Table 1.2.** Methods to compare protein cavities (continued).

| Year | Name | Detection[a] | Principle | Scoring | Datasets |
|------|------|-----------|-----------|---------|----------|
| 2013 | Apoc[51] | CAVITA-TOR,[51] LIGSITE Ligand | Seed alignment by comparing secondary structures, optimized by solving linear assignment problem | Composite of vector orientation, distance, properties | Diverse, similar ligand recognition sites |
| 2013 | TrixP[168] | DoGSite | Search for common shape and triplets of points by bitmap indexing | Composite of matches count, angle between vectors, mismatches penalty | Diverse, 8 difficult cases, protease, estrogen receptor, HIV reverse transcriptase |
| 2014 | *e*MatchSite[52] | *e*FindSite[169] | Template-based alignment optimized by Hungarian algorithm | Machine learning score: RMSD, residue, properties | Cofactors, steroid sites |
| 2014 | RAPMAD[49] | LIGSITE | Comparison of 14 pairwise distance histograms, one for each property | Jensen-Shannon divergence of histograms | Cofactor sites, proteases, diverse |
| 2015 | IsoMIF[170] | GetCleft[170] | Clique detection in graphs of interaction grid points | Tanimoto of descriptors of matched points | Cofactors, steroid sites |
| 2016 | G-LoSA[60] | Ligand | Clique detection in graphs of atoms | Feature-weighted count of matches | Diverse, Ca+ sites, similar ligands recognition sites, protein-protein interfaces |
| 2016 | SiteHopper[171,172] | Ligand | Comparison of surface atoms by Gaussian shapes matching | Weighted combination of Shape and color Tanimoto | Diverse using binding affinities |

**Table 1.2.** Methods to compare protein cavities (continued).
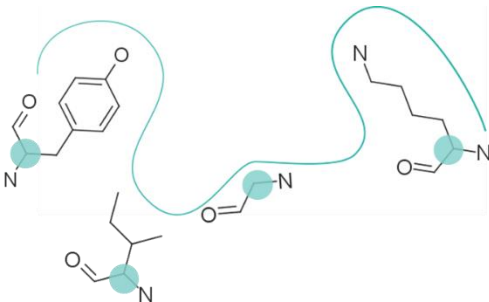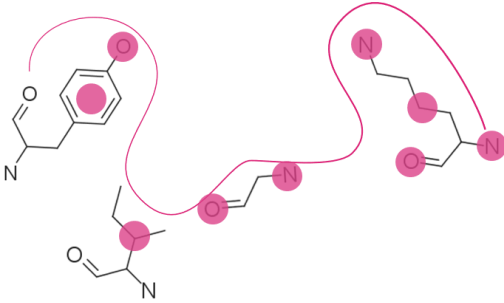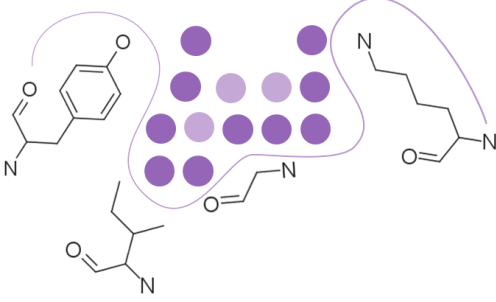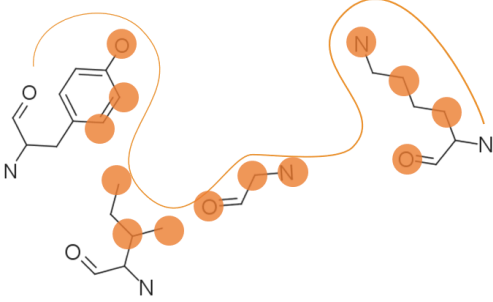
| Year | Name | Detection[a] | Principle | Scoring | Datasets |
|------|------|-----------|-----------|---------|----------|
| 2019 | DeepDrug3D[173] | Ligand | Convolutional neural network model | Binary classification | Cofactors, steroids sites, proteases |
| 2020 | DeeplyTough[55] | Fpocket Ligand | Convolutional neural network model | Binary classification | Cofactor sites, Diverse and using binding affinities |
| 2021 | PocketShape[174] | Ligand | Initial alignment optimized by Hungarian algorithm | Composite of matches, orientation of residues | Diverse SCOP classes, kinases |
| 2021 | Site2Vec[175] | Ligand | Machine learning (random forest) on autoencoder-generated descriptor | Binary classification | Cofactors, steroid sites, diverse |

## 1.3.1. Pocket representation

Once pockets are delimited, features are selected by considering different aspects. This step aims at focusing on the relevant characteristics that explain ligand recognition, while decreasing the so considered "unnecessary" information. Our brains will perform the same exercise on everyday life's objects, for example if we are asked to compare two cars: we might decompose the information into major aspects such as the brand, design, color, motor, etc. Interestingly, different people will focus on different combinations of these aspects resulting in different decision-making. For pocket modeling, there is the general knowledge that the attributes (size, physicochemical properties, flexibility) of residues flanking the site and their relative 3D location explain the specific recognition of ligands.[31,33,50,176] Therefore, site comparison methods approximate these residues into various representations which differ at three levels: (i) the discretization of the residues, (ii) the viewpoint and (iii) the chemical features.

Firstly, possible representations (**Table 1.3**), from coarse-grained to more detailed, are an atom (typically the Cα or Cβ) describing an entire residue (e.g., Apoc), a group of pseudocenters or vectors associated to residue fragments (e.g., CavBase), 3D voxels or surface grid points (e.g., DeepDrug3D) and all atoms cloud (e.g., Ramensky *et al.*). The resolution of the representation determines how local the subsequent comparison can be. For example, rigid matching of atoms which are 7 Å apart in a query pocket can only be associated to similarly spaced atoms in the reference pocket, therefore excluding a pertinent association of smaller areas. Resolution also influences sensitivity to chemical and coordinates variations (**Figure 1.3**). Coarse-grained representations are less sensitive to variations in atomic coordinates but are more perceptive of changes in chemical properties such as single residue mutations. They offer a better signal to noise ratio at the cost of information. In grid/polyhedron-based approaches, the grid resolution (often 0.5 to 1.5 Å)/number of triangles are adjusted to capture the shape of the site while compromising between precision and computing.[82,170] Although small changes of residues are reflected in detailed representations, they can be perceived to a lesser extent since drowned in many other information. Detection of such details are highly influenced by the assignment of chemical features and the performance of the search algorithm. Noticeably, some methods have adopted a mix representation scheme, where gross representations are used for a faster search and whereas finer representations are involved in the scoring.[41]

**Table 1.3.** Discretization of the residues to represent a protein cavity

| Representation | Illustration[a] | Methods |
|---|---|---|
| Single points (e.g. alpha carbon) |  | APoc, *e*MatchSite, Feldman *et al.* (PSILO®), FuzCav, G-LoSA, PocketAlign[b], SiteAlign[b], SMAP, SOIPPA |
| Pseudocenters |  | BSSF, CavBase[b], KRIPO, PocketAlign[b], PocketMatch, RAPMAD, Site2Vec, SiteEngine, SuMo, TrixP[b] |
| Surface points, surface patches, volume points, polyhedron |  | CavBase[b], DeepDrug3D, DeeplyTough, IsoMiF, Patch-Surfer, Shaper, SiteAlign[b], TrixP[b], VolSite |
| All atoms (non-hydrogen) |  | Binkowski *et al.*, Brakoulias *et al.*, Milletti *et al.*, P.A.R.I.S, ProBiS, SiteHopper, TISPA |

[a] The protein cavity is delimited by a few residues (hydrogen atom are not shown). Representative points at different resolutions are depicted as colored spheres. [b] Some methods use mixed representations; in PocketAlign, several schemes are proposed.
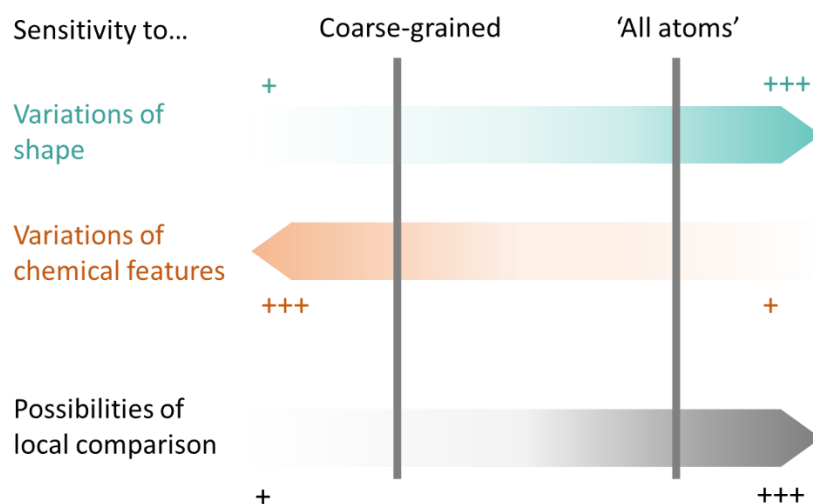
**Figure 1.3.** Sensitivity of coarse-grained or 'all atoms' cavity representations to variations in atomic coordinates, chemical features and subsequent applications (+: low, +++: high).

Secondly, most methods adopt the protein perspective by considering atoms or pseudocenters at the protein surface (e.g. FuzCav, SMAP). A few stand out by projecting these protein patterns into the ligand space, where polyhedron, voxels or points are annotated with the properties of nearest or well-oriented protein features (e.g., IsoMIF, SiteAlign) (**Figure 1.4**). Such discretization aims at offering a good balance between information completeness while handling variations in atomic coordinates and features. However, it is important to recall that grid-based representations are affected by the centroid location and axes orientation during the grid initialization. As a result, the distribution of feature types might change between different 3D models of the same protein (a pharmacophoric feature might move in adjacent voxels or not represented at all), particularly when a voxel is associated to only one feature at a time.
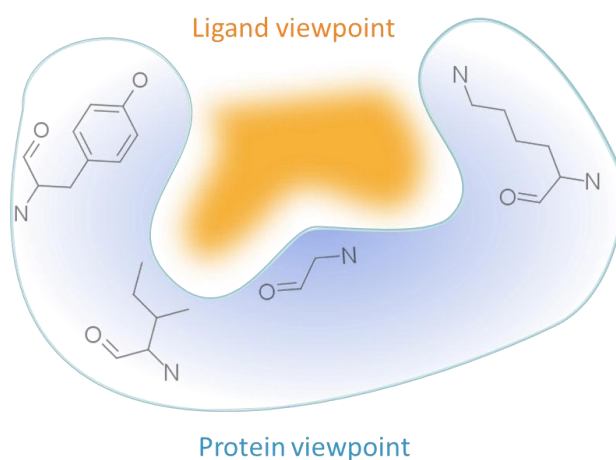


**Figure 1.4.** Protein cavity representation according to the protein or the ligand perspective. Representation of the protein side occupy larger surface to compare.

Finally, besides the two aspects described above, methods differ in their definition of chemical and geometric features. For example, Binkowski *et al.* do not consider the chemical type of atoms but showed that the relative position of the surface atoms describing the shape of the pocket already contain some discriminative information.[159,177] However, shape only information is insufficient, hence it is not surprising that almost all the state-of-the-art site comparison methods annotate surface coordinates atoms with pharmacophoric features to improve discrimination between redundant areas. In coarse-grained representations, $C\alpha/C\beta$ atoms are annotated according to the chemical groups of their residues. For instance, APoc defined eight exclusive chemical groups, allowing a residue to belong to only one.[51] Searching for identity of chemical features between the query and reference pockets with such representations do not account for the interchanging role that fragments in different amino acids can perform: the hydroxyl group of serine and tyrosine are h-bond donor or acceptor whereas tyrosine additionally displays an aromatic feature as a phenylalanine; yet serine and tyrosine belong to different classes. To correct this effect, residues are assigned multiple classes (e.g. Feldman *et al.*, SiteAlign).[44,53] Alternatively, single or group of atoms defining pseudocenters are annotated according to their interaction capacities (e.g. a histidine side chain is represented by h-bond donor-acceptor and aromatic pseudocenters in CavBase). Commonly, five to eight pharmacophoric features are defined (KRIPO, SiteEngine, VolSite),[41,45,82] up to more than 40 atom types (Ramensky *et al.*, PocketFEATURE).[59,165] Other possible chemical attributes are partial charges used in P.A.R.I.S (sup-CK) or SiteEngine scoring,[42,163] atomic density in SuMo[157] or atom types in Brakoulias *et al.*[158] Definition of many feature types might improve the description of the site with precision but might at the same time hinder remote similarity detection by narrowing the applicability domain of the method. Aside chemical features, geometrical patterns are sometimes considered: CavBase and RAPMAD indicate the directionality of polar features by vectors,[41,49] SuMo considers the directionality of the patterns toward the cavity by scalar triple product,[157] SOIPPA assign normal vectors to local surfaces,[161] TrixP and SiteAlign consider distances to fixed points.[44,168]

In a nutshell, there are various ways to represent a protein cavity. Challenges reside in finding a good balance between comprehensive representation of features to ensure reliability and loose representation enabling to detect remote similarities. While the absence of pocket attributes cannot be recovered at the later comparison step, too many attributes may constitute difficulties to the search algorithm in separating the signal from the noise.

## 1.3.2. Search algorithms

Following the selection of features characterizing the cavities, similarity is estimated by algorithms that search for common patterns shared between two sites. First, representations of the protein cavities are converted or organized into comparable and computer-friendly objects that can be processed automatically. There are a variety of search algorithms to this end, which can be categorized according to their inputs, procedure, and visual interpretability (**Figure 1.5**).
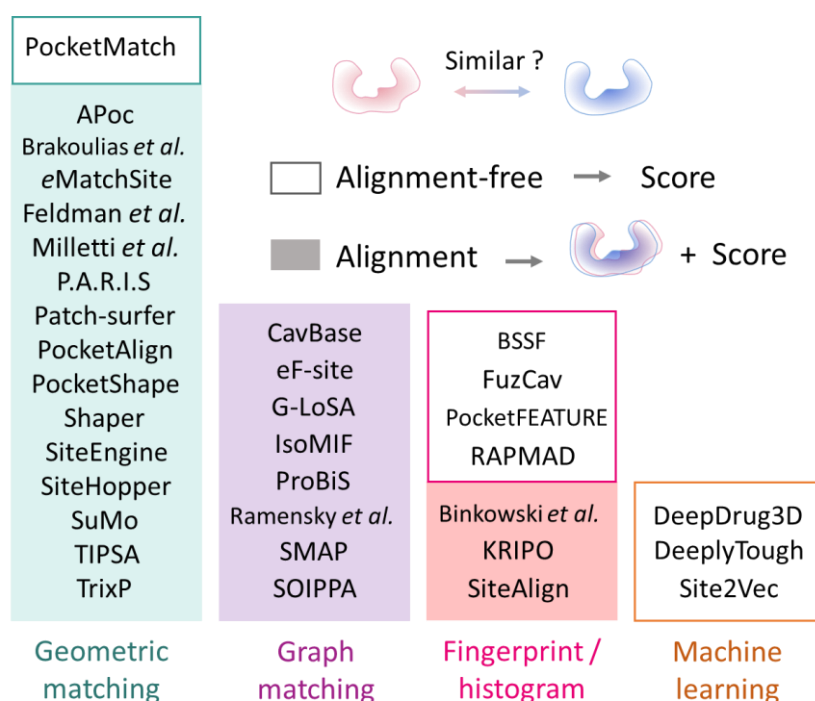


**Figure 1.5.** Classification of state-of-the-art methods for protein pockets comparison. Alignment-based methods (colored background) compute a transformation (rotation, translation) to superpose the query to the target site whereas alignment-free methods (white background) do not provide visual superposition.

The first category of algorithms searches for geometric (e.g. pairwise distances, angles, shape) and chemical (identical or compatible types) constraints to match. It is not sound to be expecting a perfect match, given the errors in 3D structure resolution, the flexibility nature of proteins, the aim to find unobvious similarities. Therefore, a certain margin of geometric errors is always tolerated. PocketMatch compares set of distances belonging to 90 combinations of atom types and properties to establish correspondences between two pockets and keep the solution maximizing the number of correspondences.[43] Global alignment methods (P.A.R.I.S, SiteHopper, Shaper) try to maximize the overlap between two cavities. A seed alignment is initialized, for example by superposing centroids or

principal axes of the two sites, then optimized.[82,163,171] SiteHopper and Shaper rely on the OpenEye tool ROCS (OpenEye Scientific Software, Santa Fe, USA), where atoms/points are represented by smooth Gaussians to enable fuzzy shape comparison.[82,171] A different approach for global optimization is to establish seed correspondences—APoc compares local protein fragments, secondary structures, Milletti et al. associate points based on their circular fingerprints' similarity, *e*MatchSite relates C$\alpha$ according to seven residue-level scores, Patch-Surfer compares the patch surface properties by 3D functions—then solves assignment problems by the Hungarian or other combinatorial optimization algorithms.[51,52,162,166] PocketAlign is based on a similar approach using BLOSSUM62 weights when generating local seed alignments, that are later extended to the full structures.[164]

Alternatively, some methods partition the pocket by considering a few points each time. Given that at least three points are necessary to superpose two objects without ambiguity, those methods enumerate triplets (Brakoulias *et al.*, Feldman *et al.*, SiteEngine, SuMo, TrixP) or quadruplets (TIPSA) of feature points in the query to iteratively search for equivalent cliques in the target.[42,53,157,158,167,168] The formation of the n-tuples can be customized to avoid promiscuous sets. In TrixP, triangles solely made of hydrophobic features are not considered. A match can signify a simple correspondence of identical chemical types and pairwise distances (SiteEngine, TIPSA) or of additional properties such as vector angles, local shape (TrixP). Aligning all possible combinations is costly in time, hence SiteEngine and TrixP respectively employ hashing and bitmap indexing allowing a 'search IN' for faster identification of similar patterns.

In the second category, selected points form the nodes of a graph. According to the cavity representation, each node is annotated by a property and the edges by their lengths. Comparing two cavities results in comparing two graphs to extract the (maximum) common subgraphs. To achieve this end, a product graph is built, by associating similar nodes (property comparison) and edges of almost equal distances, tolerating a certain deviation. Cliques are identified in this association graph to derive pairs of equivalent points that can be used to superpose the two cavities. CavBase, G-LoSA, ProBiS, etc. (**Figure 1.5**) are based on this principle. Differences between methods arise from the graph construction (minimal and maximal distances to consider adjacent nodes), distance tolerances, and the definition of a property match (identity or compatibility). For example, G-LoSA tolerates three different distance deviations (1.5, 2.0 and 2.5 Å) and further evaluates the alignment of local triangles within each clique of more than four nodes.[60] Clique detection is computationally expensive, particularly with dense graphs (e.g. 0.5 Å grid spacing in IsoMIF).[170] Therefore, it requires practically efficient solutions such as the Bron–Kerbosch algorithm and improved variants.[178,179]

Methods in the third category generally adopt a global vision of the protein sites. They consider a pocket as a fixed-length fingerprint or histograms, where comparing two pockets is calculating the similarity or distances between their fingerprints/histograms. BSSF, FuzCav and KRIPO respectively compute

couple or triplets of pharmacophoric features separated by binned distances. While the two former count the number of occurrences of each combination, bits are activated in KRIPO when a combination occurs. Later, KRIPO fuzzifies its fingerprints to account for the neighborhood phenomena.[45] SiteAlign also compare fingerprints, but contrarily to the other methods, the fingerprint of the query pocket is iteratively generated, as it derives from properties of the cavity projected on a rotated/translated 80-face polyhedron.[44] Since the site is discretized and a finite number of geometric transformations are sampled, the performance of the search depends on the resolution of the steps, at the cost of the computing time. Finally, Binkowski *et al.* and RAPMAD compare distributions of pairwise distances between the pocket features.[49,159] RAPMAD generates 14 histograms, one for each of the seven pharmacophoric features, considering two centroids. The idea behind these implementations is that similar binding sites will exhibit similar set of distances. However, these methods may suffer from matching redundant distances that do not superpose geometrically. The advantage of fingerprints/histograms is to enable faster comparison, without the computationally expensive alignment. Still, KRIPO and Binkowski *et al.* generate an alignment independently of the comparison procedure for visual inspections, SiteAlign as part of its search procedure.

Finally, the recent regain of interest for deep neural networks on chemical information favors the emergence of data-driven methods for binding site comparison. Typically, binary classification models are created to discriminate between similar and dissimilar pairs of pockets. Site2Vec transform the features representing a cavity into a fixed-length vector that can feed a random forest classifier. DeepDrug3D and DeeplyTough discretize the 3D space of the pocket as voxels, and logically train a convolutional neural network (CNN) model.[55,173] Besides the dependency to sufficiently diverse training datasets for a generalized model, these approaches suffer from interpretability of the predictions. Interestingly, DeepDrug3D exploits the activation map to visually highlight areas that largely contribute to the classification.

The above-summarized methods use only the protein information for comparison. Provided the pocket is delimited, they have a larger scope that reaches deorphanization of targets. When bound ligands are available, comparing the protein-ligand interactions can be an efficient alternative, particularly when the goal is to reproduce existing binding modes. Likewise, dedicated methods are based on graph alignment (e.g.Grim) or fingerprints comparison (e.g. TIFP).[180]

## 1.3.3. Local comparison of protein cavities

Looking for an average match that maximizes the overlap between entire cavities is not forcibly the right solution to similarity estimation. Local comparison is a popular term, often used to differentiate full protein structural comparison from protein site comparison. Here, we refer to truly local comparison

of protein pockets (**Figure 1.6**), i.e. subpockets of approximately 3-to-4 Å radius (for reference, approximately the shortest distance between a chain of four atoms connected by simple bonds). Enabling local similarity detection is relevant for drug design applications since a few similar subpockets between two targets may suffice for a same ligand to bind. This observation was applied to explain the binding of cyclooxygenase type 2 inhibitors to carbonic anhydrase.[46]

Logically, methods that can operate locally have implemented detailed site representation and/or adequate algorithms that partition the cavity during the search. In the G-LoSA example, global matches are decomposed into local subsites to generate other solutions. Local comparison can also be achieved by providing subpockets as input to the search algorithm. KRIPO enables to compare subpockets delimited by fragmented ligands.[45] While the search algorithms are a major factor in detecting subtle common motifs, how pocket similarities are quantified is equally important, since generalizing the score over the full pockets might hinder any local similarity as well.
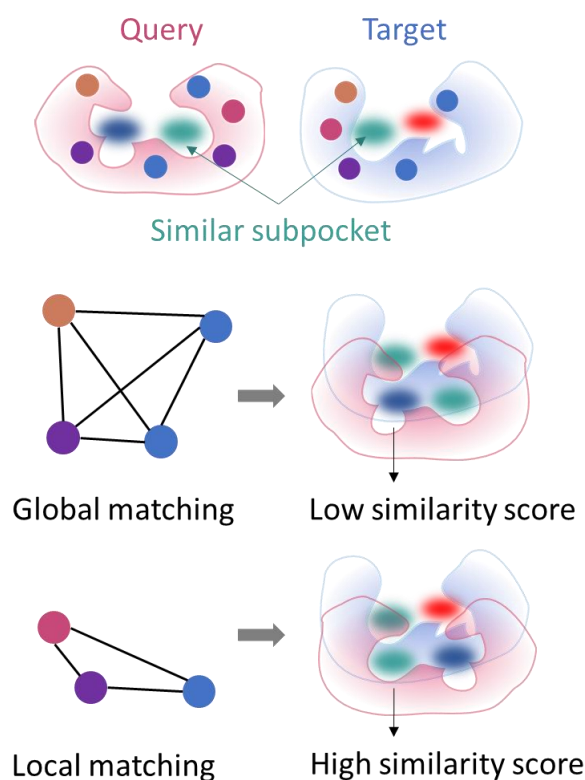


**Figure 1.6.** Global versus local pattern comparison.

Local comparison is notably suitable to handle cases of conformational change upon ligand binding.[58] By analogy to ligand versus fragment promiscuity, comparing smaller cavity regions is likely to be more redundant at the proteome scale than comparing full cavities, enabling to catch similarities between remote proteins but at the same time yielding possible unspecific matches. Finally, successful discrimination requires a robust scoring scheme.

## 1.3.4. Scoring functions

Scoring functions serve two purposes. They quantify the final output generated by the search algorithm. In many cases (e.g., alignment-based), they are also used to guide the search and prioritize one among several possible solutions. It is not uncommon to use distinct scoring functions for the search and final quantification.[42] Consequently, a method may implement an accurate representation and efficient search algorithm but fail to accurately predict similarity levels if the scoring functions are incorrect. Some analogy can be made with the problem of pose sampling and ranking in docking, leading to rescoring efforts.[181] Aspects to consider when defining a scoring function for site comparison are (i) the discriminative potential, (ii) the minimal and maximal boundaries, (iii) the broadness, (iv) the sensitivity to the size of the cavities, (v) the interpretability. The very simple and intuitive scoring scheme counts the number of common patterns between two pockets (Brakoulias *et al.*).[158] However, bigger sites would tend to score higher as the chances for a match increase. To avoid this bias, methods account for the size of the pockets using metrics such as the proportion of aligned features with respect to the query/target size (FuzCav, PocketMatch), Tanimoto indices (IsoMIF, KRIPO, TIPSA, Shaper) and Tversky indices (Shaper). SiteHopper adopts a linear combination of Tanimoto measures for shape and chemical features matching. Almost all alignment-based geometric matching methods aim at minimizing the root mean square deviation (RMSD) of superposition candidates or with respect to a cutoff (Brakoulias *et al.*, SuMo, etc.). In some cases, the RMSD is also a composite of the final score (Milletti *et al.*, PocketAlign). In the same way, CavBase R2 score accounts for the RMSD of peudocenters when scoring the overlap of the surface grid points. Implementing successive scores (Binkowski *et al.*, ProBiS) enables the user to apply a custom filter according to the desired application. For instance, SiteEngine proposes a hierarchical workflow where a gross evaluation allows to quickly filter out bad solutions before applying a finer rescoring on promising matches. Instead of reporting similarities, some methods rather measure the distances between pockets (SiteAlign)—the lower, the better. BSSF and RAPMAD, which compare histograms, respectively report the Kolmogorov-Smirnov and the Jensen-Shannon divergences. Scoring functions can be more complex, often at the cost of interpretability (Feldman *et al.*, *e*MatchSite, P.A.R.I.S).

Weights are used to give more or less importance to different variables (types of features, geometric patterns) but their assignment are at best subjective,[60,166,168] intuitive such as inverse of feature frequency, or adapted from sequence alignment methods (BLOSSUM, PSSM).[161,164,182,183] Proportioning penalties of mismatches with respect to the positive contributions of the matches as in TrixP is tricky and might better or worsen the discrimination performance in noisy representations. Fingerprint comparison is delicate, when bins are counts or integer descriptors with variable ranges, or when comparing two pockets of different sizes. Descriptors are normalized,[44] or the scores are corrected to account for the increase of activated bits with respect to the size of the cavity.[45] Finally, the

commutativity of the score should be regarded, to ensure a consistent output whatever the reference/query order.

A few studies[44,47,51,54,56,82,161] have assessed the significance (Z-score, P-values) of their scoring by analyzing random distributions or robustness to variations in the cavities (simulated data, molecular dynamic simulations). While these studies offer a certain overview on possible scoring thresholds in screening settings, we draw attention to their biases to used datasets.

## 1.4.  Retrospective evaluations and datasets

To demonstrate their applicability, methods for comparing protein pockets have been evaluated for their ability to (i) discriminate between similar and dissimilar binding sites (classification), (ii) retrieve similar pairs seeded in decoys (enrichment), or (iii) cluster proteins belonging to the same families according to other classifications (e.g. SCOP, functional annotations).[160,184,185] The availability of structural data impacts the design of evaluation datasets.

As for any benchmarking study, the quality of the dataset is instrumental to the reliability of the conclusions. Popular computational approaches such as molecular docking benefit from well-established standards and datasets.[186,187] Predicting the binding affinity of molecules to a target can be directly verified by experimental measures in many circumstances. Contrarily, pocket similarity cannot be measured experimentally. Instead, similarity prediction suggests hypotheses such as the recognition of similar ligands or the catalysis of the same reaction, which are then confronted to *in vitro* experiments. What is conveyed here is that there is not a straight line between predictions and verifications since ligand recognition involves other parameters likely not evaluated by site comparison methods, such as the pocket flexibility, the influence of disregarded parts of the protein (residues outside the cavity), the ligand conformations and energetics. Indeed, the ligand may bind to different proteins in different conformations and using different interaction patterns.[58]

Nevertheless, many available datasets are used with the assumption that similar pockets are those binding to identical or similar ligands, and *vice versa* (APoc set, Kahraman *et al.*, TOUGH-M1, TOUCH-C1, Barelier et al., **Table 1.4**).[51,173,177,188,189]

These include proteins belonging to the same family for the easiest ones, and unrelated proteins for the most difficult datasets. In these cases, unrelated proteins are predicted by other computational approaches (sequence alignment, global structural comparison). Besides the discussions above, one issue encountered with these definitions is how to set the similarity cutoff to group proteins and ligands. Chen *et al.* (Vertex) dataset defines similar pairs as pockets in PDB proteins sharing at least three

submicromolar ligands according to ChEMBL while dissimilar pairs share at least three ligands with large affinity variations going from one target to the other.[171] Although giving a different perspective, this dataset is imbalanced as the similar pairs (n = 6598) largely outnumbered the dissimilar pairs (n = 379). Still, the main concern is the ChEMBL ligands used for annotation not necessarily be targeting the PDB binding sites that are finally compared. Generally, datasets relying on ligand binding information suffer from data incompleteness.[190,191] Dissimilar pairs are based on limited available/accessible binding information, because all ligands have not been tested against all targets. Otherwise, some pairs labeled as 'dissimilar' might have fallen into the 'similar' classes.

Given the bias in the PDB data towards some protein-cofactors complexes and well-studied protein families, methods have been extensively evaluated on nucleotide-binding pockets. Similarly, intrafamily retrieval of proteases, kinases or steroid-binding sites were widely studied.[41,162,192] Alternatively, other datasets proposed pairs of similar and dissimilar sites based on their functional annotations (UniProt, Enzyme Classification number)[185] and fold (SCOP,[160] CATH[184]) starting from the non-redundant sc-PDB database to reduce these biases.[44,47,193]

The ProSPECCTs benchmarking work intended to propose guidelines for methods evaluation while revealing common issues.[63] Many datasets are too easy or do not correspond to realistic challenges. Compilation of difficult cases, drawn from experimental observations are provided but such examples are rare.[43,47,168] Finally, the most effective evaluations are prospective applications in research.

**Table 1.4.** Common datasets used in benchmarking studies for pocket comparison.

| Purpose | Name | Content | # Positive (# Negatives) |
|---|---|---|---|
| Pairs of cavities from dissimilar proteins binding identical or similar ligands (positives) and dissimilar ligands (negatives)<br><br>Vertex: positives are pairs of sites in proteins sharing 3 high affinity ligands (potency < 100 nM) vs. pairs of sites in proteins sharing 3 ligands with divergent affinities | APoc set[51] | Various | 38 066 (38 066) |
| | Barelier et al.[188] | Various | 62 |
| | Homogeneous[163] | Various | 100 |
| | Kahraman et al. / extended[163,177] | Cofactor sites | 100 / 972 |
| | TOUGH-M1[189] | Various | 505 116 (556 810) |
| | TOUGH-C1[173] | Nucleotides, heme, steroid sites | 2218 |
| | Vertex[171] | Various | 6598 (379) |

| Pairs of cavities associated to the same (positives) or different (negatives) functions and fold class | sc-PDB-derived[47] | Various | 769 (769) |
|---|---|---|---|
| Intra-family classification | Proteases, kinases, GPCRs, Estrogen receptors[41,44,82,162,192] | | - |
| Difficult cases | Difficult cases[43,47] | Diverse from experimental validations | 8 |
| Successful applications | ProSPECCTs D7[63] | Diverse from experimental validations | 115 (56 284) |
| Structures of identical sequences | ProSPECCTs D1[63] | Various | 13 430 (92 846) |
| | ProSPECCTs D1.2[63] | Various | 241 (1784) |
| NMR structures | ProSPECCTs D2[63] | Various | 7729 (100 512) |
| Synthetic set: random mutations | ProSPECCTs D3 and D4[63] | Various | 13 430 (67 150) |
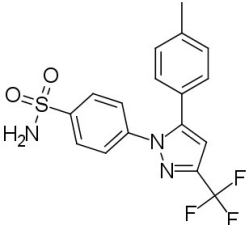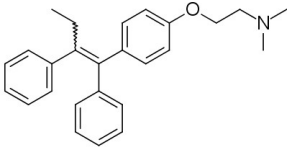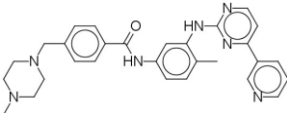
# 1.5. Applications in medicinal chemistry and practical considerations

Protein cavities comparison have been used alongside with other computational methods to predict or explain the binding of small molecules to different targets. Many of these success stories are described in a recent review.[18] Following secondary targets prediction, structural information (e.g. bound ligands) are used as hints to efficiently explore the chemical space for faster hit identification. Proposed putative hits are directly tested experimentally or serve for designing focused screening libraries. The most striking examples involve unrelated targets. For example, the graph matching method CavBase was successful in detecting the subpockets similarity between cyclooxygenase type 2 (COX-2) and human carbonic anhydrase (CA), supporting the nanomolar inhibition of CA by COX-2 inhibitors.[46] Other literature examples involving diverse methods are summarized **Table 1.5**. Practically, inspection of aligned features or manual selection, in addition to the high similarity scores and rankings were carried out, highlighting the advantage of alignment-based methods. Other computational studies by docking and molecular dynamics simulations are used complementarily.[194] Ligand induced fit of the protein might hinder the detection of hidden similarity, hence the exploration of several query and target

structures when available.[195,196] Although several studies are rather explanation of *in vitro*/clinical observations[46,197] than fully blind predictions or involve targets that were already known to share common characteristics (evolutionary conservation, cofactor ATP or NAD sites, kinases polypharmacology),[196,198–201] the detected similarities/divergences were to be proved and provided new insights. Strikingly, pocket comparison has enabled new discoveries with limited to no preliminary information. All together, these case studies demonstrated how the analysis of cavity similarities can benefit drug design.

**Table 1.5.** Examples of binding site comparison applications relevant to medicinal chemistry.[18]

| Year | Methods (Study)[a] | Primary target | Secondary target | Compound / affinity to secondary target |
|------|------|------|------|------|
| 2004 | CavBase (C)[46] | Cyclooxygenase type 2 (COX-2) | Human carbonic anhydrase (CA) |  Celecoxib $IC_{50}$ = 21 nM |
| 2006 | CavBase (P)[202] | Querying SARS-Cov M[pro] to a database of amino acids-bound subpockets for peptide design | | Design of a focused library of peptides ~7 – 20 µM |
| 2007 | SOIPPA (E)[197] | Estrogen receptor alpha (ERα) | Sarcoplasmic Reticulum (SR) Ca2+ ion channel ATPase (SERCA, putative) |  Tamoxifen Inhibits thapsigargin (SERCA inhibitor) effects |
| 2009 | Brakoulias *et al.* (C)[201] | Rationalization of cross-reactivity of kinase inhibitors | |  Imatinib |

[a] Type of study: (C) confirmation, (E) explanation of experimental or clinical observations, (P) prediction of new findings.

**Table 1.5.** Examples of binding site comparison applications relevant to medicinal chemistry.[18] (continued)
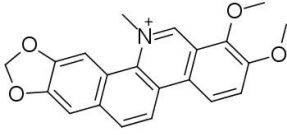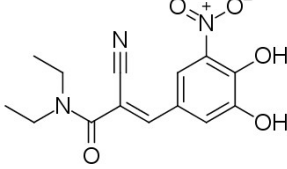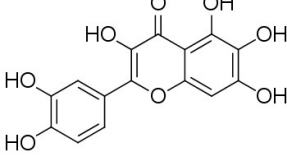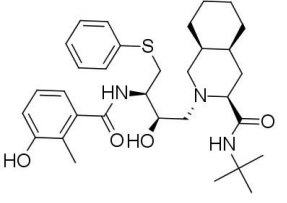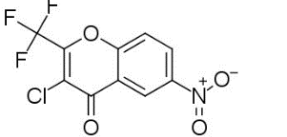
| Year | Methods (Study)[a] | Primary target | Secondary target | Compound / affinity to secondary target |
|---|---|---|---|---|
| 2009 | CPASS[203] (P)[195] | Bcl-2 apoptosis protein Bcl-xL | *S. typhimurium* type III Secretion System Needle Protein (PrgI) |  Chelerythrine 2D NMR binding analysis |
| 2009 | SOIPPA (P)[198] | Catechol-O-methyltransferase (COMT) | *M. tuberculosis* enoyl-acyl carrier protein reductase (InhA) |  Entacapone $MIC_{99} = 260\ \mu M$ |
| 2010 | SiteAlign (P)[199] | Pim-1 kinase | Synapsin I |  Quercetagetin $IC_{50} = 0.15\ \mu M$ |
| 2011 | SMAP (P)[204] | HIV-1 protease | Epidermal growth factor receptor (EGFR) |  Nelfinavir High micromolar |
| 2012 | PSSC[205] (P)[206] | Monoamine oxidase (MAO) | Lysine-specific demethylase 1 (LSD1) |  Namoline $IC_{50} = 51\ \mu M$ |
| 2013 | SiteEngine (P)[207] | Template ubiquitin (Ub)-binding interfaces | Discovery of new Ub-binding domain: ALIX-V | ALIX-V:mono-Ub MST Kd = 119 μM |

**Table 1.5.** Examples of binding site comparison applications relevant to medicinal chemistry.[18] (continued)
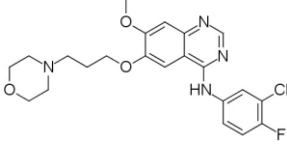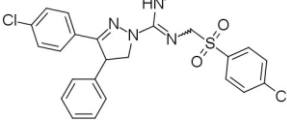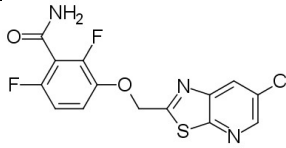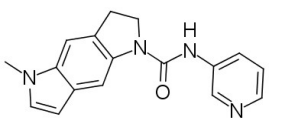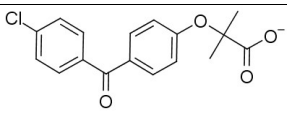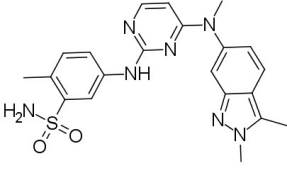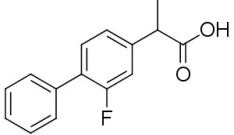
| Year | Methods (Study)[a] | Primary target | Secondary target | Compound / affinity to secondary target |
|------|------|------|------|------|
| 2014 | SMAP (P)[194] | Epidermal growth factor receptor (EGFR) | β-secretase (BACE-1) | Gefitinib  $IC_{50} = 20\ \mu M$ |
| 2015 | KRIPO (E)[208] | Cannabinoid receptor 1 (CB1R) | Adenine nucleotide translocase 1 (ANT1) | Ibipinabant  Inhibition of ADP/ATP exchange |
| 2015 | PocketFEA-TURE (E)[196,209] | *S.aureus* FtsZ (SaFtsZ) | Selectivity of PC190723 to SaFtsZ vs. other species FtsZ and mutants SaFtsZ | |
| 2015 | PocketMatch (C)[210,211] | Serotonin metabotropic receptors: 5-HT$_{2B}$R 5-HT$_{2C}$R | Ionotropic α7 nicotinic acetylcholine receptor (nAChR) | SB-206553  $EC_{50} = 1.5\ \mu M$ |
| 2015 | PSIM[212] (P)[213] | Peroxisome proliferator-activated receptor gamma (PPARγ) | Cyclooxygenase type 1 (COX-1) | Fenofibric acid  $IC_{50} = 950\ \mu M$ |
| 2015 | TM-align[27] (P)[200] | Tyrosine kinase family members | Acetylcholinestera-se (AchE) | Pazopanib  $IC_{50} = 0.93\ \mu M$ |

**Table 1.5.** Examples of binding site comparison applications relevant to medicinal chemistry.[18] (continued)

| Year | Methods (Study)[a] | Primary target | Secondary target | Compound / affinity to secondary target |
|------|---------|----------------|------------------|------------------------------------------|
| 2019 | VolSite-Shaper (P)[67] | Cyclooxygenase type 1 (COX-1) | Cinnamoylesterase |  Flurbiprofen Allosteric inhibition (IC50 ~400 µM) |

[a] Type of study: (C) confirmation, (E) explanation of experimental or clinical observations, (P) prediction of new findings.

## 1.6. Conclusions

This chapter have presented the current state of protein site comparison applied to small molecule drug design. As one of the computer-aided drug design strategies, assessing the similarity of protein pockets constitutes a unique way to analyze structural information, hence complement other well-spread approaches. The repertoire of available methods is diverse with respect to the detection and representation of cavities, the search algorithms, the scoring functions. All of these aspects must somehow be coordinated to achieve the best performance. Still, limitation of experimental data and bias in datasets constitute major obstacles to properly evaluate such methods. In reality, estimating protein site similarity is context-dependent for different considered pairs, and for different studies. The importance of matched features is influenced by the chemical context and physicochemical considerations of the targets, making it hard to predict subtle and specific similarities from generalized principles. One holy grail of computational chemists is to repurpose existing drugs proposed by structure-based experiments. Although this pursuit appears at best hardly probable due to the optimization of drugs to their targets,[214,215] protein sites comparison have demonstrated its effective contribution to medicinal chemistry projects, from the elucidation of previous biological observations to generation of new hypotheses supported by experimentally validation. The majority of the-state-of-the-art methods are based on superposition of the compared structures. Alignment allows visual inspection and increase the possibilities of applications. Typically, pocket-bound ligands in the reference frame can be transposed to the target pocket and serve as starting point for ligand generation.

Improvement of the algorithmic efficiency of methods alongside with technological progress would enable to better follow the current growth of publicly-available protein structures.
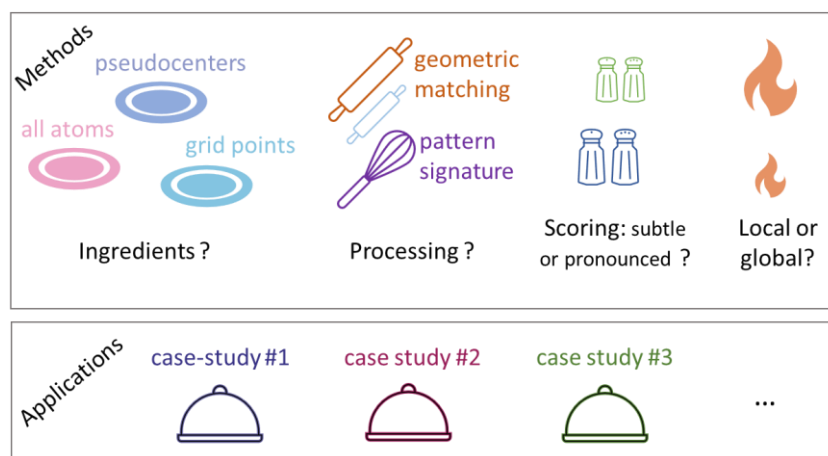


For each case study (meal), might correspond a different combination of methods (recipe).

## 1.7. References

1.  Illergård, K.; Ardell, D. H.; Elofsson, A. Structure Is Three to Ten Times More Conserved than Sequence - A Study of Structural Response in Protein Cores. *Proteins Struct. Funct. Bioinforma.* **2009**, *77*, 499–508.

2.  McCoy, A. J. Solving Structures of Protein Complexes by Molecular Replacement with Phaser. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2007**, *63*, 32–41.

3.  Ilari, A.; Savino, C. Protein Structure Determination by X-Ray Crystallography; Keith, J. M., Ed.; Humana Press: Totowa, NJ, 2008; pp 63–87.

4.  Fraser, J. S.; van den Bedem, H.; Samelson, A. J.; Lang, P. T.; Holton, J. M.; Echols, N.; Alber, T. Accessing Protein Conformational Ensembles Using Room-Temperature X-Ray Crystallography. *Proc. Natl. Acad. Sci.* **2011**, *108*, 16247–16252.

5.  Cavalli, A.; Salvatella, X.; Dobson, C. M.; Vendruscolo, M. Protein Structure Determination from NMR Chemical Shifts. *Proc. Natl. Acad. Sci.* **2007**, *104*, 9615–9620.

6.  Zivanov, J.; Nakane, T.; Forsberg, B. O.; Kimanius, D.; Hagen, W. J.; Lindahl, E.; Scheres, S. H. New Tools for Automated High-Resolution Cryo-EM Structure Determination in RELION-3. *Elife* **2018**, *7*, 1–22.

7.  Yip, K. M.; Fischer, N.; Paknia, E.; Chari, A.; Stark, H. Atomic-Resolution Protein Structure Determination by Cryo-EM. *Nature* **2020**, *587*, 157–161.

8.  Henderson, R.; Baldwin, J. M.; Ceska, T. A.; Zemlin, F.; Beckmann, E.; Downing, K. H. Model for the Structure of Bacteriorhodopsin Based on High-Resolution Electron Cryo-Microscopy. *J. Mol. Biol.* **1990**, *213*, 899–929.

9.  Stahlberg, H.; Walz, T. Molecular Electron Microscopy: State of the Art and Current Challenges. *ACS Chem. Biol.* **2008**, *3*, 268–281.

10. Liang, J.; Woodward, C.; Edelsbrunner, H. Anatomy of Protein Pockets and Cavities:

Measurement of Binding Site Geometry and Implications for Ligand Design. *Protein Sci.* **1998**, *7*, 1884–1897.

11. Rognan, D. Structure-Based Approaches to Target Fishing and Ligand Profiling. *Mol. Inform.* **2010**, *29*, 176–187.

12. Lewis, R. A. [8] Clefts and Binding Sites in Protein Receptors; 1991; Vol. 202, pp 126–156.

13. Berman, H. M. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.

14. Burley, S. K.; Bhikadiya, C.; Bi, C.; Bittrich, S.; Chen, L.; Crichlow, G. V.; Christie, C. H.; Dalenberg, K.; Di Costanzo, L.; Duarte, J. M.; Dutta, S.; Feng, Z.; Ganesan, S.; Goodsell, D. S.; Ghosh, S.; Green, R. K.; Guranović, V.; Guzenko, D.; Hudson, B. P.; Lawson, C. L.; Liang, Y.; Lowe, R.; Namkoong, H.; Peisach, E.; Persikova, I.; Randle, C.; Rose, A.; Rose, Y.; Sali, A.; Segura, J.; Sekharan, M.; Shao, C.; Tao, Y.-P.; Voigt, M.; Westbrook, J. D.; Young, J. Y.; Zardecki, C.; Zhuravleva, M. RCSB Protein Data Bank: Powerful New Tools for Exploring 3D Structures of Biological Macromolecules for Basic and Applied Research and Education in Fundamental Biology, Biomedicine, Biotechnology, Bioengineering and Energy Sciences. *Nucleic Acids Res.* **2021**, *49*, D437–D451.

15. Bhagavat, R.; Sankar, S.; Srinivasan, N.; Chandra, N. An Augmented Pocketome: Detection and Analysis of Small-Molecule Binding Pockets in Proteins of Known 3D Structure. *Structure* **2018**, *26*, 499-512.e2.

16. Pérot, S.; Sperandio, O.; Miteva, M. A.; Camproux, A. C.; Villoutreix, B. O. Druggable Pockets and Binding Site Centric Chemical Space: A Paradigm Shift in Drug Discovery. *Drug Discov. Today* **2010**, *15*, 656–667.

17. Vulpetti, A.; Kalliokoski, T.; Milletti, F. Chemogenomics in Drug Discovery: Computational Methods Based on the Comparison of Binding Sites. *Future Medicinal Chemistry*. October 2012, pp 1971–1979.

18. Ehrt, C.; Brinkjost, T.; Koch, O. Impact of Binding Site Comparisons on Medicinal Chemistry and Rational Molecular Design. *J. Med. Chem.* **2016**, *59*, 4121–4151.

19. Crystallography: Protein Data Bank. *Nat. New Biol.* **1971**, *233*, 223–223.

20. Rossmann, M. G.; Blow, D. M. The Detection of Sub-Units within the Crystallographic Asymmetric Unit. *Acta Crystallogr.* **1962**, *15*, 24–31.

21. Taylor, W. R.; Orengo, C. A. Protein Structure Alignment. *J. Mol. Biol.* **1989**, *208*, 1–22.

22. Nussinov, R.; Wolfson, H. J. Efficient Detection of Three-Dimensional Structural Motifs in Biological Macromolecules by Computer Vision Techniques. *Proc. Natl. Acad. Sci.* **1991**, *88*, 10495–10499.

23. Fischer, D.; Wolfson, H.; Nussinov, R. Spatial, Sequence-Order-Independent Structural Comparison of α/β Proteins: Evolutionaiy Implications. *J. Biomol. Struct. Dyn.* **1993**, *11*, 367–380.

24. Vriend, G.; Sander, C. Detection of Common Three-Dimensional Substructures in Proteins. *Proteins Struct. Funct. Genet.* **1991**, *11*, 52–58.

25. Rao, S. T.; Rossmann, M. G. Comparison of Super-Secondary Structures in Proteins. *J. Mol. Biol.* **1973**, *76*, 241–256.

26. Rossmann, M. G.; Argos, P. Exploring Structural Homology of Proteins. *J. Mol. Biol.* **1976**, *105*, 75–95.

27. Zhang, Y.; Skolnick, J. TM-Align: A Protein Structure Alignment Algorithm Based on the TM-Score. *Nucleic Acids Res.* **2005**, *33*, 2302–2309.

28.     Fetrow, J. S.; Godzik, A.; Skolnick, J. Functional Analysis of the Escherichia Coli Genome Using the Sequence-to-Structure-to-Function Paradigm: Identification of Proteins Exhibiting the Glutaredoxin/Thioredoxin Disulfide Oxidoreductase Activity 1 1Edited by F. E. Cohen. *J. Mol. Biol.* **1998**, *282*, 703–711.

29.     Fetrow, J. S.; Skolnick, J. Method for Prediction of Protein Function from Sequence Using the Sequence-to-Structure-to-Function Paradigm with Application to Glutaredoxins/Thioredoxins and T 1 Ribonucleases 1 1Edited by F. Cohen. *J. Mol. Biol.* **1998**, *281*, 949–968.

30.     Artymiuk, P. J.; Poirrette, A. R.; Grindley, H. M.; Rice, D. W.; Willett, P. A Graph-Theoretic Approach to the Identification of Three-Dimensional Patterns of Amino Acid Side-Chains in Protein Structures. *J. Mol. Biol.* **1994**, *243*, 327–344.

31.     Fischer, D.; Wolfson, H.; Lin, S. L.; Nussinov, R. Three-Dimensional, Sequence Order-Independent Structural Comparison of a Serine Protease against the Crystallographic Database Reveals Active Site Similarities: Potential Implications to Evolution and to Protein Folding. *Protein Sci.* **1994**, *3*, 769–778.

32.     Russell, R. B.; Sasieni, P. D.; Sternberg, M. J. E. Supersites within Superfolds. Binding Site Similarity in the Absence of Homology. *J. Mol. Biol.* **1998**, *282*, 903–918.

33.     Fischer, D.; Norel, R.; Wolfson, H.; Nussinov, R. Surface Motifs by a Computer Vision Technique: Searches, Detection, and Implications for Protein-Ligand Recognition. *Proteins Struct. Funct. Genet.* **1993**, *16*, 278–292.

34.     Wallace, A. C.; Borkakoti, N.; Thornton, J. M. Tess: A Geometric Hashing Algorithm for Deriving 3D Coordinate Templates for Searching Structural Databases. Application to Enzyme Active Sites. *Protein Sci.* **1997**, *6*, 2308–2323.

35.     Wallace, A. C.; Laskowski, R. A.; Thornton, J. M. Derivation of 3D Coordinate Templates for Searching Structural Databases: Application to Ser-His-Asp Catalytic Triads in the Serine Proteinases and Lipases. *Protein Sci.* **1996**, *5*, 1001–1013.

36.     Hendlich, M.; Rippmann, F.; Barnickel, G. LIGSITE: Automatic and Efficient Detection of Potential Small Molecule-Binding Sites in Proteins. *J. Mol. Graph. Model.* **1997**, *15*, 359–363.

37.     Levitt, D. G.; Banaszak, L. J. POCKET: A Computer Graphics Method for Identifying and Displaying Protein Cavities and Their Surrounding Amino Acids. *J. Mol. Graph.* **1992**, *10*, 229–234.

38.     Volkamer, A.; Kuhn, D.; Rippmann, F.; Rarey, M. DoGSiteScorer: A Web Server for Automatic Binding Site Prediction, Analysis and Druggability Assessment. *Bioinformatics* **2012**, *28*, 2074–2075.

39.     Krivák, R.; Hoksza, D. P2Rank: Machine Learning Based Tool for Rapid and Accurate Prediction of Ligand Binding Sites from Protein Structure. *J. Cheminform.* **2018**, *10*, 39.

40.     Le Guilloux, V.; Schmidtke, P.; Tuffery, P. Fpocket: An Open Source Platform for Ligand Pocket Detection. *BMC Bioinformatics* **2009**, *10*, 1–11.

41.     Schmitt, S.; Kuhn, D.; Klebe, G. A New Method to Detect Related Function among Proteins Independent of Sequence and Fold Homology. *J. Mol. Biol.* **2002**, *323*, 387–406.

42.     Shulman-Peleg, A.; Nussinov, R.; Wolfson, H. J. Recognition of Functional Sites in Protein Structures. *J. Mol. Biol.* **2004**, *339*, 607–633.

43.     Yeturu, K.; Chandra, N. PocketMatch: A New Algorithm to Compare Binding Sites in Protein Structures. *BMC Bioinformatics* **2008**, *9*, 543.

44.     Schalon, C.; Surgand, J. S.; Kellenberger, E.; Rognan, D. A Simple and Fuzzy Method to Align and Compare Druggable Ligand-Binding Sites. *Proteins Struct. Funct. Genet.* **2008**, *71*, 1755–

1778.

45.     Wood, D. J.; Vlieg, J. De; Wagener, M.; Ritschel, T. Pharmacophore Fingerprint-Based Approach to Binding Site Subpocket Similarity and Its Application to Bioisostere Replacement. *J. Chem. Inf. Model.* **2012**, *52*, 2031–2043.

46.     Weber, A.; Casini, A.; Heine, A.; Kuhn, D.; Supuran, C. T.; Scozzafava, A.; Klebe, G. Unexpected Nanomolar Inhibition of Carbonic Anhydrase by COX-2-Selective Celecoxib: New Pharmacological Opportunities Due to Related Binding Site Recognition. *J. Med. Chem.* **2004**, *47*, 550–557.

47.     Weill, N.; Rognan, D. Alignment-Free Ultra-High-Throughput Comparison of Druggable Protein-Ligand Binding Sites. *J. Chem. Inf. Model.* **2010**, *50*, 123–135.

48.     Xiong, B.; Wu, J.; Burk, D. L.; Xue, M.; Jiang, H.; Shen, J. BSSF: A Fingerprint Based Ultrafast Binding Site Similarity Search and Function Analysis Server. *BMC Bioinformatics* **2010**, *11*, 47.

49.     Krotzky, T.; Grunwald, C.; Egerland, U.; Klebe, G. Large-Scale Mining for Similar Protein Binding Pockets: With RAPMAD Retrieval on the Fly Becomes Real. *J. Chem. Inf. Model.* **2015**, *55*, 165–179.

50.     Kellenberger, E.; Schalon, C.; Rognan, D. How to Measure the Similarity Between Protein Ligand-Binding Sites? *Curr. Comput. Aided-Drug Des.* **2008**, *4*, 209–220.

51.     Gao, M.; Skolnick, J. APoc: Large-Scale Identification of Similar Protein Pockets. *Bioinformatics* **2013**, *29*, 597–604.

52.     Brylinski, M. EMatchSite: Sequence Order-Independent Structure Alignments of Ligand Binding Pockets in Protein Models. *PLoS Comput. Biol.* **2014**, *10*, e1003829.

53.     Feldman, H. J.; Labute, P. Pocket Similarity: Are α Carbons Enough? *J. Chem. Inf. Model.* **2010**, *50*, 1466–1475.

54.     Konc, J.; Janežič, D. ProBiS Algorithm for Detection of Structurally Similar Protein Binding Sites by Local Structural Alignment. *Bioinformatics* **2010**, *26*, 1160–1168.

55.     Simonovsky, M.; Meyers, J. DeeplyTough: Learning Structural Comparison of Protein Binding Sites. *J. Chem. Inf. Model.* **2020**, *60*, 2356–2366.

56.     Xie, L.; Xie, L.; Bourne, P. E. A Unified Statistical Model to Support Local Sequence Order Independent Similarity Searching for Ligand-Binding Sites and Its Application to Genome-Based Drug Discovery. *Bioinformatics* **2009**, *25*, i305–i312.

57.     Levitt, M.; Gerstein, M. A Unified Statistical Framework for Sequence Comparison and Structure Comparison. *Proc. Natl. Acad. Sci.* **1998**, *95*, 5913–5920.

58.     Kalliokoski, T.; Olsson, T. S. G.; Vulpetti, A. Subpocket Analysis Method for Fragment-Based Drug Discovery. *J. Chem. Inf. Model.* **2013**, *53*, 131–141.

59.     Ramensky, V.; Sobol, A.; Zaitseva, N.; Rubinov, A.; Zosimov, V. A Novel Approach to Local Similarity of Protein Binding Sites Substantially Improves Computational Drug Design Results. *Proteins Struct. Funct. Bioinforma.* **2007**, *69*, 349–357.

60.     Lee, H. S.; Im, W. G-LoSA: An Efficient Computational Tool for Local Structure-Centric Biological Studies and Drug Design. *Protein Sci.* **2016**, *25*, 865–876.

61.     Nisius, B.; Sha, F.; Gohlke, H. Structure-Based Computational Analysis of Protein Binding Sites for Function and Druggability Prediction. *J. Biotechnol.* **2012**, *159*, 123–134.

62.     Volkamer, A.; von Behren, M. M.; Bietz, S.; Rarey, M. Prediction, Analysis, and Comparison of Active Sites. In *Applied Chemoinformatics*; Wiley-VCH Verlag GmbH & Co. KGaA:

Weinheim, Germany, 2018; pp 283–311.

63. Ehrt, C.; Brinkjost, T.; Koch, O. A Benchmark Driven Guide to Binding Site Comparison: An Exhaustive Evaluation Using Tailor-Made Data Sets (ProSPECCTs). *PLoS Comput. Biol.* **2018**, *14*, 1–50.

64. Naderi, M.; Lemoine, J. M.; Govindaraj, R. G.; Kana, O. Z.; Feinstein, W. P.; Brylinski, M. Binding Site Matching in Rational Drug Design: Algorithms and Applications. *Brief. Bioinform.* **2019**, *20*, 2167–2184.

65. Yan, C.; Wu, F.; Jernigan, R. L.; Dobbs, D.; Honavar, V. Characterization of Protein–Protein Interfaces. *Protein J.* **2008**, *27*, 59–70.

66. Gao, M.; Skolnick, J. Structural Space of Protein–Protein Interfaces Is Degenerate, Close to Complete, and Highly Connected. *Proc. Natl. Acad. Sci.* **2010**, *107*, 22517–22522.

67. Da Silva, F.; Bret, G.; Teixeira, L.; Gonzalez, C. F.; Rognan, D. Exhaustive Repertoire of Druggable Cavities at Protein-Protein Interfaces of Known Three-Dimensional Structure. *J. Med. Chem.* **2019**, *62*, 9732–9742.

68. Laskowski, R. A.; Luscombe, N. M.; Swindells, M. B.; Thornton, J. M. Protein Clefts in Molecular Recognition and Function. *Protein Sci.* **1996**, *5*, 2438–2452.

69. Laskowski, R. A. SURFNET: A Program for Visualizing Molecular Surfaces, Cavities, and Intermolecular Interactions. *J. Mol. Graph.* **1995**, *13*, 323–330.

70. Wu, Q.; Peng, Z.; Zhang, Y.; Yang, J. COACH-D: Improved Protein–Ligand Binding Sites Prediction with Refined Ligand-Binding Poses through Molecular Docking. *Nucleic Acids Res.* **2018**, *46*, W438–W442.

71. Ben Chorin, A.; Masrati, G.; Kessel, A.; Narunsky, A.; Sprinzak, J.; Lahav, S.; Ashkenazy, H.; Ben-Tal, N. ConSurf-DB: An Accessible Repository for the Evolutionary Conservation Patterns of the Majority of PDB Proteins. *Protein Sci.* **2020**, *29*, 258–267.

72. Berezin, C.; Glaser, F.; Rosenberg, J.; Paz, I.; Pupko, T.; Fariselli, P.; Casadio, R.; Ben-Tal, N. ConSeq: The Identification of Functionally and Structurally Important Residues in Protein Sequences. *Bioinformatics* **2004**, *20*, 1322–1324.

73. Goldenberg, O.; Erez, E.; Nimrod, G.; Ben-Tal, N. The ConSurf-DB: Pre-Calculated Evolutionary Conservation Profiles of Protein Structures. *Nucleic Acids Res.* **2009**, *37*, D323–D327.

74. Glaser, F.; Pupko, T.; Paz, I.; Bell, R. E.; Bechor-Shental, D.; Martz, E.; Ben-Tal, N. ConSurf: Identification of Functional Regions in Proteins by Surface-Mapping of Phylogenetic Information. *Bioinformatics* **2003**, *19*, 163–164.

75. Brylinski, M.; Skolnick, J. A Threading-Based Method (FINDSITE) for Ligand-Binding Site Prediction and Functional Annotation. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 129–134.

76. Roy, A.; Yang, J.; Zhang, Y. COFACTOR: An Accurate Comparative Algorithm for Structure-Based Protein Function Annotation. *Nucleic Acids Res.* **2012**, *40*, W471–W477.

77. Capra, J. A.; Laskowski, R. A.; Thornton, J. M.; Singh, M.; Funkhouser, T. A. Predicting Protein Ligand Binding Sites by Combining Evolutionary Sequence Conservation and 3D Structure. *PLoS Comput. Biol.* **2009**, *5*, e1000585.

78. McGreig, J. E.; Uri, H.; Antczak, M.; Sternberg, M. J. E.; Michaelis, M.; Wass, M. N. 3DLigandSite: Structure-Based Prediction of Protein–Ligand Binding Sites. *Nucleic Acids Res.* **2022**, *50*, W13–W20.

79. Fogha, J.; Diharce, J.; Obled, A.; Aci-Sèche, S.; Bonnet, P. Computational Analysis of

Crystallization Additives for the Identification of New Allosteric Sites. *ACS Omega* **2020**, *5*, 2114–2122.

80. Da Silva, F.; Desaphy, J.; Rognan, D. IChem: A Versatile Toolkit for Detecting, Comparing, and Predicting Protein–Ligand Interactions. *ChemMedChem* **2018**, *13*, 507–510.

81. Volkamer, A.; Griewel, A.; Grombacher, T.; Rarey, M. Analyzing the Topology of Active Sites: On the Prediction of Pockets and Subpockets. *J. Chem. Inf. Model.* **2010**, *50*, 2041–2052.

82. Desaphy, J.; Azdimousa, K.; Kellenberger, E.; Rognan, D. Comparison and Druggability Prediction of Protein–Ligand Binding Sites from Pharmacophore-Annotated Cavity Shapes. *J. Chem. Inf. Model.* **2012**, *52*, 2287–2299.

83. Oliveira, S. H. P.; Ferraz, F. A. N.; Honorato, R. V; Xavier-Neto, J.; Sobreira, T. J. P.; de Oliveira, P. S. L. KVFinder: Steered Identification of Protein Cavities as a PyMOL Plugin. *BMC Bioinformatics* **2014**, *15*, 1–8.

84. Semwal, R.; Aier, I.; Varadwaj, P. K.; Antsiperov, S. PROcket, an Efficient Algorithm to Predict Protein Ligand Binding Site. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer International Publishing, 2019; Vol. 11465 LNBI, pp 453–461.

85. Marchand, J.-R.; Pirard, B.; Ertl, P.; Sirockin, F. CAVIAR: A Method for Automatic Cavity Detection, Description and Decomposition into Subcavities. *J. Comput. Aided. Mol. Des.* **2021**, *35*, 737–750.

86. Saberi Fathi, S. M.; Tuszynski, J. A. A Simple Method for Finding a Protein's Ligand-Binding Pockets. *BMC Struct. Biol.* **2014**, *14*, 1–9.

87. Kleywegt, G. J.; Alwyn Jones, T. Detection, Delineation, Measurement and Display of Cavities in Macromolecular Structures. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **1994**, *50*, 178–185.

88. Petřek, M.; Otyepka, M.; Banáš, P.; Košinová, P.; Koča, J.; Damborský, J. CAVER: A New Tool to Explore Routes from Protein Clefts, Pockets and Cavities. *BMC Bioinformatics* **2006**, *7*, 316.

89. Huang, B.; Schroeder, M. LIGSITEcsc: Predicting Ligand Binding Sites Using the Connolly Surface and Degree of Conservation. *BMC Struct. Biol.* **2006**, *6*, 1–11.

90. Weisel, M.; Proschak, E.; Schneider, G. PocketPicker: Analysis of Ligand Binding-Sites with Shape Descriptors. *Chemistry Central Journal*. 2007, pp 1–17.

91. Kalidas, Y.; Chandra, N. PocketDepth: A New Depth Based Algorithm for Identification of Ligand Binding Sites in Proteins. *J. Struct. Biol.* **2008**, *161*, 31–42.

92. Tripathi, A.; Kellogg, G. E. A Novel and Efficient Tool for Locating and Characterizing Protein Cavities and Binding Sites. *Proteins Struct. Funct. Bioinforma.* **2010**, *78*, 825–842.

93. Kawabata, T. Detection of Multiscale Pockets on Protein Surfaces Using Mathematical Morphology. *Proteins Struct. Funct. Bioinforma.* **2010**, *78*, 1195–1211.

94. Till, M. S.; Ullmann, G. M. McVol - A Program for Calculating Protein Volumes and Identifying Cavities by a Monte Carlo Algorithm. *J. Mol. Model.* **2010**, *16*, 419–429.

95. Binkowski, T. A. CASTp: Computed Atlas of Surface Topography of Proteins. *Nucleic Acids Res.* **2003**, *31*, 3352–3355.

96. Tian, W.; Chen, C.; Lei, X.; Zhao, J.; Liang, J. CASTp 3.0: Computed Atlas of Surface Topography of Proteins. *Nucleic Acids Res.* **2018**, *46*, W363–W367.

97. Peters, K. P.; Fauck, J.; Frömmel, C. The Automatic Search for Ligand Binding Sites in Proteins of Known Three-Dimensional Structure Using Only Geometric Criteria. *J. Mol. Biol.* **1996**, *256*,

201–213.

98. Tan, K. P.; Varadarajan, R.; Madhusudhan, M. S. DEPTH: A Web Server to Compute Depth and Predict Small-Molecule Binding Cavities in Proteins. *Nucleic Acids Res.* **2011**, *39*, W242–W248.

99. Yu, J.; Zhou, Y.; Tanaka, I.; Yao, M. Roll: A New Algorithm for the Detection of Protein Pockets and Cavities with a Rolling Probe Sphere. *Bioinformatics* **2009**, *26*, 46–52.

100. Ho, B. K.; Gruswitz, F. HOLLOW: Generating Accurate Representations of Channel and Interior Surfaces in Molecular Structures. *BMC Struct. Biol.* **2008**, *8*, 1–6.

101. Kawabata, T.; Go, N. Detection of Pockets on Protein Surfaces Using Small and Large Probe Spheres to Find Putative Ligand Binding Sites. *Proteins Struct. Funct. Bioinforma.* **2007**, *68*, 516–529.

102. Xie, L.; Bourne, P. E. A Robust and Efficient Algorithm for the Shape Description of Protein Structures and Its Application in Predicting Ligand Binding Sites. *BMC Bioinformatics* **2007**, *8*, 1–13.

103. Glaser, F.; Morris, R. J.; Najmanovich, R. J.; Laskowski, R. A.; Thornton, J. M. A Method for Localizing Ligand Binding Pockets in Protein Structures. *Proteins Struct. Funct. Genet.* **2006**, *62*, 479–488.

104. Brady, G. P.; Stouten, P. F. W.; Jr, G. P. B.; Stouten, P. F. W. Fast Prediction and Visualization of Protein Binding Pockets with PASS. *J. Comput. Aided. Mol. Des.* **2000**, *14*, 383–401.

105. Smart, O. S.; Neduvelil, J. G.; Wang, X.; Wallace, B. A.; Sansom, M. S. P. HOLE: A Program for the Analysis of the Pore Dimensions of Ion Channel Structural Models. *J. Mol. Graph.* **1996**, *14*, 354–360.

106. Zhu, H.; Pisabarro, M. T. MSPocket: An Orientation-Independent Algorithm for the Detection of Ligand Binding Pockets. *Bioinformatics* **2011**, *27*, 351–358.

107. Tseng, Y. Y.; Dupree, C.; Chen, Z. J.; Li, W.-H. SplitPocket: Identification of Protein Functional Surfaces and Characterization of Their Spatial Patterns. *Nucleic Acids Res.* **2009**, *37*, W384–W389.

108. Ngan, C.; Hall, D. R.; Zerbe, B.; Grove, L. E.; Kozakov, D.; Vajda, S. FTSite: High Accuracy Detection of Ligand Binding Sites on Unbound Protein Structures. *Bioinformatics* **2012**, *28*, 286–287.

109. Halgren, T. A. Identifying and Characterizing Binding Sites and Assessing Druggability. *J. Chem. Inf. Model.* **2009**, *49*, 377–389.

110. Ghersi, D.; Sanchez, R. EasyMIFs and SiteHound: A Toolkit for the Identification of Ligand-Binding Sites in Protein Structures. *Bioinformatics* **2009**, *25*, 3185–3186.

111. Harris, R.; Olson, A. J.; Goodsell, D. S. Automated Prediction of Ligand-Binding Sites in Proteins. *Proteins Struct. Funct. Bioinforma.* **2007**, *70*, 1506–1517.

112. Laurie, A. T. R.; Jackson, R. M. Q-SiteFinder: An Energy-Based Method for the Prediction of Protein-Ligand Binding Sites. *Bioinformatics* **2005**, *21*, 1908–1916.

113. An, J.; Totrov, M.; Abagyan, R. Pocketome via Comprehensive Identification and Classification of Ligand Binding Envelopes. *Mol. Cell. Proteomics* **2005**, *4*, 752–761.

114. An, J.; Totrov, M.; Abagyan, R. Comprehensive Identification of "Druggable" Protein Ligand Binding Sites. *Genome Inform.* **2004**, *15*, 31–41.

115. Ruppert, J.; Welch, W.; Jain, A. N. Automatic Identification and Representation of Protein Binding Sites for Molecular Docking. *Protein Sci.* **1997**, *6*, 524–533.

116. Goodford, P. J. A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules. *J. Med. Chem.* **1985**, *28*, 849–857.

117. Schneider, S.; Zacharias, M. Combining Geometric Pocket Detection and Desolvation Properties to Detect Putative Ligand Binding Sites on Proteins. *J. Struct. Biol.* **2012**, *180*, 546–550.

118. Morita, M.; Nakamura, S.; Shimizu, K. Highly Accurate Method for Ligand-Binding Site Prediction in Unbound State (Apo) Protein Structures. *Proteins Struct. Funct. Bioinforma.* **2008**, *73*, 468–479.

119. Tuzmen, C.; Erman, B. Identification of Ligand Binding Sites of Proteins Using the Gaussian Network Model. *PLoS One* **2011**, *6*, e16474.

120. Santana, C. A.; Silveira, S. D. A.; Moraes, J. P. A.; Izidoro, S. C.; de Melo-Minardi, R. C.; Ribeiro, A. J. M.; Tyzack, J. D.; Borkakoti, N.; Thornton, J. M. GRaSP: A Graph-Based Residue Neighborhood Strategy to Predict Binding Sites. *Bioinformatics* **2020**, *36*, i726–i734.

121. Wong, G. Y.; Leung, F. H. F.; Ling, S. S. H. Identification of Protein-Ligand Binding Site Using Multi-Clustering and Support Vector Machine. In *IECON 2016 - 42nd Annual Conference of the IEEE Industrial Electronics Society*; IEEE, 2016; pp 939–944.

122. Krivák, R.; Hoksza, D. Improving Protein-Ligand Binding Site Prediction Accuracy by Classification of Inner Pocket Points Using Local Features. *J. Cheminform.* **2015**, *7*, 12.

123. Nayal, M.; Honig, B. On the Nature of Cavities on Protein Surfaces: Application to the Identification of Drug-Binding Sites. *Proteins Struct. Funct. Bioinforma.* **2006**, *63*, 892–906.

124. Yan, X.; Lu, Y.; Li, Z.; Wei, Q.; Gao, X.; Wang, S.; Wu, S.; Cui, S. PointSite: A Point Cloud Segmentation Tool for Identification of Protein Ligand Binding Atoms. *J. Chem. Inf. Model.* **2022**, *62*, 2835–2845.

125. Aggarwal, R.; Gupta, A.; Chelur, V.; Jawahar, C. V; Priyakumar, U. D. DeepPocket: Ligand Binding Site Detection and Segmentation Using 3D Convolutional Neural Networks. *J. Chem. Inf. Model.* **2021**.

126. Kandel, J.; Tayara, H.; Chong, K. T. PUResNet: Prediction of Protein-Ligand Binding Sites Using Deep Residual Neural Network. *J. Cheminform.* **2021**, *13*, 65.

127. Mylonas, S. K.; Axenopoulos, A.; Daras, P. DeepSurf: A Surface-Based Deep Learning Approach for the Prediction of Ligand Binding Sites on Proteins. *Bioinformatics* **2021**, *37*, 1681–1690.

128. Kozlovskii, I.; Popov, P. Spatiotemporal Identification of Druggable Binding Sites Using Deep Learning. *Commun. Biol.* **2020**, *3*, 618.

129. Jiang, M.; Li, Z.; Bian, Y.; Wei, Z. A Novel Protein Descriptor for the Prediction of Drug Binding Sites. *BMC Bioinformatics* **2019**, *20*, 478.

130. Jian, J.; Elumalai, P.; Pitti, T.; Wu, C. Y.; Tsai, K.-C.; Chang, J.-Y.; Peng, H.-P.; Yang, A.-S. Predicting Ligand Binding Sites on Protein Surfaces by 3-Dimensional Probability Density Distributions of Interacting Atoms. *PLoS One* **2016**, *11*, e0160315.

131. Edelsbrunner, H.; Kirkpatrick, D.; Seidel, R. On the Shape of a Set of Points in the Plane. *IEEE Trans. Inf. Theory* **1983**, *29*, 551–559.

132. Jones, J. E. E. On the Determination of Molecular Fields.—I. From the Variation of the Viscosity of a Gas with Temperature. *Proc. R. Soc. London. Ser. A, Contain. Pap. a Math. Phys. Character* **1924**, *106*, 441–462.

133. Jones, J. E. E. On the Determination of Molecular Fields. —II. From the Equation of State of a Gas. *Proc. R. Soc. London. Ser. A, Contain. Pap. a Math. Phys. Character* **1924**, *106*, 463–477.

134. Qi, C. R.; Su, H.; Mo, K.; Guibas, L. J. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. *2016 Fourth Int. Conf. 3D Vis.* **2016**, 601–610.

135. Degac, J.; Winter, U.; Helms, V. Graph-Based Clustering of Predicted Ligand-Binding Pockets on Protein Surfaces. *J. Chem. Inf. Model.* **2015**, *55*, 1944–1952.

136. Huang, B. MetaPocket: A Meta Approach to Improve Protein Ligand Binding Site Prediction. *Omi. A J. Integr. Biol.* **2009**, *13*, 325–330.

137. Zhang, Z.; Li, Y.; Lin, B.; Schroeder, M.; Huang, B. Identification of Cavities on Protein Surface Using Multiple Computational Approaches for Drug Binding Site Prediction. *Bioinformatics* **2011**, *27*, 2083–2088.

138. Hajduk, P. J.; Huth, J. R.; Tse, C. Predicting Protein Druggability REVIEWS. Drug Discov. Today, 10,1675–1682. *Drug Discov. Today* **2005**, *10*, 1675–1682.

139. Cheng, A. C.; Coleman, R. G.; Smyth, K. T.; Cao, Q.; Soulard, P.; Caffrey, D. R.; Salzberg, A. C.; Huang, E. S. Structure-Based Maximal Affinity Model Predicts Small-Molecule Druggability. *Nat. Biotechnol.* **2007**, *25*, 71–75.

140. Fauman, E. B.; Rai, B. K.; Huang, E. S. Structure-Based Druggability Assessment — Identifying Suitable Targets for Small Molecule Therapeutics. *Curr. Opin. Chem. Biol. 15*, 463–468.

141. Hussein, H. A.; Geneix, C.; Petitjean, M.; Borrel, A.; Flatters, D.; Camproux, A. Global Vision of Druggability Issues : Applications and Perspectives. *Drug Discov. Today* **2017**, *22*, 404–415.

142. Volkamer, A.; Kuhn, D.; Grombacher, T.; Rippmann, F.; Rarey, M. Combining Global and Local Measures for Structure-Based Druggability Predictions. *J. Chem. Inf. Model.* **2012**, *52*, 360–372.

143. Perola, E.; Herman, L.; Weiss, J. Development of a Rule-Based Method for the Assessment of Protein Druggability. *J. Chem. Inf. Model.* **2012**, *52*, 1027–1038.

144. Sheridan, R. P.; Maiorov, V. N.; Holloway, M. K.; Cornell, W. D.; Gao, Y.-D. Drug-like Density: A Method of Quantifying the "Bindability" of a Protein Target Based on a Very Large Set of Pockets and Drug-like Ligands from the Protein Data Bank. *J. Chem. Inf. Model.* **2010**, *50*, 2029–2040.

145. Krasowski, A.; Muthas, D.; Sarkar, A.; Schmitt, S.; Brenk, R. DrugPred: A Structure-Based Approach To Predict Protein Druggability Developed Using an Extensive Nonredundant Data Set. *J. Chem. Inf. Model.* **2011**, *51*, 2829–2842.

146. Borrel, A.; Regad, L.; Xhaard, H.; Petitjean, M.; Camproux, A. PockDrug: A Model for Predicting Pocket Druggability That Overcomes Pocket Estimation Uncertainties. *J. Chem. Inf. Model.* **2015**, *55*, 882–895.

147. Edfeldt, F. N. B.; Folmer, R. H. A.; Breeze, A. L. Fragment Screening to Predict Druggability (Ligandability) and Lead Discovery Success. *Drug Discov. Today* **2011**, *16*, 284–287.

148. Tran-Nguyen, V.-K. K.; Da Silva, F.; Bret, G.; Rognan, D. All in One: Cavity Detection, Druggability Estimate, Cavity-Based Pharmacophore Perception, and Virtual Screening. *J. Chem. Inf. Model.* **2019**, *59*, 573–585.

149. Rooklin, D.; Wang, C.; Katigbak, J.; Arora, P. S.; Zhang, Y. AlphaSpace: Fragment-Centric Topographical Mapping to Target Protein-Protein Interaction Interfaces. *J. Chem. Inf. Model.* **2015**, *55*, 1585–1599.

150. Coleman, R. G.; Sharp, K. A. Travel Depth, a New Shape Descriptor for Macromolecules: Application to Ligand Binding. *J. Mol. Biol.* **2006**, *362*, 441–458.

151. Miao, Z.; Westhof, E. RBscore&NBench : A High-Level Web Server for Nucleic Acid Binding Residues Prediction with a Large-Scale Benchmarking Database. *Nucleic Acids Res.* **2016**, *44*,

W562–W567.

152. Trabuco, L. G.; Lise, S.; Petsalaki, E.; Russell, R. B. PepSite: Prediction of Peptide-Binding Sites from Protein Surfaces. *Nucleic Acids Res.* **2012**, *40*, W423–W427.

153. Sehnal, D.; Vařeková, R. S.; Berka, K.; Pravda, L.; Navrátilová, V.; Banáš, P.; Ionescu, C. M.; Otyepka, M.; Koča, J. MOLE 2.0: Advanced Approach for Analysis of Biomacromolecular Channels. *J. Cheminform.* **2013**, *5*, 1–13.

154. Benkaidali, L.; Andre, F.; Maouche, B.; Siregar, P.; Benyettou, M.; Maurel, F.; Petitjean, M. Computing Cavities, Channels, Pores and Pockets in Proteins from Non-Spherical Ligands Models. *Bioinformatics* **2014**, *30*, 792–800.

155. Parca, L.; Gherardini, P. F.; Helmer-Citterich, M.; Ausiello, G. Phosphate Binding Sites Identification in Protein Structures. *Nucleic Acids Res.* **2011**, *39*, 1231–1242.

156. Kinoshita, K.; Furui, J.; Nakamura, H. Identification of Protein Functions from a Molecular Surface Database, EF-Site. *J. Struct. Funct. Genomics* **2002**, *2*, 9–22.

157. Jambon, M.; Imberty, A.; Deléage, G.; Geourjon, C. A New Bioinformatic Approach to Detect Common 3D Sites in Protein Structures. *Proteins Struct. Funct. Bioinforma.* **2003**, *52*, 137–145.

158. Brakoulias, A.; Jackson, R. M. Towards a Structural Classification of Phosphate Binding Sites in Protein-Nucleotide Complexes: An Automated All-against-All Structural Comparison Using Geometric Matching. *Proteins Struct. Funct. Bioinforma.* **2004**, *56*, 250–260.

159. Binkowski, T. A.; Joachimiak, A. Protein Functional Surfaces: Global Shape Matching and Local Spatial Alignments of Ligand Binding Sites. *BMC Struct. Biol.* **2008**, *8*, 45.

160. Andreeva, A.; Kulesha, E.; Gough, J.; Murzin, A. G. The SCOP Database in 2020: Expanded Classification of Representative Family and Superfamily Domains of Known Protein Structures. *Nucleic Acids Res.* **2020**, *48*, D376–D382.

161. Xie, L.; Bourne, P. E. Detecting Evolutionary Relationships across Existing Fold Space, Using Sequence Order-Independent Profile–Profile Alignments. *Proc. Natl. Acad. Sci.* **2008**, *105*, 5441–5446.

162. Milletti, F.; Vulpetti, A.; F., M.; A., V. Predicting Polypharmacology by Binding Site Similarity: From Kinases to the Protein Universe. *J. Chem. Inf. Model.* **2010**, *50*, 1418–1431.

163. Hoffmann, B.; Zaslavskiy, M.; Vert, J.; Stoven, V. A New Protein Binding Pocket Similarity Measure Based on Comparison of Clouds of Atoms in 3D: Application to Ligand Prediction. *BMC Bioinformatics* **2010**, *11*, 99.

164. Yeturu, K.; Chandra, N. PocketAlign A Novel Algorithm for Aligning Binding Sites in Protein Structures. *J. Chem. Inf. Model.* **2011**, *51*, 1725–1736.

165. Liu, T.; Altman, R. B. Using Multiple Microenvironments to Find Similar Ligand-Binding Sites: Application to Kinase Inhibitor Binding. *PLoS Comput. Biol.* **2011**, *7*, e1002326.

166. Sael, L.; Kihara, D. Detecting Local Ligand-Binding Site Similarity in Nonhomologous Proteins by Surface Patch Comparison. *Proteins Struct. Funct. Bioinforma.* **2012**, *80*, 1177–1195.

167. Ellingson, L.; Zhang, J. Protein Surface Matching by Combining Local and Global Geometric Information. *PLoS One* **2012**, *7*.

168. von Behren, M. M.; Volkamer, A.; Henzler, A. M.; Schomburg, K. T.; Urbaczek, S.; Rarey, M. Fast Protein Binding Site Comparison via an Index-Based Screening Technology. *J. Chem. Inf. Model.* **2013**, *53*, 411–422.

169. Brylinski, M.; Feinstein, W. P. EFindSite: Improved Prediction of Ligand Binding Sites in

Protein Models Using Meta-Threading, Machine Learning and Auxiliary Ligands. *J. Comput. Aided. Mol. Des.* **2013**, *27*, 551–567.

170. Chartier, M.; Najmanovich, R. Detection of Binding Site Molecular Interaction Field Similarities. *J. Chem. Inf. Model.* **2015**, *55*, 1600–1615.

171. Chen, Y. C.; Tolbert, R.; Aronov, A. M.; McGaughey, G.; Walters, W. P.; Meireles, L. Prediction of Protein Pairs Sharing Common Active Ligands Using Protein Sequence, Structure, and Ligand Similarity. *J. Chem. Inf. Model.* **2016**, *56*, 1734–1745.

172. Batista, J.; Hawkins, P. C.; Tolbert, R.; Geballe, M. T. SiteHopper - a Unique Tool for Binding Site Comparison. *J. Cheminform.* **2014**, *6*, P57.

173. Pu, L.; Govindaraj, R. G.; Lemoine, J. M.; Wu, H. C.; Brylinski, M. Deepdrug3D: Classification of Ligand-Binding Pockets in Proteins with a Convolutional Neural Network. *PLoS Comput. Biol.* **2019**, *15*, e1006718.

174. Li, S.; Cai, C.; Gong, J.; Liu, X.; Li, H. A Fast Protein Binding Site Comparison Algorithm for Proteome-wide Protein Function Prediction and Drug Repurposing. *Proteins Struct. Funct. Bioinforma.* **2021**, *89*, 1541–1556.

175. Bhadra, A.; Yeturu, K. Site2Vec: A Reference Frame Invariant Algorithm for Vector Embedding of Protein–Ligand Binding Sites. *Mach. Learn. Sci. Technol.* **2021**, *2*, 015005.

176. Haupt, V. J.; Daminelli, S.; Schroeder, M. Drug Promiscuity in PDB: Protein Binding Site Similarity Is Key. *PLoS One* **2013**, *8*.

177. Kahraman, A.; Morris, R. J.; Laskowski, R. A.; Thornton, J. M. Shape Variation in Protein Binding Pockets and Their Ligands. *J. Mol. Biol.* **2007**, *368*, 283–301.

178. Bron, C.; Kerbosch, J. Algorithm 457: Finding All Cliques of an Undirected Graph. *Commun. ACM* **1973**, *16*, 575–577.

179. Johnston, H. C. Cliques of a Graph-Variations on the Bron-Kerbosch Algorithm. *Int. J. Comput. Inf. Sci.* **1976**, *5*, 209–238.

180. Desaphy, J.; Raimbaud, E.; Ducrot, P.; Rognan, D. Encoding Protein–Ligand Interaction Patterns in Fingerprints and Graphs. *J. Chem. Inf. Model.* **2013**, *53*, 623–637.

181. Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and Scoring in Virtual Screening for Drug Discovery: Methods and Applications. *Nat. Rev. Drug Discov.* **2004**, *3*, 935–949.

182. Henikoff, S.; Henikoff, J. G. Position-Based Sequence Weights. *J. Mol. Biol.* **1994**, *243*, 574–578.

183. Henikoff, S.; Henikoff, J. G. Amino Acid Substitution Matrices from Protein Blocks. *Proc. Natl. Acad. Sci.* **1992**, *89*, 10915–10919.

184. Sillitoe, I.; Bordin, N.; Dawson, N.; Waman, V. P.; Ashford, P.; Scholes, H. M.; Pang, C. S. M.; Woodridge, L.; Rauer, C.; Sen, N.; Abbasian, M.; Le Cornu, S.; Lam, S. D.; Berka, K.; Varekova, I. H.; Svobodova, R.; Lees, J.; Orengo, C. A. CATH: Increased Structural Coverage of Functional Space. *Nucleic Acids Res.* **2021**, *49*, D266–D273.

185. Bateman, A. UniProt: A Worldwide Hub of Protein Knowledge. *Nucleic Acids Res.* **2019**, *47*, D506–D515.

186. Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.* **2012**, *55*, 6582–6594.

187. Tran-Nguyen, V.-K.; Jacquemard, C.; Rognan, D. LIT-PCBA: An Unbiased Data Set for

Machine Learning and Virtual Screening. *J. Chem. Inf. Model.* **2020**, *60*, 4263–4273.

188. Barelier, S.; Sterling, T.; O'Meara, M. J.; Shoichet, B. K. The Recognition of Identical Ligands by Unrelated Proteins. *ACS Chem. Biol.* **2015**, *10*, 2772–2784.

189. Govindaraj, R. G.; Brylinski, M. Comparative Assessment of Strategies to Identify Similar Ligand-Binding Pockets in Proteins. *BMC Bioinformatics* **2018**, *19*, 1–17.

190. Mestres, J.; Gregori-Puigjané, E.; Valverde, S.; Solé, R. V. Data Completeness—the Achilles Heel of Drug-Target Networks. *Nat. Biotechnol.* **2008**, *26*, 983–984.

191. Hu, Y.; Bajorath, J. High-Resolution View of Compound Promiscuity. *F1000Research* **2013**, *2*, 144.

192. Kuhn, D.; Weskamp, N.; Hüllermeier, E.; Klebe, G. Functional Classification of Protein Kinase Binding Sites Using Cavbase. *ChemMedChem* **2007**, *2*, 1432–1447.

193. Desaphy, J.; Bret, G.; Rognan, D.; Kellenberger, E. Sc-PDB: A 3D-Database of Ligandable Binding Sites-10 Years On. *Nucleic Acids Res.* **2015**, *43*, D399–D404.

194. Niu, M.; Hu, J.; Wu, S.; Zhang, X.; Xu, H.; Zhang, Y.; Zhang, J.; Yang, Y. Structural Bioinformatics-Based Identification of EGFR Inhibitor Gefitinib as a Putative Lead Compound for BACE. *Chem. Biol. Drug Des.* **2014**, *83*, 81–88.

195. Shortridge, M. D.; Powers, R. Structural and Functional Similarity between the Bacterial Type III Secretion System Needle Protein PrgI and the Eukaryotic Apoptosis Bcl-2 Proteins. *PLoS One* **2009**, *4*, e7442.

196. Miguel, A.; Hsin, J.; Liu, T.; Tang, G.; Altman, R. B.; Huang, K. C. Variations in the Binding Pocket of an Inhibitor of the Bacterial Division Protein FtsZ across Genotypes and Species. *PLOS Comput. Biol.* **2015**, *11*, e1004117.

197. Xie, L.; Wang, J.; Bourne, P. E. In Silico Elucidation of the Molecular Mechanism Defining the Adverse Effect of Selective Estrogen Receptor Modulators. *PLoS Comput. Biol.* **2007**, *3*, e217.

198. Kinnings, S. L.; Liu, N.; Buchmeier, N.; Tonge, P. J.; Xie, L.; Bourne, P. E. Drug Discovery Using Chemical Systems Biology: Repositioning the Safe Medicine Comtan to Treat Multi-Drug and Extensively Drug Resistant Tuberculosis. *PLoS Comput. Biol.* **2009**, *5*, e1000423.

199. de Franchi, E.; Schalon, C.; Messa, M.; Onofri, F.; Benfenati, F.; Rognan, D. Binding of Protein Kinase Inhibitors to Synapsin I Inferred from Pair-Wise Binding Site Similarity Measurements. *PLoS One* **2010**, *5*, e12214.

200. Yang, Y.; Li, G.; Zhao, D.; Yu, H.; Zheng, X.; Peng, X.; Zhang, X.; Fu, T.; Hu, X.; Niu, M.; Ji, X.; Zou, L.; Wang, J. Computational Discovery and Experimental Verification of Tyrosine Kinase Inhibitor Pazopanib for the Reversal of Memory and Cognitive Deficits in Rat Model Neurodegeneration. *Chem. Sci.* **2015**, *6*, 2812–2821.

201. Kinnings, S. L.; Jackson, R. M. Binding Site Similarity Analysis for the Functional Classification of the Protein Kinase Family. *J. Chem. Inf. Model.* **2009**, *49*, 318–329.

202. Al-Gharabli, S. I.; Ali Shah, S. T.; Weik, S.; Schmidt, M. F.; Mesters, J. R.; Kuhn, D.; Klebe, G.; Hilgenfeld, R.; Rademann, J. An Efficient Method for the Synthesis of Peptide Aldehyde Libraries Employed in the Discovery of Reversible SARS Coronavirus Main Protease (SARS-Cov Mpro) Inhibitors. *ChemBioChem* **2006**, *7*, 1048–1055.

203. Powers, R.; Copeland, J. C.; Germer, K.; Mercier, K. A.; Ramanathan, V.; Revesz, P. Comparison of Protein Active Site Structures for Functional Annotation of Proteins and Drug Design. *Proteins Struct. Funct. Bioinforma.* **2006**, *65*, 124–135.

204. Xie, L.; Evangelidis, T.; Xie, L.; Bourne, P. E. Drug Discovery Using Chemical Systems

Biology: Weak Inhibition of Multiple Kinases May Contribute to the Anti-Cancer Effect of Nelfinavir. *PLoS Comput. Biol.* **2011**, *7*, e1002037.

205. Dekker, F. J.; Koch, M. A.; Waldmann, H. Protein Structure Similarity Clustering (PSSC) and Natural Product Structure as Inspiration Sources for Drug Development and Chemical Genomics. *Curr. Opin. Chem. Biol.* **2005**, *9*, 232–239.

206. Willmann, D.; Lim, S.; Wetzel, S.; Metzger, E.; Jandausch, A.; Wilk, W.; Jung, M.; Forne, I.; Imhof, A.; Janzer, A.; Kirfel, J.; Waldmann, H.; Schüle, R.; Buettner, R. Impairment of Prostate Cancer Cell Growth by a Selective and Reversible Lysine-Specific Demethylase 1 Inhibitor. *Int. J. Cancer* **2012**, *131*, 2704–2709.

207. Keren-Kaplan, T.; Attali, I.; Estrin, M.; Kuo, L. S.; Farkash, E.; Jerabek-Willemsen, M.; Blutraich, N.; Artzi, S.; Peri, A.; Freed, E. O.; Wolfson, H. J.; Prag, G. Structure-Based in Silico Identification of Ubiquitin-Binding Domains Provides Insights into the ALIX-V:Ubiquitin Complex and Retrovirus Budding. *EMBO J.* **2013**, *32*, 538–551.

208. Schirris, T. J. J.; Ritschel, T.; Herma Renkema, G.; Willems, P. H. G. M.; Smeitink, J. A. M.; Russel, F. G. M. Mitochondrial ADP/ATP Exchange Inhibition: A Novel off-Target Mechanism Underlying Ibipinabant-Induced Myotoxicity. *Sci. Rep.* **2015**, *5*, 14533.

209. Haydon, D. J.; Stokes, N. R.; Ure, R.; Galbraith, G.; Bennett, J. M.; Brown, D. R.; Baker, P. J.; Barynin, V. V; Rice, D. W.; Sedelnikova, S. E.; Heal, J. R.; Sheridan, J. M.; Aiwale, S. T.; Chauhan, P. K.; Srivastava, A.; Taneja, A.; Collins, I.; Errington, J.; Czaplewski, L. G. An Inhibitor of FtsZ with Potent and Selective Anti-Staphylococcal Activity. *Science (80-. ).* **2008**, *321*, 1673–1675.

210. Dunlop, J.; Lock, T.; Jow, B.; Sitzia, F.; Grauer, S.; Jow, F.; Kramer, A.; Bowlby, M. R.; Randall, A.; Kowal, D.; Gilbert, A.; Comery, T. A.; LaRocque, J.; Soloveva, V.; Brown, J.; Roncarati, R. Old and New Pharmacology: Positive Allosteric Modulation of the A7 Nicotinic Acetylcholine Receptor by the 5-Hydroxytryptamine 2B/C Receptor Antagonist SB-206553 (3,5-Dihydro-5-Methyl- N -3-Pyridinylbenzo[1,2- B :4,5- b ']Di Pyrrole-1(2 H )-Carboxamide). *J. Pharmacol. Exp. Ther.* **2009**, *328*, 766–776.

211. Möller-Acuña, P.; Contreras-Riquelme, J. S.; Rojas-Fuentes, C.; Nuñez-Vivanco, G.; Alzate-Morales, J.; Iturriaga-Vásquez, P.; Arias, H. R.; Reyes-Parada, M. Similarities between the Binding Sites of SB-206553 at Serotonin Type 2 and Alpha7 Acetylcholine Nicotinic Receptors: Rationale for Its Polypharmacological Profile. *PLoS One* **2015**, *10*, e0134444.

212. Spitzer, R.; Cleves, A. E.; Varela, R.; Jain, A. N. Protein Function Annotation by Local Binding Site Surface Similarity. *Proteins Struct. Funct. Bioinforma.* **2014**, *82*, 679–694.

213. Cleves, A. E.; Jain, A. N. Chemical and Protein Structural Basis for Biological Crosstalk between PPARα and COX Enzymes. *J. Comput. Aided. Mol. Des.* **2015**, *29*, 101–112.

214. Edwards, A. What Are the Odds of Finding a COVID-19 Drug from a Lab Repurposing Screen? *J. Chem. Inf. Model.* **2020**, *60*, 5727–5729.

215. Lewis, R. A. Best Practices for Repurposing Studies. *J. Comput. Aided. Mol. Des.* **2021**, *35*, 1189–1193.

# CHAPTER 2

Development of a new method for local
comparison of protein pockets

## 2.1.   Scope, motivations, and novelty

In the previous chapter, we learned why protein pocket comparisons are useful and important in drug design. We navigated through a broad range of state-of-the-art methods, which differ in how they simultaneously represent, compare pockets and score their similarity. We think that the variety of methods is an asset with respect to the difficulty in estimating pocket similarities and the quite different applicability domains. The current work was initiated with this in mind. We strikingly observed the underrepresentation of local comparison algorithms, which to our perspective, are suitable for comparing pockets of different sizes. Thus, small protein areas that can bind fragment-sized moieties (subpockets) can be appropriately compared to an entire pocket. The subsequent possibilities for drug design looked promising.

By building on a previous work in our lab where a protein pocket is represented as a three-dimensional (3D) cloud of annotated points (VolSite,[1] see **Chapter 1**), we aimed at exploring image recognition approaches. Computer vision algorithms have been used in the field for decades, particularly in alignment-based approaches.[2–7]

Herein, we introduced for the first time the application of sampling-based point cloud registration (PCR) to the binding site comparison problem. PCR is originally applied to millions of points which represent the surface of any kind of objects (tables, buildings, scenes, etc.). More information is given in **section 2.2.** We later found that at the time of this study, PCR only started being applied to ligand surfaces comparison[8] while the shape descriptor has been used for classification of entire protein structures.[9,10] Independently, the choice of this algorithm was motivated by the resemblance between the standard 3D image inputs and our pocket representation. Both are ensemble of 3D points with annotations: RGB color for the first and distinct pharmacophoric properties for the second. However, the small-size (a few hundred of points), sparseness, grid regularity, volumetric nature of the pocket clouds instead of surfaces, and the definition of pocket edges questioned the applicability of PCR to our problem.

To delineate the two problems, common tasks of PCR would superpose objects which are known to share overlapping areas. In the binding site comparison case, whether there is any overlapping area is an additional variable to be estimated.

In this chapter, we have prototyped, optimized, and benchmarked a point cloud registration algorithm to compare protein pockets. The open-source method has been publicly released at https://github.com/kimeguida/ProCare.

## 2.2. Previous work

This section only aims at summarizing the knowledge relevant to this chapter. For more details, we refer the reader to the original papers.

### 2.2.1. Source of druggable protein-ligand complexes

The screening Protein Data Bank (sc-PDB)[11] is a public database of curated protein-ligand complexes, compiled by our laboratory. It was first released in 2006 and updated along the years.[11–13] It aims at providing a non-redundant subset of the PDB, useful to find relevant starting data for structure-based screening. Structures are selected according to several filtering rules: structure resolution, consistency of annotations from different sources, nature of amino acids, the presence of pharmacological and buried ligands. A careful treatment (e.g. ionization and protonation states of both protein and ligand atoms, keeping bound water molecules) finally yields protein chain and ligand structures available in MOL2 formats, offering the advantage of atom type information and connectivity table. The database actually provides more materials and services than stated. The 2016 archive consisted of ~16,000 unique protein-ligand X-ray structures made of 4755 unique proteins, and 6326 unique ligands. The recent 2022 archive (Bret *et al.*, unpublished data) consists of ~37,000 unique protein-ligand complexes (X-ray, NMR, cryo-EM), 7105 unique proteins and 13993 unique ligands. We draw attention to the fact that protein-ligand redundancy was removed by binding mode analysis. In other words, only a representative protein-ligand complex is kept out of the available PDB copies, even when residues slightly deviate due to local flexibility. Outcomes for binding site comparison may be a loss of information. Nonetheless, this database is a sufficient data source to evaluate and apply our method.

### 2.2.2. Point cloud registration

In computer vision, point cloud registration is the process of finding a transformation, i.e., the rotation, translation and scaling that adequately superpose two overlapping clouds. It falls within the general registration problem, whose applications span object reconstruction in robotics, medical imaging, photography, cinematography, etc. Objects are modeled as two-dimensional (2D) or three-dimensional (3D) color images when associated with a depth (RGB-D).[14] The depth information is the distance between each pixel and a fixed reference, the camera. Hence, the 3D shapes of objects are characterized. These data points are collected via range imaging techniques such as LIDAR (light detection and ranging), tomography scanning, structured-light 3D scanners, time of flight 3D scanners, and represented as point clouds, or processed into meshes and voxels by appropriate methods (**Figure 2.1**). It is interesting to note that point clouds are unstructured and unordered data, without neighborhood

information, and describing the surface of objects (i.e. what the camera can see). Contrarily, the point clouds of protein cavities are volumetric data (i.e. any position in the cavity is independent of the viewpoint), obtained first via voxelization.
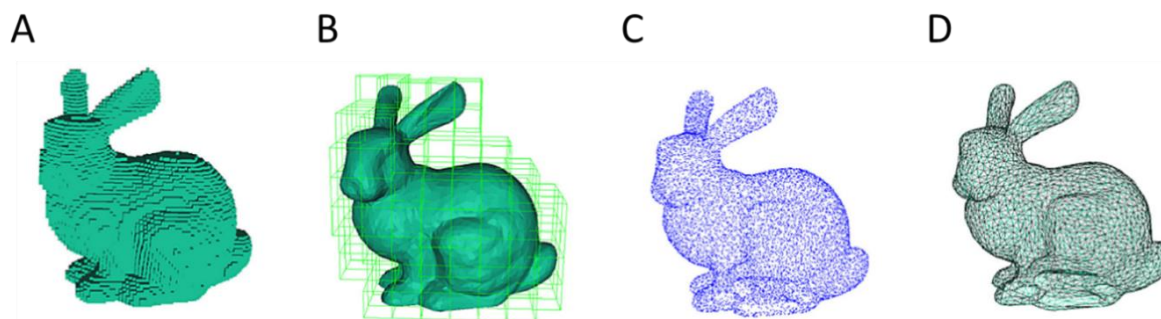


**Figure 2.1.** Examples of different 3D representations. The Stanford Bunny model in A) voxel, B) sparse voxel octree, C) point cloud, and D) mesh. Adapted from Fahim et al (2021).[15]

There are two scenarios of (point cloud) registration. In the first case, a set of correspondences between the two models is known. In that respect, the registration task consists of finding the best alignment that minimizes the superposition error. Estimating a transformation is a non-trivial exercise, influenced by the presence of noise and the planarity of the sets.[16] This is to account for when developing alignment-based binding site comparison methods, where scoring and chances to detect similarity rely on proposed superposition. In linear algebra, solutions to various definitions of the orthogonal Procrustes problem are searched.[16,17] The Kabsch algorithm is popular in the structural biology field to estimate a proper rotation.[2,18] Translation is estimated by alignment of centroids. This singular value decomposition-based solution was first introduced by Schönemann (1966), later proposed by Arun et al. (1987) and other studies.[16,19] In 1991, Umeyama refined the Arun's solution to handle noisy data.[20] This implementation is used in our method. Other solutions have been reported, based on orthonormal matrices, or quaternions where both rotation and translation are calculated.[21–23]

In the second registration scenario, there is no prior knowledge of equivalent points. It is a variable to be estimated. Correspondence estimation is one of the fundamental problems in computer vision. The iterative closest point (ICP)[24,25] is a well know algorithm which repeatedly, associates the closest points in the Euclidian space as correspondences and estimates a transformation until convergence. This solution is not efficient and is sensitive to the initial guess, i.e. a good alignment is obtained provided a good initial orientation. Also, ICP is prone to be trapped in a local minimum. To solve this issue, other methods were implemented for global optimization of the alignment.[26–28] Alternatively, shape descriptors were developed to systematically recognize similar local areas in objects, including machine-learning-based approaches.[15,29–32] In our studies, data-driven approaches were first disregarded due to the amount of data available and the quest for interpretability. Geometry-based approaches seemed

suitable for our goals and were therefore investigated. Major open source and maintained packages for point cloud processing and registration are listed **Table 2.1**.

**Table 2.1.** Community open-source packages for point cloud processing and registration

| Name | Source | Language |
| --- | --- | --- |
| CloudCompare | cloudcompare.org | C++ |
| Open3D | www.open3d.org | C++, Python |
| OpenCV | opencv.org | C++, Python, Java, MATLAB |
| Point Cloud Library PCL | pointclouds.org | C++ |

At the time of this study, PCL has not been maintained for a while whereas its reimplementation Open3D was being actively improved and offered two programming language interfaces. Hence, Open3D was prioritized for our method development.

## 2.3. A computer vision approach to align and compare protein cavities: Application to fragment-based drug design

This section was integrally published in:

The open source code is available at: https://github.com/kimeguida/ProCare
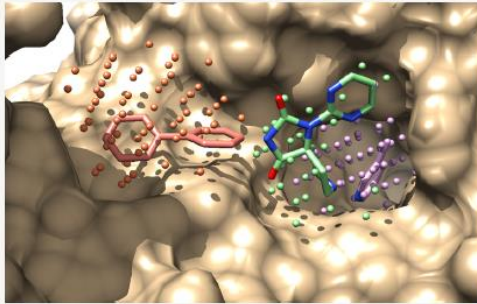
### A Computer Vision Approach to Align and Compare Protein Cavities: Application to Fragment-Based Drug Design

Merveille Eguida and Didier Rognan*

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

**ABSTRACT:** Identifying local similarities in binding sites from distant proteins is a major hurdle to rational drug design. We herewith present a novel method, borrowed from computer vision, adapted to mine fragment subpockets and compare them to whole ligand-binding sites. Pockets are represented by pharmacophore-annotated point clouds mimicking ideal ligands or fragments. Point cloud registration is used to find the transformation enabling an optimal overlap of points sharing similar topological and pharmacophoric neighborhoods. The method (ProCare) was calibrated on a large set of druggable cavities and applied to the comparison of fragment subpockets to entire cavities. A collection of 33,953 subpockets annotated with their bound fragments was screened for local similarity to cavities from recently described protein X-ray structures. ProCare was able to detect local similarities between remote pockets and transfer the corresponding fragments to the query cavity space, thereby proposing a first step to fragment-based design approaches targeting orphan cavities.

## 2.3.1. Abstract

Identifying local similarities in binding sites from distant proteins is a major hurdle to rational drug design. We herewith present a novel method, borrowed from computer vision, adapted to mine fragment subpockets and compare them to whole ligand-binding sites. Pockets are represented by pharmacophore-annotated point clouds mimicking ideal ligands or fragments. Point cloud registration is used to find the transformation enabling an optimal overlap of points sharing similar topological and pharmacophoric neighborhoods. The method (ProCare) was calibrated on a large set of druggable cavities, and applied to the comparison of fragment subpockets to entire cavities. A collection of 33,953 subpockets annotated with their bound fragments was screened for local similarity to cavities from recently described protein X-ray structures. ProCare was able to detect local similarities between remote pockets and transfer the corresponding fragments to the query cavity space, thereby proposing a first step to fragment-based design approaches targeting orphan cavities.

## 2.3.2. Introduction

Three-dimensional (3D) structures of protein-ligand complexes are the corner stones of structure-based rational approaches to ligand design.[1] Among the many computational methods[2] to infer putative relationships between ligand and target spaces, detection and pairwise comparison of protein-ligand binding sites have gained considerable popularity in the last decade.[3-5] Potential cavities can be first detected at the surface of macromolecules using a myriad of computational tools,[5] classically grouped in three categories: geometry-based (e.g. CavBase,[6] VolSite,[7] Fpocket[8]), energy-based (e.g. GRID,[9] Q-SiteFinder[10]) and evolutionary-based (e.g. SURFNET-ConSurf[11]), although some methods may combine different approaches (e.g. Ligsitecsc[12], SiteMap[13]). Whereas geometry-based approaches rely on the prior calculation of the target's molecular surface to identify accessible pockets, energy-based methods compute interaction energies on a 3D lattice between the target protein and several probe atoms. Last, evolutionary-based tools require a multiple sequence or structural alignment of targets from the same family to pinpoint evolutionary conserved motifs that can be linked to the recognition of specific ligand structures. Interestingly, structural druggability or ligandability,[14] the propensity to accommodate high-affinity drug-like ligands, can be computed on the fly using machine-learning models[8, 7] trained on sets of known druggable and undruggable sites. Once pockets have been detected, they can be systematically compared at a high-throughput to detect global similarities even in absence of fold conservation.[5] Many descriptors (fingerprints, distance counts, pharmacophoric triplets, grid points, point clouds, graphs, and shapes) of protein-ligand binding pockets can be used by geometric hashing[15] or clique detection[6] algorithms to find the most prominent shared features guiding the structural alignment of protein cavities.

Following the basic principle that similar cavities recognize similar ligands, protein-ligand binding site comparison methods have been successfully used in many drug discovery scenarios: (i) assigning a function from a target's 3D structure,[16-18] (ii) finding hits for a novel target,[19] (iii) prioritizing compound library design,[20] (iv) repurposing ancient drugs for new targets,[21-23] (v) explaining the poypharmacological profile of known drugs,[24] (vi) predicting unexpected off-targets[25-28] and extending potential binding sites to new areas of target space.[29-30] A practical guide to navigate across all available methods and benchmarking data sets has been recently described.[31]

Most of above-described methods consider pocket similarity from a global and not a local point of view. In other words, current methods usually estimate the similarity between whole 3D objects (pockets) without specifically rewarding the microenvironments (subpockets) responsible for that similarity. For related protein pairs (e.g. serine/threonine protein kinases, aminergic G protein-coupled receptors), a good alignment and similarity estimate will be found. However, current methods will generally fail to find correspondences between binding pockets from totally unrelated proteins. The consequences are two-fold. First, the proposed initial 3D alignment of both pockets will prioritize global properties (e.g.

molecular shape, principle axes and moments of inertia) over particular microenvironments. A wrong preliminary misalignment will therefore not be corrected after refinement and will lead to erroneous similarity estimates. Second, inferring ligand information from pocket similarity searches (e.g. merging ligand coordinates from one reference pocket to a target cavity) will address the entire ligand structure as a whole, without any obvious clues about which ligand substructure ideally fits which subpocket. Therefore, most existing computational methods are well suited to repurpose existing ligands for new pockets,[21-23] but not to prioritize ligand fragments for specific protein subsites, a very important process in fragment-based drug discovery.[32]

Fewer examples of subpocket comparisons are available to date.[6, 33-40] Existing approaches follow a common flowchart made of four steps: (i) fragmentation of protein-bound PDB ligands into smaller pieces; (ii) registration of protein-ligand non covalent interactions; (iii) definition of protein microenvironments interacting with above-reported ligand chemical moieties; (iv) mathematical representation of the microenvironment into a graph, pharmacophore or fingerprint; (v) pairwise similarity calculation between a reference and a query microenvironment.

Reported methods differ in the level of ligand fragmentation (few connected atoms,[33] chemical group,[34] fragment[35-39]), the atomic definition of protein microenvironments (atom[33] or residue[35] based, surface feature pseudoatoms[21, 37, 39]), the computational representation of the subpocket (graph,[33, 36-38] fingerprint[34, 39]), the alignment method (clique detection,[6] rigid-body transformation,[34] rmsd alignment[35]) and the scoring function (simple Tanimoto or cosine metric,[36-39] shape and/or pharmacophore overlap,[33-34, 38] rmsd of key atoms[35]) to estimate pairwise pocket similarity. To the best of our knowledge, only retrospective validation of subpocket comparisons have been proposed, one of the most impressive being the *a posteriori* molecular explanation to the unexpected cross-reactivity of cyclooxygenase-2 inhibitors with human carbonic anhydrase.[21] Moreover, most approaches are focusing on fragment-bound sub-cavities and cannot easily predict local similarities between the whole of a novel cavity and a collection of microenvironments. Last, the lack of availability of most methods (KRIPO[36] being a noticeable exception) hampers the usage of above-described tools.

There is therefore still a need for novel computational methods, notably those relying on novel cavity representations and alternative alignment methods, applicable at a high throughout to compare entire cavities to fragment-annotated protein microenvironment collections. Following the above guidelines, we herewith present a novel pocket comparison method (ProCare: Protein Cavity registration), particularly adapted to detect local similarity between entire cavities and fragment subpockets, that significantly differs from existing computational tools. ProCare utilizes the concept of point cloud registration, widely used in computer vision to compare and align 2D/3D images. We first describe the implementation of the method to align and compare entire cavities. After parameter optimization and fine-tuning a scoring function to evaluate pocket similarity, we then apply the new method to the

comparison of fragment subpockets to full cavities, thereby enabling to fill new binding pockets with complementary fragments.

### 2.3.3. Results and discussion

In computer vision, pattern recognition, and robotics, point cloud registration[41,42,43] is the process of finding the best spatial transformation (e.g., scaling, rotation and translation) that aligns two point clouds (**Figure 1**).



**Figure 1.** Schematic representation of point cloud registration. The red cloud is rotated and translated along its three main axes until the optimal alignment to the green cloud is found.

Since this concept may not be familiar to medicinal chemists, we here provide a brief summary of the underlying principles and algorithms. The basic principle behind registration of two clouds of points (cloud 1 and cloud 2) requires to first identifying pairs of equivalent points. Two points, respectively in cloud 1 and cloud 2, will be considered equivalent if they are sharing a similar microenvironment, in other words a similar topological arrangement of their neighboring points. Because the aim is to match two geometrical shapes, the environment of a point is herein described by a histogram of angular values called fast point feature histogram or FPFH (see Computational methods). For example, one can imagine discriminating between carbon atoms in 2D representations of cyclobutyl and cyclohexyl moieties, as we would do for the corners of a square and a hexagon, respectively. Since each descriptor of the FPFH is a "count" of a certain angle value range, the similarity of two FPFHs can be estimated via a simple Euclidian distance. However, the FPFH although complex, cannot avoid ambiguities in detecting correspondences, especially when there exist irrelevant points (called outliers) that should not be considered. A solution to rule out outlier points is the Random Sample Consensus (RANSAC) algorithm[44-45]. At each RANSAC iteration, a few points are randomly sampled in cloud 1, their

corresponding points in cloud 2 are assigned, the relevance of these correspondences is verified by comparing the topological distances and finally a rotation/translation is estimated to align the sampled sets. This preliminary alignment, based on only a few points, is then refined with an iterative closest point (ICP) method. ICP is an iterative algorithm[46] that minimizes the overall root-mean square deviation between corresponding points in both clouds.

Interestingly, point cloud registration has rarely been used to overlay molecular surfaces of proteins[47-48] and ligands.[49] With respect to previous approaches using recognition algorithms to compare protein cavities,[50-51] we here take advantage of our previous work describing a protein pocket by a point cloud located in ligand space.[7] The cloud is described as an ensemble of 3D points regularly filling the pocket, each point having a specific pharmacophoric property ("color") complementary to that of the nearby protein environment.[7] The cloud is therefore bigger (200-300 points), regular and complementary in shape and pharmacophoric properties to flanking protein residues. We will first demonstrate the proof-of-concept of applying this computational method to the problem of protein cavity alignments, next fine tune a set of parameters enabling an optimal performance on a large dataset of known cavities, and then propose a physicochemically relevant score to quantify the alignment and pocket similarity. Last, we will apply the optimized method to the specific problem of finding local similarities between fragment subpockets and whole cavities.

**ProCare implementation and parameter optimization**

Preliminary attempts suggested that many parameters of point cloud registration strongly influence the quality of the alignment. We therefore systematically studied 15 key parameters (**Table 1**, Computational methods) by enumerating 157,465 parameter combinations in order to consider their effect of as well as their interdependencies. To test all these conditions, a very simple data set of five similar pairs completed by five dissimilar cavity pairs (EASY1 set; **Table S1**, Computational methods) was designed, just to filter out those parameter combinations that failed in either producing any kind of alignment (fitness = 0), or could not perfectly discriminate similar from dissimilar pairs (ROC AUCs < 1). These two simple filters enabled to decrease the number of potential combinations from 157,465 to 20,181 (**Figure 2**).

**Alignment conditions** 157465

Fitness > 0

□ Filtered with EASY1 set
■ Filtered with BO1 set

79698

ROC AUC = 1

20181

RMSD < 2 Å

314

**Best alignment condition:
Top 5 ROC AUC, manual inspection**
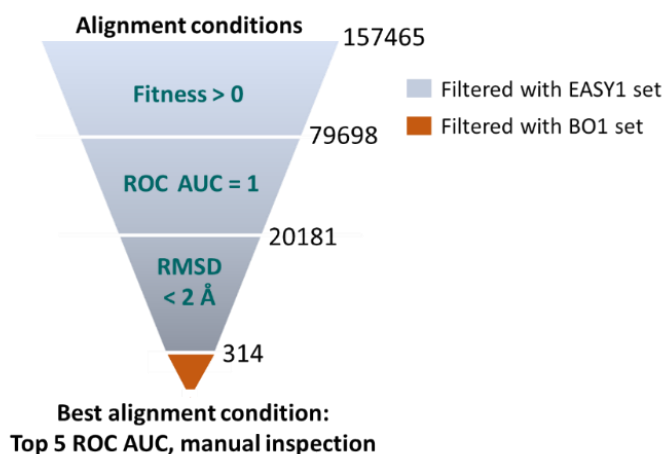
**Figure 2.** Selection procedure to determine the best alignment parameters. 157,465 different conditions (a set of parameters) were initially enumerated and non-relevant conditions filtered-out with the EASY1 set. The 314 remaining alignment conditions were evaluated with the BO1 set and the best one selected by its discrimination performance (high ROC AUC) and manual inspection.

For the remaining possibilities, the output transformation matrices were applied to the protein coordinates of the similar pairs to ensure whether the corresponding protein structures were correctly aligned (rmsd on backbone heavy atoms < 2 Å) or not. A total number of 314 combinations (0.2 % of the total number) still fulfilled the above-described requirements. In order to benchmark the 314 remaining alignment conditions, we designed a larger and much more diverse data set (BO1 set, **Tables S2 and S3**, see Computational methods) of similar pairs and dissimilar pairs of cavities starting from the sc-PDB archive of 16,034 druggable-protein-ligand complexes.[52] The BO1 data set consists of 766 pairs of non-redundant VolSite cavities (383 similar pairs, 383 dissimilar pairs) covering 507 different proteins (460 in the set of similar, 178 in the set of dissimilar), 62 different sets of Uniprot functional annotations for similar pairs and 38 for dissimilar pairs (**Figure S1**).

The 314 pre-selected conditions were used to align cavity pairs from the BO1 set. The area under the ROC curve (ROC AUC) of a binary classification (similar, dissimilar) was calculated to rank each condition using three possible scoring functions (*ph4-strict*, *ph4-rules* and *ph4-ext*) differing by the fuzziness of allowed pharmacophoric matches (see Computational methods). We finally selected the best alignment condition (see parameters in **Table S4**) that yielded a ROC AUC value of 0.87 (CI = [0.85;0.89]), based on the *ph4-ext* scoring. Although the current approach was successful in aligning and ranking cavity pairs from a large and diverse data set, we observed that some pairs of similar cavities still remained misaligned (see example in **Figure S2**). Constraining the alignment to consider both shape and color might solve the problem. However, the existing colored-ICP algorithm[53] which aims at optimizing both geometric (shape) and photometric (colors) terms is not suited here for two reasons: (i)

ICP requires a starting point close to the optimal solution, meaning that ICP would not rescue initial FPFH feature-based misalignments; (ii) the meaning and assignment of color in a pharmacophoric context do not correspond to that utilized in image processing (RGB primary colors). Using the optimal set of parameters on the BO1 set, but refining the rough RANSAC alignment with the FPFH-colored-icp method confirmed our initial hypothesis, as the corresponding AUC (ROC AUC = 0.83; CI = [0.81;0.86]) was inferior to that reported above. We have therefore implemented a new descriptor to improve the correspondences estimation during the feature-based alignment.

**Improvement of the method with histograms encoding shape and pharmacophoric properties**

In light of the interesting results we previously obtained with the FPFH-icp routine and regarding the misalignment issues that arose, we have modified the FPFH descriptor implemented by default (Computational methods). Similarly to the way that shape information is binned to form a normalized 33-bin histogram, we encoded the distribution of eight pharmacophoric features (**Table 2**; Computational methods) in the neighborhood of a point into an eight-bin histogram, each bin corresponding to one of the eight pharmacophoric features. The final 41-bin histogram, termed c-FPFH (see Computational methods) was next utilized to improve RANSAC preliminary alignments of BO1 cavity pairs. Obtained results were compared to that obtained using the standard FPFH descriptor and to the alignments obtained our previously-reported Shaper[7] tool that uses a smooth Gaussian function to optimize the shape overlap of cavity points. Using the *ph4-ext* scoring function to score alignment of BO1 cavity pairs, the novel c-FPFH appears clearly superior to the standard one (c-PFPH, ROC AUC= 0.93, CI = [0.91;0.94]; FPFH, ROC AUC = 0.87) in discriminating similar from dissimilar pairs (**Figure 3**). The performance of the novel descriptor was almost similar to that obtained with the state-of-the art Shaper alignment tool (ROC AUC = 0.92, CI = [0.90; 0.93]) on the same data set. The Shaper method[7] was used here as a baseline alignment method for two reasons: (i) it has been favorably evaluated by independent groups[31, 54] on different benchmarking datasets featuring various applicability domains and comparison scenarios[31,], (ii) it is the only tool that can unambiguously be compared to ProCare because they use an identical input (two point clouds) for generating and scoring cavity alignments. Observed differences are therefore directly explained by different alignment qualities, the scoring function used by both methods remaining comparable.
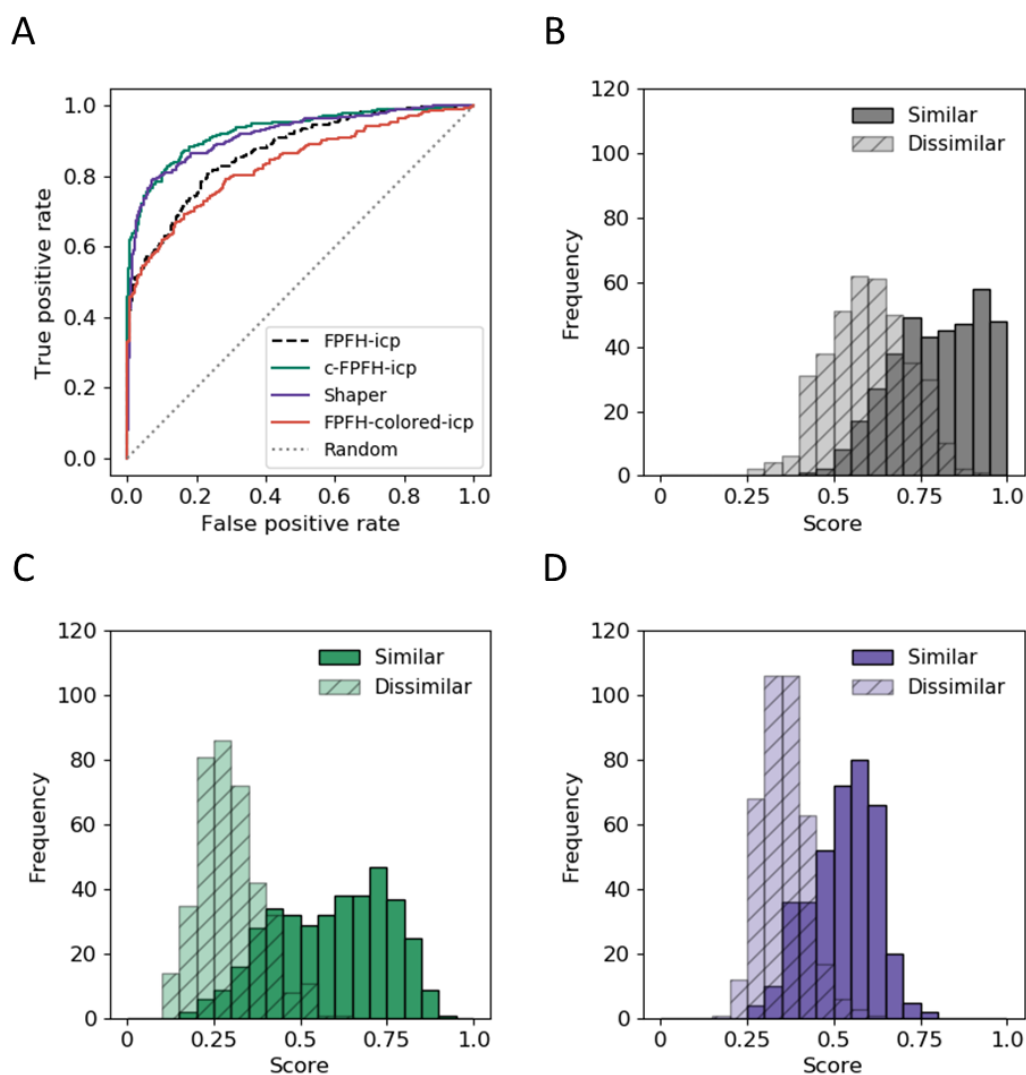
**Figure 3.** Evaluation of ProCare scoring in comparing cavities from the BO1 set. **A**) Receiver operating characteristics (ROC) plot in ranking BO1 cavity pairs with the *ph4-ext* scoring function, using ProCare (standard FPFH descriptor, new c-FPFH descriptor) and Shaper; **B**) Distribution of *ph4-ext* scores after ProCare overlay with FPFH-icp refinement; **C**) Distribution of ph4-ext scores after ProCare overlay with c-FPFH-icp refinement; **D)** Distribution of scores after Shaper overlay.

The improvement of the discrimination with c-FPFH descriptors is due to the correction of alignment errors previously reported, which are consequently reflected on scores. Differences in the ranking between methods is partially explained by misalignment of some similar pairs, and by the different fuzziness level of the utilized scoring functions. In quite a few cases, alignments of similar cavities were well approximated when evaluating the consequent alignment of the corresponding proteins, while the scores were inferior to the median score obtained for similar pairs. For those misaligned pairs, we did not find any correlation between alignment scores and chemical similarity of the cavity-bound ligands

(Tversky on Morgan fingerprint and MCS uniformly ranged from 0 to 1). Another reason for misalignments is the difference in shape (globular vs. planar) observed between the two cavities, rendering neighborhood similarities of randomly sampled points difficult to catch. Of course, we cannot exclude the possibility to have wrongly annotated BO1 pairs, particularly those predicted dissimilar. However, observing a similarity between binding sites of functionally unrelated proteins is a very rare event[38] so that, even if present in the data set, such cases are negligible.

**Statistical evaluation of ProCare score distributions**

The ability of the method combining c-FPFH descriptors for aligning and *ph4-ext* for scoring, was first assessed by its ability to discriminate similar and dissimilar cavities of the BO1 set, using incremental variations of the *ph4-ext* score (from here on ProCare score). The optimal discriminative power (recall = precision = F-measure = 0.85) is obtained at a threshold value of 0.39 for the investigated data set (**Figure 4A**). To check whether this threshold value is data set-dependent, we next generated a background distribution of 2.5 million alignments (510 non-redundant BO1 cavities vs. 4,223 sc-PDB cavities). 100 statistically representative samples of 100,000 values each, could be fitted to a generalized extreme value (GEV) distribution (**Figure 4B**) according to the Kolmogorov-Smirnov test (D = 0.046, P-value = 0.0292, a =0.02) with a probability density function of the type:

$$f(x) = \exp(-(1 + kz)^{-1/k}) \, (1 + kz)^{-1-1/k} \quad k \neq 0 \tag{1}$$

$$f(x) = \exp(-z - \exp(-z)) \quad\quad\quad k = 0$$

with $k = -0.15024$, $s = 0.08338$, $m = 0.24475$, $z = \frac{x-\mu}{\sigma}$

The significance level *p* of the detected similarity represents the probability of obtaining the same or higher similarity score $Z > z$ by chance is:

$$p(Z > z) = 1 - \exp\left(-(1 + kz)^{-\frac{1}{k}}\right) \quad\quad k \neq 0 \tag{2}$$

$$p(Z > z) = 1 - \exp\left(-\exp(-z)\right) \quad\quad k=0$$

From the background distribution, a statistically significant threshold for the ProCare score was set at a value of 0.47, which corresponds to a *p*-value of 0.05. At this threshold, the classification of the previous BO1 set yields to a lower recall (0.72) but a much better precision (0.95). From here on, ProCare will be used with the above-reported best set of parameters, combining c-FPFH descriptors for aligning and *ph4-ext* for scoring pocket alignments.
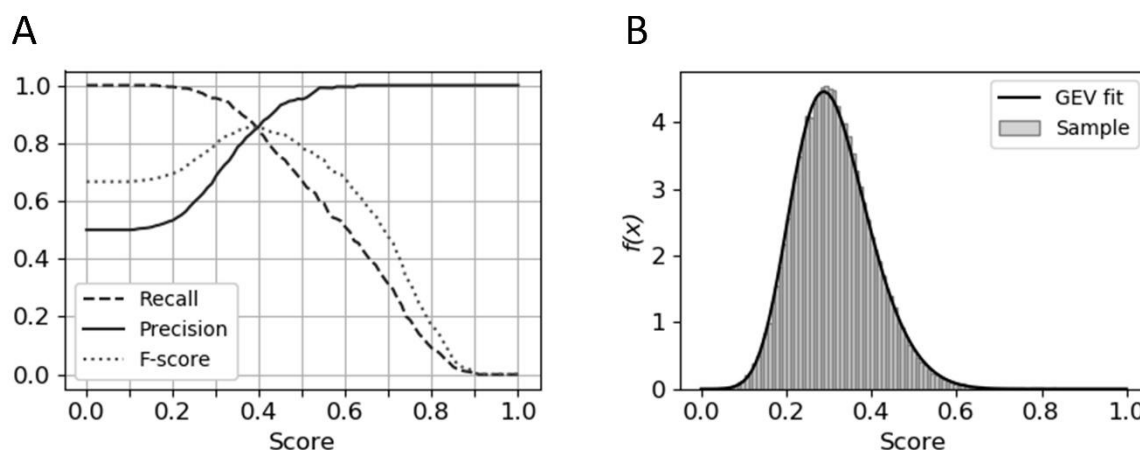
**Figure 4.** Statistical evaluation and sensitivity of ProCare to variations in atomic coordinates. **A)** Variation of statistical parameters (recall, precision, F-measure) of a binary classification model (similar/dissimilar) of BO1 cavity pairs for increasing ProCare similarity score thresholds; **B)** Fitting randomly sampled ProCare scores to a generalized extreme value (GEV) distribution. Repeated random samples (n = 100) showed to be representative of the whole population of scores (Scipy combined p-value for the 100 Kolmogorov-Smirnov p-values with Fisher's method: 0.90). GEV parameters were estimated with EasyFit.[55]

**Benchmarking ProCare versus state-of-the art methods in a medicinal chemistry context**

A fair comparison of a novel algorithm to state-of-the art competing methods is a difficult exercise because of the many sources of possible biases that can directly influence pocket similarity assessments:[31] data set assembly, pocket definition, scoring metrics, purpose (e.g. off-target prediction, polypharmacology, drug repurposing, target's function assessment). We herewith made the choice of a classical medicinal chemistry scenario: Do two pockets bind to the same ligands (chemotypes) or not? For that purpose, we revisited the recently published Vertex dataset[56] comprising 6,598 positive and 379 negative protein pairs defined from 6,029 protein structures. Interestingly, pairs were chosen depending on the availability (or not) of common high-affinity ligands (potency $\leq$ 100 nM). However, the published data set was strongly imbalanced (positive pairs >> negative pairs) and required some filtering (see Computational methods) to reach an equivalent numbers of 338 positive and 338 negative pairs (**Table S5**). Six publicly available methods (FuzCav,[57] Kripo,[36] PocketMatch,[58] ProBiS,[59] Shaper,[7] SiteAlign[60]; see Computational methods for more details), considered as state-of-the art cavity comparison tools by independent groups,[31, 54] were compared to the herein presented method for their ability to discriminate positive from negative pairs by the simple estimation of their ligand-binding pocket similarity (**Figure. 5**).

As a general trend, methods mapping physicochemical and/or pharmacophoric properties onto binding site atoms (FuzCav, PocketMatch, SiteAlign, KRIPO) outperformed the two methods (ProCare, Shaper)

relying on descriptors mapped onto pseudoligand atomic coordinates. This observation is easily explained by the design of the Vertex dataset that assigns positive pairs to very similar proteins of the same target family (e.g. Ser/Thr protein kinase, protease) sharing high sequence and structure homologies. However, these tools exhibit at least one drawback that does not exists with ProCare. First, alignment-independent methods (FuzCav, PocketMatch) are very fast and accurate but produce results that are hard to interpret since no protein overlay is generated. From a medicinal chemistry perspective, the absence of protein alignment prevents transferring a ligand from a reference pocket to another one and thereby hinders a structure-based hit to lead optimization. Second, the SiteAlign technology, although very precise, is very slow (ca 30 sec./comparison) and presents a limited applicability domain to short lists of proteins, unless executed in a distributed parallel computing environment. ProBiS allows a precise classification of positive and negative pairs but at the cost of a low completeness (only 64% of pairs could be treated, **Figure 5**).
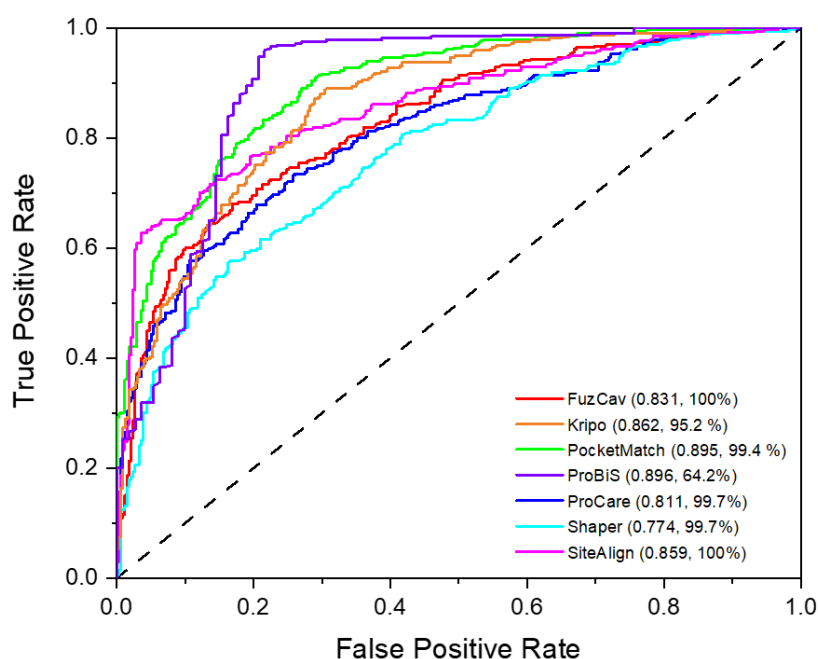


**Figure 5.** Receiver operating characteristics (ROC) plot for ranking 676 protein pairs (Vertex set: 338 positive, 338 negative) by decreasing pocket similarity, according to six different methods. Area under the ROC curve and completeness (% of successfully processed pairs) are indicated in brackets for each method.

Last, the KRIPO method that relies on known-protein ligand interactions to generate binding site descriptors failed in producing results for 5% of test cases and cannot be used for apo-proteins. ProCare therefore constitutes a widely applicable, robust approach to detect binding site similarity, as it is the only method cumulating high speed (a few sec/comparison), good precision (ROCAUC = 0.81),

interpretability (aligned proteins, list of distances between matched residues) and large applicability domain (ligand-bound and ligand-free protein structures).

**Detecting similarity between fragment subpockets and whole protein cavities**

As demonstrated in the previous section, point cloud registration can be successfully applied to align and compare entire protein cavities. Is it still applicable to smaller objects (fragment-binding sites), a notoriously difficult problem in cavity comparisons?[5] To answer this question, we systematically aligned cavity pairs from the Frag-Lig set[61] (**Table S6**; Computational methods) in which the same protein is bound to either a drug-like ligand or a substructural fragment of the later ligand (see Computational Methods). A correct subpocket to full cavity alignment can therefore be easily deduced after applying the ProCare transformation matrix to the corresponding protein-fragment complex and computing two properties: (i) the rmsd of the fragment-bound protein to the full ligand-bound target, (ii) the similarity of interactions observed between the full cavity and either the merged fragment or the reference full drug-like ligand.

Examination of pocket sizes, expressed as the number of points in the corresponding clouds, confirmed that the fragment-bound subpockets are much smaller than the entire cavities to which the corresponding full ligands bind to (**Figure S3**). In 91% of the cases, a structural alignment of both protein structures, performed by the combinatorial extension (CE) method,[62] yields to a rmsd on C-alpha atoms below 2 Å, illustrating that no major conformational changes occurs at the protein level upon ligand binding, when compared to the original fragment-bound protein structure (**Figure 6A**). In this context, ProCare clearly outperforms Shaper in proposing reliable alignments (rmsd of protein backbone atoms $\leq 2$ Å) in 42% of cases vs. 34% for the Gaussian-based Shaper method (**Figure 6A**). For those structurally well-aligned pockets, the ProCare score was higher than the previously defined threshold (score 0.47, $p$-value = 0.05) in 98% of the cases, suggesting that scores obtained by aligning full cavities can be translated to the comparison of pockets of very different sizes.

We next looked whether the better alignments proposed by ProCare, corresponds to a better positioning of the fragments after rotation/translation to the full cavity. Since fragments were not always real substructures of the full drug-like ligand counterpart (but sometimes just bioisosteric substructural parts), we could not compute rms deviations on fragment atomic coordinates. We therefore estimated the similarity of interactions between the fragment subpocket and either the ProCare-aligned fragment or the native drug-like ligand, using a Tanimoto coefficient calculated on molecular interaction fingerprints (IFP).[63]
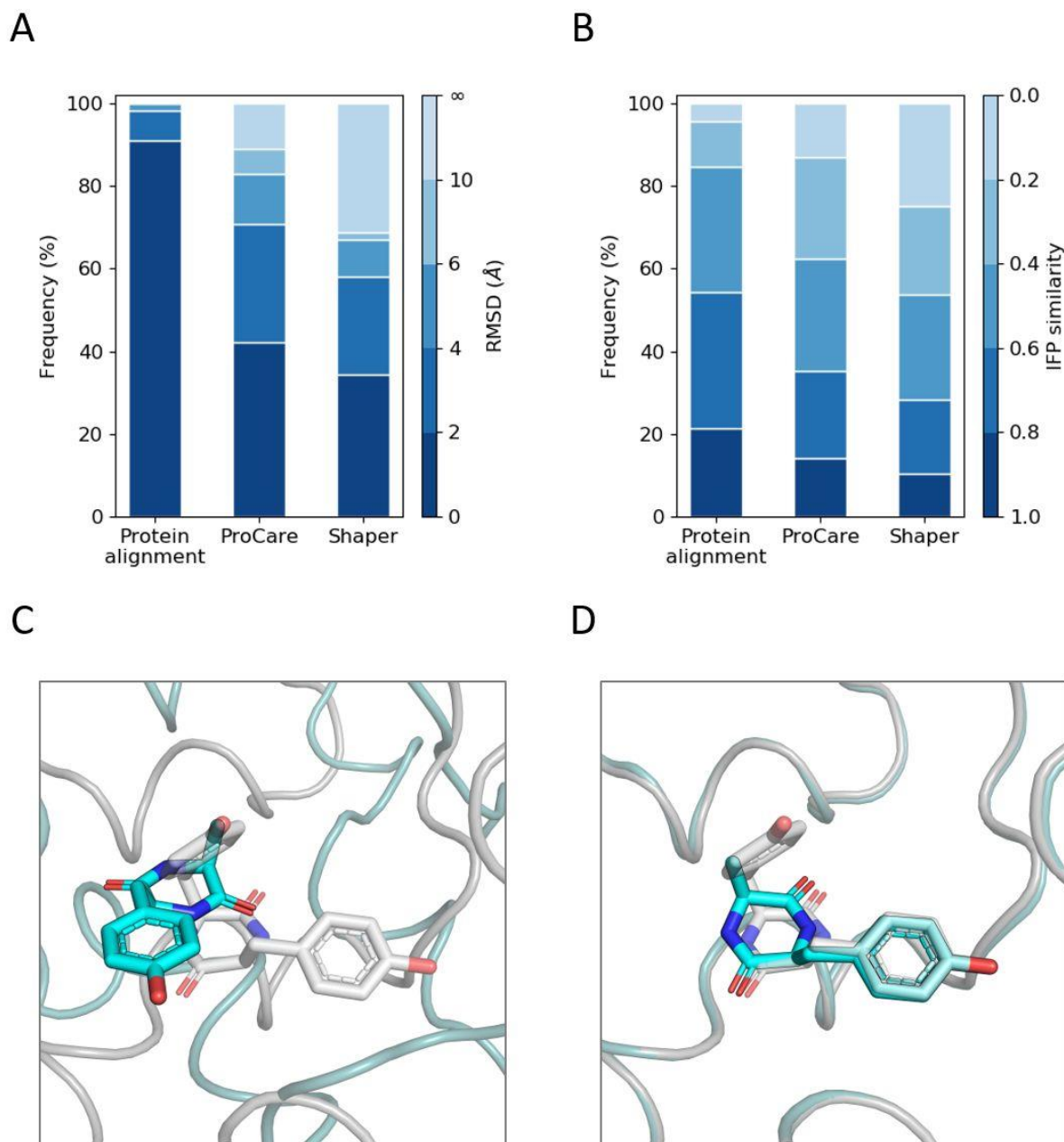
**Figure 6.** Evaluation of ProCare alignment of fragment supockets to full cavities. **A)** Proportion of pairs of proteins poses yielding rmsd on main chain atoms falling into the following intervals (Å) [0;2[, [2;4[, [4;6[, [6;10[, [10;∞[ after applying the transformation matrix derived from ProCare and Shaper alignments. The values were compared to the original structural alignments of the proteins obtained by the CE algorithm;[62] **B)** Proportion of pairs of fragment poses yielding IFP similarity with their paired ligands which falls into the following intervals [0;0.2], ]0.2;0.4], ]0.4;0.6], ]0.6;0.8], ]0.8;1.0]; **C)** Example of Shaper misalignment of cavities from cytochrome P121 bound to fragment 1G9 (PDB ID 4IQ7) and ligand YTT (PDB ID 3G5H; rmsd of proteins backbone heavy atoms: 22 Å; rmsd of ligands matching substructure: 5.4 Å); **D)** ProCare correct alignment of the same cavity pair (rmsd of proteins backbone heavy atoms: 0.45 Å; rmsd of ligands matching substructure: 0.59 Å).

Considering a conserved binding mode for IFP similarities higher than 0.6,[63] the CE structural alignment indicates that the fragment binding mode is conserved in the full ligand in 53% of cases (**Figure 6B**). Provided with this baseline, ProCare succeeded in correctly positioning the fragment in the full pocket in 35% of cases whereas Shaper was only successful in 28% of cases (**Figure 6B**), thereby confirming that the better cavity alignments provided by ProCare also translates into better poses of the corresponding fragment. In many examples, Shaper misalignments were indeed rescued by the herein described point cloud registration (**Figures 6C, D**).

**Virtual screening of fragment subpockets to assist fragment-based drug design: a first proof-of-concept**

We next extended the concept of fragment positioning inferred from binding sites alignments, to pairs of unrelated proteins. In this fragment-based drug design exercise, we took high-resolution X-ray structures of protein-ligand complexes recently disclosed for the first time in the Protein Data Bank, and checked whether screening a collection of fragment subpockets for similarity to the novel query cavities (**Table 3**), could help reconstitute, even partly, the masked query-bound ligands.

**Table 3.** Binding site comparison of three protein-ligand complexes recently released in the PDB.

| Target | PDB ID | Ligand[a] | Resolution, Å | Release date | Cavity size[b] |
|---|---|---|---|---|---|
| M5 muscarinic receptor | 6OL9 | 0HK | 2.5 | 2019-12-11 | 99 |
| TNF-alpha trimer | 6OOY | A7M | 2.5 | 2019-12-25 | 208 |
| Cysteinyl leukotriene receptor 2 | 6RZ8 | KNZ | 2.7 | 2019-12-11 | 241 |

[a] Ligand chemical component HET code

[b] number of cavity points. The volume of cavity (in $Å^3$) is the number of points x 3.375 (third power of the grid resolution in Å)

A collection of 33,953 fragment subpockets was obtained by fragmenting all sc-PDB-bound ligands (sc-PDB fragment set, Computational methods) using a previously reported protocol,[64] while keeping protein-bound 3D coordinates. The fragment subpocket collection was then screened for ProCare similarity to the three novel cavities whose structure had recently been disclosed and therefore not present in the sc-PDB archive. After point cloud registration, the corresponding fragments were merged into the coordinate frame of the query cavity using the optimal transformation matrix, and filtered according to two criteria: (i) compliance to the fragment rule-of-three[65] (hence, our fragmentation protocol may find no possible fragmentation of the sc-PDB ligand), (ii) ProCare score > 0.47. Remaining fragments hits were then ranked by a composite score (FragScore, eq. 3) taking into account

both pocket similarity and interaction fingerprint similarity when comparing selected fragments with the masked ligand co-crystallized with the target query.

$$FragScore = Procare_{score} + IFP_{sim} + \frac{1}{2} IFP\_polar_{sim}$$

(3)

where $IFP_{sim}$ is the similarity of full interaction fingerprints

and $IFP\_polar_{sim}$ is similarity of polar interaction fingerprints

The first query cavity is small-sized (335 Å$^3$) and was retrieved from the recently published muscarinic M5 receptor structure bound the tiotropium inverse agonist.[66] It is intended to be an easy challenge since the same ligand bound to three related muscarinic receptor subtypes (M1, M3 and M4) in five sc-PDB entries. Therefore, this first query was meant as a quality control of the ProCare alignment protocol and subsequent scoring function. Hence, three tiotropium-based fragments are ranked among the top 33$^{th}$ fragments (**Table S7**) and nicely posed with respect to the true M5-bound tiotopium pose (**Figure 7A**, **Table 4**). Interestingly, highly ranked fragments derived from ligands bound to unrelated proteins (e.g. Hemolymph juvenile hormone binding protein, PDB ID: 3AOS, Ligand HET: JH2; Histone deacetylase-like amidohydrolase, PDB ID: 1ZZ1, Ligand HET: SHH; **Figure 7B, Table 4**) nicely overlaps M5-bound tiotropium and suggest suitable starting points for fragment growing and/or linking.



A

FragID: 5CXV_OHK_1_1
Rank = 1, FragScore = 1.61

B

FragID: 3A0S_JH2_1_1
Rank = 10, FragScore =1.10

FragID: 1ZZ1_SHH_1_1
Rank = 11, FragScore = 1.10

C

FragID: 4KZ0_1UJ_1_1
Rank = 1, FragScore = 1.48

D

FragID: 3R04_UNQ_1_1
Rank = 7, FragScore = 1.28

E

FragID: 3N6U_TSU_1_1
Rank = 45, FragScore = 1.16

F

FragID: 1Q43_CMP_1_1
Rank = 22, FragScore = 1.15

FragID: 3QCH_NX2_1_2
Rank = 7, FragScore = 1.25

FragID: 1RUO_CMP_1_1
Rank = 1, FragScore = 1.43

G

FragID: 3F9M_MRK_1_1
Rank = 2, FragScore = 1.38

FragID: 3DT1_P40_1_3
Rank = 10, FragScore = 1.21

**Figure 7.** ProCare positioning of sc-PDB fragments in novel cavities. Atoms are colored using a cpk color-coding (nitrogen: blue; oxygen: red; sulfur; yellow; carbon of fragment: cyan/rosy salmon, green; carbon of true ligand, white). **A-B)** Placing a fragment derived from a muscarinic M1 receptor-bound ligand (PDB ID: 5CXV; HET: 0HK), and a hemolymph juvenile hormone binding protein-bound ligand (PDB ID: 3AOS; HET:JH2) in the muscarinic M5 receptor cavity (PDB ID 6OL9); **C-E)** Placing a fragment derived from a phosphatidylinositol 4,5-bisphosphate 3-kinase-bound ligand (PDB ID: 4KZ0; HET: 1UJ), a protein kinase Pim1-bound ligand (PDB ID: 3R04; HET: UNQ), and a LysR type regulator-bound ligand (PDB ID: 3N6U; HET: NSU) in the TNF-alpha trimer cavity (PDB ID 6OOY); **F-G)** Placing fragments derived from a catabolite gene activator protein-bound ligand (PDB ID: 1RU0; HET: CMP), a potassium/sodium hyperpolarization-activated cyclic nucleotide-gated channel 2-bound ligand (PDB ID: 1Q43; HET: CMP), a receptor-type tyrosine-protein phosphatase gamma-bound ligand (PDB ID: 3QCH; HET: NX2), a glucokinase-bound ligand (PDB ID: 3F9M; HET: MRK) and a MAP kinase 14-bound ligand (PDB ID: 3DT1; HET: P40) in the cysteinyl leukotriene receptor 2 cavity (PDB ID 6RZ8).

Of course, visual inspection of the merged fragments into the query cavity space remains necessary to optimize fragment hits (e.g. JH2 fragment lacks the necessary ammonium group for $\pi$-cation interaction to Tyr481) for the intended cavity. The second query cavity (681 $\text{Å}^3$) is present at the interface of an asymmetrical tumor necrosis factor-alpha (TNF-alpha) trimer. This unique inhibitor-bound TNF conformation has very recently been reported[67] and has no comparable structure in the sc-PDB archive. Nevertheless, several sc-PDB fragments (e.g. 4KZ0_1UJ, 3R04_UNQ; see list of top 100 scorers in **Table S8**) selected from unrelated proteins, appear among the top ProCare scorers, and are true bioisosteres of the benzimidazole moiety of the TNF-alpha inhibitor (**Figure 7C-D, Table 4**). The ProCare poses of the selected fragments nicely overlaps that of the true ligand, and recapitulates aromatic interactions exhibited by the bicyclic benzimidazole ring and a hydrogen bond to Tyr151 side chain of the TNF-alpha cavity. Likewise, the disubstituted aromatic substituent of the true TNF-alpha inhibitor is also mimicked by one of the top scoring aromatic fragment (3N6U_NSU, **Figure 7E, Table 4**).

**Table 4.** Selection of top-scoring fragments for three novel cavities.

| Target[a] | Fragment | | | | | | |
|---|---|---|---|---|---|---|---|
| | Name[b] | Rank | FragScore[c] | Procare | p-value | IFP$_{sim}$ | IFP_polar$_{sim}$ |
| 6OL9 | 5CVX_OHK_1_1 | 1 | 1.61 | 0.82 | 2.04e-12 | 0.53 | 0.50 |
| | 3AOS_JH2_1_1 | 10 | 1.10 | 0.57 | 0.006 | 0.53 | 0.00 |
| | 1ZZ1_SSH_1_1 | 11 | 1.10 | 0.56 | 0.008 | 0.54 | 0.00 |
| 6OOY | 4KZ0_1UJ_1_1 | 1 | 1.48 | 0.57 | 0.006 | 0.67 | 0.50 |
| | 3R04_UNQ_1_1 | 7 | 1.28 | 0.65 | 1.63e-04 | 0.46 | 0.33 |
| | 3N6U_TSU_1_1 | 45 | 1.16 | 0.64 | 7.89e-04 | 0.36 | 0.33 |
| 6RZ8 | 1RUO-CMP_1_1 | 1 | 1.43 | 0.55 | 0.010 | 0.43 | 0.50 |
| | 3F9M_MRK_1_1 | 2 | 1.38 | 0.57 | 0.006 | 0.64 | 0.00 |
| | 3QCH_NX2_1_2 | 7 | 1.25 | 0.52 | 0.020 | 0.73 | 0.00 |
| | 3DT1_P40_1_3 | 10 | 1.21 | 0.57 | 0.006 | 0.64 | 0.00 |
| | 1Q43_CMP_1_1 | 22 | 1.15 | 0.47 | 0.054 | 0.43 | 0.50 |

[a] Targets are named according to their PDB identifier (6OL9, M5 muscarinic receptor; 6OOY, TNF-alpha trimer, 6RZ8, Cysteinyl leukotriene receptor 2)

[b] Fragment name (PDB_HET_C_M) is inferred from the cognate target PDB identifier (PDB), the corresponding ligand chemical component (HET), the target cavity identifier (C), and the fragment number (N).

[c] The Fragscore is computed according to eq. 3

The last query used for this preliminary proof-of-concept comes from the structure of an antagonist-bound cysteinyl leukotriene type 2 receptor (CysLTR2, PDB ID 6RZ8).[68] Again, this structure has no similar homologue in the sc-PDB archive, such that the ProCare search for potential subpocket matching has no obvious bias. The CysLTR2 pocket is wider (813 $\text{Å}^3$) than the two previous ones, and is fully occupied by a high molecular weight ligand (ONO-2080365, HET: KNZ) filling three separate subsites, thereby challenging ProCare for finding local similarity to each of the three subpockets and finding appropriate fragments. The benzoxazine dicarboxylic acid-binding subpocket in CysLTR2 is found similar to that of two adenosine-3',5'-cyclic-monophosphate (cAMP) pockets from unrelated proteins (catabolite gene activator protein, PDB ID: 1RUO; Potassium/sodium hyperpolarization-activated cyclic nucleotide-gated channel 2, PDB ID: 1Q43) with the cyclic phosphate group mimicking each of the two carboxylic acids of the CysLTR2 antagonist (**Figure 7F, Table 4**) and interacting with a basic residue (Arg82 for 1RUO, Arg591 for 1Q43) that drives the subpocket similarity to the CysLTR2 cavity (**Figure S4**). Local similarity to the central phenoxy-binding subsite is also found in a subpocket from a receptor tyrosine phosphatase (PDB ID: 3QCH, **Figure 7F, Table 4**) with a nice overlap of the corresponding dichlorophenyl fragment to the fluorophenyl substructure of the CysLTR2 ligand. Another fragment mimicking both the benzoxazine and the central fluophenyl CysLTR2 antagonist is selected by ProCare from remote pocket similarity to that of a glucokinase pocket (PDB ID 3F9M, **Figure 7G, Table 4**). Last the hydrophobic CysLTR2 subsite accommodating the terminal difluorophenyl ring of the bound inhibitor is found similar to that of a MAP kinase 14 subpocket (PDB ID: 3DT1) with a nice overlap of the cognate phenyl fragment to the terminal aromatic ring of the CysLTR2 ligand (**Figure 7G, Table 4**). Altogether, ProCare managed to find subpocket similarity between each of the three CysLTR2 subsites with totally unrelated subpockets and proposes reliable fragments for a structure-based fragment linking strategy (see the list of 100 top fragments in **Table S9**). Importantly, subpocket similarity and fragment posing were found for very different reasons ranging from salt bridge mimicry to the conservation of hydrogen bonds and hydrophobic/aromatic interactions.

We acknowledge that the empirical FragScore, used in the present exercise, can only be used in case the query cavity is already filled with a ligand. It enables to retrieve either apolar/aromatic fragments exhibiting a high interaction fingerprint similarity score (IFP), or polar/charged fragments with a high polar interaction fingerprint similarity value (IFP_polar). Cavity pairwise similarity, expressed by the ProCare score remains however the main driver for fragment selection, and can be used to query cavities in the apo-state. The accompanying *p*-value gives a statistical support to the predictions and can be used as a surrogate to the ProCare similarity value.

## 2.3.4. Conclusions

We herewith present a novel computational method, inspired from computer vision, to align and compare protein cavities. Cavities are represented as 3D point clouds annotated by pharmacophoric properties mimicking that of an ideal ligand, and aligned by the point cloud registration. Importantly, ProCare takes advantage of a novel point feature histogram to encode cavity microenvironments, thereby favoring the overlay of supockets sharing similar geometrical and physicochemical properties. The new method is able to align either entire pockets, subpockets, and compares subsites to full cavities. It exhibits a comparable performance to state-of-the-art methods when tested across a variety of benchmarking data sets. A key feature of ProCare is its unique ability to detect local similarities and thereby compare cavities of quite different sizes (e.g. fragment-bound subpockets vs. full ligand-bound cavities). We herewith provide the proof-of-concept of its application in a fragment-based drug design scenario in which cavities from recently described X-ray structures have been compared to a collection of fragment-bound subpockets. Local similarities undetectable with standard cavity comparison tools are found by ProCare, and enable after cavity overlay, to directly locate the corresponding fragments in the query cavity. Interestingly, proposed fragments are derived from remote targets that are totally different from the query, and proved to be identical or bioisosteric to susbtructures of the unmasked query cavity-bound ligand. Of course, designing a full ligand still requires to either grow and/or link ProCare-aligned fragments with any of existing computational fragment linking tool.[69-72] Nevertheless, the novel method enables to elaborate a fragment-based drug design strategy from the simple knowledge of a cavity 3D structure, by simple detection of local similarities to a large collection of fragment-bound subpockets.

In its current implementation, ProCare can still be optimized with respect to speed and completeness. A pairwise similarity search can be achieved in a couple of seconds, but the cpu cost could be significantly reduced by optimizing the nearest neighbor search and excluding irrelevant points in the preliminary RANSAC alignment procedure. Moreover, usage of the RANSAC algorithm does not guarantee to find the best possible solution to the registration. Deterministic algorithms able to find the absolute minimum have recently been proposed[73] and should be tested further on. Last, the method could also be applied to align ligands to cavity points, and propose a computer vision approach to the protein-ligand docking problem. ProCare is freely available upon request to authors.

## 2.3.5. Computational methods

**Data Sets**

***EASY1 set.*** This data set consists of five pairs of known similar cavities and five pairs of known dissimilar cavities (**Table S1**). Protein-ligand X-ray structures were extracted from the sc-PDB database (http://bioinfo-pharma.u-strasbg.fr/scPDB)[52]. Cavities were computed from ligand-free sc-PDB protein input (mol2 file format) with using default parameters of the VolSite[7] algorithm within the IChem v. 5.2.9 toolkit.[74] Cavity points, located on a 1.5-Å three-dimensional (3D) lattice and annotated by pharmacophoric properties,[7] were placed within 6 Å of heavy atoms of the corresponding hidden ligand, and visually checked with Pymol v.2.1.0.[75]

***BO1 diverse set***. Starting from all 16,034 sc-PDB protein-ligand complexes, unique proteins were retrieved and clustered according to UniProt[76] keywords. Proteins without keywords (cluster "No Keywords") and singletons were discarded. For each cluster, the proteins sequences in fasta format were retrieved from the UniprotKB API and gathered to form a multi-fasta alignment file of the cluster. In case several isoforms were available for one protein, only the first one (default) has been considered. Then, multiple sequence alignments were performed with Clustal Omega[77] via the EMBL-EBI web services REST API[78]using default parameters, and outputted in ClustalW format. The Percent Identity Matrix (PIM) files were processed to retrieve pairs of proteins having different Uniprot AC and a sequence identity between 50 and 100%. For enzymes (Function-Keywords containing one of the 6 enzyme classes), the Enzyme Classification (E.C.) number was fetched from UniprotKB and additional filtering was performed to discard pairs having different E.C. numbers and pairs in which at least one partner is not annotated with E.C. number (e.g. TrEMBL entries). At this stage, PDB atomic coordinates of ligand-bound protein chains were extracted and structurally aligned with Sybyl-X 2.1.1[79] ("biopolymer align_structure" method, default parameters). Pairs of proteins for which the root-mean square deviation (rmsd) of main chain coordinates is higher than 5 Å were discarded. For 30 pairs, a manual structural alignment was performed with Maestro v.11.9.011[75] to rescue SYBYL misalignments.

For each of the remaining 643 pairs, corresponding cavities were computed from the position of their bound ligands, as described above for the EASY1 set. The transformation matrix used to align the proteins was applied to their corresponding cavities using the realign module of the IChem toolkit. Pairs of cavity points were next analyzed for co-localization, by measuring all possible pairwise distances. A pair was kept if three conditions were verified: (i) at least 45% of all pairwise distances are lower than 10 Å, (ii) any cavity point in one pair member has more than 50 unique neighbors (d < 1.5 Å) in the cognate pair member; (iii) bound ligands according to Morgan fingerprints (radius = 2)[80] were not identical (Tanimoto coefficient Tc ≠ 1). Finally, a set of 383 pairs (**Table S2**) was annotated as "similar". An equally-sized set of dissimilar pairs (**Table S3**) was defined from the above described clustering of

UniprotKB keywords, as protein pairs sharing no single functional keyword and different ligands HET codes with a chemical similarity, expressed by a Tanimoto coefficient on Morgan fingerprints (radius = 2), below 0.4. Finally, an equivalent number of 383 dissimilar pairs was retrieved randomly from that list, with the constraint that the distribution of differences in cavity volumes between dissimilar pairs matches that of similar pairs.

***Vertex Set.*** The dataset was retrieved from the original publication[56] and comprises 6,598 positive and 379 negative protein pairs defined from 6,029 protein structures. Positive and negative labels were originally assigned as whether the pair share high affinity common ligands (potency < 100 nM) or not. The full dataset provides a total of 1,564,605 putative matches, considering multiple structures (e.g. 5 PDB entries for human CDK5) and all possible bound ligands for a single protein structure. Since the dataset is very imbalanced, a post-processing step was conducted to achieve an equivalent number of positive and negative labels. For each possible protein pair, the chemical 2D similarity of their ligands was computed from RDKit Morgan fingerprints (radius = 2) and the pair with the highest ligand similarity saved as representative sample (for positive pairs, $0.4 \leq$ ligand similarity $\leq 0.7$). For each remaining pair, the corresponding pockets were identified with the VolSite module of IChem, leading to a final set of 338 negative and 841 positive pairs out of which 338 were randomly retrieved to achieve an equivalent number of positive and negative samples (**Table S5**).

***Frag-Lig set.*** This data set is a subset of the previously reported PDBmob data set,[61] and consists of 578 pairs of cavities from the same protein (same Uniprot AC), bound to a drug-like ligand and a substructural fragment of the latter ligand. The data set provides already aligned protein-ligand/fragment complexes for each target set. For each unique protein of the PDBmob data set, all possible pairs of protein-fragment and protein-druglike ligand were formed. The Tversky similarity of the paired fragments and ligands were calculated using RDKit Morgan fingerprints (radius = 2) and maximum common substructures (RDKit FindMCS default parameters). A first selection conserved pairs with both similarity metrics superior to 0.6. The corresponding cavities were computed with IChem VolSite using default parameters. For fragment-bound structures, only the close vicinity (4 Å) of the fragment was considered for cavity detection (VolSite *CAVITY_4* output). For ligand-bound structures, the entire cavity (VolSite *CAVITY_ALL* output) was retrieved. This preliminary list was then filtered to remove drug-like-bound cavities of smaller volume that that of the fragment counterpart. Then, fragment/ligand occupancy in their cognate cavities was inspected to ascertain that any heavy atom has a cavity point within a 2 Å distance. Last cavity overlap (fragment-bound vs. ligand-bound) was computed by estimating the number of fragment-bound cavity points with a close neighbor ($\leq$ 2Å) in druglike-bound cavity points. Only pairs with 100% overlap were finally retained to yield 578 pairs (**Table S6**). For

each pair, atomic coordinates of the fragment-protein complex were randomly translated by 10 Å along the three axes x-y-z and rotated by 180° along the x-axis, in order to put reference and target complexes in different coordinate frames.

***sc-PDB fragment set.*** For each of the 16,034 entries of the sc-PDB data set,[52] the corresponding 3D structure of the ligand was fragmented using a previously-described protocol[64] in three steps. First, a ring perception algorithm is used to detect aromatic and aliphatic rings of the ligand. Second, acyclic atoms are then parsed to assign either a linker or substituent label, as whether to the corresponding bonds are connecting two rings or not. Linker atoms are left unchanged. In case of substituent atoms, single bonds involving the closest apolar carbon (in terms of bond distance) to any ring are later cleaved at the condition that the cleaved bond is at least three bonds away from the cyclic root atom. Third, fragments are kept at the condition that they make at least 4 interactions (including ≥1 polar or aromatic) with the target. The fragment set contains 33,953 fragments out of which 7,294 are unique. For each of the 33,953 protein-bound fragments, the 4 Å-surrounding cavity was computed in IChem VolSite as described above.

**Point Cloud registration**

The herein described method relies on Open3D v.0.5.0,[81] a library for point cloud processing. The library is available in C++ programming language but provides a python interface with pybind11, and allows parallel computing via the OpenMP environment. For the sake of efficiency, Open3D was compiled and installed from source in conda environment following the provided guidelines. Protein cavity files computed with VolSite (mol2 format) were converted into PCD (Point Cloud Data) file format version 0.7. The Header was kept as default unless the "WIDTH" and "POINT" sections that were updated with the cavity size (number of cavity points). The "DATA ascii" section contained the x, y, z coordinates of the mol2 file and a fourth column assigning a color to each of the eight VolSite pharmacophoric properties.[7] Normal vectors and fast point feature histograms (FPFH)[82] were computed for the source cloud and the target cloud. A first rough alignment was performed based on FPFH descriptors with the Random Sample Consensus (RANSAC) method[45] in an iterative way (*registration_ransac_based_on_feature_matching* function). The rough alignment was subsequently refined with an Iterative Closest Point algorithm[46] (*registration_icp* function) starting from the transformation matrix of the rough alignment. Alternative to registration_icp is registration_colored_icp, which is a function considering the color of points to compute transformation matrices. We further implemented a new descriptor, the colored-FPFH (c-FPFH). c-FPFH consists of 41 bins: the 33 FPFH bins, with eight additional normalized bins accounting for the distribution of the eight colors (pharmacophoric properties) in the neighborhood of the point (**Figure 8**).
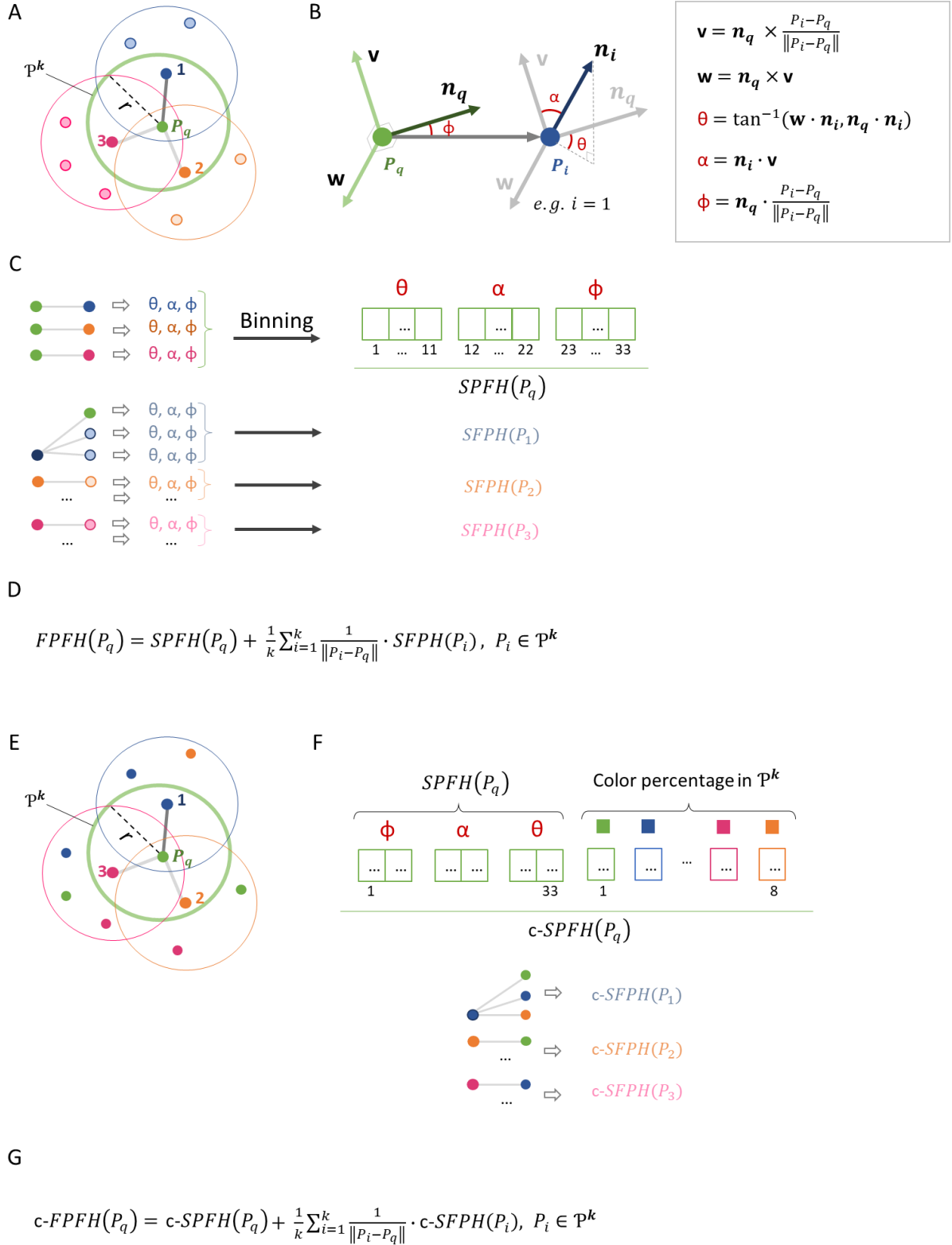
**Figure 8.** Fast point feature histogram (FPFH) and colored fast point feature histogram (c-FPFH) computation. **A)** Simplified schematic representation of a cloud of points. The neighborhood is perceived without considering the points colors. Considering a point $P_q$ (green) whose FPFH is to be

computed, its neighbor points $\mathcal{P}^k = \{1, 2, 3\}$ within a radius $r$ are determined (green circle). For each neighbor in $\mathcal{P}^k$, their respective neighbors are also determined within the radius $r$; **B)** Between a point and each of its neighbor, an ensemble of $\theta$, $\alpha$ and $\varphi$ angular values are computed to reflect the local environment of each point; **C)** Each of the $\theta$, $\alpha$ and $\varphi$ computed values for the point Pq and its normal $n_q$ are respectively binned into 11-bin histograms with regular intervals deduced from minimal and maximal distances. The resulting 33-bin histogram forms the simplified point feature histogram (SPFH) of the point Pq. Similarly, the SPFH is computed for each point in $\mathcal{P}^k$; **D)** The FPFH of the point $P_q$ is the sum of its SPFH and the distance-weighted average of its neighbors' SPFHs; **E)** Simplified schematic representation of a cloud of points with perception of point colors. Considering a point $P_q$ (green) whose c-FPFH is to be computed, its neighbor points $\mathcal{P}^k = \{1, 2, 3\}$ within a radius $r$ are determined (green circle). For each neighbor in $\mathcal{P}^k$, their respective neighbors are also determined within radius $r$; **F)** The 33-bin histogram SPFH is computed for the point $P_q$, in addition to eight bins coding for the eight pharmacophoric features respectively, encompassing the percentage of each pharmacophoric feature in $\mathcal{P}^k$. The final 41-bin histogram forms the c-SPFH of the point $P_q$. Similarly, the c-SPFH is computed for each point in $\mathcal{P}^k$; **G)** The c-FPFH of the point $P_q$ is the sum of its c-SPFH and the distance-weighted average of its neighbors' c-SPFHs.

**ProCare parameters**

A set of values were rationally defined for 15 Open3D parameters (**Table 1**). A combination of these values led to 157,464 different alignment conditions.

All possible combinations were tested on the EASY1 data set and their performance evaluated in three steps. First, parameter sets having rough and refined alignment fitness values higher than 0 were retrieved and their corresponding alignments were rescored with the above-described *ph4-strict* scoring scheme.

**Table 1.** Open3D parameters values for ProCare alignment (default values are underlined)

| Parameter | Tested values |
|---|---|
| RANSAC cycle number of validations, *rn* | 2; <u>4</u>; 5 |
| RANSAC maximum number of validations, *rv* | 50; <u>500</u> |
| RANSAC maximum number of iterations, *ri* | 50,000; 100,000; <u>4,000,000</u> |
| Rough alignment transformation estimation type, *gt* | <u>TransformationEstimationPointToPoint</u> |
| Rough alignment distance threshold in Å, *gd* | 0.75; <u>1.20</u>; 1.50 |
| Checkers similarity threshold, *cs* | 0.90; <u>0.96</u>; 1.00 |

| | |
|---|---|
| ICP alignment transformation estimation type, *it* | TransformationEstimationPointToPoint; <u>TransformationEstimationPointToPlane</u> |
| ICP alignment distance threshold in Å, *id* | 0.75; 1.50; 3.00; <u>6.00</u> |
| ICP maximum iterations, *ii* | <u>30</u>; 100; 500 |
| ICP relative fitness threshold, *if* | $10^{-7}$; <u>$10^{-6}$</u>; $10^{-5}$ |
| ICP relative RMSE threshold, *ir* | $10^{-7}$; <u>$10^{-6}$</u>; $10^{-5}$ |
| Nearest neighbor search radius for normals in Å, *nr* | 1.6; 3.1; <u>10</u> |
| Maximum number of neighbors for normal, *nm* | <u>30</u>; 471[a] |
| Nearest neighbor search radius for FPFH in Å, *r* | <u>2</u>; 3.1; 4.6 |
| Maximum neighbors for FPFH, *fm* | <u>100</u>; 135[a] |

[a] Theoretical maximal value for 1.5 Å-regularly spaced point sets.

Second, the area under the receiver operating characteristic (ROC) curve was assessed using either the Tanimoto or the Tversky metric to rank alignment similarity values. Corresponding parameter sets were conserved only if the ROC AUC was equal to 1. Finally, the target protein structures were aligned with UCSF Chimera v.1.12[83] using the cavity transformation matrix previously generated by ProCare for three EASY1 pairs (HIV protease: residues 1-99, 100-198; beta-2 adrenergic receptor: residues 1-202, 363-44; cyclin-dependent kinase 2: 2c6t-residues 1-35, 45-148; 1dm2 residues 1-35, 36-139, 140-272). Only parameter sets leading to a mean rmsd (backbone heavy atoms) below 2 Å were kept for further analysis on the BO1 data set.

**ProCare scoring**

The quality of the alignment was estimated by two scores (fitness, RMSE) in Open3D. The fitness score (eq. 4) measures the overlap of source and target clouds as the ratio of the number of inlier correspondences (points in the source cloud that are fitted to the target cloud, based on a nearest neighbor search on coordinates after transformation) to the total number of points in the source cloud.

$$fitness = \frac{Number\ of\ inlier\ correspondences}{Total\ number\ of\ points\ in\ source\ cloud} \tag{4}$$

RMSE (eq. 5) is the root-mean square error between corresponding pairs of points in source and target clouds.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Ps_i - Pt_i)^2}, \qquad Ps \in \mathbb{P}_{source}, \qquad Pt \in \mathbb{P}_{target}$$

$$(5)$$

We then implemented 3 additional scoring functions (*ph4-strict*, *ph4-rules*, *ph4-ext*) to evaluate the alignment of pharmacophoric properties. The *ph4-strict* scoring method, relies on the ball-tree algorithm implemented in scikit-learn,[84] and searches for the nearest neighbor point in the largest cavity for each point of the smallest cavity, within a maximum distance d ( d = 1.5 Å by default). Three similarity indices (Tanimoto, Tversky, Wei) are computed for each alignment (eq. 6-8).

$$Tanimoto = \frac{c}{a+b}$$

$$(6)$$

$$Tversky(\alpha, \beta) = \frac{c}{\alpha(a-c) + \beta(b-c)+c}, \ (\alpha=0.95, \ \mathit{6}=0.05)$$

$$(7)$$

$$Wei = \sum_{i}^{properties} \frac{1}{f_i}\cdot\frac{c_i}{a}, \qquad i \in \{CA, CZ, O, N, OD1, NZ, OG, DU\}$$

$$(8)$$

Where *c* is the number of fitted points of identical pharmacophoric properties,

*a* and *b* are number of points of the smallest and the largest cavity, respectively,

$c_i$ is the number of points of property *i* aligned,

$f_i$ is the average frequency of points with property *i* in all sc-PDB cavities.

The *ph4-rules* scoring function is defined as the *ph4-strict*, with *c* equal to the number of fitted points of similar pharmacophoric properties (**Table 2**). The *ph4-ext* scoring function is defined as the *ph4-strict*, with *c* as the number of points in the smallest cloud which has a point of the same property of any of its neighbors in the target cloud. As for the *ph4-strict* scoring method, the Tanimoto, Tversky and frequency-weighted metrics are calculated.

**Table 2.** Pharmacophoric matching rules used by the ph4-rules scoring function.

| Property | Definition | Compatible pharmacophoric properties |
|---|---|---|
| CA | Hydrophobic | CA, CZ |
| CZ | Aromatic | CZ, CA |
| N | H-bond donor | N, NZ, OG |
| NZ | Positive | NZ, N, OG |
| O | H-bond acceptor | O, OD1, OG |
| OD1 | Negative | OD1, O, OG |
| OG | H-bond acceptor & donor | OG, N, O, OD1, NZ |
| DU | Dummy atom | DU |

**Shaper comparisons**

Starting the from the same set of point clouds, Shaper[7] relies on the OpenEye ShapeTK toolkit[85] and a smooth Gaussian function to maximize the overlap of both cavity shapes and colors (pharmacophoric properties). The alignment between cavities A and B was scored as the higher of two Tversky metrics (eq. 9-10).

$$S_{A,B} = \frac{O_{A,B}}{0.95\ I_A + 0.05\ I_B + O_{A,B}} \tag{9}$$

$$S_{A,B} = \frac{O_{A,B}}{0.05\ I_A + 0.95\ I_B + O_{A,B}} \tag{10}$$

where $O_{A,B}$ is the overlap between colors of cavities A and B, and I non-overlapped colors of each entity A and B.

**FuzCav comparisons**

FuzCav is an alignment-independent ultra-fast pocket similarity tool[57] relying on generic 4833-integer vector registering counts of all possible pharmacophoric triplets from the C-α atomic coordinates of binding site-lining residues. The code was retrieved from authors' website[86] and used with default parameters on binding sites (mol2 file format) deduced from atomic coordinates of the bound ligand, selecting any amino acid for which one heavy atom is present in a 6.5-Å radius sphere centered on the geometric barycenter of ligand heavy atoms. Similarity between two pockets was estimated from the Hamming distance between the two compared fingerprints.

**KRIPO comparisons**

KRIPO discretizes protein- bound ligands into small fragments and further describe their binding subpockets by 3-point pharmacophore fuzzy fingerprints.[36] Similarity between two fingerprints is estimated by a modified Tanimoto coefficient taking into account the mean density of each bit string. The code (version 1.0.1, released date: 2018-03-28) was downloaded from https://github.com/3D-e-Chem/kripo. For purposes of comparing to other methods, default parameters were used to compute fingerprints without fragmentation using ligand expo sdf files.[87] Lastly, fingerprints similarities were computed with Kripodb using default parameters and setting the score cutoff to 0.

**PocketMatch comparisons**

PocketMatch[58] describes a binding pocket as a set of 90 lists of sorted distances between three sets of critical atoms (Cα, Cβ and centroid of the side chain) of any cavity-lining residue classified in five groups according to their physicochemical properties. Similarity between two binding sites is scored as the net average of the number of matching distances in the 90 lists as a fraction of the total number of distance elements in the bigger set. The program (version 2.1) was retrieved from authors' website[88] and used with default parameters from ligand-binding sites in regular PDB file format. Similarity between two pockets was estimated using the P_max_OP score.

**ProBiS comparisons**

ProBiS detects structurally similar sites on protein surfaces by local surface structure alignment using a fast maximum clique algorithm.[59] The program (version 2.4.7) was downloaded from the authors' web site.[89] Starting from protein-ligand PDB files, default settings were used at the exception of the distance used to define binding site atoms from ligand atomic coordinates which was raised from 3.0 (default value to 6.5). Similarity between two pockets was estimated using the alignment score.

**SiteAlign comparisons**

SiteAlign[60] is an alignment-dependent algorithm describing a pocket by eight topological and physicochemical attributes, projected from the Cα-atom of cavity-lining residues to an 80 triangle-discretized polyhedron placed at the center of the binding site, thus defining a cavity fingerprint of 640 integers. 3-D alignment is performed by moving the sphere within the target binding site while keeping the query sphere fixed. After each move, the distance of the newly described cavity descriptor is compared to that of the query, the best alignment being that minimizing the distance between both cavity fingerprints. The program (version 4.0) was retrieved from authors' website[86] and used with default parameters from ligand-binding sites in regular mol2 file format. Similarity between two pockets was estimated as 1 minus the d2 score.

**ProCare running times**

Cavity alignments were run on a 64-bit Intel Core i5-4590 @ 3.30 GHz processor with 4 threads, 16 Go RAM. Average running time of a pair-wise comparison is 2.17 s.

**Statistical analysis**

Data analysis was performed with in-house python scripts. The 90 % confidence intervals CI = [$i_{upper}$ ; $i_{lower}$] for area under the ROC curve were obtained with 5,000 bootstrap samples, where $i_{upper}$ and $i_{lower}$ were calculated with the NumPy[90] package to be the 95th and the 5th percentiles. Sampling fitting to the generalized extreme value (GEV) distribution and statistical tests were performed with EasyFit[55] and Scipy.[91]

## 2.3.6. Associated content

**Supporting Information**

The supporting information is available free of charge on the ACS Publications website at DOI: https://dx.doi.org/10.1021/acs.jmedchem.0c00422.

Properties of the BO1 data set of 766 protein-ligand cavity pairs; Example of misalignment for a pair of similar cavities from the BO1 set; Distribution of pocket size for fragments (light blue) and full cavities (dark blue); ProCare overlay of cavities from unrelated targets; EASY1 set of similar and dissimilar pairs; List of BO1 similar pairs; List of BO1 dissimilar pairs; Optimal parameters to align cavities from the BO1 set; Revised Vertex dataset of 338 positive and 338 negative pairs; Frag-Lig set of 578 pairs of protein-fragment and related protein-ligand and complexes; Fragment hits for the muscarinic M5 receptor (PDB ID 6OL9); Fragment hits for the TNF-alpha (PDB ID 6OOY); Fragment hits for the cysteinyl leukotriene receptor 2 (PDB ID 6RZ8).

**Abbreviations used**

3D, three-dimensional; PDB, Protein Data Bank; rmsd, root-mean square deviation; RANSAC, Random Sample Consensus; ICP, iterative closest point algorithm; FPFH, Fast Point Feature Histogram; CI, Confidence Interval.

## 2.3.7. References

1. Goodsell, D. S.; Zardecki, C.; Di Costanzo, L.; Duarte, J. M.; Hudson, B. P.; Persikova, I.; Segura, J.; Shao, C.; Voigt, M.; Westbrook, J. D.; Young, J. Y.; Burley, S. K., Rcsb Protein Data Bank: Enabling Biomedical Research and Drug Discovery. *Protein Sci.,* **2020,** *29*, 52-65.

2. Rognan, D., Chemogenomic Approaches to Rational Drug Design. *Br. J. Pharmacol.,* **2007,** *152*, 38-52.

3. Perot, S.; Sperandio, O.; Miteva, M. A.; Camproux, A. C.; Villoutreix, B. O., Druggable Pockets and Binding Site Centric Chemical Space: A Paradigm Shift in Drug Discovery. *Drug Discov. Today,* **2010,** *15*, 656-667.

4. Volkamer, A.; Griewel, A.; Grombacher, T.; Rarey, M., Analyzing the Topology of Active Sites: On the Prediction of Pockets and Subpockets. *J. Chem. Inf. Model.,* **2010,** *50*, 2041-2052.

5. Ehrt, C.; Brinkjost, T.; Koch, O., Impact of Binding Site Comparisons on Medicinal Chemistry and Rational Molecular Design. *J. Med. Chem.,* **2016,** *59*, 4121-4151.

6. Schmitt, S.; Kuhn, D.; Klebe, G., A New Method to Detect Related Function among Proteins Independent of Sequence and Fold Homology. *J. Mol. Biol.,* **2002,** *323*, 387-406.

7. Desaphy, J.; Azdimousa, K.; Kellenberger, E.; Rognan, D., Comparison and Druggability Prediction of Protein-Ligand Binding Sites from Pharmacophore-Annotated Cavity Shapes. *J. Chem. Inf. Model.,* **2012,** *52*, 2287-2299.

8. Le Guilloux, V.; Schmidtke, P.; Tuffery, P., Fpocket: An Open Source Platform for Ligand Pocket Detection. *BMC Bioinformatics,* **2009,** *10*, 168.

9. Goodford, P. J., A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules. *J. Med. Chem.,* **1985,** *28*, 849-857.

10. Laurie, A. T.; Jackson, R. M., Q-Sitefinder: An Energy-Based Method for the Prediction of Protein-Ligand Binding Sites. *Bioinformatics,* **2005,** *21*, 1908-1916.

11. Glaser, F.; Morris, R. J.; Najmanovich, R. J.; Laskowski, R. A.; Thornton, J. M., A Method for Localizing Ligand Binding Pockets in Protein Structures. *Proteins,* **2006,** *62*, 479-488.

12. Huang, B.; Schroeder, M., Ligsitecsc: Predicting Ligand Binding Sites Using the Connolly Surface and Degree of Conservation. *BMC Struct. Biol.,* **2006,** *6*, 19.

13. Halgren, T. A., Identifying and Characterizing Binding Sites and Assessing Druggability. *J. Chem. Inf. Model.,* **2009,** *49*, 377-389.

14. Vukovic, S.; Huggins, D. J., Quantitative Metrics for Drug-Target Ligandability. *Drug Discov. Today,* **2018,** *23*, 1258-1266.

15. Shulman-Peleg, A.; Nussinov, R.; Wolfson, H. J., Recognition of Functional Sites in Protein Structures. *J. Mol. Biol.,* **2004,** *339*, 607-633.

16. Kinoshita, K.; Nakamura, H., Identification of Protein Biochemical Functions by Similarity Search Using the Molecular Surface Database EF-Site. *Protein Sci.,* **2003,** *12*, 1589-1595.

17. Tseng, Y. Y.; Li, W. H., Classification of Protein Functional Surfaces Using Structural Characteristics. *Proc. Natl. Acad. Sci. U. S. A.,* **2012,** *109*, 1170-1175.

18. Konc, J.; Hodoscek, M.; Ogrizek, M.; Trykowska Konc, J.; Janezic, D., Structure-Based Function Prediction of Uncharacterized Protein Using Binding Sites Comparison. *PLoS Comput. Biol.,* **2013,** *9*, e1003341.

19. Willmann, D.; Lim, S.; Wetzel, S.; Metzger, E.; Jandausch, A.; Wilk, W.; Jung, M.; Forne, I.; Imhof, A.; Janzer, A.; Kirfel, J.; Waldmann, H.; Schule, R.; Buettner, R., Impairment of Prostate Cancer Cell Growth by a Selective and Reversible Lysine-Specific Demethylase 1 Inhibitor. *Int. J. Cancer,* **2012,** *131*, 2704-2709.

20. Al-Gharabli, S. I.; Shah, S. T.; Weik, S.; Schmidt, M. F.; Mesters, J. R.; Kuhn, D.; Klebe, G.; Hilgenfeld, R.; Rademann, J., An Efficient Method for the Synthesis of Peptide Aldehyde Libraries Employed in the Discovery of Reversible Sars Coronavirus Main Protease (Sars-Cov Mpro) Inhibitors. *Chembiochem,* **2006,** *7*, 1048-1055.

21. Weber, A.; Casini, A.; Heine, A.; Kuhn, D.; Supuran, C. T.; Scozzafava, A.; Klebe, G., Unexpected Nanomolar Inhibition of Carbonic Anhydrase by Cox-2-Selective Celecoxib: New Pharmacological Opportunities Due to Related Binding Site Recognition. *J. Med. Chem.,* **2004,** *47*, 550-557.

22. Kinnings, S. L.; Liu, N.; Buchmeier, N.; Tonge, P. J.; Xie, L.; Bourne, P. E., Drug Discovery Using Chemical Systems Biology: Repositioning the Safe Medicine Comtan to Treat Multi-Drug and Extensively Drug Resistant Tuberculosis. *PLoS Comput. Biol.,* **2009,** *5*, e1000423.

23. Yang, Y. L.; Li, G. H.; Zhao, D. Y.; Yu, H. Y.; Zheng, X. L.; Peng, X. D.; Zhang, X.; Fu, T.; Hu, X. Q.; Niu, M. S.; Ji, X. F.; Zou, L. B.; Wang, J., Computational Discovery and Experimental Verification of Tyrosine Kinase Inhibitor Pazopanib for the Reversal of Memory and Cognitive Deficits in Rat Model Neurodegeneration. *Chem. Sci.,* **2015,** *6*, 2812-2821.

24. Xie, L.; Li, J.; Xie, L.; Bourne, P. E., Drug Discovery Using Chemical Systems Biology: Identification of the Protein-Ligand Binding Network to Explain the Side Effects of CETP Inhibitors. *PLoS Comput. Biol.,* **2009,** *5*, e1000387.

25. De Franchi, E.; Schalon, C.; Messa, M.; Onofri, F.; Benfenati, F.; Rognan, D., Binding of Protein Kinase Inhibitors to Synapsin I Inferred from Pair-Wise Binding Site Similarity Measurements. *PLoS One,* **2010,** *5*, e12214.

26. Cleves, A. E.; Jain, A. N., Chemical and Protein Structural Basis for Biological Crosstalk between PPARalpha and Cox Enzymes. *J. Comput.-Aided. Mol. Des.,* **2015,** *29*, 101-112.

27. Kakisaka, M.; Sasaki, Y.; Yamada, K.; Kondoh, Y.; Hikono, H.; Osada, H.; Tomii, K.; Saito, T.; Aida, Y., A Novel Antiviral Target Structure Involved in the RNA Binding, Dimerization, and Nuclear Export Functions of the Influenza a Virus Nucleoprotein. *PLoS Pathog.,* **2015,** *11*, e1005062.

28. Schirris, T. J.; Ritschel, T.; Herma Renkema, G.; Willems, P. H.; Smeitink, J. A.; Russel, F. G., Mitochondrial ADP/ATP Exchange Inhibition: A Novel Off-Target Mechanism Underlying Ibipinabant-Induced Myotoxicity. *Sci. Rep.,* **2015,** *5*, 14533.

29. Babu, M.; Beloglazova, N.; Flick, R.; Graham, C.; Skarina, T.; Nocek, B.; Gagarinova, A.; Pogoutse, O.; Brown, G.; Binkowski, A.; Phanse, S.; Joachimiak, A.; Koonin, E. V.; Savchenko, A.; Emili, A.; Greenblatt, J.; Edwards, A. M.; Yakunin, A. F., A Dual Function of the CRISPR-Cas System in Bacterial Antivirus Immunity and DNA Repair. *Mol. Microbiol.,* **2011,** *79*, 484-502.

30. Da Silva, F.; Bret, G.; Teixeira, L.; Gonzalez, C. F.; Rognan, D., Exhaustive Repertoire of Druggable Cavities at Protein-Protein Interfaces of Known Three-Dimensional Structure. *J. Med. Chem.,* **2019,** *62*, 9732-9742.

31. Ehrt, C.; Brinkjost, T.; Koch, O., A Benchmark Driven Guide to Binding Site Comparison: An Exhaustive Evaluation Using Tailor-Made Data Sets (Prospeccts). *PLoS Comput. Biol.,* **2018,** *14*, e1006483.

32. Erlanson, D. A.; Fesik, S. W.; Hubbard, R. E.; Jahnke, W.; Jhoti, H., Twenty Years On: The Impact of Fragments on Drug Discovery. *Nat. Rev. Drug. Discov.,* **2016,** *15*, 605-619.

33. Ramensky, V.; Sobol, A.; Zaitseva, N.; Rubinov, A.; Zosimov, V., A Novel Approach to Local Similarity of Protein Binding Sites Substantially Improves Computational Drug Design Results. *Proteins,* **2007,** *69*, 349-357.

34. Wallach, I.; Lilien, R. H., Prediction of Sub-Cavity Binding Preferences Using an Adaptive Physicochemical Structure Representation. *Bioinformatics,* **2009,** *25*, I296-I304.

35. Durrant, J. D.; Friedman, A. J.; McCammon, J. A., Crystaldock: A Novel Approach to Fragment-Based Drug Design. *J. Chem. Inf. Model.,* **2011,** *51*, 2573-2580.

36. Wood, D. J.; de Vlieg, J.; Wagener, M.; Ritschel, T., Pharmacophore Fingerprint-Based Approach to Binding Site Subpocket Similarity and Its Application to Bioisostere Replacement. *J. Chem. Inf. Model.,* **2012,** *52*, 2031-2043.

37. Jalencas, X.; Mestres, J., Chemoisosterism in the Proteome. *J. Chem. Inf. Model.,* **2013,** *53*, 279-292.

38. Kalliokoski, T.; Olsson, T. S. G.; Vulpetti, A., Subpocket Analysis Method for Fragment-Based Drug Discovery. *J. Chem. Inf. Model.,* **2013,** *53*, 131-141.

39. Tang, G. W.; Altman, R. B., Knowledge-Based Fragment Binding Prediction. *PLoS Comput Biol,* **2014,** *10*, e1003589.

40. Bartolowits, M.; Davisson, V. J., Considerations of Protein Subpockets in Fragment-Based Drug Design. *Chem. Biol. Drug Des.,* **2016,** *87*, 5-20.

41. Cheng, L.; Chen, S.; Liu, X. Q.; Xu, H.; Wu, Y.; Li, M. C.; Chen, Y. M., Registration of Laser Scanning Point Clouds: A Review. *Sensors-Basel,* **2018,** *18*, 1641.

42. Chui, H. L.; Rangarajan, A., A New Point Matching Algorithm for Non-Rigid Registration. *Comput. Vis. Image Und.,* **2003,** *89*, 114-141.

43. Jian, B.; Vemuri, B. C., Robust Point Set Registration Using Gaussian Mixture Models. *IEEE Trans. Pattern Anal. Mach. Intell.,* **2011,** *33*, 1633-1645.

44. http://pointclouds.org/documentation/tutorials/random_sample_consensus.php (accessed January 23, 2020).

45. Fischler, M. A.; Bolles, R. C., Random Sample Consensus - A Paradigm for Model-Fitting with Applications to Image-Analysis and Automated Cartography. *Commun. ACM,* **1981,** *24*, 381-395.

46. Besl, P. J.; Mckay, N. D., A Method for Registration of 3-D Shapes. *IEEE Trans. Pattern Anal. Mach. Intell,* **1992,** *14*, 239-256.

47. Bertolazzi, P.; Guerra, C.; Liuzzi, G., A Global Optimization Algorithm for Protein Surface Alignment. *BMC Bioinformatics,* **2010,** *11*, 488.

48. Polychronidou, E.; Kalamaras, I.; Agathangelidis, A.; Sutton, L. A.; Yan, X. J.; Bikos, V.; Vardi, A.; Mochament, K.; Chiorazzi, N.; Belessi, C.; Rosenquist, R.; Ghia, P.; Stamatopoulos, K.; Vlamos, P.; Chailyan, A.; Overby, N.; Marcatili, P.; Hatzidimitriou, A.; Tzovaras, D., Automated Shape-Based Clustering of 3D Immunoglobulin Protein Structures in Chronic Lymphocytic Leukemia. *BMC Bioinformatics,* **2018,** *19*, 414.

49. Douguet, D.; Payan, F., Sensaas (Sensitive Surface as a Shape): Utilizing Open-Source Algorithms for 3D Point Cloud Alignment of Molecules. *arXiv:1908.11267,* **2019**.

50. Hoffmann, B.; Zaslavskiy, M.; Vert, J. P.; Stoven, V., A New Protein Binding Pocket Similarity Measure Based on Comparison of Clouds of Atoms in 3D: Application to Ligand Prediction. *BMC Bioinformatics,* **2010,** *11*, 99.

51. Milletti, F.; Vulpetti, A., Predicting Polypharmacology by Binding Site Similarity: From Kinases to the Protein Universe. *J. Chem. Inf. Model.,* **2010,** *50*, 1418-1431.

52. Desaphy, J.; Bret, G.; Rognan, D.; Kellenberger, E., sc-PDB: A 3D-Database of Ligandable Binding Sites-10 Years On. *Nucleic Acids Res.,* **2015,** *43*, D399-404.

53. Park, J.; Zhou, Q. Y.; Koltun, V., Colored Point Cloud Registration Revisited. *2017 IEEE Int. Conf. Comput. Vis. (ICCV),* **2017**, 144-152.

54. Simonovsky, M.; Meyers, J., DeeplyTough: Learning Structural Comparison of Protein Binding Sites. *J Chem Inf Model,* **2020**, 60, 2356-2366.

55. Easyfit Version 5.6, Mathwave Technologies, http://www.mathwave.com/ (accessed March 02, 2020).

56. Chen, Y. C.; Tolbert, R.; Aronov, A. M.; McGaughey, G.; Walters, W. P.; Meireles, L., Prediction of Protein Pairs Sharing Common Active Ligands Using Protein Sequence, Structure, and Ligand Similarity. *J. Chem. Inf. Model.,* **2016,** *56*, 1734-1745.

57. Weill, N.; Rognan, D., Alignment-Free Ultra-High-Throughput Comparison of Druggable Protein-Ligand Binding Sites. *J. Chem. Inf. Model.,* **2010,** *50*, 123-135.

58. Yeturu, K.; Chandra, N., Pocketmatch: A New Algorithm to Compare Binding Sites in Protein Structures. *BMC Bioinformatics,* **2008,** *9*, 543.

59. Konc, J.; Janezic, D., Probis Algorithm for Detection of Structurally Similar Protein Binding Sites by Local Structural Alignment. *Bioinformatics,* **2010,** *26*, 1160-1168.

60. Schalon, C.; Surgand, J. S.; Kellenberger, E.; Rognan, D., A Simple and Fuzzy Method to Align and Compare Druggable Ligand-Binding Sites. *Proteins,* **2008,** *71*, 1755-1778.

61. Drwal, M. N.; Bret, G.; Perez, C.; Jacquemard, C.; Desaphy, J.; Kellenberger, E., Structural Insights on Fragment Binding Mode Conservation. *J. Med. Chem.,* **2018,** *61*, 5963-5973.

62. Shindyalov, I. N.; Bourne, P. E., Protein Structure Alignment by Incremental Combinatorial Extension (CE) of the Optimal Path. *Protein Eng.,* **1998,** *11*, 739-747.

63. Marcou, G.; Rognan, D., Optimizing Fragment and Scaffold Docking by Use of Molecular Interaction Fingerprints. *J. Chem. Inf. Model.,* **2007,** *47*, 195-207.

64. Desaphy, J.; Rognan, D., sc-PDB-Frag: A Database of Protein-Ligand Interaction Patterns for Bioisosteric Replacements. *J. Chem. Inf. Model.,* **2014,** *54*, 1908-1918.

65. Congreve, M.; Carr, R.; Murray, C.; Jhoti, H., A 'Rule of Three' for Fragment-Based Lead Discovery? *Drug Discov. Today,* **2003,** *8*, 876-877.

66. Vuckovic, Z.; Gentry, P. R.; Berizzi, A. E.; Hirata, K.; Varghese, S.; Thompson, G.; van der Westhuizen, E. T.; Burger, W. A. C.; Rahmani, R.; Valant, C.; Langmead, C. J.; Lindsley, C. W.; Baell, J. B.; Tobin, A. B.; Sexton, P. M.; Christopoulos, A.; Thal, D. M., Crystal Structure of the M5 Muscarinic Acetylcholine Receptor. *Proc. Natl. Acad. Sci. U. S. A.,* **2019,** *116*, 26001-26007.

67. O'Connell, J.; Porter, J.; Kroeplien, B.; Norman, T.; Rapecki, S.; Davis, R.; McMillan, D.; Arakaki, T.; Burgin, A.; Fox Iii, D.; Ceska, T.; Lecomte, F.; Maloney, A.; Vugler, A.; Carrington, B.; Cossins, B. P.; Bourne, T.; Lawson, A., Small Molecules That Inhibit TNF Signalling by Stabilising an Asymmetric Form of the Trimer. *Nat. Commun.,* **2019,** *10*, 5795.

68. Gusach, A.; Luginina, A.; Marin, E.; Brouillette, R. L.; Besserer-Offroy, É.; Longpré, J.-M.; Ishchenko, A.; Popov, P.; Patel, N.; Fujimoto, T.; Maruyama, T.; Stauch, B.; Ergasheva, M.; Romanovskaia, D.; Stepko, A.; Kovalev, K.; Shevtsov, M.; Gordeliy, V.; Han, G. W.; Katritch, V.; Borshchevskiy, V.; Sarret, P.; Mishin, A.; Cherezov, V., Structural Basis of Ligand Selectivity and Disease Mutations in Cysteinyl Leukotriene Receptors. *Nat. Commun.,* **2019,** *10*, 5573.

69. Lauri, G.; Bartlett, P. A., Caveat - A Program to Facilitate the Design of Organic-Molecules. *J Comput.Aided Mol. Des,* **1994,** *8*, 51-66.

70. Maass, P.; Schulz-Gasch, T.; Stahl, M.; Rarey, M., Recore: A Fast and Versatile Method for Scaffold Hopping Based on Small Molecule Crystal Structure Conformations. *J. Chem. Inf. Model.,* **2007,** *47*, 390-399.

71. Wang, L.-H.; Evers, A.; Monecke, P.; Naumann, T., Ligand Based Lead Generation - Considering Chemical Accessibility in Rescaffolding Approaches Via Brood. *J. Cheminform.,* **2012,** *4*, O20.

72. Imrie, F.; Bradley, A. R.; van der Schaar, M.; Deane, C. M., Deep Generative Models for 3D Linker Design. *J. Chem. Inf. Model.,* **2020,** *60*, 1983-1995.

73. Yang, J.; Li, H.; Campbell, D.; Jia, Y., Go-Icp: A Globally Optimal Solution to 3D ICP Point-Set Registration. *IEEE Trans. Pattern Anal. Mach. Intell.,* **2016,** *38*, 2241-2254.

74. Da Silva, F.; Desaphy, J.; Rognan, D., Ichem: A Versatile Toolkit for Detecting, Comparing, and Predicting Protein-Ligand Interactions. *ChemMedChem,* **2018,** *13*, 507-510.

75. Shrödinger LLC, New York, NY 10036-4041, U.S.A.

76. The Uniprot Consortium, Uniprot: A Worldwide Hub of Protein Knowledge. *Nucleic Acids Res.,* **2019,** *47*, D506-D515.

77. Sievers, F.; Wilm, A.; Dineen, D.; Gibson, T. J.; Karplus, K.; Li, W.; Lopez, R.; McWilliam, H.; Remmert, M.; Soding, J.; Thompson, J. D.; Higgins, D. G., Fast, Scalable Generation of High-Quality Protein Multiple Sequence Alignments Using Clustal Omega. *Mol. Syst. Biol.,* **2011,** *7*, 539.

78. Madeira, F.; Park, Y. M.; Lee, J.; Buso, N.; Gur, T.; Madhusoodanan, N.; Basutkar, P.; Tivey, A. R. N.; Potter, S. C.; Finn, R. D.; Lopez, R., The EMBLl-EBI Search and Sequence Analysis Tools APIs in 2019. *Nucleic Acids Res.,* **2019,** *47*, W636-W641.

79. Certara USA, Inc., Princeton, NJ 08540, U.S.A.

80. Rdkit: Open-Source Cheminformatics Software, https://www.rdkit.org/ (accessed March 07, 2020).

81. Open3D - A Modern Library for 3D Data Processing, http://www.open3d.org/ (accessed January 23, 2020).

82. Rusu, R. B.; Blodow, N.; Beetz, M., Fast Point Feature Histograms (FPFH) for 3D Registration. *IEEE Int. Conf. Robot.,* **2009**, 1848-1853.

83. Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E., UCSF Chimera - A Visualization System for Exploratory Research and Analysis. *J. Comput. Chem.,* **2004,** *25*, 1605-1612.

84. Scikit-Learn: Machine Learning in Python, Https://Scikit-Learn.Org/Stable/ (accessed January 23, 2020).

85. Openeye Scientific Software, Santa Fe, NM 87508, U.S.A.

86. Structural Chemogenomics Group, Laboratoire d'Innovation Thérapeutique, UMR 7200 CNRS-Université de Strasbourg, http://bioinfo-pharma.u-strasbg.fr/labwebsite/download.Html (accessed May 15, 2020).

87. Ligand Expo, http://ligand-expo.rcsb.org (accessed  April 24, 2020).

88. PocketMatch, http://proline.physics.Iisc.ernet.in/pocketmatch/ (accessed May 15, 2020).

89. In Silico Laboratory for Innovation in drug Discovery, http://insilab.org/probis-algorithm/ (Accessed May 18, 2020).

90. NumPy, https://numpy.org/ (Accessed January 23, 2020).

91. SciPy, Https://www.scipy.org/ (Accessed March 02, 2020).

## 2.3.8. Supporting information for *A computer vision approach to align and compare protein cavities: Application to fragment-based drug design*

**Figure S1.** Properties of the BO1 data set of 766 protein-ligand cavity pairs

**Figure S2.** Example of misalignment for a pair of similar cavities from the BO1 set.

**Figure S3.** Distribution of pocket size for fragments (light blue) and full cavities (dark blue). Size is expressed as the number of points (voxel centers) encompassing the pocket placed in a 2 Å-regular 3D lattice.

**Figure S4.** ProCare overlay of cavities from unrelated targets.

**Table S1.** EASY1 set of similar and dissimilar pairs

**Table S2.** List of BO1 similar pairs

**Table S3.** List of BO1 dissimilar pairs

**Table S4.** Optimal parameters to align cavities from the BO1 set.

**Table S5.** Revised Vertex dataset of 338 positive and 338 negative pairs

**Table S6.** Frag-Lig set of 578 pairs of protein-fragment and related protein-ligand and complexes.

**Table S7.** Fragment hits for the muscarinic M5 receptor (PDB ID 6OL9).

**Table S8.** Fragment hits for the TNF-alpha (PDB ID 6OOY).
**Table S9.** Fragment hits for the cysteinyl leukotriene receptor 2 (PDB ID 6RZ8).
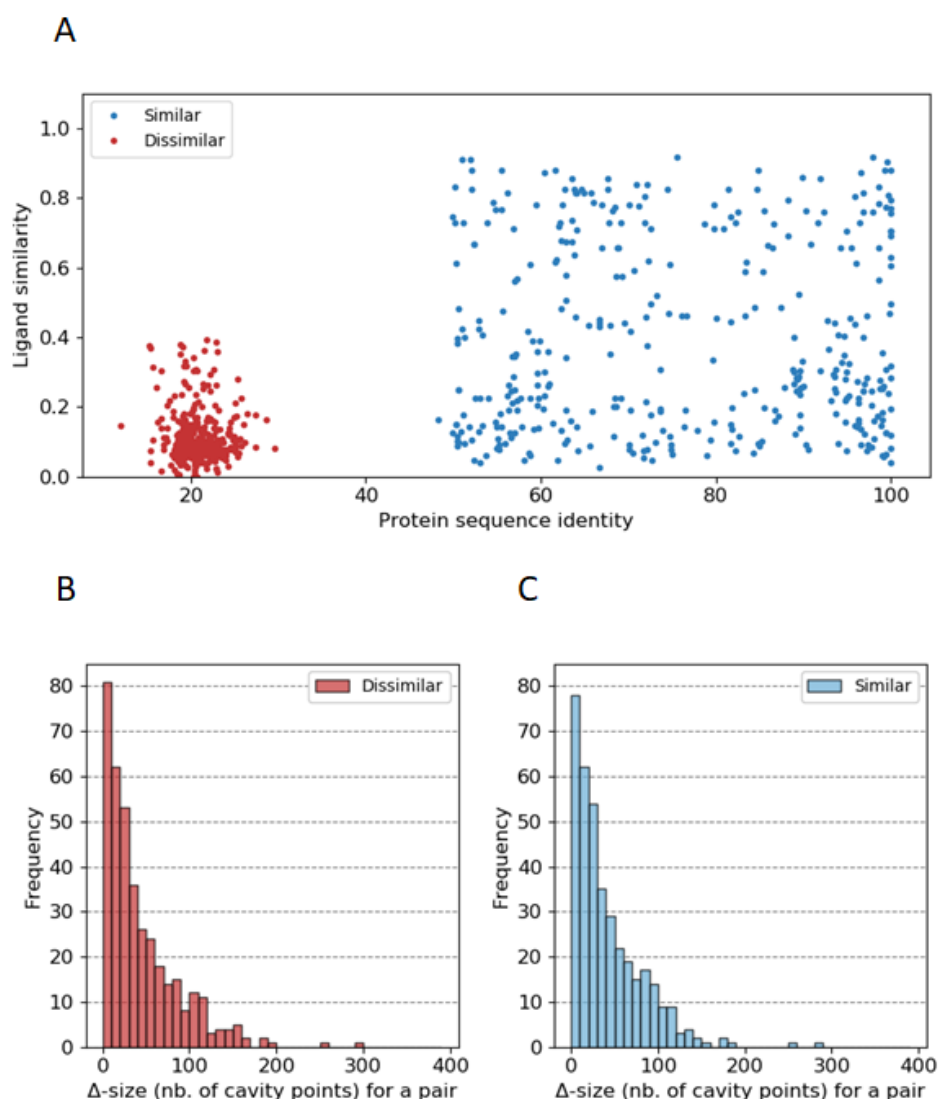
A



B



C



**Figure S1.** Properties of the BO1 data set of 766 protein-ligand cavity pairs (383 similar, 383 dissimilar). Because the notion of similarity and dissimilarity of protein pockets is context-dependent, we defined two similar cavities as deriving from pairs of different proteins (different Uniprot accession numbers) that are similar in terms of sequence (50-100% identity), structure (rmsd on backbone atoms ≤ 5 Å) and functions (Uniprot keywords annotation). No constraint was applied on the bound-ligand chemical similarity, so that different cases are represented (0 ≤ chemical similarity < 1; see Computational methods for similarity calculation). Conversely, pairs of dissimilar cavities were formed from the same target space, but need to be different in terms of function and bound ligands (0 ≤ chemical similarity ≤ 0.4) in order to rule out potential wrong class annotations. The final sets of similar and dissimilar cavities have comparable distribution of size (i.e. number of points) difference between members of each pair, with the aim of eliminating possible biases in results due to alignment of differently-sized objects.

**A)** Protein-bound ligand similarity (Tanimoto coefficient from Morgan fingerprints) vs. protein sequence identity (PIM of Clustal Omega alignment with default parameters) for similar (blue) and dissimilar pairs (redbrick); **B)** Distribution of the difference in the size of cavity point clouds for dissimilar pairs; **C)** Distribution of the difference in the size of cavity point clouds for similar pairs.
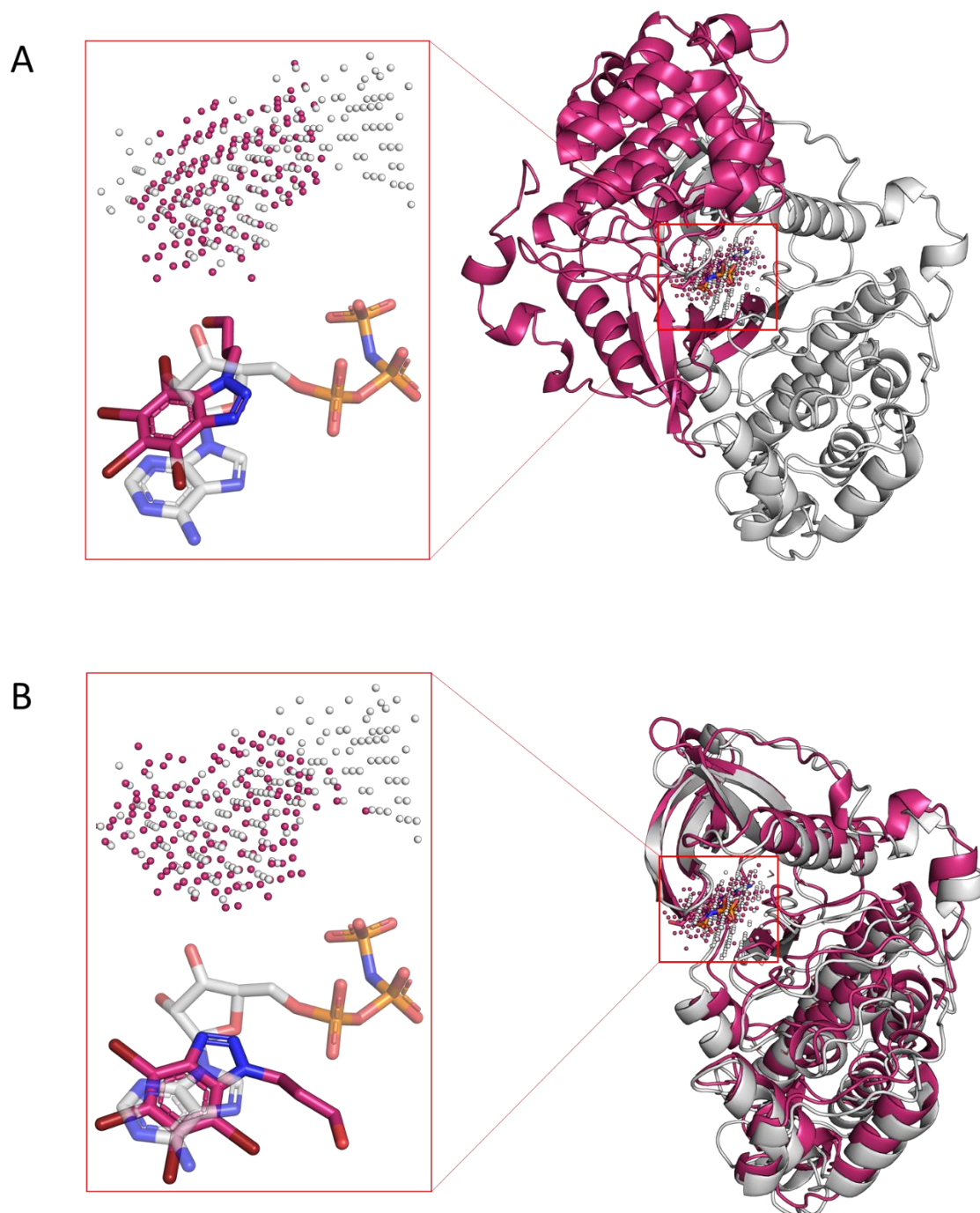


**Figure S2.** Example of misalignment for a pair of similar cavities from the BO1 set. **A)** ProCare FPFH-icp alignment of 3-(4,5,6,7-tetrabromo-1H-benzotriazol-1-yl)propan-1-ol cavity in casein kinase II subunit alpha' (PDB ID: 3OFM, HET: 4B0) to phosphoaminophosphonic acid-adenylate ester cavity in

casein kinase II subunit alpha (PDB ID: 3U87, HET:ANP). The ProCare transformation matrix was applied to ligand and protein atomic coordinates and showed misalignment of proteins (rmsd of protein backbone heavy atoms: 43 Å); **B)** ProCare c-FPFH-icp correct alignment (rmsd of proteins backbone heavy atoms: 3.1 Å) of the same pair.
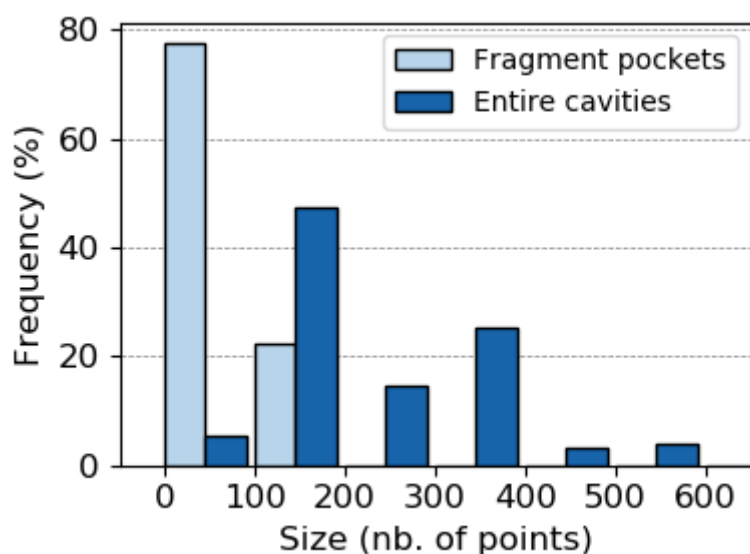


**Figure S3.** Distribution of pocket size for fragments (light blue) and full cavities (dark blue). Size is expressed as the number of points (voxel centers) encompassing the pocket placed in a 1.5 Å-regular 3D lattice.
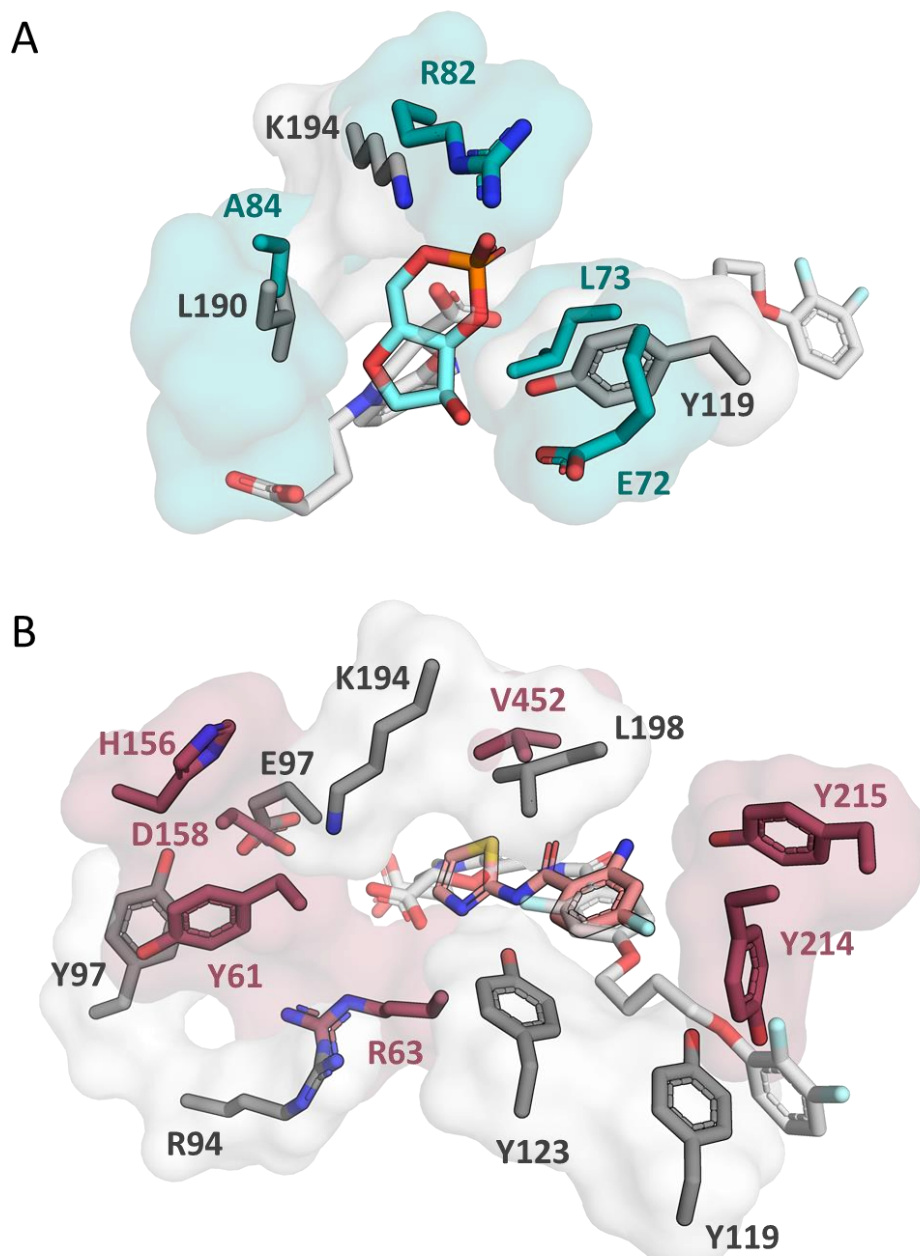
**Figure S4.** ProCare overlay of cavities from unrelated targets. **A)** Alignment of a phosphoribosyl-binding subpocket in catabolite activator protein CAP (PDB ID: 1RUO, HET: CMP) to full ONO-2080365 binding site in cysteinyl leukotriene receptor 2 CYSTLR2 (PDB ID: 6RZ8, HET: KNZ). The derived transformation matrix was applied to the corresponding fragments and protein atomic coordinates. In both proteins, basic residues (K194 in CYSTLR2 and R82 in CAP) interacting with acidic groups in ligands were matched. Hydrogen-bond acceptors (Y119 in CYSTLR2 and E72 in CAP), aliphatic hydrophobic residues (L190 in CYSTLR2 and A84 in CAP) are also matched; **B)** Alignment of N-(1,3-thiazol-2-yl)benzamide moiety binding environment in glucokinase (PDB ID: 3F9M, HET: MRK) to full ONO-2080365 binding site in cysteinyl leukotriene receptor 2 CYSTLR2 (PDB ID: 6RZ8, HET: KNZ). The derived transformation matrix was applied to the corresponding fragments and protein atomic coordinates. In both proteins, aromatic residues (Y97, Y119, Y123 in CYSTLR2 and Y61, Y214,

Y215 in glucokinase), basic residues (K194 in CYSTLR2 and H156 in glucokinase), acidic residues (E97 in CYSTLR2 and D158 in glucokinase), aliphatic hydrophobic residues (L198 in CYSTLR2 and V452 in CAP) were matched**.**

**Table S1.** EASY1 set of five similar and five dissimilar pairs

| | β2-adrenergic receptor | | Estrogen receptor α | | Cyclin-dependent kinase 2 | | HIV-1 protease | | Glutamate receptor 2 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 2RH1 | 5D6L | 2OUZ | 3ERT | 2C6T | 1DM2 | 1C6X | 2B7Z | 1FTL | 1LB9 |
| 2RH1 / 5D6L | | 🟩 | 🟥 | | | 🟥 | | 🟥 | | |
| 2OUZ / 3ERT | | | | 🟩 | | 🟥 | | | | |
| 2C6T / 1DM2 | | | | | 🟩 | | | 🟥 | | |
| 1C6X / 2B7Z | | | | | | | | 🟩 | | |
| 1FTL / 1LB9 | | | | | | | | | | 🟩 |

Five targets are represented by two protein-ligand complexes, each (PDB identifiers given as column and row names). The pairs of similar and dissimilar cavities are displayed by green and red boxes, respectively.

**Table S2.** List of BO1 similar pairs

The list of the 383 pairs of similar cavities is available as supporting information at:
https://doi.org/10.1021/acs.jmedchem.0c00422: jm0c00422_si_001.pdf

**Table S3.** List of BO1 dissimilar pairs

The list of the 383 pairs of dissimilar cavities is available as supporting information at:
https://doi.org/10.1021/acs.jmedchem.0c00422: jm0c00422_si_001.pdf

**Table S4.** Optimal Open3D parameters to align cavities from the BO1 set.

| Parameter | Value |
|---|---|
| RANSAC cycle number of validations, *rn* | 4 |
| RANSAC maximum number of validations, *rv* | 500 |
| RANSAC maximum number of iterations, *ri* | 4,000,000 |
| Rough alignment transformation estimation type, *gt* | TransformationEstimationPointToPoint |
| Rough alignment distance threshold in Å, *gd* | 1.5 |
| Checkers similarity threshold, *cs* | 0.9 |
| ICP alignment transformation estimation type, *it* | TransformationEstimationPointToPoint |
| ICP alignment distance threshold in Å, *id* | 3 |
| ICP maximum iterations, *ii* | 100 |
| ICP relative fitness threshold, *if* | 10-6 |
| ICP relative RMSE threshold, *ir* | 10-6 |
| Nearest neighbor search radius for normals in Å, *nr* | 3.1 |
| Maximum number of neighbors for normal, *nm* | 471 |
| Nearest neighbor search radius for FPFH in Å, *fr* | 3.1 |
| Maximum neighbors for FPFH, *fm* | 135 |

**Table S5.** Revised Vertex dataset of 338 positive and 338 negative pairs

The list of the pairs of similar and dissimilar cavities is available as supporting information at:
https://doi.org/10.1021/acs.jmedchem.0c00422: jm0c00422_si_001.pdf

**Table S6.** Frag-Lig set of 578 pairs of protein-fragment and related protein-ligand and complexes.

The list of the pairs of cavities is available as supporting information at:
https://doi.org/10.1021/acs.jmedchem.0c00422: jm0c00422_si_001.pdf

**Table S7.** Fragment hits for the muscarinic M5 receptor (PDB ID 6OL9).

| FragID[a] | Protein name | Rank | ProCaRe[b] | IFP[c] | IFP_polar[d] | FragScore[e] |
|---|---|---|---|---|---|---|
| 5CXV_0HK_1_1 | Muscarinic acetylcholine receptor m1 | 1 | 0.82 | 0.54 | 0.50 | 1.61 |
| 1N43_BTN_1_1 | Streptavidin | 2 | 0.48 | 0.54 | 0.50 | 1.27 |
| 1UMK_FAD_1_1 | Nadh-cytochrome b5 reductase | 3 | 0.60 | 0.42 | 0.50 | 1.26 |
| 1YRO_GDU_2_2 | Alpha-lactalbumin | 4 | 0.48 | 0.46 | 0.50 | 1.19 |
| 1C0I_BE2_2_1 | D-amino acid oxydase | 5 | 0.63 | 0.27 | 0.50 | 1.16 |
| 4U16_OHK_1_1 | Muscarinic acetylcholine receptor m3 | 6 | 0.72 | 0.43 | 0.00 | 1.15 |
| 3HV6_R39_1_1 | Mitogen-activated protein kinase 14 | 7 | 0.56 | 0.33 | 0.50 | 1.14 |
| 3RPE_FAD_1_1 | Modulator of drug activity b | 8 | 0.62 | 0.25 | 0.50 | 1.12 |
| 3U2L_FAD_1_1 | Fad-linked sulfhydryl oxidase alr | 9 | 0.61 | 0.25 | 0.50 | 1.11 |
| 3AOS_JH2_1_1 | Hemolymph juvenile hormone binding protein | 10 | 0.57 | 0.53 | 0.00 | 1.10 |
| 1ZZ1_SHH_1_1 | Histone deacetylase-like amidohydrolase | 11 | 0.56 | 0.54 | 0.00 | 1.10 |
| 4BMZ_MTA_1_1 | Mta/sah nucleosidase | 12 | 0.47 | 0.36 | 0.50 | 1.08 |
| 2YG3_FAD_2_3 | Putrescine oxidase | 13 | 0.49 | 0.33 | 0.50 | 1.07 |
| 1S3V_TQD_1_2 | Dihydrofolate reductase | 14 | 0.53 | 0.54 | 0.00 | 1.07 |
| 3HZG_FAD_1_1 | Thymidylate synthase thyx | 15 | 0.48 | 0.33 | 0.50 | 1.07 |
| 1QJX_W02_1_3 | None | 16 | 0.58 | 0.45 | 0.00 | 1.04 |
| 2EIX_FAD_2_1 | Nadh-cytochrome b5 reductase | 17 | 0.47 | 0.31 | 0.50 | 1.03 |
| 3QCI_NX3_1_2 | Receptor-type tyrosine-protein phosphatase gamma | 18 | 0.64 | 0.38 | 0.00 | 1.03 |
| 3G5E_Q74_1_1 | Aldose reductase | 19 | 0.52 | 0.50 | 0.00 | 1.02 |
| 4U15_0HK_1_1 | Muscarinic acetylcholine receptor m3 | 20 | 0.67 | 0.36 | 0.00 | 1.02 |
| 3VLN_ASC_1_1 | Glutathione s-transferase omega-1 | 21 | 0.54 | 0.23 | 0.50 | 1.02 |
| 4H96_14Q_1_3 | Dihydrofolate reductase | 22 | 0.55 | 0.31 | 0.33 | 1.02 |
| 4B1I_A8P_1_2 | Poly(adp-ribose) glycohydrolase | 23 | 0.52 | 0.25 | 0.50 | 1.02 |
| 3VTB_TKA_1_1 | Vitamin d3 receptor | 24 | 0.52 | 0.50 | 0.00 | 1.02 |
| 3ETE_NDP_11_3 | Glutamate dehydrogenase | 25 | 0.48 | 0.29 | 0.50 | 1.01 |
| 2BF4_FAD_1_1 | Nadph-cytochrome p450 reductase | 26 | 0.51 | 0.25 | 0.50 | 1.01 |
| 4JJU_1MB_2_1 | Genome polyprotein | 27 | 0.51 | 0.50 | 0.00 | 1.01 |
| 2PDG_47D_1_1 | Aldose reductase | 28 | 0.62 | 0.38 | 0.00 | 1.01 |
| 4AA0_AA0_1_3 | Mitogen-activated protein kinase 14 | 29 | 0.48 | 0.27 | 0.50 | 1.01 |
| 1VOT_HUP_1_1 | Acetylcholinesterase | 30 | 0.54 | 0.47 | 0.00 | 1.00 |
| 1OE0_TTP_2_2 | Deoxyribonucleoside kinase | 31 | 0.60 | 0.15 | 0.50 | 1.00 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 3PX3_T3Q_2_1 | N-methyltransferase | 32 | 0.49 | 0.25 | 0.50 | 0.99 |
| 4U14_0HK_1_1 | Muscarinic acetylcholine receptor m3 | 33 | 0.56 | 0.43 | 0.00 | 0.99 |
| 4GCA_2X9_1_2 | Aldose reductase | 34 | 0.61 | 0.38 | 0.00 | 0.99 |
| 2G27_4LG_2_1 | Renin | 35 | 0.56 | 0.42 | 0.00 | 0.98 |
| 3GHR_LDT_1_2 | Aldose reductase | 36 | 0.59 | 0.38 | 0.00 | 0.98 |
| 2PDX_ZST_1_2 | Aldose reductase | 37 | 0.54 | 0.44 | 0.00 | 0.98 |
| 4KNI_E1E_1_2 | Carbonic anhydrase 2 | 38 | 0.54 | 0.43 | 0.00 | 0.97 |
| 1LCZ_BH7_2_1 | Streptavidin | 39 | 0.55 | 0.42 | 0.00 | 0.97 |
| 4GDA_BTN_1_1 | Streptavidin | 40 | 0.52 | 0.45 | 0.00 | 0.97 |
| 3E93_19B_1_4 | Mitogen-activated protein kinase 14 | 41 | 0.64 | 0.33 | 0.00 | 0.97 |
| 3PX2_T3Q_2_1 | N-methyltransferase | 42 | 0.48 | 0.23 | 0.50 | 0.96 |
| 3OU7_SAM_2_1 | Sam-dependent methyltransferase | 43 | 0.52 | 0.44 | 0.00 | 0.95 |
| 2BAB_FAD_1_3 | Putative aminooxidase | 44 | 0.52 | 0.43 | 0.00 | 0.95 |
| 2PDB_ZST_1_2 | Aldose reductase | 45 | 0.49 | 0.46 | 0.00 | 0.95 |
| 5PAH_LDP_1_1 | Phenylalanine 4-monooxygenase | 46 | 0.55 | 0.15 | 0.50 | 0.95 |
| 1QIW_DPD_2_2 | Calmodulin | 47 | 0.55 | 0.40 | 0.00 | 0.95 |
| 3G70_A5T_1_3 | Renin | 48 | 0.56 | 0.38 | 0.00 | 0.95 |
| 3LBO_LDT_1_2 | Aldose reductase | 49 | 0.56 | 0.38 | 0.00 | 0.95 |
| 4A6D_SAM_1_1 | Hydroxyindole o-methyltransferase | 50 | 0.48 | 0.47 | 0.00 | 0.94 |
| 4GBD_MCF_1_2 | Methylthioadenosine deaminase | 51 | 0.47 | 0.47 | 0.00 | 0.94 |
| 4XUG_F9F_1_1 | Tryptophan synthase alpha chain | 52 | 0.47 | 0.47 | 0.00 | 0.94 |
| 3G72_A6T_1_3 | Renin | 53 | 0.48 | 0.46 | 0.00 | 0.94 |
| 1AH4_NAP_1_3 | Aldose reductase | 54 | 0.48 | 0.29 | 0.33 | 0.94 |
| 2PD9_FID_1_1 | Aldose reductase | 55 | 0.50 | 0.44 | 0.00 | 0.94 |
| 2HVO_ZST_1_2 | Aldose reductase | 56 | 0.51 | 0.43 | 0.00 | 0.94 |
| 3N7H_DE3_1_1 | Odorant binding protein | 57 | 0.51 | 0.43 | 0.00 | 0.94 |
| 2HNZ_PC0_1_2 | Reverse transcriptase/ribonuclease h | 58 | 0.47 | 0.46 | 0.00 | 0.93 |
| 4JUA_TZD_1_1 | Benzoylformate decarboxylase | 59 | 0.63 | 0.14 | 0.33 | 0.93 |
| 4BFP_SWY_2_4 | Tankyrase-2 | 60 | 0.48 | 0.45 | 0.00 | 0.93 |
| 3T7R_6PP_1_1 | Putative methyltransferase | 61 | 0.48 | 0.29 | 0.33 | 0.93 |
| 2CND_FAD_1_1 | Nadh-dependent nitrate reductase | 62 | 0.62 | 0.31 | 0.00 | 0.93 |
| 2FZ9_ZST_1_2 | Aldose reductase | 63 | 0.50 | 0.43 | 0.00 | 0.93 |
| 3LCC_SAH_1_1 | Putative methyl chloride transferase | 64 | 0.50 | 0.43 | 0.00 | 0.93 |
| 1T64_TSN_2_1 | Histone deacetylase 8 | 65 | 0.51 | 0.41 | 0.00 | 0.92 |
| 4EMD_C5P_1_2 | 4-diphosphocytidyl-2-c-methyl-d-erythritol kinase | 66 | 0.59 | 0.08 | 0.50 | 0.92 |
| 4R5W_XAV_2_1 | Poly [adp-ribose] polymerase 1 | 67 | 0.56 | 0.36 | 0.00 | 0.92 |
| 2PD5_ZST_1_2 | Aldose reductase | 68 | 0.49 | 0.43 | 0.00 | 0.92 |
| 2IU8_UD1_1_2 | Udp-3-o-[3-hydroxymyristoyl] glucosamine n-acyltransferase | 69 | 0.59 | 0.08 | 0.50 | 0.92 |
| 2I65_NAD_2_1 | Adp-ribosyl cyclase 1 | 70 | 0.62 | 0.30 | 0.00 | 0.92 |
| 3UFL_508_1_2 | Beta-secretase 1 | 71 | 0.65 | 0.27 | 0.00 | 0.92 |
| 4UM3_09R_17_2 | Acetylcholine binding protein | 72 | 0.61 | 0.31 | 0.00 | 0.92 |
| 1G3M_PCQ_1_1 | Estrogen sulfotransferase | 73 | 0.58 | 0.33 | 0.00 | 0.92 |
| 3L8S_BFF_1_2 | Mitogen-activated protein kinase 14 | 74 | 0.49 | 0.43 | 0.00 | 0.92 |
| 1M51_TSX_1_2 | Phosphoenolpyruvate carboxykinase | 75 | 0.58 | 0.33 | 0.00 | 0.91 |
| 4YFY_0FX_1_3 | Viof | 76 | 0.70 | 0.21 | 0.00 | 0.91 |
| 2A8Y_MTA_7_1 | 5'-methylthioadenosine phosphorylase | 77 | 0.48 | 0.43 | 0.00 | 0.91 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 3QCL_NXV_1_2 | Receptor-type tyrosine-protein phosphatase gamma | 78 | 0.49 | 0.42 | 0.00 | 0.91 |
| 3NJQ_NJQ_2_1 | ORF 17 | 79 | 0.52 | 0.38 | 0.00 | 0.91 |
| 4A79_P1B_2_1 | Amine oxidase [flavin-containing] B | 80 | 0.66 | 0.25 | 0.00 | 0.91 |
| 4BU9_08C_2_2 | Tankyrase-2 | 81 | 0.54 | 0.36 | 0.00 | 0.91 |
| 1HQT_NAP_1_3 | Aldehyde reductase | 82 | 0.56 | 0.22 | 0.25 | 0.91 |
| 2FZ8_ZST_1_1 | Aldose reductase | 83 | 0.53 | 0.38 | 0.00 | 0.91 |
| 2O5D_VR1_2_1 | HCV | 84 | 0.48 | 0.18 | 0.50 | 0.90 |
| 3G1O_RF1_1_1 | Transcriptional regulatory repressor protein (tetr-family) ethr | 85 | 0.50 | 0.40 | 0.00 | 0.90 |
| 3NWE_662_1_3 | Catechol o-methyltransferase | 86 | 0.54 | 0.36 | 0.00 | 0.90 |
| 2JGS_BTN_3_1 | Circular permutant of avidin | 87 | 0.55 | 0.35 | 0.00 | 0.90 |
| 1O5P_CHR_1_2 | Neocarzinostatin | 88 | 0.56 | 0.33 | 0.00 | 0.90 |
| 1IKV_EFZ_1_2 | Pol polyprotein | 89 | 0.65 | 0.25 | 0.00 | 0.90 |
| 1SM4_FAD_2_3 | Chloroplast ferredoxin-nadp+ oxidoreductase | 90 | 0.48 | 0.25 | 0.33 | 0.89 |
| 3W2E_FAD_1_1 | Nadh-cytochrome b5 reductase 3 | 91 | 0.56 | 0.33 | 0.00 | 0.89 |
| 2Q96_A18_1_2 | Methionine aminopeptidase | 92 | 0.66 | 0.23 | 0.00 | 0.89 |
| 3QCM_NXW_1_3 | Receptor-type tyrosine-protein phosphatase gamma | 93 | 0.53 | 0.36 | 0.00 | 0.89 |
| 4I5X_FLF_1_1 | Aldo-keto reductase family 1 member b10 | 94 | 0.50 | 0.39 | 0.00 | 0.89 |
| 1FRB_ZST_1_2 | Fr-1 protein | 95 | 0.51 | 0.38 | 0.00 | 0.89 |
| 2V8P_CDP_3_1 | 4-diphosphocytidyl-2-c-methyl-d-erythritol kinase | 96 | 0.56 | 0.08 | 0.50 | 0.88 |
| 4M7V_RAR_1_3 | Dihydrofolate reductase | 97 | 0.48 | 0.40 | 0.00 | 0.88 |
| 3TVX_PNX_1_1 | Camp-specific 3' | 98 | 0.53 | 0.36 | 0.00 | 0.88 |
| 1PAX_DHQ_1_1 | Poly(adp-ribose) polymerase | 99 | 0.52 | 0.36 | 0.00 | 0.88 |
| 3O8H_O8H_1_1 | Transcriptional regulatory repressor protein (tetr-family) ethr | 100 | 0.48 | 0.23 | 0.33 | 0.88 |

[a] Fragment name (PDB_HET_C_M) is inferred from the cognate target PDB identifier (PDB), the corresponding ligand chemical component (HET), the target cavity identifier (C), and the fragment number (N).

[b] cavity similarity score, computed by ProCare, between the fragment-bound subpocket and the query target cavity

[c] Interaction fingerprint similarity, computed with IChem, between the subpocket-fragment interaction fingerprint and the query target-ligand interaction fingerprint

[d] Interaction fingerprint similarity (polar interactions only), computed with IChem, between the subpocket-fragment interaction fingerprint and the query target-ligand interaction fingerprint

[e] FragScore = ProCare + IFP + 0.5*(IFP_polar)

**Table S8.** Fragment hits for the TNF-alpha (PDB ID 6OOY).

| FragID[a] | Protein name | Rank | ProCaRe[b] | IFP[c] | IFP_polar[d] | FragScore[e] |
|---|---|---|---|---|---|---|
| 4KZ0_1UJ_1_1 | Phosphatidylinositol 4,5-bisphosphate 3-kinase | 1 | 0.57 | 0.67 | 0.50 | 1.48 |
| 4CCB_OFG_1_4 | Alk tyrosine kinase receptor | 2 | 0.72 | 0.67 | 0.00 | 1.39 |
| 4NQM_Y1Z_1_3 | Bromodomain-containing protein 4 | 3 | 0.60 | 0.57 | 0.33 | 1.34 |
| 3K3K_A8S_1_1 | Abscisic acid receptor pyr1 | 4 | 0.51 | 0.64 | 0.33 | 1.32 |
| 1VRT_NVP_1_2 | HIV-1 reverse transcriptase | 5 | 0.63 | 0.50 | 0.33 | 1.30 |
| 3K90_A8S_1_1 | Abscisic acid receptor pyr1 | 6 | 0.55 | 0.57 | 0.33 | 1.29 |
| 3R04_UNQ_1_1 | Proto-oncogene serine/threonine-protein kinase pim-1 | 7 | 0.65 | 0.46 | 0.33 | 1.28 |
| 1LW0_NVP_1_2 | HIV-1 reverse transcriptase | 8 | 0.61 | 0.50 | 0.33 | 1.27 |
| 4IWC_1GV_2_1 | Estrogen receptor | 9 | 0.73 | 0.55 | 0.00 | 1.27 |
| 4OTY_LUR_2_1 | Prostaglandin g/h synthase 2 | 10 | 0.62 | 0.64 | 0.00 | 1.26 |
| 3UMW_596_1_2 | Proto-oncogene serine/threonine-protein kinase pim-1 | 11 | 0.70 | 0.56 | 0.00 | 1.25 |
| 1LWE_NVP_1_2 | HIV-1 reverse transcriptase | 12 | 0.58 | 0.50 | 0.33 | 1.25 |
| 2L85_L85_1_1 | Creb-binding protein | 13 | 0.58 | 0.50 | 0.33 | 1.25 |
| 3TUC_FPW_1_2 | Tyrosine-protein kinase syk | 14 | 0.67 | 0.57 | 0.00 | 1.24 |
| 4NYW_2O3_1_2 | Creb-binding protein | 15 | 0.53 | 0.54 | 0.33 | 1.24 |
| 3BTO_SSB_1_1 | Liver alcohol dehydrogenase | 16 | 0.63 | 0.60 | 0.00 | 1.23 |
| 1YDT_IQB_1_1 | C-AMP-dependent protein kinase | 17 | 0.61 | 0.50 | 0.25 | 1.23 |
| 4HXM_1A8_1_1 | Bromodomain-containing protein 4 | 18 | 0.57 | 0.53 | 0.25 | 1.23 |
| 1NDE_MON_1_3 | Estrogen receptor beta | 19 | 0.63 | 0.60 | 0.00 | 1.23 |
| 4F9W_LM4_3_3 | Mitogen-activated protein kinase 14 | 20 | 0.67 | 0.56 | 0.00 | 1.22 |
| 4DFL_0K0_1_1 | Tyrosine-protein kinase syk | 21 | 0.56 | 0.50 | 0.33 | 1.22 |
| 1Q3E_PCG_2_1 | Potassium/sodium hyperpolarization-activated cyclic nucleotide-gated channel 2 | 22 | 0.57 | 0.40 | 0.50 | 1.22 |
| 3RR3_FLR_3_2 | Prostaglandin g/h synthase 2 | 23 | 0.68 | 0.54 | 0.00 | 1.22 |
| 4IV2_1GR_2_1 | Estrogen receptor | 24 | 0.64 | 0.57 | 0.00 | 1.21 |
| 1JLQ_SBN_1_1 | HIV-1 reverse transcriptase | 25 | 0.59 | 0.62 | 0.00 | 1.21 |
| 3EVC_SAH_1_1 | RNA-directed rna polymerase ns5 | 26 | 0.49 | 0.38 | 0.67 | 1.21 |
| 2NNL_ERD_1_1 | Dihydroflavonol 4-reductase | 27 | 0.54 | 0.50 | 0.33 | 1.20 |
| 4CFL_8DQ_1_1 | Brd4 protein | 28 | 0.57 | 0.47 | 0.33 | 1.20 |
| 3RIN_I2O_1_2 | Mitogen-activated protein kinase 14 | 29 | 0.61 | 0.43 | 0.33 | 1.20 |
| 3K3J_I46_2_1 | Mitogen-activated protein kinase 14 | 30 | 0.73 | 0.46 | 0.00 | 1.20 |
| 3CX5_SMA_2_1 | Cytochrome b-c1 complex subunit 1 | 31 | 0.62 | 0.41 | 0.33 | 1.19 |
| 3GB2_G3B_1_2 | Glycogen synthase kinase-3 beta | 32 | 0.57 | 0.46 | 0.33 | 1.19 |
| 1S1X_NVP_1_2 | HIV-1 reverse transcriptase | 33 | 0.60 | 0.43 | 0.33 | 1.19 |
| 4G1W_G1W_1_1 | Mitogen-activated protein kinase 8 | 34 | 0.58 | 0.62 | 0.00 | 1.19 |
| 3V49_PK0_1_1 | Androgen receptor | 35 | 0.56 | 0.47 | 0.33 | 1.19 |
| 4F9Y_GG5_1_3 | Mitogen-activated protein kinase 14 | 36 | 0.77 | 0.42 | 0.00 | 1.18 |
| 2X0W_X0W_1_1 | Cellular tumor antigen p53 | 37 | 0.55 | 0.47 | 0.33 | 1.18 |
| 2XIZ_XIZ_1_1 | Proto-oncogene serine/threonine protein kinase pim-1 | 38 | 0.68 | 0.50 | 0.00 | 1.18 |
| 4NG5_PFB_4_1 | Alcohol dehydrogenase e chain | 39 | 0.55 | 0.64 | 0.00 | 1.18 |
| 3Q7D_NPX_1_1 | Prostaglandin g/h synthase 2 | 40 | 0.67 | 0.50 | 0.00 | 1.17 |
| 4PWD_NVP_1_2 | HIV-1 reverse transcriptase | 41 | 0.58 | 0.43 | 0.33 | 1.17 |
| 2JJ3_JJ3_2_1 | Estrogen receptor beta | 42 | 0.64 | 0.53 | 0.00 | 1.17 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 2L1R_SXK_1_2 | Troponin c | 43 | 0.71 | 0.46 | 0.00 | 1.17 |
| 3K14_535_1_1 | 2-c-methyl-d-erythritol 2 | 44 | 0.50 | 0.67 | 0.00 | 1.17 |
| 3N6U_TSU_1_1 | Lysr type regulator of tsambcd | 45 | 0.64 | 0.36 | 0.33 | 1.16 |
| 2XIY_XIY_1_1 | Proto-oncogene serine/threonine protein kinase pim-1 | 46 | 0.70 | 0.46 | 0.00 | 1.16 |
| 4I5H_G17_1_2 | Mitogen-activated protein kinase 1 | 47 | 0.60 | 0.56 | 0.00 | 1.16 |
| 4ZHX_C1V_1_2 | 5'-amp-activated protein kinase catalytic subunit alpha-2 | 48 | 0.71 | 0.44 | 0.00 | 1.16 |
| 2CLF_F6F_1_1 | Tryptophan synthase alpha chain | 49 | 0.54 | 0.62 | 0.00 | 1.15 |
| 4OKT_198_1_1 | Androgen receptor | 50 | 0.48 | 0.67 | 0.00 | 1.15 |
| 3MSS_MS7_4_2 | Tyrosine-protein kinase ABL1 | 51 | 0.60 | 0.55 | 0.00 | 1.15 |
| 3LP1_NVP_2_2 | HIV-1 reverse transcriptase | 52 | 0.55 | 0.43 | 0.33 | 1.15 |
| 2WUZ_TPF_2_1 | Lanosterol 14-alpha-demethylase | 53 | 0.52 | 0.63 | 0.00 | 1.15 |
| 2WMW_ZYW_1_1 | Serine/threonine-protein kinase chk1 | 54 | 0.65 | 0.50 | 0.00 | 1.15 |
| 5DQ8_FLF_2_1 | Transcriptional enhancer factor tef-4 | 55 | 0.57 | 0.57 | 0.00 | 1.14 |
| 4F4P_0SB_1_2 | Tyrosine-protein kinase syk | 56 | 0.64 | 0.50 | 0.00 | 1.14 |
| 3I0R_RT3_1_1 | Reverse transcriptase/ribonuclease h | 57 | 0.60 | 0.55 | 0.00 | 1.14 |
| 4EH4_0OL_2_1 | Mitogen-activated protein kinase 14 | 58 | 0.48 | 0.67 | 0.00 | 1.14 |
| 1OUK_084_1_3 | Mitogen-activated protein kinase 14 | 59 | 0.70 | 0.44 | 0.00 | 1.14 |
| 4PH9_IBP_1_1 | Prostaglandin g/h synthase 2 | 60 | 0.53 | 0.62 | 0.00 | 1.14 |
| 2UZT_SS3_1_2 | Camp-dependent protein kinase | 61 | 0.71 | 0.43 | 0.00 | 1.14 |
| 2RTP_IMI_1_1 | Streptavidin | 62 | 0.51 | 0.46 | 0.33 | 1.14 |
| 4ANQ_VGH_1_2 | Alk tyrosine kinase receptor | 63 | 0.69 | 0.44 | 0.00 | 1.14 |
| 2Q2Y_MKR_2_1 | Kinesin-like protein kif11 | 64 | 0.52 | 0.62 | 0.00 | 1.14 |
| 1BDB_NAD_1_3 | Cis-biphenyl-2 | 65 | 0.51 | 0.50 | 0.25 | 1.14 |
| 4IUI_1GQ_1_2 | Estrogen receptor | 66 | 0.49 | 0.64 | 0.00 | 1.14 |
| 3SRS_M23_1_2 | Dihydrofolate reductase | 67 | 0.60 | 0.54 | 0.00 | 1.14 |
| 4OJB_198_1_1 | Androgen receptor | 68 | 0.64 | 0.50 | 0.00 | 1.14 |
| 3IW2_EKO_1_1 | XAA-PRO Dipeptidase | 69 | 0.55 | 0.58 | 0.00 | 1.13 |
| 3IW7_IPK_1_1 | Mitogen-activated protein kinase 14 | 70 | 0.70 | 0.43 | 0.00 | 1.13 |
| 4KQK_PCR_1_1 | Nicotinate-nucleotide--dimethylbenzimidazole phosphoribosyltransferase | 71 | 0.56 | 0.57 | 0.00 | 1.13 |
| 1PMU_9HP_1_1 | Mitogen-activated protein kinase 10 | 72 | 0.67 | 0.46 | 0.00 | 1.13 |
| 3NC2_QUZ_1_1 | Ketohexokinase | 73 | 0.53 | 0.60 | 0.00 | 1.13 |
| 4GE7_0K5_1_1 | Kynurenine/alpha-aminoadipate aminotransferase | 74 | 0.53 | 0.43 | 0.33 | 1.13 |
| 2I0V_6C3_1_1 | Cfms tyrosine kinase | 75 | 0.50 | 0.63 | 0.00 | 1.13 |
| 4OLM_198_1_1 | Androgen receptor | 76 | 0.57 | 0.56 | 0.00 | 1.13 |
| 3ZSI_52P_1_1 | Mitogen-activated protein kinase 14 | 77 | 0.63 | 0.50 | 0.00 | 1.13 |
| 4IVY_1GT_1_1 | Estrogen receptor | 78 | 0.62 | 0.50 | 0.00 | 1.12 |
| 3F8C_HT1_1_3 | Transcriptional regulator | 79 | 0.65 | 0.31 | 0.33 | 1.12 |
| 4ERF_0R3_1_3 | E3 ubiquitin-protein ligase mdm2 | 80 | 0.62 | 0.50 | 0.00 | 1.12 |
| 3PVW_QRX_1_1 | Beta-adrenergic receptor kinase 1 | 81 | 0.49 | 0.46 | 0.33 | 1.12 |
| 2ITP_AEE_1_1 | Epidermal growth factor receptor precursor | 82 | 0.62 | 0.33 | 0.33 | 1.12 |
| 3BQR_4RB_1_1 | Death-associated protein kinase 3 | 83 | 0.72 | 0.40 | 0.00 | 1.12 |
| 4CFK_LY2_1_1 | Brd4 protein | 84 | 0.55 | 0.40 | 0.33 | 1.12 |
| 2YFE_YFE_2_1 | Peroxisome proliferator-activated receptor gamma | 85 | 0.48 | 0.64 | 0.00 | 1.12 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 3KDT_7HA_2_3 | Peroxisome proliferator-activated receptor alpha | 86 | 0.59 | 0.36 | 0.33 | 1.12 |
| 3ZLS_92P_1_1 | Dual specificity mitogen-activated protein kinase kinase 1 | 87 | 0.58 | 0.54 | 0.00 | 1.11 |
| 2QHN_582_1_1 | Serine/threonine-protein kinase chk1 | 88 | 0.48 | 0.47 | 0.33 | 1.11 |
| 4EOS_1RO_1_3 | Cyclin-dependent kinase 2 | 89 | 0.64 | 0.31 | 0.33 | 1.11 |
| 3KPK_FAD_1_3 | Sulfide-quinone reductase | 90 | 0.54 | 0.40 | 0.33 | 1.11 |
| 3C5U_P41_2_1 | Mitogen-activated protein kinase 14 | 91 | 0.57 | 0.54 | 0.00 | 1.11 |
| 2M56_CAM_1_1 | Camphor 5-monooxygenase | 92 | 0.53 | 0.58 | 0.00 | 1.11 |
| 4FJ2_NAP_3_3 | 17beta-hydroxysteroid dehydrogenase | 93 | 0.48 | 0.47 | 0.33 | 1.11 |
| 1UUM_AFI_2_2 | Dihydroorotate dehydrogenase | 94 | 0.67 | 0.44 | 0.00 | 1.11 |
| 2A4Z_BYM_1_1 | Phosphatidylinositol-4 | 95 | 0.55 | 0.56 | 0.00 | 1.11 |
| 2X2K_X2K_1_1 | Proto-oncogene tyrosine-protein kinase receptor ret | 96 | 0.54 | 0.57 | 0.00 | 1.11 |
| 2YIS_I46_2_1 | Mitogen-activated protein kinase 14 | 97 | 0.69 | 0.42 | 0.00 | 1.11 |
| 1C0T_BM1_1_1 | HIV-1 reverse transcriptase | 98 | 0.60 | 0.50 | 0.00 | 1.10 |
| 3Q95_ESL_1_1 | Estrogen receptor | 99 | 0.57 | 0.53 | 0.00 | 1.10 |
| 3L8S_BFF_1_2 | Mitogen-activated protein kinase 14 | 100 | 0.60 | 0.50 | 0.00 | 1.10 |

[a] Fragment name (PDB_HET_C_M) is inferred from the cognate target PDB identifier (PDB), the corresponding ligand chemical component (HET), the target cavity identifier (C), and the fragment number (N).

[b] cavity similarity score, computed by ProCare, between the fragment-bound subpocket and the query target cavity

[c] Interaction fingerprint similarity, computed with IChem, between the subpocket-fragment interaction fingerprint and the query target-ligand interaction fingerprint

[d] Interaction fingerprint similarity (polar interactions only), computed with IChem, between the subpocket-fragment interaction fingerprint and the query target-ligand interaction fingerprint

[e] FragScore = ProCare + IFP + 0.5*(IFP_polar)

**Table S9.** Fragment hits for the cysteinyl leukotriene receptor 2 (PDB ID 6RZ8)

| FragID[a] | Protein name | Rank | ProCaRe[b] | IFP[c] | IFP_polar[d] | FragScore[e] |
|---|---|---|---|---|---|---|
| 1RUO_CMP_1_1 | Catabolite gene activator protein | 1 | 0.55 | 0.38 | 1.00 | 1.43 |
| 3F9M_MRK_1_1 | Glucokinase | 2 | 0.57 | 0.56 | 0.50 | 1.38 |
| 2XBJ_XBJ_1_2 | Serine/threonine-protein kinase chk2 | 3 | 0.64 | 0.67 | 0.00 | 1.30 |
| 4FCQ_2N6_1_1 | Heat shock protein hsp 90-alpha | 4 | 0.52 | 0.78 | 0.00 | 1.30 |
| 2RHT_C1E_1_1 | 2-hydroxy-6-oxo-6-phenylhexa-2 | 5 | 0.51 | 0.27 | 1.00 | 1.29 |
| 3UIV_308_1_1 | None | 6 | 0.61 | 0.64 | 0.00 | 1.25 |
| 3QCH_NX2_1_2 | Receptor-type tyrosine-protein phosphatase gamma | 7 | 0.52 | 0.73 | 0.00 | 1.25 |
| 1YW2_PGJ_1_2 | Mitogen-activated protein kinase 14 | 8 | 0.67 | 0.55 | 0.00 | 1.22 |
| 3MTF_A3F_2_1 | Activin receptor type-1 | 9 | 0.55 | 0.50 | 0.33 | 1.22 |
| 3DT1_P40_1_3 | Mitogen-activated protein kinase 14 | 10 | 0.57 | 0.64 | 0.00 | 1.21 |
| 2ZB3_NDP_1_3 | Prostaglandin reductase 2 | 11 | 0.49 | 0.32 | 0.80 | 1.20 |
| 1YC3_4BC_1_3 | Heat shock protein hsp 90-alpha | 12 | 0.54 | 0.42 | 0.50 | 1.20 |
| 2Q2Y_MKR_2_1 | Kinesin-like protein kif11 | 13 | 0.47 | 0.71 | 0.00 | 1.19 |
| 1MX5_HTQ_3_1 | None | 14 | 0.58 | 0.36 | 0.50 | 1.19 |
| 3FL9_TOP_2_2 | Dihydrofolate reductase (dhfr) | 15 | 0.53 | 0.40 | 0.50 | 1.18 |
| 4Z35_ON7_1_1 | Lysophosphatidic acid receptor 1 | 16 | 0.51 | 0.67 | 0.00 | 1.18 |
| 1E06_IPB_2_1 | None | 17 | 0.51 | 0.67 | 0.00 | 1.17 |
| 1HPZ_AAP_1_2 | Pol polyprotein | 18 | 0.47 | 0.70 | 0.00 | 1.17 |
| 4MF1_29Y_1_2 | Tyrosine-protein kinase itk/tsk | 19 | 0.60 | 0.57 | 0.00 | 1.17 |
| 3CW9_01A_2_3 | 4-chlorobenzoyl coa ligase | 20 | 0.52 | 0.47 | 0.33 | 1.16 |
| 3VRY_B43_1_3 | Tyrosine-protein kinase hck | 21 | 0.62 | 0.54 | 0.00 | 1.15 |
| 1Q43_CMP_1_1 | Potassium/sodium hyperpolarization-activated cyclic nucleotide-gated channel 2 | 22 | 0.47 | 0.43 | 0.50 | 1.15 |
| 1EET_BFU_1_2 | Hiv-1 reverse transcriptase | 23 | 0.48 | 0.67 | 0.00 | 1.15 |
| 3TKU_M77_1_1 | Serine/threonine-protein kinase mrck beta | 24 | 0.48 | 0.41 | 0.50 | 1.15 |
| 2C3I_IYZ_1_3 | Pimtide | 25 | 0.54 | 0.60 | 0.00 | 1.14 |
| 2YI5_YI5_1_2 | Heat shock protein hsp 90-alpha | 26 | 0.49 | 0.40 | 0.50 | 1.14 |
| 4IWQ_1FV_1_2 | Serine/threonine-protein kinase tbk1 | 27 | 0.52 | 0.62 | 0.00 | 1.14 |
| 1RD4_L08_1_2 | Integrin alpha-l | 28 | 0.51 | 0.62 | 0.00 | 1.13 |
| 3IW7_IPK_1_1 | Mitogen-activated protein kinase 14 | 29 | 0.63 | 0.50 | 0.00 | 1.13 |
| 3L8S_BFF_1_2 | Mitogen-activated protein kinase 14 | 30 | 0.59 | 0.54 | 0.00 | 1.13 |
| 3ULE_C69_1_1 | Actin-related protein 3 | 31 | 0.56 | 0.44 | 0.25 | 1.12 |
| 3OAF_OAG_1_1 | Dihydrofolate reductase | 32 | 0.53 | 0.58 | 0.00 | 1.12 |
| 4LGH_0JN_2_2 | Proto-oncogene tyrosine-protein kinase src | 33 | 0.62 | 0.50 | 0.00 | 1.12 |
| 3HQ5_GKK_1_2 | None | 34 | 0.57 | 0.55 | 0.00 | 1.12 |
| 3CX5_SMA_2_1 | Cytochrome b-c1 complex subunit 1 | 35 | 0.55 | 0.56 | 0.00 | 1.10 |
| 3ZSG_T75_1_3 | Mitogen-activated protein kinase 14 | 36 | 0.52 | 0.58 | 0.00 | 1.10 |
| 2IOK_IOK_1_3 | None | 37 | 0.60 | 0.50 | 0.00 | 1.10 |
| 3MSS_MS7_4_1 | None | 38 | 0.54 | 0.56 | 0.00 | 1.10 |
| 4LH7_1X8_1_1 | Dna ligase | 39 | 0.51 | 0.33 | 0.50 | 1.09 |
| 2RTF_BTN_1_1 | Streptavidin | 40 | 0.48 | 0.62 | 0.00 | 1.09 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 3N4M_CMP_1_1 | Catabolite gene activator | 41 | 0.51 | 0.42 | 0.33 | 1.09 |
| 3T94_MTA_4_1 | 5'-methylthioadenosine phosphorylase (mtap) | 42 | 0.62 | 0.46 | 0.00 | 1.08 |
| 4L4B_CAM_1_1 | Camphor 5-monooxygenase | 43 | 0.48 | 0.60 | 0.00 | 1.08 |
| 3GL2_D3M_2_1 | Ddmc | 44 | 0.51 | 0.57 | 0.00 | 1.08 |
| 4NG5_PFB_4_1 | Alcohol dehydrogenase e chain | 45 | 0.58 | 0.50 | 0.00 | 1.08 |
| 3OZU_X89_1_3 | None | 46 | 0.54 | 0.54 | 0.00 | 1.08 |
| 3VS4_VSF_1_3 | Tyrosine-protein kinase hck | 47 | 0.51 | 0.57 | 0.00 | 1.08 |
| 3GC7_B45_1_1 | Mitogen-activated protein kinase 14 | 48 | 0.54 | 0.55 | 0.00 | 1.08 |
| 4H38_0YX_1_2 | Undecaprenyl pyrophosphate synthase | 49 | 0.62 | 0.45 | 0.00 | 1.08 |
| 4F4P_0SB_1_2 | Tyrosine-protein kinase syk | 50 | 0.49 | 0.58 | 0.00 | 1.08 |
| 1VRU_AAP_1_1 | Hiv-1 reverse transcriptase | 51 | 0.62 | 0.45 | 0.00 | 1.07 |
| 2EXC_JNK_1_1 | Mitogen-activated protein kinase 10 | 52 | 0.54 | 0.53 | 0.00 | 1.07 |
| 4Z34_ON7_1_2 | Lysophosphatidic acid receptor 1 | 53 | 0.53 | 0.55 | 0.00 | 1.07 |
| 1ZUC_T98_1_1 | Progesterone receptor | 54 | 0.64 | 0.43 | 0.00 | 1.07 |
| 4C66_H4C_1_2 | Bromodomain-containing protein 4 | 55 | 0.57 | 0.50 | 0.00 | 1.07 |
| 4OTY_LUR_2_1 | Prostaglandin g/h synthase 2 | 56 | 0.53 | 0.53 | 0.00 | 1.07 |
| 3RUK_AER_3_2 | Steroid 17-alpha-hydroxylase/17 | 57 | 0.56 | 0.50 | 0.00 | 1.06 |
| 1CR6_CPU_1_2 | None | 58 | 0.47 | 0.42 | 0.33 | 1.06 |
| 4G27_PHU_1_1 | None | 59 | 0.60 | 0.45 | 0.00 | 1.06 |
| 2ZDT_46C_1_1 | Mitogen-activated protein kinase 10 | 60 | 0.52 | 0.54 | 0.00 | 1.06 |
| 4LGG_VGG_1_2 | Proto-oncogene tyrosine-protein kinase src | 61 | 0.55 | 0.50 | 0.00 | 1.05 |
| 2G76_NAD_1_3 | D-3-phosphoglycerate dehydrogenase | 62 | 0.49 | 0.23 | 0.67 | 1.05 |
| 3SRS_M23_1_2 | Dihydrofolate reductase | 63 | 0.60 | 0.45 | 0.00 | 1.05 |
| 1VRT_NVP_1_2 | None | 64 | 0.59 | 0.46 | 0.00 | 1.05 |
| 1JHV_PCR_1_1 | None | 65 | 0.60 | 0.45 | 0.00 | 1.05 |
| 1IKY_MSD_1_1 | Pol polyprotein | 66 | 0.55 | 0.50 | 0.00 | 1.05 |
| 3V66_D3A_1_1 | None | 67 | 0.55 | 0.50 | 0.00 | 1.05 |
| 2XAE_2XA_3_3 | Kinesin-like protein kif11 | 68 | 0.59 | 0.46 | 0.00 | 1.05 |
| 2UZT_SS3_1_1 | Camp-dependent protein kinase | 69 | 0.56 | 0.23 | 0.50 | 1.04 |
| 4F84_SAM_1_1 | Geranyl diphosphate 2-c-methyltransferase | 70 | 0.50 | 0.29 | 0.50 | 1.04 |
| 4BIE_IE6_1_2 | Mitogen-activated protein kinase kinase kinase 5 | 71 | 0.63 | 0.42 | 0.00 | 1.04 |
| 2J7Y_E3O_1_1 | Estrogen receptor beta | 72 | 0.51 | 0.53 | 0.00 | 1.04 |
| 2XYX_Z00_1_2 | None | 73 | 0.68 | 0.36 | 0.00 | 1.04 |
| 5KCP_PFB_2_1 | Alcohol dehydrogenase e chain | 74 | 0.50 | 0.54 | 0.00 | 1.04 |
| 2ZB1_GK4_1_2 | Mitogen-activated protein kinase 14 | 75 | 0.54 | 0.50 | 0.00 | 1.04 |
| 4G2I_0VQ_1_1 | Vitamin d3 receptor | 76 | 0.54 | 0.50 | 0.00 | 1.04 |
| 3TQ9_MTX_1_2 | Dihydrofolate reductase | 77 | 0.57 | 0.46 | 0.00 | 1.04 |
| 3CD2_MTX_1_2 | Dihydrofolate reductase | 78 | 0.66 | 0.38 | 0.00 | 1.04 |
| 2QBM_CAM_1_1 | Cytochrome p450-cam | 79 | 0.50 | 0.54 | 0.00 | 1.04 |
| 4O1Y_NLA_1_1 | None | 80 | 0.50 | 0.54 | 0.00 | 1.04 |
| 5DP2_NAP_1_3 | Curf | 81 | 0.48 | 0.35 | 0.40 | 1.04 |
| 1GUF_NDP_1_2 | Enoyl-[acyl-carrier-protein] reductase [nadph | 82 | 0.53 | 0.17 | 0.67 | 1.03 |
| 1J3J_CP6_1_2 | Bifunctional dihydrofolate reductase-thymidylate synthase | 83 | 0.53 | 0.50 | 0.00 | 1.03 |

| | | | | | |
|---|---|---|---|---|---|
| 4HW7_64M_1_3 | Macrophage colony-stimulating factor 1 receptor | 84 | 0.59 | 0.44 | 0.00 | 1.03 |
| 3DXM_N24_1_1 | Actin-related protein 3 | 85 | 0.62 | 0.42 | 0.00 | 1.03 |
| 2AYR_L4G_1_2 | Estrogen receptor | 86 | 0.58 | 0.45 | 0.00 | 1.03 |
| 2ZM4_KSM_1_2 | Proto-oncogene tyrosine-protein kinase lck | 87 | 0.57 | 0.45 | 0.00 | 1.03 |
| 2IZI_BTN_1_1 | Streptavidin | 88 | 0.53 | 0.50 | 0.00 | 1.03 |
| 1I7I_AZ2_1_1 | Peroxisome proliferator activated receptor gamma | 89 | 0.53 | 0.50 | 0.00 | 1.03 |
| 1LW0_NVP_1_2 | None | 90 | 0.49 | 0.54 | 0.00 | 1.03 |
| 1C1C_612_1_2 | Hiv-1 reverse transcriptase (a-chain) | 91 | 0.52 | 0.50 | 0.00 | 1.02 |
| 3EEL_53T_2_3 | Dihydrofolate reductase | 92 | 0.57 | 0.45 | 0.00 | 1.02 |
| 3SR5_Q12_1_2 | Dihydrofolate reductase | 93 | 0.52 | 0.50 | 0.00 | 1.02 |
| 3Q2A_PAB_2_1 | None | 94 | 0.57 | 0.44 | 0.00 | 1.02 |
| 4BBE_3O4_2_2 | Tyrosine-protein kinase jak2 | 95 | 0.52 | 0.50 | 0.00 | 1.02 |
| 3EWK_FAD_1_3 | Sensor protein | 96 | 0.50 | 0.35 | 0.33 | 1.02 |
| 3W16_P9J_1_1 | Aurora kinase a | 97 | 0.55 | 0.47 | 0.00 | 1.01 |
| 2CF6_NAP_1_3 | Cinnamyl alcohol dehydrogenase | 98 | 0.52 | 0.29 | 0.40 | 1.01 |
| 4FAK_SAM_1_1 | Ribosomal rna large subunit methyltransferase h | 99 | 0.47 | 0.54 | 0.00 | 1.01 |
| 4MEO_25V_1_1 | Bromodomain-containing protein 4 | 100 | 0.60 | 0.42 | 0.00 | 1.01 |

[a] Fragment name (PDB_HET_C_M) is inferred from the cognate target PDB identifier (PDB), the corresponding ligand chemical component (HET), the target cavity identifier (C), and the fragment number (N).

[b] cavity similarity score, computed by ProCare, between the fragment-bound subpocket and the query target cavity

[c] Interaction fingerprint similarity, computed with IChem, between the subpocket-fragment interaction fingerprint and the query target-ligand interaction fingerprint

[d] Interaction fingerprint similarity (polar interactions only), computed with IChem, between the subpocket-fragment interaction fingerprint and the query target-ligand interaction fingerprint

[e] FragScore = ProCare + IFP + 0.5*(IFP_polar)

# 2.4. Critical evaluation of ProCare

## 2.4.1. ProCare algorithm

Several implementations of ProCare were attempted to improve the method, although incremental. We quickly remind the comparison procedure to serve a basis for discussion here: (1) N (with N > 2) points are randomly sampled in pocket #1 and associated with their nearest neighbor in pockets #2 according to the Euclidian distance their descriptors; (2) conservation of pairwise distances between all points in #1 versus #2 is checked (topological verification); (3) an initial alignment is estimated on the N pairs of points, (4) the alignment is refined with ICP and (5) the final alignment is scored.

First, it was intriguing that when optimizing the set of alignment parameters, we found that sampling N=4 points was yielding better alignment and discrimination, compared to sampling three and five points. This value is consistent with what Open3D authors experienced on their image inputs. Our hypothesis is that although sampling three points is sufficient to estimate a transformation, it is more permissive and yields to false-positive topological verification. Contrarily, comparing five points would impose more constraints, so that the topological verification is harder to pass. In this sense, we implemented two variants to study this effect and avoid the non-deterministic aspect of the algorithm. In the first variant, all the points in pocket #1 are sampled simultaneously. This variant was unsuccessful unless identical pockets are compared, therefore useless. In the second variant, the set of equivalent points is progressively increased by adding a pair of points that satisfies the topological verification of the set. This variant was successful only for very similar pockets (e.g. different PDB structures of the same protein), therefore unapplicable for detecting remote similarities. These studies shed light on the importance of the initial correspondences.

Since points are associated to their nearest neighbor in the descriptor space, a point is always associated to another, even if the similarity of the descriptors is meaningless. Applying a distance cutoff is not a systematic solution and is prone to be dependent on the dataset. In a new version where the sampled points are tracked, we observed that the distance ranges leading to a good alignment is hardly distinguishable from the distance ranges leading to a bad alignment (**Figure 2.2**).
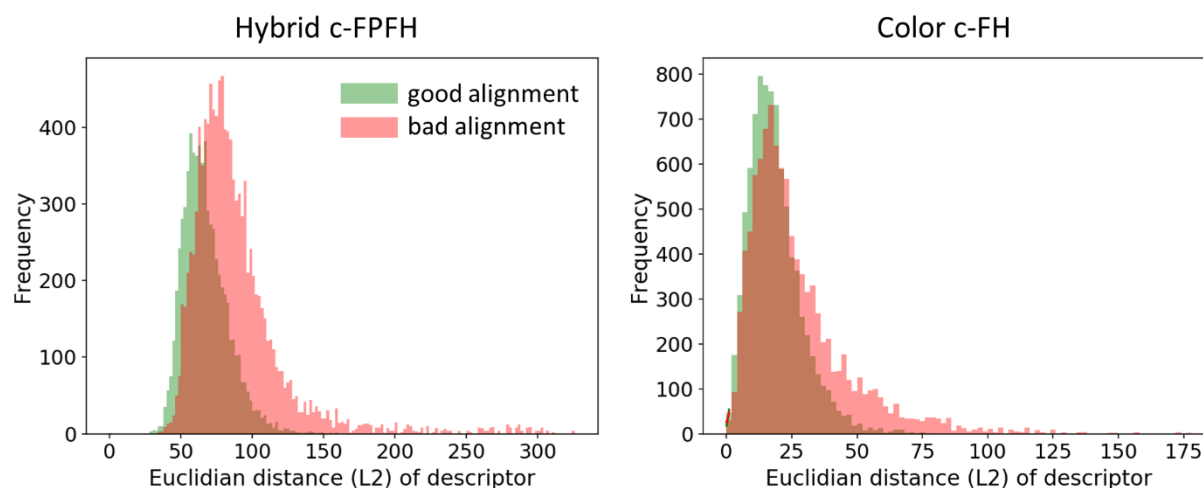
**Figure 2.2.** Euclidian distance between descriptors of RANSAC-sampled and topologically valid points (N = 4 pairs) leading to good and bad alignments. ~31,000 sc-PDB subpockets were translated and realigned on their corresponding pockets. The transformation matrix was applied to the co-crystal ligand and the RMSD between the original position and the new position after alignment is reported. A good alignment refers to RMSD ≤ 0.5 Å, a bad alignment to RMSD ≥ 4 Å. Alignments were proposed using two sets of pocket descriptors (c-FPFH and c-FH).

Analysis of the RANSAC-equivalenced pairs showed that sometimes, redundant pairs of points are sampled whereas a proper rotation requires three different pairs. Not surprisingly, sampling less than three different pairs led to more misalignments (**Figure 2.3**). As a result, redundancy of the correspondences during the procedure should be used as quality filter to decrease the chances of misalignment.
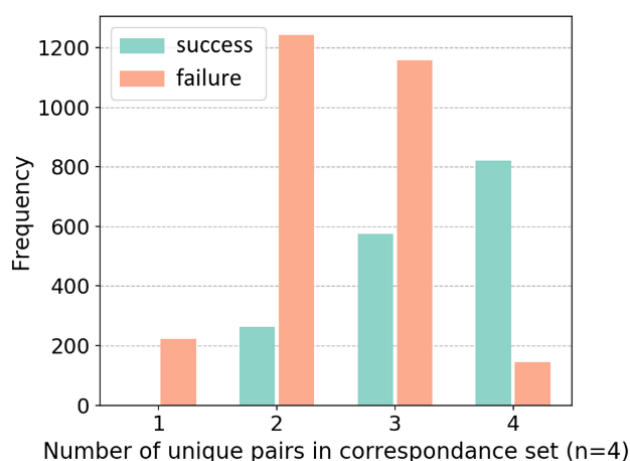


**Figure 2.3.** RANSAC correspondences used for transformation estimation. Sampling four different pairs increase the chances of a good alignment. sc-PDB subpockets were translated into different coordinate frames and realigned to their corresponding pockets. A good alignment (success) refers to RMSD ≤ 0.5 Å with respect to cocrystal coordinates, a bad alignment (failure) to RMSD ≥ 4 Å.

The grid-based arrangement of points implies a fixed number of possible pairwise distances. Therefore, it is also possible that random matches verify topological constraints and serve as wrong initial alignment. Our hypothesis is that points occupying the core of the cavity are hard to differentiate due the regular repartition of neighbor points around them. Possible improvements pertain to the metric (e.g. using L1 distance), optimizing the weights of the shape and color bins in the descriptor, evaluating other descriptors. In this regard, later applications showed that the color part of the c-FPFH descriptor (c-FH), encoding the relative distribution of pharmacophoric features around each point, showed equivalent discrimination performance as c-FPFH. Interestingly, c-FH alignments tend to be more refined than c-FPFH alignments, when comparing pocket to pocket or subpocket to pocket. In future prospective applications, a 'divide and conquer' mode is possible, by performing shape-only, color-only and hybrid descriptor-based alignments.

VolSite cavity descriptions are noisy with respect to pharmacophoric annotation. Statistics on the sc-PDB revealed that the hydrophobic points (CA) are present in a large proportion (ca. 40%), compared to the other pharmacophoric features.[1] Thus, it was not surprising that they also contribute more to the proposed alignments and might erroneously increase the similarity score. However, not considering the CA feature is not applicable for highly hydrophobic pockets and generally led to poor discrimination. The same conclusions were derived for the dummy (DU) feature. Contrarily, some features such as negatively ionized OD1 are rare (ca. 5% of all annotations). Given that only one pharmacophoric feature is assigned to a point, a residue might be present in the site, yet not represented in the cavity cloud if a different residue is closest to the point. For example, this was observed in the hinge area of some protein kinase structures. Some features clusters in patches, others are isolated—but important points.

Scoring is the final step of the comparison. At that stage, it is not possible to rescue an alignment solution that has not been previously explored. The scoring scheme should be robust enough to discriminate relevant from noisy similarity estimates. Several scoring schemes were evaluated, some of them are alignment-free. We showed that pairwise comparison of point descriptors in the two pockets can discriminate similar from dissimilar pairs in the BO1 dataset (**Figure 2.4**) and can be used as an additional filter.
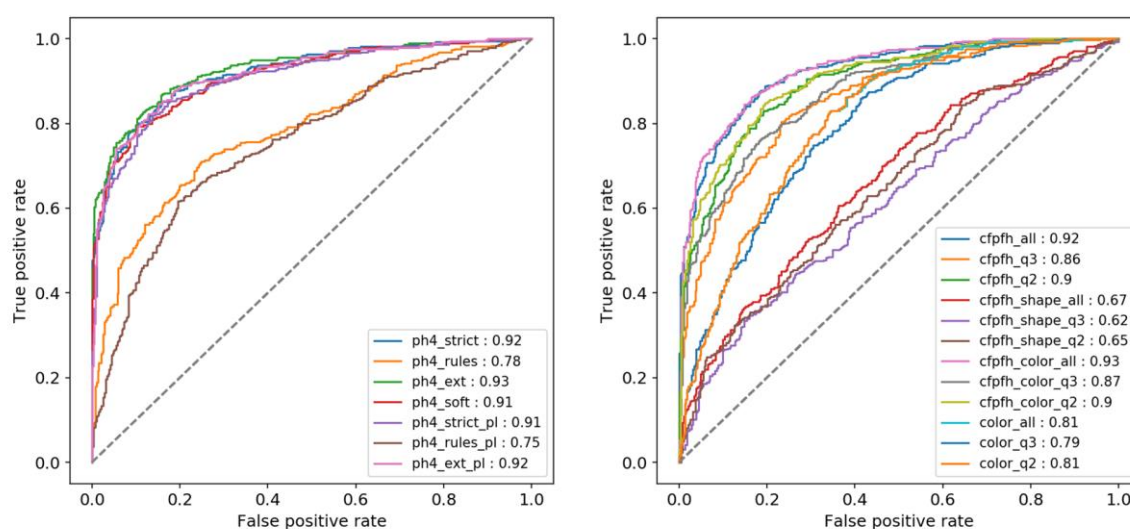
**Figure 2.4.** Scoring scheme optimization. The receiver operating characteristics curve of (left) alignment-based and (right) alignment-free scoring on the BO1 dataset. The ph4-strict, ph4-rules and ph4-ext were previously described. The ph4-soft is the ph4-strict without distance cutoff. ph4-strict_pl, ph4-ext_pl, ph4-ext_pl are the piece-wise linear implementation of their counterparts (intervals are below 0.75 Å, between 0.75 and 1.5 Å, beyond 1.5 Å). Alignment-free scoring are the mean pairwise points descriptor distances in the compared pockets, with the idea that similar pockets would share more similar points in the descriptor space, lowering the average distance; 'all', 'q2' and 'q3' denote the use of all, above median and above third quartile distances.

In future developments to rescue wrong initializations, we suggest the generation of multiple alignment solutions during the sampling and the use of a pharmacophoric scoring as a convergence criterion instead of current color-agnostic fitness score.

## 2.4.2. Sensitivity to protein fold and coordinate deviations

Finding the right balance between detection of subtle changes in a cavity while enabling remote similarity detection is one of the challenges to binding site comparison tools.

The dependency of ProCare to the protein structure/fold has been assessed on the radical SAM superfamily (RSS) of proteins, described by Holliday *et al.*[33] This family of proteins covers 63785 different sequences, 1500 protein architectures, and 150 folds, all of them having converged to form a catalytic site using S-adenosylmethionine (SAM) in a radical enzymatic mechanism. The RSS dataset

used here is composed of 15 representative proteins of known X-ray structures describing nine different classes varying in folds and catalyzing different enzymatic reactions. Pairwise comparison of SAM binding cavities was achieved with the current method and compared to that obtained with 6 other cavity comparison tools (FuzCav, KRIPO, PocketMatch, ProBiS, Shaper, SiteAlign) representative of the current state of a recent review from an independent group.[34] Seven RSS subgroups (L1, L2, L11, L13, L15, L16, L19) are represented by a single protein structure whereas two subgroups (L6 and L17) are described by five and three different proteins, respectively. Using default parameters and developer-suggested thresholds for distinguishing similar from dissimilar cavities, we first derived a 15*15 cavity similarity matrix and computed the proportion of cavity pairs still considered similar by each of the investigated tool (**Figure 2.5**).
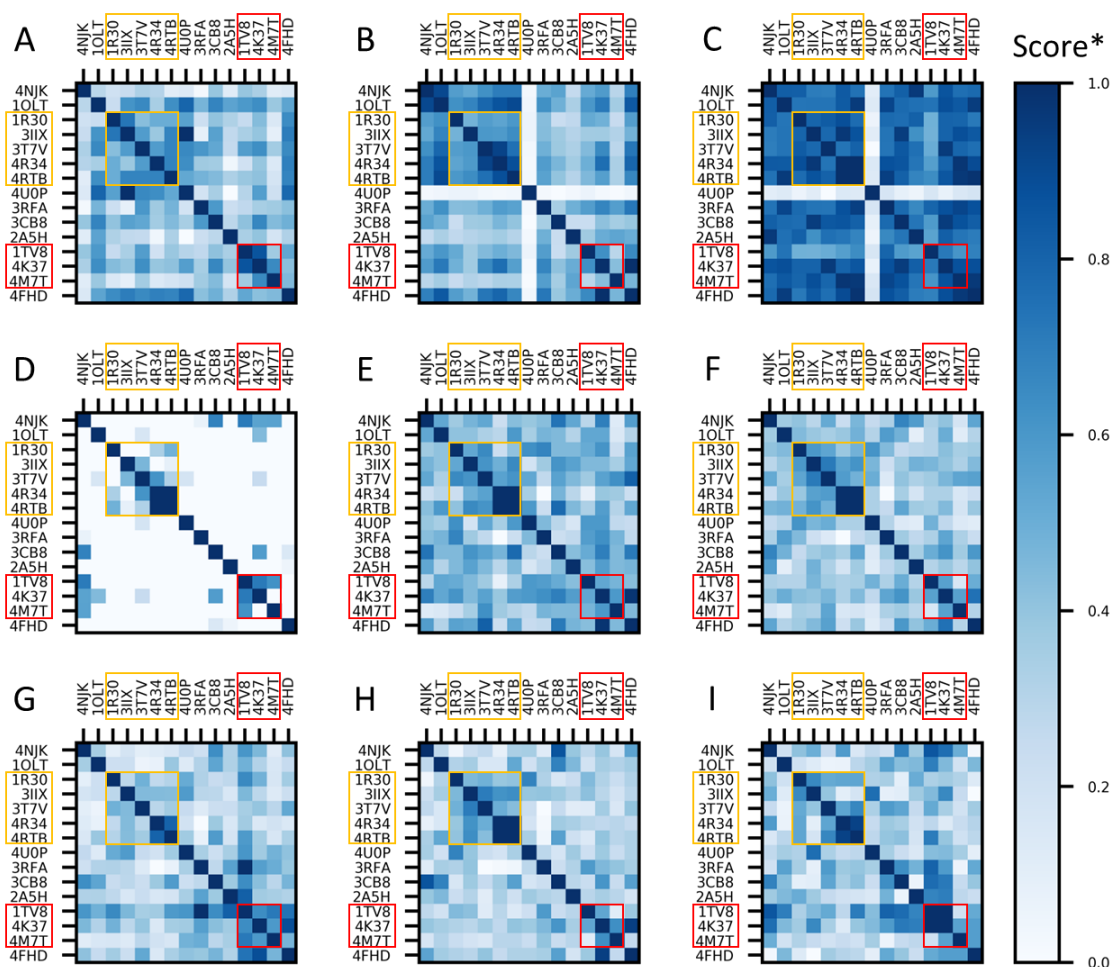


**Figure 2.5.** Pairwise binding site comparison of 15 Radical SAM Superfamily entries. Nine subgroups are represented: L1 (PDB ID: 4NJK), L2 (PDB ID: 1OLT), L6 (PDB IDs: 1R30, 3IIX, 3T7V, 4R34, 4RTB), L11 (PDB ID: 4U0P), L13 (PDB ID: 3RFA), L15 (PDB ID: 3CB8), L16 (PDB ID: 2A5H), L17 (PDB IDs: 1TV8, 4K37, 4M7T), L19 (PDB ID: 4FHD). Score* is a normalized score: score* = (score_method -min_score_method) / (max_score_method - min_score_method). Self-comparisons (diagonal of the matrix) were automatically assigned a maximum score of 1. L6 and L17 subgroups are

encircled in yellow and red respectively. A) FuzCav. B) KRIPO. C) PocketMatch. D) ProBiS; alignment_score was used when z-score is higher than the default 1 or was set to 0 when no alignment was produced. E) ProCare with cavity detected at 4 Å. F) ProCare with cavity detected at 6 Å. G) Shaper with cavity detected at 4 Å. H) Shaper with cavity detected at 6 Å. I) SiteAlign; distances d were normalized to d' and converted into a similarity score 1-d'.

Obtained results are hardly interpretable because very much dependent of pocket definition and threshold values to estimate pairwise similarities. ProCare estimates that 55% of all pairs are still similar despite the very different protein folds and structures, a proportion higher than that obtained by three tools (FuzCav, ProBiS, SiteAlign), almost similar to KRIPO (63%), but lower than the performance reached by the two best tools (Shaper, PocketMatch; 91% for both methods). The latter two tools outperforming ProCare in this benchmarking exercise might however be too promiscuous and not specific enough. We then examined whether all compared cavity comparison tools were equally able to predict higher similarity values for intra-class than for inter-class comparisons.

Indeed, some tools are not well suited for finer comparisons. On the one hand, PocketMatch (C) is not specific enough to discriminate among RSS classes. On the other hand, ProBiS (D) fails in detecting inter-class pocket similarities. KRIPO (B), although partially clustering entries for L6 and L17 subgroups did not succeed in finding any similarity between one entry (4U0P) and the 14 others. Altogether, ProCare (E, F) as well as two other tools (FuzCav (A), SiteAlign (I)) provide the best compromise between selectivity and precision. It affords high similarity values throughout the matrix but enables a clear distinction of the two subclasses represented by more than one entry. As to be expected, pocket definition (size of the binding site) has a clear impact on the heat maps produced by a single tool. Since this definition varies from a method to another one and cannot always be homogenized, a truly unbiased comparison of all methods presented here remains difficult, notably for this dataset for which no experimental data can support (or not) the predicted similarity estimation.

To be robust, methods need to be insensitive to variations in atomic coordinates of the pocket, frequently observed upon ligand binding and experimental details of the structural determination method (e.g. X-ray diffraction, single-particle cryo-electron microscopy, homology modeling). We therefore designed two data sets (MD-PLA2, Holo-Apo) to assess ProCare robustness to align and score identical cavities exhibiting small to large variations in atomic coordinates. In the first set (MD-PLA2), the phospholipase A2-atropine complex was subjected to a 10 ns molecular dynamics (MD) simulation in explicit water, and 1000 MD snapshots of the atropine-bound cavity were retained for pairwise similarity calculations. The second set (Holo-Apo) is composed of 10 pairs of pockets in a ligand-bound (holo) and ligand-free (apo) form, showing from small (rmsd < 1.0 Å) to large (rmsd > 4.0 Å) variations in the atomic coordinates of cavity-lining heavy atoms.

For both sets, ProCare still detected cavity similarity up to variations in atomic coordinates located in a grey zone around 2.5-3.0 Å RMSD of heavy atoms (**Figure 2.6**), which is in line with the usually admitted 2.0 Å RMSD in posing ligands by molecular docking.
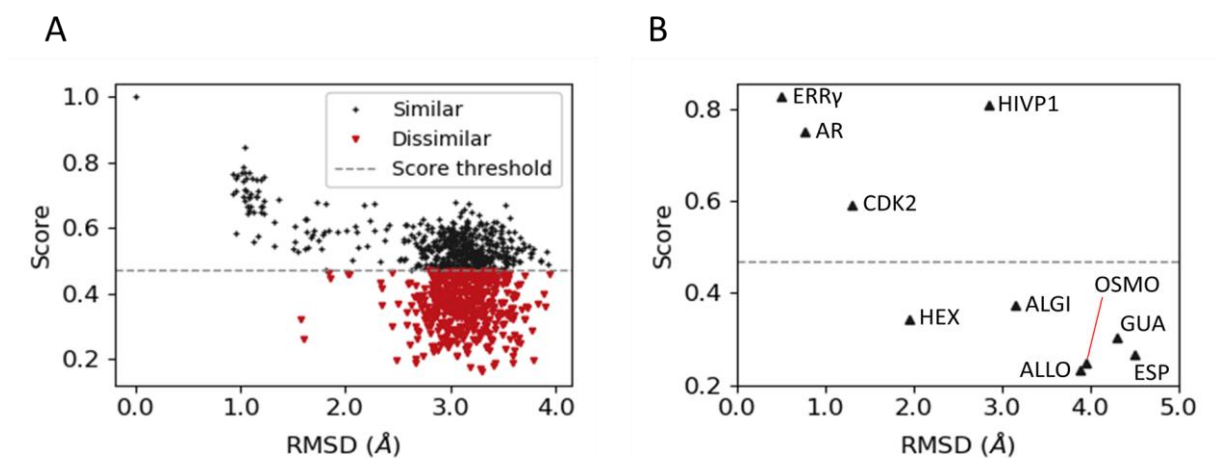


**Figure 2.6.** Sensitivity of the ProCare score to variations in atomic coordinates. A) Atomic coordinates variations of the pocket (RMSD on heavy atoms to the first snapshot), induced by molecular dynamics simulation of phospholipase A2 in complex with atropine (PDB ID: 2ARM). A score of 0.47 (dotted line) corresponds a statistically significant threshold (p-value = 0.05) to discriminate similar from dissimilar cavities; B) Sensitivity of the ProCare score to ligand-induced variations in atomic coordinates of pockets (RMSD on heavy atoms) of the Holo-Apo set (cell division protein kinase 2, CDK2, PDB IDs: 1DM2, 2JGZ; HIV-1 protease, HIVP1, PDB IDs: 1QBS, 1HHP; estrogen-related receptor gamma, ERRγ, PDB IDs: 2ZKC, 2ZBS; aldose reductase, AR, PDB IDs: 1ADS, 2NVD; hexokinase, Hexo, PDB IDs: 2E2O, 2E2N; alginate-binding protein, ALGI, PDB IDs:1Y3N, 1Y3Q; Osmo-protection protein, OSMO, PDB IDs: 1SW2, 1SW5; D-allose binding protein, ALLO, PDB IDs:1RPJ, 1GUD; guanylate kinase, GUA, PDB IDs: 1EX7, 1EX6; 5-enolpyruvylshikimate-3-phosphate synthase, ESP, PDB IDs:1RF4, 1RF5). A score of 0.47 (dotted line) corresponds a statistically significant threshold (p-value = 0.05) to discriminate similar from dissimilar cavities.

## 2.4.3. Local comparisons

Local comparison of cavities is desired for unobvious similarity detection. Herein, there are three levels of definition. Firstly, local comparison denotes the specific positioning of a small pocket (subpocket) with respect to a larger pocket. Secondly, when comparing two cavities independently of their sizes, locality refers to specific partial alignment when applicable. Finally, the third level pertains the scoring scheme. ProCare allows the three levels of local comparison by local description around each point,

point-to-point correspondences, and a symmetrical scoring scheme accounting for the size of the pockets.

We however draw attention to the fact that detection of partial overlapping areas relies on the positions of sampled points. When the alignment is constrained on sampled points that are spread in large cavities, the resulting comparison can only be global. Contrarily, sampling a few clustered points would enable partial alignment when applicable.

### 2.4.4. Computing time

The ProCare algorithm can be optimized with respect to the alignment speed. ProCare was implemented based on existing package that allows multithreading. Interestingly, compilation of a non-parallelized version improved the alignment time by a factor two. This is not surprising, given the number of points treated. The alignment time is largely dominated by the number of RANSAC iterations until convergence. Implementing the different improvement proposals discussed above might yield a quicker convergence. Finally, ProCare core is available in both C++ and Python, but the execution tools were provided in Python only. Developing a full C++ tool might also speed up the comparisons.

## 2.5. References

1. Desaphy, J.; Azdimousa, K.; Kellenberger, E.; Rognan, D. Comparison and Druggability Prediction of Protein–Ligand Binding Sites from Pharmacophore-Annotated Cavity Shapes. *J. Chem. Inf. Model.* **2012**, *52*, 2287–2299.

2. Kabsch, W. A Solution for the Best Rotation to Relate Two Sets of Vectors. *Acta Crystallogr. Sect. A* **1976**, *32*, 922–923.

3. Taylor, W. R.; Orengo, C. A. Protein Structure Alignment. *J. Mol. Biol.* **1989**, *208*, 1–22.

4. Fischer, D.; Norel, R.; Wolfson, H.; Nussinov, R. Surface Motifs by a Computer Vision Technique: Searches, Detection, and Implications for Protein-Ligand Recognition. *Proteins Struct. Funct. Genet.* **1993**, *16*, 278–292.

5. Shulman-Peleg, A.; Nussinov, R.; Wolfson, H. J. Recognition of Functional Sites in Protein Structures. *J. Mol. Biol.* **2004**, *339*, 607–633.

6. Milletti, F.; Vulpetti, A.; F., M.; A., V. Predicting Polypharmacology by Binding Site Similarity: From Kinases to the Protein Universe. *J. Chem. Inf. Model.* **2010**, *50*, 1418–1431.

7. Ellingson, L.; Zhang, J. Protein Surface Matching by Combining Local and Global Geometric

Information. *PLoS One* **2012**, *7*.

8. Douguet, D.; Payan, F. Sensaas: Shape-Based Alignment by Registration of Colored Point-Based Surfaces. *Mol. Inform.* **2020**, *39*, 1–13.

9. Polychronidou, E.; Kalamaras, I.; Agathangelidis, A.; Sutton, L. A.; Yan, X. J.; Bikos, V.; Vardi, A.; Mochament, K.; Chiorazzi, N.; Belessi, C.; Rosenquist, R.; Ghia, P.; Stamatopoulos, K.; Vlamos, P.; Chailyan, A.; Overby, N.; Marcatili, P.; Hatzidimitriou, A.; Tzovaras, D. Automated Shape-Based Clustering of 3D Immunoglobulin Protein Structures in Chronic Lymphocytic Leukemia. *BMC Bioinformatics* **2018**, *19*.

10. Langenfeld, F.; Axenopoulos, A.; Benhabiles, H.; Daras, P.; Giachetti, A.; Han, X.; Hammoudi, K.; Kihara, D.; Lai, T. M.; Melkemi, M.; Mylonas, S. K.; Terashi, G.; Wang, Y.; Windal, F.; Montes, M. *SHREC ' 19 Protein Shape Retrieval Contest*; 2019.

11. Desaphy, J.; Bret, G.; Rognan, D.; Kellenberger, E. Sc-PDB: A 3D-Database of Ligandable Binding Sites-10 Years On. *Nucleic Acids Res.* **2015**, *43*, D399–D404.

12. Kellenberger, E.; Muller, P.; Schalon, C.; Bret, G.; Foata, N.; Rognan, D. Sc-PDB: An Annotated Database of Druggable Binding Sites from the Protein Data Bank. *J. Chem. Inf. Model.* **2006**, *46*, 717–727.

13. Meslamani, J.; Rognan, D.; Kellenberger, E. Sc-PDB: A Database for Identifying Variations and Multiplicity of "druggable" Binding Sites in Proteins. *Bioinformatics* **2011**, *27*, 1324–1326.

14. Huang, X.; Mei, G.; Zhang, J.; Abbas, R. A Comprehensive Survey on Point Cloud Registration. **2021**, 1–17.

15. Fahim, G.; Amin, K.; Zarif, S. Single-View 3D Reconstruction: A Survey of Deep Learning Methods. *Comput. Graph.* **2021**, *94*, 164–190.

16. Eggert, D. W.; Lorusso, A.; Fisher, R. B. Estimating 3-D Rigid Body Transformations: A Comparison of Four Major Algorithms. *Mach. Vis. Appl.* **1997**, *9*, 272–290.

17. Schönemann, P. H. A Generalized Solution of the Orthogonal Procrustes Problem. *Psychometrika* **1966**, *31*, 1–10.

18. Kabsch, W. A Discussion of the Solution for the Best Rotation to Relate Two Sets of Vectors. *Acta Crystallogr. Sect. A* **1978**, *34*, 827–828.

19. Arun, K. S.; Huang, T. S.; Blostein, S. D. Least-Squares Fitting of Two 3-D Point Sets. *IEEE Trans. Pattern Anal. Mach. Intell.* **1987**, *PAMI-9*, 698–700.

20. Umeyama, S. Least-Squares Estimation of Transformation Parameters between Two Point Patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **1991**, *13*, 376–380.

21. Horn, B. K. P.; Hilden, H. M.; Negahdaripour, S. Closed-Form Solution of Absolute Orientation Using Orthonormal Matrices. *J. Opt. Soc. Am. A* **1988**, *5*, 1127.

22. Horn, B. K. P. Closed-Form Solution of Absolute Orientation Using Unit Quaternions. *J. Opt. Soc. Am. A* **1987**, *4*, 629.

23. Walker, M. W.; Shao, L.; Volz, R. A. Estimating 3-D Location Parameters Using Dual Number Quaternions. *CVGIP Image Underst.* **1991**, *54*, 358–367.

24.     Besl, P. J.; McKay, N. D. A Method for Registration of 3-D Shapes. *IEEE Trans. Pattern Anal. Mach. Intell.* **1992**, *14*, 239–256.

25.     Chen, Y.; Medioni, G. Object Modelling by Registration of Multiple Range Images. *Image Vis. Comput.* **1992**, *10*, 145–155.

26.     Yang, J.; Li, H.; Campbell, D.; Jia, Y. Go-ICP: A Globally Optimal Solution to 3D ICP Point-Set Registration. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 2241–2254.

27.     Campbell, D.; Petersson, L. GOGMA: Globally-Optimal Gaussian Mixture Alignment. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE, **2016**; Vol. 2016-Decem, pp 5685–5694.

28.     Mellado, N.; Aiger, D.; Mitra, N. J. Super 4PCS Fast Global Pointcloud Registration via Smart Indexing. *Comput. Graph. Forum* **2014**, *33*, 205–215.

29.     Rusu, R. B.; Blodow, N.; Beetz, M. Fast Point Feature Histograms (FPFH) for 3D Registration. *2009 IEEE Int. Conf. Robot. Autom.* **2009**, 3212–3217.

30.     Chua, C. S.; Jarvis, R. Point Signatures: A New Representation for 3D Object Recognition. *Int. J. Comput. Vis.* **1997**, *25*, 63–85.

31.     Drost, B.; Ulrich, M.; Navab, N.; Ilic, S. Model Globally, Match Locally: Efficient and Robust 3D Object Recognition. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*; IEEE, **2010**; pp 998–1005.

32.     Zeng, A.; Song, S.; Niessner, M.; Fisher, M.; Xiao, J.; Funkhouser, T. 3DMatch: Learning Local Geometric Descriptors from RGB-D Reconstructions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE, **2017**; Vol. 2017-Janua, pp 199–208.

33.     Holliday, G. L.; Akiva, E.; Meng, E. C.; Brown, S. D.; Calhoun, S.; Pieper, U.; Sali, A.; Booker, S. J.; Babbitt, P. C. *Atlas of the Radical SAM Superfamily: Divergent Evolution of Function Using a "Plug and Play" Domain*, 1st ed.; Elsevier Inc., **2018**; Vol. 606.

34.     Ehrt, C.; Brinkjost, T.; Koch, O. A Benchmark Driven Guide to Binding Site Comparison: An Exhaustive Evaluation Using Tailor-Made Data Sets (ProSPECCTs). *PLoS Comput. Biol.* **2018**, *14*, 1–50.

# CHAPTER 3

ProCare validation: fragment repurposing and secondary target prediction

# 3.1. Unexpected similarity between HIV-1 reverse transcriptase and tumor necrosis factor binding sites revealed by computer vision

This section was integrally published in:

Merveille Eguida and Didier Rognan. *J. Cheminform.* 2021, 13, 90.

**RESEARCH ARTICLE**                                    **Open Access**

## Unexpected similarity between HIV-1 reverse transcriptase and tumor necrosis factor binding sites revealed by computer vision

Merveille Eguida and Didier Rognan*

**Abstract**

Rationalizing the identification of hidden similarities across the repertoire of druggable protein cavities remains a major hurdle to a true proteome-wide structure-based discovery of novel drug candidates. We recently described a new computational approach (ProCare), inspired by numerical image processing, to identify local similarities in fragment-based subpockets. During the validation of the method, we unexpectedly identified a possible similarity in the binding pockets of two unrelated targets, human tumor necrosis factor alpha (TNF-α) and HIV-1 reverse transcriptase (HIV-1 RT). Microscale thermophoresis experiments confirmed the ProCare prediction as two of the three tested and FDA-approved HIV-1 RT inhibitors indeed bind to soluble human TNF-α trimer. Interestingly, the herein disclosed similarity could be revealed neither by state-of-the-art binding sites comparison methods nor by ligand-based pairwise similarity searches, suggesting that the point cloud registration approach implemented in ProCare, is uniquely suited to identify local and unobvious similarities among totally unrelated targets.

**Keywords:** Binding sites, Similarity, Point cloud registration

## 3.1.1. Abstract

Rationalizing the identification of hidden similarities across the repertoire of druggable protein cavities remains a major hurdle to a true proteome-wide structure-based discovery of novel drug candidates. We recently described a new computational approach (ProCare), inspired by numerical image processing, to identify local similarities in fragment-based subpockets. During the validation of the method, we unexpectedly identified a possible similarity in the binding pockets of two unrelated targets, human tumor necrosis factor alpha (TNF-α) and HIV-1 reverse transcriptase (HIV-1 RT). Microscale thermophoresis experiments confirmed the ProCare prediction as two of the three tested and FDA-approved HIV-1 RT inhibitors indeed bind to soluble human TNF-α trimer. Interestingly, the herein disclosed similarity could be revealed neither by state-of-the-art binding sites comparison methods nor by ligand-based pairwise similarity searches, suggesting that the point cloud registration approach implemented in ProCare, is uniquely suited to identify local and unobvious similarities among totally unrelated targets.

**Keywords:** binding sites, similarity, point cloud registration

## 3.1.2. Introduction

Among the many possible approaches to structure-based drug design [1, 2], inferring novel ligand properties from the large-scale comparison of their possible binding pockets gains popularity as the repertoire of protein cavities of known three-dimensional (3D) structures (pocketome) is constantly increasing, thereby offering unique opportunities to design ligands while simultaneously considering multiple targets [3]. The term 'pocketome' was first coined in 2004 by An et al. [4] to describe the universe of cavities located at the surface of macromolecules and capable of binding low molecular-weight ligands. A systematic survey of currently available three-dimensional structures [5], suggests that its size is estimated to ca. 250,000 pockets [6] out of which 10-15% are accommodating true drug-like compounds [7, 8]. Pocket locations can be inferred from the position of already-bound molecules or predicted on the fly, by one of the many available cavity detection methods [3, 9]. The pockeome space can then be searched by numerous computational tools [10] for similarity to any query cavity to predict evolutionary relationships and protein-ligand interactions [3]. The later application is notably of paramount importance to the drug discovery field as it may reveal hidden relationships for guiding the design of safer drug candidates with a precise control of selectivity [3] with respect to either on-targets (polypharmacology approach) [11] or off-targets (side effects mitigation) [12], in a time and cost-effective manner [13].

Currently available methods are generally able to detect global similarities between two druggable pockets from different proteins, and therefore permit to transfer drug-like compounds from one target space to another [3]. Identifying more subtle local similarities at the level of fragment-bound pockets remains a much more difficult problem [14] as it requires a searchable archive of fragment-bound subpockets [15–17] and a computational method focusing on local subpocket descriptors. Consequently, there are still very few reports of experimentally verified subpocket similarity examples that have enabled the transfer of chemical fragments across unrelated proteins [18]. To fill the need for local similarity searching methods while comparing pockets of different sizes, we developed a novel method (ProCare) [17] relying on point cloud registration, a numerical image processing to find a spatial transformation (*e.g.,* scaling, rotation and translation) that aligns two point clouds [19, 20]. ProCare uses as input a point cloud representation of the protein pocket or subpockets, where each point is annotated by eight possible pharmacophoric features (hydrophobic, aromatic, H-bond donor, H-bond acceptor, H-bond donor and acceptor, positive, negative, dummy) complementary to that of the pocket microenvironment [21]. Since ProCare uses local descriptors to compare and align binding subpockets, the method is particularly suited to fragment-based design strategies aimed at positioning fragments in subpockets of any druggable cavity.

While validating the method by focused benchmarking studies [17], we noticed some unexpected local similarity between subpockets from two unrelated proteins with 23% sequence identity: human tumor

necrosis factor alpha  (TNF-α) trimer [22] and human immunodeficiency virus type 1 reverse transcriptase (HIV-1 RT) [23]. On the one hand side, TNF-α is a homotrimeric pro-inflammatory cytokine involved in autoimmune disorders such as rheumatoid arthritis and Crohn's disease [24]. It is currently targeted by monoclonal antibodies preventing its recognition by TNF-α receptors (TNFR1 and TNFR2). To date, no small molecule TNF-α inhibitor has reached the market [22]. On the other side, HIV-1 RT is an enzyme used by the HIV virus to replicate its genome by first generating a complementary DNA from the viral RNA template. HIV-1 RT can be blocked by many marketed drugs [25] binding to either the catalytic site (nucleoside inhibitors, e.g. zidovudine) or a remote allosteric pocket (non-nucleoside inhibitors, e.g. efavirenz).

To exclude potential artifacts or biases and provide a strong statistical support to this initial prediction, we here systematically compared the inner cavity of three inhibitor-bound TNF-α trimer structures with 122 non-nucleoside inhibitor-bound HIV-1 RT X-ray structures. In a large majority of pairwise comparisons, the corresponding subpockets were deemed similar, a prediction that could be confirmed by biophysical experiments evidencing a direct micromolar binding of non-nucleoside HIV-1 RT inhibitors to human soluble TNF-α. Interestingly, this unexpected similarity could not be recovered by state-of-the-art cavity comparisons tools suggesting the unique ability of ProCare to delineate subtle local relationships between unrelated target cavities.

## 3.1.3. Results and discussion

Identifying similarity between pockets from different proteins suggests that the latter might bind to similar molecules. Although molecular recognition is a dynamic and complex process, the above hypothesis is worth investigating in drug design for hit discovery or for potential off-targets detection. We previously described ProCare [17], a novel computational method relying on a point cloud registration algorithm [19, 20] to assess the similarity between protein pockets. ProCare computes and uses local descriptors, which makes it particularly suitable for detecting local similarities among cavities of different sizes. Typically, ProCare aligns the cavities, described by a cloud of 3D points labeled with pharmacophoric features, by comparing the point descriptors and then derives a similarity score. In the current study (see flowchart in **Figure 1**), ProCare was used to detect local similarities between the full cavity of the target protein (here the inner core of the TNF- α trimer) and a collection of 31,570 subpockets from the sc-PDB dataset [8], a repository of 16,034 protein-ligand complexes of known three-dimensional structure for which the ligand is a pharmacological agent bound to a druggable cavity. First, the full cavity of the target protein is computed with the in-house VolSite algorithm [21] and represented by a cloud of pharmacophore-annotated points (**Figure 1**). In parallel, the collection of

subpocket point clouds is generated after fragmentation of each protein-bound sc-PDB ligand and consideration of the immediate vicinity (4 Å) of generated fragments. Last, the ProCare method aligns and computes the pairwise similarity between the target point cloud, and that from subpockets from the sc-PDB archive (**Figure 1**). When a statistically significant similarity is found between a subpocket and the target cavity, the transformation matrix used for the previous alignment is then applied to the corresponding and hidden bound fragment that is directly positioned in the target cavity. In absence of major clashes, the corresponding fragment can therefore be used for a fragment growing or linking strategy or even directly tested for binding to the target.

While benchmarking the ProCare method, we noticed unexpected high similarities (ProCare score > 0.47; p-value < 0.05) between the core pocket at the interface of an inhibitor-bound asymmetric human TNF-α trimer (PDB ID 6OOY) [22], and several non-nucleoside binding sites of inhibitor-bound HIV-1 RT (**Supporting Table S1**). Notably, seven subpockets from the HIV-1 RT were ranked among the 100 top scoring subpockets, with high ProCare similarity scores (ranging from 0.67 to 0.72) corresponding to very low p-values (from $2.5 \times 10^{-4}$ to $2.1 \times 10^{-5}$).
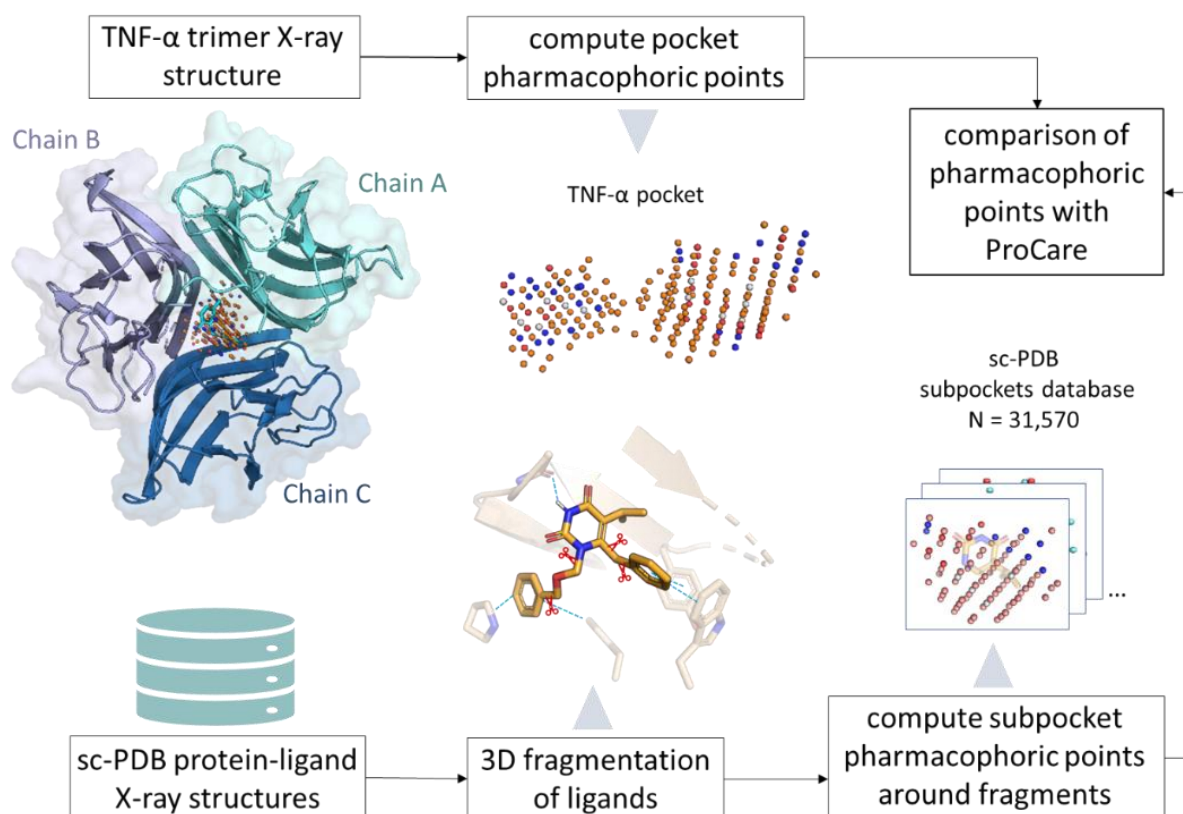


**Fig. 1 Virtual screening of sc-PDB subpockets for similarity to the core cavity TNF-α.** The inner pocket of TNF-α (PDB ID 6OOY) is converted as a cloud of points with pharmacophoric properties (orange: hydrophobic and aromatic, blue: H-bond donor and positive ionizable, red: H-bond acceptor, H-bond donor and acceptor, and negative ionizable, white: dummy) and compared to the corresponding point clouds originating from fragment-bound subpockets of sc-PDB ligands.

151

To assess that the predicted similarity between these unrelated binding sites was not fortuitous, we computed the Receiver-Operating Characteristic (ROC) curve of a binary classifier for which all cavities of a single sc-PDB target (**Table 1**) are artificially annotated as positives, the rest being defined as negatives. For each target, the ROC curve was defined from the full list of sorted ProCare similarity scores by plotting the true positive rate versus the false positive rate at different threshold settings (**Supporting Figure S1**). The area under the ROC curve (ROCAUC) provides a statistical estimation of the accuracy of the classifier to discriminate positives from negatives and therefore predict whether the samples from one particular target are similar (or not) to the TNF-α cavity (**Table 1**).

**Table 1** Area under the ROC curve of pairwise ProCare similarity scores.[a]

| Target | Site | Number of subpockets[b] | ROCAUC |
|---|---|---|---|
| HIV-1 RT | non-nucleoside | 195 (122) | 0.84 |
| β2 adrenergic receptor | orthosteric | 14 (14) | 0.35 |
| Carbonic anhydrase II | catalytic | 183 (137) | 0.38 |
| Cyclin-dependent kinase 2 | catalytic | 461 (274) | 0.63 |
| Heat shock protein 90α | catalytic | 214 (117) | 0.64 |
| Thrombin | catalytic | 253 (126) | 0.35 |

[a] For each target, the similarity scores of the corresponding subpockets (actives) and decoys (any other subpocket) to the TNF-α query (PDB ID 6OOY) are used to compute the area under the ROC curve.

[b] Total number of subpockets for the corresponding target. The number of PDB entries are in brackets.

Making the hypothesis that the HIV-1 RT non-nucleoside binding pocket is similar to that of TNF-α, the ProCare score nicely discriminates positives (HIV-1 RT) from decoys (all other sc-PDB cavities) with a ROCAUC value (0.84) well above the threshold corresponding to a random classification, ROCAUC=0.50). Repeating the same exercise with five randomly picked targets (β2 adrenergic receptor, carbonic anhydrase II, cyclin-dependent kinase 2, heat shock protein 90α, and thrombin) lead to much poorer ROC AUC values close or even inferior to random classifications (**Table 1**). To further exclude a potential bias in the ProCare alignment/scoring method due to the reference TNF-α structure (PDB ID 6OOY) and give a stronger statistical support to our prediction, we systematically compared two additional binding sites (PDB IDs 6OOZ, 6OP0) from available asymmetric human TNF-α X-ray structures [22] to that of 122 HIV-1 RT structures bound to non-nucleoside inhibitors.

**Exhaustive comparison of TNF-α trimer and HIV-1 reverse transcriptase binding sites.** A ProCare similarity matrix was built by comparing cavities of three asymmetric TNF-α structures (PDB identifiers 6OOY, 6OOZ and 6OP0) co-crystallized with benzimidazole inhibitors to the 195 subpockets from 122

non-nucleoside HIV-1 RT inhibitors binding sites (**Supporting Table S2**; **Figure 2**) available in the sc-PDB. We observed that 76% of all pairwise comparisons were scored higher than the previously statistically determined ProCare similarity score threshold of 0.47 [17] (**Figure 2A**).
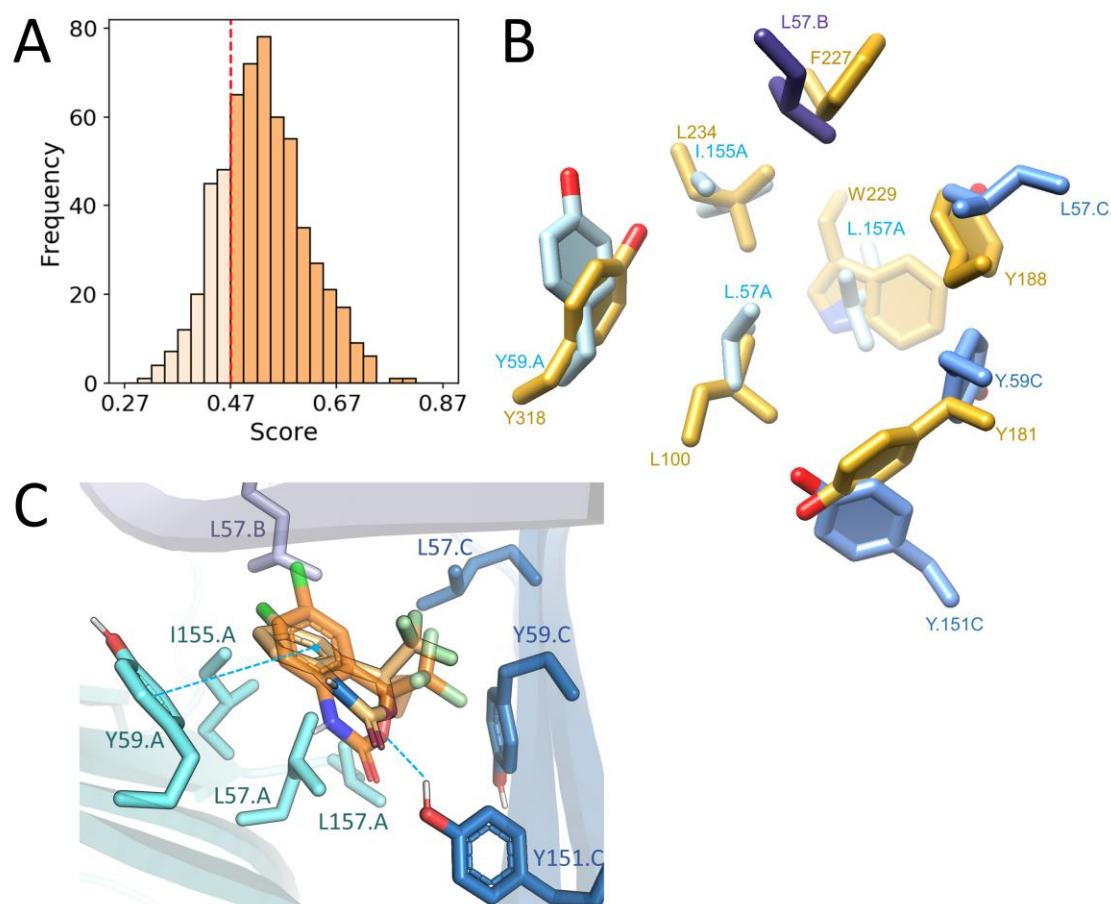


**Fig. 2 Comparison of TNF-α trimer and HIV-1 RT binding sites with ProCare.** (**A**) Distribution of pairwise similarity scores (n = 195 x 3). Entries scoring above 0.47 (p-value=0.05; threshold marked by the red dashed line) are considered similar according to a previous statistical analysis of 2 million pairwise alignments [17]. (**B**) Aligned residues of TNF-α (chain A: cyan, chain B: dark slate blue, chain C: cornflower blue; PDB code: 6OOZ) to HIV-1 RT (orange, PDB code: 1FKO) after rotation and translation of HIV-1 RT protein with the resulting ProCare alignment matrix. (**C**) ProCare alignment of efavirenz main fragment (light orange) in the TNF-α trimer pocket and PLANTS docking (transparent orange) in the TNF-α trimer pocket (PDB code: 6OOZ). Edge-to-face aromatic interaction with TYR59 of TNF-α chain A and hydrogen bond with TYR151 of TNF-α chain C are depicted by blue dashed lines.

To exclude the possibility that the predicted similarity is caused by peculiar mutations of the HIV-1 RT non-nucleoside biding site, we also compared pairwise similarities for both wild type and mutated HIV-

1 RT pockets, but did not observe significant differences in the percentage of HIV-1 RT pockets predicted similar to that of TNF-α (74% and 82% of similar pockets for wild type and mutants, respectively). We thus conclude that the predicted similarity between pockets from these two unrelated targets is independent on the chosen PDB structures and is not biased by mutations in the HIV-1 RT binding site. Since ProCare yields a transformation matrix to align the compared objects (subpockets onto the target pockets), we herein provided the visual analysis for one entry (efavirenz-bound subpocket) aligned to the TNF-α structure 6OOZ. Pairs of residues of equivalent interaction properties (aromatic, hydrogen bond donor and acceptor, hydrophobic) respectively in TNF-α and HIV-1 RT were nicely matched (**Figure 2B**) demonstrating that the similarity caught with the point clouds is truly present at the residue level. Matched TNF-α/HIV-1 RT residues were: LEU57.A/LEU100; TYR59.A/TYR318; ILE155.A/LEU234; LEU157.A/TRP229; LEU57.B/PHE227; LEU57.C/TYR188; TYR59.C/TYR181 and TYR151.C/TYR181. Inspection of the matched pharmacophoric points that are contributing to the ProCare score showed a mixed contribution of aromatic, hydrogen bond donor and hydrophobic points (**Supporting Figure S2**) in agreement with the aligned residues (**Figure 2B**) and the statistics of the contributions of the eight pharmacophoric features to the detected similarity (**Supporting Figure S3**). Furthermore, efavirenz was docked into TNF-α binding site 6OOZ with PLANTS [26] after validation of the docking protocol by self-docking of the cocrystallized ligand UCB-5307 in 6OOZ (RMSD of top-ranked pose by ChemPLP to crystal coordinates: 0.47 Å, ChemPLP score of -124.79). The ProCare-aligned efavirenz fragment (**Figure 3B**) in TNF-α fitted well with one of the PLANTS docking solutions (ranked 3rd/10 with a ChemPLP score of -79.32), corresponding to a RMSD of 1.8 Å of efavirenz main fragment heavy atoms to the ProCare pose (**Figure 2C**). Aside the potential hydrophobic interactions in the TNF-α binding site, efavirenz docking pose displayed an edge-to-face aromatic interaction with residue TYR59.A and a hydrogen bond with TYR151.C. Interestingly, efavirenz bound to HIV-1 RT protein structure (1FKO) exhibits an edge-to-face aromatic interaction with residue TYR318 [27] (**Supporting Figure S4A**) that was matched by ProCare to TYR59.A in TNF-α (**Figure 2B**). Both TYR59.A and TYR151.C are key residues [22] involved in the micromolar and nanomolar binding of the co-crystallized ligands UCB-6876, UCB-5307 and UCB-9260 (**Figure 3**) in the TNF-α structures 6OOY, 6OOZ, 6OP0; the interaction between TYR151.C residue and the benzimidazole moiety being a hydrogen bond (**Supporting Figure S4B**). Altogether, these observations are strongly suggesting that subpockets in the non-nucleoside binding site of HIV-1 RT are similar to the TNF-α trimer cavity.
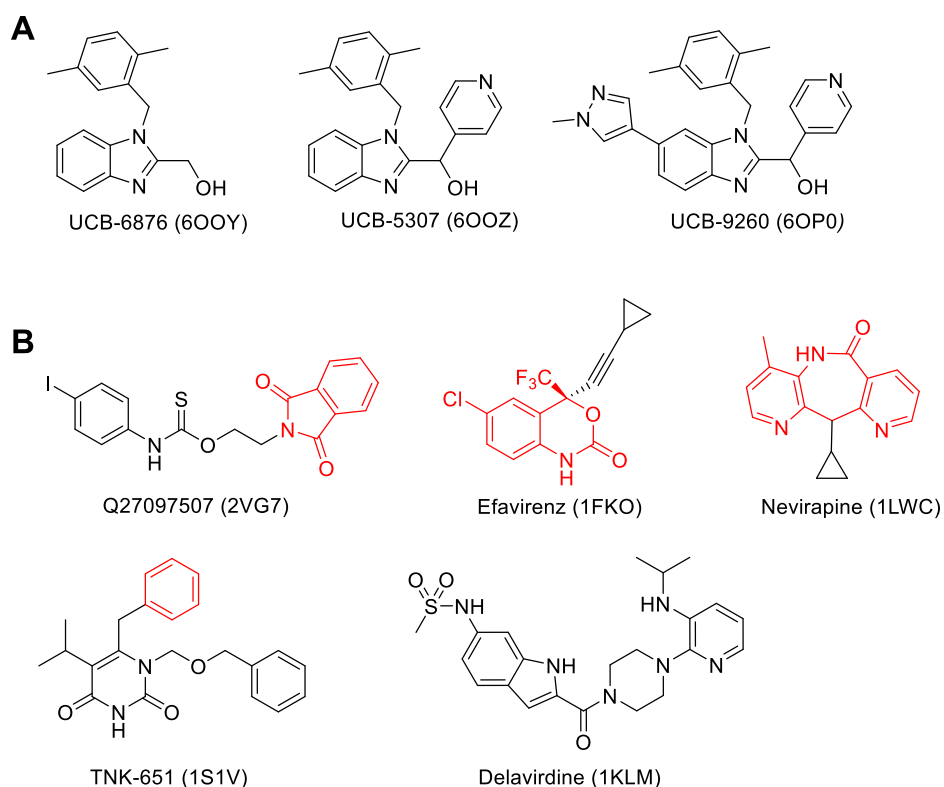
**Fig. 3 Structures of TNF-α and HIV-1 RT non-nucleoside inhibitors.** (**A**) TNF-α inhibitors and (**B**) HIV-1 RT non-nucleoside inhibitors (PDB entries between brackets). Red substructures indicate the main fragment binding to the HIV-1 RT subpocket found similar to the TNF-α cavity.

Assuming that similar binding sites should accommodate similar ligands, HIV-1 RT non-nucleoside inhibitors should therefore bind to TNF-α. In order to prioritize HIV-1 RT inhibitors for experimental validation of our hypothesis, we checked which inhibitors were bound to the HIV-RT subpockets that are predicted by ProCare as the most similar to the TNF-α cavity (**Table 2**).

Among the corresponding inhibitors, two compounds (Q27097507, TNK6-51) are not commercially available and were not considered. However, two easily purchasable FDA-approved drugs (efavirenz, nevirapine; **Figure 3**) are almost entirely buried in the HIV-1 RT subpockets found similar to the TNF-α cavity, exhibit a size and molecular volume similar to that of two TNF-α inhibitors (UCB-6876 and UCB-5307; **Figure 3**) and were therefore selected for biological evaluation. In addition, we also considered a third marketed inhibitor (delavirdine; **Table 2**, **Figure 3**) whose pocket was found much less similar to that of TNF-α, although just above the 0.47 ProCare similarity threshold.

**Table 2** Bound inhibitors of the HIV-1 reverse transcriptase cavities found similar to TNF-α cavities.

| HIV-RT Inhibitor[a] | HIV1-RT PDB entry | TNF-α PDB entry | ProCare score | Rank |
|---|---|---|---|---|
| NNI (Q27097507) | 2VG7 | 6OOZ | 0.810 | 1 |
| EFZ (Efavirenz) | 1FKO | 6OOZ | 0.773 | 2 |
| NVP (Nevirapine) | 1LWC | 6OOZ | 0.737 | 3 |
| TNK (TNK-651) | 1S1V | 6OOZ | 0.731 | 4 |
| NVP (Nevirapine) | 2HNY | 6OOZ | 0.729 | 5 |
| … | | … | … | … |
| SPP (Delavirdine)[b] | 1KLM | 6OOZ | 0.484 | 408 |

[a] PDB chemical component identifier (Name in brackets).

[b] After manual fragmentation, a higher ProCare score (0.599) was obtained for the subpocket of delavirdine's fragment #2 (**Supporting Figure S5**) against 6OOY pocket (**Supporting Table S3**).

**Non-nucleoside HIV-1 RT inhibitors bind to human TNF-α.** Three different non-nucleoside FDA-approved drugs (nevirapine, efavirenz and delavirdine) were tested for direct binding to a fluorescent-labelled TNF-α trimer by microscale thermophoresis (MST), a robust and sensitive biophysical method to detect and quantify molecular interactions in solution [28, 29]. The MST signal is based on ligand-dependent temperature-induced changes (thermophoresis, temperature-related fluorescence intensity) of the fluorescence emission of the labelled protein target. The 17.3 kDa homotrimeric TNF-α that spontaneously assemble in solution [30, 31] was therefore labelled by a RED-fluorescent probe for MST experiments in presence of increasing concentrations of the three HIV-1 RT inhibitors (**Figure 4**).

MST traces in presence of efavirenz and delavirdine showed distinct states (from bound to unbound), indicating a direct interaction of these two inhibitors with TNF-α (**Figure 4A, B**). Dissociation constants ($K_D$) could be derived for the two corresponding complexes and estimated to $24 \pm 8$ µM (Efavirenz) and $39 \pm 9$ µM (Delavirdine), respectively (**Figure 4A, B**). The measured dissociation constants for the two HIV-1 RT inhibitors are in the same range of magnitude than that of UCB-6876 ($K_D = 22$ µM) [22], one of the three TNF-α inhibitors used as a reference for this study.
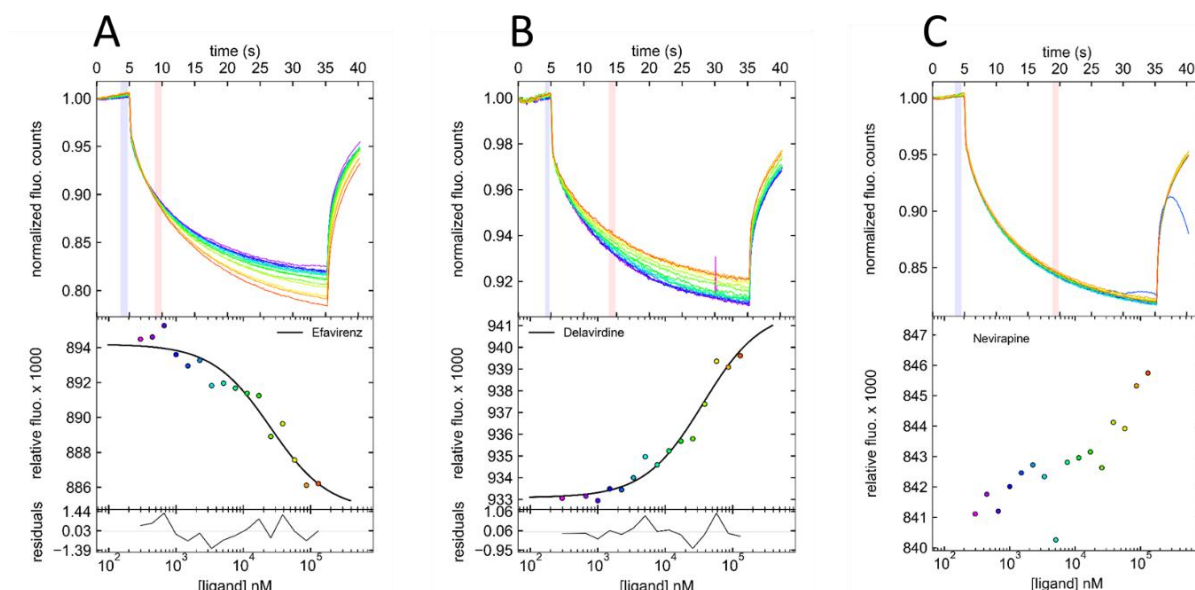
**Fig. 4 Microscale thermophoresis (MST) demonstrates a direct interaction between HIV-1 RT inhibitors and RED fluorescent-tagged TNF-α.** For analysis, the change in thermophoresis is expressed as the change in the normalized fluorescence ($\Delta F_{norm}$), which is defined as $F_{hot}/F_{cold}$ (*F*-values correspond to average fluorescence values between defined areas marked by the red and blue cursors). Titration of the non-fluorescent ligand results in a gradual change in thermophoresis, which is plotted as $\Delta F_{norm}$ to yield a binding curve, which can be fitted to derive binding constants. **(A)** Experimental MST traces of efavirenz ($K_D = 24 \pm 8$ µM); **(B)** Experimental MST traces of delavirdine ($K_D = 39 \pm 9$ µM); **(C)** Experimental MST traces of nevirapine. Only the best MST traces (highest signal to noise ratio) are shown here. Values for all experiments conducted according to different experimental protocols are listed in **Supporting Table S4**.

Contrarily to our prediction, no thermophoresis signal could be detected in presence of nevirapine (**Figure 4C**) indicating no binding of this inhibitor to TNF-α, at least in our experimental settings. The herein observations were insensitive to experimental protocols (buffer composition, solubilizing agents, incubation time, MST power; **Supporting Table S4**).

In absence of X-ray structures of TNF-α bound to efavirenz and delavirdine, we cannot rule out the possibility that both inhibitors bind to a different pocket than that highlighted in the current computational study. This hypothesis is however very unlikely for two reasons: (i) no other cavity than that occurring at the inner core of the multimeric TNF-α could be detected among the currently existing 33 structures available in the Protein Data Bank; (ii) all non-covalent small molecular weight inhibitors co-crystallized with TNF-α dimeric or trimeric forms [32–35] are exactly bound at the central pocket examined in this study.

We should recall here that none of the HIV-1 RT inhibitors has been optimized for binding to TNF-α and is directly repurposable for treating TNF-α -dependent autoimmune disorders. However, we do think that efavirenz may be optimized to a much more potent HIV-1 RT inhibitor by following a strategy similar to that reported to modify the 22 μM TNF-α inhibitor UCB-6876 to a 9 nM lead (UCB-5307; **Figure 3**) by just occupying a side pocket formed by the three TYR199 side chains of the TNF-α homotrimer with a pyridyl ring [22]. Structure-guided efavirenz optimization for TNF-α binding is therefore possible by appropriate trimming of unnecessary cyclopropylethynyl substituent and occupation of the above-described potency subpocket.

**The similarity between TNF-α trimer and HIV-1 reverse transcriptase binding sites is not obvious.** To demonstrate whether the herein disclosed similarity between the human TNF-α trimer and the HIV-1 RT non-nucleoside binding sites is obvious, we performed the same set of pairwise binding site comparisons, as that previously reported for ProCare (**Figure 2**), with state-of-the-art methods [10] developed either in-house (FuzCav [36], Shaper [21] and SiteAlign [37]) or by third parties (G-LoSA [38], KRIPO [15] and ProBiS [39]). The binding site perception, comparison algorithm and scoring function is specific to each method. Some methods (FuzCav, SiteAlign) consider entire cavities while some others utilize either fragment-bound subpockets (KRIPO, Shaper) or local protein descriptors (G-LoSA). To make the comparison consistent, the same set of atomic coordinates were compared, a binding site being defined by the protein PDB identifier, the ligand PDB HET record (three alphanumeric character describing non-standard PDB residues), chain identifiers and list of amino acids lining the cavity. The only exception was for the KRIPO method, which used all the chains available in the PDB entry, but still corresponding to the same tuple (PDB, HET) as for the other methods. For each method, the distribution (**Figure 5**) and percentage of pairwise comparisons scored above the developer's recommended similarity threshold (**Table 3**) were reported.

**Table 3** Comparison of three TNF-α and 122 HIV-1 RT non-nucleoside binding sites by state-of-the-art cavity comparison methods.

| Method | Score threshold[a] | Metric | Success rate[b] |
|---|---|---|---|
| G-LoSA | 0.59 | GA-score | 35.2 |
| KRIPO | 0.50 | Modified Tanimoto coefficient | 5.8 |
| Shaper | 0.44 | ColorRefTversky | 1.4 |
| SiteAlign | 0.6, 0.2 | d1 and d2 distances[c] | 0.3 |
| FuzCav | 0.16 | Tanimoto coefficient | 0 |
| ProBiS | 2 | Z-score[d] | 0 |
| ProCare | 0.47 | ProCare score | 76.6 |

[a] Developer's recommended similarity/distance threshold for estimating two binding sites similar.

Strikingly, only the G-LoSA method relying on a graph-based local alignment of cavity-lining amino acids, managed to find some similarity between the two sets of binding sites, however with reduced success rate (35.2%) when compared to the ProCare algorithm (76.6 % success rate; **Table 3**). We acknowledge that the developer's recommended thresholds may be biased toward peculiar datasets. However, all methods compared herein were subjected to the same protocol and we do think that the threshold scores are appropriate indicators in a virtual screening setting where there is no room for a one-by-one case study of each pairwise comparison.
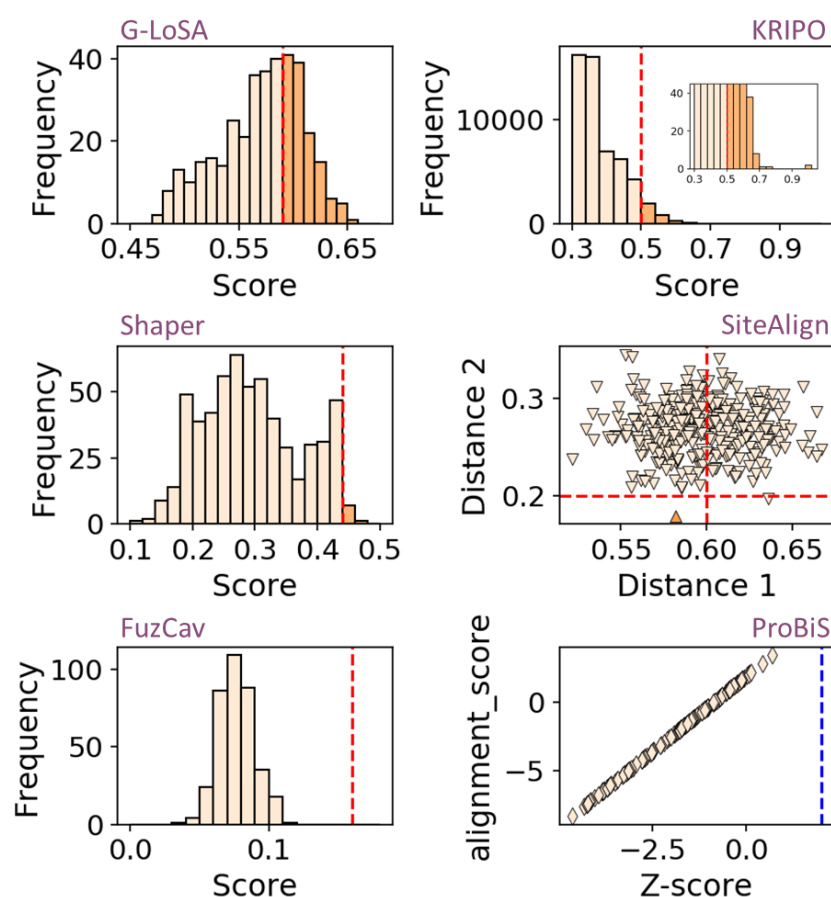


**Fig. 5 Score distribution of pairwise comparisons between binding sites of TNF-α trimer and HIV-1 reverse transcriptase.** Binding sites in asymmetric structures of TNF-α trimer (n=3) were compared to binding sites of non-nucleoside inhibitors in HIV-1 reverse transcriptase (sc-PDB set, n=122). Pairs with similarity measures scored above each method-specific threshold (red dashed line) were considered

similar. For SiteAlign comparisons, pairs are considered similar in case the two distances (distance 1, distance 2) are below the recommended cut-off. For ProBiS, the threshold above which an alignment is considered significant is marked by the blue dashed line.

The herein reported binding of some HIV-1 RT non-nucleoside inhibitors to human TNF-α remains unobvious to many binding site comparison algorithms. Would this unexpected feature be better captured by remote ligand similarities? To investigate this question, we compared 2D and 3D descriptors of the corresponding inhibitors (**Figure 6**).
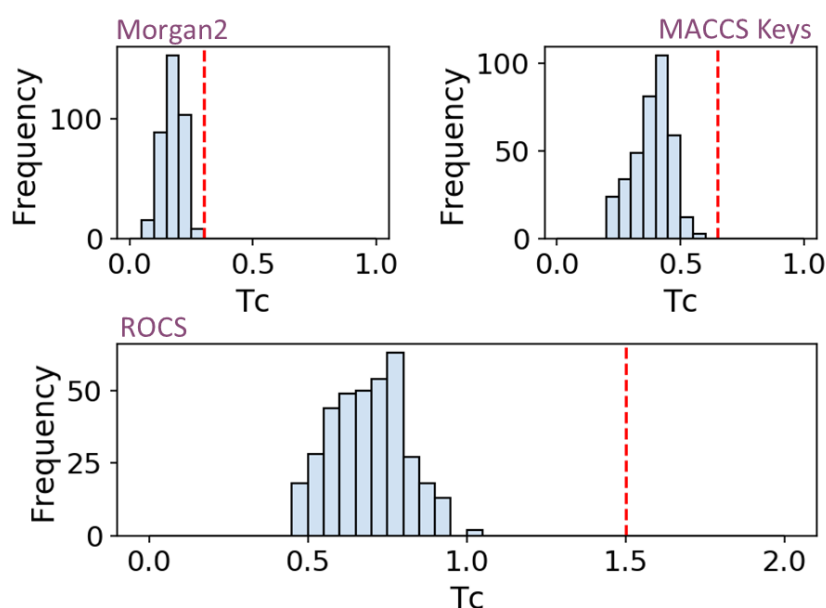


**Fig. 6  Pairwise similarity between inhibitors of TNF-α trimer and non-nucleoside inhibitors of HIV-1 reverse transcriptase.** Recently described TNF-α trimer inhibitors (n=3) were compared to non-nucleoside inhibitors of HIV-1 RT (sc-PDB set, n=122). Pairs with similarity measures scored above each descriptor-specific threshold (red dashed line) were considered similar. (**Top left**) 2D similarity estimated by a Tanimoto metric using Morgan2 circular fingerprint, (**Top right**) 2D similarity estimated by a Tanimoto metric using 166 MACCS public keys. (**Bottom**) 3D shape comparison (ROCS) estimated by the TanimotoCombo metric.

Neither comparing 2D fingerprints nor 3D shapes would have confidently suggested possible binding of HIV-1 RT inhibitors to TNF-α trimer (**Figure 6**) since none of the considered ligand pairs exhibit a pairwise similarity above an acceptable threshold (Morgan2 circular fingerprint: 0.30 [40]; 166 public MACCS keys: 0.65 [40], TanimotoCombo ROCS 3D similarity: 1.5 [41, 42]). We should precise here that 3D similarities were inferred from PDB protein-bound ligand X-ray structures and that alternative conformations might be selected by the two targets, although the very rigid efavirenz does indeed bind

to the two proteins of interest albeit with different affinities (TNF-α, $K_D$=24 μM; HIV-1RT, ChEMBL median $IC_{50}$= 20 nM). Extending 2D fingerprint comparisons to additional 2,361 HIV-1 RT inhibitors (**Supporting Table S5**) from the ChEMBL database [43], did not change our conclusion since only 0.71% and 0.09% of the corresponding pairs were found similar using Morgan2 and 166 public MACCS keys, respectively  (data not shown).

## 3.1.4. Conclusions

Herein, we describe a systematic comparison of fragment-bound subpockets from *a priori* unrelated targets (TNF-α, HIV-1 RT) but predicted to share local similarities according to our recently-developed ProCare point cloud registration method.  The computational prediction was verified by microscale thermophoresis experiments evidencing the micromolar binding of some but not all HIV-1 RT non-nucleoside inhibitors to human soluble TNF-α. Interestingly, the ProCare prediction could not be revealed by other state-of-the-art cavity or ligand similarity search methods. Point cloud registration, a computational method frequently used for digital image processing in robotics and medical imaging, enables the detection of subtle and local protein similarities thanks to a powerful description of subpocket microenvironments. The ProCare method appears as a promising idea generator for drug repurposing and fragment-based ligand design since it is able to pick starting ligands at a proteomic scale.

## 3.1.5. Methods

**Preparation of protein and ligand structures**

*TNF-α structures.* The recently described asymmetric structures of the human TNF-α trimer bound to different inhibitors were retrieved from the RCSB Protein Data Bank (PDB) homepage (https://www.rcsb.org) [44] using the following identifiers: 6OOY, 6OP0, 6OOZ [22]. The PDB structures were protonated with Protoss [45] v.4.0, then split into protein, ligands and water molecules and finally converted into mol2 format with Sybyl-X v.2.1.1 (Certara USA, Inc., Princeton, NJ 08540). The binding sites ('SITE') were defined as any protein residue with at least one heavy atom closer than 6.5 Å from any ligand heavy atom and saved in mol2 and pdb formats. The ligands were converted into sdf format with OpenEye Python toolkits v.2020.0.4 (OpenEye Scientific Software, Santa Fe, U.S.A.). Cavities were detected with IChem v.5.2.9 VolSite utility [21] (cavity_all output) using default parameters. The cavity points are labeled with eight possible pharmacophoric features (hydrophobic, aromatic, H-bond donor, H-bond acceptor, H-bond donor and acceptor, positive, negative, dummy) that are complementary to the features of the nearest protein atom. If no protein atom is found within a 4 Å distance of a cavity point, the latter is assigned a dummy property.

*HIV-1 reverse transcriptase PDB structures.* Starting from the PDB structure 1VRT as a reference, a search was performed in the RCSB PDB (https://www.rcsb.org) [44] to retrieve all structures with strict matching ("Structure Similarity" query in the PDB). After visual check, 122 entries already available in the sc-PDB repository (http://bioinfo-pharma.u-strasbg.fr/scPDB) [8] and for which the ligand is a non-nucleoside inhibitor were kept. The remaining PDB structures were protonated with Protoss [45] v4.0. The list of the PDB identifiers and Uniprot accession numbers is reported **Supporting Table S2**. According to the sc-PDB preparation rules, the binding sites ('SITE') were defined as described above. Protein, ligand and binding site 'SITE' structures were directly retrieved in mol2 file format from the sc-PDB archive. The corresponding 122 ligands were 3D-fragmented with the IChem v.5.2.9 [49] fragmentation utility [47] and the complementary VolSite [21] cavity points, computed at 4 Å around each fragment were finally saved. The ligands were converted into sdf format as described above.

**Preparation of HIV-1 reverse transcriptase ChEMBL ligands**

Bioassay information were first retrieved from the ChEMBL [43] dataset (Release 28; https://www.ebi.ac.uk/chembl) by querying the general keyword 'reverse transcriptase' and retaining ChEMBL target identifiers (CHEMBL247, CHEMBL4296301, CHEMBL2366516) corresponding to HIV-1 RT. Ligands with a measured sub-micromolar half-maximal inhibitory concentration (IC50) against the HIV1-RT single target were defined here as inhibitors (**Supporting Table S5**). The

corresponding SMILES strings were retrieved and further processed with RDKit (Open-source cheminformatics; http://www.rdkit.org) v.2019.03.4.0 to remove redundancy.

**Preparation of sc-PDB fragments and subpockets**

Ligands coordinates from the sc-PDB (http://bioinfo-pharma.u-strasbg.fr/scPDB) [46] v.2016 archive were fragmented in 3D with the IChem v.5.2.9 fragmentation utility [47]. Fragmentations occurs in the binding sites so that only the main fragments interacting sufficiently (four interactions of which at least one is polar) with their target proteins were kept. Finally, the cavity pharmacophoric points cloud were computed at 4 Å from the fragments center to describe the protein subpocket, using the IChem v.5.2.9 VolSite utility ("cavity_4" output). VolSite cavities exhibiting less than three points were removed. A total of 31,570 valid fragment-bound subpockets were finally obtained.

**Cavity similarities**

*ProCare.* ProCare [17] v.0.1.1 pairwise comparison were performed on cavities computed with the VolSite module [21] in IChem v5.2.9 [49]. Entire cavities ("cavity_all" output) were calculated for TNF-α structures whereas only cavity points closer than 4.0 Å from any fragmented ligand center ("cavity_4" output) were considered for sc-PDB subpockets. VolSite cavity points were directly used for point cloud registration staring with determination of colored fast point feature histograms (c-FPFH) as previously described [17]. Finally, the respective set of c-FPFH descriptors for the two cavities were compared to each other using a RANSAC algorithm [19, 20] followed by refinement with default parameters [17]. Alignments results were scored with the default ProCare scoring function [17] which evaluates with a Tversky metric the proportion of aligned points of the same pharmacophoric features. In agreement with our previous study [17] where the similarity threshold of 0.47 (p-value of 0.05) was statistically determined, pockets scoring above 0.47 were considered similar.

*FuzCav.* FuzCav [36], an alignment-free method, was used to compare the binding site 'SITE' (mol2 format) entries of TNF-α dataset to the binding sites of HIV-1 RT sc-PDB dataset. Each binding site was tagged to compute a 4,833 bit-string that count all possible pharmacophoric triplets based on the atomic coordinates of Cα atoms lining the binding cavity. The pairwise comparisons of the fingerprints were evaluated with the default similarity score, with a threshold set at a value of 0.16 to distinguish similar from dissimilar binding sites.

*G-LoSA*. G-LoSA [38] v.2.2 is an alignment tool that was used with the binding sites 'SITE' pdb files. G-LoSA computes a set of inter-structural Cα pair distances to derive a graph, which will later be subjected to maximum clique search. The default G-LoSA score (GA-score) was used to evaluate the alignments. A threshold value of 0.59, recommended by the authors [38] and corresponding to a p-value of 0.05, was used to distinguish similar from dissimilar binding sites.

*KRIPO*. PDB ligands structural information were downloaded from Ligand Expo (http://ligand-expo.rcsb.org/) and prepared according to the KRIPO procedure (https://github.com/3D-e-Chem/kripo). Then KRIPO [15] v.1.0.1 was used with the list of prepared PDB structures for the pharmacophore fuzzy fingerprints calculations, using default parameters (fragmentation procedure activated). The pairwise similarities of the fingerprints were estimated with kripodb (v.3.0.0) using the modified Tanimoto coefficient as similarity metric. A threshold value of 0.50 was used to distinguish similar from dissimilar binding sites.

*ProBiS*. In a first place, the surface information (srf files) was computed for each prepared PDB structures with the default parameters referenced in the manual (3.0 Å to the ligand). ProBiS [39] requires a list of ligand HET code and residue number for each PDB entries. That list was provided to ensure that the ligands/sites considered are the same as in the binding site datasets used for other methods. Then, the alignment and comparison of the srf files were executed with default parameters, except for the Z-score that was set to a high negative value (-9999) as suggested by the authors to enforce the output of all results. Similarity between two binding sites was evaluated by the alignment score and Z-score. A threshold Z-score value of 2.0 was used to distinguish significant from irrelevant binding site alignments.

*SiteAlign*. For each entry, the list of natural amino acids in the 'SITE' mol2 files were provided as input. SiteAlign [37] v.4.0 describes a binding site by a polyhedron of 80 discretized triangles annotated with eight possible pharmacophoric features projected from cavity-lining C-α atoms. This results in a fingerprint of 640 integers. The pairwise comparisons imply aligning the corresponding polyhedron and computing the d1 and d2 distances of the fingerprints. The distance thresholds of d1=0.6 and d2=0.2 were applied respectively, to discriminate similar from dissimilar binding sites.

*Shaper*. Shaper [21] v.1.0 uses the same input files (VolSite cavities in mol2 file format) as ProCare. Shaper is an alignment method based on the OpenEye ShapeTK toolkit (OpenEye Toolkits 2020.2.0,

OpenEye Scientific Software, Santa Fe) to maximize the overlap of shape and pharmacophoric features of the two compared cavities, thanks to a smooth Gaussian function. The alignments were realized with default settings and scored with a Tversky metric putting more weight on the reference cavity (RefTve). A threshold RefTve value of 0.44 (p-value = 0.005) was used to distinguish similar from dissimilar binding sites.

**Ligand similarities**

*Ligand 2D similarity.* Morgan fingerprints on the one hand, and 166 public MACSS keys on the other hand were computed on the PDB ligands (sdf format) and ChEMBL ligands (SMILES strings) with RDKit (Open-source cheminformatics; http://www.rdkit.org) python package v.2019.03.4.0 using default parameters (radius = 2 Å for the Morgan fingerprints). The Tanimoto coefficients of the pairwise TNF-α ligands/HIV-1 RT ligands fingerprints comparison were reported. Cut-off values of 0.30 (Morgan fingerprints) and 0.65 (MACCS keys) were used to discriminate chemically similar from dissimilar ligands.

*Ligand 3D similarity.* sc-PDB HIV-1 RT inhibitors were compared to TNF-α inhibitors with OpenEye ROCS v.3.2.0.4 and scored by decreasing Tanimoto similarity metric accounting for both shape and chemical features overlap (TanimotoCombo). A TanimotoCombo cut-off value of 1.5 was used to discriminate chemically similar from dissimilar ligands.

**Docking**

TNF-α X-ray structure 6OOZ was prepared as described above (see TNF-α structures). 6OOZ co-crystallized ligand on the one hand, delavirdine, efavirenz and nevirapine as well as their main fragments on the other hand were drawn with MarvinSketch v.16.10.17 (ChemAxon Ltd, 1031 Budapest, Hungary) and saved into 2D sdf format. They were ionized with Filter v.2.5.1.4 (OpenEye Scientific Software, Santa Fe, U.S.A.) using customized parameters (**Supporting Table S6**). Then Corina v.3.40 (Molecular Networks GmbH, 90411 Nürnberg, Germany) was used to generate a starting 3D conformation for each inhibitor. The prepared molecules were docked into the target 6OOZ with PLANTS v.1.2 [26] using the following configuration: the grid was set at 13 Å from the binding site center; poses were searched 'speed1' settings to generate a maximum of 10 poses per ligand using a clustering rmsd of 2 Å. Solutions were scored with the default ChemPLP scoring function [26]. The docking protocol was validated by computing the rmsd between of the docked 6OOZ ligand coordinates and the X-ray coordinates. Results were processed and rescored by computing the interaction fingerprint (IFP) similarity (Tanimoto metric)

[48] between X-ray and docking poses. The IFPs were computed with IChem v.5.2.9 IFP module. All poses were visually inspected using Maestro v.2019-3 (Schrödinger, New York, NY 10036-4041).

**Chemicals and biologicals**

Nevirapine (catalog #S1742), efavirenz (catalog #S4685) and delavirdine mesylate (catalog #S6452) were purchased from Selleck Chemicals (https://www.selleckchem.com/). Soluble human TNF-α (catalog # Z01001) was purchased from GenScript (http://www.genscript.com).

**Binding of HIV-1 RT inhibitors to human TNF-α (Microscale thermophoresis)**

Human TNF-α was labeled using the RED-NHS 2nd generation labeling kit (NanoTemper Technologies GmbH) using a protein concentration of 10 µM and a molar dye-to-protein ratio ~ 3:1. A label/protein ratio of 0.4 was determined using photometry at 650 and 280 nm. Compounds efavirenz, delavirdine and nevirapine were initially dissolved in DMSO to afford stock solutions of 10 mM. These were then diluted to initial concentrations of 260 µM into 20 mM K-phosphate pH 7.4, 150 mM NaCl ensuring a final concentration of DMSO of 2.6 %. These compounds were serially diluted 2:1 in buffer 20 mM K-phosphate pH 7.4, 150 mM NaCl, 2.6 % DMSO producing ligand concentrations ranging from 260 µM to 594 nM (16 titration points). For MST measurements, each ligand dilution was mixed with 1 volume of fluorescently-labelled TNF-α at 680 nM in 20 mM K-phosphate pH 7.4, 150 mM NaCl, 0.02% Tween-20, which leads to a final concentration of TNF-α of 340 nM and final ligand concentrations at half of the ranges above. The final buffer is 20 mM K- phosphate pH 7.4, 150 mM NaCl, 0.01% Tween-20 and 1.3 % DMSO. After a 15-min incubation at room temperature in the dark, followed by centrifugation at 13,000 g for 3 min, each solution was filled into Monolith NT Premium capillaries (NanoTemper Technologies GmbH). Thermophoresis was measured at 25°C with 40% LED power and 20%, 40% and 80% MST power using a Monolith NT.115 (NanoTemper Technologies GmbH). Data were analyzed in the NT Analysis software version 1.5.41 (NanoTemper Technologies GmbH).

## 3.1.6. Associated contents

**List of abbreviations**

2D: two-dimension

3D: three-dimension

AUC: Area Under the Curve

c-FPFH: colored Fast Point Feature Histogram

DMSO: Dimethyl Sulfoxide

HIV: Human Immunodeficiency Virus

MST: Microscale Thermophoresis

PDB: Protein Data Bank

RANSAC: Random Sample Consensus

RMSD: Root Mean Square Deviation

ROC: Receiver Operating Characteristics

RT: Reverse Transcriptase

TNF: Tumor Necrosis Factor

**Supplementary information**

**Figure S1**: Receiver operating characteristic (ROC) curves derived from ProCare similarity scores. **Figure S2**: ProCare alignment of efavirenz main fragment subpocket onto TNF-α trimer pocket. **Figure S3**: Contributions of the eight pharmacophoric features to the ProCare similarity score between HIV-1 RT and TNF-α. **Figure S4**: Non-covalent interactions between efavirenz and HIV-1 RT, and between UCB-5307 and TNF-α trimer. **Figure S5**: Manual fragmentation of delavirdine in three fragments (#1 to #3). **Table S1**: sc-PDB subpockets sorted by decreased ProCare similarity to the inner cavity of human TNF-α. **Table S2**: PDB entries describing non-nucleoside inhibitors bound to HIV-1 reverse transcriptase. **Table S3**: Comparison of delavirdine subpockets, resulting from manual fragmentation, with TNF-α trimer pockets. **Table S4**: Dissociation constant (KD) of three HIV-1 RT inhibitor binding to human soluble TNF-α, according to MST experimental conditions. **Table S5**: CHEMBL entries describing HIV-1 RT non-nucleoside inhibitors. **Table S6**: Customized rules for OpenEye Filter ionization.

**Availability of data and materials**

*Data.* Input and results data are available at https://github.com/kimeguida/ProCare_TNF.

*Software.* ProCare, version 0.1.1 and 0.1.0, https://github.com/kimeguida/ProCare; Fuzcav, http://bioinfo-pharma.u-strasbg.fr/labwebsite/downloads/FuzCav.tgz; G-LoSA, version 2.2, https://compbio.lehigh.edu/GLoSA; KRIPODB, version 3.0.0, http://3d-e-chem.github.io/kripodb; KRIPO, version 1.0.1, https://github.com/3D-e-Chem/kripo; ProBiS, http://insilab.org/probis-algorithm/; SiteAlign, version 4.0, http://bioinfo-pharma.u-strasbg.fr/labwebsite/downloads/SiteAlign-4.0.tgz; Shaper, version 1.0, http://bioinfo-pharma.u-strasbg.fr/labwebsite/downloads/Shaper.tgz; RDKit python package, version 2019.03.4.0, https://www.rdkit.org/; ROCS, version 3.2.0.4, https://www.eyesopen.com/rocs; IChem, version 5.2.9, http://bioinfo-pharma.u-

strasbg.fr/labwebsite/downloads/IChem_v.5.2.9.tgz: Python OpenEye toolkits version 2020.0.4; FILTER, version 2.5.1.4, https://www.eyesopen.com/filter; PLANTS version 1.2, http://www.tcd.uni-konstanz.de/plants_download; Python package Matplotlib version 3.0.2; Maestro vesion 2019-3, https://www.schrodinger.com/products/maestro; Pymol version 2.1, https://pymol.org/2; Sybyl-X v.2.1.1, https://www.certara.com/sybyl-x-software; MarvinSketch version 16.10.17, https://chemaxon.com/products/marvin;

**Competing interests**

The authors declare no competing interests.

**Author contributions**

Conceived and design experiments: ME, DR. Performed the experiments: ME, Analyzed the data: ME, DR. Wrote the paper: ME, DR.

## 3.1.7. References

1.   Sliwoski G, Kothiwale S, Meiler J, Lowe EW (2014) Computational Methods in Drug Discovery. Pharmacol Rev 66:334–395. https://doi.org/10.1124/pr.112.007336
2.   Rognan D (2017) The impact of in silico screening in the discovery of novel and safer drug candidates. Pharmacol Ther 175:47–66. https://doi.org/10.1016/j.pharmthera.2017.02.034
3.   Ehrt C, Brinkjost T, Koch O (2016) Impact of Binding Site Comparisons on Medicinal Chemistry and Rational Molecular Design. J Med Chem 59:4121–4151. https://doi.org/10.1021/acs.jmedchem.6b00078
4.   An J, Totrov M, Abagyan R (2004) Comprehensive identification of "druggable" protein ligand binding sites. Genome Inform 15 2:31–41
5.   Berman HM (2000) The Protein Data Bank. Nucleic Acids Res 28:235–242. https://doi.org/10.1093/nar/28.1.235
6.   Bhagavat R, Sankar S, Srinivasan N, Chandra N (2018) An Augmented Pocketome: Detection and Analysis of Small-Molecule Binding Pockets in Proteins of Known 3D Structure. Structure 26:499-512.e2. https://doi.org/10.1016/j.str.2018.02.001
7.   Kufareva I, Ilatovskiy A V., Abagyan R (2012) Pocketome: an encyclopedia of small-molecule binding sites in 4D. Nucleic Acids Res 40:D535–D540. https://doi.org/10.1093/nar/gkr825
8.   Desaphy J, Bret G, Rognan D, Kellenberger E (2014) sc-PDB: a 3D-database of ligandable binding sites—10 years on. Nucleic Acids Res 43:D399–D404. https://doi.org/10.1093/nar/gku928
9.   Pérot S, Sperandio O, Miteva MA, et al (2010) Druggable pockets and binding site centric chemical space: A paradigm shift in drug discovery. Drug Discov Today 15:656–667. https://doi.org/10.1016/j.drudis.2010.05.015
10.  Ehrt C, Brinkjost T, Koch O (2018) A benchmark driven guide to binding site comparison: An exhaustive evaluation using tailor-made data sets (ProSPECCTs). PLoS Comput Biol 14:1–50. https://doi.org/10.1371/journal.pcbi.1006483
11.  Besnard J, Ruda GF, Setola V, et al (2012) Automated design of ligands to polypharmacological profiles. Nature 492:215–220. https://doi.org/10.1038/nature11691
12.  Jenkinson S, Schmidt F, Rosenbrier Ribeiro L, et al (2020) A practical guide to secondary pharmacology in drug discovery. J Pharmacol Toxicol Methods 105:106869. https://doi.org/10.1016/j.vascn.2020.106869
13.  Talevi A, Bellera CL (2020) Challenges and opportunities with drug repurposing: finding strategies to find alternative uses of therapeutics. Expert Opin Drug Discov 15:397–401. https://doi.org/10.1080/17460441.2020.1704729
14.  Milletti F, Vulpetti A, F. M, A. V (2010) Predicting Polypharmacology by Binding Site Similarity: From Kinases to the Protein Universe. J Chem Inf Model 50:1418–1431. https://doi.org/10.1021/ci1001263
15.  Wood DJ, Vlieg J De, Wagener M, Ritschel T (2012) Pharmacophore fingerprint-based approach to binding site subpocket similarity and its application to bioisostere replacement. J Chem Inf Model 52:2031–2043. https://doi.org/10.1021/ci3000776
16.  Kalliokoski T, Olsson TSG, Vulpetti A (2013) Subpocket analysis method for fragment-based drug discovery. J Chem Inf Model 53:131–141. https://doi.org/10.1021/ci300523r
17.  Eguida M, Rognan D (2020) A Computer Vision Approach to Align and Compare Protein Cavities: Application to Fragment-Based Drug Design. J Med Chem 63:7127–7142. https://doi.org/10.1021/acs.jmedchem.0c00422
18.  Zhou H, Cao H, Skolnick J (2021) FRAGSITE: A Fragment-Based Approach for Virtual Ligand Screening. J Chem Inf Model acs.jcim.0c01160. https://doi.org/10.1021/acs.jcim.0c01160

19. Rusu RB, Cousins S (2011) 3D is here: Point Cloud Library (PCL). In: 2011 IEEE International Conference on Robotics and Automation. IEEE, pp 1–4

20. Zhou Q-Y, Park J, Koltun V (2018) Open3D: A Modern Library for 3D Data Processing. arXiv:180109847. https://doi.org/10.1007/s00104-009-1793-x

21. Desaphy J, Azdimousa K, Kellenberger E, Rognan D (2012) Comparison and Druggability Prediction of Protein–Ligand Binding Sites from Pharmacophore-Annotated Cavity Shapes. J Chem Inf Model 52:2287–2299. https://doi.org/10.1021/ci300184x

22. O'Connell J, Porter J, Kroeplien B, et al (2019) Small molecules that inhibit TNF signalling by stabilising an asymmetric form of the trimer. Nat Commun 10:5795. https://doi.org/10.1038/s41467-019-13616-1

23. Kohlstaedt L, Wang J, Friedman J, et al (1992) Crystal structure at 3.5 A resolution of HIV-1 reverse transcriptase complexed with an inhibitor. Science (80- ) 256:1783–1790. https://doi.org/10.1126/science.1377403

24. Brenner D, Blaser H, Mak TW (2015) Regulation of tumour necrosis factor signalling: live or let die. Nat Rev Immunol 15:362–374. https://doi.org/10.1038/nri3834

25. Jochmans D (2008) Novel HIV-1 reverse transcriptase inhibitors. Virus Res 134:171–185. https://doi.org/10.1016/j.virusres.2008.01.003

26. Korb O, Stützle T, Exner TE (2009) Empirical Scoring Functions for Advanced Protein−Ligand Docking with PLANTS. J Chem Inf Model 49:84–96. https://doi.org/10.1021/ci800298z

27. Ren J, Milton J, Weaver KL, et al (2000) Structural Basis for the Resilience of Efavirenz (DMP-266) to Drug Resistance Mutations in HIV-1 Reverse Transcriptase. Structure 8:1089–1094. https://doi.org/10.1016/S0969-2126(00)00513-X

28. Wienken CJ, Baaske P, Rothbauer U, et al (2010) Protein-binding assays in biological liquids using microscale thermophoresis. Nat Commun. https://doi.org/10.1038/ncomms1093

29. Jerabek-Willemsen M, André T, Wanner R, et al (2014) MicroScale Thermophoresis: Interaction analysis and beyond. J Mol Struct 1077:101–113. https://doi.org/10.1016/j.molstruc.2014.03.009

30. Daub H, Traxler L, Ismajli F, et al (2020) The trimer to monomer transition of Tumor Necrosis Factor-Alpha is a dynamic process that is significantly altered by therapeutic antibodies. Sci Rep 10:9265. https://doi.org/10.1038/s41598-020-66123-5

31. Corti A, Fassina G, Marcucci F, et al (1992) Oligomeric tumour necrosis factor α slowly converts into inactive forms at bioactive levels. Biochem J 284:905–910. https://doi.org/10.1042/bj2840905

32. Blevitt JM, Hack MD, Herman KL, et al (2017) Structural Basis of Small-Molecule Aggregate Induced Inhibition of a Protein–Protein Interaction. J Med Chem 60:3511–3517. https://doi.org/10.1021/acs.jmedchem.6b01836

33. Xiao H-Y, Li N, Duan JJW, et al (2020) Biologic-like In Vivo Efficacy with Small Molecule Inhibitors of TNFα Identified Using Scaffold Hopping and Structure-Based Drug Design Approaches. J Med Chem 63:15050–15071. https://doi.org/10.1021/acs.jmedchem.0c01732

34. Dietrich JD, Longenecker KL, Wilson NS, et al (2021) Development of Orally Efficacious Allosteric Inhibitors of TNFα via Fragment-Based Drug Design. J Med Chem 64:417–429. https://doi.org/10.1021/acs.jmedchem.0c01280

35. McMillan D, Martinez-Fleites C, Porter J, et al (2021) Structural insights into the disruption of TNF-TNFR1 signalling by small molecules stabilising a distorted TNF. Nat Commun 12:582. https://doi.org/10.1038/s41467-020-20828-3

36. Weill N, Rognan D (2010) Alignment-free ultra-high-throughput comparison of druggable protein-ligand binding sites. J Chem Inf Model 50:123–135. https://doi.org/10.1021/ci900349y

37. Schalon C, Surgand JS, Kellenberger E, Rognan D (2008) A simple and fuzzy method to align

and compare druggable ligand-binding sites. Proteins Struct Funct Genet 71:1755–1778. https://doi.org/10.1002/prot.21858

38. Lee HS, Im W (2016) G-LoSA: An efficient computational tool for local structure-centric biological studies and drug design. Protein Sci 25:865–876. https://doi.org/10.1002/pro.2890

39. Konc J, Janežič D (2010) ProBiS algorithm for detection of structurally similar protein binding sites by local structural alignment. Bioinformatics 26:1160–1168. https://doi.org/10.1093/bioinformatics/btq100

40. Maggiora G, Vogt M, Stumpfe D, Bajorath J (2014) Molecular Similarity in Medicinal Chemistry. J Med Chem 57:3186–3204. https://doi.org/10.1021/jm401411z

41. Lo YC, Senese S, Damoiseaux R, Torres JZ (2016) 3D Chemical Similarity Networks for Structure-Based Target Prediction and Scaffold Hopping. ACS Chem Biol 11:2244–2253. https://doi.org/10.1021/acschembio.6b00253

42. Rush TS, Grant JA, Mosyak L, Nicholls A (2005) A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. J Med Chem 48:1489–1495. https://doi.org/10.1021/jm040163o

43. Gaulton A, Bellis LJ, Bento AP, et al (2012) ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 40:D1100–D1107. https://doi.org/10.1093/nar/gkr777

44. Burley SK, Bhikadiya C, Bi C, et al (2021) RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. Nucleic Acids Res 49:D437–D451. https://doi.org/10.1093/nar/gkaa1038

45. Bietz S, Urbaczek S, Schulz B, Rarey M (2014) Protoss: a holistic approach to predict tautomers and protonation states in protein-ligand complexes. J Cheminform 6:12. https://doi.org/10.1186/1758-2946-6-12

46. Desaphy J, Bret G, Rognan D, Kellenberger E (2015) Sc-PDB: A 3D-database of ligandable binding sites-10 years on. Nucleic Acids Res 43:D399–D404. https://doi.org/10.1093/nar/gku928

47. Desaphy J, Rognan D (2014) Sc-PDB-Frag: A database of protein-ligand interaction patterns for bioisosteric replacements. J Chem Inf Model 54:1908–1918. https://doi.org/10.1021/ci500282c

48. Marcou G, Rognan D (2007) Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. J Chem Inf Model 47:195–207. https://doi.org/10.1021/ci600342e

49. Da Silva F, Desaphy J, Rognan D (2018) IChem: A Versatile Toolkit for Detecting, Comparing, and Predicting Protein–Ligand Interactions. ChemMedChem 13:507–510. https://doi.org/10.1002/cmdc.201700505

## 3.1.8. Supplementary information for *Unexpected similarity between HIV-1 reverse transcriptase and tumor necrosis factor binding sites revealed by computer vision*

**Figure S1.** Receiver operating characteristic (ROC) curves of ProCare similarity scores.

**Figure S2.** ProCare alignment of efavirenz main fragment subpocket onto TNF-α trimer pocket.

**Figure S3.** Contributions of the eight pharmacophoric features to the ProCare similarity score between HIV-1 RT and TNF-α.

**Figure S4.** Non-covalent interactions between efavirenz and HIV-1 RT; and between UCB-5307 and TNF-α trimer.

**Figure S5.** Manual fragmentation of delavirdine in three fragments (#1 to #3).


**Table S1.** sc-PDB subpockets sorted by decreased ProCare similarity to the inner cavity of human TNF-α.

**Table S2.** PDB entries describing non-nucleoside inhibitors bound to HIV-1 reverse transcriptase.

**Table S3.** Comparison of delavirdine subpockets, resulting from manual fragmentation, with TNF-α trimer pockets.

**Table S4.** Dissociation constant ($K_D$) of three HIV-1 RT inhibitor binding to human soluble TNF-α, according to MST experimental conditions.

**Table S5.** CHEMBL entries describing HIV-1 RT non-nucleoside inhibitors.

**Table S6.** Customized rules for OpenEye Filter ionization.

**Figure S1.** Receiver operating characteristic (ROC) curves derived from ProCare similarity scores between sc-PDB subpockets and three TNF-α cavities (6OOY, 6OOZ, 6OP0). For each target (HIV-1 RT, HIV-1 reverse transcriptase; ADRB2, β2 adrenergic receptor; CAH2, carbonic anhydrase; CDK2, cyclin-dependent kinase 2; HSP90A, heat shock protein 90α; THRB, thrombin), the hypothesis is made that its cavity is similar to that of TNF-α and the area under the ROC curve of the corresponding classification is computed. The diagonal black dashed line corresponds to the performance of a random classifier (ROCAUC = 0.50). Number of subpockets for each target is given in brackets. **(A)** 6OOY query, **(B)** 6OOZ query and **(C)** 6OP0 query.
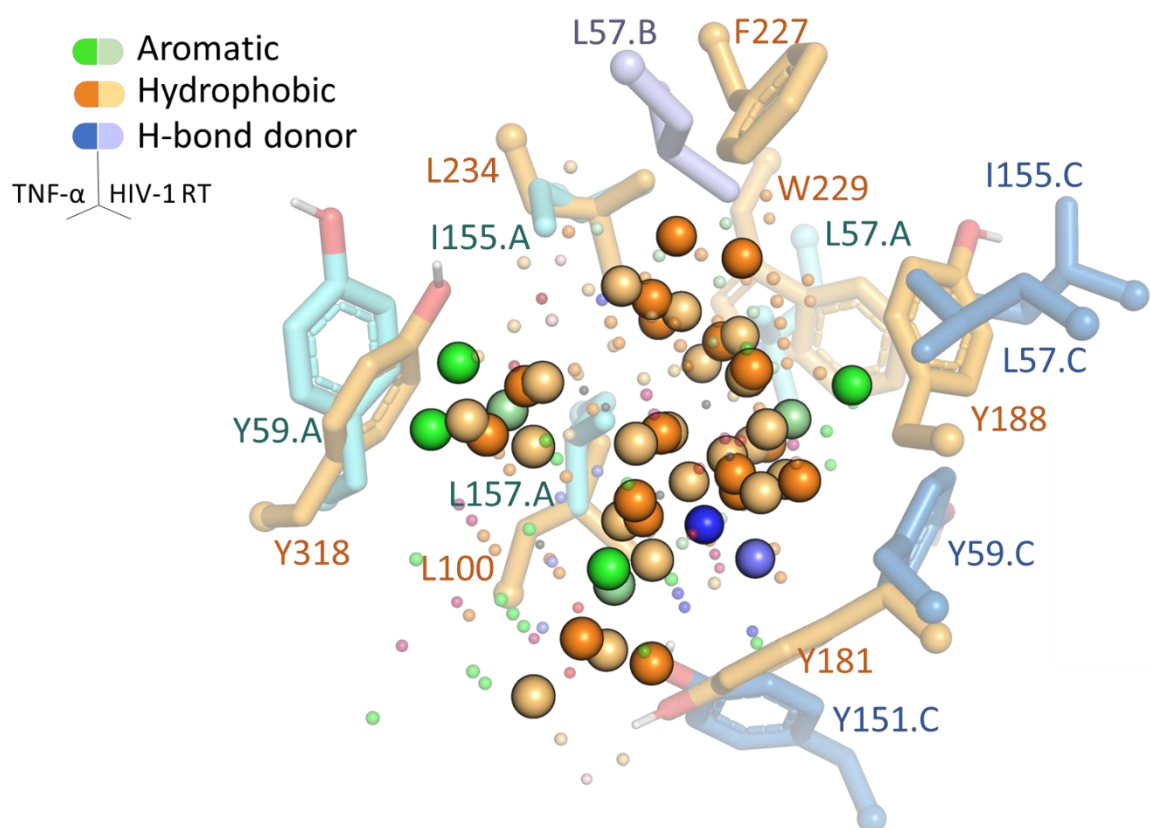
**Figure S2.** ProCare alignment of efavirenz main fragment subpocket (PDB code: 1FKO, HET code: EFZ) onto TNF-α trimer pocket (PDB code: 6OOZ, HET code: A6Y). Matched pharmacophoric points are depicted with dark-colored (TNF-α) and light-colored (HIV-1 RT) large spheres. Small spheres represent pharmacophoric points not considered by the best ProCare alignment.
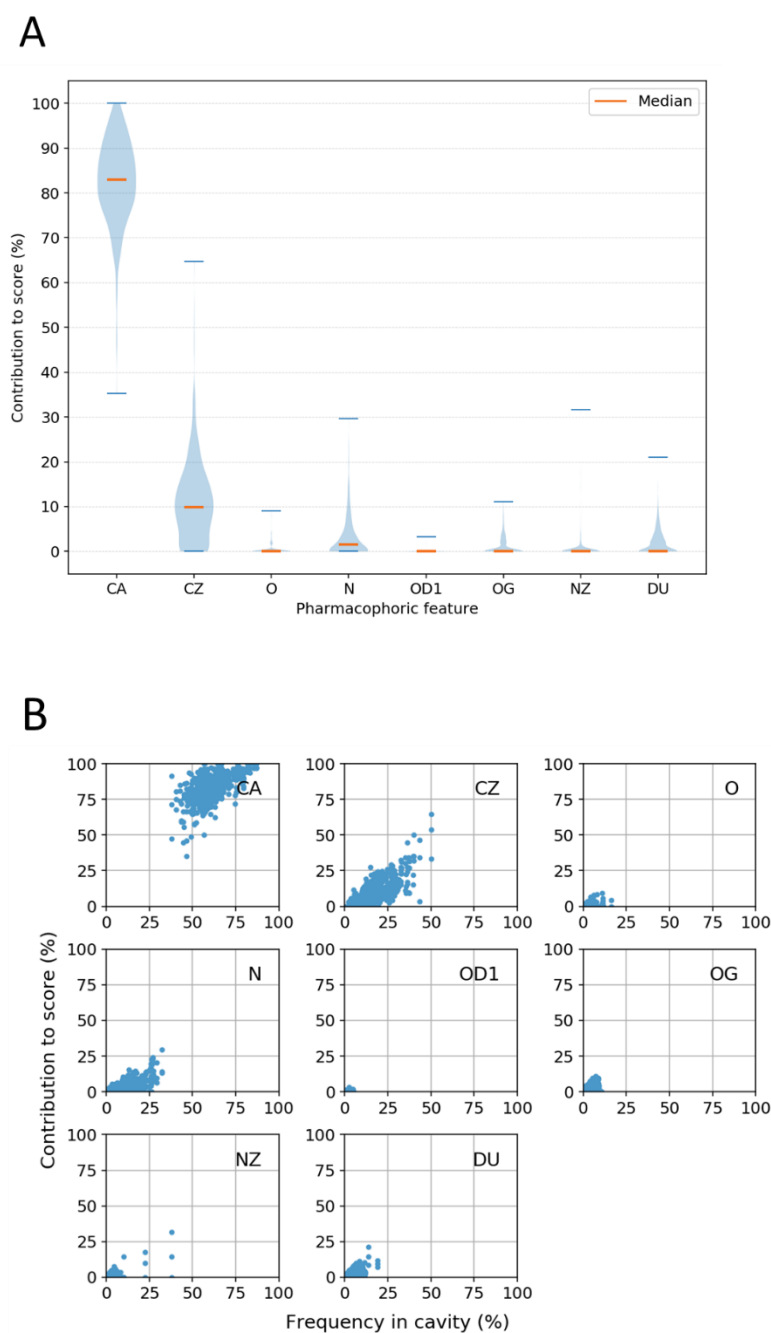
A



B



**Figure S3.** Contributions of the eight pharmacophoric features to the ProCare similarity score between HIV-1 RT (PDB ID 1FKO) and TNF-α (PDB ID 6OOZ). CA: hydrophobic, CZ: aromatic, O: h-bond acceptor, N: h-bond donor, OD1: negative, OG: h-bond acceptor and donor, NZ: positive, DU: dummy. **(A)** Aromatic pharmacophoric features are contributing more to the similarity between TNF-α trimer pockets (N=3) and HIV-1 RT subpockets (N = 195) although they are less frequent in the HIV-1 RT subpockets than hydrophobic points **(B).**
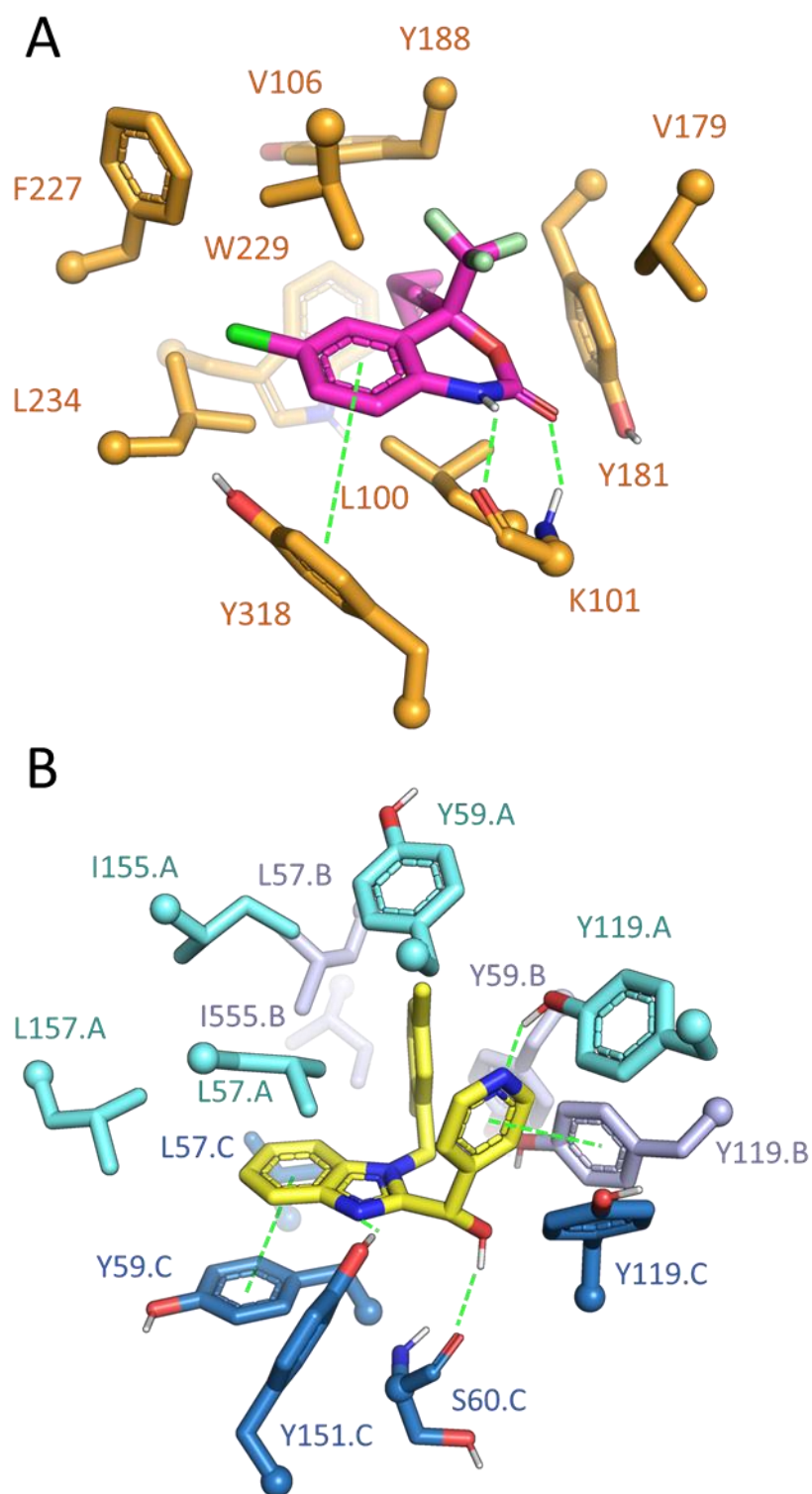
**Figure S4.** Non-covalent interactions between **(A)** efavirenz and HIV-1 RT (PDB ID 1FKO, HET code: EFZ) and **(B)** UCB-5307 and TNF-α trimer (PDB ID 6OOZ, HET code: A6Y).
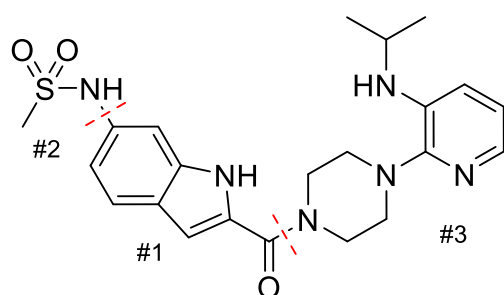
**Figure S5.** Manual fragmentation of delavirdine (PDB code: 1KLM, HET code: SPP) in three fragments (#1 to #3).

**Table S1.** sc-PDB subpockets sorted by decreased ProCare similarity to the inner cavity of human TNF-α (PDB code: 6OOY)

| Cavity ID[a] | Protein name (Uniprot) | Score[b] | Rank |
|---|---|---|---|
| 4f9y_GG5_1_3 | mitogen-activated protein kinase 14 | 0.7679 | 1 |
| 1mz9_VDY_2_1 | cartilage oligomeric matrix protein | 0.7673 | 2 |
| 1oz1_FPH_1_2 | mitogen-activated protein kinase 14 | 0.7634 | 3 |
| 3g9n_J88_1_2 | mitogen-activated protein kinase 10 | 0.7527 | 4 |
| 4fyn_0VE_1_2 | tyrosine-protein kinase syk | 0.7504 | 5 |
| 4kb8_1QN_3_1 | casein kinase i isoform delta | 0.7358 | 6 |
| 3k3j_I46_2_1 | mitogen-activated protein kinase 14 | 0.7338 | 7 |
| 2xj1_XJ1_1_2 | serine/threonine-protein kinase pim-1 | 0.7303 | 8 |
| 4tuv_CPZ_1_1 | cytochrome p450 119 | 0.7303 | 9 |
| 1mr9_ACO_3_2 | streptogramin a acetyltransferase | 0.7301 | 10 |
| 2fze_APR_1_1 | alcohol dehydrogenase class-3 | 0.728 | 11 |
| 4iwc_1GV_2_1 | estrogen receptor | 0.726 | 12 |
| 1ncr_W11_1_2 | human rhinovirus 16 | 0.7256 | 13 |
| 2ykm_YKN_1_2 | HIV-1 reverse transcriptase | 0.7242 | 14 |
| 4a7c_E46_1_1 | serine/threonine-protein kinase pim-1 | 0.7234 | 15 |
| 4wm7_W11_1_2 | capsid protein vp0 | 0.7216 | 16 |
| 3q2a_PAB_2_1 | toluene-4-monooxygenase system, hydroxylase component subunit alpha | 0.7209 | 17 |
| 3bqr_4RB_1_1 | death-associated protein kinase 3 | 0.7193 | 18 |
| 4ccb_OFG_1_4 | alk tyrosine kinase receptor | 0.7191 | 19 |
| 3fc1_52P_1_1 | mitogen-activated protein kinase 14 | 0.7137 | 20 |
| 2uzt_SS3_1_2 | camp-dependent protein kinase catalytic subunit alpha | 0.7126 | 21 |
| 4ewq_MWL_2_3 | mitogen-activated protein kinase 14 | 0.7122 | 22 |
| 4zhx_C1V_1_2 | 5'-amp-activated protein kinase catalytic subunit alpha-2 | 0.7122 | 23 |
| 4ogi_R78_2_2 | bromodomain-containing protein 4 | 0.7115 | 24 |
| 3roc_29A_1_2 | mitogen-activated protein kinase 14 | 0.7093 | 25 |
| 2l1r_SXK_1_2 | troponin c, slow skeletal and cardiac muscles | 0.7071 | 26 |

177

| | | | |
|---|---|---|---|
| 4h98_14Q_2_3 | dihydrofolate reductase | 0.7059 | 27 |
| 1lwc_NVP_1_1 | HIV-1 reverse transcriptase | 0.7024 | 28 |
| 3iw7_IPK_1_1 | mitogen-activated protein kinase 14 | 0.7024 | 29 |
| 2prh_238_2_2 | dihydroorotate dehydrogenase (quinone), mitochondrial | 0.7023 | 30 |
| 3hll_I45_1_1 | mitogen-activated protein kinase 14 | 0.7023 | 31 |
| 2vg7_NNI_1_1 | HIV-1 reverse transcriptase | 0.7022 | 32 |
| 1lwc_NVP_1_2 | HIV-1 reverse transcriptase | 0.6985 | 33 |
| 1ouk_084_1_3 | mitogen-activated protein kinase 14 | 0.6981 | 34 |
| 2xiy_XIY_1_1 | serine/threonine-protein kinase pim-1 | 0.6981 | 35 |
| 4k33_ACP_1_2 | fibroblast growth factor receptor 3 | 0.698 | 36 |
| 3umw_596_1_2 | serine/threonine-protein kinase pim-1 | 0.6978 | 37 |
| 4nkw_PLO_4_1 | steroid 17-alpha-hydroxylase/17,20 lyase | 0.6975 | 38 |
| 2iok_IOK_1_3 | estrogen receptor | 0.6969 | 39 |
| 2qd9_LGF_1_2 | mitogen-activated protein kinase 14 | 0.6967 | 40 |
| 2hnd_NVP_1_1 | HIV-1 reverse transcriptase | 0.6964 | 41 |
| 5av4_GEN_1_2 | death-associated protein kinase 1 | 0.6964 | 42 |
| 4zth_GG5_1_2 | mitogen-activated protein kinase 14 | 0.6961 | 43 |
| 3vs2_VSB_2_3 | tyrosine-protein kinase hck | 0.6951 | 44 |
| 4r3c_GG5_1_3 | mitogen-activated protein kinase 14 | 0.6951 | 45 |
| 4q5h_ANP_1_2 | protein kinase ospg | 0.6939 | 46 |
| 5awm_ANP_1_2 | stress-activated protein kinase jnk | 0.6936 | 47 |
| 4anq_VGH_1_2 | alk tyrosine kinase receptor | 0.6933 | 48 |
| 1mp0_NAD_2_2 | alcohol dehydrogenase class-3 | 0.6922 | 49 |
| 4anv_751_1_1 | phosphatidylinositol 4,5-bisphosphate 3-kinase catalytic subunit gamma isoform | 0.6918 | 50 |
| 2bxo_OPB_1_2 | albumin | 0.6914 | 51 |
| 1nav_IH5_1_1 | thyroid hormone receptor alpha | 0.6905 | 52 |
| 3fyw_XCF_1_3 | dihydrofolate reductase | 0.6903 | 53 |
| 3lp0_NVP_2_1 | HIV-1 reverse transcriptase | 0.6897 | 54 |
| 1adc_PAD_1_2 | alcohol dehydrogenase e chain | 0.6892 | 55 |
| 2yis_I46_2_1 | mitogen-activated protein kinase 14 | 0.689 | 56 |
| 4uun_NAI_2_2 | l-lactate dehydrogenase | 0.6882 | 57 |
| 4fl2_ANP_1_2 | tyrosine-protein kinase syk | 0.6876 | 58 |
| 4kb8_1QN_3_2 | casein kinase i isoform delta | 0.6874 | 59 |
| 1pjc_NAD_1_2 | alanine dehydrogenase | 0.687 | 60 |
| 3hl7_I47_1_2 | mitogen-activated protein kinase 14 | 0.687 | 61 |
| 4l0q_NAD_1_2 | alcohol dehydrogenase class-3 | 0.6864 | 62 |
| 4loo_SB4_1_2 | mitogen-activated protein kinase 14 | 0.6864 | 63 |
| 3go6_ADP_1_2 | ribokinase | 0.6833 | 64 |
| 3wze_BAX_1_1 | vascular endothelial growth factor receptor 2 | 0.6833 | 65 |
| 2xiz_XIZ_1_1 | serine/threonine-protein kinase pim-1 | 0.6822 | 66 |
| 3rsr_N5P_1_1 | ribonucleoside-diphosphate reductase large chain 1 | 0.6816 | 67 |
| 1jkl_ANP_1_2 | death-associated protein kinase 1 | 0.6807 | 68 |
| 3bxz_ADP_2_2 | protein translocase subunit seca | 0.6807 | 69 |

| | | | |
|---|---|---|---|
| 3uyt_0CK_3_3 | casein kinase i isoform delta | 0.6807 | 70 |
| 1qiw_DPD_2_2 | calmodulin | 0.6806 | 71 |
| 4mbl_26L_1_2 | serine/threonine-protein kinase pim-1 | 0.6802 | 72 |
| 4kbc_1QJ_1_1 | casein kinase i isoform delta | 0.6797 | 73 |
| 1vru_AAP_1_2 | HIV-1 reverse transcriptase | 0.6791 | 74 |
| | | | |
| 1f0y_NAD_2_2 | hydroxyacyl-coenzyme a dehydrogenase, mitochondrial | 0.6786 | 75 |
| 4hds_IPH_1_1 | n(1)-alpha-phosphoribosyltransferase | 0.6786 | 76 |
| 4mzu_COA_22_2 | wxcm-like protein | 0.6786 | 77 |
| 3gc7_B45_1_1 | mitogen-activated protein kinase 14 | 0.6782 | 78 |
| 3qf9_NM8_1_2 | serine/threonine-protein kinase pim-1 | 0.678 | 79 |
| 3rr3_FLR_3_2 | prostaglandin g/h synthase 2 | 0.678 | 80 |
| 4ix6_ADP_1_2 | protein kinase domain-containing protein | 0.6771 | 81 |
| 1pf9_ADP_2_2 | 60 kda chaperonin | 0.6761 | 82 |
| 2bu7_TF3_1_2 | [pyruvate dehydrogenase (acetyl-transferring)] kinase isozyme 2, mitochondrial | 0.6761 | 83 |
| 5ani_ES4_1_1 | cyclin-dependent kinase 2 | 0.6761 | 84 |
| 1tuv_VK3_1_1 | probable quinol monooxygenase ygin | 0.6754 | 85 |
| 2zm1_KSF_1_2 | tyrosine-protein kinase lck | 0.6752 | 86 |
| 3bea_IXH_1_3 | angiopoietin-1 receptor | 0.6752 | 87 |
| 4hur_ACO_3_1 | virginiamycin a acetyltransferase | 0.6752 | 88 |
| 5dr2_ATP_1_2 | aurora kinase a | 0.6752 | 89 |
| 5dgz_L20_1_1 | serine/threonine-protein kinase pim-1 | 0.674 | 90 |
| 4iu7_1GM_1_1 | estrogen receptor | 0.6736 | 91 |
| 2pnu_ENM_1_1 | androgen receptor | 0.6736 | 92 |
| 3hvc_GG5_1_3 | mitogen-activated protein kinase 14 | 0.6736 | 93 |
| 3wwm_ADP_1_2 | [lysw]-aminoadipate kinase | 0.6736 | 94 |
| 3znr_NU9_1_3 | histone deacetylase 7 | 0.6736 | 95 |
| 3q7d_NPX_1_1 | prostaglandin g/h synthase 2 | 0.6726 | 96 |
| 3fkn_FKN_1_1 | mitogen-activated protein kinase 14 | 0.6723 | 97 |
| 4dgm_AGI_1_1 | casein kinase ii subunit alpha | 0.6721 | 98 |
| 4i5h_G17_1_1 | mitogen-activated protein kinase 1 | 0.6721 | 99 |
| 3t9i_3T9_1_1 | serine/threonine-protein kinase pim-1 | 0.6715 | 100 |
| ... | ... | ... | … |
| 5je3_SAH_2_2 | class I sam-dependent methyltransferase | 0.0000 | 31570 |

[a] Cavity ID (PDB_HET_C_M) is inferred from the cognate target PDB identifier (PDB), the corresponding ligand chemical component (HET), the target cavity identifier (C), and the fragment number (N).

[b] ProCare similarity score. A value above 0.47 corresponds to statistically significant similarity (p-value < 0.05) between the pair of pockets under investigation [17].

**Table S2.** PDB entries describing non-nucleoside inhibitors bound to HIV-1 reverse transcriptase

The list of the 122 HIV-RT entries is available as supporting information at:

https://doi.org/10.1186/s13321-021-00567-3: 13321_2021_567_MOESM1_ESM.pdf

**Table S3.** Comparison of delavirdine subpockets, resulting from manual fragmentation, with TNF-α trimer pockets.

| PDB/HET code | Fragment # | TNF-α PDB entry | ProCare score | Rank[a] |
|---|---|---|---|---|
| 1KLM/SPP | 1 | 6OOY | 0.328 | 588 |
| 1KLM/SPP | 2 | 6OOY | 0.599 | 113 |
| 1KLM/SPP | 3 | 6OOY | 0.283 | 593 |
| 1KLM/SPP | 1 | 6OOZ | 0.361 | 581 |
| 1KLM/SPP | 2 | 6OOZ | 0.570 | 174 |
| 1KLM/SPP | 3 | 6OOZ | 0.416 | 549 |
| 1KLM/SPP | 1 | 6OP0 | 0.342 | 586 |
| 1KLM/SPP | 2 | 6OP0 | 0.534 | 272 |
| 1KLM/SPP | 3 | 6OP0 | 0.130 | 594 |

[a] Rank after adding delavirdine fragment scores to the ProCare screening results that yielded a total of 594 pairwise scores.

**Table S4.** Dissociation constant ($K_D$) of three HIV-1 RT inhibitor binding to human soluble TNF-α, according to MST experimental conditions.

| HIV-1 RT Inhibitor | TNF concentration nM | DMSO concentration (%) in MST buffer | Tween-20 concentration in MST buffer | Incubation time min | MST power % | $K_D \pm CI$[a] μM |
|---|---|---|---|---|---|---|
| efavirenz | 220 | 5.0 | 0.05 | 5 | 40 | $45 \pm 9$ |
| efavirenz | 220 | 5.0 | 0.05 | 5 | 80 | $47 \pm 12$ |
| efavirenz | 220 | 5.0 | 0.01 | 5 | 80 | $26 \pm 5$ |
| efavirenz | 220 | 5.0 | 0.01 | 30 | 80 | $27 \pm 6$ |
| efavirenz | 220 | 2.5 | 0.01 | 30 | 80 | $11 \pm 3$ |
| efavirenz | 170 | 1.3 | 0.01 | 20 | 40 | $17 \pm 5$ |
| efavirenz | 170 | 1.3 | 0.01 | 20 | 80 | $24 \pm 4$ |
| efavirenz | 340 | 1.3 | 0.01 | 15 | 40 | $24 \pm 8$[b] |
| efavirenz | 340 | 1.3 | 0.01 | 15 | 80 | $38 \pm 5$ |
| delavirdine | 220 | 5.0 | 0.05 | 5 | 40 | $203 \pm 143$ |

| | | | | | | |
|---|---|---|---|---|---|---|
| delavirdine | 220 | 5.0 | 0.01 | 5 | 40 | 84 ± 57 |
| delavirdine | 170 | 1.3 | 0.01 | 5 | 40 | 90 ± 50 |
| delavirdine | 170 | 1.3 | 0.01 | 60 | 20 | 81 ± 31 |
| delavirdine | 170 | 1.3 | 0.01 | 15 | 20 | 69 ± 23 |
| delavirdine | 340 | 1.3 | 0.01 | 15 | 20 | 39 ± 9[b] |
| delavirdine | 340 | 1.3 | 0.01 | 120 | 20 | 56 ± 20 |
| | | | | | | |
| nevirapine | 220 | 5.0 | 0.05 | 5 | 40 | no signal |
| nevirapine | 340 | 1.3 | 0.01 | 20 | 40 | no signal |

[a] CI: 68.3% confidence interval

[b] MST measure with the highest signal to noise ratio

**Table S5.** CHEMBL entries describing HIV-1 RT non-nucleoside inhibitors.

Available at https://github.com/kimeguida/ProCare_TNF

**Table S6.** Customized rules for OpenEye Filter ionization.

MIN_MOLWT      1        "Minimum molecular weight"
MAX_MOLWT     15000       "Maximum molecular weight"
MIN_NUM_HVY   0        "Minimum number of heavy atoms"
MAX_NUM_HVY   2500       "Maximum number of heavy atoms"
MIN_RING_SYS   0     "Minumum number of ring systems"
MAX_RING_SYS   50     "Maximum number of ring systems"
MIN_RING_SIZE   0     "Minimum atoms in any ring system"
MAX_RING_SIZE   200     "Maximum atoms in any ring system"
MIN_CON_NON_RING   0     "Minimum number of connected non-ring atoms"
MAX_CON_NON_RING   190     "Maximum number of connected non-ring atoms"
MIN_FCNGRP      0     "Minimum number of functional groups"
MAX_FCNGRP      70     "Maximum number of functional groups"
MIN_UNBRANCHED   0     "Minimum number of connected unbranched non-ring atoms"
MAX_UNBRANCHED   130     "Maximum number of connected unbranched non-ring atoms"
MIN_CARBONS      0     "Minimum number of carbons"
MAX_CARBONS      410     "Maximum number of carbons"
MIN_HETEROATOMS   0     "Minimum number of heteroatoms"
MAX_HETEROATOMS   140     "Maximum number of heteroatoms"
MIN_Het_C_Ratio   0.04     "Minimum heteroatom to carbon ratio"
MAX_Het_C_Ratio   40.0     "Maximum heteroatom to carbon ratio"
MIN_HALIDE_FRACTION     0.0     "Minimum Halide Fraction"
MAX_HALIDE_FRACTION     0.99     "Maximum Halide Fraction"
#count ring degrees of freedom = (#BondsInRing) - 4 - (RigidBondsInRing) - (BondsSharedWithOtherRings)
#must be >= 0, from JCAMD 14:251-265,2000.
ADJUST_ROT_FOR_RING     true     "BOOLEAN for whether to estimate degrees of freedom in rings"
MIN_ROT_BONDS   0     "Minimum number of rotatable bonds"
MAX_ROT_BONDS   160     "Maximum number of rotatable bonds"
MIN_RIGID_BONDS   0     "Minimum number of rigid bonds"

```
MAX_RIGID_BONDS    550      "Maximum number of rigid bonds"
MIN_HBOND_DONORS  0      "Minimum number of hydrogen-bond donors"
MAX_HBOND_DONORS  90      "Maximum number of hydrogen-bond donors"
MIN_HBOND_ACCEPTORS  0      "Minimum number of hydrogen-bond acceptors"
MAX_HBOND_ACCEPTORS  130    "Maximum number of hydrogen-bond acceptors"
MIN_LIPINSKI_DONORS  0      "Minimum number of hydrogens on O & N atoms"
MAX_LIPINSKI_DONORS  60      "Maximum number of hydrogens on O & N atoms"
MIN_LIPINSKI_ACCEPTORS  0      "Minimum number of oxygen & nitrogen atoms"
MAX_LIPINSKI_ACCEPTORS  140      "Maximum number of oxygen & nitrogen atoms"
MIN_COUNT_FORMAL_CRG    0      "Minimum number formal charges"
MAX_COUNT_FORMAL_CRG    40      "Maximum number of formal charges"
MIN_SUM_FORMAL_CRG    -20      "Minimum sum of formal charges"
MAX_SUM_FORMAL_CRG    20       "Maximum sum of formal charges"
MIN_CHIRAL_CENTERS    0    "Minimum chiral centers"
MAX_CHIRAL_CENTERS    100    "Maximum chiral centers"
MIN_XLOGP      -30.0      "Minimum XLogP"
MAX_XLOGP      60.85      "Maximum XLogP"
#choices are insoluble<poorly<moderately<soluble<very<highly
MIN_SOLUBILITY      insoluble "Minimum solubility"
PSA_USE_SandP    false    "Count S and P as polar atoms"
MIN_2D_PSA      0.0      "Minimum 2-Dimensional (SMILES) Polar Surface Area"
MAX_2D_PSA      2050.0    "Maximum 2-Dimensional (SMILES) Polar Surface Area"
AGGREGATORS    false      "Eliminate known aggregators"
PRED_AGG      false    "Eliminate predicted aggregators"
#secondary filters (based on multiple primary filters)
GSK_VEBER      false      "PSA>140 or >10 rot bonds"
MAX_LIPINSKI   5      "Maximum number of Lipinski violations"
MIN_ABS      0.01      "Minimum probability F>10% in rats"
PHARMACOPIA    false    "LogP > 5.88 or PSA > 131.6"
ALLOWED_ELEMENTS  H,C,N,O,F,P,S,Cl,Br,I,B
ELIMINATE_METALS Sc,Ti,V,Cr,Mn,Fe,Co,Ni,Cu,Zn,Y,Zr,Nb,Mo,Tc,Ru,Rh,Pd,Ag,Cd
```
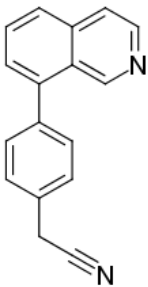
## 3.2. Scope and critical evaluation of the study
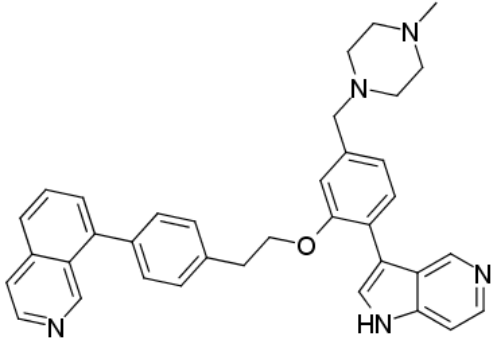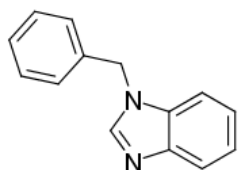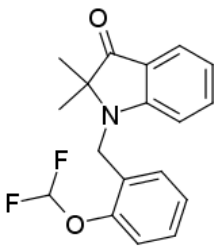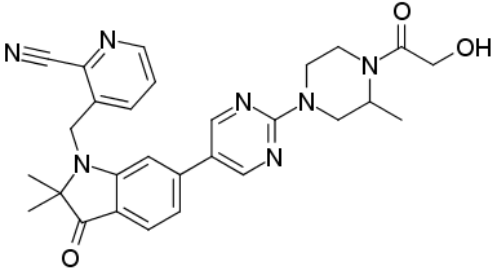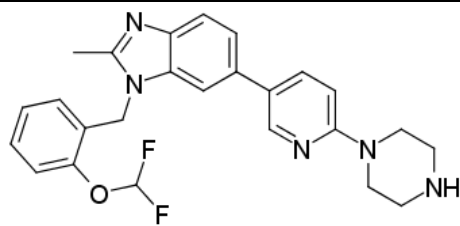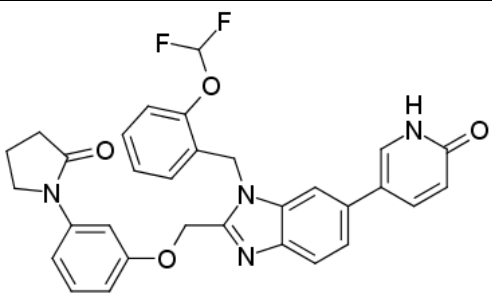
In **Chapter 2**, we presented ProCare and its possible applications. Whereas benchmarking via retrospective studies is necessary to validate an approach, it is important to delineate the actual applicability in real-life cases—what the method was developed for in the first place. This study aimed at evaluating whether ProCare can predict similarity between structurally and functionally remote pockets and transfer a binding fragment from one pocket to the other. The choice of the target, human tumor necrosis factor-alpha (TNF-α),[1,2] was motivated by its unavailability in the sc-PDB database and its importance in human diseases.
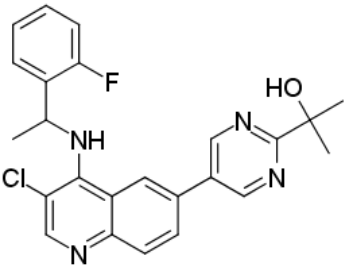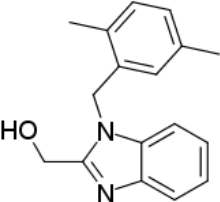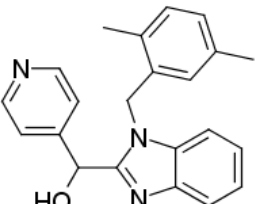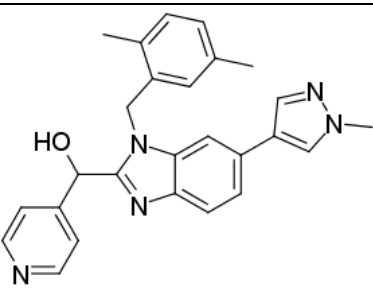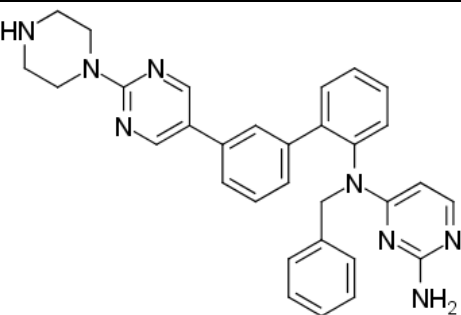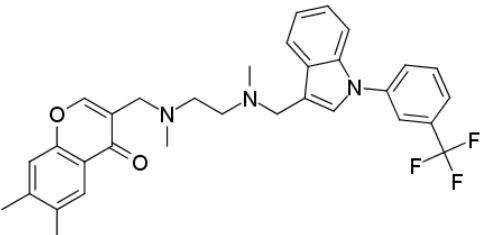
TNF-α is a pro-inflammatory cytokine, released by the immune system for infection signaling. It binds to and activate one of its two receptors, TNF receptor 1 (TNFR1).[1] Targeting TNF-α has been a successful strategy to treat autoimmune diseases such as rheumatoid arthritis, psoriasis, or inflammatory Bowel disease such as Crohn's disease or ulcerative colitis. Approved and commercialized inhibitors are monoclonal anti-human TNF-α antibodies (e.g. Infliximab) or chimeric proteins mimicking TNFR (e.g. Etanercept).[2] Due to the challenges of biologics regarding administration, immunogenicity and other side effects, drug design efforts are made to develop small molecule inhibitors.[3] Among the strategies, some small molecules in the clinical phases disrupt TNF-α pathways (e.g. p38 inhibitors). Others directly target the trimeric interface (**Table 3.1**). We should recall that, among published inhibitors accessible in ChEMBL (https://www.ebi.ac.uk/chembl) for instance, not all were co-crystallized with TNF-α or released in the Protein Data Bank (PDB).[4–6]

Out of the 35 TNF-α homotrimer, dimer and monomer structures in the PDB, one third were released in the last two years, after the generation of the hypothesis leading to this work. Some of these structures are complexes with small molecules inducing some asymmetric shape of the trimeric TNF-α and disrupting its downstream effects. The most recent asymmetric trimeric complexes (PDB ID 6OOY, 6OOZ, 6OP0) at the time of the study were selected.

**Table 3.1.** Small molecules binding TNF-α trimer interface and available in the PDB (on 06/27/2022).

| PDB ID (Resol.) | Ligand | Binding Affinity in nM (assay, measure) | Release date |
|---|---|---|---|
| 6X81 (2.81 Å) |  | 2 700 (SPR $K_D$)[7] | 2021-01-13 |

| | | | |
|---|---|---|---|
| 6X82 (2.75 Å) |  | 2.4 (SPR $K_D$)[7] | 2021-01-13 |
| 6X83 (2.83 Å) |  | 300 000 (SPR $K_D$)[7] | 2021-01-13 |
| 6X85 (2.85 Å) |  | 19 000 (SPR $K_D$)[7] | 2021-01-13 |
| 6X86 (2.93 Å) |  | 7.3 (SPR $K_D$)[7] | 2021-01-13 |
| 7KP9 (2.15 Å) |  | N/A | 2021-01-13 |
| 7KPA (2.3 Å) 7KPB (3 Å) |  | 8.1 (SPR $K_D$)[8] | 2021-01-13 |

| | | | |
|---|---|---|---|
| 7JRA<br>(2.1 Å) | | 47<br>(TNF-α HTRF<br>$IC_{50}$)[9] | 2020-12-09 |
| 6OOY<br>(2.5 Å) | | 22 000<br>(SPR $K_D$)[10] | 2019-12-25 |
| 6OOZ<br>(2.8 Å) | | 9<br>(SPR $K_D$)[10] | 2019-12-25 |
| 6OP0<br>(2.55 Å) | | 13.8<br>(SPR $K_D$)[10] | 2019-12-25 |
| 5MU8<br>(3 Å) | | 1 200<br>(TNF-α-TNFR1<br>HTRF $IC_{50}$)[11] | 2017-03-29 |
| 2AZ5<br>(2.1 Å) | | 22 000<br>(TNF-α-TNFR1<br>HTRF $IC_{50}$)[12] | 2005-11-29 |

TNF-α pocket is highly hydrophobic (55 % of IChem cavity points are hydrophobic) and might falsely match with other hydrophobic pockets. For example, high similarity was predicted for estrogen receptor subpockets without further investigations. Interestingly, ProCare aligned four polar features as well, associated to a triangle of distinct TNF-α/HIVRT protein residues, hence excluding the possibility of unspecific matches. Subpocket-based alignment of corresponding HIVRT fragments (derived from nevirapine and efavirenz) superposed to docking solutions encouraged us to continue the study whereas the nevirapine butterfly shape nicely matched the benzimidazole ligands of TNF-α. As discussed in the previous chapters, there is no experimental measure and not one definition of pocket similarity. Herein, 'similar' subpockets means 'capable of binding the same molecules, by exhibiting some features that can result in favorable energetic contributions. As binding occurs due to contributions other than enthalpy, absence of experimental binding data would have resulted in limited to no conclusions in our experimental design. Other factors are the assay settings or solubility problems. Contrarily, identifying at least one example is enough to prove the above proposition as it is a matter of possibility instead of systematic observation. Accordingly, we made no effort to evaluate TNF-α inhibitors on HIVRT.

Prior to the direct binding microscale thermophoresis (MST) experiments, efavirenz and delavirdine showed to interact *in vitro* with TNF-α in differential scanning fluorimetry (nanoDSF) assays while the nevirapine hypothesis failed. We note that intact (and not the corresponding fragments) efavirenz and nevirapine were tested whereas the hypothesis was derived from comparing their fragments subpockets. The additional moieties might perturb predicted interactions or rather add positive contribution to the binding. Nonetheless, a global *a posteriori* comparison with whole HIVRT pocket enclosing efavirenz yielded scores above the similarity threshold, albeit with a different alignment. Several attempts to access the SPR assay and have a basis for direct comparison with UCB TNF-α inhibitors[10] by contacting the authors remained unsuccessful.

Given the importance of TNF-α, we further assessed the effects of the three HIVRT inhibitors on the ability of TNF-α to binds to its receptor TNFR1. While the detected signals were consistent with the MST results (signal for delavirdine and efavirenz, no effect with nevirapine), they were weak ($< 30\%$ inhibition at $100\,\mu M$, **Figure 3.1**). Further investigations with or without crystal structure of complexes, which are out of the scope of this thesis, would provide more insights.
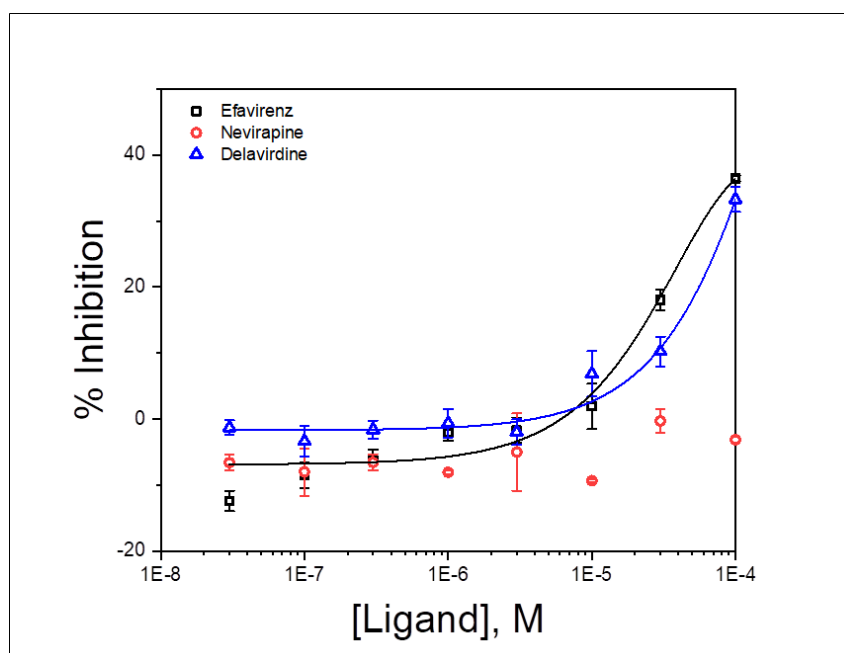
**Figure 3.1.** Inhibition of 0.1 nM [$^{125}$I]-TNF-α binding to human TNF receptor type 1 (TNFR1) in U-937 cells,[13] by three HIV-1 reverse transcriptase inhibitors (Eurofins Discovery assay #76). Results are mean ± SEM of two experiments.

**What did we learn about the method?** Visualization of aligned features in the protein pockets provides additional insights. When prioritizing pocket matches, attention must be paid on the size and feature composition of the subpockets to decrease the chances of false positives. Because ProCare score was made symmetrical and adapted to compare pockets of different sizes, smallest pockets would tend to have higher scores when the latter are highly hydrophobic. Additional experiments such as docking or molecular dynamic simulations might be useful to provide different perspectives.

## 3.3.   References

1.   Kalliolias, G. D.; Ivashkiv, L. B. TNF Biology, Pathogenic Mechanisms and Emerging Therapeutic Strategies. *Nat. Rev. Rheumatol.* **2016**, *12*, 49–62.

2.   Palladino, M. A.; Bahjat, F. R.; Theodorakis, E. A.; Moldawer, L. L. Anti-TNF-α Therapies: The next Generation. *Nat. Rev. Drug Discov.* **2003**, *2*, 736–746.

3.   Dömling, A.; Li, X. TNF-α: The Shape of Small Molecules to Come? *Drug Discov. Today* **2022**, *27*, 3–7.

4.   Mouhsine, H.; Guillemain, H.; Moreau, G.; Fourati, N.; Zerrouki, C.; Baron, B.; Desallais, L.; Gizzi, P.; Ben Nasr, N.; Perrier, J.; Ratsimandresy, R.; Spadoni, J. L.; Do, H.; England, P.;

Montes, M.; Zagury, J. F. Identification of an in Vivo Orally Active Dual-Binding Protein-Protein Interaction Inhibitor Targeting TNFα through Combined in Silico/in Vitro/in Vivo Screening. *Sci. Rep.* **2017**, *7*, 1–10.

5.    Chen, S.; Feng, Z.; Wang, Y.; Ma, S.; Hu, Z.; Yang, P.; Chai, Y.; Xie, X. Discovery of Novel Ligands for TNF-α and TNF Receptor-1 through Structure-Based Virtual Screening and Biological Assay. *J. Chem. Inf. Model.* **2017**, *57*, 1101–1111.

6.    Sun, W.; Wu, Y.; Zheng, M.; Yang, Y.; Liu, Y.; Wu, C.; Zhou, Y.; Zhang, Y.; Chen, L.; Li, H. Discovery of an Orally Active Small-Molecule Tumor Necrosis Factor-α Inhibitor. *J. Med. Chem.* **2020**, *63*, 8146–8156.

7.    Dietrich, J. D.; Longenecker, K. L.; Wilson, N. S.; Goess, C.; Panchal, S. C.; Swann, S. L.; Petros, A. M.; Hobson, A. D.; Ihle, D.; Song, D.; Richardson, P.; Comess, K. M.; Cox, P. B.; Dombrowski, A.; Sarris, K.; Donnelly-Roberts, D. L.; Duignan, D. B.; Gomtsyan, A.; Jung, P.; Krueger, A. C.; Mathieu, S.; McClure, A.; Stoll, V. S.; Wetter, J.; Mankovich, J. A.; Hajduk, P. J.; Vasudevan, A.; Stoffel, R. H.; Sun, C. Development of Orally Efficacious Allosteric Inhibitors of TNFα via Fragment-Based Drug Design. *J. Med. Chem.* **2021**, *64*, 417–429.

8.    Lightwood, D. J.; Munro, R. J.; Porter, J.; McMillan, D.; Carrington, B.; Turner, A.; Scott-Tucker, A.; Hickford, E. S.; Schmidt, A.; Fox, D.; Maloney, A.; Ceska, T.; Bourne, T.; O'Connell, J.; Lawson, A. D. G. A Conformation-Selective Monoclonal Antibody against a Small Molecule-Stabilised Signalling-Deficient Form of TNF. *Nat. Commun.* **2021**, *12*.

9.    Xiao, H.-Y. Y.; Li, N.; Duan, J. J. W.; Jiang, B.; Lu, Z.; Ngu, K.; Tino, J.; Kopcho, L. M.; Lu, H.; Chen, J.; Tebben, A. J.; Sheriff, S.; Chang, C. Y.; Yanchunas, J.; Calambur, D.; Gao, M.; Shuster, D. J.; Susulic, V.; Xie, J. H.; Guarino, V. R.; Wu, D.-R. R.; Gregor, K. R.; Goldstine, C. B.; Hynes, J.; Macor, J. E.; Salter-Cid, L.; Burke, J. R.; Shaw, P. J.; Dhar, T. G. M. M. Biologic-like in Vivo Efficacy with Small Molecule Inhibitors of TNFα Identified Using Scaffold Hopping and Structure-Based Drug Design Approaches. *J. Med. Chem.* **2020**, *63*, 15050–15071.

10.   O'Connell, J.; Porter, J.; Kroeplien, B.; Norman, T.; Rapecki, S.; Davis, R.; McMillan, D.; Arakaki, T.; Burgin, A.; Fox III, D.; Ceska, T.; Lecomte, F.; Maloney, A.; Vugler, A.; Carrington, B.; Cossins, B. P.; Bourne, T.; Lawson, A. Small Molecules That Inhibit TNF Signalling by Stabilising an Asymmetric Form of the Trimer. *Nat. Commun.* **2019**, *10*, 5795.

11.   Blevitt, J. M.; Hack, M. D.; Herman, K. L.; Jackson, P. F.; Krawczuk, P. J.; Lebsack, A. D.; Liu, A. X.; Mirzadegan, T.; Nelen, M. I.; Patrick, A. N.; Steinbacher, S.; Milla, M. E.; Lumb, K. J. Structural Basis of Small-Molecule Aggregate Induced Inhibition of a Protein–Protein Interaction. *J. Med. Chem.* **2017**, *60*, 3511–3517.

12.   He, M. M.; Smith, A. S.; Oslob, J. D.; Flanagan, W. M.; Braisted, A. C.; Whitty, A.; Cancilla, M. T.; Wang, J.; Lugovskoy, A. A.; Yoburn, J. C.; Fung, A. D.; Farrington, G.; Eldredge, J. K.; Day, E. S.; Cruz, L. A.; Cachero, T. G.; Miller, S. K.; Friedman, J. E.; Choong, I. C.; Cunningham, B. C. Small-Molecule Inhibition of TNF-α. *Science.* **2005**, *310*, 1022–1025.

13.   Brockhaus, M.; Schoenfeld, H. J.; Schlaeger, E. J.; Hunziker, W.; Lesslauer, W.; Loetscher, H. Identification of Two Types of Tumor Necrosis Factor Receptors on Human Cell Lines by Monoclonal Antibodies. *Proc. Natl. Acad. Sci.* **1990**, *87*, 3127–3131.

# CHAPTER 4

Pocket-focused library design

# 4.1. Scope and motivations

Compound library compilation is among the very first steps in a structure-based virtual screening campaign. Classically, lists of compounds from chemical vendors of choice are merged and filtered according to the project specifications. The size of such libraries can range from a few thousands to billions. Yet, a finite number of molecules are to be screened, and it is at best hoped that the library covers areas in the chemical space where potential hits are. This assumption is a necessary condition for the success of the screening, even before considering the performance of the methods to prioritize the best compounds. Among the possible strategies to efficiently explore the chemical space, the brute force approach consists of screening the largest possible diverse library, acknowledging the computing resources and prioritization efforts it demands.[1] Alternative ways use available information on the target, like pharmacophore of known ligands or deconstruction-recombination of inhibitors to build a target-focused library of smaller size, faster to screen and with expected higher hit rate.[2] We herein propose a semi-automatic workflow to generate molecule ideas for a given target by borrowing and linking bound fragments from available protein-bound ligands when their protein subpockets are locally similar to the target cavity. Accordingly, the POEM (Pocket-Oriented Elaboration of Molecule) computational workflow was developed. It is applicable even when only the apo structure of the target (without known binding ligand) is available.

The research questions raised by this methodology lays in combining two approximations: (i) the fragment still binds to the same subpocket as the corresponding substructure in the fully enumerated molecule; (ii) the fragment pose is not altered by linking to another fragment. Fragment-based drug design efforts demonstrated that linking two fragments does not always ensure conservation of their initial binding mode in the newly formed ligand; reversely, it has been shown experimentally that ligands deconstruction generates  fragments that do not necessarily bind to the same pocket as in the original ligands.[3] Therefore, POEM rationally relies on the proportion that escape these considerations. This study does not aim at answering the binding mode conservation questions in themselves but rather to propose a reasonable and useful tool to support hit discovery.

POEM was evaluated on three targets (**Table 4.1**): (1) cyclin-dependent kinase 8 (CDK8) for which ligands are known, allowing both retrospective and prospective studies, (2) the quinolinate synthase (NadA), a metalloprotein with Fe/S cluster in a narrow binding site for which no inhibitors are known and (3) the WD40 domain of leucine-rich repeat kinase 2 (LRRK2) whose pocket appears hardly druggable with no available ligands. With these applications, we aspire to validate and show the capacities and the limits of the approach.

**Table 4.1.** Characteristics of targets in POEM case-studies.

| Target | Pocket | Volume ($\text{Å}^3$)[a] | Pharmacological ligands | Prosthetic group |
|--------|--------|-------------|------------------------|------------------|
| CDK8 | catalytic | 891 | yes | No |
| NadA | catalytic | 213 | No | [4Fe-4S] |
| LRRK2 WDR | scaffold | 1411 | No | No |

[a] Pocket volume measured by the VolSite module of IChem v.5.2.9.

## 4.2. Target-focused library design by pocket-applied computer vision and fragment deep generative linking

This project was pursued as a collaboration with Pr M. Hibert who, together with his team, were investigating the protein CDK8 inhibitors.

### 4.2.1. Biological relevance of CDK8 in drug discovery and structural aspects

Cyclin-dependent kinase 8 (CDK8) is serine/threonine protein kinase (EC 2.7.11.22) which catalyzes the transfer of the gamma phosphate of ATP to hydroxyl groups of specific serine or threonine residues in peptide substrates. Many human diseases are associated with kinases as phosphorylation is a post-translational modification involved in several cellular processes. CDK8 belongs to the cyclin-dependent kinase (CDK) family whose members are conserved in eucaryotes and were originally known to play a role in the regulation of the cell cycle (CDK1, CDK2, CDK4 and CDK6). As part of the coactivator Mediator complex, CDK8 however regulates the transcription activities of RNA polymerase II, the multiprotein complex that transcribes deoxyribonucleic acid (DNA) into ribonucleic acid (RNA). Consequently, disrupting CDK8 functions would affect RNA polymerase II-dependent genes expression required for cell life. The CDK8 gene is located on chromosome 13q, a large portion of which was identified as overexpressed in colon cancers.[4–6] Studies have demonstrated that inhibition of CDK8 activity through CDK8 gene silencing or small molecule inhibitors decreased proliferation of β-catenin-dependent colon cancer cell lines.[4,7] CDK8 oncogenic role was also shown in other cancers (melanoma, gastric, breast, and ovarian cancers),[8–11] positioning CDK8 as a potential drug target.

Recently, a few selective CDK8 inhibitors have been positioned as potential therapeutics for the Diamond-Blackfan anemia[12,13] (DBA, ORPHA code: 124), a rare orphan disease. DBA is a ribosomopathy that affects the bone marrow which fails to produce mature and fully functional red blood cells in sufficient quantity. While the incidence is estimated to 1:150,000 in Europe, patients usually rely on red blood cells transfusion and/or corticosteroid treatments and are subjected to the related consequences (iron chelation therapy to prevent hemochromatosis, steroids adverse effects).[14] Although the underlying mechanisms are not well known and the potential drug targets are still to be fully validated,[15] some doors are open for exploration.

CDK8 is composed of 464 amino acids and exists as two possible isoforms by alternative splicing. These isoforms differ by deletion of residue K370 in isoform 2 (https://www.uniprot.org/uniprot/P49336#expression). The sequence adopts the protein kinase-like (PKL) fold, mostly-β-stranded N-lobe connected to the mostly-α-helical C-lobe via the hinge region (**Supporting information**). Structural motifs of kinases are well characterized and shared by all eucaryotic/eucaryotic-like protein kinases (ePK/ELK).[16] The ATP site sits between the N-lobe and the

C-lobe, flanked by the glycine-rich loop (G-loop or P-loop) in the top, the catalytic loop containing the HRD motif and the activation loop (A-loop or T-loop) in the bottom, the αC-helix on the right, while the adenine head interacts with the hinge.[17] An important pattern is the DFG (DMG in CDK8) motif of the A-loop whose *open* conformation (Phe/Met making hydrophobic contact with αC-helix) indicates the active state of the kinase, while the *close* conformation marks the inactive state.[17] Kinase inhibitors are classified according to the their binding site and bound-kinase state (Type I to VI). Type I inhibitors bind to the catalytic site in active conformation, while type II inhibitors bind to the inactive DMG-out conformation.[18] More information about kinase domains and their regulations are available in the literature.[19] To be active, kinases of the CDK family associates with other protein partners, mainly cyclins. CDK8 interacts with cyclin C. To this date (17/04/2022), only 31 structures of CDK8-CyclinC are available in the Protein Data Bank (PDB) in contrast to some other CDKs (e.g. 427 CDK2 entries in the PDB). Among these structures, one PDB entry corresponds to the apo-protein, 20 relates to complexes with type I inhibitors (DMG 'in'), and ten with type II inhibitors binding to the back pocket (DMG 'out') (**Supporting information**).

The following section (4.2.2 – 4.2.9) has been revised and published in:

The open source code is available at: https://github.com/kimeguida/POEM
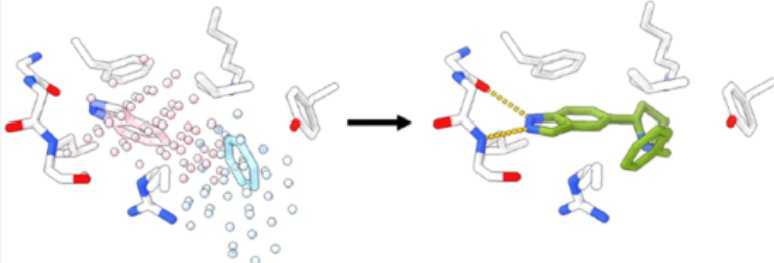
# Target-Focused Library Design by Pocket-Applied Computer Vision and Fragment Deep Generative Linking

Merveille Eguida, Christel Schmitt-Valencia, Marcel Hibert, Pascal Villa, and Didier Rognan*

ACCESS | Metrics & More | Article Recommendations | Supporting Information

**ABSTRACT:** We here describe a computational approach (POEM: Pocket Oriented Elaboration of Molecules) to drive the generation of target-focused libraries while taking advantage of all publicly available structural information on protein–ligand complexes. A collection of 31 384 PDB-derived images with key shapes and pharmacophoric properties, describing fragment-bound microenvironments, is first aligned to the query target cavity by a computer vision method. The fragments of the most similar PDB subpockets are then directly positioned in the query cavity using the corresponding image transformation matrices. Lastly, suitable connectable atoms of oriented fragment pairs are linked by a deep generative model to yield fully connected molecules. POEM was applied to generate a library of 1.5 million potential cyclin-dependent kinase 8 inhibitors. By synthesizing and testing as few as 43 compounds, a few nanomolar inhibitors were quickly obtained with limited resources in just two iterative cycles.

## 4.2.2. Abstract

Choosing the most appropriate chemical space is key to successfully screen compound libraries for early drug discovery. We here describe a novel computational approach, inspired from fragment-based design, to drive the generation of target-focused libraries while taking advantage of all publicly available structural information on protein-ligand complexes. The query target cavity, represented by an image with key shape and pharmacophoric properties, is first aligned by a computer vision method to a collection of 31 384 images describing fragment-bound microenvironments (subpockets) from the Protein Data Bank. The fragments of the most similar PDB subpockets are then directly positioned in the query cavity using the corresponding image transformation matrices. Last, suitable connectable atoms of oriented fragment pairs are linked by a deep generative model to yield fully connected molecules. As a first proof of concept, the method was applied to generate a library of 1.5 million potential cyclin-dependent kinase 8 (CDK8) inhibitors. After appropriate filtering, as few as 43 compounds were purchased or synthesized, and tested for *in vitro* competitive CDK8 inhibition. Several nanomolar inhibitors were quickly obtained with limited resources in just two iterative cycles. The approach is applicable to any druggable cavity of known three-dimensional structure, irrespective of prior ligand information.

## 4.2.3. Introduction

Fragment-based drug design (FBDD)[1] has gain considerable popularity in the last 20 years for identifying new lead compounds and guiding the optimization towards drug candidates, even up to the market with four recently approved drugs.[2] Common FBDD programs starts by screening libraries of low molecular weight compound (fragments)[3] by multiple biophysical methods such as nuclear magnetic resonance spectroscopy (NMR), surface plasmon resonance (SPR), isothermal titration calorimetry (ITC) or mass spectroscopy (MS) to cite just a few.[4] Key advantages of FBDD with respect to biochemical high-throughput screening (HTS) are the sampling of a much larger chemical space as well as higher hit rates, even for difficult targets for which other approaches failed. Despite low affinities, fragment hits can be progressed to leads by linking, merging or growing approaches.[5] Although not necessary, it is usually advisable to start from high quality X-ray diffraction data to position fragment hits in their cognate target.[6] Even if FDBB is now widely used for hit identification, not all targets and fragments are suitable to X-ray diffraction. One the one hand, some targets still proved to be hard to isolate, purify in large scale and produce high-quality crystals for X-ray diffraction. On the other hand, some fragments cannot be detected by the latter technique because of poor physicochemical properties or too low affinities. In such cases, computational approaches are the only alternatives to predict the most viable positions of fragment hits identified experimentally[7] or to identify new hits by *in silico* screening.[8]

Three computational approaches can be used to predict the relative orientation of a fragment in a target cavity: molecular docking, functional group mapping and deconstruction-reconstruction. Molecular docking[9] is by far the most popular structure-based approach and aims at identifying both the bound conformation and the orientation of the ligand in a target cavity from their respective stereochemical and topological complementarities. Although it has mostly been applied to drug-like compounds, docking can be used to pose fragments with an accuracy comparable to that of lead-like compounds.[10-11] Docking is the computational method that is the closest to experimental fragment screening, and can be directly applied to any fragment library. In addition to potential hit identification, the fragment position in the target cavity is also predicted. Unfortunately, scoring weak-binding fragments remains a challenge and requires an efficient post-processing, e.g. knowledge-based protein-ligand interaction rescoring.[12-14]

Functional group mapping[15] uses probe atoms or groups to map a protein cavity at their preferential location. Probes can be positioned according to protein-ligand interaction energies at regular points of a three-dimensional (3D) lattice[16-17] or by molecular dynamics (MD) sampling.[18] Interestingly, exhaustive all-atom MD better captures protein flexibility and solvation issues, and may also unmask transient cavities hidden to conventional docking protocol. Key drawback is the computational burden limiting a

wide applicability for virtual screening. Moreover, reconstructing a fully connected ligand from several discontinuous propensity maps is not straightforward.

Last, deconstruction-reconstruction approaches[19] aim at computationally splitting protein-bound ligand X-ray structures into fragments according to well-known retrosynthetic organic chemistry rules.[20-21] Resulting fragments can then be recombined into new chemical entities while taking into account the protein environment. The method still suffers from the tricky recombination step (linking, merging, scaffold hopping)[22] that may disturb the original fragment binding modes or generate conformational strains. Interestingly, deep generative models[23-25] for linking disconnected fragments have shown some promises as they learn from millions of existing bioactive ligands. Deconstruction-reconstruction is mainly target-specific and applicable to targets for which numerous co-crystallized ligands are already available, although docking poses may be used in principle.

None of the above-reported method really takes profit of the increasing amount of structural data on protein-ligand complexes and their druggable pockets.[26] Since low molecular weight fragments have been shown to bind to preferential protein microenvironments regardless of their evolutionary relationship,[27] a FBDD approach considering the whole universe of druggable ligands and pockets is desired. Capitalizing on our recent numerical image processing tool to describe and align protein cavities,[28] we here propose to pose fragments according to the local similarity of their respective subpockets to the target cavity. Applying the transformation matrix leading to the optimal subpocket-cavity alignment, the corresponding fragments are directly positioned into the target cavity and connected, under topological constraints, by a deep generative linker to yield fully connected molecules. Applying the method to the catalytic site of human cyclin dependent kinase 8 (CDK8), a focused library of 1.5 million chemical entities could be quickly generated. Interestingly, most newly generated compounds exhibited unprecedented structures. *In vitro* biological evaluation of 43 carefully selected compounds identified several nanomolar inhibitors within just two design iterations and limited experimental efforts.

## 4.2.4. Results and discussion

**Setting the scene**

We herein present a novel method to design target cavity-focused libraries based on predicted similarities between the target cavity and a library of PDB fragment-bound subpockets (**Figure 1**). The underlying idea is to locate the most complementary fragments in the target cavity based on the estimated similarity of their corresponding subpockets, and then to link the prepositioned fragments into drug-like compounds using a deep generative linker. Accordingly, this approach can be implemented even in the absence of known ligands for the target protein. To assess its applicability and limits in a real-life drug design project, the method is here applied to CDK8, a target of pharmaceutical interest[29] and known X-ray structure.[30] In the following sections, we will describe, step by step, each part of the workflow until the experimental validation of newly generated inhibitors.
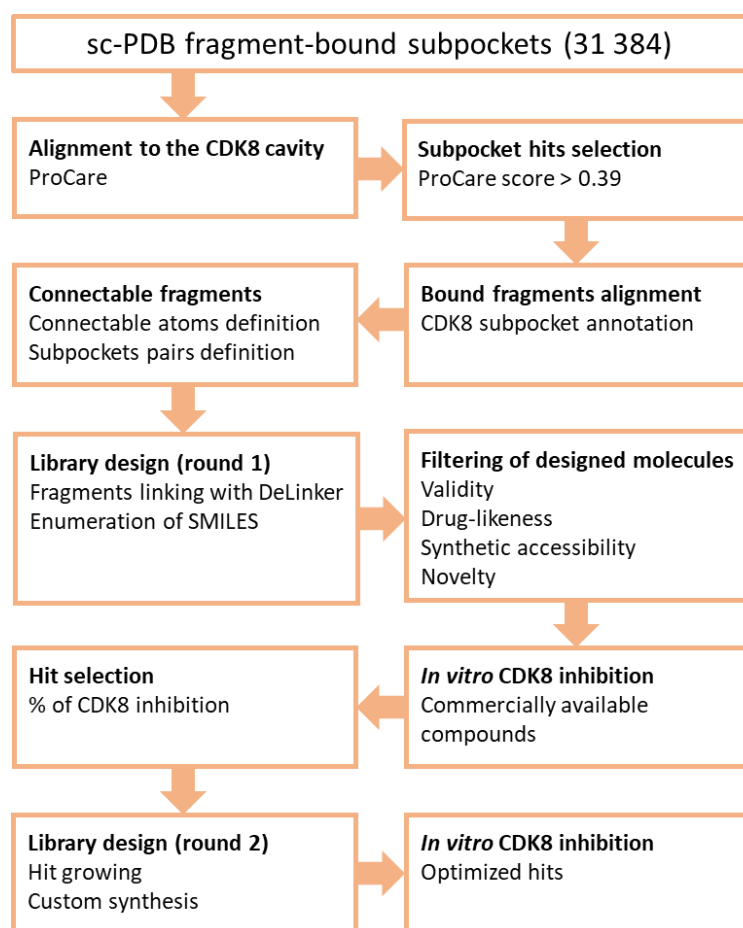


**Figure 1.** Overall workflow of the computational method including *in vitro* experimental validation.

**Alignment of fragments to the target cavity**

Subpockets, defined as the immediate protein environment around bound fragments of druggable protein-ligand complexes (sc-PDB dataset),[31] were compared and aligned to the ATP pocket of CDK8 with the aim to use the hidden bound fragments for library design. The rationale of this implementation is that according to the similarity principle, fragments originating from similar subpockets are likely to reproduce favorable interactions with the target pocket. The term 'fragment' here refers to the molecular moieties obtained after interaction-aware 3D fragmentation of ligands bound to proteins so that each fragment exhibits at least one polar interaction and at least four interactions with its target.[32] The query CDK8 pocket and the sc-PDB subpockets are represented as a cloud of 1.5 Å-spaced points annotated by eight pharmacophoric properties (hydrophobic, aromatic, H-bond acceptor, H-bond donor, H-bond acceptor and donor, positive ionizable, negative ionizable, null).[33] The term 'pocket' describes the full druggable cavity available at the surface of the protein while a subpocket is defined from its bound fragment. Since we aimed at targeting the ATP binding site in its type-I 'DMG in' conformation, the druggable pockets were first detected from 19 available CDK8 structures (**Table S1**). The largest pocket (830.3 Å$^3$) selected as representative was retrieved from the 5HBH[30] PDB entry (**Figure 2**). This pocket incorporates regions around the hinge, the gatekeeper F97, whereas on the opposite side extends to a solvent exposed area near the αD helix. It covers the DMG motif and reaches the αC-helix (**Figure 2A**). It thus spans several already described kinase subpockets: the adenine pocket, the front pockets FP-I and FP-II, the back pockets BP-I-A and BP-I-B in the gate area.[34] The 31 384 sc-PDB subpockets were compared and aligned to the CDK8 cavity with the in-house ProCare method (**Figure S1**).[28] Briefly, ProCare finds the best possible local alignment of cavity-defining points using a point cloud registration algorithm[35-36] and scores the alignment according to the overlap of pharmacophoric properties of the aligned points. According to a preliminary study on the set of CDK8 structures, the original ProCare alignment fingerprint was modified to account only for the spatial distribution of pharmacophoric features (**Figure S2-S3**), a modification leading to a better alignment of CDK8 subpockets and fragments to the corresponding full cavities.
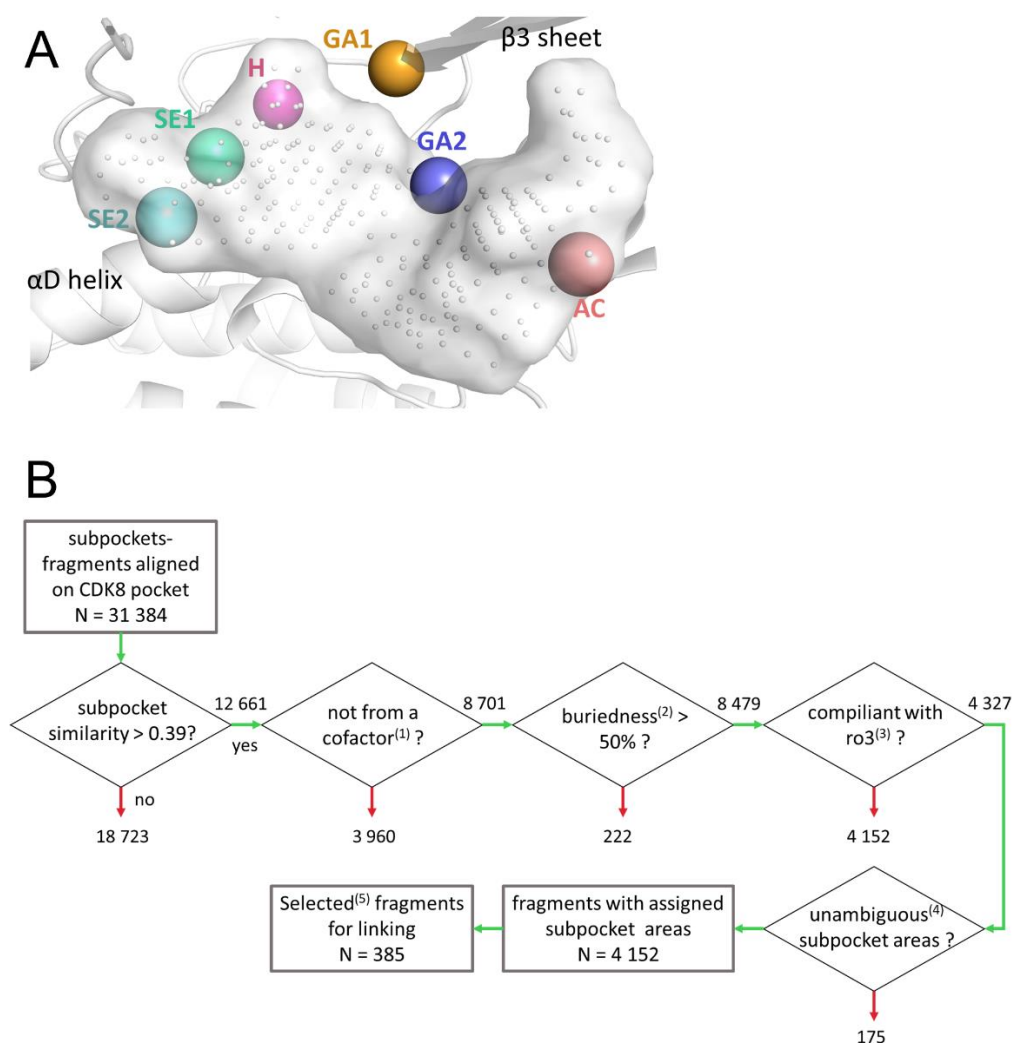
**Figure 2.** Seed fragments selection to fill the CDK8 query cavity. **A)** Description of the reference CDK8 pocket (PDB ID: 5HBH). Cavity points (grey dots, 246 points) delineate a ligand-accessible envelope (solid surface, 830.3 Å$^3$) and areas (hinge, H; gate area 1, GA1; gate area 2, GA2; solvent-exposed area 1, SE1; solvent-exposed area 2, SE2; αC area, AC) according to the distance to key CDK8 atoms (spheres). **B)** Fragments selection workflow. (1) A list of cofactors (PDB HET code) is provided in the sc-PDB database. (2) Fragments buriedness is approximated as the percentage of heavy atoms within 1.5 Å of one CDK8 cavity point. (3) fragment rule-of-three:[37] molecular weight ≤ 300 g.mol-1, logP ≤ 3, H-bond donor count ≤ 3 and H-bond acceptor count ≤ 3. (4) ambiguous annotation denotes assignment of two or more incompatible areas (Methods section) out of the six possible areas. (5) All annotated fragments from H, GA1, SE2 areas and a random sampling of 100 fragments from GA2 were selected.

Once transformation matrices of the alignment of sc-PDB subpockets to the target cavity were obtained, the same rotation/translation matrices were applied to the corresponding sc-PDB fragments to position them in the CDK8 cavity. Posed fragments were then filtered according to five criteria (**Figure 2B**). Fragments originating from subpockets exhibiting a similarity score to the CDK8 pocket above a threshold value of 0.39 (previously shown to optimally discriminate known similar from known

dissimilar binding sites)[28] were first selected, leading to a set of 12 661 fragments. Remaining fragments were further pruned according to three criteria: (i) belonging to a cofactor (therefore avoiding purine-base fragments), (ii) insufficient buriedness in the target cavity, (iii) no compliance to the fragment rule-of-three.[37] Remaining fragments were then annotated by one of the six CDK8 areas in which they were positioned: hinge (H), gate (GA1, GA2), solvent-accessible (SE1, SE2), αC helix (AC) (**Table 1**, **Figure 3**). 4 152 fragments could be unambiguously assigned to one CDK8 area: H (1.4%), GA1 (2.7%), GA2 (22.5%), SE1 (61.9%), SE2 (2.8%) and AC (8.7%) (**Figure 3A**).

**Table 1.** Annotation of the CDK8 target cavity by key pharmacophoric atoms.

| Area | Label | Key CDK8 atoms | KLIFS subpockets[a] |
|------|-------|----------------|--------------------|
| Hinge area | H | Asp98.O, Ala100.N, Ala100.O | AP |
| Gate area 1 | GA1 | Phe97.CA (gatekeeper residue) | AP, BP-I-A, BP-I-B |
| Gate area 2 | GA2 | Lys52.NZ | AP, FP-I, FP-II |
| Solvent-accessible area 1 | SE1 | Arg366.CZ | - |
| Solvent-accessible area 2 | SE2 | His106.CE1 | - |
| αC helix area | AC | Ser62.CA | - |

[a] Full or partial overlap with KLIFS[34] subpockets: AP: adenine pocket, BP: back pocket, FP: front pocket
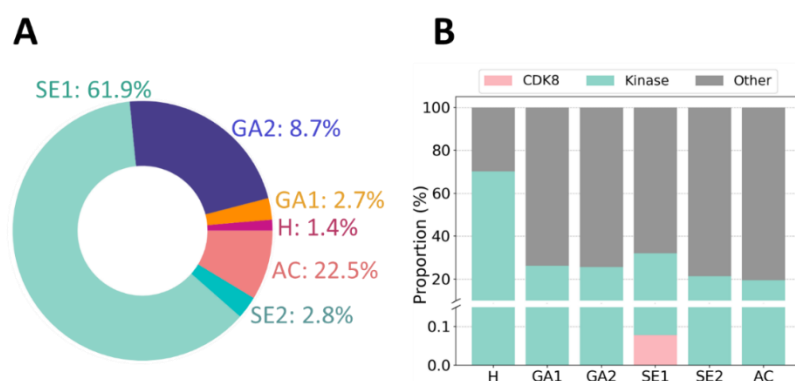


**Figure 3**. CDK8 subpocket occupancy of sc-PDB fragments. **A)** Assignment of CDK8 pocket areas to 4 152 sc-PDB fragments. **B)** Origin of sc-PDB fragments per area.

We next analyzed the origin of the sc-PDB ligands these fragments were derived from. As to be expected, 70% of fragments assigned to the hinge area (H) come from protein kinase inhibitors, the

remaining 30% originating from a ligand co-crystallized with a protein that belong to a non-kinase family (**Figure 3B**). However, it should be noted that fragments from known CDK8 inhibitors were not selected as occupying the hinge region. Two simple reasons explain this absence: (i) the seven CDK8 ligands in the sc-PDB dataset are type II inhibitors binding to a DMG-out conformation and occupy the back pocket, (ii) the only CDK8 ligand (3RGF) that binds to the hinge could not be fragmented by our protocol and therefore did not pass our filters. The other areas (GA1, GA2, SE1, SE2, AC) were assigned fragments from both kinase (~25%) and non-kinase ligands (~75%). While the initial sc-PDB subpocket database contains 16% of entries from protein kinases, the enrichment observed for hinge-selected fragments (4.4) is logically due to the specific stereoelectronic features of the hinge area, notably the hydrogen bonding capacity of Asp98 and Ala100 backbone heteroatoms imposing complementary features on the ligand side. To limit the size of the library, all fragments were not considered for full enumeration of complete molecules. Whereas all fragments bound to H (n=57), GA1 (n=111) and SE2 (n=117) subpockets were selected, only 100 GA2-bound fragments were randomly chosen. Duplicates, in other words 2D identical fragments were kept as they do not originate from the same 3D subpocket, therefore resulted in different alignments that may differently impact molecules design. Comprehensive statistics of the pairwise fragment similarity (**Figure S4**) and the observed distribution of their physicochemical properties (**Figure S5**) clearly evidence their chemical diversity. 385 fragments were selected at this stage for the next linking stage.

**Round-1 library generation**

The DeLinker deep generative model[23] was used to link the above-selected fragments. Briefly, DeLinker uses a graph-based deep generative model, trained on the ZINC[38] or PDBbind[39] databases, to expand bond by bond the two fragments to be connected until final SMILES strings are generated by a variational autoencoder while keeping 3D constraints through a set of distances and angles between connectable atoms.[23] In the current work, all possible connectable atoms of hinge-annotated fragments (H) were used as seeds to find potential connectable atoms in fragments filling three remaining subpockets (GA1, GA2, SE2) (**Figure S6**).

An atom is considered connectable if it is a heavy atom covalently bonded to a hydrogen, that bond being used as exit vector for the linking. Pairs of atoms belonging to different fragments are then associated by restricting the angle between the exit vectors and distances between the corresponding heavy atoms (see Methods) in order to avoid pointless connections and lower the number of combinations (**Figure S7**). Starting from 385 fragments, 1 517 488 SMILES strings were generated by linking fragment pairs with DeLinker. 15% of the proposed solutions were discarded since they correspond to uncomplete molecules where the SMILES consisted of a linker moiety attached to only one of the two fragments (**Figure 4**).
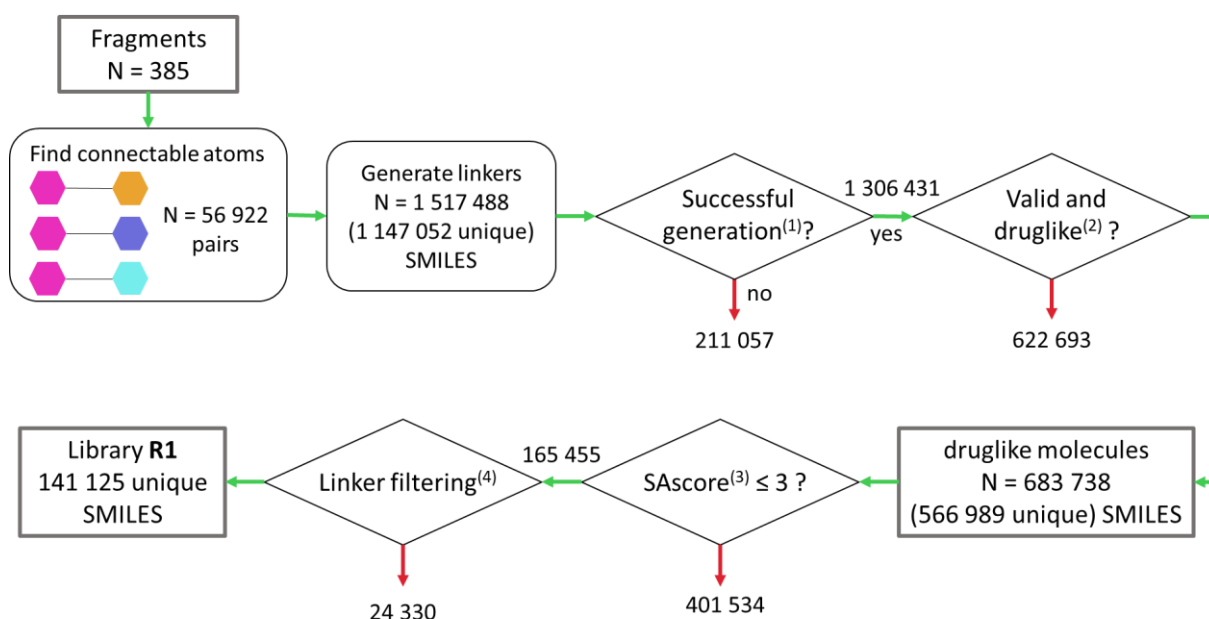
**Figure 4**. Focused library design via linking selected fragments. Fragments aligned in the H area were paired with fragments from GA1, GA2 and SE2 areas. SMILES were generated by linking fragment pairs with DeLinker[23] and filtered to compose the first-round library R1. (1) Successful linking signifies that both fragments have been attached to the linker whereas cases where only one of the fragments was linked were considered unsuccessful. (2) Druglikeness is defined by customized OpenEye Filter rules available in **Table S2**. (3) Synthetic accessibility score.[40] (4) Filter to remove unwanted aliphatic linkers.

The remaining molecules were filtered for drug-likeness (**Table S2**) resulting in 566 989 unique SMILES. Although the redundant SMILES per pair of connectable atoms were removed during the linking process, duplicated molecules still arose when connecting the same 3D fragments via equivalent exit atoms (symmetry cases) or connecting the same duplicated fragments originating from different subpockets. After keeping only molecules that are likely to be synthesized (SAscore[40] $\leq$ 3), only those having a linker compliant with defined rules (**Figure S8**) were finally kept. The remaining 141 125 molecules composed the first-round R1 library (**Figure 4**). A majority of the generated molecules arose from combining the hinge and the solvent-exposed SE2 fragments which account for more than 50% of the sets (**Figure 5**).
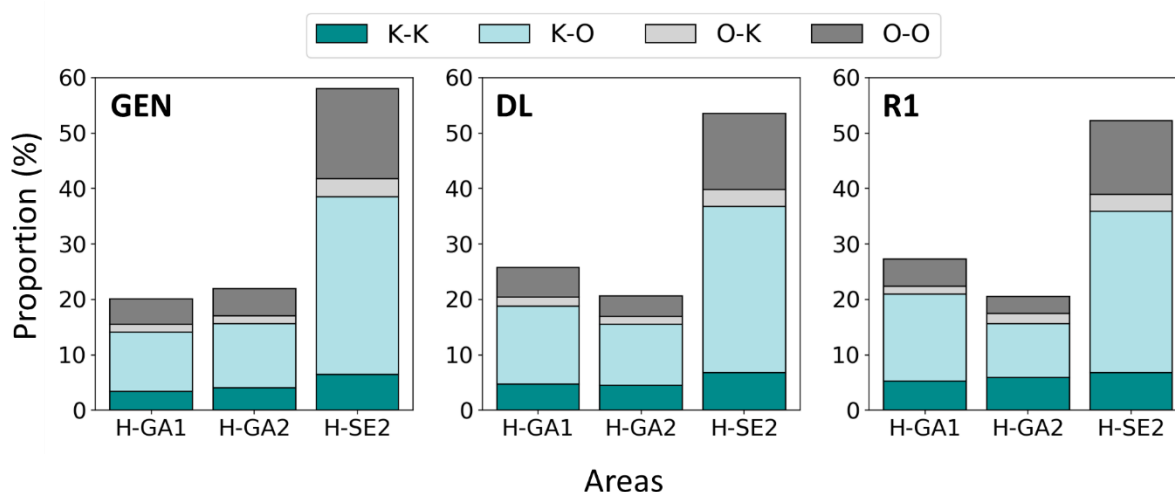
**Figure 5**. Protein origin of fragments pairs in newly generated molecules. From left to right, the full set after cleaning unsuccessful generation out (GEN), the drug-like subset (DL) and round-1 library (R1). The distribution is given for the combinations annotated by the targeted CDK8 area (H, hinge; GA1, gate area 1; GA2, gate area 2, SE2, solvent-exposed area 2) and color-coded according to the protein origin (co-crystallized target) of the two connected fragments (K, protein kinase; O, other; K-K, both fragments were derived from a protein kinase structure; K-O, H-fragment derived from a protein kinase and the other fragment from a non-kinase protein structure; O-K, H-fragment derived from a non-protein kinase and the other fragment from a kinase protein structure; O-O, both fragments were derived from a non-kinase protein structure).

Indeed, the average number of generated SMILES strings per pair of H-SE fragments is higher than for the two other areas, a consequence of having more pairs of connectable atoms and more generated linkers per connectable atoms for the H-SE subpockets. While it was expected that kinase-derived fragments would contribute to most of the generated molecules, only 14% of SMILES strings were generated by linking two kinase-bound fragments. Interestingly, around 26% of the molecules were made of two fragments originating from a non-kinase protein. Interestingly, the observed proportions do not vary between the full set, the drug-like subset and the R1 set (**Figure 5**). Most of the generated molecules (> 90 %) were already compliant with the Lipinski's rule of five (**Figure S9**). Albeit two fragments were assembled, many generated molecules still remained in the fragment space with around 10 % of SMILES strings being compliant with the fragment rule-of-three[37] (**Figure S9**). Filtering the designed molecules to R1 library members did not bias our selection towards molecules with particular properties as the distribution of the molecular properties, although reported individually, remained comparable among the sets (full, drug-like and R1; **Figure S9**). To give insights on the chemical space covered by R1 library members, we further assessed its overlap with either a broad purpose bioactive chemical space[41] (1.7 million ChEMBL compounds) or a recently described kinase-focused ligand space (6.7 million KinFragLib library members).[42] 259 unique R1 library molecules were exactly found in ChEMBL among which only a few have been assayed against protein kinases, while only five R1 library

compounds were identical to KinFragLib molecules. Considering similarity, only 0.85% and 13% of R1 library members were found similar to KingFragLib and ChEMBL molecules, respectively, according to a Tanimoto coefficient, computed from Morgan2 fingerprints higher than 0.60. The herein proposed computational workflow is therefore able to generate really new chemical entities, the chemical diversity of the generated molecules stemming from the diversity of the seed fragments pool, the connectivity and the possible linkers.

As a first validation of the structure-based workflow, we verified whether the drug-like subset contains molecules highly similar to 302 submicromolar human CDK8 inhibitors retrieved from the ChEMBL database. Using the similarity search protocol described in the methods section, we found 44 molecules that matched with 35 unique known CDK8 inhibitors (representing three series of congeneric molecules). While these molecules were built with fragments from all possible areas, most of them were assembled from hinge-fragments originally co-crystallized with protein kinases, linked to fragments originally co-crystallized with non-kinase proteins.

**The round-1 library contains novel and potent CDK8 inhibitors**

To identify chemically novel hits, we filtered first-round R1 library members by dissimilarity (Tanimoto coefficient < 0.5, RDKit7 fingerprints) to all CDK8 compounds available in ChEMBL[41] and to all seed sc-PDB fragments. Hits were then searched for availability among 8.2 million commercially available drug-like compounds (**Table S3**) to select 37 compounds that are identical or very similar (Tanimoto coefficient > 0.90, RDKit7 fingerprints) to their queries (**Table S4**). These compounds were purchased and tested for CDK8 inhibition in a homogeneous time-resolved fluorescence (HTRF) assay aimed at measuring the FRET signal between a fluorescent-labelled ATP competitive inhibitor and the fluorescent-tagged CDK8 soluble kinase (see Methods). Six out of the 37 tested molecules (compounds **9**, **11**, **12**, **29**, **32**, **37**) inhibited the CDK8 kinase by more than 50% at the single concentration of 10 μM (**Figure 6**). Notably two related compounds (**12** and **37**), exhibiting more than 80% inhibition were assembled from the same pair of 3D fragments by just inverting the ester linkage (**Figure 6**). They differ from the original R1 library members by just a carbon atom (methoxy for ethoxy substitution, **Table S4**).
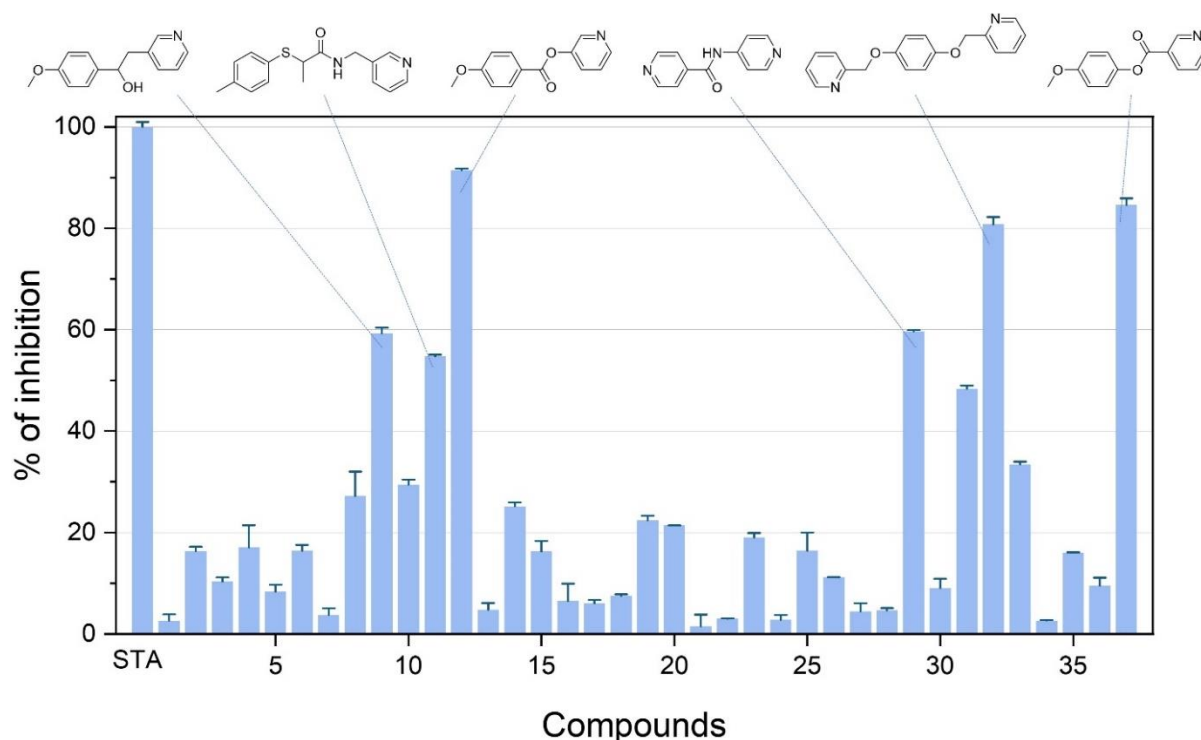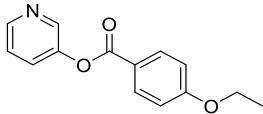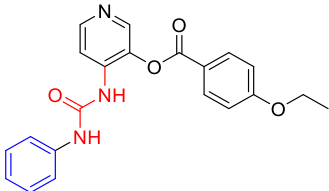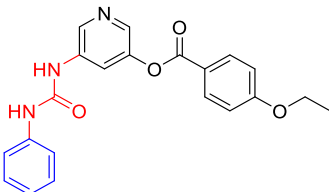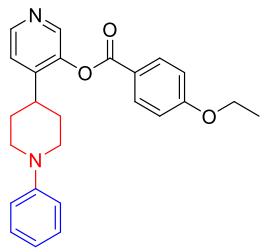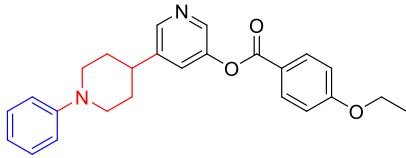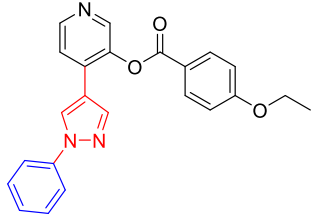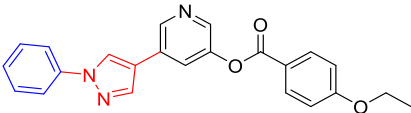
**Figure 6.** CDK8 inhibition (LanthaScreen Eu kinase competitive binding assay) by 37 commercially available compounds identical or very similar to R1 library members. Results are expressed as mean ± SEM of two independent experiments using a 10 μM concentration of competitor (STA, staurosporine control).

**Round-2 library design by fragment hit growing**

The most potent hit (**12**) from round-1 library , generated by linking a H-area pyridine fragment to a GA2-area methoxyphenyl fragment, is still a fragment-like compound (MW = 229 g.mol$^{-1}$) that can be optimized by growing towards the nearby and yet unexploited SE2 and GA1 subpockets. We thus explored the possible connections between the hinge-binding fragment of **12** and all remaining SE2 or GA1-anchored fragments, to generate a second-round library R2 of 5 700 compounds. R2 library members were filtered by physicochemical properties (number of rotatable bonds ≤ 6, no chiral centers) and synthetic accessibility (SAscore ≤ 3) to yield a final set of 151 candidates (**Table S5**). Six representative compounds (**Table 2**) were chosen for their ease of synthesis (i.e. availability of building blocks, costs of goods, number of synthetic steps) and predicted buriedness upon preliminary docking to CDK8. Three linkers (urea, piperidine, pyrazole) were chosen for their capacity to connect the H-anchoring pyridine ring to a SE2-anchored phenyl fragment. Two positions of the pyridine ring (ortho and meta position to the benzoyl ester) were predicted compatible, therefore leading to six possible analogs (**Table 2**).

**Table 2.** Round-2 library of optimized hits and their CDK8 inhibitory potency.

| Compound | Structure[a] | IC$_{50}$, nM[b] | CI 95%, nM[c] |
|---|---|---|---|
| **12** | | 376.9 | 245.2-579.5 |
| **39** | | 354.6 | 203.4-618.0 |
| **41** | | >25 000 | - |
| **44** | | 144.1 | 88.8-233.9 |
| **47** | | >25 000 | - |
| **49** | | 6.4 | 4.57-8.95 |
| **51** | | > 25 000 | - |

[a] A phenyl moiety (blue) is attached via different linkers (red) to round-1 compound **12**. [b] Inhibition of CDK8 measured in a LanthaScreen Eu kinase competitive binding assay. Results are expressed as mean ± SEM of three independent experiments. [c] confidence interval at a 95% confidence level

The six compounds were synthesized (**Scheme S1**), checked for purity (**Figures S10-S15**) and tested for *in vitro* CDK8 inhibition using the same HTRF assay as described above, to build concentration-response curves (**Figure 7**). Out of the six round-2 library compounds, three molecules (**41**, **47**, **51**) are weak CDK8 inhibitors, one compound (**39**) is equipotent to the primary hit **12**, and two analogues (**44**, **49**) exhibit a higher potency than the parent compound **12** (**Table 2**, **Figure 7**). 3,4-disubstituted pyridines (**39**, **44**, **49**) were systematically more potent than their 3,5-disubstituted congeners (**41**, **47**, **51**). Noteworthy, the single-digit nanomolar inhibitor **49** could be obtained from scratch within just two design iterations and limited experimental efforts.
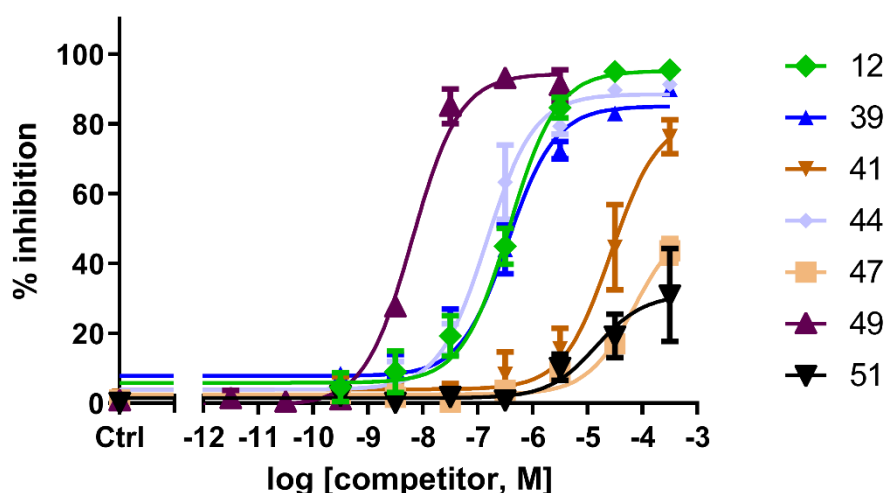


**Figure 7.** Inhibition of human CDK8 by six selected round-2 library compounds. Concentration-response curves are derived from three independent experiments with duplicates per experiment.

Its putative binding mode, deduced form molecular docking, suggests that the pyridine nitrogen atom h-bonds to the hinge backbone atoms (E98, A100) while the ethoxyphenyl and the newly introduced pyrazole moieties exhibit π- π interactions to H106 (SE2 subpocket) and the gatekeeper F97 (GA1 subpocket). Last, the terminal phenyl ring is oriented towards K52 (GA2 subpocket) for a putative π-cation interaction (**Figure 8**). While the parent hit **12** showed two possible docking poses (ethoxyphenyl towards GA2 or SE2), growing by a pyrazole prioritized the SE2 orientation, still with exhibited interactions compatible with the rationale of the initial fragment alignments.
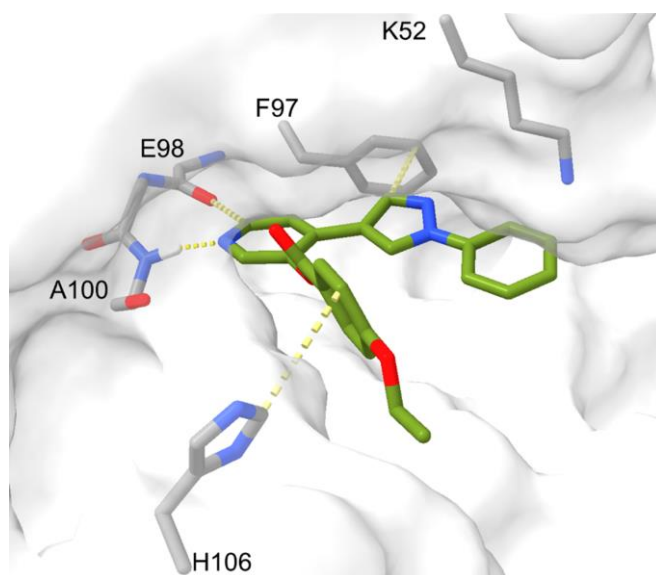
**Figure 8.** PLANTS docking pose of compound **49** (green sticks) to the catalytic site of CDK8 (PDB ID 5HBH, solid surface). H-bond to the hinge (E98, A100) and π- π interactions to F97, H106 are displayed by yellow broken bonds.

At this point, we should recall that neither early safety (e.g. kinase selectivity) nor pharmacokinetic properties (e.g. metabolic stability) have been considered in either generating or post-processing the target-focused library members. Although technically feasible, target selectivity assessment requires applying the same workflow to different cavities and prioritizing compounds generated only for the target of interest. This approach is feasible for a comparing a few targets but is rapidly impracticable at a larger scale (e.g. whole kinome). It has not been applied in the current study aimed at demonstrating the proof-of-concept of the structure-based workflow.

## 4.2.5. Conclusions

We herewith propose a novel fragment-based library design method to generate target-focused compound libraries. The originality of the approach is that seed fragments are chosen from a large repertoire of protein-bound fragmented ligand X-ray structures, and positioned in the target according to the local similarity of their protein subpocket to the target cavity. This ligand-agnostic posing protocol does not require scoring protein-ligand interactions and is fuzzy enough to transfer ligand information across unrelated target spaces. Once fragments have been posed, they are linked by a deep generative model to enumerate full molecules which are later post-processed to account for drug-likeness and synthetic accessibility. The linking step still deserves improvement, notably to enumerate candidate molecules directly in the original target 3D coordinate frame. Hence, the variational autoencoder used here generates SMILES strings and just accounts for the target binding site topology in the form of topological relationships between fragment atoms to be connected. A true 3D deep generative model[43] considering complementarity to the binding site shape and the ligand conformational freedom would be highly desirable to link subpocket-selected seed fragments. It would avoid a tedious post-processing of unrealistic solutions and the necessary docking of candidates to verify whether the starting binding hypothesis of the seed fragments is conserved.

When applied to the test case of the CDK8 kinase, the method was able to quickly suggest potential inhibitors. Within two iterations and 43 compounds, a single digit nanomolar inhibitor could be identified thereby demonstrating a first proof-of-concept of the underlying methodology. Interestingly, the method is applicable to any target of known 3D structure and does not require prior ligand knowledge.

## 4.2.6. Material and methods

**CDK8 cavity detection**

All publicly available X-ray structures of human CDK8 (UniProt accession number P49336; **Table S1**) were downloaded from the Protein Data Bank[44-45]. Type I structures (DMG-in, α-C helix-out) were put in the same coordinates frame by subsequent structural alignment to the 4F7S reference with Maestro v.2019-3 (Schrödinger, New York, NY 10036, U.S.A.) and refinement to ensure that the hinge residue Ala100 heavy atoms were fitted. Aligned structures (proteins, co-factors, ligands) were then protonated with Protoss v.4.0,[46] while optimizing the intra and inter-molecular hydrogen bond network. After discarding crystallization additives, each PDB entry was split to afford a protein (no water molecules) and a ligand in separate mol2 files using SYBYL-X 2.1.1 (Certara USA, Inc., Princeton, NJ 08540, U.S.A.). For each protein file, entire cavities ("*CAVITY_ALL*" output) were next computed with the VolSite[33] module of the IChem v.5.2.9 package,[47] using default parameters and saved as point clouds annotated by pharmacophoric features. Only cavities corresponding to the catalytic site were retained for the next steps. Upon visual inspection, the corresponding three clouds for PDB entry 5HBH were merged into a single cavity in mol2 file, yielding the reference pocket for CDK8.

**sc-PDB subpocket-fragment database**

16 034 drug-like ligands in their protein-bound X-ray structure were retrieved from the sc-PDB database[31] of druggable protein-ligand complexes and fragmented in three dimensional (3D) space within their protein binding site using the IChem fragmentation tool.[32] Only fragments exhibiting at least 4 non-covalent interactions[12] (out of which one is polar, hydrogen-bond or electrostatic interaction) with the protein target were retained. The fragments exit bonds (dummy atoms 'Z') were converted into hydrogen atoms. The immediate protein environment of each selected fragment was considered to compute VolSite point clouds, keeping only those with at least 3 points, each being closer than 4.0 Å from any fragment heavy atom ("*CAVITY_4*" output), thereby defining a subpocket point cloud in mol2 file format for 31 384 fragments.

**CDK8-focused library design**

In the first stage, 31 384 sc-PDB subpocket point clouds (**Figure S1**) were aligned to the reference 5HBH CDK8 cavity point clouds with ProCare[28] v.0.1.1 using default parameters and the c-FH color-based descriptor (**Figure S2**) corresponding to the eight terminal bins of the c-FPFH descriptor.[28] For each subpocket-cavity pair, the optimal alignment matrix was used to position the corresponding sc-

PDB fragment into the CDK8 cavity. The comparison protocol was validated by successful cross-comparison of CDK8 subpockets from type I PDB entries (**Figure S3**).

In the second stage, aligned sc-PDB fragments were filtered according to their subpocket similarity to the CDK8 cavity (ProCare score ≥ 0.39), their compliance to the fragment rule-of-three,[37] and their embedding into the CDK8 cavity such that at least half of the fragment atoms are less than 1.5 Å away to the closest CDK8 cavity point. Fragments originating from the sc-PDB list of cofactors were excluded. Resulting fragments were further annotated with the CDK8 cavity area to which they have been aligned based on their distance (closest heavy atom should be within 6 Å) to subpocket-specific preliminary defined atom centers (hinge H area, Asp98 O atom and Ala100 N and O atoms; gate area 1 GA1, Phe97 CA atom; gate area 2 GA2, Lys52 NZ atom; solvent-exposed area 1 SE1, Arg356 CZ atom; solvent-exposed area 2 SE2 subpocket, His106 CE1 atom; αC area AC, Ser62 CA atom). For selecting hinge-binding fragments, hydrogen bonds to Asp98 O or Ala100 N or O was mandatory. Since a few fragments were assigned to multiple subpockets, the following prioritization scheme was applied: H annotation takes precedence over all the other annotations, therefore a fragment interacting with the hinge centers is only annotated as such. SE1 and SE2 were defined compatible so that fragments annotated as from both areas were automatically assigned only SE2. Similarly, fragments annotated as from both AC and GA2 areas were automatically assigned only GA2. In any other case of combination (e.g. fragments annotated as from GA2 and SE1), the annotations were considered ambiguous and the fragments were discarded.

In the third stage, H fragments were defined connectable to either GA1, GA2 or SE2 fragments (in the current work, although other connections are possible). Selected fragments were converted into sdf format with OpenEye v.2.5.1.4. toolkit.[48] For each pair of fragments with hydrogen atoms connected, pairs of connectable atoms were searched based on their respective orientation as follows. A right circular cone (half-angle=$\pi/4$) is projected along the bond axis between any heavy atom $A_i$ and a bound hydrogen atom $H_i$. A connectable atom pair $A_1A_2$ is selected if heavy atoms $A_1$ and $A_2$ are located in the projection cone of their counterpart (**Figure S7**).

In the fourth stage, the recently-described DeLinker[23] deep learning method was employed to generate linkers between above-described connectable atom pairs using the default model distributed with the package and a batch size of 1. Input data were prepared as ZINC atom types features to be ready for DeLinker using the 'prepare_data' module and by setting the 'test' parameter of the 'preprocess' function to 'True' as molecules are to be found. The linker length was set to a minimum of on and a maximum of six heavy atoms. Other parameters were kept by default. Generated molecules were saved as SMILES strings and further processed to remove redundancy for each connectable atom pair. In the final stage, unsuccessful linking attempts where only a single fragment is attached to the linker were removed using the function 'get_linker' in the 'frag_utils' utility. The remaining SMILES were filtered to keep only

drug-like compounds according to in-house rules (**Table S2**). Next, the synthetic accessibility scores were computed with the the SAscore[40] method distributed with RDKit[49] to remove molecules with SAscore higher that three. Finally, molecules made of long flexible linkers were discarded according to our in-house filtering workflow (**Figure S8**), resulting in the first-round library (R1).

**Comparison with ChEMBL and KinFraglib ligands**

Standardized ChEMBL (1.7 million compounds) and KinFragLib (6.7 million) data were retrieved from the KinFragLib website.[50] Pairwise 2D fingerprint similarity to R1 molecules were assessed with RDKit[49] Morgan (radius = 2) topological fingerprint (default parameters, maximum path = 7).

**Comparison to known CDK8 inhibitors**

A search in the ChEMBL database[51, 41] for human CDK8 target assays resulted in three target report cards (CHEMBL3038474, CHEMBL5719 and CHEMBL3885556) from which bioassay data were joined and processed to keep compounds with a half maximal inhibitory concentration $IC_{50}$ inferior or equal to 1 μM. Duplicates were then removed according to and the SMILES were standardized with OpenEye Filter v.3.0.1.2 (OpenEye Scientific Software, Santa Fe, NM 87508, U.S.A.). The final list of 302 inhibitors was searched in the generated drug-like subset described above for substructure 2D similarity using both RDKit Morgan (radius = 2) and topological (maximum path = 7) fingerprints and a combination of Tanimoto (Tc) and Tversky (Tv) metrics. Pairs were reported when morgan2 Tc $\geq$ 0.6 or morgan2 Tv $\geq$ 0.8 or RDKit7 Tc $\geq$ 0.75 or RDKit7 Tv $\geq$ 0.9.

**Search for new potential CDK8 inhibitors**

R1 library members were considered as potentially new at the condition that their similarity to any of 946 unique human CDK8-tested compounds (both active and inactive) reported in ChEMBL (target card reports CHEMBL3038474, CHEMBL5719 and CHEMBL3885556) and any of the 31 384 sc-PDB fragment is inferior to 0.50 (Tanimoto coefficient from RDKit topological fingerprints). Last, the subsequent list was searched for substructure similarity (RDKit topological fingerprint Tanimoto $\geq$ 0.90) to an in-house library of 8 280 193 commercially available drug-like compounds (**Supporting Table S3**).

**Molecular docking**

Virtual hits were drawn as 2D sketches with ChemAxon MarvinSketch v.16.10.17, (ChemAxon Ltd., 1031 Budapest, Hungary) saved in sdf file format, ionized at physiological pH with OpenEye Filter v.2.5.1.4 and finally converted in 3D structures (mol2 file) with Corina v.3.40 (Molecular Networks GmbH, 90411 Nürnberg, Germany), generating all possible stereoisomers and ring conformers simultaneously. The prepared molecules were docked into the above-described CDK8 cavity using PLANTS[52] v.1.2 The search space was set at 13 Å from the binding site center with a search speed of 1 (highest accuracy). 10 poses were generated per ligand, scored by the ChemPLP scoring function and clustered using a root-mean square deviations (RMSD) of 2 Å on ligand heavy atoms. The flipped/rotated side chains were reconstructed in the protein structure for each corresponding PLANTS pose when applicable.

**Molecular data analysis**

Molecular descriptors (molecular weight (g.mol$^{-1}$), the count of heavy atoms (non-hydrogen atoms), logP, polar surface area (Å), count of H-bond acceptor, count of H-bond donor, count of rotatable bonds, count of ring systems, count of heteroatoms, bonds) were computed with RDKit. Data were processed with Python v.3.7.

**Data visualization**

Molecules were drawn in 2D with RDKit and MarvinSketch v.16.10.17, (ChemAxon Ltd., 1031 Budapest, Hungary). Three-dimensional structures were analyzed with Maestro v.2019-3 (Schrödinger, New York, NY 10036, U.S.A.) and Pymol v.2.1 (Schrödinger, New York, NY 10036, U.S.A.). Plots were generated with Matplotlib v3.0.2[53] in Python v.3.7.

**Chemistry**

All reactions were carried out under usual atmosphere unless otherwise stated. Chemicals and solvents were purchased from Enamine (LV-1035 Riga, Latvia) and were used without further purification. Yields refer to isolated compounds, estimated to be >95% pure as determined by 1H NMR or HPLC. 1H NMR spectra were recorded at 298 K on Bruker Avance III Spectrometer operating at 400 MHz. All chemical shift values δ and coupling constants J are quoted in ppm and in Hz, respectively; multiplicity (s = singulet, d = doublet, t = triplet, q = quartet, quin = quintet, sex = sextet m = multiplet, br = broad).

Preparative HPLC was performed using two methods: Method A) 2-10 min 30-70% acetonitrile, 30 ml/min ((loading pump 4 ml acetonitrile); column: YMC-ACTUS TRIART (C18; 100 mm x 20 mm; 5 µm); Method B) 2-10 min 0-50% acetonitrile, 30 ml/min ((loading pump 4 ml acetonitrile); column: SunFire C18; 100 mm x 19 mm; 5 µm)

Analytical RP-HPLC-MS was performed using Agilent Technologies 1260 Infinity LC/MSD system with DAD\ELSD Alltech 3300 and Agilent LC\MSD G6120B mass-spectrometer using the following acquisition parameters: column, Agilent Poroshell 120 SB-C18 4.6x30mm 2.7 µm with UHPLC Guard Infinity Lab Poroshell 120 SB-C18 4.6x 5mm 2.7 µm; Temperature 60° C; Mobile phase A – acetonitrile : water (99:1%), 0.1% formic acid, B – water (0.1% formic acid); Flow rate 3 ml/min; Gradient : 0.01 min –99% B, 1.5 min – 0% B, 1.73 min - 0% B, 1.74 min - 99% B; Injection volume 0.5µl; Ionization mode Electrospray ionization (ESI); Scan range m/z 83-600; DAD 215 nm, 254nm, 280 nm. Purities of all tested compounds used in the biological assays were determined by HPLC/MS using the area percentage method on the UV trace recorded at a wavelength of 254 nm. All compounds were found to have >95% purity.

*1-(3-hydroxypyridin-4-yl)3-phenylurea (38).* To a stirred solution of phenylisocyanate (0.4 g, 3.4 mmol) in DMF (5 ml) was added a solution of 4-aminopyridin-3-ol hydrochloride (0.5 g, 3.4 mmol) in DMF (5 ml) followed by the addition of triethylamine (1.4 ml, 10.2 mmol) at room temperature (r.t.). The resulting mixture was stirred at room temperature overnight. The reaction mixture was concentrated under reduced pressure and the crude residue was purified by HPLC to afford 50 mg (6%) of the 1-(3-hydroxypyridin-4-yl)-3-phenylurea **38** as a white solid which was used for the next step without further purification.

*4-(3-phenylureido)pyridin-3-yl 4-ethoxybenzoate (39).* To a stirred solution of 4-ethoxybenzoic acid (36 mg, 0.22 mmol) in DMF (2 ml), compound **38** (50 mg, 0.22 mmol), EDC (50 mg, 0.26 mmol) and DMAP (27 mg, 0.22 mmol) were added. The resulting mixture was stirred at r.t. for 16 h. After completion of the reaction, the mixture was diluted with water (7 ml) and extracted with chloroform (3x7 ml). The combined organic layers were washed with saturated aqueous NaHCO$_3$, dried over anhydrous Na$_2$SO$_4$, and concentrated under reduced pressure. The residue was purified by HPLC (method A) to afford compound **39** (40 mg, 49%) as a white solid. $^1$H NMR (400 MHz, DMSO-d6) δ 9.25 (s, 1H), 8.60 (s, 1H), 8.38 – 8.25 (m, 3H), 8.17 (d, J = 8.6 Hz, 2H), 7.43 (d, J = 8.1 Hz, 2H), 7.30 (t, J = 7.7 Hz, 2H), 7.17 (d, J = 8.7 Hz, 2H), 7.01 (t, J = 7.3 Hz, 1H), 4.18 (q, J = 7.0 Hz, 2H), 1.38 (t, J = 7.0 Hz, 3H). LC-MS (ESI) m/z 378.2 [(M+H)$^+$, calcd. C$_{21}$H$_{20}$N$_3$O$_4$, 378.1].

*1-(5-hydroxypyridin-3-yl)-3-phenylurea (40).* Compound **40** was prepared as described above for compound **38**, starting from 5-aminopyridin-3-ol hydrobromide (0.65 g, 3.4 mmol). The reaction mixture was concentrated under reduced pressure and the crude residue was purified by HPLC (method

B) to afford 60 mg (8%) of 1-(3-hydroxypyridin-5-yl)-3-phenylurea **40** as a white solid which was used for the next step without further purification.

*5-(3-phenylureido)pyridin-3-yl 4-ethoxybenzoate (41).* Compound **41** was prepared as described above for compound **39**, starting from 1-(5-hydroxypyridin-3-yl)-3-phenylurea **40** (60 mg, 0.264 mmol). The residue was purified by HPLC (method B) to afford compound **41** (36 mg, 45%) as a white solid. $^1$H NMR (400 MHz, DMSO-d6). δ 9.01 (s, 1H), 8.83 (s, 1H), 8.46 (q, J = 2.7 Hz, 1H), 8.16 (d, J = 2.7 Hz, 1H), 8.08 (td, J = 5.5, 2.2 Hz, 2H), 7.99 (t, J = 2.5 Hz, 1H), 7.45 (d, J = 7.8 Hz, 2H), 7.28 (t, J = 8.0 Hz, 2H), 7.11 (dd, J = 9.1, 2.3 Hz, 2H), 6.98 (t, J = 7.4 Hz, 1H), 4.16 (dt, J = 10.1, 6.6 Hz, 2H), 1.36 (td, J = 6.9, 2.4 Hz, 3H). LC-MS (ESI) m/z 378.2 [(M+H)$^+$, calcd. $C_{21}H_{20}N_3O_4$, 378.1].

4-(1-phenyl-3,6-dihydro-2H-pyridin-4-yl)pyridin-3-ol *(42)*. To a stirred solution of 4-iodopyridin-3-ol (0.63 g, 2.86 mmol, 1.1 eq.) and 1-phenyl-4-(4,4,5,5-tetramethyl-1,3,2-dioxaborolan-2-yl)-3,6-dihydro-2H-pyridine (0.74 g, 2.6 mmol, 1 eq.) in a mixture of 1,4-dioxane and water (20 ml, v/v=4:1), $K_2CO_3$ (1.8 g, 13 mmol, 5 eq.) was added and purged with argon for 30 min followed by the addition of Pd(dppf)Cl$_2$ (0.1 g, 0.05 eq.) and stirred at 90°C overnight. After completion, the reaction mixture was cooled to room temperature, diluted with ethyl acetate and water. The organic layer was washed with water and brine, dried over anhydrous sodium sulfate and evaporated under reduced pressure. The crude product was purified by column chromatography on silica gel (hexane/EtOAc) to afford **42** (251 mg, 38%).

*4-(1-phenyl-4-piperidyl)pyridin-3-ol (43).* Compound **42** (251 mg, 1 mmol) was dissolved in MeOH (20 ml), followed by addition of Pd (10 wt % on activated carbon, 50 mg), and then the resulting suspension was stirred at room temperature under 1 atm. hydrogen pressure overnight. The resulting reaction was filtered, concentrated under reduced pressure, and dried under vacuum, to afford **43** (201 mg, 79%) which was used for the next step without further purification.

*[4-(1-phenyl-4-piperidyl)-3-pyridyl] 4-ethoxybenzoate (44).* A solution of compound **43** (201 mg, 1 eq.), 4-ethoxybenzoic acid (131 mg, 1 eq.), Et$_3$N (0.27 ml, 2.5 eq.) and HATU (360 mg, 1.2 eq.) in dry DMSO (2 ml) was stirred at room temperature for 12h. The completion of the reaction was monitored by LCMS. The mixture was purified by HPLC (Method A) to give compound **44** (120 mg, 38% yield) as a white solid. 1H NMR (400 MHz, DMSO-d6). δ 8.46 (d, J = 5.4 Hz, 2H), 8.15 – 8.09 (m, 2H), 7.50 (d, J = 5.1 Hz, 1H), 7.22 – 7.10 (m, 4H), 6.93 (d, J = 8.2 Hz, 2H), 6.75 (t, J = 7.3 Hz, 1H), 4.16 (q, J = 6.9 Hz, 2H), 3.78 (d, J = 12.3 Hz, 2H), 2.87 – 2.79 (m, 1H), 2.63 (t, J = 10.0 Hz, 2H), 1.82 (t, J = 5.1 Hz, 4H), 1.37 (t, J = 7.0 Hz, 3H). LC-MS (ESI) m/z 403.2 [(M+H)+, calcd. $C_{25}H_{27}N_2O_3$, 403.2].

*5-(1-phenyl-3,6-dihydro-2H-pyridin-4-yl)pyridin-3-ol (45).* To a stirred solution of 5-iodopyridin-3-ol (0.63 g, 2.86 mmol, 1.1 eq.) and 1-phenyl-4-(4,4,5,5-tetramethyl-1,3,2-dioxaborolan-2-yl)-3,6-dihydro-2H-pyridine (0.74 g, 2.6 mmol, 1 eq.) in a mixture of 1,4-dioxane and water (20 ml, v/v=4:1), $K_2CO_3$ (1.8 g, 13 mmol, 5 eq.) was added and purged with argon for 30 min followed by the addition of

Pd(dppf)Cl$_2$ (0.1 g, 0.05 eq.) and stirred at 90 °C overnight. After completion, the reaction mixture was cooled to room temperature, diluted with ethyl acetate and water. The organic layer was washed with water and brine, dried over anhydrous sodium sulfate and evaporated under reduced pressure. The crude product was purified by column chromatography on silica gel (hexane/EtOAc) to afford compound **45** (326 mg, 49%).

*4-(1-phenyl-4-piperidyl)pyridin-3-ol (46).* Compound **45** (251 mg, 1 mmol) was dissolved in MeOH (20 ml), followed by addition of Pd (10 wt% on activated carbon, 50 mg), and then the resulting suspension was stirred at room temperature under 1 atm. hydrogen pressure overnight. The resulting reaction was filtered, concentrated under reduced pressure, and dried under vacuum, to afford compound **46** (220 mg, 86%) which was used for the next step without further purification.

*[5-(1-phenyl-4-piperidyl)-3-pyridyl] 4-ethoxybenzoate (47).* A solution of compound 46 (200 mg, 1 eq.), 4-ethoxybenzoic acid (131 mg, 1 eq.), Et$_3$N (0.27 mL, 2.5 eq.) and HATU (360 mg, 1.2 eq.) in dry DMSO (2 ml) was stirred at room temperature for 12h. The completion of the reaction was monitored by LCMS. The mixture was purified by HPLC (Method B) to give compound **47** (140 mg, 44% yield) as a white solid. 1H NMR (400 MHz, DMSO-d6) δ 8.48 (d, J = 1.8 Hz, 1H), 8.41 (d, J = 2.4 Hz, 1H), 8.12 – 8.05 (m, 2H), 7.71 (t, J = 2.2 Hz, 1H), 7.21 (dd, J = 8.6, 7.1 Hz, 2H), 7.15 – 7.09 (m, 2H), 6.98 (d, J = 7.8 Hz, 2H), 6.76 (t, J = 7.3 Hz, 1H), 4.16 (q, J = 7.0 Hz, 2H), 3.82 (d, J = 12.1 Hz, 2H), 2.88 – 2.71 (m, 2H), 2.54 (d, J = 1.0 Hz, 1H), 1.92 (d, J = 11.8 Hz, 2H), 1.81 (qd, J = 12.4, 3.9 Hz, 2H), 1.37 (t, J = 7.0 Hz, 3H). LC-MS (ESI) m/z 403.2 [(M+H)+, calcd. C$_{25}$H$_{27}$N$_2$O$_3$, 403.2].

*4-bromopyridin-3-yl 4-ethoxybenzoate (48).* A solution of 4-bromopyridin-3-ol (300 mg, 1.7 mmol, 1 eq.), 4-ethoxybenzoic acid (310 mg, 1.87 mmol, 1.1 eq.), DIPEA (0.89 ml, 5.1 mmol, 3 eq.) and HATU (760 mg, 2 mmol, 1.2 eq.) in DMF (10 ml) was stirred at 25°C for 16 h. The reaction mixture was poured into 50 ml of water and extracted with ethyl acetate (3x15 ml). The combined organic layers were washed with saturated ammonium chloride solution (50 ml) and brine (50 ml), dried over anhydrous sodium sulfate, and concentrated under reduced pressure to afford compound **48** as a brown solid (320 mg, purity 85%), which was used in the next step without further purification.

*4-(1-phenyl-1H-pyrazol-4-yl)pyridin-3-yl 4-ethoxybenzoate (49).* A mixture of compound **48** (200 mg, 0.62 mmol, 1 eq.), 1-(phenylpyrazol-4-yl)boronic acid (130 mg, 0.68 mmol, 1.1 eq.), cesium carbonate (400 mg, 1.24 mmol, 2 eq.) and Pd(dppf)Cl$_2$ (25 mg, 0.03 mmol, 0.05 eq.) in dioxane/water (5 ml, 10:1 v/v) was degassed and stirred at 105°C for 16 h under inert atmosphere. After cooling, the reaction mixture was poured into 30 ml of water and extracted with ethyl acetate (4x10 ml). The combined organic layers were washed with brine (20 ml), dried over anhydrous sodium sulfate, and concentrated under reduced pressure. The crude material was purified by HPLC (Method A) to afford compound **49** as a white solid (235 mg, 36% yield after 2 steps). ^{1}H NMR (400 MHz, DMSO-d6). δ 9.06 (s, 1H), 8.61 – 8.51 (m, 2H), 8.22 (d, J = 8.8 Hz, 2H), 8.14 (s, 1H), 7.88 (d, J = 5.1 Hz, 1H), 7.75 (d, J = 8.0 Hz, 2H),

7.50 (t, J = 7.8 Hz, 2H), 7.34 (t, J = 7.4 Hz, 1H), 7.16 (d, J = 8.8 Hz, 2H), 4.19 (q, J = 6.9 Hz, 2H), 1.38 (t, J = 6.9 Hz, 3H). LC-MS (ESI) m/z 386.0 [(M+H)$^+$, calcd. $C_{21}H_{20}N_3O_4$, 386.1].

*5-bromopyridin-3-yl 4-ethoxybenzoate (50).* Compound **50** was prepared as described above for compound **48**, starting from 5-bromopyridin-3-ol (300 mg, 1.7 mmol, 1 eq.) to afford a yellow solid (260 mg, purity 90%), which was used in the next step without further purification.

*5-(1-phenyl-1H-pyrazol-4-yl)pyridin-3-yl 4-ethoxybenzoate (51).* Compound **51** was prepared as described above for compound **49**, starting from 5-bromopyridin-3-yl 4-ethoxybenzoate **50** (200 mg, 0.62 mmol, 1eq.). The crude material was purified by HPLC (method B) to afford compound **51** as a white solid (50 mg, 8% yield after 2 steps). $^1$H NMR (400 MHz, DMSO-d6). δ 9.21 (s, 1H), 8.95 (d, J = 1.9 Hz, 1H), 8.44 (d, J = 2.6 Hz, 1H), 8.39 (s, 1H), 8.17 – 8.10 (m, 3H), 7.92 – 7.85 (m, 2H), 7.55 (t, J = 7.8 Hz, 2H), 7.35 (t, J = 7.2 Hz, 1H), 7.18 – 7.11 (m, 2H), 4.17 (q, J = 6.8 Hz, 2H), 1.38 (t, J = 6.8 Hz, 3H). LC-MS (ESI) m/z 386.0 [(M+H)$^+$, calcd. $C_{21}H_{20}N_3O_4$, 386.1].

**In vitro CDK8 inhibition**

Inhibitory activity of compounds was tested by using the LanthaScreen® Eu kinase binding assay optimized for CDK8/CyclinC (Invitrogen). This assay is based on the binding and displacement of an Alexa Fluor® 647-labeled ATP-competitive kinase inhibitor scaffold (kinase tracer) to the kinase. Binding of the tracer to the kinase is detected using a europium-labeled anti-tag antibody, which binds to the tagged CDK8/CyclinC. Simultaneous binding of both the tracer and antibody to the kinase results in a close proximity suitable for a high degree of FRET (fluorescence resonance energy transfer) from the europium (Eu) donor fluorophore to the Alexa Fluor® 647 acceptor fluorophore on the kinase tracer. Binding of an inhibitor to CDK8/CyclinC competes for binding with the tracer, resulting in a loss of FRET. Binding assay was performed into 384-well small volume plates (CORNING 3824) using kinase buffer provided by supplier (HEPES 50mM pH7.5, MgCl2 10mM, EGTA 1mM, Brij-35 0.01%) in a final volume of 15 µL. Briefly, 5µL of 3X compound (increasing concentrations from $3.10^{-11}$ to $3.10^{-5}$ M) prepared in kinase buffer are added to 5µL of 3X kinase/Ab solution (15nM kinase, 6nM biotin anti-His-tag antibody, 6nM Eu-streptavidin) and 5µL of 30nM kinase tracer236 (Kd 8 nM). The plate was incubated 1h at room temperature before reading with a TRF-compatible multi-well plate reader (Envision, PerkinElmer) using a classic TRF reading protocol (excitation at 337 nm; donor emission measured at 620 nm; acceptor emission measured at 665 nm). The TR-FRET signal was collected both at 665 and 620 nm, and TR-FRET ratios were calculated (acceptor signal value divided by donor signal value). $IC_{50}$ and $K_i$ values of the tested compounds were determined from competitive binding curves using GraphPad Prism software (version 6.07) as follows:

$$S = S_{min} + \frac{(S_{max} - S_{min})}{(1 + 10^{(X - logIC_{50})})}$$

S is the TR-FRET ratio value

X is the compound concentration

$$logIC_{50} = log10^{\left(logKi*\left(1+\frac{[tracer]}{K_d}\right)\right)}$$

[tracer] is the tracer concentration used in the competition assay

$K_d$ is the dissociation constant value of the tracer

### 4.2.7. Associated contents

**Safety Statement**

No unexpected or unusually high safety hazards were encountered. All experiments were conducted under ISO 9001 compliance.

**Supporting information**

Supplementary Methods section and additional figures and tables including the comparison and alignment of sc-PDB subpocket and fragments to CDK8 ATP binding site, the colored feature histogram (c-FH descriptor) used to align sc-PDB subpockets to the target cavity, the validation of the subpocket comparison protocol, the pairwise similarity of selected fragments, the properties of selected fragments, the definition of connectable fragments, the topological requirements to connect fragment atoms by a linker, the filters for DeLinker-generated linkers, the properties of generated molecules, the LC-MS analysis of compounds **39**, **41**, **44**, **47**, **49** and **51**, the synthesis of round-2 library compounds, the list of CDK8 X-ray structures, the filtering rules to select drug-like compounds, the in-house catalog of commercially available drug-like compounds, the list of 37 commercially available compounds structurally similar or identical to round-1 library members, the list of 151 round-2 library members (PDF).

Molecular formula strings–SMILES codes (CSV)

This material is available free of charge on the ACS Publications website at http://pubs.cas.org

**Abbreviations**

2D, two-dimensional; 3D, three-dimensional; AC, αC helix; CDK8, cyclin-dependent kinase 8; FBDD, fragment-based drug design; FRET, fluorescence resonance energy transfer; GA, gate area; HPLC, high performance liquid chromatography; HTRF, homogeneous time-resolved fluorescence; HTS, high-throughput screening; ITC, isothermal titration calorimetry; MD, molecular dynamics, MS, mass spectrometry; NMR, nuclear magnetic resonance spectroscopy; RMSD, root-mean square deviations; SE, solvent-exposed; SMILES, simplified molecular input line entry system; SPR, surface plasmon resonance; TR-FRET, time-resolved FRET.

## 4.2.8. References

1.  Erlanson, D. A.; Fesik, S. W.; Hubbard, R. E.; Jahnke, W.; Jhoti, H., Twenty Years On: The Impact of Fragments on Drug Discovery. *Nat. Rev. Drug. Discov.,* **2016,** *15*, 605-619.
2.  Li, Q., Application of Fragment-Based Drug Discovery to Versatile Targets. *Front. Mol. Biosci.,* **2020,** *7*, 180.
3.  Troelsen, N. S.; Clausen, M. H., Library Design Strategies to Accelerate Fragment-Based Drug Discovery. *Chem. Eur. J.,* **2020,** *26*, 11391-11403.
4.  Coyle, J.; Walser, R., Applied Biophysical Methods in Fragment-Based Drug Discovery. *SLAS Discov.,* **2020,** *25*, 471-490.
5.  Erlanson, D. A.; McDowell, R. S.; O'Brien, T., Fragment-Based Drug Discovery. *J. Med. Chem.,* **2004,** *47*, 3463-3482.
6.  Erlanson, D. A.; Davis, B. J.; Jahnke, W., Fragment-Based Drug Discovery: Advancing Fragments in the Absence of Crystal Structures. *Cell Chem. Biol.,* **2019,** *26*, 9-15.
7.  Bian, Y.; Xie, X. S., Computational Fragment-Based Drug Design: Current Trends, Strategies, and Applications. *AAPS J.,* **2018,** *20*, 59.

8. de Graaf, C.; Kooistra, A. J.; Vischer, H. F.; Katritch, V.; Kuijer, M.; Shiroishi, M.; Iwata, S.; Shimamura, T.; Stevens, R. C.; de Esch, I. J.; Leurs, R., Crystal Structure-Based Virtual Screening for Fragment-Like Ligands of the Human Histamine H(1) Receptor. *J. Med. Chem.,* **2011,** *54*, 8195-8206.

9. Brooijmans, N.; Kuntz, I. D., Molecular Recognition and Docking Algorithms. *Annu. Rev. Biophys. Biomol. Struct.,* **2003,** *32*, 335-373.

10. Sandor, M.; Kiss, R.; Keseru, G. M., Virtual Fragment Docking by Glide: A Validation Study on 190 Protein-Fragment Complexes. *J. Chem. Inf. Model.,* **2010,** *50*, 1165-1172.

11. Verdonk, M. L.; Giangreco, I.; Hall, R. J.; Korb, O.; Mortenson, P. N.; Murray, C. W., Docking Performance of Fragments and Druglike Compounds. *J. Med. Chem.,* **2011,** *54*, 5422-5431.

12. Marcou, G.; Rognan, D., Optimizing Fragment and Scaffold Docking by Use of Molecular Interaction Fingerprints. *J. Chem. Inf. Model.,* **2007,** *47*, 195-207.

13. Jacquemard, C.; Drwal, M. N.; Desaphy, J.; Kellenberger, E., Binding Mode Information Improves Fragment Docking. *J. Cheminform.,* **2019,** *11*, 24.

14. Chachulski, L.; Windshugel, B., Leads-Frag: A Benchmark Data Set for Assessment of Fragment Docking Performance. *J. Chem. Inf. Model.,* **2020,** *60*, 6544-6554.

15. Guvench, O., Computational Functional Group Mapping for Drug Discovery. *Drug Discov. Today,* **2016,** *21*, 1928-1931.

16. Kozakov, D.; Grove, L. E.; Hall, D. R.; Bohnuud, T.; Mottarella, S. E.; Luo, L.; Xia, B.; Beglov, D.; Vajda, S., The Ftmap Family of Web Servers for Determining and Characterizing Ligand-Binding Hot Spots of Proteins. *Nat. Protoc.,* **2015,** *10*, 733-755.

17. Radoux, C. J.; Olsson, T. S.; Pitt, W. R.; Groom, C. R.; Blundell, T. L., Identifying Interactions That Determine Fragment Binding at Protein Hotspots. *J. Med. Chem.,* **2016,** *59*, 4314-4325.

18. Guvench, O.; MacKerell, A. D., Jr., Computational Fragment-Based Binding Site Identification by Ligand Competitive Saturation. *PLoS Comput. Biol.,* **2009,** *5*, e1000435.

19. Chen, H.; Zhou, X.; Wang, A.; Zheng, Y.; Gao, Y.; Zhou, J., Evolutions in Fragment-Based Drug Design: The Deconstruction-Reconstruction Approach. *Drug Discov. Today,* **2015,** *20*, 105-113.

20. Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M., Recap--Retrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci.,* **1998,** *38*, 511-522.

21. Degen, J.; Wegscheid-Gerlach, C.; Zaliani, A.; Rarey, M., On the Art of Compiling and Using 'Drug-Like' Chemical Fragment Spaces. *ChemMedChem,* **2008,** *3*, 1503-1507.

22. Maass, P.; Schulz-Gasch, T.; Stahl, M.; Rarey, M., Recore: A Fast and Versatile Method for Scaffold Hopping Based on Small Molecule Crystal Structure Conformations. *J. Chem. Inf. Model.,* **2007,** *47*, 390-399.

23. Imrie, F.; Bradley, A. R.; van der Schaar, M.; Deane, C. M., Deep Generative Models for 3D Linker Design. *J. Chem. Inf. Model.,* **2020,** *60*, 1983-1995.

24. Yang, Y.; Zheng, S.; Su, S.; Zhao, C.; Xu, J.; Chen, H., Syntalinker: Automatic Fragment Linking with Deep Conditional Transformer Neural Networks. *Chem. Sci.,* **2020,** *11*, 8312-8322.

25. Imrie, F.; Hadfield, T. E.; Bradley, A. R.; Deane, C. M., Deep Generative Design with 3D Pharmacophoric Constraints. *Chem. Sci.,* **2021,** *12*, 14577-14589.

26. Bhagavat, R.; Sankar, S.; Srinivasan, N.; Chandra, N., An Augmented Pocketome: Detection and Analysis of Small-Molecule Binding Pockets in Proteins of Known 3D Structure. *Structure,* **2018,** *26*, 499-512 e492.

27. Gao, M.; Skolnick, J., A Comprehensive Survey of Small-Molecule Binding Pockets in Proteins. *PLoS Comput Biol,* **2013,** *9*, e1003302.

28. Eguida, M.; Rognan, D., A Computer Vision Approach to Align and Compare Protein Cavities: Application to Fragment-Based Drug Design. *J. Med. Chem.,* **2020,** *63*, 7127-7142.

29. Dale, T.; Clarke, P. A.; Esdar, C.; Waalboer, D.; Adeniji-Popoola, O.; Ortiz-Ruiz, M. J.; Mallinger, A.; Samant, R. S.; Czodrowski, P.; Musil, D.; Schwarz, D.; Schneider, K.; Stubbs, M.; Ewan, K.; Fraser, E.; TePoele, R.; Court, W.; Box, G.; Valenti, M.; de Haven Brandon, A.; Gowan, S.; Rohdich, F.; Raynaud, F.; Schneider, R.; Poeschke, O.; Blaukat, A.; Workman, P.; Schiemann, K.; Eccles, S. A.; Wienke, D.; Blagg, J., A Selective Chemical Probe for Exploring the Role of CDK8 and CDK19 in Human Disease. *Nat. Chem. Biol.,* **2015,** *11*, 973-980.

30. Mallinger, A.; Schiemann, K.; Rink, C.; Stieber, F.; Calderini, M.; Crumpler, S.; Stubbs, M.; Adeniji-Popoola, O.; Poeschke, O.; Busch, M.; Czodrowski, P.; Musil, D.; Schwarz, D.; Ortiz-Ruiz, M. J.; Schneider, R.; Thai, C.; Valenti, M.; Brandon, A. D.; Burke, R.; Workman, P.; Dale, T.; Wienke, D.; Clarke, P. A.; Esdar, C.; Raynaud, F. I.; Eccles, S. A.; Rohdich, F.; Blagg, J., Discovery of Potent, Selective, and Orally Bioavailable Small-Molecule Modulators of the Mediator Complex-Associated Kinases CDK8 and CDK19. *J. Med. Chem.,* **2016,** *59*, 1078-1101.

31. Desaphy, J.; Bret, G.; Rognan, D.; Kellenberger, E., sc-PDB: A 3D-Database of Ligandable Binding Sites--10 Years On. *Nucleic Acids Res.,* **2015,** *43*, D399-404.

32. Desaphy, J.; Rognan, D., sc-PDB-Frag: A Database of Protein-Ligand Interaction Patterns for Bioisosteric Replacements. *J. Chem. Inf. Model.,* **2014,** *54*, 1908-1918.

33. Desaphy, J.; Azdimousa, K.; Kellenberger, E.; Rognan, D., Comparison and Druggability Prediction of Protein-Ligand Binding Sites from Pharmacophore-Annotated Cavity Shapes. *J. Chem. Inf. Model.,* **2012,** *52*, 2287-2299.

34. van Linden, O. P.; Kooistra, A. J.; Leurs, R.; de Esch, I. J.; de Graaf, C., KLIFS: A Knowledge-Based Structural Database to Navigate Kinase-Ligand Interaction Space. *J. Med. Chem.,* **2014,** *57*, 249-277.

35. Rusu, R. B.; Cousins, S., 3D Is Here: Point Cloud Library (PCL). *IEEE Int. Conf. Robot.,* **2011**, 1, 1-4. https://doi.org/10.1109/ICRA.2011.5980567 (accessed 05-10-2022).

36. Zhou, Q.-Y.; Park, J.; Koltun, V., Open3D: A Modern Library for 3D Data Processing. *arXiv:1801.09847,* **2018** (accessed 05-10-2022).

37. Congreve, M.; Carr, R.; Murray, C.; Jhoti, H., A 'Rule of Three' for Fragment-Based Lead Discovery? *Drug Discov. Today,* **2003,** *8*, 876-877.

38. Sterling, T.; Irwin, J. J., Zinc 15--Ligand Discovery for Everyone. *J. Chem. Inf. Model.,* **2015,** *55*, 2324-2337.

39. Wang, R.; Fang, X.; Lu, Y.; Wang, S., The PDBbind Database: Collection of Binding Affinities for Protein-Ligand Complexes with Known Three-Dimensional Structures. *J. Med. Chem.,* **2004,** *47*, 2977-2980.

40. Ertl, P.; Schuffenhauer, A., Estimation of Synthetic Accessibility Score of Drug-Like Molecules Based on Molecular Complexity and Fragment Contributions. *J. Cheminform.,* **2009,** *1*, 8.

41. Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P., Chembl: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.,* **2012,** *40*, D1100-1107.

42. Sydow, D.; Schmiel, P.; Mortier, J.; Volkamer, A., KinFragLib: Exploring the Kinase Inhibitor Space Using Subpocket-Focused Fragmentation and Recombination. *J. Chem. Inf. Model.,* **2020,** *60*, 6081-6094.

43. Li, Y. B.; Pei, J. F.; Lai, L. H., Structure-Based De Novo Drug Design Using 3D Deep Generative Models. *Chem.Sci.,* **2021,** *12*, 13664-13675.

44. The Protein Data Bank, https://www.rcsb.org (accessed 05-10-2022).

45. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E., The Protein Data Bank. *Nucleic Acids Res.,* **2000,** *28*, 235-242.

46. Bietz, S.; Urbaczek, S.; Schulz, B.; Rarey, M., Protoss: A Holistic Approach to Predict Tautomers and Protonation States in Protein-Ligand Complexes. *J. Cheminform.,* **2014,** *6*, 12.

47. Da Silva, F.; Desaphy, J.; Rognan, D., Ichem: A Versatile Toolkit for Detecting, Comparing, and Predicting Protein-Ligand Interactions. *ChemMedChem,* **2018,** *13*, 507-510.

48. Openeye Scientific Sofware, Santa Fe, NM 87508, U.S.A., https://www.eyesopen.com (accessed 05-10-20222).

49. RDKit: Open-Source Cheminformatics Software, http://www.rdkit.org (accessed 05-10-2022)

50. KinFragLib, https://zenodo.org/record/3956580 (accessed 05-10-2022).

51. ChEMBL, https://www.ebi.ac.uk/chembl (accessed 05-10-2022).

52. Korb, O.; Stutzle, T.; Exner, T. E., Empirical Scoring Functions for Advanced Protein-Ligand Docking with Plants. *J. Chem. Inf. Model.,* **2009,** *49*, 84-96.

53. Matplotlib, https://matplotlib.org/3.0.2 (accessed 05-10-2022).

## 4.2.9. Supporting Information for *Target-focused library design by pocket-applied computer vision and fragment deep generative linking*

**Figure S1**. Comparison and alignment of sc-PDB subpocket and fragments to CDK8 ATP binding site.

**Figure S2**. Colored Feature Histogram (c-FH descriptor) used to align sc-PDB subpockets to the target cavity.

**Figure S3**.  Validation of the subpocket comparison protocol.

**Figure S4**.  Pairwise similarity of the 385 selected fragments.

**Figure S5**. Properties of selected fragments.

**Figure S6**. Connectable fragments are defined by connectable areas.

**Figure S7**. Topological requirements to connect fragment atoms by a linker.

**Figure S8**. Filters for DeLinker-generated linkers.

**Figure S9**. Properties of generated molecules.

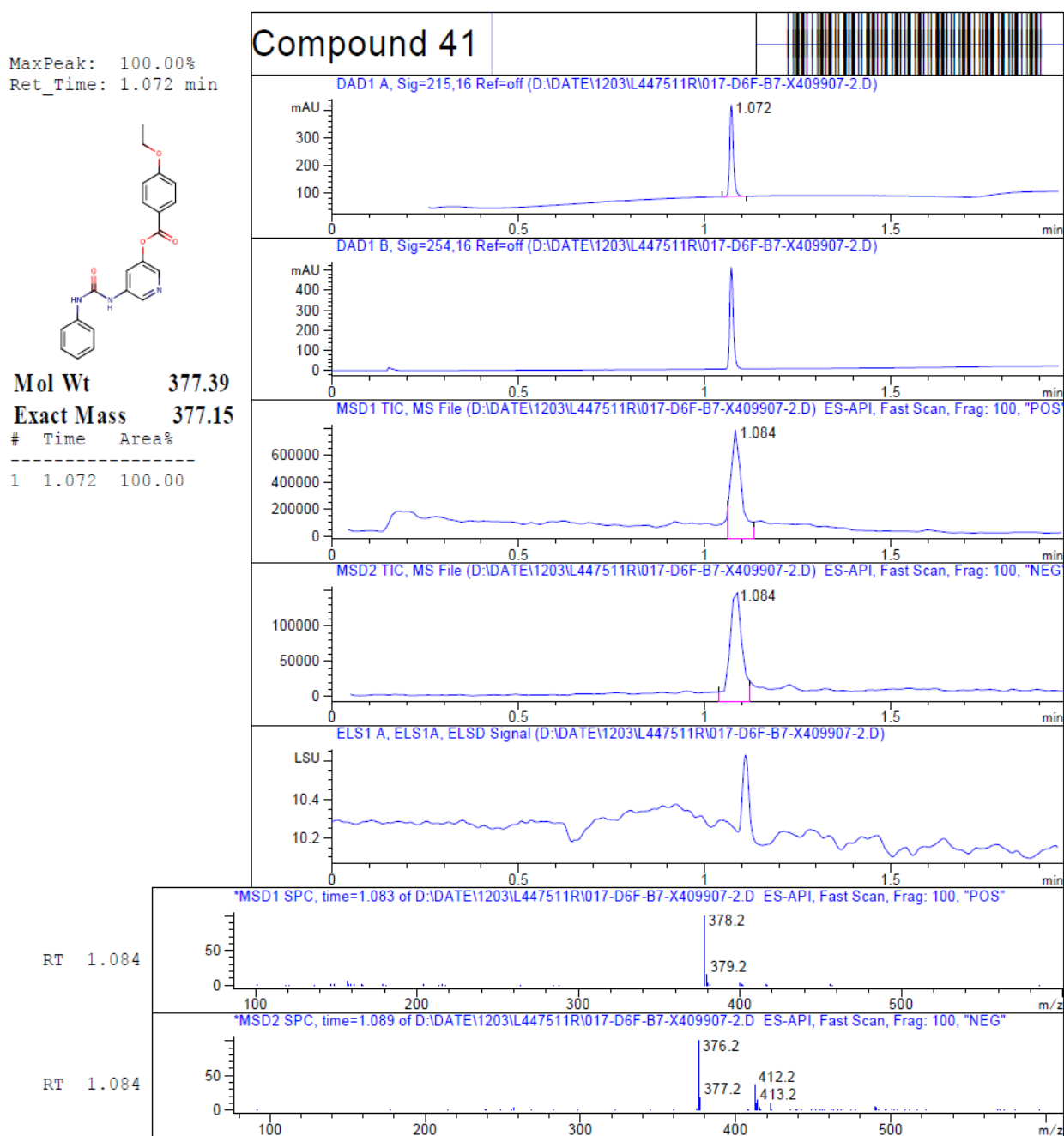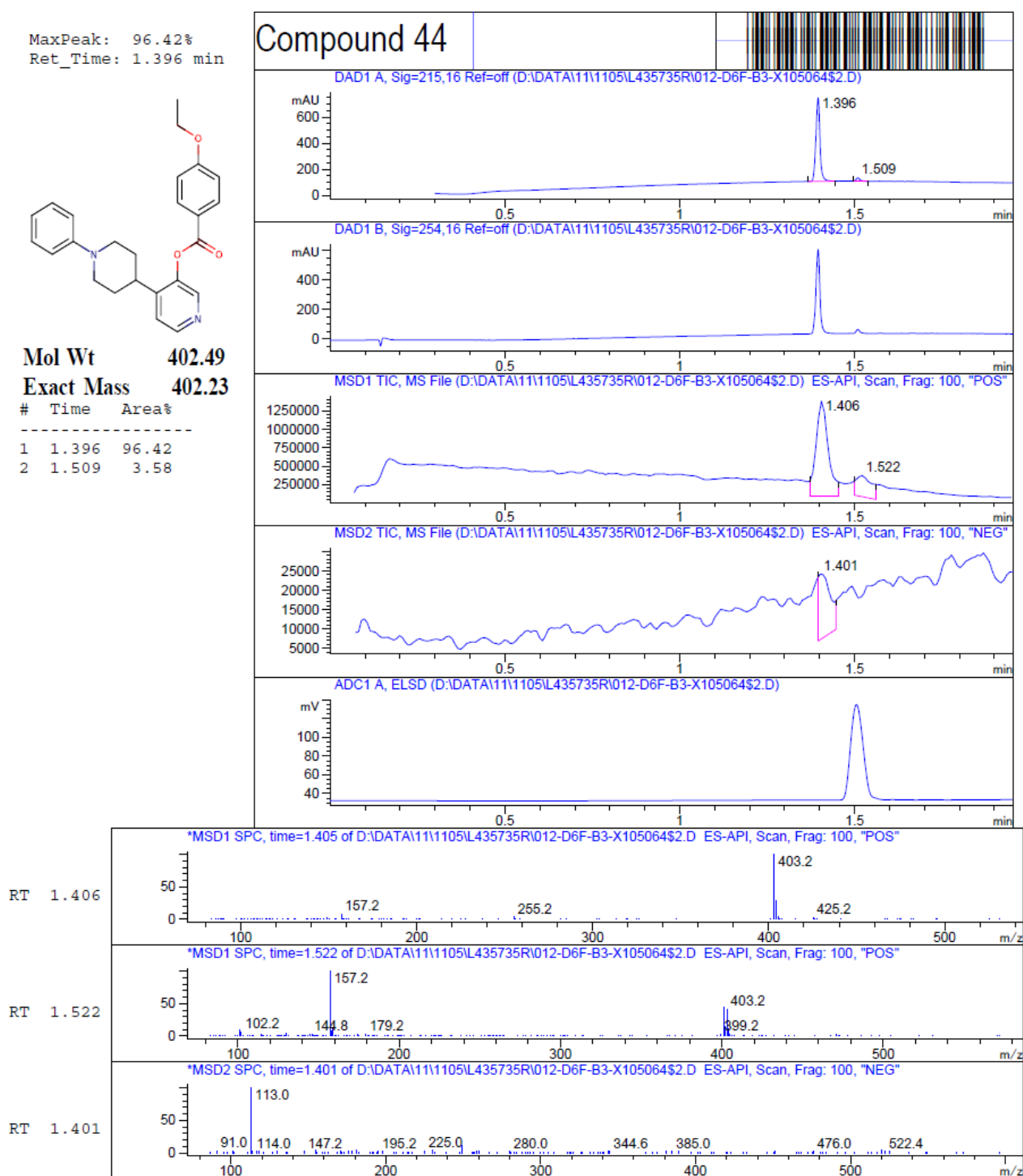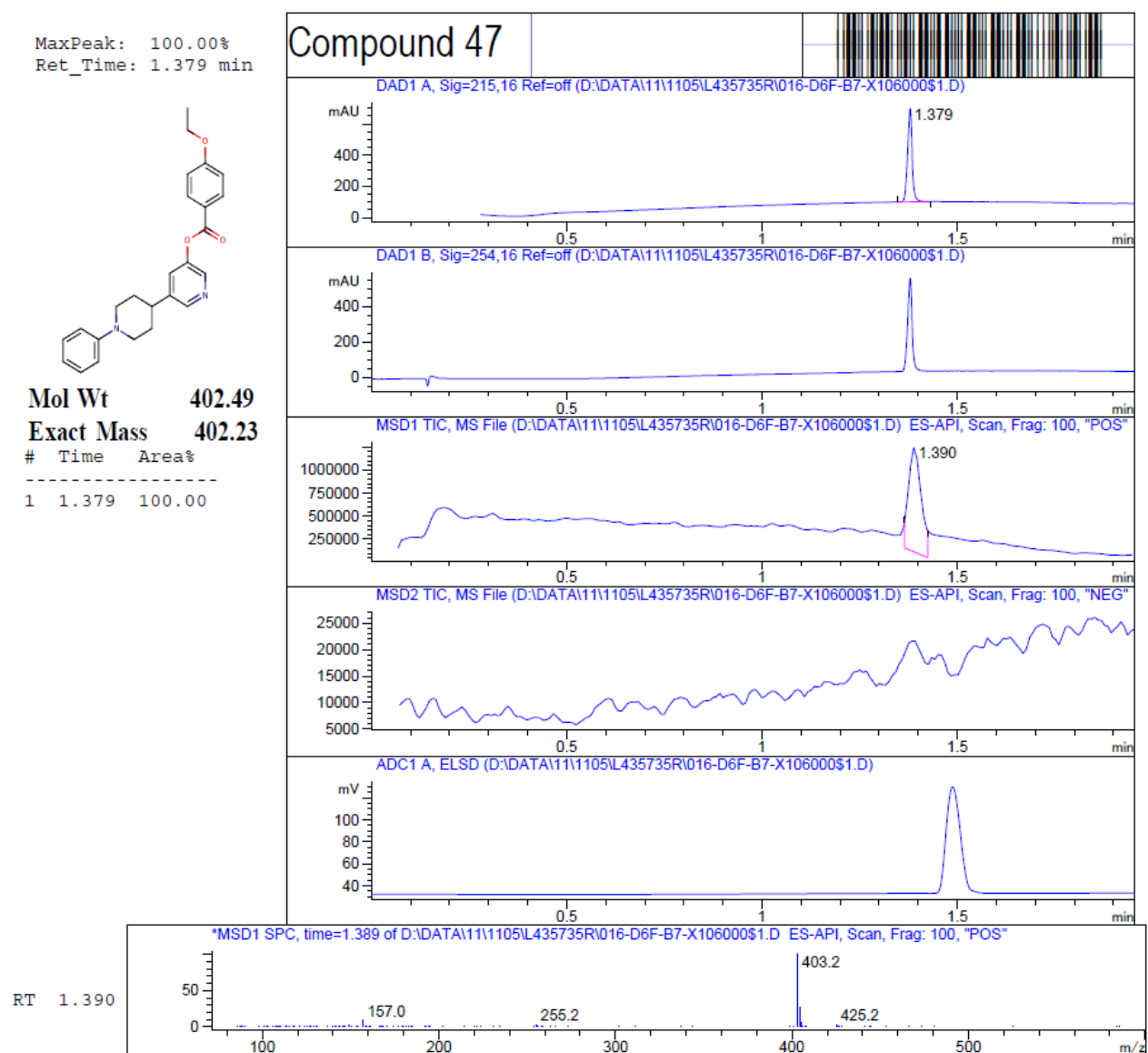**Figure S10.** LC-MS analysis of compound **39**.

**Figure S11.** LC-MS analysis of compound **41**.

**Figure S12.** LC-MS analysis of compound **44**.
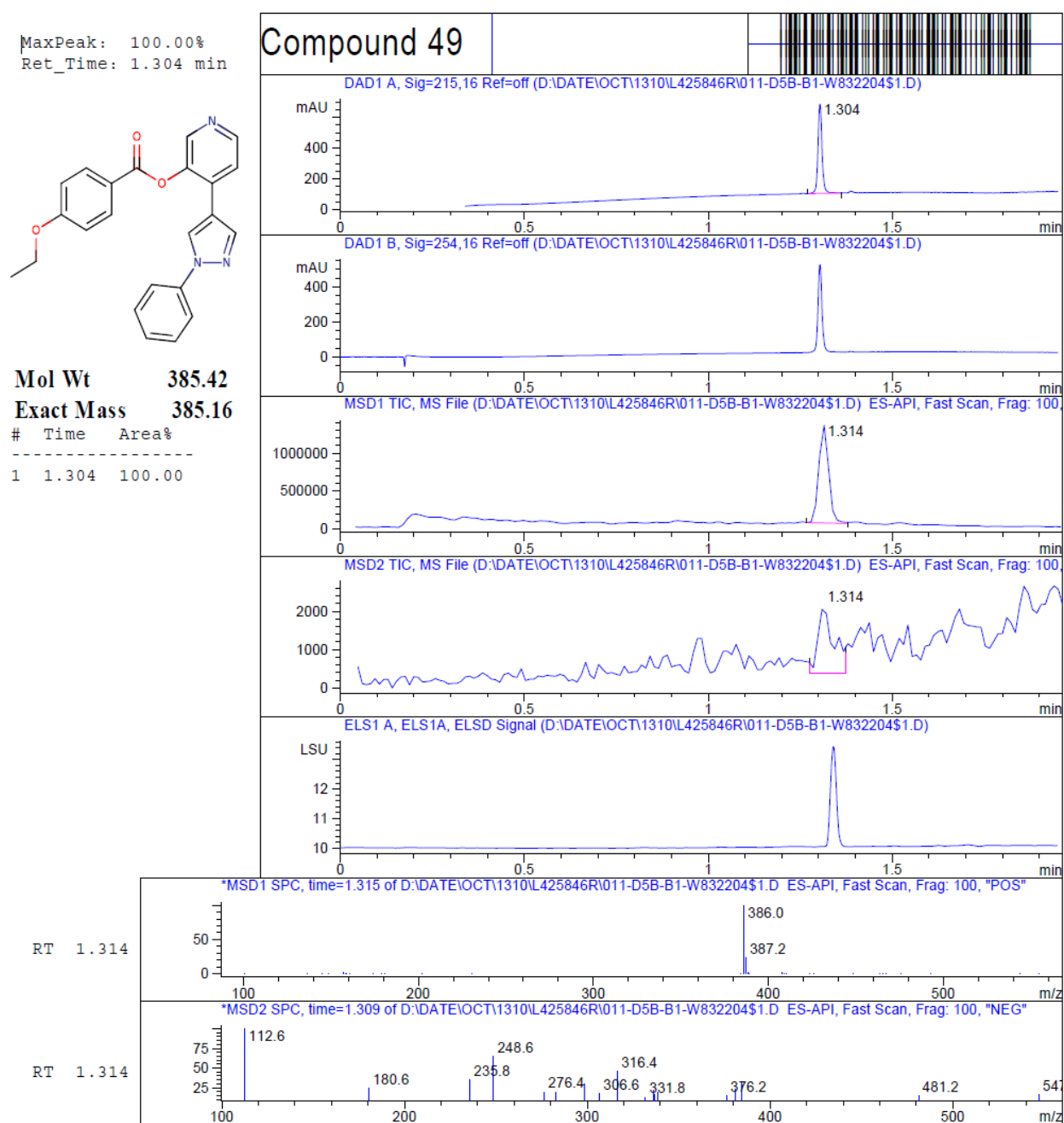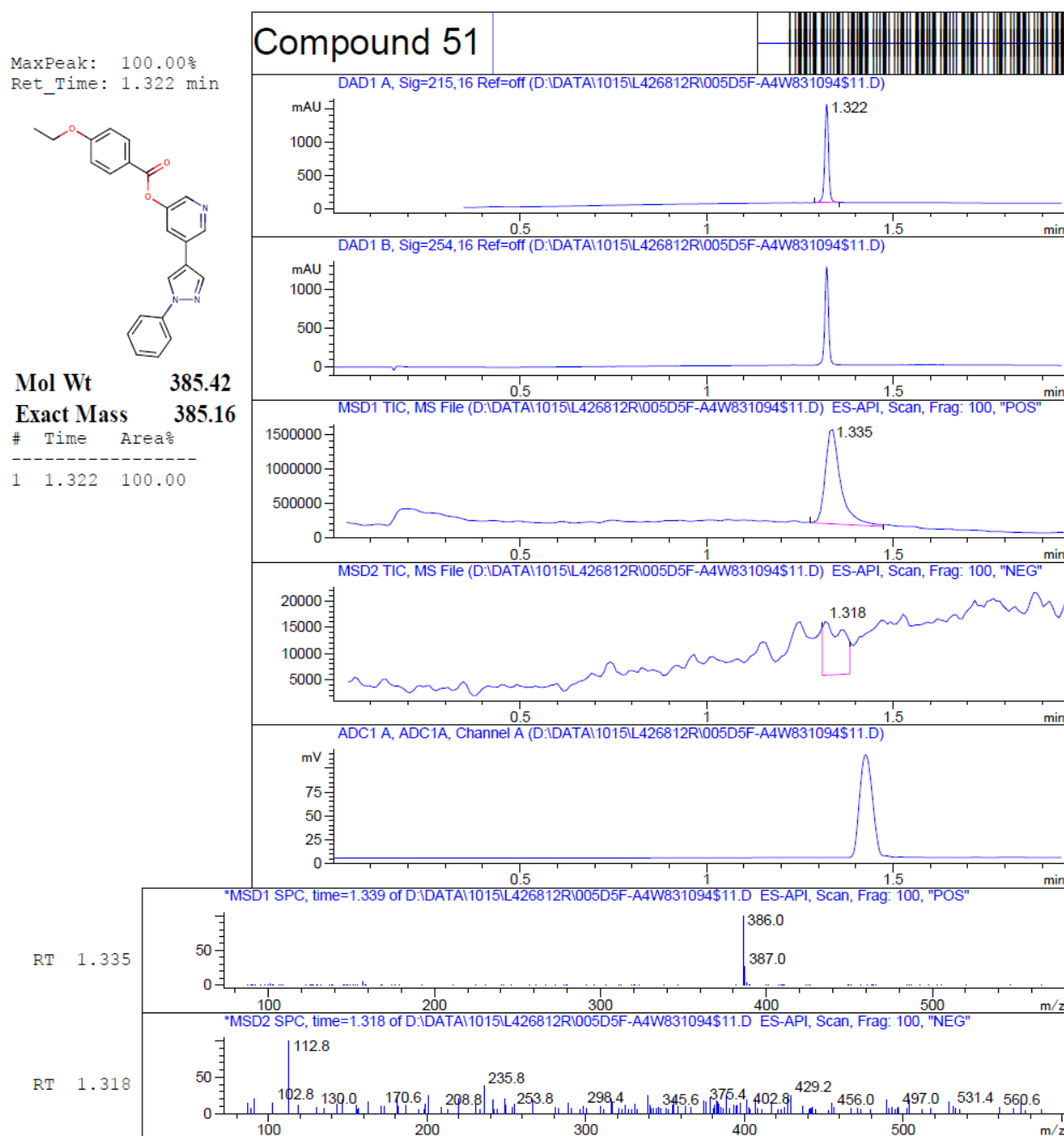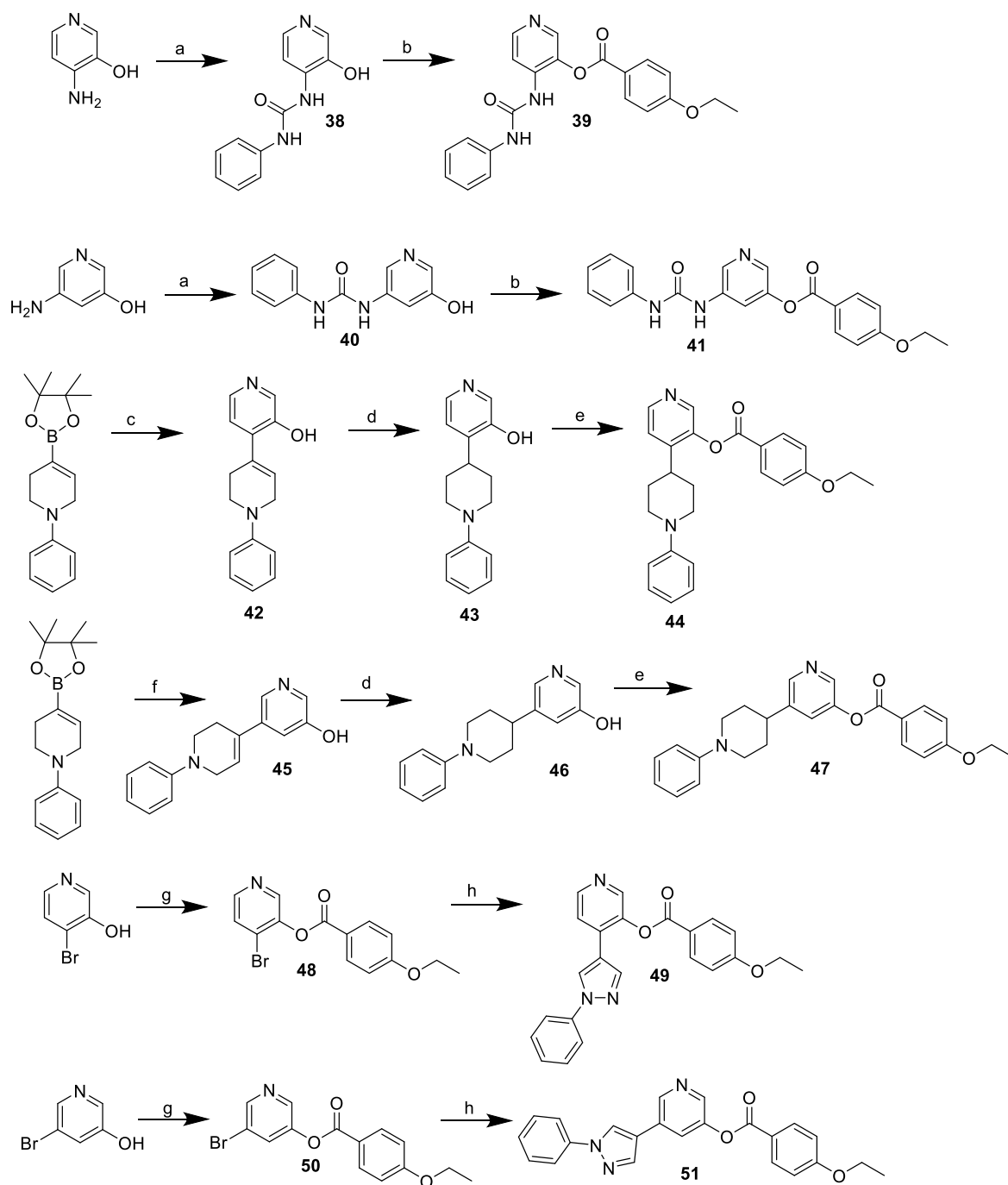
**Figure S13.** LC-MS analysis of compound **47**.

**Figure S14.** LC-MS analysis of compound **49**.

**Figure S15.** LC-MS analysis of compound **51**.

**Scheme S1.** Synthesis of round-2 library compounds.

**Table S1.** List of CDK8 X-ray structures.

**Table S2.** Filtering rules to select drug-like compounds.

**Table S3.** In-house catalog of commercially available drug-like compounds.

**Table S4.** List of 37 commercially available compounds, structurally similar or identical to round-1 library members.

**Table S5.** List of 151 round-2 library members (SMILES strings)

**A**
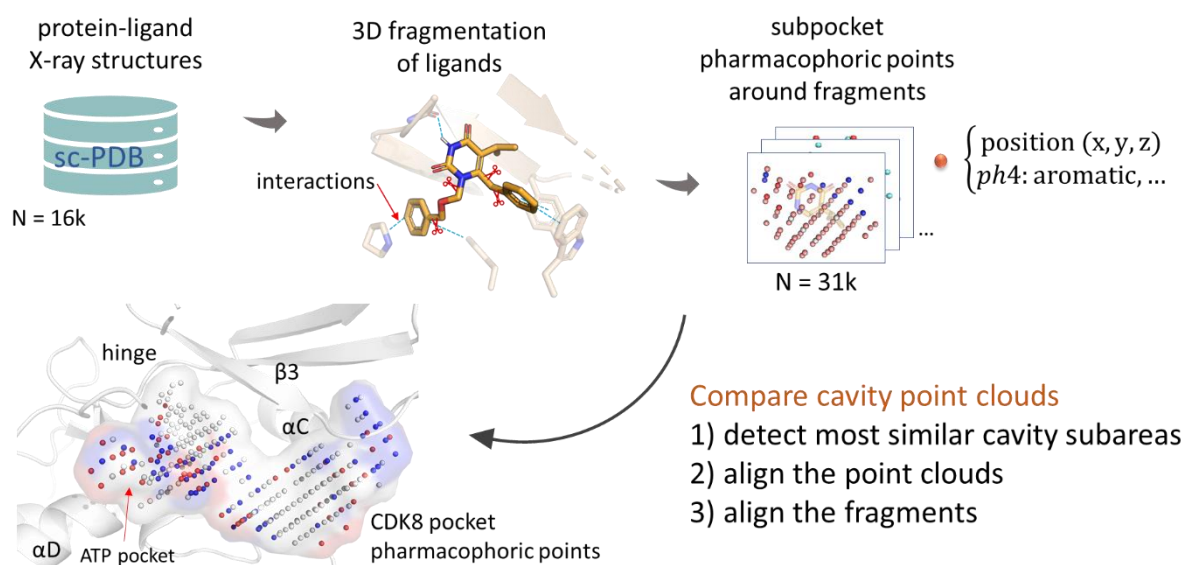


**B**
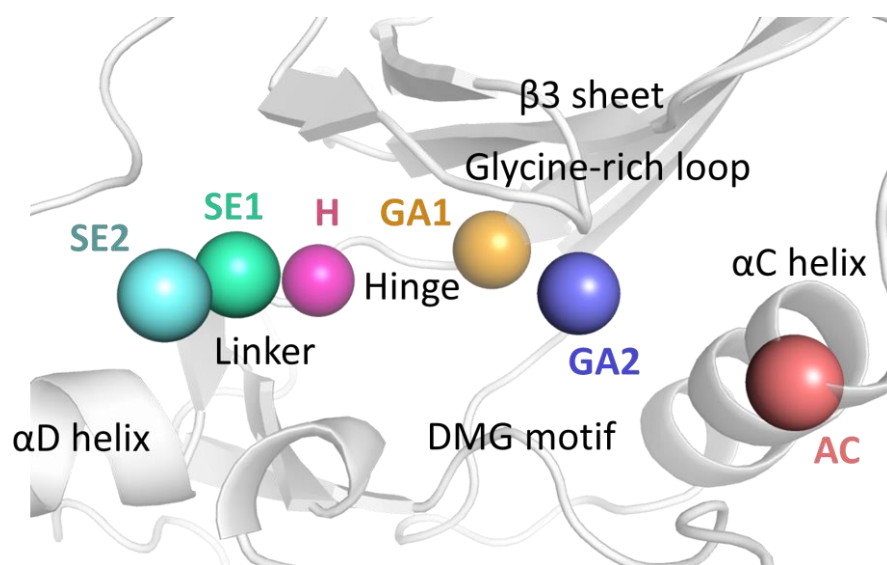


**Figure S1**. Alignment of sc-PDB[1] subpockets and fragments to CDK8 ATP binding site. **A)** Overall alignment flowchart, **B)** CDK8 areas hinge (H), gate area 1 (GA1), gate area 2 (GA2), solvent-exposed area 1 (SE1), solvent-exposed area 2 (SE2), αC area (AC).

**A)**

$\mathcal{P}^k$

$r$

$P_q$

**B)** feature proportion in $\mathcal{P}^k$

1 ... 8

c-$SFH(P_q)$

**C)**

c-$SFH(P_1)$

c-$SFH(P_2)$

c-$SFH(P_3)$

**D)**

$$\text{c-}FH(P_q) = \text{c-}SFH(P_q) + \frac{1}{k}\sum_{i=1}^{k} \frac{1}{\|P_i - P_q\|} \cdot \text{c-}SFH(P_i), \quad P_i \in \mathcal{P}^k$$
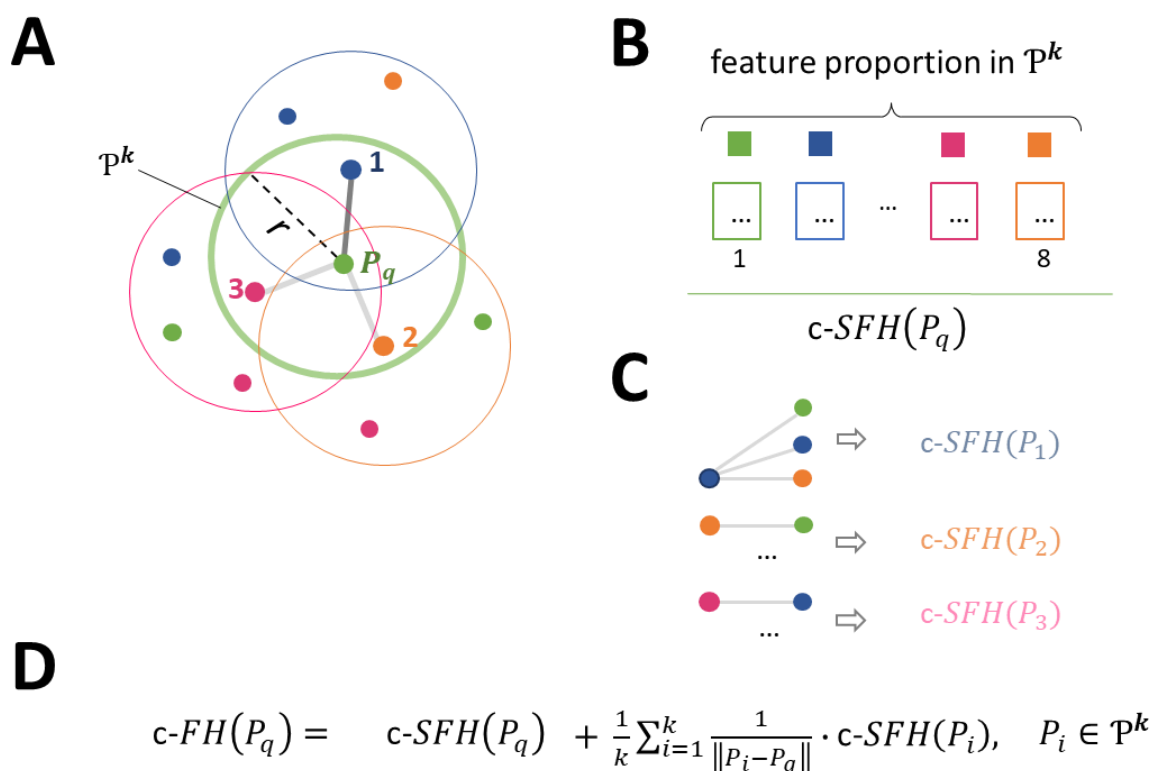
**Figure S2**. Colored Feature Histogram (c-FH descriptor)[2] used to align sc-PDB subpockets to the target cavity. **A)** Considering a point Pq (green) whose c-FH is to be computed, its neighbor points $\mathcal{P}^k$ = {1, 2, 3} within a radius r are determined (green circle). For each neighbor in $\mathcal{P}^k$, their respective neighbors are also determined within the radius r. **B)** The percentage of each of eight pharmacophoric features (hydrophobic, aromatic, H-bond donor, H-bond acceptor, H-bond acceptor and donor, positive ionizable, negative ionizable, null) is then stored into a 8-bin histogram that forms the simplified colored feature histogram (c-SFH) of the point Pq. **C)** The c-SFH is iteratively computed for each point in $\mathcal{P}^k$; **D)** The c-FH of the point Pq is the sum of its c-SFH and the distance-weighted average of its neighbors' c-FSHs.
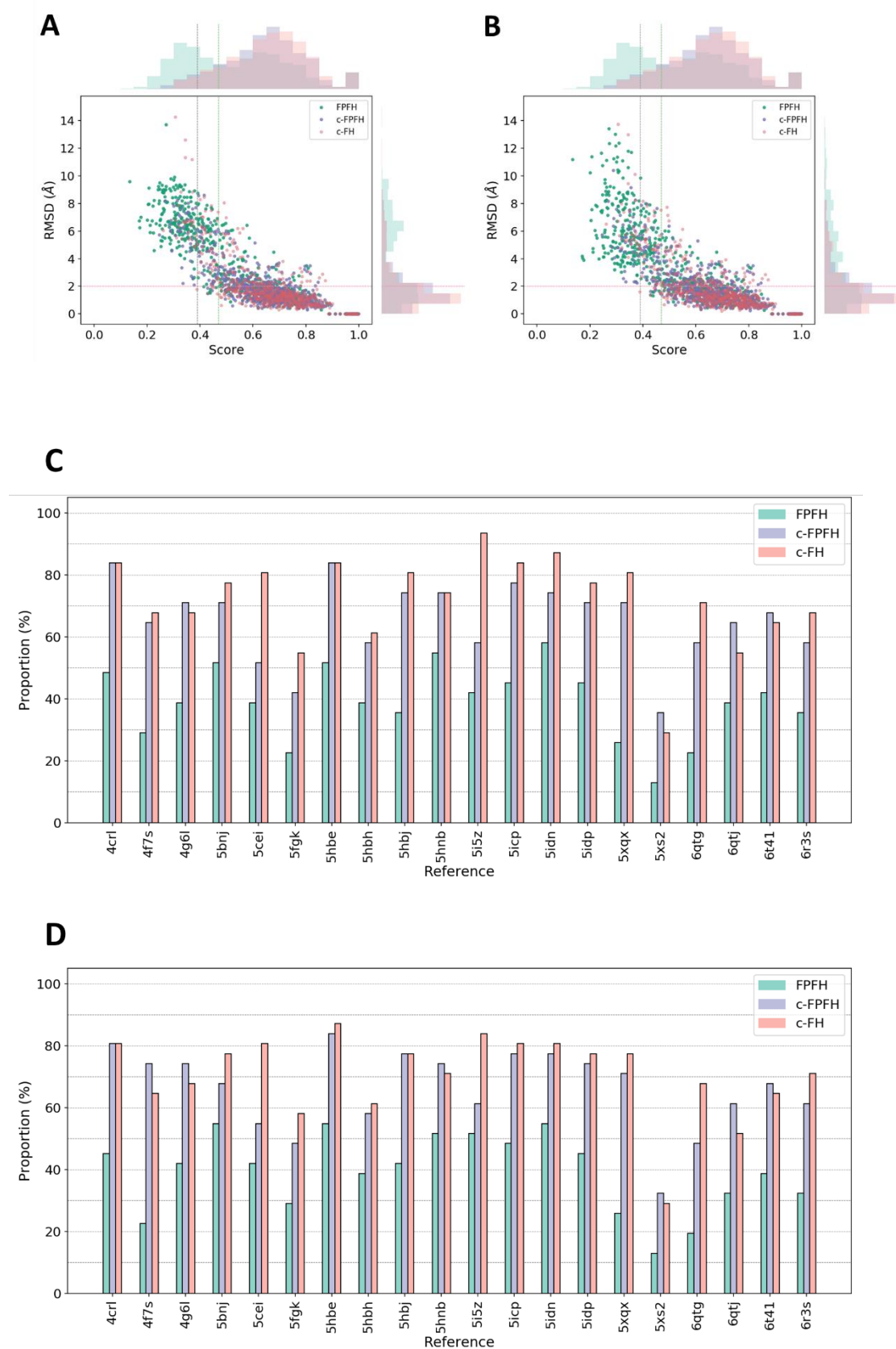
**Figure S3.** Validation of the subpocket comparison protocol. Cross-alignment of CDK8 subpockets and corresponding fragments to CDK8 full cavities. The bound inhibitors of 20 structurally-prealigned

(Maestro, Schrödinger, New York, NY 10036, U.S.A.) CDK8 PDB entries (19 type I and one apo structures) were fragmented as described in the Methods section. The immediate protein environment of each selected fragment defines the corresponding subpocket that is represented as a pharmacophore feature-annotated point cloud. After translating and rotating the subpockets and their bound fragments in a different coordinates frame, each subpocket from the 20 PDB entries was aligned to the 20 entire CDK8 cavities with ProCare using three different fingerprints: c-FPFH (colored Fast Point Feature Histogram, violet) encoding both local shape and pharmacophoric properties distributions, c-FH (colored Feature Histogram, pink) encoding local  pharmacophoric properties distributions only and FPFH (Fast Point Feature Histogram, green) encoding local shape only. The optimal transformation matrix is next applied to the accompanying subpocket-bound fragment to pose each fragment into the full cavities. **A)** Root-mean square deviation (RMSD) of ProCare-aligned subpockets from corresponding protein structure-based prealigned subpocket with respect to the ProCare score. Green dashed line: score default threshold (0.47, p-value: 0.05), grey dashed line: optimal score threshold used in this study (0.39, corresponding to the maximum F-measure discriminating known similar and known dissimilar cavities)16. **B)** RMSD of ProCare-aligned fragments from corresponding protein structure-based prealigned fragments with respect to the ProCare score. **C)** Proportion of aligned subpockets with a RMSD less than 2 Å from the corresponding protein structure-based prealigned subpocket; **D)** Proportion of aligned fragments with a RMSD of their heavy atoms less than 2 Å from the corresponding protein structure-based prealigned fragment.
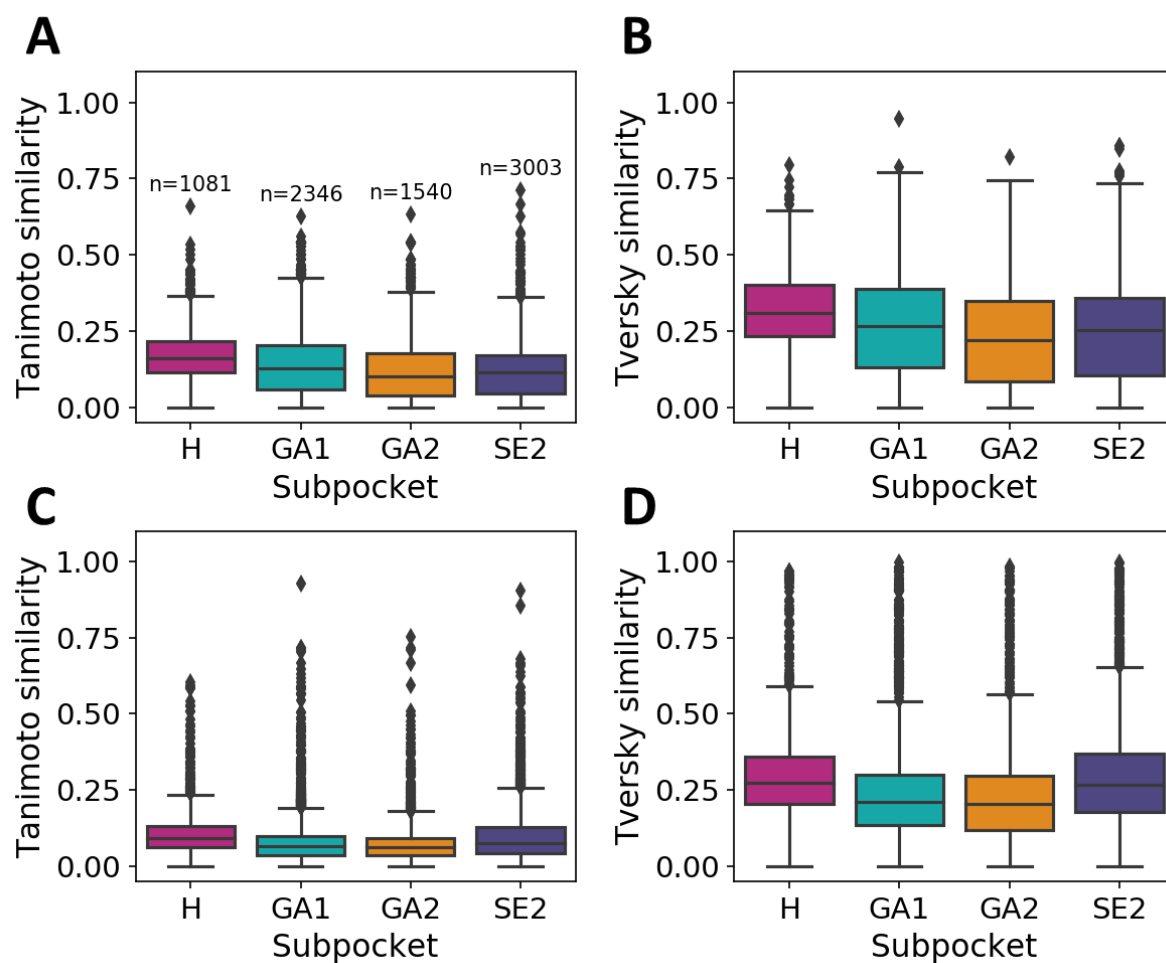
**Figure S4**. Pairwise similarity of the 385 selected fragments after removing 2D duplicates for each CDK8 area: H: hinge, GA1: gate area 1, GA2: gate area 2, SE2: solvent-exposed area 2. **A)** Tanimoto and **B)** Tversky metrics on RDKit Morgan fingerprint (radius = 2). **C)** Tanimoto and **D)** Tversky metrics on RDKit topological fingerprint (maximum path size: 7). Tversky similarity corresponds to the maximum possible, applying the largest weight (0.95) to the smallest molecule and the smallest weight (0.05) to the largest molecule. Outliers are computed to be outside the quartiles past 1.5 times the interquartile range.
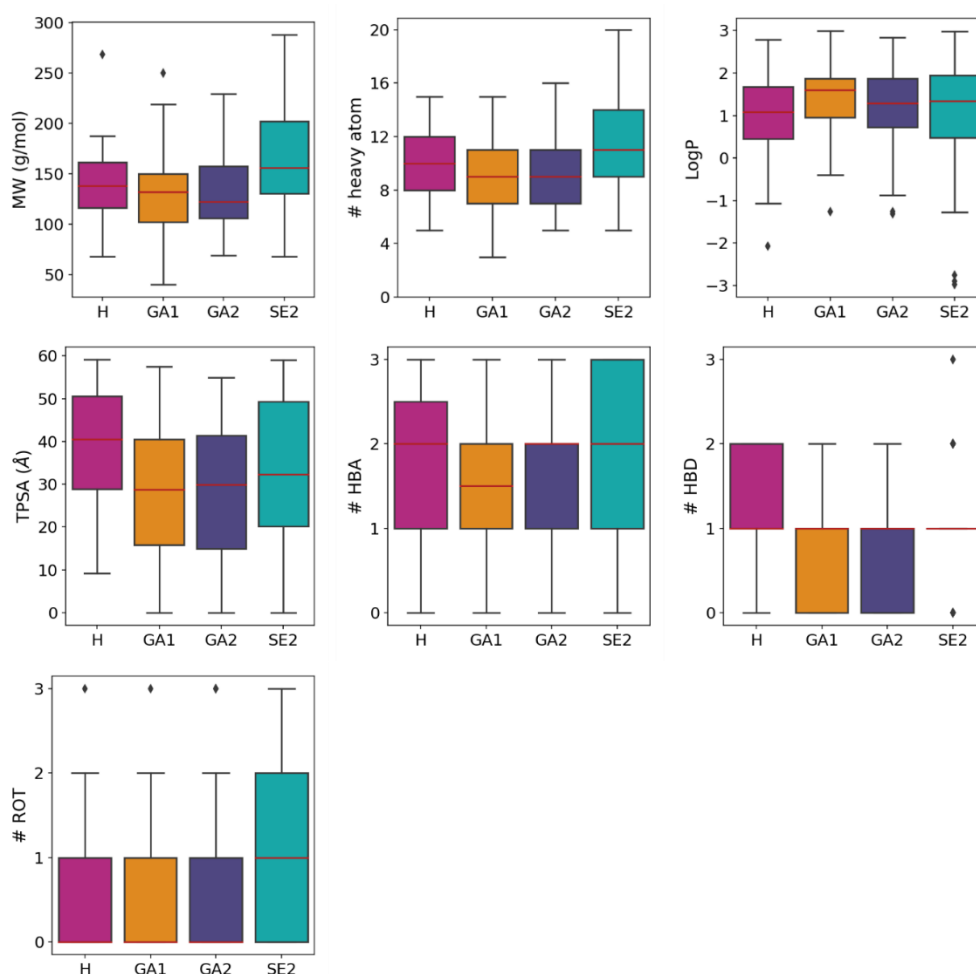
**Figure S5**. Properties of selected fragments after removing 2D duplicates for each CDK8 area: H: hinge, GA1: gate area 1, GA2: gate area 2, SE2: solvent-exposed area 2. From left to right, top to bottom, the molecular weight (g.mol⁻¹), the count of heavy atoms (non-hydrogen atoms), calculated logP, polar surface area (Å), count of H-bond acceptor, count of H-bond donor, count of rotatable bonds. The seven properties were calculated with RDKit. Outliers are computed to be outside the quartiles past 1.5 times the interquartile range.
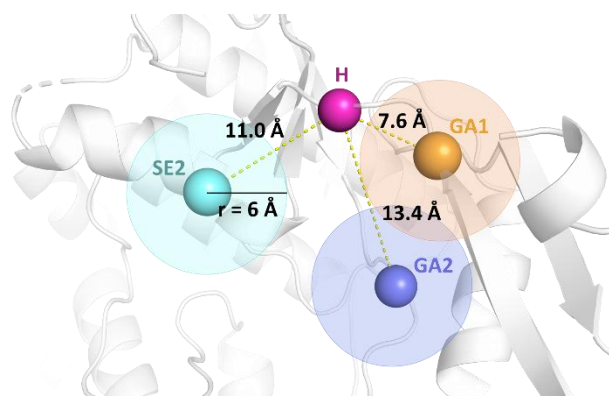


**Figure S6**. Connectable fragments are defined by connectable areas: hinge (H)-annotated fragments are paired with fragments from the gate area 1 (GA1), the gate area 2 (GA2), and the solvent-exposed area 2 (SE2). Spheres of 6 Å radius delineate each CDK8 area. Distances between area centers are reported in Å.
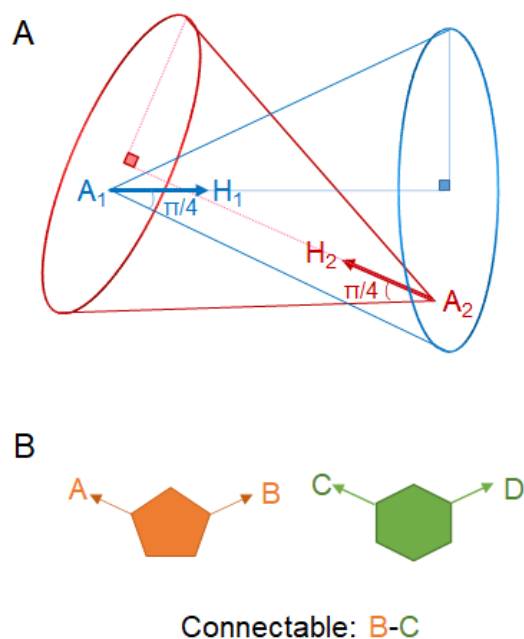
**Figure S7**. Topological requirements to connect fragment atoms by a linker. **A)** $A_1$ and $A_2$ atoms are connectable if they are bound to a hydrogen atom, are located within the projected circular cone (aperture = $\pi/2$) of their counterpart. **B)** Example fragments to be linked with linking atoms A and B for the first fragment (orange) and linking atoms C and D for the second fragment (green). Exit vectors are represented by arrows. Only atoms B and C are connectable, the connections A-C, A-D and B-D are not considered in this study.
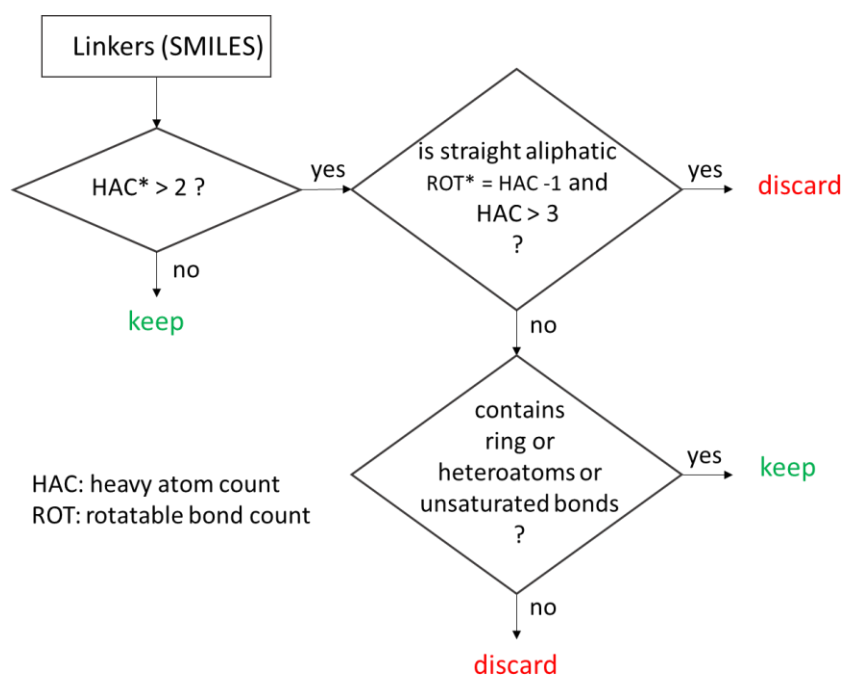


**Figure S8**. Filters for DeLinker-generated linkers. To be kept, generated linkers must be small or contain ring systems or be branched with unsaturated bonds or heteroatoms.
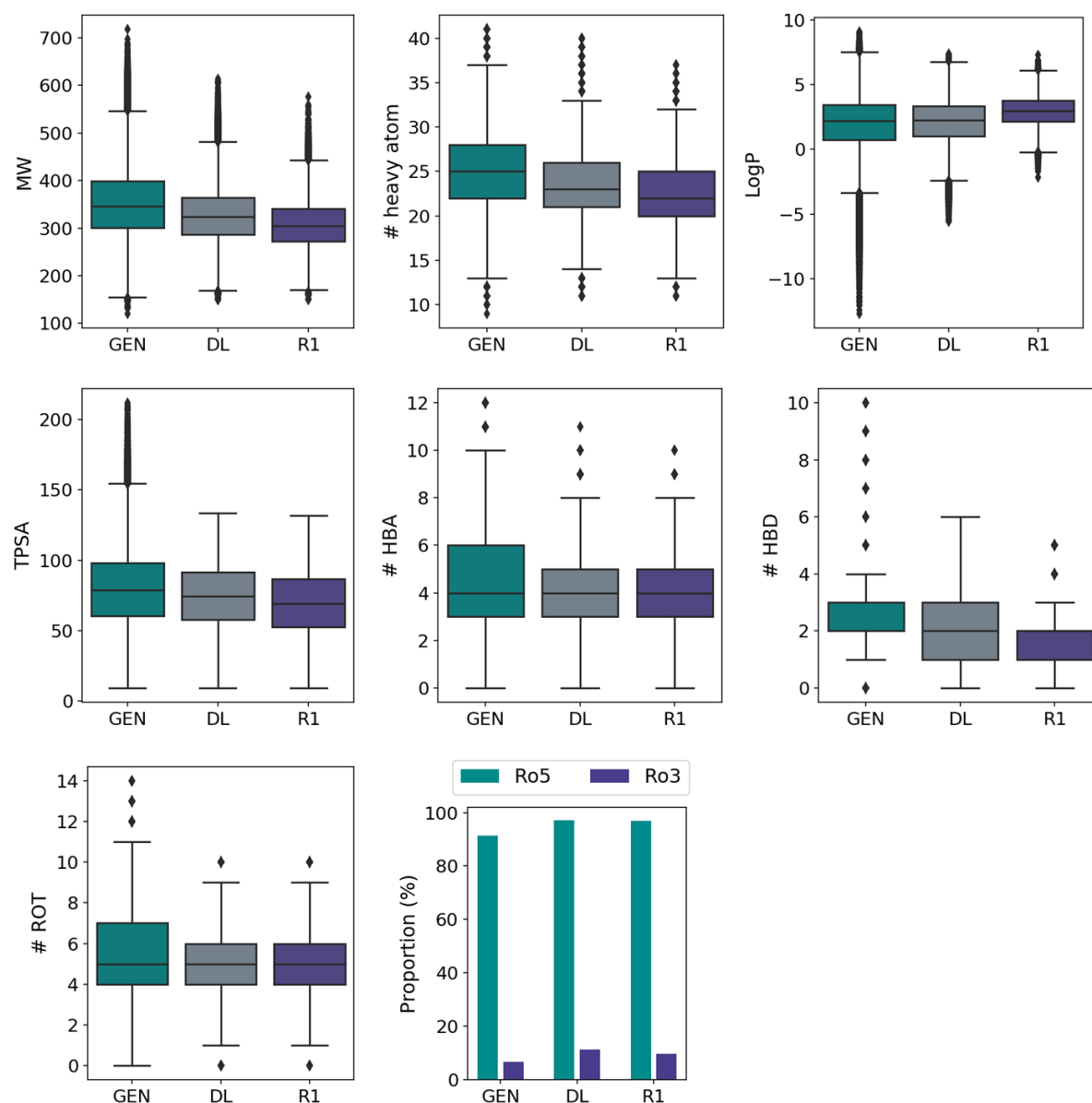
**Figure S9**. Properties of generated molecules after removing 2D duplicates. GEN: full set after removing unsuccessful generation (n=1 119 879), DL: drug-like set (n=566 989), R1: first round library (n=141 125). From left to right, top to bottom, the molecular weight (g.mol$^{-1}$), the count of heavy atoms (non-hydrogen atoms), logP, polar surface area (Å), count of H-bond acceptor, count of H-bond donor, count of rotatable bonds, proportion compliant with Lipinski's rule-of-5 and fragment rule-of-three. All properties were calculated with RDKit. Outliers are computed to be outside the quartiles past 1.5 times the interquartile range.

**Figure S10.** LC-MS analysis of compound **39.**

**Figure S11.** LC-MS analysis of compound **41**.

**Figure S12.** LC-MS analysis of compound **44.**

MaxPeak: 100.00%
Ret_Time: 1.379 min

Compound 47

Mol Wt      402.49
Exact Mass  402.23

```
#  Time   Area%
-----------------
1  1.379  100.00
```

RT  1.390

**Figure S13.** LC-MS analysis of compound **47.**

**Figure S14.** LC-MS analysis of compound **49.**

**Figure S15.** LC-MS analysis of compound **51.**

**Scheme S1.** Synthesis of round-2 library compounds[a]

**Table S1.** List of CDK8 X-ray structures (https://www.rcsb.org, accessed on June 7, 2020)

| PDB[a] | Res.[b] | Ligand[c] | Ligand SMILES | Type[d] |
|---|---|---|---|---|
| 3RGF | 2.20 | BAX | CNC(=O)c1cc(ccn1)Oc2ccc(cc2)NC(=O)Nc3ccc(c(c3)C(F)(F)F)Cl | II |
| 4CRL | 2.40 | C1I | CC12CC=C3C=C4C(C(C(CC45CCC3(C1CCC2c6ccc7ccncc7c6)O5)N(C)C)O)O | I |
| 4F6S | 2.60 | JHK | Cc1ccc(cc1)n2c(cc(n2)C(C)(C)C)NC(=O)N | II |
| 4F6U | 2.10 | HK5 | Cc1ccc(cc1)n2c(cc(n2)C(C)(C)C)NC(=O)NCCCN3CCOCC3 | II |
| 4F6W | 2.39 | 0SS | Cc1ccc(cc1)n2c(cc(n2)C(C)(C)C)NC(=O)NCCN3CCN(CC3)C(=O)Nc4cc(nn4c5ccc(cc5)C)C(C)(C)C | II |
| 4F70 | 3.00 | 0ST | Cc1ccc(cc1)n2c(cc(n2)C(C)(C)C)NC(=O)NCCN3CCOCC3 | II |
| 4F7J | 2.60 | 0SU | Cc1ccc(cc1)n2c(cc(n2)C(C)(C)C)NC(=O)NCCO | II |
| 4F7L | 2.90 | 0SO | Cc1ccc(cc1)n2c(cc(n2)C(C)(C)C)NC(=O)NCCCNC(=O)OC(C)(C)C | II |
| 4F7N | 2.65 | 0SV | Cc1ccc(cc1)n2c(cc(n2)C(C)(C)C)NC(=O)NCCCCCO | II |
| 4F7S | 2.20 | 0SW | c1ccc(cc1)CCNc2c3ccccc3ncn2 | I |
| 4G6L | 2.70 | 0SO | Cc1ccc(cc1)n2c(cc(n2)C(C)(C)C)NC(=O)NCCCNC(=O)OC(C)(C)C | N.A.[e] |
| 5BNJ | 2.64 | 4TV | Cn1cc(cn1)c2ccc(cc2)c3cncc(c3N4CCC5(CCNC5=O)CC4)Cl | I |
| 5CEI |  | 50R | CNC(=O)c1cc2c(cncc2s1)Oc3ccc(cc3)I | I |
| 5FGK | 2.36 | 5XG | c1cc2c(cc1c3cncc(c3N4CCC5(CCNC5=O)CC4)Cl)n[nH]c2N | I |
| 5HBE | 2.38 | 5Y6 | CN1c2ccc(cc2CS1(=O)=O)c3cncc(c3N4CCC5(CC4)CNC(=O)O5)Cl | I |
| 5HBH | 2.50 | 5Y7 | CN1c2ccc(cc2CS1(=O)=O)c3cncc(c3N4CCC5(CCCN5CCOC)CC4)Cl | I |
| 5HBJ | 3.00 | 5Y8 | Cn1c2ccc(cc2cn1)c3cnc(c(c3N4CCC5(CCNC5=O)CC4)Cl)N | I |
| 5HNB | 2.35 | 62M | Cc1cccc(c1)Cc2c3cc(c(cc3[nH]n2)O)C(=O)N4CCC(C4)O | I |
| 5HVY | 2.39 | 66X | CNc1nccc(n1)N2CCC(C2)NC(=O)Nc3ccc(c(c3)C(F)(F)F)CN4CCOCC4 | II |
| 5I5Z | 2.60 | 68U | CNC(=O)c1ccc2cncc(c2n1)c3ccc4c(c3)CS(=O)(=O)N4C | I |
| 5ICP | 2.18 | 69Z | Cc1ncc2n1nc(s2)C(=O)N3CCCC3c4ccc(cc4)Cl | I |
| 5IDN | 2.26 | 6A7 | Cc1c2cc(cnc2[nH]n1)C(=O)N3CCCC3c4ccc(cc4)Cl | I |
| 5IDP | 2.65 | 6A6 | c1cc(ccc1C2CCCCN2C(=O)c3ccc4c(c3)c(n[nH]4)N)F | I |
| 5XQX | 2.30 | 8CC | CNC(=O)c1cc(c[nH]1)c2ccncc2 | I |
| 5XS2 | 2.04 | 8D6 | c1cnccc1c2c[nH]c(c2Cl)C(=O)N | I |
| 6QTG | 2.70 | JH8 | CN(C)C(=O)Cn1cc(cn1)c2ccc(cc2)c3cncc4c3cccc4 | I |
| 6QTJ | 2.48 | JHK | CN(C)C(=O)Cn1cc(cn1)c2ccc(cc2)c3cncc4c3cncc4 | I |
| 6R3S | 2.19 | JRE | CC(c1c(cncc1Cl)c2cc3c(nc2)N(CCC3)C(=O)N)O | I |
| 6T41 | 2.45 | MFE | c1ccc2c(c1)c(ncn2)NCc3ccc(cc3)Cl | I |

[a] PDB identifier.

[b] Higher limit resolution, Å.

[c] Chemical component three-letter code.

[d] Structure classification.

[e] not available (ligand-free structure)

**Table S2.** Filtering rules to select drug-like compounds

```
#/********************************************************************
#Copyright (C) 2004-2020, 2020 by OpenEye Scientific Software, Inc.
#********************************************************************/
#This file defines the rules for filtering multi-structure files based on
#properties and substructure patterns.
MIN_MOLWT      200        "Minimum molecular weight"
MAX_MOLWT      600        "Maximum molecular weight"

MIN_NUM_HVY  15        "Minimum number of heavy atoms"
MAX_NUM_HVY  35        "Maximum number of heavy atoms"

MIN_RING_SYS   0      "Minumum number of ring systems"
MAX_RING_SYS   5      "Maximum number of ring systems"

MIN_RING_SIZE   0       "Minimum atoms in any ring system"
MAX_RING_SIZE   20      "Maximum atoms in any ring system"

MIN_CON_NON_RING   0      "Minimum number of connected non-ring atoms"
MAX_CON_NON_RING   15     "Maximum number of connected non-ring atoms"

MIN_FCNGRP     0     "Minimum number of functional groups"
MAX_FCNGRP     18      "Maximum number of functional groups"

MIN_UNBRANCHED  0      "Minimum number of connected unbranched non-ring atoms"
MAX_UNBRANCHED  6      "Maximum number of connected unbranched non-ring atoms"

MIN_CARBONS     7     "Minimum number of carbons"
MAX_CARBONS     35      "Maximum number of carbons"

MIN_HETEROATOMS   2      "Minimum number of heteroatoms"
MAX_HETEROATOMS   20      "Maximum number of heteroatoms"

MIN_Het_C_Ratio   0.10    "Minimum heteroatom to carbon ratio"
MAX_Het_C_Ratio   1.0     "Maximum heteroatom to carbon ratio"

MIN_HALIDE_FRACTION     0.0     "Minimum Halide Fraction"
MAX_HALIDE_FRACTION     0.5     "Maximum Halide Fraction"

#count ring degrees of freedom = (#BondsInRing) - 4 - (RigidBondsInRing) - (BondsSharedWithOtherRings)
#must be >= 0, from JCAMD 14:251-265,2000.
ADJUST_ROT_FOR_RING    true     "BOOLEAN for whether to estimate degrees of freedom in rings"

MIN_ROT_BONDS   0     "Minimum number of rotatable bonds"
MAX_ROT_BONDS   20      "Maximum number of rotatable bonds"

MIN_RIGID_BONDS   0     "Minimum number of rigid bonds"
MAX_RIGID_BONDS   35      "Maximum number of rigid bonds"

MIN_HBOND_DONORS 0      "Minimum number of hydrogen-bond donors"
MAX_HBOND_DONORS 6      "Maximum number of hydrogen-bond donors"
```

MIN_HBOND_ACCEPTORS  0      "Minimum number of hydrogen-bond acceptors"
MAX_HBOND_ACCEPTORS  8      "Maximum number of hydrogen-bond acceptors"

MIN_LIPINSKI_DONORS  0      "Minimum number of hydrogens on O & N atoms"
MAX_LIPINSKI_DONORS  5      "Maximum number of hydrogens on O & N atoms"

MIN_LIPINSKI_ACCEPTORS  0      "Minimum number of oxygen & nitrogen atoms"
MAX_LIPINSKI_ACCEPTORS  10      "Maximum number of oxygen & nitrogen atoms"

MIN_COUNT_FORMAL_CRG   0      "Minimum number formal charges"
MAX_COUNT_FORMAL_CRG   3      "Maximum number of formal charges"

MIN_SUM_FORMAL_CRG   -2      "Minimum sum of formal charges"
MAX_SUM_FORMAL_CRG   2      "Maximum sum of formal charges"

MIN_CHIRAL_CENTERS   0   "Minimum chiral centers"
MAX_CHIRAL_CENTERS   4   "Maximum chiral centers"

MIN_XLOGP      -5.0      "Minimum XLogP"
MAX_XLOGP      6.0      "Maximum XLogP"

#choices are insoluble<poorly<moderately<soluble<very<highly
MIN_SOLUBILITY      moderately      "Minimum solubility"

PSA_USE_SandP   false   "Count S and P as polar atoms"
MIN_2D_PSA      0.0      "Minimum 2-Dimensional (SMILES) Polar Surface Area"
MAX_2D_PSA      150.0   "Maximum 2-Dimensional (SMILES) Polar Surface Area"

AGGREGATORS   true      "Eliminate known aggregators"
PRED_AGG      true      "Eliminate predicted aggregators"

#secondary filters (based on multiple primary filters)
GSK_VEBER      true      "PSA>140 or >10 rot bonds"
MAX_LIPINSKI   1         "Maximum number of Lipinski violations"
MIN_ABS 0.5            "Minimum probability F>10% in rats"
PHARMACOPIA   true      "LogP > 5.88 or PSA > 131.6"

ALLOWED_ELEMENTS  H,C,N,O,F,S,Cl,Br
ELIMINATE_METALS Sc,Ti,V,Cr,Mn,Fe,Co,Ni,Cu,Zn,Y,Zr,Nb,Mo,Tc,Ru,Rh,Pd,Ag,Cd

#acceptable molecules must have <= instances of each of the patterns below

#specific, undesirable functional groups

| | | |
|---|---|---|
| RULE 0 quinone | RULE 0 | RULE 0 phosphoranes |
| RULE 0 pentafluorophenyl_esters | beta_carbonyl_quat_nitrogen | RULE 0 imidoyl_chlorides |
| | RULE 0 acylhydrazide | RULE 0 nitroso |
| RULE 0 paranitrophenyl_esters | RULE 0 cation_C_Cl_I_P_or_S | RULE 0 N_P_S_Halides |
| RULE 0 HOBT_esters | RULE 0 phosphoryl | RULE 0 carbodiimide |
| RULE 0 triflates | RULE 0 alkyl_phosphate | RULE 0 isonitrile |
| RULE 0 lawesson_s_reagent | RULE 0 phosphinic_acid | RULE 0 triacyloxime |
| RULE 0 phosphoramides | RULE 0 phosphanes | RULE 0 cyanohydrins |
| | | RULE 0 acyl_cyanides |

RULE  0  sulfonylnitrile

RULE  0  phosphonylnitrile

RULE  0  azocyanamides

RULE  0  beta_azo_carbonyl

RULE  0  polyenes

RULE  0  saponin_derivatives

RULE  1
cytochalasin_derivatives

RULE  4
cycloheximide_derivatives

RULE  1
monensin_derivatives

RULE  1
squalestatin_derivatives

#functional groups which often
eliminate compounds from
consideration

RULE  0  acid_halide

RULE  0  aldehyde

RULE  0  alkyl_halide

RULE  0  anhydride

RULE  0  azide

RULE  0  azo

RULE  0  di_peptide

RULE  0  michael_acceptor

RULE  0  beta_halo_carbonyl

RULE  0  nitro

RULE  0  oxygen_cation

RULE  0  peroxide

RULE  0  phosphonic_acid

RULE  0  phosphonic_ester

RULE  0  phosphoric_acid

RULE  0  phosphoric_ester

RULE  0  sulfonic_acid

RULE  0  sulfonic_ester

RULE  0  tricarbo_phosphene

RULE  0  epoxide

RULE  0  sulfonyl_halide

RULE  0  halopyrimidine

RULE  0  perhalo_ketone

RULE  0  aziridine

RULE  1  oxalyl

RULE  0  alphahalo_amine

RULE  0  halo_amine

RULE  0  halo_alkene

RULE  0  acyclic_NCN

RULE  0  acyclic_NS

RULE  0  SCN2

RULE  0  terminal_vinyl

RULE  0  hetero_hetero

RULE  0  hydrazine

RULE  0  N_methoyl

RULE  0  NS_beta_halothyl

RULE  0  propiolactones

RULE  0  iodoso

RULE  0  iodoxy

RULE  0  noxide


#groups of molecules

RULE  0  dye

#functional groups which are
allowed, but may not be
wanted in high quantities
#common functional groups

RULE  6  alcohol

RULE  4  alkene

RULE  4  amide

RULE  4  amino_acid

RULE  2  amine

RULE  4  primary_amine

RULE  4  secondary_amine

RULE  4  tertiary_amine

RULE  2  carboxylic_acid

RULE  6  halide

RULE  0  iodine

RULE  2  ketone

RULE  4  phenol

RULE  1  imine

RULE  1  methyl_ketone

RULE  1  alkylaniline

RULE  4  sulfonamide

RULE  1  sulfonylurea

RULE  0  phosphonamide

RULE  0  alphahalo_ketone

RULE  0  oxaziridine

RULE  1  cyclopropyl

RULE  2  guanidine

RULE  0  sulfonimine

RULE  0  sulfinimine

RULE  1  hydroxamic_acid

RULE  0  sulfinylthio

RULE  0  disulfide

RULE  0  enol_ether

RULE  0  enamine

RULE  0  organometallic

RULE  0  dithioacetal

RULE  1  oxime

RULE  0  isothiocyanate

RULE  0  isocyanate

RULE  3  lactone

RULE  3  lactam

RULE  1  thioester

RULE  1  carbonate

RULE  0  carbamic_acid

RULE  1  thiocarbamate

RULE  0  triazine

RULE  1  malonic


#other functional groups

RULE  2  alkyne

RULE  4  aniline

RULE  4  aryl_halide

RULE  2  carbamate

RULE  3  ester

RULE  5  ether

RULE  1  hydrazone

RULE  0  nonacylhydrazone

RULE  1  hydroxylamine

RULE  2  nitrile

RULE  2  sulfide

RULE  2  sulfone

RULE  2  sulfoxide

RULE  0  thiourea

RULE  1  thioamide

RULE  0  thiol

RULE  2  urea


RULE  0  hemiketal

RULE  0  hemiacetal

RULE  0  ketal

RULE  1  acetal

RULE  0  aminal

RULE  0  hemiaminal


#protecting groups

RULE  0
benzyloxycarbonyl_CBZ

RULE  0
t_butoxycarbonyl_tBOC

RULE  0
fluorenylmethoxycarbonyl_Fm
oc

RULE  1  dioxolane_5MR

RULE  1  dioxane_6MR

RULE  1
tetrahydropyran_THP

RULE  1
methoxyethoxymethyl_MEM

RULE  2  benzyl_ether

RULE  2  t_butyl_ether

RULE  0  trimethylsilyl_TMS
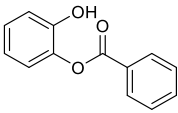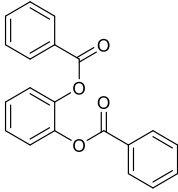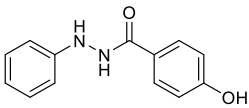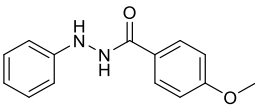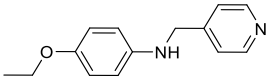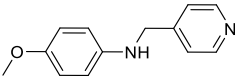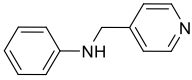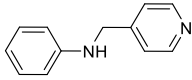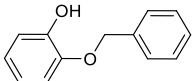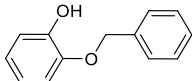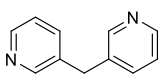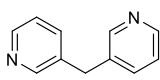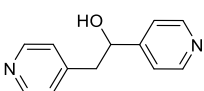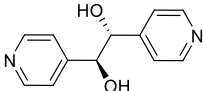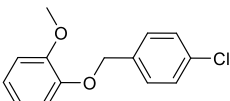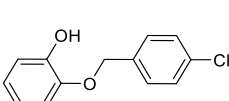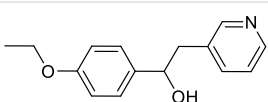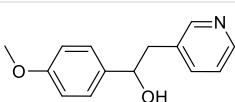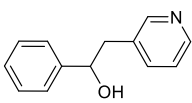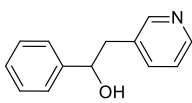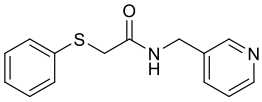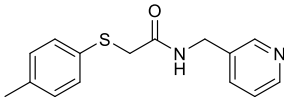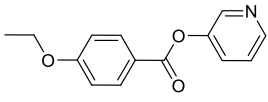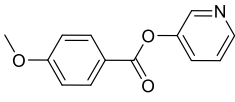
RULE  0                                  RULE  0
t_butyldimethylsilyl_TBDMS               t_butyldiphenylsilyl_TBDPS
RULE  0                                  RULE  1  phthalimides_PHT
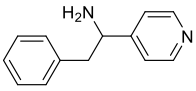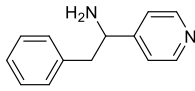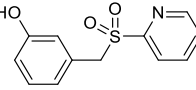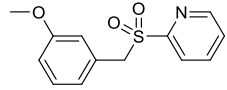triisopropylsilyl_TIPS                   RULE  2  arenesulfonyl
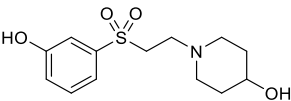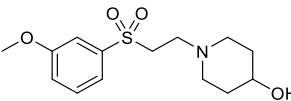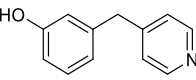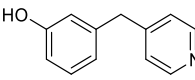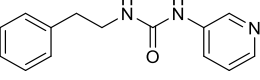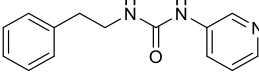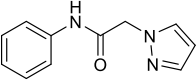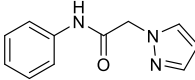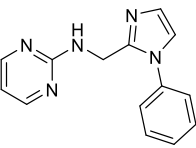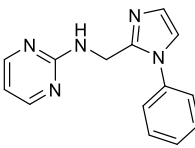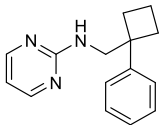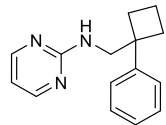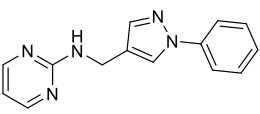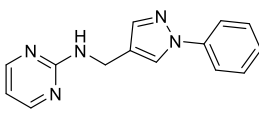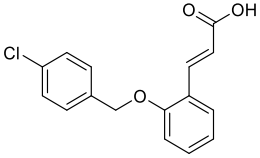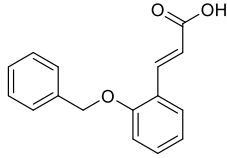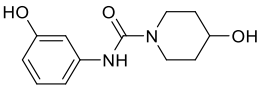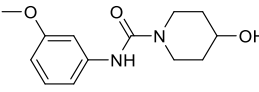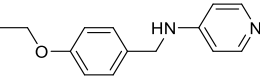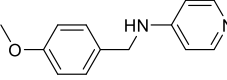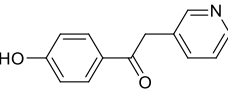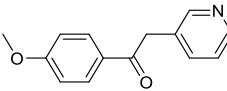
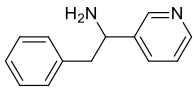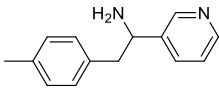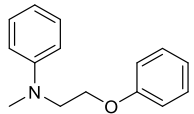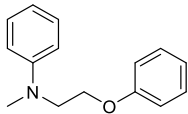**Table S3.** In-house catalog of commercially available drug-like compounds.
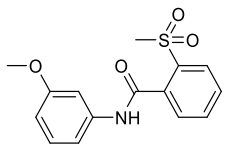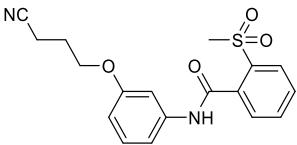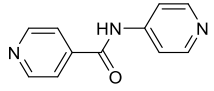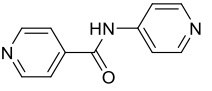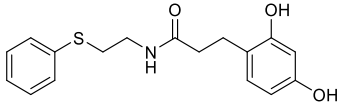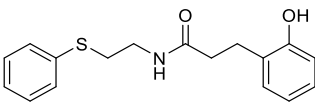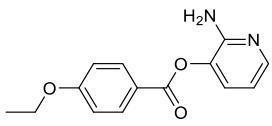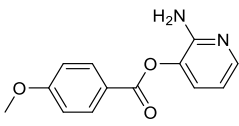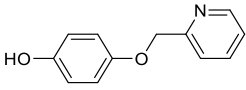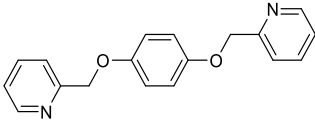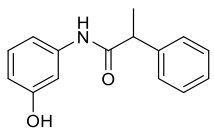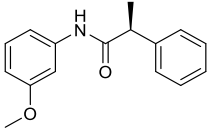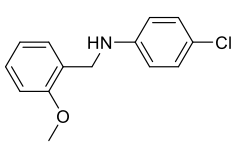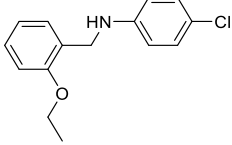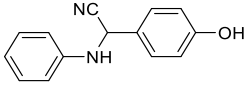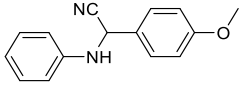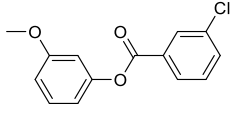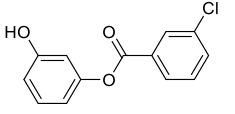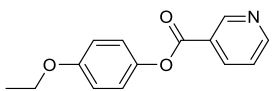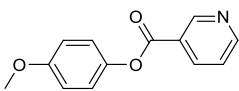
| Supplier | Compounds[a] | Cleaned[b] | Unique[c] | Drug-like[d] |
|---|---|---|---|---|
| AbamaChem | 1 496 973 | 1 389 444 | 1 355 715 | 1 112 527 |
| Alinda | 893 780 | 884 805 | 7 064 | 2 451 |
| AnalytiCon | 46 513 | 44 108 | 39 726 | 23 185 |
| Aronis | 26 848 | 26 757 | 45 | 21 |
| Asinex | 530 881 | 525 110 | 525 102 | 342 471 |
| AsisChem | 2 109 738 | 2 089 223 | 1 720 870 | 556 376 |
| BCH Research | 1 496 546 | 1 453 617 | 1 366 307 | 1 118 354 |
| Bionet | 208 417 | 207 322 | 196 734 | 84 429 |
| Cayman | 14 603 | 14 444 | 9 225 | 2 650 |
| Chembridge | 1 250 334 | 1 242 437 | 1 133 887 | 883 008 |
| ChemDiv | 1 601 806 | 1 586 112 | 1 369 021 | 817 491 |
| CNRS | 75 554 | 71 777 | 63 942 | 30 721 |
| Enamine | 2 701 170 | 2 660 152 | 2 565 862 | 1 820 949 |
| ExiMed | 60 872 | 60 708 | 3 221 | 2 484 |
| InterBioScreen | 555 658 | 545 481 | 348 868 | 174 137 |
| Intermed | 900 691 | 840 422 | 759 154 | 629 819 |
| LifeChemicals | 492 739 | 490 408 | 339 706 | 233 220 |
| Maybridge | 53 352 | 52 777 | 41 920 | 17 408 |
| Otava | 263 238 | 261 029 | 65 402 | 32 567 |
| PBMR_Labs | 1 532 541 | 1 505 095 | 427 795 | 208 920 |
| Pharmeks | 374 473 | 363 888 | 47 752 | 21 691 |
| Specs | 210 228 | 206 727 | 176 871 | 94 270 |
| Synthon_Lab | 32 275 | 32 063 | 6 374 | 2 623 |
| TimTec | 994 852 | 972 738 | 160 298 | 58 846 |
| Vitas-M | 1 413 073 | 1 383 087 | 19 535 | 9 575 |
| **Total** | **19 337 425** | **18 909 731** | **12 750 396** | **8 280 193** |

[a] compounds downloaded on June 8th 2020 from supplier websites.

[b] removal of compounds with erroneous structures and more than 2 undefined chiral centers.

[c] removal of salt-free duplicates according to canonical SMILES strings.

[d] drug-like compounds according to rules in **Table S2**.

**Table S4.** List of 37 commercially available compounds, structurally similar or identical to round-1 library members.

| Original[a] | Commercial[b] | # | ID[c] |
|---|---|---|---|
|  |  | 1 | BAS00100999 |
|  |  | 2 | BAS00127920 |
|  |  | 3 | BAS03714607 |
|  |  | 4 | BAS06103407 |
|  |  | 5 | AS-13577 |
|  |  | 6 | AS-57570 |
|  |  | 7 | AS-65001 |
|  |  | 8 | BS-4424 |
|  |  | 9 | 5238792 |
|  |  | 10 | 5238793 |
|  |  | 11 | 6387127 |
|  |  | 12 | 6736415 |

| | | | |
|---|---|---|---|
|  |  | 13 | Z1024584854 |
|  |  | 14 | Z1513812283 |
|  |  | 15 | Z166719114 |
|  |  | 16 | Z169544550 |
|  |  | 17 | Z1838235103 |
|  |  | 18 | Z229192428 |
|  |  | 19 | Z229315974 |
|  |  | 20 | Z2312274216 |
|  |  | 21 | Z236575354 |
|  |  | 22 | Z359299432 |
|  |  | 23 | Z361879486 |
|  |  | 24 | Z3899831400 |
|  |  | 25 | Z432530210 |

| | | | |
|---|---|---|---|
| | | 26 | Z513796046 |
| | | 27 | Z54748481 |
| | | 28 | Z737854118 |
| | | 29 | Z85517130 |
| | | 30 | Z91149516 |
| | | 31 | 6668547 |
| | | 32 | AE-848/02279007 |
| | | 33 | AF-407/03092027 |
| | | 34 | AH-487/42191575 |
| | | 35 | AJ-292/42152689 |
| | | 36 | AL-398/12677080 |
| | | 37 | AN-652/05929028 |

[a] Original R1 library compound. [b] Closest commercial analogue. [c] Commercial catalogue identifier.

**Table S5.** List of 151 round-2 library members (SMILES strings)

CCOc1ccc(C(=O)Oc2cncc(-c3cc(-c4ccccc4)c(C)s3)c2)cc1
CCOc1ccc(C(=O)Oc2cccnc2Cc2c[nH]cn2)cc1
CCOc1ccc(C(=O)Oc2cnccc2C(C)=Cc2ccc(C)c(C#CCO)c2)cc1
CCOc1ccc(C(=O)Oc2cncc(N3CC(c4ccccc4)C3)c2)cc1
CCOc1ccc(C(=O)Oc2cncc(-n3nnc(-c4ccccc4O)n3)c2)cc1
CCOc1ccc(C(=O)Oc2cncc(C(N)=[NH+]C(=[NH2+])Nc3ccccc3)c2)cc1
CCOc1ccc(C(=O)Oc2cnccc2C=Cc2ccc(C)c(C#CCO)c2)cc1
CCOc1ccc(C(=O)Oc2cnccc2N(CC)C(=O)N2CCc3cc(O)ccc3C2)cc1
CCOc1ccc(C(=O)Oc2cncc(-c3nnc(-c4cn[nH]c4)nn3)c2)cc1
CCOc1ccc(C(=O)Oc2cncc(NC(=[NH2+])Nc3ccccc3)c2)cc1
CCOc1ccc(C(=O)Oc2cncc(C(=O)NC(=O)Nc3ccccc3)c2)cc1
CCOc1ccc(C(=O)Oc2cnccc2C2=CCN=C2c2ccc(C)c(C#CCO)c2)cc1
CCOc1ccc(C(=O)Oc2cnccc2-c2cc(C)n(-c3ccccc3)n2)cc1
CCOc1ccc(C(=O)Oc2cnccc2N(C)C(=[NH2+])Oc2ccccc2)cc1
CCOc1ccc(C(=O)Oc2cnccc2-c2c[nH]cc2-c2ccc(O)c(O)c2)cc1
CCOc1ccc(C(=O)Oc2cnccc2NC(=O)c2ccccc2O)cc1
CCOc1ccc(C(=O)Oc2cncc(NOC(=O)n3ccnc3)c2)cc1
CCOc1ccc(C(=O)Oc2cncc(-c3cc(-c4ccccc4)co3)c2)cc1
CCOc1ccc(C(=O)Oc2cnccc2NC(=O)Nc2ccccc2O)cc1
CCOc1ccc(C(=O)Oc2cnccc2NC(=[NH2+])Oc2ccccc2)cc1
CCOc1ccc(C(=O)Oc2cnccc2-c2cc(C)c(-c3ccccc3)s2)cc1
CCOc1ccc(C(=O)Oc2cnccc2OC(=O)c2ccc3c(=O)[nH]c(C)nc3c2)cc1
CCOc1ccc(C(=O)Oc2cnccc2-c2[nH]cc(-c3ccccc3)c2N)cc1
CCOc1ccc(C(=O)Oc2cncc(-n3nnc(N4CCC(O)CC4)n3)c2)cc1
CCOc1ccc(C(=O)Oc2cncc(-c3ccn(-c4ccccc4)n3)c2)cc1
CCOc1ccc(C(=O)Oc2cnccc2-c2n[nH]nc2N2CCc3cc(O)ccc3C2)cc1
CCOc1ccc(C(=O)Oc2cnccc2-c2ccc(-c3ccccc3)o2)cc1
CCOc1ccc(C(=O)Oc2cncc(C(=O)Nc3cc(O)ccc3F)c2)cc1
CCOc1ccc(C(=O)Oc2cncc(-c3ccc(-c4ccccc4)s3)c2)cc1
CCOc1ccc(C(=O)Oc2cnccc2N(N)C(=O)Oc2ccccc2)cc1
CCOc1ccc(C(=O)Oc2cncc(C3CCN(c4ccccc4)CC3)c2)cc1
CCOc1ccc(C(=O)Oc2cncc(OC(=O)Oc3cc(O)ccc3F)c2)cc1
CCOc1ccc(C(=O)Oc2cnccc2C(=O)Oc2ccc(O)c(O)c2)cc1
CCOc1ccc(C(=O)Oc2cnccc2C(C)=Cc2ccc(C#CCO)c(C)c2)cc1
CCOc1ccc(C(=O)Oc2cncc(C(N)=[NH+]C(=O)Nc3ccccc3)c2)cc1
CCOc1ccc(C(=O)Oc2cnccc2C(=O)N(O)c2cccc(C(C)(C)C)c2)cc1
CCOc1ccc(C(=O)Oc2cncc(-c3[nH]cc(-c4ccccc4)[nH+]3)c2)cc1
CCOc1ccc(C(=O)Oc2cnccc2-c2cc(-c3ccccc3)sn2)cc1
CCOc1ccc(C(=O)Oc2cncc(NC(=O)C(=O)Oc3ccccc3)c2)cc1
CCOc1ccc(C(=O)Oc2cncc(-c3cc(C)n(-c4ccccc4)c3)c2)cc1
CCOc1ccc(C(=O)Oc2cnccc2-c2nn(N)cc2-c2ccc(O)c(O)c2)cc1
CCOc1ccc(C(=O)Oc2cncc(C(=[NH2+])NC(=O)Nc3ccccc3)c2)cc1
CCOc1ccc(C(=O)Oc2cncc(CSC(=[NH2+])N3CCC(O)CC3)c2)cc1
CCOc1ccc(C(=O)Oc2cncc(C(=O)N(C)c3cc(O)ccc3F)c2)cc1
CCOc1ccc(C(=O)Oc2cnccc2-c2ccn(-c3ccccc3)n2)cc1
CCOc1ccc(C(=O)Oc2cncc(NC3=[NH+]CC=C3c3ccccc3)c2)cc1
CCOc1ccc(C(=O)Oc2cnccc2-c2ccccc2-c2c(Cl)cccc2Cl)cc1
CCOc1ccc(C(=O)Oc2cncc(-c3cnn(-c4ccccc4)n3)c2)cc1
CCOc1ccc(C(=O)Oc2cnccc2NC(=[NH2+])c2ccc(O)c(O)c2)cc1
CCOc1ccc(C(=O)Oc2cncc(NC(=[NH2+])Nc3ccccc3O)c2)cc1

CCOc1ccc(C(=O)Oc2cnccc2-c2nc(C)c(-c3ccccc3)s2)cc1
CCOc1ccc(C(=O)Oc2cccnc2-c2ccccc2-c2ccccc2O)cc1
CCOc1ccc(C(=O)Oc2cnccc2-c2n[nH]c(N3CCC(O)CC3)n2)cc1
CCOc1ccc(C(=O)Oc2cncc(ON=C(S)n3ccnc3)c2)cc1
CCOc1ccc(C(=O)Oc2cncc(NC(=O)Oc3ccccc3)c2)cc1
CCOc1ccc(C(=O)Oc2cnccc2NC(=O)Nc2ccc(O)c(O)c2)cc1
CCOc1ccc(C(=O)Oc2cnccc2N2CC(c3ccccc3)C2)cc1
CCOc1ccc(C(=O)Oc2cnccc2OC(=O)Nc2ccccc2O)cc1
CCOc1ccc(C(=O)Oc2cnccc2N=C(O)c2ccc(C)c(C#CCO)c2)cc1
CCOc1ccc(C(=O)Oc2cncc(-c3cc(C)n(-c4ccccc4)n3)c2)cc1
CCOc1ccc(C(=O)Oc2cnccc2C(=O)N2CC(c3cnc[nH]3)C2)cc1
CCOc1ccc(C(=O)Oc2cncc(C3=NC(c4ccccc4)=[NH+]C3)c2)cc1
CCOc1ccc(C(=O)Oc2cncc(-c3sc(-c4ccccc4)cc3N)c2)cc1
CCOc1ccc(C(=O)Oc2cnccc2Cc2ccc3c(=O)[nH]c(C)nc3c2)cc1
CCOc1ccc(C(=O)Oc2cnccc2C2=CC(c3ccccc3O)=CC2)cc1
CCOc1ccc(C(=O)Oc2cnccc2-c2csc(-n3ccnc3)n2)cc1
CCOc1ccc(C(=O)Oc2cnccc2C=C(C)c2ccc(O)c(O)c2)cc1
CCOc1ccc(C(=O)Oc2cnccc2OC(=O)Oc2c(O)ccc3c2OCO3)cc1
CCOc1ccc(C(=O)Oc2cnccc2C=Cc2ccc(C#CCO)c(C)c2)cc1
CCOc1ccc(C(=O)Oc2cnccc2NC(=O)Nc2ccccc2)cc1
CCOc1ccc(C(=O)Oc2cnccc2-c2cnn(-c3ccccc3O)c2C)cc1
CCOc1ccc(C(=O)Oc2cnccc2C(N)=[NH+]C(=O)Nc2ccccc2)cc1
CCOc1ccc(C(=O)Oc2cnccc2C(C#N)=Cc2ccc(O)c(O)c2)cc1
CCOc1ccc(C(=O)Oc2cnccc2NC(=O)NC(=O)c2ccccc2)cc1
CCOc1ccc(C(=O)Oc2cnccc2OC(=O)C(=O)Nc2c(O)ccc3c2OCO3)cc1
CCOc1ccc(C(=O)Oc2cnccc2OC(=O)Oc2ccccc2)cc1
CCOc1ccc(C(=O)Oc2cnccc2SOC(=O)N2CCC(O)CC2)cc1
CCOc1ccc(C(=O)Oc2cncc(OC(=O)Oc3ccccc3)c2)cc1
CCOc1ccc(C(=O)Oc2cncc(-c3ccn(-c4ccccc4)c3)c2)cc1
CCOc1ccc(C(=O)Oc2cnccc2-c2cccc(N3CCC(O)CC3)c2)cc1
CCOc1ccc(C(=O)Oc2cnccc2C(=O)NC(=O)c2ccccc2O)cc1
CCOc1ccc(C(=O)Oc2cnccc2C(=O)Sc2ccccc2)cc1
CCOc1ccc(C(=O)Oc2cnccc2CSC(=[NH2+])N2CCC(O)CC2)cc1
CCOc1ccc(C(=O)Oc2cncc(C=Cc3ccccc3O)c2)cc1
CCOc1ccc(C(=O)Oc2cncc(NC(=O)Nc3ccccc3O)c2)cc1
CCOc1ccc(C(=O)Oc2cnccc2-c2cnn(-c3ccccc3)c2)cc1
CCOc1ccc(C(=O)Oc2cncc(-c3coc(-c4ccccc4)n3)c2)cc1
CCOc1ccc(C(=O)Oc2cnccc2-c2cc(C)n(-c3ccccc3O)n2)cc1
CCOc1ccc(C(=O)Oc2cnccc2C2=CCN=C2c2ccc(C#CCO)c(C)c2)cc1
CCOc1ccc(C(=O)Oc2cncc(C=NC(=[NH2+])Nc3ccccc3)c2)cc1
CCOc1ccc(C(=O)Oc2cnccc2-c2csc(-c3ccccc3)n2)cc1
CCOc1ccc(C(=O)Oc2cncc(NC(=O)c3cc(O)ccc3F)c2)cc1
CCOc1ccc(C(=O)Oc2cnccc2N(C)C(=[NH2+])N2CCc3cc(O)ccc3C2)cc1
CCOc1ccc(C(=O)Oc2cnccc2C=Cc2ccccc2O)cc1
CCOc1ccc(C(=O)Oc2cccnc2Cc2cnc(C)nc2N)cc1
CCOc1ccc(C(=O)Oc2cncc(NC(=O)Nc3ccccc3)c2)cc1
CCOc1ccc(C(=O)Oc2cnccc2C2=NCC=C2c2ccc(O)c(O)c2)cc1
CCOc1ccc(C(=O)Oc2cnccc2C(=O)Nc2ccc3c(=O)[nH]c(C)nc3c2)cc1
CCOc1ccc(C(=O)Oc2cncc(-c3cc(-c4ccccc4)n[nH]3)c2)cc1
CCOc1ccc(C(=O)Oc2cnccc2-c2cccc(-c3c(O)ccc4c3OCO4)c2)cc1
CCOc1ccc(C(=O)Oc2cnccc2NC(=O)Nc2cc(O)ccc2F)cc1
CCOc1ccc(C(=O)Oc2cnccc2-c2csc(-c3c(O)ccc4c3OCO4)n2)cc1
CCOc1ccc(C(=O)Oc2cccnc2Cc2ccccc2O)cc1

CCOc1ccc(C(=O)Oc2cnccc2C(=O)Nc2ccc(C#CCO)c(C)c2)cc1
CCOc1ccc(C(=O)Oc2cnccc2C2=NCC=C2C2CCCC2)cc1
CCOc1ccc(C(=O)Oc2cnccc2CC(=[NH2+])N2CCc3cc(O)ccc3C2)cc1
CCOc1ccc(C(=O)Oc2cncc(C(=O)NC(=O)Nc3cn[nH]c3)c2)cc1
CCOc1ccc(C(=O)Oc2cncc(-c3nc(-n4ccnc4)n[nH]3)c2)cc1
CCOc1ccc(C(=O)Oc2cnccc2C(=[NH2+])NC(=O)Nc2ccccc2)cc1
CCOc1ccc(C(=O)Oc2cnccc2-c2nccc(-c3c(O)ccc4c3OCO4)n2)cc1
C=C(C(=[NH2+])N1CCc2cc(O)ccc2C1)c1ccncc1OC(=O)c1ccc(OCC)cc1
CCOc1ccc(C(=O)Oc2cnccc2-c2nc(-c3ccccc3)sc2C)cc1
CCOc1ccc(C(=O)Oc2cnccc2OC(=O)Nc2cc(O)ccc2F)cc1
CCOc1ccc(C(=O)Oc2cncc(-c3nc(-c4ccccc4)cs3)c2)cc1
CCOc1ccc(C(=O)Oc2cncc(NC(=S)c3ccccc3)c2)cc1
CCOc1ccc(C(=O)Oc2cnccc2-c2ncn(-c3ccccc3)n2)cc1
CCOc1ccc(C(=O)Oc2cncc(NNC(=O)n3ccnc3)c2)cc1
CCOc1ccc(C(=O)Oc2cnccc2C=Cc2ccc3c(=O)[nH]c(C)nc3c2)cc1
CCOc1ccc(C(=O)Oc2cccnc2C(=[NH2+])Nc2ccccc2O)cc1
CCOc1ccc(C(=O)Oc2cncc(C(=[NH2+])Nc3cc(O)ccc3F)c2)cc1
CCOc1ccc(C(=O)Oc2cncc(SC(=O)Oc3ccccc3)c2)cc1
CCOc1ccc(C(=O)Oc2cnccc2-c2n[nH]cc2-c2ccc(O)c(O)c2)cc1
CCOc1ccc(C(=O)Oc2cnccc2NC(=O)Oc2ccccc2)cc1
CCOc1ccc(C(=O)Oc2cncc(-n3nc(-c4ccccc4)[nH]3)c2)cc1
CCOc1ccc(C(=O)Oc2cncc(C(=[NH2+])N=Cc3ccccc3O)c2)cc1
CCOc1ccc(C(=O)Oc2cncc(-c3ccc(N4CCC(O)CC4)cc3)c2)cc1
CCOc1ccc(C(=O)Oc2cnccc2-c2ccc(-c3c(O)ccc4c3OCO4)o2)cc1
CCOc1ccc(C(=O)Oc2cnccc2OC(=O)c2ccc(C#CCO)c(C)c2)cc1
CCOc1ccc(C(=O)Oc2cnccc2C(=O)Nc2ccc(O)c(O)c2)cc1
CCOc1ccc(C(=O)Oc2cnccc2C(=O)N2CC(n3ccnc3)C2)cc1
CCOc1ccc(C(=O)Oc2cnccc2C=Cc2ccc(O)c(O)c2)cc1
CCOc1ccc(C(=O)Oc2cnccc2NC(=O)N2CCc3cc(O)ccc3C2)cc1
CCOc1ccc(C(=O)Oc2cnccc2-c2nnc(N3CCC(O)CC3)o2)cc1
CCOc1ccc(C(=O)Oc2cnccc2C(=O)Nc2ccccc2)cc1
CCOc1ccc(C(=O)Oc2cnccc2N(C)C(=O)Oc2c(C)cccc2C)cc1
CCOc1ccc(C(=O)Oc2cnccc2OC(=O)c2ccc(C)c(C#CCO)c2)cc1
CCOc1ccc(C(=O)Oc2cnccc2C=C(C#N)c2ccc(C#CCO)c(C)c2)cc1
CCOc1ccc(C(=O)Oc2cncc(-c3cc(-c4ccccc4O)nnn3)c2)cc1
CCOc1ccc(C(=O)Oc2cnccc2C(O)=Cc2ccc(C)c(C#CCO)c2)cc1
CCOc1ccc(C(=O)Oc2cncc(-c3cccc(N4CCC(O)CC4)c3)c2)cc1
CCOc1ccc(C(=O)Oc2cnccc2C(=O)NC(=O)c2ccccc2)cc1
CCOc1ccc(C(=O)Oc2cncc(OC(=O)Nc3ccccc3O)c2)cc1
CCOc1ccc(C(=O)Oc2cnccc2C(=[NH2+])Nc2ccc(O)c(O)c2)cc1
CCOc1ccc(C(=O)Oc2cnccc2C(=O)N(C)c2ccc3c(=O)[nH]c(C)nc3c2)cc1
CCOc1ccc(C(=O)Oc2cnccc2-c2cccc(-c3cnc[nH]3)c2)cc1
CCOc1ccc(C(=O)Oc2cncc(N=[SH]c3ccccc3)c2)cc1
CCOc1ccc(C(=O)Oc2cnccc2C=C(C#N)c2ccc(C)c(C#CCO)c2)cc1
CCOc1ccc(C(=O)Oc2cnccc2NC(=O)Oc2c(C)cccc2C)cc1
CCOc1ccc(C(=O)Oc2cnccc2-c2cc(C)n(-c3cc(O)ccc3F)n2)cc1
CCOc1ccc(C(=O)Oc2cnccc2-c2nc[nH]c2N2CCc3cc(O)ccc3C2)cc1
CCOc1ccc(C(=O)Oc2cncc(C(=O)Nc3ccccc3)c2)cc1

## 4.3. Identifying the first inhibitors of a bacterial quinolinate synthase

### 4.3.1. Project description and structural aspects

Quinolate synthase (NadA) is a mainly-prokaryotic enzyme that catalyzes the formation of quinolinic acid (**Figure 4.1**), a precursor for the essential cofactor NAD.[20] Because of its role and its absence in eucaryotes, it appears as an interesting potential target for selective antibacterial design. To date, there is no pharmacological inhibitor of this enzyme.[21] Known ligands are either substrate analogs or derivatives of reaction intermediates. This project was started in collaboration with a Biology team at the Grenoble University (Dr. S. Ollagnier de Choudens, Laboratoire de Chimie et Biologie des Métaux, UMR5249) with the goal of identifying selective pharmacological inhibitors of NadA. Previous studies have characterized the structure of bacterial NadA. The catalytic site adopts an active open or close conformation and contains a [4Fe-4S] cluster necessary for its activity.[22] We thought that the small cavity of NadA ($< 300$ Å$^3$) constitutes a challenge for classical virtual screening approaches and offers a difficult case study to evaluate the POEM workflow.
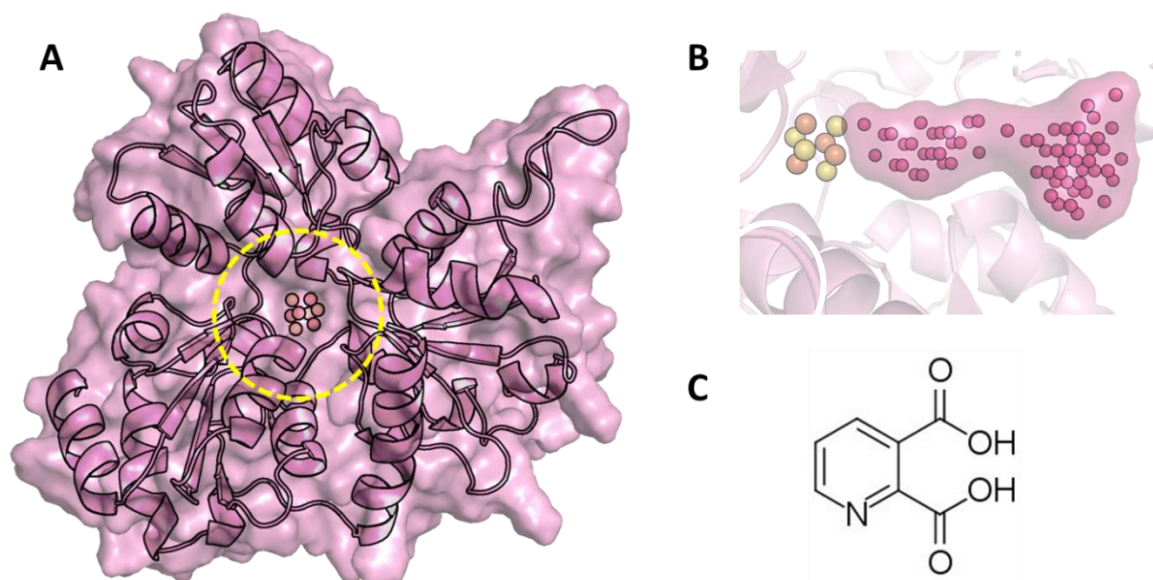


**Figure 4.1.** Structure of *Thermotoga maritima* quinolinate synthase. A) Cartoon and surface representation (PDB ID: 4P3X). The [4Fe-4S] cluster is depicted by spheres in the catalytic pocket (yellow circle). B) VolSite cavity represented by warm pink spheres annotated by one of eight possible pharmacophoric features (hydrophobic, aromatic, h-bond acceptor, h-bond donor, h-bond acceptor or donor, negative or positive ionizable, dummy). The envelope available for inhibitor binding is represented by a solid surface, illustrating the narrowness of the pocket. C) 2D structure of quinolinic acid.

## 4.3.2. Materials and methods

We aimed at designing molecules that can bind to *Helicobacter Pylori* NadA (*hp*NadA) catalytic site. Since no structure is available for that target, a homology model was built with Swiss-model[23] using an open-conformation 3D structure of *Thermotoga maritima* MSB8 (PDB ID: 4P3X) as template. Although sequence alignment with ClustalO yielded 35% identity, the binding site is generally conserved with a few amino acid changes (**Annex 4.1**). The structure was protonated with Protoss v.4.[24] The cavity points were computed with IChem VolSite[25] v.5.2.9 and pruned to avoid areas behind the iron-sulfur cluster (**Figure 4.1**).

The NadA cavity was compared to 31 384 sc-PDB subpockets and the cognate fragments were transferred into the target cavity using ProCare[26] v.0.1.2 with the three alignment descriptors (color c-FH, shape FPFH and hybrid c-FPFH), as described in **section 4.2**.

## 4.3.3. Results and discussion

Following the subpockets comparison to the *hp*NadA pocket, we first observed that the number of subpockets candidates decreases by two third in comparison with CDK8 but this is not surprising knowing of overrepresented protein families in the PDB. However, it raises questions on the chances to generate hit ideas. After removing a majority of cofactor-derived moieties, four to eight hundred fragments (including 2D duplicates) were considered for each descriptor. Consistent with previous observations, that shape-only descriptor yielded the fewest propositions. Given the small volume occupied by the pocket points (~200 $\text{Å}^3$), it was not possible to join fragments occupying adjacent subpockets as they often overlap. Fragments that could be subjected to linking were imidazole derivatives and benzene. We then pursued a different strategy where transferred fragments that occupy the entire cavity were directly considered as putative hits. To this end two selections were visually checked: (i) consensus fragments whose subpockets scored over the previously validated similarity threshold of 0.47 for all the three descriptors (n=186) and (ii) those who in addition to being compliant with rule (i) exhibited a buriedness over 50% into the target cavity cloud (n=39) (**Figure 4.2**).
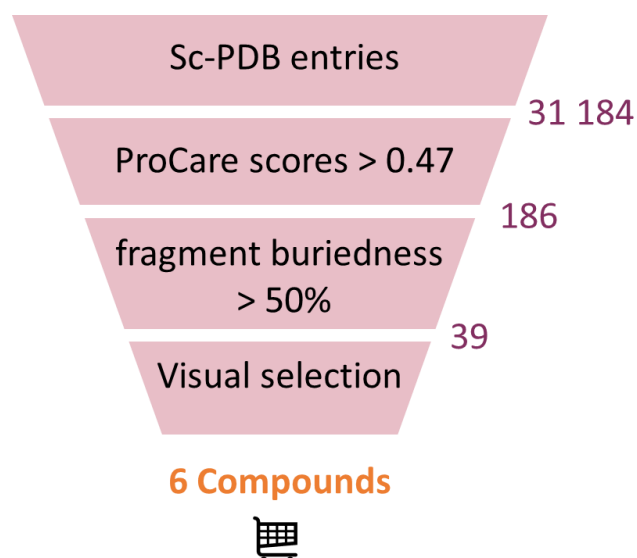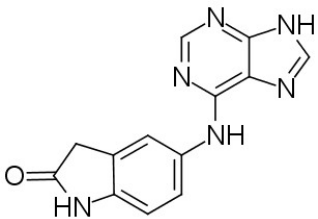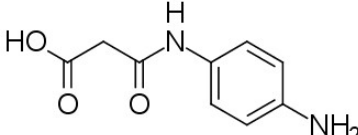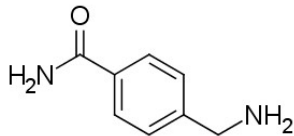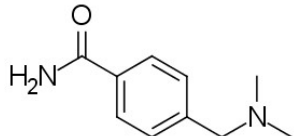
**Figure 4.2.** Selection of NadA virtual hits.

It is important to note that a scoring-based consensus does not necessarily mean that the fragments adopt the same alignment. Therefore, pose-based consensus (RMSD < 3 Å) was used as additional filter. Associated points of the same pharmacophoric features between the target pocket and the fragment subpocket were computed and visually analyzed alongside the fragments. Preference was given to fragments whose pharmacophoric features match that of cavity points. Fragments that orient lone pairs toward the [4Fe-4S] cluster susceptible to coordinate the later (e.g., moieties containing nitrogen, oxygen, sulfur atoms) were discarded, in order to increase NadA specific binding. Finally, after visual check of all ProCare poses, six compounds identical or very similar (Morgan2 Tanimoto > 0.48) to predicted hits were purchased for future *in vitro* evaluation (**Table 4.2**).

**Table 4.2.** Structure of six commercially available virtual hit selected for experimental validation.

| Identifier | Supplier | Structure |
|---|---|---|
| Z769001730 | Enamine |  |
| CAY10009880-1 | Biomol |  |

| | | |
|---|---|---|
| Z5109253219 | Enamine |  |
| Z104484866 | Enamine |  |
| EN300-17770 | Enamine |  |
| EN300-27258 | Enamine |  |

## 4.3.4. Conclusion

In this study, we attempted to design a focused library for identifying pharmacological ligands of *Helicobacter Pylori* NadA catalytic site. As a second case study to validate POEM, the target pocket was narrow and contains an iron-sulfur cluster, adding difficulty to the application. The dimensions of the cavity did not facilitate linking fragments occupying adjacent subpockets but instead suggested to use directly proposed fragments as putative hits. By not applying a generative linking, a lower number of molecule ideas was expected, decreasing the chances to identify actual hits. In computational screening, final selection of virtual hits is often subjective. The current study did not escape this rule. In this scenario, mapping aligned cavity points to the fragment atoms offered a supplemental quality check out of which six hits were prioritized to test their ability to inhibit *in vitro* the catalytic activity of the enzyme (ongoing work).

## 4.4. Hit prediction for the WD40 domain of leucine-rich repeats kinase 2

### 4.4.1. Project description and structural aspects

This project was started as part of the CACHE (Critical Assessment of Computational Hit-finding Experiments) international challenge.[27] It aims at publicly benchmarking computational methods ability to predict hits for relevant targets by confronting predictions to experimental validations. For this first round whose production phase occurred from March 9th to May 9th of 2022, the WD40 repeats (WDR) domain of the human leucine-rich repeats kinase 2 (LRRK2) was chosen. Mutations in the LRRK2 gene are commonly associated with Parkinson's disease whether it was inherited or appeared sporadically.[28] To this current date, therapeutics in preclinical or more advanced phases against LRRK2 are either small molecules inhibiting the kinase domain or biologics.[28–30] The WDR domain, a β-propeller of seven blades (**Figure 4.3**), was shown to mediate LRRK2 protein-protein interactions with microtubules and vesicles trafficking in neurons.[31] Therefore, it appears as a promising drug target.[32] The goal of this challenge is to target the core cavity (**Figure 4.3**) with small molecules. The first experimental results of our predictions are expected no earlier than this fall, hence we will discuss here the problems and solutions encountered by applying POEM to this target.



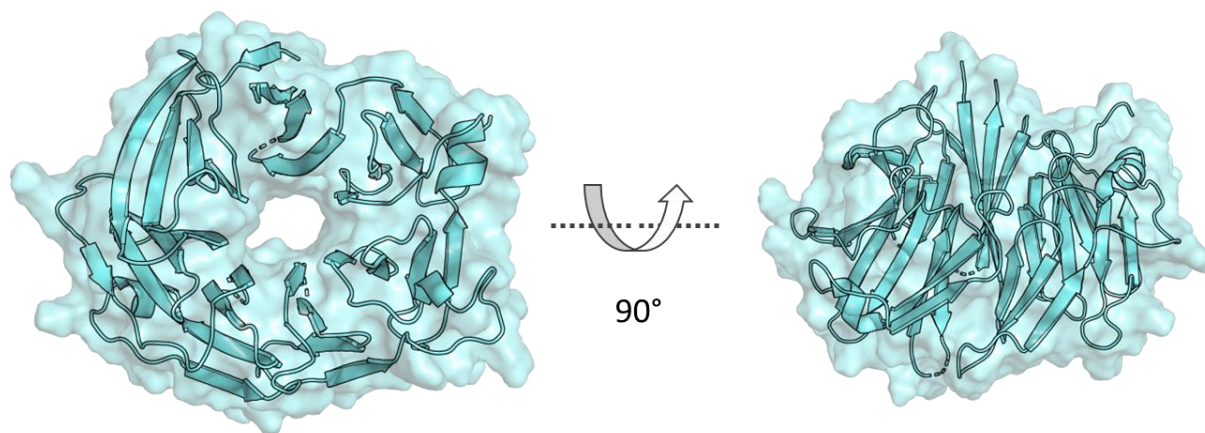**Figure 4.3.** Structure of LRRK2 WD40 domain. Cartoon and surface representation of PDB entry 6DLO; (left) top view showing the core cavity, (right) side view.

### 4.4.2. Materials and Methods

**Structures preparation**

The dimeric structure of LRRK2 WDR (PDB ID: 6DLO, X-ray resolution: 2.7 Å)[33] indicated as starting structure was downloaded from the PDB (https://www.rcsb.org), as well as the monomeric full length

cryo-EM structure 7LHT (3.5 Å). The WDR chains were extracted, aligned to be in the same coordinates frame with Maestro v.2019-3 (Schrödinger, New York, NY 10036, U.S.A.), protonated with Protoss v.4[24] and converted into mol2 format with SYBYL-X v.2.1.1 (Certara USA, Inc., Princeton, NJ 08540, U.S.A.). The pocket point clouds were generated with IChem VolSite v.5.2.9.[25]

**sc-PDB fragments and subpockets v.2022**

Starting from the latest sc-PDB v. 2022 release, IChem fragments and subpockets were prepared from the protein-ligand complexes as described in **section 4.2**. Additionally, fragments originating from 3D RECAP fragmentation[34] (in-house implementation) were added, removing 3D duplicates with IChem fragments—duplicates are the same fragments (by topological fingerprints) occupying the same subpocket of the same PDB entry. Exit dummy atoms resulting from the fragmentation were converted into hydrogen atoms with SYBYL-X v.2.1.1 (Certara Inc., Princeton, U.S.A.). Computed subpockets with IChem VolSite[25] v.5.2.9 were filtered as previously, by discarding those with less than 3 points. The new version (v.2022) of the sc-PDB subpocket-fragment database consists of 107 828 entries, three times more than the previous 2016 version.[35]

**Pocket comparison**

sc-PDB subpockets were compared to the WDR cavity with ProCare[26] v.0.1.2, using the 3 descriptors (color c-FH, shape FPFH and hybrid c-FPFH) and default scoring scheme. The alignment matrices obtained were next applied to the corresponding fragments to pose them in the target cavity. Aligned target/query cavity points were extracted with ProCare tools.

**Interactions detection**

Protein-fragment interactions (h-bond, ionic, aromatic, hydrophobic) were detected with IChem[36] v.5.2.9 IFP module with default angle and distance parameters. Interaction triplets were detected with INTS module.

**Buriedness**

Fragments buriedness in the WDR pocket were computed with the IChem 5.2.9 Utils module.

**sc-PDB entries annotation**

Protein annotations of sc-PDB[35] entries (name, Uniprot[37] accession, function keywords) were extracted via the RCSB PDB application programming interface (API) with inhouse scripts. The chain identifier associated to the ligand in the PDB (author chain) was corrected from the mmCIF file of the entry, to finally assign the correct assembly ID.

**Target enrichment**

For each target represented by their Uniprot accession (polyprotein are disregarded) the enrichment rate was calculated as the proportion of their PDB entries for which a subpocket scored higher than the selection threshold ($N_{top}$) relative to the initial number in the sc-PDB database ($N_{total}$):

$$r\ (\%) = \frac{N_{top}}{N_{total}} \times 100 \qquad \text{eq. 4.1}$$

**Search in commercial libraries**

The Enamine REAL diverse set of 38 million molecules was downloaded (https://enamine.net/compound-collections/real-compounds/real-compound-libraries, accessed on April 20th 2022) and filtered for druglikness (**Section 4.2, Supporting Table S2**) with OpenEye Filter v.3.0.1.2 (OpenEye Scientific Software, Santa Fe, NM 87508, U.S.A.) yielding 24 million druglike molecules. Similarly, the in stock list from MCULE database (https://mcule.com/database/) was prepared as backup, yielding 2.3 million druglike molecules. These compounds were compared to the designed molecules using RDKit v.2019.03.4.0 (http://www.rdkit.org) Morgan2 fingerprint. Pairs were considered similar when the Tanimoto metric was higher than 0.7.

**Docking**

Hits candidates were ionized at physiological pH with OpenEye Filter v.3.0.1.2 and finally converted in 3D structures (mol2 file) with Corina v.3.40 (Molecular Networks GmbH, 90411 Nürnberg, Germany). Possible stereoisomers and ring conformers were generated simultaneously. The prepared molecules were docked into the WD40 cavity with PLANTS[38] v.1.2. The search space was set at 20 Å from the binding site center with a search speed of 1 (highest accuracy). Ten poses ranked by the ChemPLP scoring function were generated per ligand. A root-mean square deviations (RMSD) of 2 Å on ligand heavy atoms was used to cluster solutions. The flipped/rotated side chains were considered in the protein structure for each corresponding PLANTS pose.

**Shape-based alignment of molecules**

Commercial compounds found similar to potential hits were aligned with OpenEye ROCS v.3.0.1.2 (OpenEye Scientific Software, Santa Fe, NM 87508, U.S.A.) to the pair of seed fragments, optimizing the shape and chemical features overlap by conformational search. The alignments were ranked with the Tanimoto combo score.

### 4.4.3. Results and discussions

**Choice of the WDR structure**

The starting WDR structure 6DLO is a dimer with missing loops at both the top and down sides of the mouth surface (**Figure 4.4**). For this study, the chain A was selected over chain B as it was missing less residues, after careful alignment and inspection. Contrarily, the low-resolution cryo-EM structure (7LHT) was not missing residues. Consequently, VolSite cavity points extended towards the loop region modifying the shape of the cloud (**Figure 4.4**). We expected this to affect alignment of the subpockets. Whether these extra cavity points are important is unknown, in the absence of any structure with bound ligands. Unresolved loops due to high flexibility does not exclude that those residues might play a crucial role for ligand binding. One particularity of these pockets is their high proportion of h-bond donor features (30%). The two other most abundant features in similar proportions were hydrophobic and undetermined dummy features. Although the pockets of these two structures were found similar (highest ProCare Score: 0.70), the two pockets were kept for parallel library design.
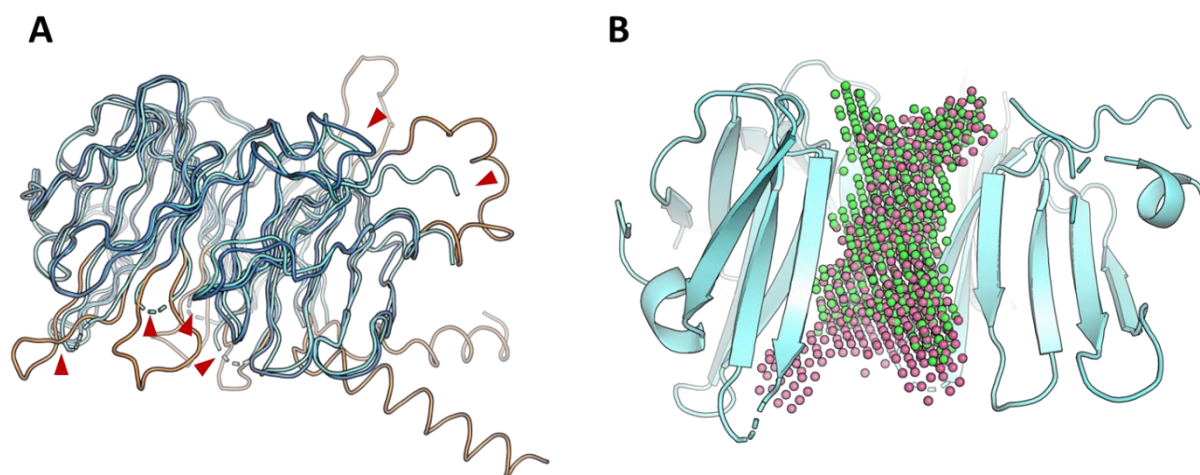


**Figure 4.4.** Overlay of WDR-LRRK2 protein structures and cavities. A) The 6DLO X-ray structure (light blue) is missing loops (orange, red arrows) present in the cryo-EM structure 7LHT (dark blue). B) The VolSite cavity points of 7LHT structure (warm pink) extended to the outer bottom region of the core, compared to the 6DLO cavity (green).

**Fragments selection**

The first step for elaborating molecules is the selection of seed fragments. Distributions of ProCare similarity scores showed similar trends for the color and hybrid histograms (c-FH and c-FPFH), compared to the shape descriptor. This observation is in accordance with all previous studies. Given the high proportion of polar features in the pockets, alignments by the color descriptor were chosen. Consistently, only subpockets scoring over the similarity threshold of 0.47 were considered, yielding

two lists (6DLO and 7LHT) of ~2 700 (2.5%) entries. No ligand is yet known for this target, therefore co-factors were kept at this stage, since they can provide useful information. Four different analyses were carried. Firstly, the 294 fragments common to the two lists were inspected. When considering the coherence of the alignments, this count decreased to 64. It appeared that top or bottom sides of the cavity were differently prioritized for alignment to the two templates. These differences are probably due to the extension of the cavity points toward flexible loops in one of the pockets but might also be related to the random sampling procedure in the ProCare method suggesting other alternatives for alignment. Secondly, we checked for the fragments buriedness. Even if they were not optimally positioned, a clear distinction between buried and accessible fragments is to be expected. However, the cylinder-shaped cavity yielded poor buriedness, that could not be interpreted. The third source of information was enrichment in certain targets. High rates were obtained by kinase-bound nucleotide-like fragments. The fourth and final analysis to prioritize a few fragments for linking was to assess their likelihood to interact with surrounding protein residues. Given the approximation in the fragments positioning regarding interaction detection with the target, we did not initially consider interactions with target residues according to strict angle/distance rules. Fragments atoms were converted into equivalent pharmacophoric features (more description in **Chapter 5**) as the pocket. Keeping fragments having at least half of their polar features identical to and within 3 Å of an aligned cavity point in the target (threshold set by retrospective analysis of the fragments in their original pockets) led to 389 non-cofactor fragments for 6DLO, and 1016 for 7LHT. According to the previous conclusions, a few co-factor-derived fragments were added by visual selection to compile two final lists of 412 and 1048 candidates to be linked for 6DLO and 7LHT respectively.

**Library enumeration and virtual hit selection**

Linking fragments requires to cluster them by target areas and to identify connectable areas. To this end, we defined a procedure to automatically identify areas where selected fragments were frequently aligned (the consensus from the two templates were used). Target cavity points that were aligned by more than 25-30% of subpocket hits defined two main areas. The first area is located around residue Y2249 (bottom side) and curiously overlap with a hotspot detected by the fragment-hotspot tool[39] of Cambridge Crystallographic Data Centre (https://fragment-hotspot-maps.ccdc.cam.ac.uk). The second area lay at the opposite side, around M2301 (top side), a conserved motif across species (**Figure 4.5**).[33]
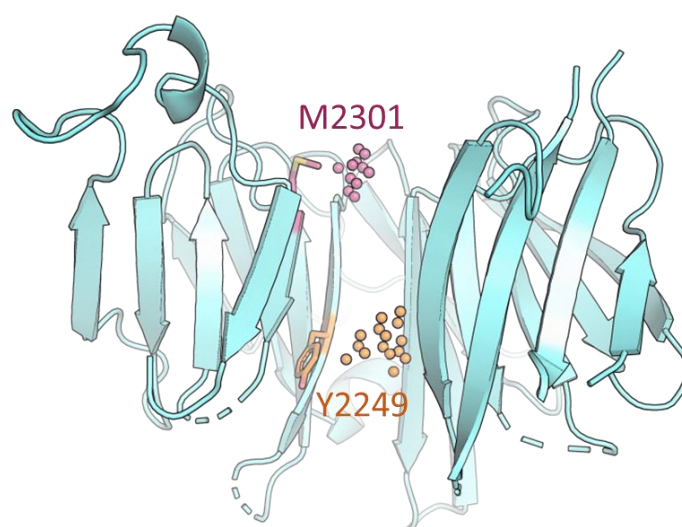
**Figure 4.5.** Frequently aligned areas of WDR-LRRK2 to sc-PDB subpockets. Two areas were defined for fragment annotation prior to linking: around M3201 on top side and Y2249 on bottom side. PDB entry: 6DLO.

Fragments were assigned areas based on their distance to the consistent points, and those more or less equally distant were assigned 'middle' area (**Table 4.3**). Given the high number of fragments in the 7LHT *bottom* area, we could apply additional filtering by keeping fragments that exhibit at least one polar interaction (IChem IFP module[36]) with the target.

**Table 4.3.** Assignment of pocket areas to aligned fragments.

| PDB reference | top | middle | bottom |
|---|---|---|---|
| 6DLO | 134 | 34 | 244 |
| 7LHT | 51 | 63 | 934 (195)[a] |

[a] 934 fragments were assigned to bottom area, a sampling based on detected polar interaction with WDR reduced the list to 195 fragments.

It is not realistic that a high-affinity ligand would specifically bind right in the middle of the cylindric pocket. However, to evaluate the automatic design, we did not bias the selection of the fragments. In the current case, there is not a clear definition of the binding site. Available β-propeller structures showed that molecular partners bind at the very outer surface[32] (www.rcsb.org), but it is unclear whether the top or bottom side should be prioritized. A few studies suggested that one side (top) might be more prone to protein-protein interactions.[33,40]

While investigating the two sides, four connectivity schemes were defined to generate molecules of acceptable sizes: *top-top*, *top-middle, bottom-bottom*, *bottom-middle* (**Figure 4.6**). Identifying connectable atoms among seed fragment pairs is not a simple combinatory problem because it also aims at avoiding geometrically irrelevant connections while calibrating the size of the final library.

In addition to rules implemented in the CDK8 study, pairs of connectable fragments must display a cumulative size of 13 to 25 heavy atoms. This prevents from connecting two very small fragments. For future applications, a filter can be applied to the fragments database prior to alignment. Almost colinear and overlapping fragments planes are not desirable since that would require distorted linkers. Subsequently, fragment pairs displaying a least 3 pairwise distances between 0 and 2 Å were discarded. Theses implementations clearly improved the list of fragments to be linked.
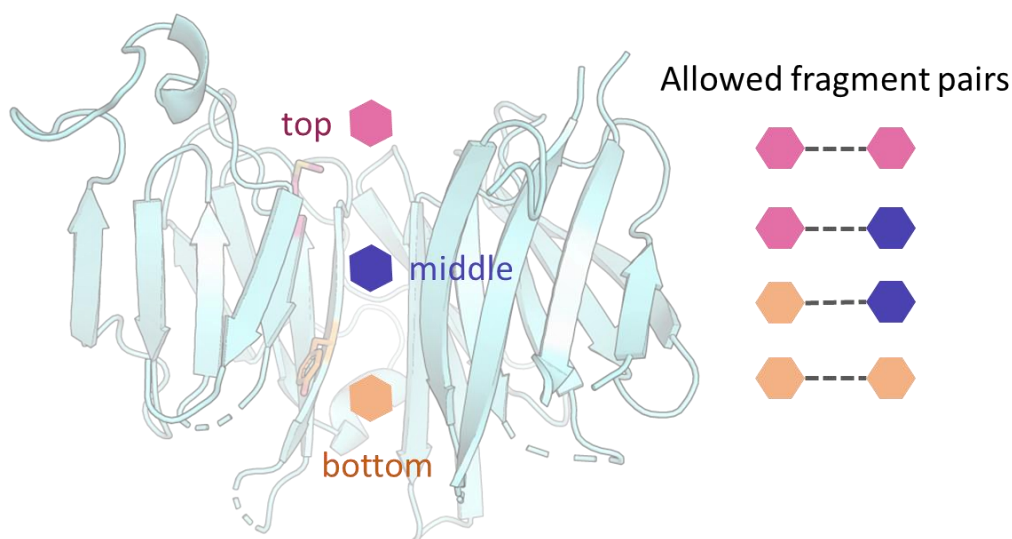


**Figure 4.6.** POEM connectable areas in the WDR-LRRK2. PDB entry: 6DLO.

The DeLinker generative program[41] was applied to generate 900 k (7LHT) and 1.9 million (6DLO) complete molecules, out of which 400 k and 123 k were druglike with decent linkers and synthetic accessibility. The two lists shared 2316 molecules. To achieve a list of ca. 150 commercial compounds (as requested by the CACHE challenge organizers), POEM 6DLO virtual hits were searched in the druglike diverse set of Enamine REAL database (Morgan2 Tanimoto > 0.7) to retrieve similar compounds and a backup list was compiled from MCULE in stock database using the 7LHT virtual hits as queries (Morgan2 Tanimoto > 0.8). The most similar compounds were then subjected to a series of filters (removing chiral compounds and molecules with more than six rotatable bonds) and last clustered according to their Bemis-Murcko scaffolds (Agnes method, Pipeline Pilot, Dassault Systèmes, France). Finally, 100 compounds were prioritized for the synthesis costs, as estimated by Enamine (**Figure 4.7**).
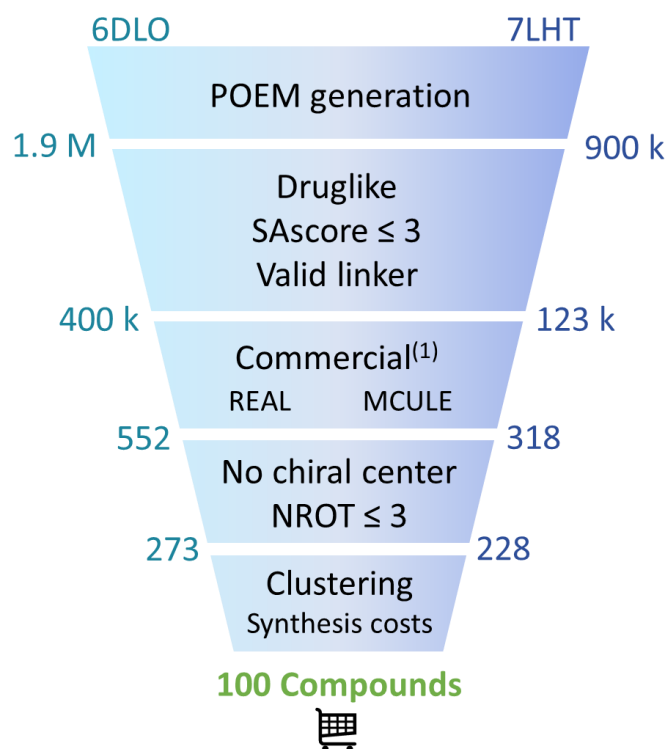
**Figure 4.7.** Library design and virtual hits selection for WD40-LRRK2. (1) Similar compounds to generated molecules were retrieved with Morgan2 Tanimoto > 0.7 (6DLO vs. Enamine REAL diverse set) and Morgan2 Tanimoto > 0.8 (7LHT vs. MCULE *in stock* set).

At this stage, docking of virtual hits showed no privileged subpockets (*top*/*bottom*/*middle*) and could not be used for interpretation. Likewise, ROCS similarity searches could not be exploited as well since shape and chemical property alignment of commercial compounds onto fragments showed that generated conformers do not always overlap with the two original fragments when the linker induced incompatible conformation.

## 4.4.4. Conclusion

The CACHE challenge offers a fully blind case study to the practicability and reliability of POEM to generate pocket-focused molecule ideas. Starting from a hardly druggable target with very little information, we adapted the workflow to assemble molecules thought to have chances to bind to the target. The fragments selection and linking protocol included new steps to rule out unreasonable fragment pair combination. Under different project constraints (e.g. timing), other studies such as molecular dynamic simulations despite its limitations could have helped to model the shape of the pocket, providing different starting structures for screening.

# 4.5. Critical evaluation of the three POEM validation studies

In these projects, we aimed at validating POEM, a new workflow to generate a library of molecules tailored to a target pocket, by linking pre-positioned 3D fragments from protein-ligand X-ray structures according to their subpocket resemblance with the target pocket.

## 4.5.1. Novelty

Although the POEM idea falls within the concept of target-based *de novo* drug design since the 1990s,[42] it differs from existing methods by a combination of several aspects: (1) no reference ligand is required for the target while some methods (e.g., BREED[43], KinFragLib[44]) rely on reference protein-ligand complexes for molecular hybridization, (2) pairs of fragments are directly used for elaboration, as opposed to (grid-based) sampling of atoms as in BUILDER,[45] CONCEPTS[46] or Ramensky *et al.*,[47] (3) the fragments templates are derived from existing protein-ligand complexes in their X-ray conformation, instead of using a library of template fragments as in LUDI,[48] LigBuilder,[49] or FastGrow,[50] (4) fragments are positioned according to the similarity of their subpocket to the target cavity and are not scored by any energy criteria (e.g., GroupBuild[51], LUDI[48]), (5) the fragments linking is based on a 3D-constrained variational autoencoder to generate potential linker graphs, instead of strict topological generators guided by explicit bond and torsion angle ranges.[52] The closest implementations to POEM are the work by Moriaud *et al.*[53] and Durrant *et al.*,[54] suggesting building block fragments to link on the basis of their environment similarity with the target site, albeit with a different site representation and comparison algorithm.[55] Moreover, the latter methods do not enumerate fully connected molecules from the position of seed fragments.

## 4.5.2. Fragment database: ligand deconstruction

The ligand fragmentation protocol influences the content of the designed library in different manners: the subpockets definition, alignment, and linker generation. To study these effects, a different 3D fragmentation scheme based on RECAP retrosynthetic rules[34] was implemented in our lab as alternative to IChem to reproduce the CDK8 case study. The IChem fragmentation[56] method used here breaks single bonds more or less around rings and discards acyclic structures. Substituents or linker groups are kept attached to the core ring. To ensure that the fragments reflect the pharmacophoric features of the subpocket, only those interacting with lining residues were used. However, we draw special attention to the cases where the presence of some chemical groups on the fragments, not particularly involved in interactions with the original target, may be rather making bad contacts once aligned to the target cavity.

Clashes were also observed due to the subpocket only partially overlapping with the fragment, typically a subpocket missing points in areas of low buriedness according to VolSite implementation. We solved these issues by either computing clashes with the target upon alignment, or by estimating the embedding of the fragments in the subpockets/pockets. As a solution to avoid useless fragments or substituents and reduce the chances of bad contacts, the fragment-subpocket database can be improved by scoring the matching between pharmacophoric features of the fragment atoms (more details in **Chapter 5**) and the subpocket points. The high occurrence of certain fragments such as adenine (17% of IChem fragments) prompts to analyze fragment-subpocket redundancy in the database. Finally, analysis of the fragment space coverage with respect to commercial fragment databases or deconstructed compounds in public repositories would provide useful information regarding prospective applicability.

## 4.5.3. Fragments positioning

We purposely linked the direct ProCare-based alignment of the fragments to demonstrate it already contained rich information across different target families for molecule design. However, the fragments position can be optimized in the pocket prior to the linking procedure. For instance, we achieved this goal using OpenEye Szybki energy refinement (OpenEye Scientific Software, Santa Fe, NM 87508, U.S.A.). Indeed, on the CDK8 case, 71% of selected fragments have deviated by more than 2 Å upon optimization, effect that can affect the linker generation. Another idea would be to redock the selected fragments into the target pocket. In either case, only solutions close to the original subpocket-based fragment positioning should be considered to not entirely lose the pocket comparison logic. We recall that such optimizations are subjected to a force-field implementation and add complexity to the workflow. While the binding of close conformations (RMSD-based) of a fragment to structurally distant pockets still remains a rare event,[57] we interestingly observed cases where the same fragments originally bound to different proteins were closely aligned (fragments RMSD < 3 Å) into the CDK8 pocket. On the other hand, the same fragments from different protein subpockets were aligned at different locations as well. We cannot computationally assess the accuracy of these predictions, but it can simply be explained by the dissimilarity between these original subpockets. We underline that this is consistent with the well-known promiscuity of fragments in experimental screenings.[58,59] The issue observed was when the same fragments from the same subpocket in the same protein align to different target pocket areas. This highlights the noises in the subpocket definition and sampling effects in the comparison algorithm discussed in **Chapter 2**.

### 4.5.4. Fragments linking

The deep generative linking algorithm (DeLinker[41]) employed in the current version offers the advantage of being flexible. Indeed, the positions of the fragment rely on the performance of the pocket alignment. Even assuming that the pocket alignment is perfect (which is clearly not the case), it should not be expected that the fragments would systematically adopt the exact pose nor the same conformation upon binding in its new pocket. Therefore, it is not sound to use torsion-based linking approaches. Previous attempts with stricter methods such as ReCore[60] (BioSolveIT GmbH, Sankt Augustin, Germany) on carefully chosen examples led to unsuccessful linking. Although DeLinker attempts to propose linkers likely to match inputs 3D constraints, final molecules are enumerated as SMILES strings, thereby losing the initial target coordinates frame. To assess that the linking procedure is still compatible with the initial fragment poses proposed by ProCare, each enumerated compound must be generated in 3D (using the RDKit routine of DeLinker or other conformer generators) and docked or aligned to the cavity of interest. This workaround being impractical at a high-throughput level, development a true 3D linking method from ProCare fragment poses would constitute a true added value to the current POEM workflow.

While pairing, all fragments were treated equally, without considering their relative buriedness and solvent accessibility in the target. Given the enthalpic nature of fragments binding,[61] connecting two loosely buried fragments decreases the chances to observe the same binding mode in the obtained molecules. The consistency between the poses of the fragments and that of the fully enumerated molecule is a bottleneck for fragment based approaches.[3,62] The designed linker can as well induce changes in the binding mode but these are hard to predict prior to complete enumeration of the molecule. This effect was hypothesized by docking in the second round of the CDK8 study while docking first round experimental hits showed consistent poses with predicted binding subpockets.

### 4.5.5. Synthetic accessibility

The synthetic accessibility is the most crucial characteristic of the library members as nice-looking molecules predicted to interact with the target are useless unless they can be synthetized for experimental assays. Although estimating synthesis hardness with the knowledge-based Ertl and Schuffenhauer method,[63] we were herein limited by available commercial compounds highly similar to designed molecules, at least to evaluate the workflow as quickly as possible. In future production use, it is highly desirable to increase the proportion of really synthesizable molecules via retrosynthetic rules even if challenges regarding rewards and chemical conditions optimization still remain. To achieve this goal, designed molecules can be fragmented and analyzed according to predefined reactions, availability and

cost of building blocks. Another benefit of a such filtering is the reduction of the library size and easier prioritization of virtual hits.

## 4.5.6. Chemical diversity

One of the important characteristics of a library are the diversity of the molecules. There are different definitions of diversity but for the sake of simplicity, we will only refer to the Bemis-Murcko scaffolds.[64] Here, the diversity of the designed library is a consequence of both the diversity of the original fragments pool, fragments connectivity and the diversity of the generated linkers. The problem is almost combinatorial. Theoretically, starting from a pool of F different (two-dimensional based identity) fragments, an average C connectable atoms per fragment and L possible linkers, the maximum size N of the library is :

$$N = F^2 \times C^2 \times L \qquad \text{eq. 4.2}$$

In the CDK8 study, around 200 different fragments representing a hundred scaffolds were used. Interestingly, few fragments are shared between the four pocket areas, reducing the combinations. Not surprisingly, the most promiscuous fragments were benzene and substituted phenols as a consequence of practices in small molecule ligand design and the fragmentation approach.

## 4.5.7. Computing time

We report here the most time-consuming steps in the design process (**Table 4.4**). Filtering and data processing were instantaneous to a few minutes-lasting.

**Table 4.4.** Running time of different POEM steps.

| Step | Resources | Average time |
|---|---|---|
| Pocket-fragment alignment with ProCare | Intel® Xeon® Silver 4114 CPU @ 2.20GHz 1 thread, 4 Go Computer cluster | 1 s – per pair of fragments |
| Identification of connectable atoms | Intel® Core™ i5-4590 4 threads, 16 Go | 0.19 s – per pair of fragments |

Local

| | | |
|---|---|---|
| Linking with DeLinker | Intel® Xeon® Silver 4114 CPU @ 2.20GHz + NVIDIA Tesla K80 GPU, 24 Go Computer cluster | 20 s – per pair of atoms |

### 4.5.8. Towards a fully automated method?

This POEM approach is not fully automatized. The definition of 'linkable fragments' is left to the appreciation of the user with respect to the pairs of subpockets to connect. The relative orientation of fragments exit vectors is also a tunable parameter although an aperture of $\pi/2$ have shown to be consistent. The present workflow offers enough flexibility to adapt to the target specifications. Throughout these three studies, the fragments selection was the most difficult step. We hope that these studies, supported by experimental validation, as well as considerations for improvement discussed here will provide a strong basis for decision making.

## 4.6. References

1. Lyu, J.; Wang, S.; Balius, T. E.; Singh, I.; Levit, A.; Moroz, Y. S.; O'Meara, M. J.; Che, T.; Algaa, E.; Tolmachova, K.; Tolmachev, A. A.; Shoichet, B. K.; Roth, B. L.; Irwin, J. J. Ultra-Large Library Docking for Discovering New Chemotypes. *Nature* **2019**, *566*, 224–229.

2. John Harris, C.; D. Hill, R.; W. Sheppard, D.; J. Slater, M.; F.W. Stouten, P. The Design and Application of Target-Focused Compound Libraries. *Comb. Chem. High Throughput Screen.* **2011**, *14*, 521–531.

3. Kozakova, D.; Hall, D. R.; Jehle, S.; Luo, L.; Ochiana, S. O.; Jones, E. V.; Pollastri, M.; Allen, K. N.; Whitty, A.; Vajda, S. Ligand Deconstruction: Why Some Fragment Binding Positions Are Conserved and Others Are Not. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112*, E2585–E2594.

4. Firestein, R.; Bass, A. J.; Kim, S. Y.; Dunn, I. F.; Silver, S. J.; Guney, I.; Freed, E.; Ligon, A. H.; Vena, N.; Ogino, S.; Chheda, M. G.; Tamayo, P.; Finn, S.; Shrestha, Y.; Boehm, J. S.; Jain, S.; Bojarski, E.; Mermel, C.; Barretina, J.; Chan, J. A.; Baselga, J.; Tabernero, J.; Root, D. E.; Fuchs, C. S.; Loda, M.; Shivdasani, R. A.; Meyerson, M.; Hahn, W. C. CDK8 Is a Colorectal Cancer Oncogene That Regulates β-Catenin Activity. *Nature* **2008**, *455*, 547–551.

5. Oliveira, D. M.; Santamaria, G.; Laudanna, C.; Migliozzi, S.; Zoppoli, P.; Quist, M.; Grasso, C.; Mignogna, C.; Elia, L.; Faniello, M. C.; Marinaro, C.; Sacco, R.; Corcione, F.; Viglietto, G.; Malanga, D.; Rizzuto, A. Identification of Copy Number Alterations in Colon Cancer from Analysis of Amplicon-Based next Generation Sequencing Data. *Oncotarget* **2018**, *9*, 20409–20425.

6. Tsafrir, D.; Bacolod, M.; Selvanayagam, Z.; Tsafrir, I.; Shia, J.; Zeng, Z.; Liu, H.; Krier, C.; Stengel, R. F.; Barany, F.; Gerald, W. L.; Paty, P. B.; Domany, E.; Notterman, D. A. Relationship of Gene Expression and Chromosomal Abnormalities in Colorectal Cancer. *Cancer Res.* **2006**, *66*, 2129–2137.

7. Dale, T.; Clarke, P. A.; Esdar, C.; Waalboer, D.; Adeniji-Popoola, O.; Ortiz-Ruiz, M. J.; Mallinger, A.; Samant, R. S.; Czodrowski, P.; Musil, D.; Schwarz, D.; Schneider, K.; Stubbs, M.; Ewan, K.; Fraser, E.; TePoele, R.; Court, W.; Box, G.; Valenti, M.; De Haven Brandon, A.; Gowan, S.; Rohdich, F.; Raynaud, F.; Schneider, R.; Poeschke, O.; Blaukat, A.; Workman, P.; Schiemann, K.; Eccles, S. A.; Wienke, D.; Blagg, J. A Selective Chemical Probe for Exploring the Role of CDK8 and CDK19 in Human Disease. *Nat. Chem. Biol.* **2015**, *11*, 973–980.

8. Kapoor, A.; Goldberg, M. S.; Cumberland, L. K.; Ratnakumar, K.; Segura, M. F.; Emanuel, P. O.; Menendez, S.; Vardabasso, C.; LeRoy, G.; Vidal, C. I.; Polsky, D.; Osman, I.; Garcia, B. A.; Hernando, E.; Bernstein, E. The Histone Variant MacroH2A Suppresses Melanoma Progression through Regulation of CDK8. *Nature* **2010**, *468*, 1105–1109.

9. Porter, D. C.; Farmaki, E.; Altilia, S.; Schools, G. P.; West, D. K.; Chen, M.; Chang, B.-D.; Puzyrev, A. T.; Lim, C.; Rokow-Kittell, R.; Friedhoff, L. T.; Papavassiliou, A. G.; Kalurupalle, S.; Hurteau, G.; Shi, J.; Baran, P. S.; Gyorffy, B.; Wentland, M. P.; Broude, E. V.; Kiaris, H.; Roninson, I. B. Cyclin-Dependent Kinase 8 Mediates Chemotherapy-Induced Tumor-Promoting Paracrine Activities. *Proc. Natl. Acad. Sci.* **2012**, *109*, 13799–13804.

10. Kim, M.-Y.; Han, S. I.; Lim, S.-C.; Lim. Roles of Cyclin-Dependent Kinase 8 and β-Catenin in the Oncogenesis and Progression of Gastric Adenocarcinoma. *Int. J. Oncol.* **2011**, *38*, 1375–

1383.

11. Xu, W.; Ji, J.-Y. Dysregulation of CDK8 and Cyclin C in Tumorigenesis. *J. Genet. Genomics* **2011**, *38*, 439–452.

12. Flygare, J.; Johansson, L.; Lundbäck, T. Compounds for Treatment of Hypoproliferative Disorders. WO2017076968A1, 2017.

13. Chen, J.; Siva, K.; Rzymski, T.; Johansson, L.; Lundbäck, T.; Nunez Villacis, L.; Ek, F.; Wang, B.; George, A. J.; Wan, Y.; Shi, L.; Ilsley, M.; Subramaniam, A.; Jain, M.; Debnath, S.; Ghani Alattar, A.; Axelsson, H.; Mazan, M.; Majewska, E.; Tedgard, U. R.; Wlodarski, M. W.; Gustavsson, A.-L.; Olsson, R.; Mikula, M.; Zhu, X.; Brzózka, K.; Hannan, R.; Flygare, J. Small Molecule Screens Identify CDK8-Inhibitors As Candidate Diamond-Blackfan Anemia Drugs. *Blood* **2018**, *132*, 753–753.

14. Vlachos, A.; Muir, E. How I Treat Diamond-Blackfan Anemia. *Blood* **2010**, *116*, 3715–3723.

15. Wilkes, M. C.; Siva, K.; Chen, J.; Varetti, G.; Youn, M. Y.; Chae, H.; Ek, F.; Olsson, R.; Lundbäck, T.; Dever, D. P.; Nishimura, T.; Narla, A.; Glader, B.; Nakauchi, H.; Porteus, M. H.; Repellin, C. E.; Gazda, H. T.; Lin, S.; Serrano, M.; Flygare, J.; Sakamoto, K. M. Diamond Blackfan Anemia Is Mediated by Hyperactive Nemo-like Kinase. *Nat. Commun.* **2020**, *11*.

16. Fabbro, D.; Cowan-Jacob, S. W.; Moebitz, H. Ten Things You Should Know about Protein Kinases: IUPHAR Review 14. *Br. J. Pharmacol.* **2015**, *172*, 2675–2700.

17. Johnson, L. N.; Noble, M. E. M.; Owen, D. J. Active and Inactive Protein Kinases: Structural Basis for Regulation. *Cell* **1996**, *85*, 149–158.

18. Roskoski, R. Classification of Small Molecule Protein Kinase Inhibitors Based upon the Structures of Their Drug-Enzyme Complexes. *Pharmacol. Res.* **2016**, *103*, 26–48.

19. McClendon, C. L.; Kornev, A. P.; Gilson, M. K.; Taylor, S. S. Dynamic Architecture of a Protein Kinase. *Proc. Natl. Acad. Sci.* **2014**, *111*, E4623–E4631.

20. Ollagnier-de Choudens, S.; Loiseau, L.; Sanakis, Y.; Barras, F.; Fontecave, M. Quinolinate Synthetase, an Iron-Sulfur Enzyme in NAD Biosynthesis. *FEBS Lett.* **2005**, *579*, 3737–3743.

21. Chan, A.; Clémancey, M.; Mouesca, J.-M.; Amara, P.; Hamelin, O.; Latour, J.-M.; Ollagnier de Choudens, S. Studies of Inhibitor Binding to the [4Fe-4S] Cluster of Quinolinate Synthase. *Angew. Chemie Int. Ed.* **2012**, *51*, 7711–7714.

22. Cherrier, M. V.; Chan, A.; Darnault, C.; Reichmann, D.; Amara, P.; Ollagnier De Choudens, S.; Fontecilla-Camps, J. C. The Crystal Structure of Fe4S4 Quinolinate Synthase Unravels an Enzymatic Dehydration Mechanism That Uses Tyrosine and a Hydrolase-Type Triad. *J. Am. Chem. Soc.* **2014**, *136*, 5253–5256.

23. Waterhouse, A.; Bertoni, M.; Bienert, S.; Studer, G.; Tauriello, G.; Gumienny, R.; Heer, F. T.; de Beer, T. A. P.; Rempfer, C.; Bordoli, L.; Lepore, R.; Schwede, T. SWISS-MODEL: Homology Modelling of Protein Structures and Complexes. *Nucleic Acids Res.* **2018**, *46*, W296–W303.

24. Bietz, S.; Urbaczek, S.; Schulz, B.; Rarey, M. Protoss: A Holistic Approach to Predict Tautomers and Protonation States in Protein-Ligand Complexes. *J. Cheminform.* **2014**, *6*, 12.

25. Desaphy, J.; Azdimousa, K.; Kellenberger, E.; Rognan, D. Comparison and Druggability

Prediction of Protein–Ligand Binding Sites from Pharmacophore-Annotated Cavity Shapes. *J. Chem. Inf. Model.* **2012**, *52*, 2287–2299.

26. Eguida, M.; Rognan, D. A Computer Vision Approach to Align and Compare Protein Cavities: Application to Fragment-Based Drug Design. *J. Med. Chem.* **2020**, *63*, 7127–7142.

27. Ackloo, S.; Al-awar, R.; Amaro, R. E.; Arrowsmith, C. H.; Azevedo, H.; Batey, R. A.; Bengio, Y.; Betz, U. A. K.; Bologa, C. G.; Chodera, J. D.; Cornell, W. D.; Dunham, I.; Ecker, G. F.; Edfeldt, K.; Edwards, A. M.; Gilson, M. K.; Gordijo, C. R.; Hessler, G.; Hillisch, A.; Hogner, A.; Irwin, J. J.; Jansen, J. M.; Kuhn, D.; Leach, A. R.; Lee, A. A.; Lessel, U.; Morgan, M. R.; Moult, J.; Muegge, I.; Oprea, T. I.; Perry, B. G.; Riley, P.; Rousseaux, S. A. L.; Saikatendu, K. S.; Santhakumar, V.; Schapira, M.; Scholten, C.; Todd, M. H.; Vedadi, M.; Volkamer, A.; Willson, T. M. CACHE (Critical Assessment of Computational Hit-Finding Experiments): A Public–Private Partnership Benchmarking Initiative to Enable the Development of Computational Methods for Hit-Finding. *Nat. Rev. Chem.* **2022**, *6*, 287–295.

28. Tolosa, E.; Vila, M.; Klein, C.; Rascol, O. LRRK2 in Parkinson Disease: Challenges of Clinical Trials. *Nat. Rev. Neurol.* **2020**, *16*, 97–107.

29. Kluss, J. H.; Lewis, P. A.; Greggio, E. Leucine-Rich Repeat Kinase 2 (LRRK2): An Update on the Potential Therapeutic Target for Parkinson's Disease. *Expert Opin. Ther. Targets* **2022**, *26*, 537–546.

30. Jennings, D.; Huntwork-Rodriguez, S.; Henry, A. G.; Sasaki, J. C.; Meisner, R.; Diaz, D.; Solanoy, H.; Wang, X.; Negrou, E.; Bondar, V. V.; Ghosh, R.; Maloney, M. T.; Propson, N. E.; Zhu, Y.; Maciuca, R. D.; Harris, L.; Kay, A.; LeWitt, P.; King, T. A.; Kern, D.; Ellenbogen, A.; Goodman, I.; Siderowf, A.; Aldred, J.; Omidvar, O.; Masoud, S. T.; Davis, S. S.; Arguello, A.; Estrada, A. A.; de Vicente, J.; Sweeney, Z. K.; Astarita, G.; Borin, M. T.; Wong, B. K.; Wong, H.; Nguyen, H.; Scearce-Levie, K.; Ho, C.; Troyer, M. D. Preclinical and Clinical Evaluation of the LRRK2 Inhibitor DNL201 for Parkinson's Disease. *Sci. Transl. Med.* **2022**, *14*, 1–18.

31. Deniston, C. K.; Salogiannis, J.; Mathea, S.; Snead, D. M.; Lahiri, I.; Matyszewski, M.; Donosa, O.; Watanabe, R.; Böhning, J.; Shiau, A. K.; Knapp, S.; Villa, E.; Reck-Peterson, S. L.; Leschziner, A. E. Structure of LRRK2 in Parkinson's Disease and Model for Microtubule Interaction. *Nature* **2020**, *588*, 344–349.

32. Schapira, M.; Tyers, M.; Torrent, M.; Arrowsmith, C. H. WD40 Repeat Domain Proteins: A Novel Target Class? *Nat. Rev. Drug Discov.* **2017**, *16*, 773–786.

33. Zhang, P.; Fan, Y.; Ru, H.; Wang, L.; Magupalli, V. G.; Taylor, S. S.; Alessi, D. R.; Wu, H. Crystal Structure of the WD40 Domain Dimer of LRRK2. *Proc. Natl. Acad. Sci.* **2019**, *116*, 1579–1584.

34. Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAP - Retrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511–522.

35. Desaphy, J.; Bret, G.; Rognan, D.; Kellenberger, E. Sc-PDB: A 3D-Database of Ligandable Binding Sites—10 Years On. *Nucleic Acids Res.* **2014**, *43*, D399–D404.

36. Da Silva, F.; Desaphy, J.; Rognan, D. IChem: A Versatile Toolkit for Detecting, Comparing, and Predicting Protein–Ligand Interactions. *ChemMedChem* **2018**, *13*, 507–510.

37.     Bateman, A. UniProt: A Worldwide Hub of Protein Knowledge. *Nucleic Acids Res.* **2019**, *47*, D506–D515.

38.     Korb, O.; Stützle, T.; Exner, T. E. PLANTS: Application of Ant Colony Optimization to Structure-Based Drug Design BT  - Ant Colony Optimization and Swarm Intelligence; Dorigo, M., Gambardella, L. M., Birattari, M., Martinoli, A., Poli, R., Stützle, T., Eds.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2006; pp 247–258.

39.     Radoux, C. J.; Olsson, T. S. G.; Pitt, W. R.; Groom, C. R.; Blundell, T. L. Identifying Interactions That Determine Fragment Binding at Protein Hotspots. *J. Med. Chem.* **2016**, *59*, 4314–4325.

40.     Sprague, E. R. Structure of the C-Terminal Domain of Tup1, a Corepressor of Transcription in Yeast. *EMBO J.* **2000**, *19*, 3016–3027.

41.     Imrie, F.; Bradley, A. R.; van der Schaar, M.; Deane, C. M. Deep Generative Models for 3D Linker Design. *J. Chem. Inf. Model.* **2020**, *60*, 1983–1995.

42.     Schneider, G.; Fechner, U. Computer-Based de Novo Design of Drug-like Molecules. *Nat. Rev. Drug Discov.* **2005**, *4*, 649–663.

43.     Pierce, A. C.; Rao, G.; Bemis, G. W. BREED: Generating Novel Inhibitors through Hybridization of Known Ligands. Application to CDK2, P38, and HIV Protease. *J. Med. Chem.* **2004**, *47*, 2768–2775.

44.     Sydow, D.; Schmiel, P.; Mortier, J.; Volkamer, A. KinFragLib: Exploring the Kinase Inhibitor Space Using Subpocket-Focused Fragmentation and Recombination. *J. Chem. Inf. Model.* **2020**, *60*, 6081–6094.

45.     Lewis, R. A.; Roe, D. C.; Huang, C.; Ferrin, T. E.; Langridge, R.; Kuntz, I. D. Automated Site-Directed Drug Design Using Molecular Lattices. *J. Mol. Graph.* **1992**, *10*, 66–78.

46.     Pearlman, D. A.; Murcko, M. A. CONCEPTS: New Dynamic Algorithm Forde Novo Drug Suggestion. *J. Comput. Chem.* **1993**, *14*, 1184–1193.

47.     Ramensky, V.; Sobol, A.; Zaitseva, N.; Rubinov, A.; Zosimov, V. A Novel Approach to Local Similarity of Protein Binding Sites Substantially Improves Computational Drug Design Results. *Proteins Struct. Funct. Bioinforma.* **2007**, *69*, 349–357.

48.     Bshm, H.-J. The Computer Program LUDI: A New Method for the de Novo Design of Enzyme Inhibitors. *J. Comput. Aided. Mol. Des.* **1992**, *6*, 61–78.

49.     Wang, R.; Gao, Y.; Lai, L. LigBuilder: A Multi-Purpose Program for Structure-Based Drug Design. *J. Mol. Model.* **2000**, *6*, 498–516.

50.     Penner, P.; Martiny, V.; Gohier, A.; Gastreich, M.; Ducrot, P.; Brown, D.; Rarey, M. Shape-Based Descriptors for Efficient Structure-Based Fragment Growing. *J. Chem. Inf. Model.* **2020**, *60*, 6269–6281.

51.     Rotstein, S. H.; Murcko, M. A. GroupBuild: A Fragment-Based Method for de Novo Drug Design. *J. Med. Chem.* **1993**, *36*, 1700–1710.

52.     Lewis, R. A. Automated Site-Directed Drug Design: A Method for the Generation of General Three-Dimensional Molecular Graphs. *J. Mol. Graph.* **1992**, *10*, 131–143.

53.     Moriaud, F.; Doppelt-Azeroual, O.; Martin, L.; Oguievetskaia, K.; Koch, K.; Vorotyntsev, A.;

Adcock, S. A.; Delfaud, F. Computational Fragment-Based Approach at PDB Scale by Protein Local Similarity. *J. Chem. Inf. Model.* **2009**, *49*, 280–294.

54. Durrant, J. D.; Friedman, A. J.; McCammon, J. A. CrystalDock: A Novel Approach to Fragment-Based Drug Design. *J. Chem. Inf. Model.* **2011**, *51*, 2573–2580.

55. Jambon, M.; Imberty, A.; Deléage, G.; Geourjon, C. A New Bioinformatic Approach to Detect Common 3D Sites in Protein Structures. *Proteins Struct. Funct. Bioinforma.* **2003**, *52*, 137–145.

56. Desaphy, J.; Rognan, D. Sc-PDB-Frag: A Database of Protein-Ligand Interaction Patterns for Bioisosteric Replacements. *J. Chem. Inf. Model.* **2014**, *54*, 1908–1918.

57. Kalliokoski, T.; Olsson, T. S. G.; Vulpetti, A. Subpocket Analysis Method for Fragment-Based Drug Discovery. *J. Chem. Inf. Model.* **2013**, *53*, 131–141.

58. Drwal, M. N.; Bret, G.; Kellenberger, E. Multi-Target Fragments Display Versatile Binding Modes. *Mol. Inform.* **2017**, *36*, 1–7.

59. Giordanetto, F.; Jin, C.; Willmore, L.; Feher, M.; Shaw, D. E. Fragment Hits: What Do They Look Like and How Do They Bind? *J. Med. Chem.* **2019**, *62*, 3381–3394.

60. Maass, P.; Schulz-Gasch, T.; Stahl, M.; Rarey, M. Recore: A Fast and Versatile Method for Scaffold Hopping Based on Small Molecule Crystal Structure Conformations. *J. Chem. Inf. Model.* **2007**, *47*, 390–399.

61. Ferenczy, G. G.; Keseru, G. M. On the Enthalpic Preference of Fragment Binding. *Medchemcomm* **2016**, *7*, 332–337.

62. Erlanson, D. A.; Fesik, S. W.; Hubbard, R. E.; Jahnke, W.; Jhoti, H. Twenty Years on: The Impact of Fragments on Drug Discovery. *Nat. Rev. Drug Discov.* **2016**, *15*, 605–619.

63. Ertl, P.; Schuffenhauer, A. Estimation of Synthetic Accessibility Score of Drug-like Molecules Based on Molecular Complexity and Fragment Contributions. *J. Cheminform.* **2009**, *1*, 1–11.

64. Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.

# Annex 4

**Annex 4.1.** Sequence alignment of *Thermotoga maritima, Helicobacter Pylori* and *Mycobacterium leprae* quinolate synthase.



⋆ Cavity residue

# CHAPTER 5

Perspectives: from cavities to ligands

# 5.1. Context

At the earlier phases of drug discovery programs, structure-based virtual screening is one of the deployed strategies if the target structure is available and a binding pocket characterized. It popularized since it aims at identifying initial hits with minimal cost and experimental efforts.[1] Starting from a carefully designed virtual library, a few-steps workflow is often implemented to progressively filter bad propositions out and focus more computational resources on promising compounds. At the later stages, heavier computational methods such as binding free energy calculations (e.g., MM-GBSA, FEP) which consider the bound and unbound states of the receptor-ligand complexes in simulated dynamics can be carried on a few candidates for final prioritization.[2] Contrarily, the initial steps of the workflow require faster methods which can process many molecules in a comparatively short space of time.

Three-dimensional (3D) pharmacophore screening is adapted to this task, is intuitive to human understanding and can be fuzzy enough to escape problems known to structure-based methods (target flexibility, target-dependent parametrization, accuracy of scoring functions in ranking).[3–6] According to the International Union of Pure and Applied Chemistry (IUPAC), a pharmacophore is "an ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target and to trigger (or to block) its biological response".[7] Pure ligand-based pharmacophores are generated from a set of known ligands that exhibited the investigated biological activity[8,9] but are quickly limited by two factors: (i) diversity of the training set, (ii) absence of the receptor constraints, (iii) inapplicability to apo target structures for which no bound ligand is available. When protein-ligand complexes are available, 3D structure-based pharmacophore incorporate interaction and hindrance information to select or exclude features but are concerned by the limitations stated above in (i) and (iii).[8,9] Still, orphan proteins would benefit from pharmacophore modelling that relies on the protein structure only. The prediction of areas in apo proteins, that are favorable or that would highly contribute to binding (hotspot) is performed by analyzing properties (molecular fields, pharmacophoric features) at atomic level on 3D lattice (e.g., GRID,[10] SuperStar,[11] VolSite[12]), at fragment level (e.g. FTMAP[13]) or processing predictions of other methods (Radoux *et al.* based on GRID).[14] Attributes are defined by interaction potentials with probes (e.g., FLAP[4]) or empirically by analyzing the relative position of the cavity features (HS-Pharm,[15] Snooker,[16] VolSite[12]). Some methods integrate pharmacophoric patterns from molecular dynamics trajectories (GRAIL,[17] MCSS,[18] SILCS[19]). Following the pharmacophores definition, small molecules are screened by confronting the ligand to the target space, either by fingerprint comparison (FLAP) or by 3D alignment (LigandScout,[20] PHASE,[21] Shaper2[22]).[23] In most cases, the generation of multiple conformations of the ligands are required prior to the screening but some methods can generate them on the fly.[9] Strikingly, several of the available methods to achieve pharmacophore modeling and screening are part of commercial software without free academic license: e.g., Radoux *et al.*[14] (The Cambridge Crystallographic Data Centre, Cambridge, UK), FLAP[4] (Molecular Discovery, Borehamwood, UK), LigandScout[20] (Inte:Ligand, Vienna, Austria),

Catalyst[24] (Dassault Systèmes Biovia, Velizy-Villacoublay, France), Molecular Operating Environment (Chemical Computing group, Montréal, Canada), PHASE[21] (Schrödinger, New York, USA).

The idea that VolSite cavities[12] mimic some ligand features in the volumetric ligand space led to the definition of pharmacophores and alignment-based screening in a recent study of my host laboratory.[22] By default, VolSite cavities are dense (~300 points) but remain comparable to ligand atoms (~30). The ideal method would be able to pick the relevant areas from these dense clouds and match them to consistent ligand features. Previous attempts by global shape matching (Shaper) failed to reproduce known X-ray poses.[22] Indeed, visual inspection of hundreds of cavities showed that VolSite points are spread to areas not occupied by ligand atoms, which add complexity to the search. Reducing the cavity by selecting or grouping points that would match with the ligand features led to: (i) a visually interpretable pharmacophore that can serve for many purposes, and (ii) an improvement of the subsequent alignments. However, we herein wished to overcome two limitations :

(a) the resulting VolSite-derived pharmacophores were defined by empirical rules parametrized on a few cases and which might not generalize on certain targets,

(b) the alignments were optimized and scored in the receptor binding site by potential energy minimization using the MMFF94 force field[25] in OpenEye Szybki (OpenEye Scientific Software, Santa Fe, USA).

As a continuation of our previous work[26] in **Chapter 2** and inspired by the machine-learning-based pharmacophore modelling method HS-Pharm,[15] we herein aimed at developing a purely topological tool for ligand-cavity alignment and a model for denoising VolSite cavities.

## 5.2. Materials and methods

**Datasets**

The sc-PDB database[27] of curated protein-ligand complexes were used in versions 2016 (16 150 entries) and 2022 (37 922 entries, Bret *et al.,* unpublished). Entries were protonated according to Protoss v.4 rules and saved into TRIPOS mol2 format.

The sc-PDB diverse set was compiled from the sc-PDB 2016.[22] Following the pairwise comparisons of the complexes interaction graphs using IChem Grim,[28] the agglomerative clustering of the similarity (GrimScore) matrix with a threshold of 0.70 was applied to obtain 176 protein-ligand complexes exhibiting diverse and non-redundant interaction patterns.

**Representations of protein cavities**

Protein cavities were represented by four images (**Figure 5.1**):

- *VolSite cloud of points ('cavity ALL')*, the default VolSite implementation described in **Chapters 1-4**.[12,29]

- *VolSite pharmacophores ('cavity pharm')* obtained by recently described post-processing rules.[22] Briefly, a set of 213 protein-ligand complexes were used to learn the properties of an ideal pharmacophore defined by the ligand atoms. The 'cavity ALL' points were then pruned according to these rules and refined by considering the directionality of polar interactions and sufficient hydrophobic neighborhood for this feature. Points not fulfilling these rules were removed. In a later stage, the remaining points were hierarchically clustered to yield cavities of less than 50 points (version used in this work). Contrarily to default VolSite cavities, 'cavity pharm' are assigned seven possible VolSite properties (hydrophobic, aromatic, h-bond donor, h-bond acceptor, h-bond acceptor, and donor, positive ionizable, negative ionizable) and an additional 'metal' property.

- *Projected points ('cavity projected')* obtained by projecting cavity-lining atoms into the ligand space instead of sampling a grid. The 'cavity projected' points were generated by first delimitating the protein heavy atoms within 3.5 Å from any 'cavity ALL' point, keeping track of the residues they originate from. The centroid of the cavity was calculated as the center of mass of these atoms. In a similar fashion to KRIPO pharmacophores,[30] these atoms were defined as 'root' and projected (3.5 – 4 Å from the root) into the cavity space by ensuring that the angle point-root-centroid falls within 90°. Aromatic rings were represented by their center of mass. Points were annotated by seven features to be complementary to the properties of the protein atom they originate from according to VolSite rules (hydrophobic, aromatic, h-bond donor, h-bond acceptor, h-bond acceptor, and donor, positive ionizable, negative ionizable).

- *VolSite simplified cloud of points ('cavity pruned')* generated from the 'cavity ALL' by keeping only points of identical features within d Å, d ∈ {1.5, 2} from the ligand interacting atoms. Interactions were detected with IChem[29]. This representation mimics the ideal pharmacophoric points that match the ligand features and geometry.
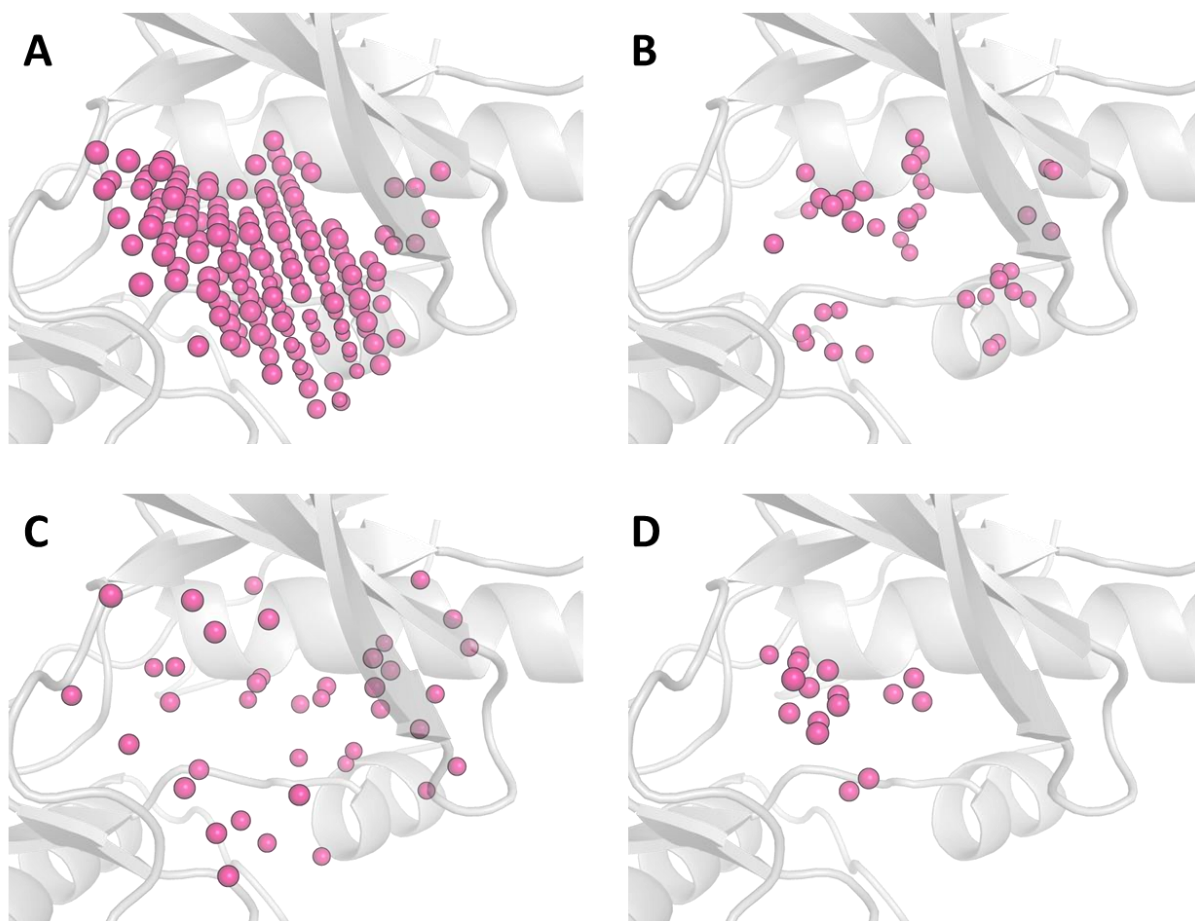


**Figure 5.1.** Different representations of a protein cavity. Spheres represent the cavity points of eight possible features: hydrophobic, aromatic, h-bond acceptor, h-bond donnor, h-bond acceptor or donor, positive, negative, dummy. A) VolSite 'cavity ALL', B) VolSite 'cavity pharm', C) 'cavity projected', D) VolSite 'cavity pruned' determined at 1.5 Å from the ligand. PDB entry: 4CCB. For this entry, the number of points were respectively 164, 38, 40, and 16 in cavities A) to D).

**Representation of ligands**

Ligands in TRIPOS (Certara, Princeton, USA) mol2 format were processed to assign pharmacophoric features to atomic positions, according to their connectivity and atom types. Briefly aliphatic carbon, sulfur, halogen atoms were assigned hydrophobic features if not bounded to any heteroatom. Aromatic features were defined by aromatic atoms (C.ar and N.ar atom types). Aromatic-labelled points were by

extension also annotated as hydrophobic. Nitrogen and oxygen atoms were assigned h-bond donor feature if they are connected to hydrogen atoms, otherwise h-bond acceptor. Positions which satisfy both h-bond donor and acceptor were additionally annotated 'donor and acceptor' features (e.g. sp3 oxygen connected to a hydrogen atom). Positively charged heteroatoms were assigned 'positive' features and h-bond donor if applicable, whereas negatively charged heteroatoms were annotated with 'negative' feature and H-bond acceptor. A particular treatment was applied to ring systems to cluster their atoms of the same feature into their center of mass. Atoms that could not be assigned any feature were disregarded. According to these rules, multiple features can be assigned to the same position. We later refer to this representation as 'lig pharm' (**Figure 5.2**).
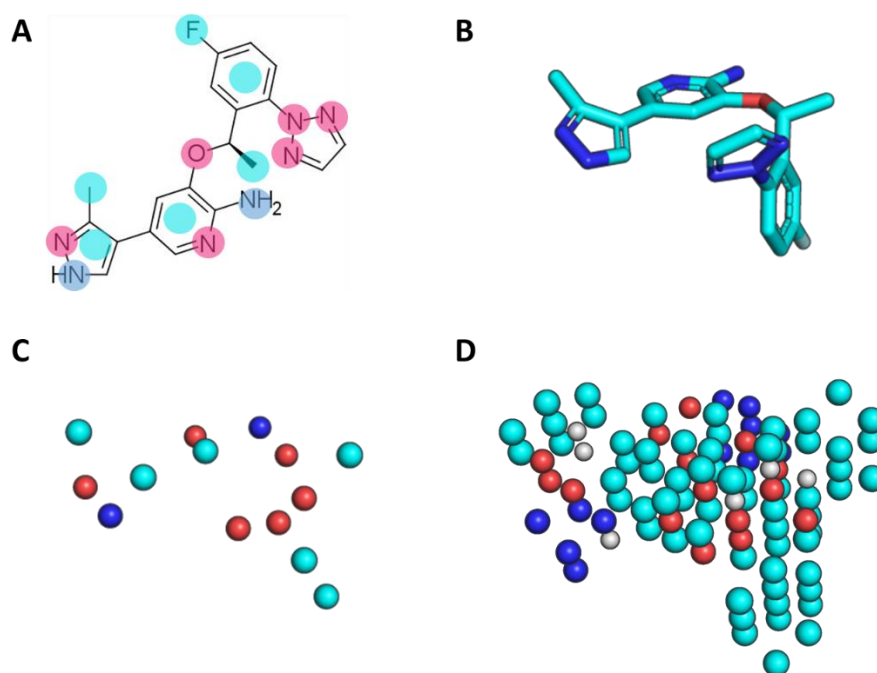


**Figure 5.2.** Representations of a ligand. A) two-dimensional structure highlighting pharmacophoric features by dots. B) X-ray conformation processed into 3D pharmacophoric representations: C) 'lig pharm' and D) 'ligvoxel+'. PDB entry: 4CCB. Red points correspond to h-bond acceptor or negative ionizable, blue h-bond donor or positive, cyan aromatic or hydrophobic, white dummy.

An augmented representation of the ligands was generated by extending the 'lig pharm' points ('ligvoxel+'). The 'lig pharm' was put into a 3D grid of step r (r = 1 and 1.5 Å). Then, each voxel of the grid two-step away of a point (scanning through the x, y, z axes direction) was represented by its centroid and annotated by the features of the closest point. If annotation is ambiguous, compatibility rules are checked to prioritize one feature (e.g. aromatic will be preferred over hydrophobic, positive ionizable over h-bond donor, negative over h-bond acceptor) or 'dummy' is assigned in case of incompatibility (e.g., aromatic versus h-bond acceptor). In the version discussed here, only one feature is hence assigned per position (**Figure 5.2**).

**Point cloud registration**

Ligands and 'ligvoxel+' were translated (10 Å) and rotated (180° flip along the x axis) into different coordinate frames. Then ProCare (default parameters)[26] was used to realign the 'ligvoxel+' to the VolSite cavity for each entry. The resulting transformation matrices are applied to align the corresponding ligands. The root mean square deviations (RMSD) to the ligand X-ray positions were reported considering symmetry.

**Graph matching**

The protein cavity and ligand pharmacophoric points were represented as two separated graphs of all pairwise connections. Points, annotated by the same sets of pharmacophoric features formed the nodes. Edges were labelled by the Euclidian distance between these nodes. A product graph was built by comparing all possible combinations of nodes and edges in the two graphs, while tolerating a distance deviation of d = 2 Å by default (d is an adjustable parameter) and a strict match of the nodes' pharmacophoric features. Then using the Bron-Kerbosh algorithm,[31] all maximal cliques were found in the association graph. From the pairs of corresponding points between the cavity and the ligand representation obtained, a transformation matrix was applied to align the ligand representation points and atoms onto the cavity frame. The translation vectors were estimated by aligning the centroids of the correspondence sets, and the rotation matrices by the Kabsch algorithm[32] implemented in SciPy v.1.7.2.[33] Several scoring schemes were hierarchically implemented: the size of the clique nodes (**eq 5.1**), the root mean square error of the clique (**eq 5.2**), the coverage of the aligned ligand atoms (**eq 5.3**), a pharmacophoric score (**eq 5.4**), and a combo score (**eq 5.5**).

$$S = |M| \qquad\qquad \text{eq 5.1}$$

where $M$ is the set of the maximal clique pairs of nodes.

$$RMSE = \sqrt{\frac{\sum_i^N (P_i^C - P_i^L)^2}{N}} \qquad\qquad \text{eq 5.2}$$

where $P_i^C$ and $P_i^L$ are respectively the cartesian coordinates of the points in the cavity and aligned ligand representation for each correspondence $i$.

$$coverage = \frac{|A_b|}{|A|} \qquad\qquad \text{eq 5.3}$$

where $A_b$ is the set of aligned ligand atoms buried in the cavity cloud within 2 Å of any cavity points and A is the set of all ligand atoms.

$$ph4_{score} = \sum_{i}^{E} w_i \qquad\qquad \text{eq 5.4}$$

where $w_i$ are the weights of the edges $E$ of the maximal clique. $w_i$ is arbitrarily set to 1 when a pair of polar features is involved, 0.5 when the edge connects hydrophobic nodes. In a different setting, $w_i$ corresponds to the inverse of the frequency of the point feature in the sc-PDB.

$$combo_{score} = \frac{ph4_{score} \times coverage}{RMSE} \qquad\qquad \text{eq. 5.5}$$

The RMSD of the aligned ligands to the X-ray pose were reported considering the symmetry.

**Cavity point descriptors**

Staring from the VolSite 'cavity ALL' of the sc-PDB v. 2022, a set of descriptors were computed for each point:

    (a) FP1 is an 8-bit fingerprint which encodes the VolSite physicochemical features of the point (hydrophobic, aromatic, h-bond donor, h-bond acceptor, h-bond acceptor and donor, positive ionizable, negative ionizable, dummy). Additionally, some points can activate more than one bit by compatibility rules: aromatic points are additionally considered hydrophobic, negative are h-bond donor, positive are h-bond acceptors, acceptor-donor additionally activates both the donor and acceptor bits.

    (b) FP2 is the 12-bit fingerprint encoding the buriedness of the point. A set of 114 regular rays of equally-spaced solid angles (22.5°) and 8 Å length were projected from the point in focus. Then, the buriedness is estimated as the number of rays that pass less than 1.5 Å away from any protein atom. The buriedness values were binned from the lowest value 0 to the highest value 114 with an increment of 10 units. The corresponding bit of the point buriedness is activated.

    (c) FP3 is a 24-position fingerprint counting each of the eight pharmacophoric features in three concentric neighborhoods of 1.5, 3 and 4.5 Å distance from the point.

    (d) FP4 is a 288-bin histogram which encodes the proportion of points for each combination of pharmacophoric features and buriedness intervals, in the three concentric neighborhoods.

    (e) descriptor FP5 is the Euclidean distance of the point to the centroid of the cavity.

Accordingly, a total of 333 descriptors were obtained for each point.

**Cavity point prediction**

A thousand of 'cavity ALL' entries were randomly extracted from the sc-PDB as application test set. Then, the remaining entries (36 922) were processed to positively label points within 1.5 Å to a ligand atom interacting with the protein according to IChem[29] and of the same pharmacophore feature ('lig

pharm'). Any other point is labeled as negative. Points from all cavities were pooled to generate a set for each of the seven VolSite features (dummy points were disregarded) and the data was balanced by randomly sampling the same number of negative and positive in each set. It was verified that the sampling did not overrepresent particular PBD entries.

Random Forest models were trained to classify VolSite 'cavity ALL' points as *interacting* (positive class) or *non-interacting* (negative class), using the 333 descriptors. The above-described data of labelled points were split into a training (75%) and external test set (25%). The training set was subjected to a five-fold cross-validation (CV) using the Scikit-learn classifier with 100 trees and a number of splits equal to the square root of the number of descriptors. The final model trained on all the training set was applied to the external test set. The prediction accuracy (**eq 5.6**) of the CV training, CV test and external test sets, as well as the features importance were reported. The sensitivity, specificity and balanced accuracy were reported on the application set (**eq 5.7-5.9**).

$$ACC = \frac{TN + TP}{TN + TP + FN + FP} \qquad \text{eq 5.6}$$

where TN is the number of true negatives, TP true positives, FN false negatives and FP false positives.

$$sensitivity = \frac{TP}{TP + FN} \qquad \text{eq 5.7}$$

$$specificity = \frac{TN}{TN + FP} \qquad \text{eq 5.8}$$

$$BA = \frac{sensitivity + specificity}{2} \qquad \text{eq 5.9}$$

The models were saved and applied to the 1000 cavities in the application test set to save the predicted positive points for each cavity in mol2 files. A baseline model was implemented with the Gaussian naïve Bayes classifier in Scikit-learn, trained on the training set and evaluated on the external test set.

**Scripts and packages**

Inhouse scripts were used to process entries and analyze results in Python 3.7, using the following main packages and their dependencies: Matplotlib v.3.0.2, NetworkX v.2.6.3, NumPy 1.16.2, ProCare v.0.1.2, Scikit-learn v.0.24.2, SciPy v.1.7.2, maximal_clique (https://gist.github.com/abhin4v/8304062) after validation on easy synthetic data. The RMSD of ligands were computed with OpenEye Python API (OpenEye Scientific Software, Santa Fe, USA) when symmetry is considered.

# 5.3. Discussions and perspectives

We herein present preliminary results and discuss future directions.

Being able to properly align ligand atoms to the VolSite cavities, by solely considering topological and pharmacophoric features can offer an interesting alternative to the docking problem. A recent method was proposed to achieve this goal, relying on Gaussian shape (OpenEye Shape TK) alignment of the ligands on empirically-pruned VolSite pharmacophores, followed by energy minimization refinement (OpenEye Szybki).[22] Previous attempts of shape-only (topological and pharmacophoric) alignments failed to propose solutions close the X-ray pose of the ligands. It was therefore considered to apply a different algorithmic paradigm, such as discrete geometric pattern matching, instead of global shape matching. The point cloud registration approach implemented in ProCare,[26] and graph matching were investigated. The success of an alignment depends on three factors: finding the right correspondences, estimating a correct rotation and translation, top-scoring the right solutions. To evaluate the algorithms in their initial developments, the X-ray conformation of the ligands were used.

## 5.3.1. Point cloud registration of ligands to protein cavities

VolSite cavities are grid-based sampled points which adopt a regular disposition and do not correspond to mol2 atom types and relative positions in the ligand conformations. Contrary to the homogeneous comparison of protein cavity clouds, the solid point-to-point comparison of ligand features to protein cavity clouds requires a conversion into comparable objects, where the ligand space is similarly represented as the target space. This was achieved at two levels: (i) the featurization of ligand atoms into pharmacophoric types and (ii) a geometric transformation into grid voxels.

At step (i), ligand atoms were converted into seven possible VolSite features according to their atom types ('lig pharm', see **methods section**). Since ProCare first searches for initial alignment by associating the nearest neighbors according to the shape-pharmacophoric histograms (c-FPFH),[26] we first analyzed whether the c-FPFH of the 'lig pharm' and the cavities 'cavity ALL' (**Figures 5.1**, **5.2**) can establish good correspondences. A good correspondence is a pair of ligand and cavity point, which are each other's nearest neighbor in the c-FPFH descriptor space and are less than 2 Å apart in the X-ray pose. A minimum of three good correspondences are necessary to estimate a rotation. Analysis of the 16,000 sc-PDB v.2016 entries showed that only 25 % of the ligands were assigned more than 3 correspondences to theoretically enable a good alignment (**Figure 5.3**).
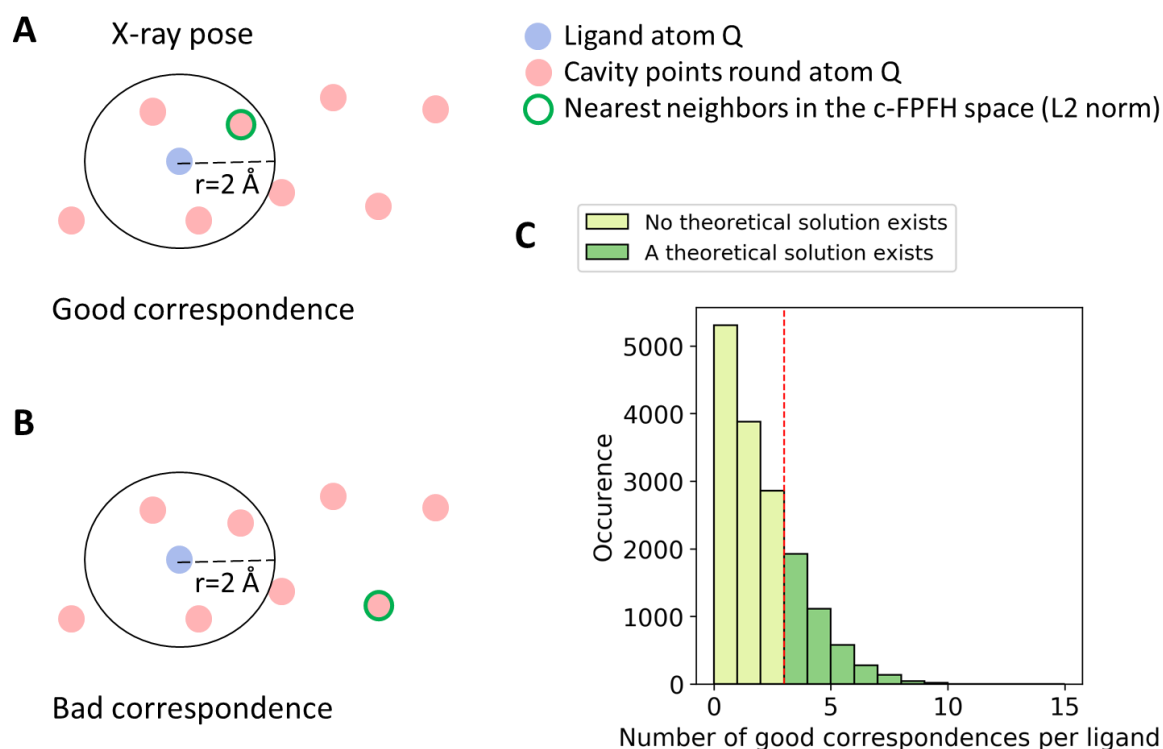
**Figure 5.3.** Analysis of the chances to correctly align ligand atoms to VolSite cavities. A) Description of a good c-FPFH-based correspondence, B) bad correspondence. C) Distribution of the count of good correspondences for each sc-PDB ligands.

These results were still encouraging since some ligands already contain shape and feature information, but not surprising as the few atoms of the ligand (10-30) could not properly describe a local shape and property neighborhood experienced in the cavities of more than 100 points. To apply the ProCare method, in step (ii) the ligand features were augmented in a grid by occupying the adjacent voxels of each atom along the x, y and z axes ('ligvoxel+', **Figure 5.2**). Starting from a different coordinates frame, the 176 'ligvoxel+' of the sc-PDB diverse set were realigned to the cavities 'cavity ALL' with ProCare default parameters and the resulting alignment matrices were used to align the corresponding ligands. **Figure 5.4** shows that 30 % of the ligands (53 entries) were aligned closed to their X-ray pose (RMSD $\leq 2$ Å) using the FPFH[34] descriptor. Increasing the grid resolution to 1 Å to better sample shapes did not improve the results. This posing approach is clearly less accurate than that achieved by state-of-the-art docking tools that commonly pose ca. 75% of ligands within 2 Å RMSD.[35,36]
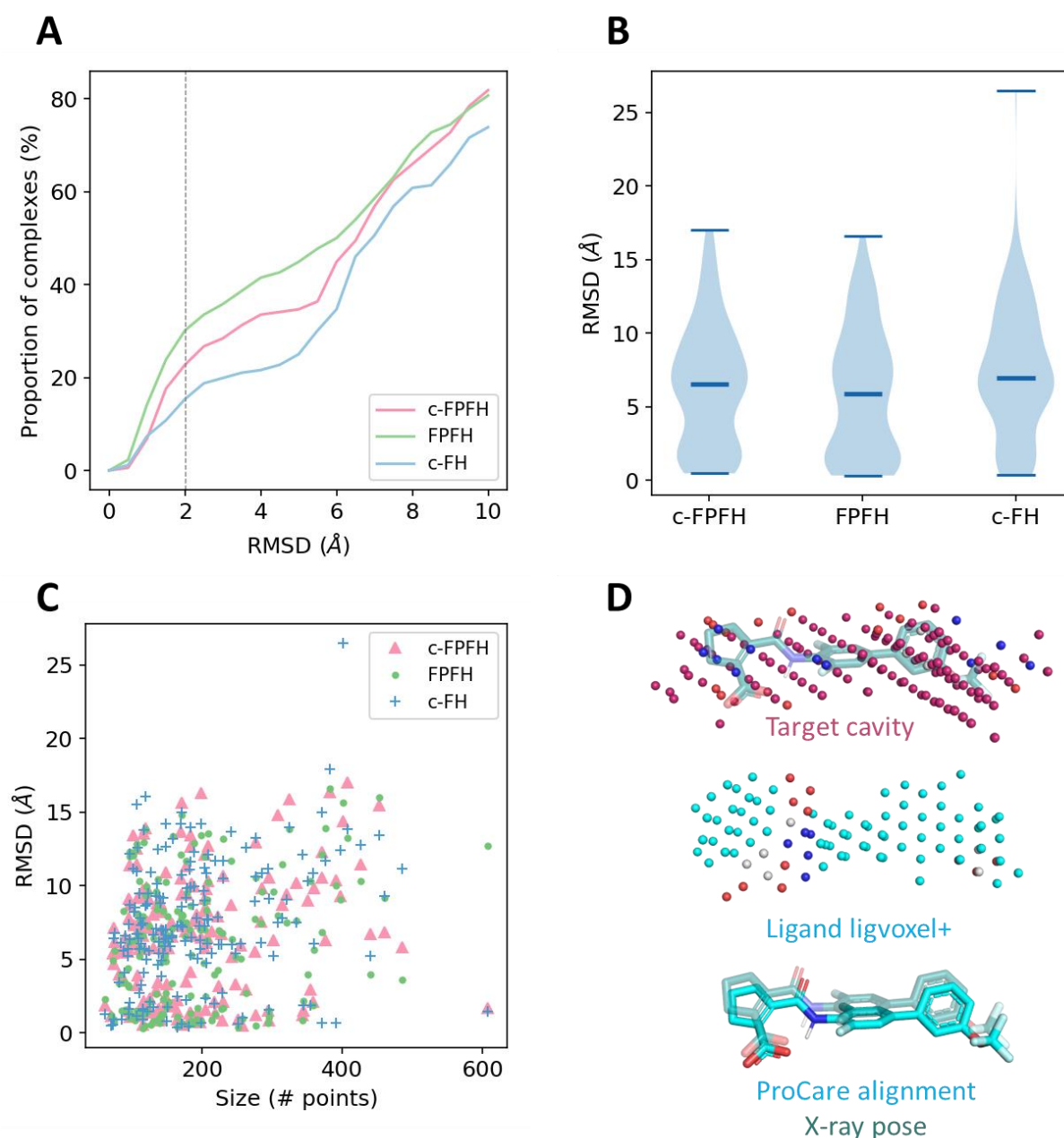
**Figure 5.4.** Prediction of the X-ray pose of the 176 sc-PDB diverse ligands by point cloud registration (ProCare)[26] to the target cavities. A) Cumulative percentage of ligands aligned within a threshold RMSD to the X-ray pose. At a threshold of 2 Å, 30%, 22% and 15% of the ligands were correctly aligned by the respective descriptors: shape-only FPFH,[34] hybrid c-FPFH[26] and features-only c-FH.[26] B) Violin plot distribution of the RMSDs showed a median value around 6 Å. C) The distribution of the RMSDs with respect of the size of the VolSite cavities does not show a size bias. D) Example (PDB ID: 2FPT) of c-FPFH correct alignment of the ligvoxel+ (cyan cloud) to the target cavity (warm pink cloud, masked ligand was shown for illustration but not used in the alignment) resulting in a good overlay with the ligand X-ray pose (transparent dark cyan), RMSD: 0.94 Å.

In summary, this section described the first attempts to align ligands to VolSite cavities using point cloud registration. Although originally skeptical about this approach, we showed that a minimal information is encoded in the ligand atoms and cavity points to allow relevant matches. Contrarily to the cavity-to-cavity comparison where the feature-only descriptor c-FH yielded equal to better results in some cases, the shape information of the FPFH seem to be crucial for the ligand-to-cavity alignment. Possible reasons for failure are:

(i)     the assignment of the pharmacophoric features to the ligand atoms 'lig pharm',

(ii)    the accuracy of the ligand representation 'ligvoxel+',

(iii)   the presence of noise in VolSite cavities while features are more uniformly grouped and the local shapes more rounded in the augmented ligands,

(iv)    the inadequacy of the c-FPFH descriptors to properly capture resemblances in this setting.

The above-derived conclusions can also be biased by highly represented ligands (e.g., nucleotide derivatives) in the sc-PDB diverse set. A proper study requires to compare the performances to other methods such as shape alignment and docking, starting from multiple conformations on several datasets. However, besides the poor performance, a practical limitation of this approach is the computing time. It takes approximately 1 to 2 seconds to align a single ligand conformation to a cavity point cloud. Therefore, its usage in large scale screening is hardly appealing, unless it would provide particular solutions unseen by other methods.

To escape the reasons evoked above (ii and iv), we applied a graph matching algorithm to the problem.

## 5.3.2. Graph matching of ligands to protein cavities

Contrarily to the ProCare approach where the exploration of the solutions is partially related to the transformation estimation (iterative closest point refinement), the search for common subgraph is independent of alignment estimation. However, graph matching algorithms are known to be computationally costly. Thus, we sequentially investigated the following aspects:

(i)     the ligand to cavity alignment speed,

(ii)    the identification of correct correspondences,

(iii)   the top-scoring of good solutions, and

(iv)    the estimation of rotation/translation.

Graphs of the two ligand representations ('lig pharm' and 'ligvoxel+') were compared to the graphs of the protein representations ('cavity ALL', 'cavity pharm' and 'cavity projected') following the algorithm described in the Methods section. Initial tests on three entries (PDB IDs: 2RH1, 2FV9, 3DKC) ruled out any comparison with the entire cavity 'cavity ALL' in a setting where all pairwise distances were investigated (**Table 5.1**). Restricting the graph definition to a certain interval of distances (1.5 -

4.1 Å) and to the connection of a certain nodes (e.g., discard hydrophobic-hydrophobic connections) would reduce the graph density and faster the search. This resulted in an acceptable running time (~1 s) to compare the 'lig pharm' to the 'cavity ALL'. We note that adjusting these parameters require an extensive study to generalize to many cases.

**Table 5.1.** Computing time of protein cavity and ligand pharmacophore graphs comparison.

| Representations | | Ligand | |
|---|---|---|---|
| | | *lig pharm* | *ligvoxel+* |
| Protein | *cavityALL* | 9 s (1 s)[a] | > 4 min |
| | *cavity pharm* | 0.5 s | 29 s |
| | *cavity projected* | 0.3 s | 0.3 s |

Green cells were considered acceptable time. The maximal time observed was reported.

[a] Reducing the graph density improved the running time.

In the next step, the performance of the algorithm to identify good correspondences (pairs of cavity and ligand points of the same feature, within 2 Å distance from the X-ray pose) was investigated on the 176 entries of the sc-PDB diverse set. Encouragingly, at least one good set of correspondences of more than three pairs could be found for 151 entries (86 %). However, for a successful comparison, these cliques must be top-ranked among many decoys (400 to 700 000 cliques). To this end, different scoring schemes (**eq 5.1-5.5**) were tested unsuccessfully. It proved hard to discriminate the correct cliques from the irrelevant ones by considering the size of the cliques and geometric constraints such as the *coverage* and RMSE after alignment. Comparison of the 'lig pharm' to the 'cavity pharm' and 'cavity projected' led to the same conclusions. Given that the RMSE and *coverage* are dependent on the alignment, we investigated the accuracy of the transformation estimation. In this final step, the ligand 'lig pharm' were transferred from their X-ray frame into a different coordinate frame and realigned to the 'cavity ALL' using the retrospectively known correspondences from the initial X-ray poses. Rotation and translation were estimated and applied to the ligand atoms using the Kabsch[32] implementation in SciPy (see **Methods**). The RMSD to the X-ray poses showed that even when knowing the pairs of points to associate, the estimation of rotation and translation barely yielded alignments within 2 Å from the X-ray poses (median RMSD: 5.5 Å), irrespective of the size of the ligands (**Figure 5.5A-C**). In contrast, the quality of the cavity delimitation with respect to the ligand might affect the propension to obtain good alignments in prospective searches where the cavity does not cover all the ligand substructures (**Figure 5.5D**). Visualization of several entries showed sub-optimal rotation estimation (**Figure 5.5E**). The reasons of these results are under investigation. Possible hypotheses are the planarity of the points, the bijectivity of the correspondences, or the use of other optimization algorithms to find correspondences. However, a spectacular improvement should not be expected: the topology of the

cavity and the ligands are not identical; therefore, transformation estimation, which is a minimization problem would always yield residuals that are not null, but rather the best compromises.
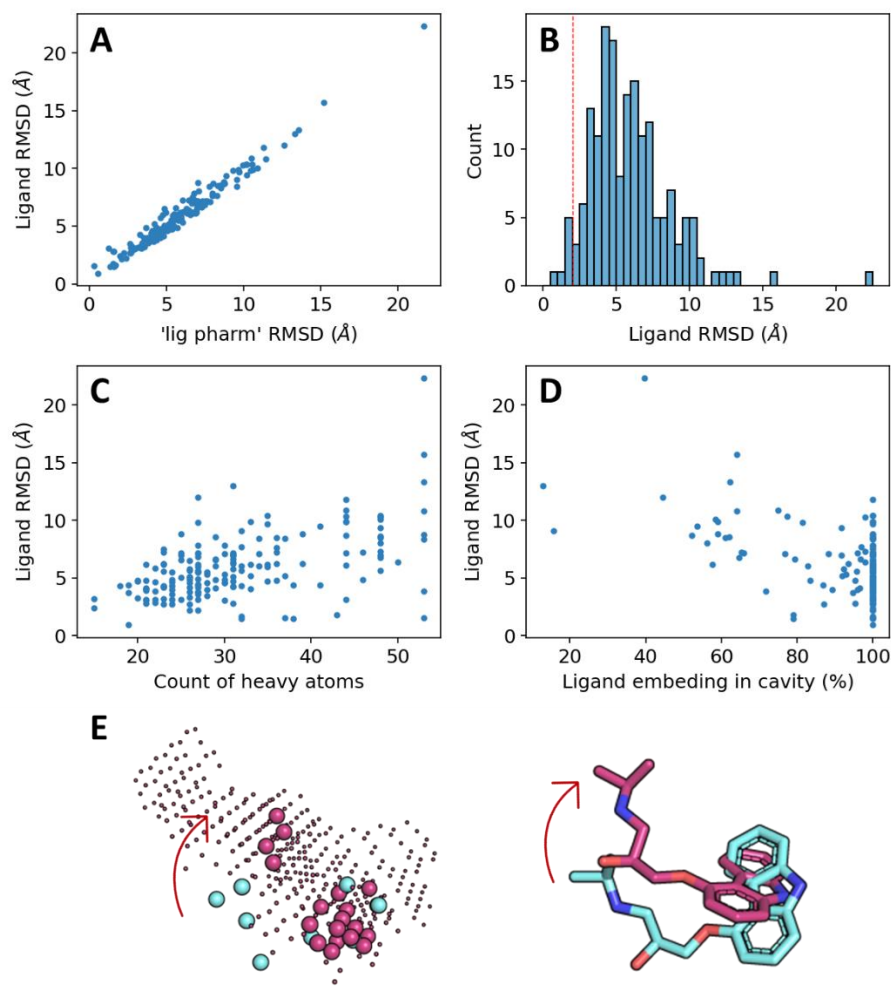


**Figure 5.5.** Best alignments of 'lig pharm' to 'cavity ALL' obtained using the X-ray correspondences for the 176 complexes in the sc-PDB diverse set. Obtained alignments were applied to the corresponding ligands, to compare their poses to the X-ray coordinates. A) The ligand RMSDs correlates well with the 'lig pharm' RMSD (Pearson r = 0.98), therefore only the ligands were further analyzed. B) Distribution of the ligands RMSDs: only 3% have a RMSD below 2 Å. C) Ligand RMSD with respect to ligand size. D) Ligands RMSD with respect to the quality of the cavity, defined as the percentage of ligand heavy atoms within 2 Å to a cavity point in the X-ray pose. E) Example of alignment estimation (PDB ID: 2RH1) showing the 'lig pharm' (light blue) superposed to the cavities (warm pink) using the correspondences (big spheres). Sub-optimal rotation of the tail (red arrow) led to a RMSD of 2.83 Å to the X-ray pose.

In a nutshell, this study revealed four key points:

- the ligand and cavity representations contain exploitable information for their point-to-point comparison and superposition by graph matching,
- the graph definition should be optimized to allow millisecond comparison of ligand features to entire VolSite cavities, otherwise other cavity representations should be used,
- the graph search can identify good point-to-point correspondences between the cavity and the ligand,
- a robust scoring needs to be investigated to top-rank the correct poses and later for discriminating between active and inactive molecules,
- the alignment estimation needs to be improved.

### 5.3.3. Prediction of pharmacophoric points from the apo target cavity

Predicting key points from the VolSite cavities can be valuable for different applications: better definition of the binding site for cavity-to-cavity comparisons (**Chapter 2**), improvement of ligand-to-cavity comparisons (**sections 5.3.1** and **5.3.2** above), and rescoring of docking poses. By defining *important* points as those that match with the interacting ligand features in proximity (modeled by 'cavity pruned'), Random Forest (RF) models were trained to discriminate the important from the so called *unused* points using a set of 333 descriptors. The datasets from the sc-PDB 2022 were prepared and split into training, external test and application set as described in the **Methods** section. A sample was used to train the model and a remaining sample which did not see the model was used for evaluation. As the number of entries were balanced in the negative and positives classes, the accuracy was reported in these earlier analyses. Initial models trained on a balanced ensemble of the seven features data (randomly sampling 6120 from each feature data) yielded a poor accuracy below 0.7 on the external test set. Contrarily, training a separate model for each feature (then using FP2 to FP5, **Methods**) improved the accuracy on the external test set although the models clearly overfitted the training sets (**Table 5.2**).

**Table 5.2.** Accuracy of pharmacophoric points predictions.

| Feature[a] | Dataset size[b] (# unique PDB entries) | ACC ± std (5-fold CV) | | ACC |
|---|---|---|---|---|
| | | Training (60 %) | Test (15 %) | Ext. test (25 %) |
| CA | 254 416 (35 734) | 1 ± 0 | 0.742 ± 0.003 | 0.740 |
| CZ | 14 856 (9124) | 1 ± 0 | 0.781 ± 0.007 | 0.797 |
| N | 103 126 (30 550) | 1 ± 0 | 0.733 ± 0.004 | 0.729 |
| NZ | 6120 (4312) | 1 ± 0 | 0.784 ± 0.007 | 0.782 |
| O | 158 094 (33 011) | 1 ± 0 | 0.734 ± 0.003 | 0.735 |
| OD1 | 57 070 (19 098) | 1 ± 0 | 0.766 ± 0.006 | 0.769 |
| OG | 69 346 (25 675) | 1 ± 0 | 0.694 ± 0.004 | 0.699 |

[a] VolSite pharmacophoric features: CA: hydrophobic, CZ: aromatic, N: h-bond donor, NZ: positive, O: h-bond acceptor, OD1: negative, OG: h-bond donor or acceptor.

[b] The number of points (positive and negative classes).

In contrast to the RF models, the baseline models obtained from the Bayesian classifier yielded lower accuracy values (0.63) on the external test set. To verify the relevance of the predictions, randomly shuffling the content of the descriptors and of the classes in the training set respectively led to an accuracy of 0.5 on the external test set. Finally, the obtained models were applied to 1000 VolSite 'cavityALL' cavities from the application set. For each cavity, the seven RF models were applied, and points predicted to be *important* were saved into a new cavity file. On average, more than two third of the cavities' points were trimmed independently of their original frequency (**Figure 5.6A**). Analysis of the true positive rates showed that the few positive points are often kept (few loss) while improvements are to be made on removing more negative points (**Figure 5.6B-C**). Still, the observed accuracy values were encouraging. We pay careful attention to these metrics as negative points clearly outnumbered positive points.
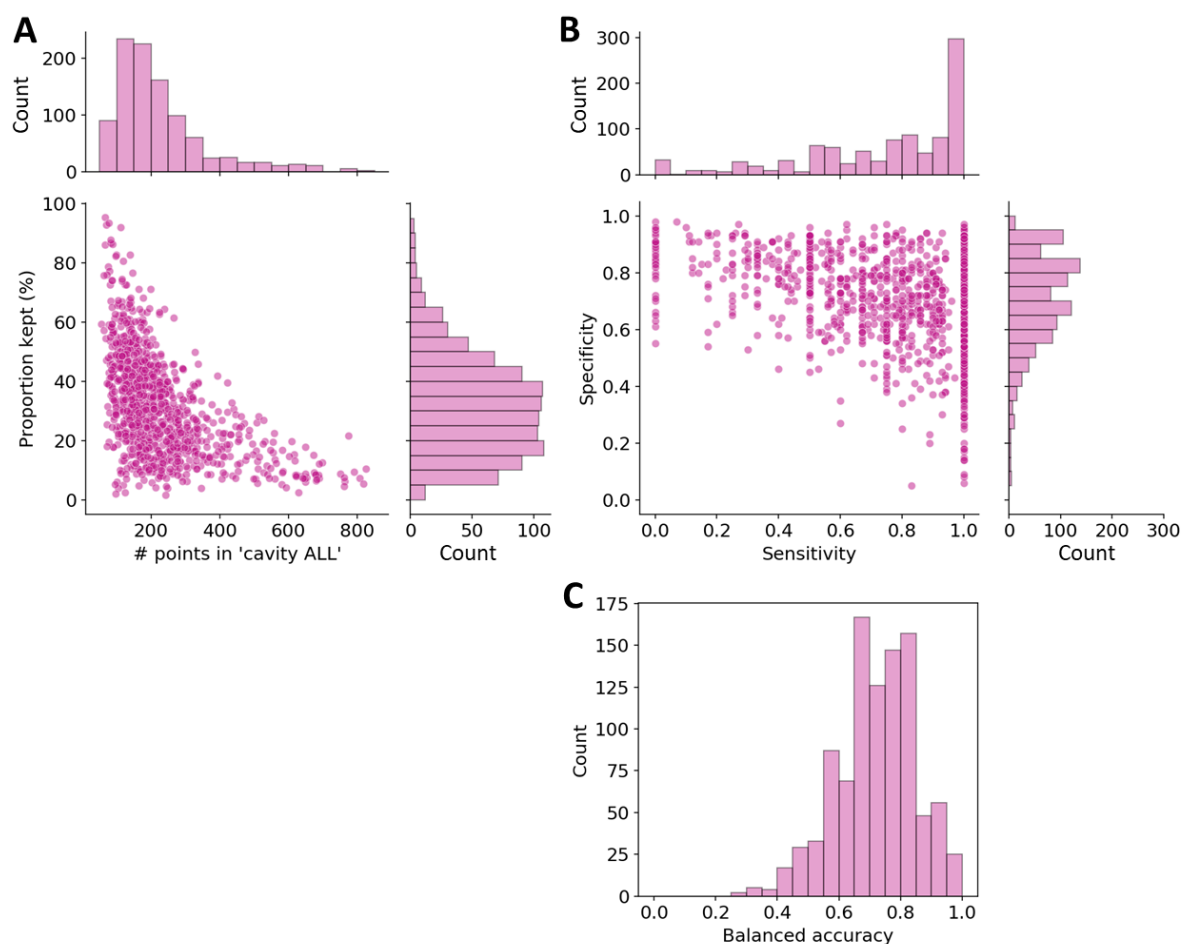
**Figure 5.6.** Statistics of predicting interacting points in VolSite cavities from the application set. A) Proportion of points kept (predicted as *important*) with respect to the number of points in the cavity. B) Specificity (true negative rate, eq 5.8) of the predictions with respect to the sensitivity (true positive rate, eq 5.7). C) Balanced accuracy (eq. 5.9).

Examples of predictions on three different proteins are shown in **Figure 5.7.** The first two examples illustrate cases where the models restricted the cavity points to fit the X-ray ligand. In the last example, the *important* points were not correctly defined (at least according to that ligand).
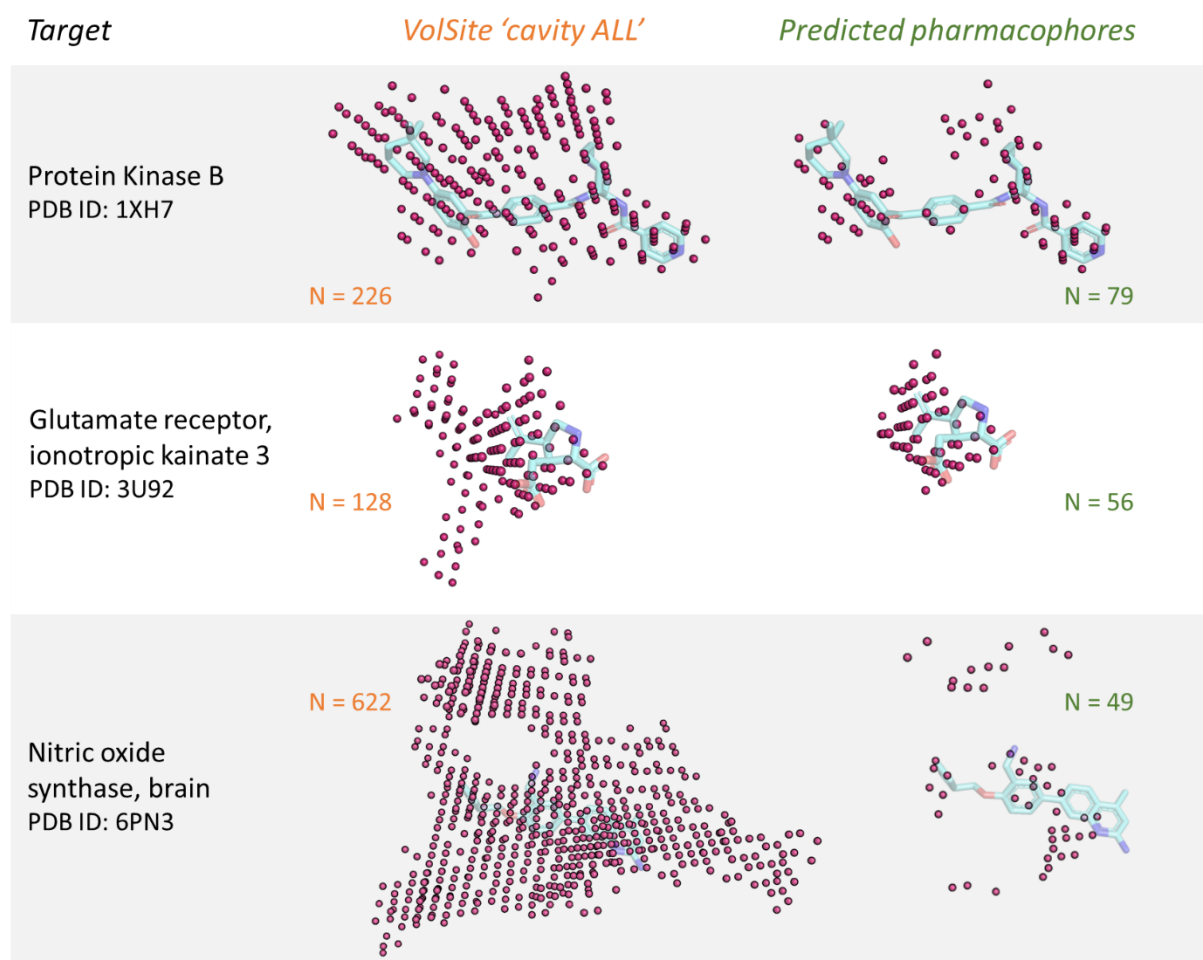
**Figure 5.7.** Prediction of *important* points in VolSite cavities from the application set. Seven RF machine learning models were trained and applied to classify interacting (kept) and non-interacting (removed) points, taking as input the VolSite cavitied detected on the apo structures (clouds on the left) and yielding a pruned cavity (cloud on the right). The masked X-ray ligand is illustrated in the background (transparent blue).

In summary, the results presented herein were the first steps towards the development of a machine learning model to discriminate between interacting and non-interacting cavity points. These initial results are encouraging to pursue a thorough study. Due to the bias in the PDB towards certain protein cavities (e.g., Adenine-binding, phosphate sites), the predictive models might achieve better results on related cavities (e.g. protein kinases, ATP sites). The data splitting should account for the distribution of the protein families instead of the PDB IDs. Other splitting scenarios are possible (e.g., time-split). Different baseline models will be implemented for comparison, while assessing the sensitivity and precision of the predictions. Finally, the applicability of the models should be assessed on proteins in complex with different congeneric ligands that might exhibit different binding modes, as well as new target structures.

296

## 5.4. References

1. Rognan, D. The Impact of in Silico Screening in the Discovery of Novel and Safer Drug Candidates. *Pharmacol. Ther.* **2017**, *175*, 47–66.

2. Schindler, C. E. M. M.; Baumann, H.; Blum, A.; Böse, D.; Buchstaller, H.-P.; Burgdorf, L.; Cappel, D.; Chekler, E.; Czodrowski, P.; Dorsch, D.; Eguida, M. K. I.; Follows, B.; Fuchß, T.; Grädler, U.; Gunera, J.; Johnson, T.; Jorand Lebrun, C.; Karra, S.; Klein, M.; Knehans, T.; Koetzner, L.; Krier, M.; Leiendecker, M.; Leuthner, B.; Li, L.; Mochalkin, I.; Musil, D.; Neagu, C.; Rippmann, F.; Schiemann, K.; Schulz, R.; Steinbrecher, T.; Tanzer, E.; Unzue Lopez, A.; Viacava Follis, A.; Wegener, A.; Kuhn, D. Large-Scale Assessment of Binding Free Energy Calculations in Active Drug Discovery Projects. *J. Chem. Inf. Model.* **2020**, *60*, 5457–5474.

3. Carlson, H. A.; Masukawa, K. M.; Rubins, K.; Bushman, F. D.; Jorgensen, W. L.; Lins, R. D.; Briggs, J. M.; McCammon, J. A. Developing a Dynamic Pharmacophore Model for HIV-1 Integrase. *J. Med. Chem.* **2000**, *43*, 2100–2114.

4. Baroni, M.; Cruciani, G.; Sciabola, S.; Perruccio, F.; Mason, J. S. A Common Reference Framework for Analyzing/Comparing Proteins and Ligands. Fingerprints for Ligands and Proteins (FLAP): Theory and Application. *J. Chem. Inf. Model.* **2007**, *47*, 279–294.

5. Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and Scoring in Virtual Screening for Drug Discovery: Methods and Applications. *Nat. Rev. Drug Discov.* **2004**, *3*, 935–949.

6. Sanders, M. P. A.; McGuire, R.; Roumen, L.; de Esch, I. J. P.; de Vlieg, J.; Klomp, J. P. G.; de Graaf, C. From the Protein's Perspective: The Benefits and Challenges of Protein Structure-Based Pharmacophore Modeling. *Medchemcomm* **2012**, *3*, 28–38.

7. Wermuth, C. G.; Ganellin, C. R.; Lindberg, P.; Mitscher, L. A. Glossary of Terms Used in Medicinal Chemistry (IUPAC Recommendations 1998). *Pure Appl. Chem.* **1998**, *70*, 1129–1143.

8. Yang, S.-Y. Pharmacophore Modeling and Applications in Drug Discovery: Challenges and Recent Advances. *Drug Discov. Today* **2010**, *15*, 444–450.

9. Schaller, D.; Šribar, D.; Noonan, T.; Deng, L.; Nguyen, T. N.; Pach, S.; Machalz, D.; Bermudez, M.; Wolber, G. Next Generation 3D Pharmacophore Modeling. *WIREs Comput. Mol. Sci.* **2020**, *10*, 1–20.

10. Goodford, P. J. A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules. *J. Med. Chem.* **1985**, *28*, 849–857.

11. Verdonk, M. L.; Cole, J. C.; Taylor, R. SuperStar: A Knowledge-Based Approach for Identifying Interaction Sites in Proteins. *J. Mol. Biol.* **1999**, *289*, 1093–1108.

12. Desaphy, J.; Azdimousa, K.; Kellenberger, E.; Rognan, D. Comparison and Druggability Prediction of Protein–Ligand Binding Sites from Pharmacophore-Annotated Cavity Shapes. *J. Chem. Inf. Model.* **2012**, *52*, 2287–2299.

13. Brenke, R.; Kozakov, D.; Chuang, G.-Y.; Beglov, D.; Hall, D.; Landon, M. R.; Mattos, C.; Vajda, S. Fragment-Based Identification of Druggable 'Hot Spots' of Proteins Using Fourier Domain Correlation Techniques. *Bioinformatics* **2009**, *25*, 621–627.

14. Radoux, C. J.; Olsson, T. S. G.; Pitt, W. R.; Groom, C. R.; Blundell, T. L. Identifying Interactions That Determine Fragment Binding at Protein Hotspots. *J. Med. Chem.* **2016**, *59*, 4314–4325.

15. Barillari, C.; Marcou, G.; Rognan, D. Hot-Spots-Guided Receptor-Based Pharmacophores (HS-Pharm): A Knowledge-Based Approach to Identify Ligand-Anchoring Atoms in Protein Cavities and Prioritize Structure-Based Pharmacophores. *J. Chem. Inf. Model.* **2008**, *48*, 1396–1410.

16. Sanders, M. P. A.; Verhoeven, S.; de Graaf, C.; Roumen, L.; Vroling, B.; Nabuurs, S. B.; de Vlieg, J.; Klomp, J. P. G. Snooker: A Structure-Based Pharmacophore Generation Tool Applied

to Class A GPCRs. *J. Chem. Inf. Model.* **2011**, *51*, 2277–2292.

17. Schuetz, D. A.; Seidel, T.; Garon, A.; Martini, R.; Körbel, M.; Ecker, G. F.; Langer, T. GRAIL: GRids of PhArmacophore Interaction FieLds. *J. Chem. Theory Comput.* **2018**, *14*, 4958–4970.

18. Miranker, A.; Karplus, M. Functionality Maps of Binding Sites: A Multiple Copy Simultaneous Search Method. *Proteins Struct. Funct. Genet.* **1991**, *11*, 29–34.

19. Yu, W.; Lakkaraju, S. K.; Raman, E. P.; Fang, L.; MacKerell, A. D. Pharmacophore Modeling Using Site-Identification by Ligand Competitive Saturation (SILCS) with Multiple Probe Molecules. *J. Chem. Inf. Model.* **2015**, *55*, 407–420.

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

21. Dixon, S. L.; Smondyrev, A. M.; Knoll, E. H.; Rao, S. N.; Shaw, D. E.; Friesner, R. A. PHASE: A New Engine for Pharmacophore Perception, 3D QSAR Model Development, and 3D Database Screening: 1. Methodology and Preliminary Results. *J. Comput. Aided. Mol. Des.* **2006**, *20*, 647–671.

22. Tran-Nguyen, V.-K. K.; Da Silva, F.; Bret, G.; Rognan, D. All in One: Cavity Detection, Druggability Estimate, Cavity-Based Pharmacophore Perception, and Virtual Screening. *J. Chem. Inf. Model.* **2019**, *59*, 573–585.

23. Wolber, G.; Seidel, T.; Bendix, F.; Langer, T. Molecule-Pharmacophore Superpositioning and Pattern Matching in Computational Drug Design. *Drug Discov. Today* **2008**, *13*, 23–29.

24. Barnum, D.; Greene, J.; Smellie, A.; Sprague, P. Identification of Common Functional Configurations Among Molecules. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 563–571.

25. Halgren, T. a. Merck Molecular Force Field. *J. Comput. Chem.* **1996**, *17*, 490–519.

26. Eguida, M.; Rognan, D. A Computer Vision Approach to Align and Compare Protein Cavities: Application to Fragment-Based Drug Design. *J. Med. Chem.* **2020**, *63*, 7127–7142.

27. Desaphy, J.; Bret, G.; Rognan, D.; Kellenberger, E. Sc-PDB: A 3D-Database of Ligandable Binding Sites-10 Years On. *Nucleic Acids Res.* **2015**, *43*, D399–D404.

28. Desaphy, J.; Raimbaud, E.; Ducrot, P.; Rognan, D. Encoding Protein–Ligand Interaction Patterns in Fingerprints and Graphs. *J. Chem. Inf. Model.* **2013**, *53*, 623–637.

29. Da Silva, F.; Desaphy, J.; Rognan, D. IChem: A Versatile Toolkit for Detecting, Comparing, and Predicting Protein–Ligand Interactions. *ChemMedChem* **2018**, *13*, 507–510.

30. Wood, D. J.; Vlieg, J. De; Wagener, M.; Ritschel, T. Pharmacophore Fingerprint-Based Approach to Binding Site Subpocket Similarity and Its Application to Bioisostere Replacement. *J. Chem. Inf. Model.* **2012**, *52*, 2031–2043.

31. Bron, C.; Kerbosch, J. Algorithm 457: Finding All Cliques of an Undirected Graph. *Commun. ACM* **1973**, *16*, 575–577.

32. Kabsch, W. A Discussion of the Solution for the Best Rotation to Relate Two Sets of Vectors. *Acta Crystallogr. Sect. A* **1978**, *34*, 827–828.

33. Jones, E.; Oliphant, T.; Peterson, P.; others. SciPy: Open Source Scientific Tools for Python.

34. Rusu, R. B.; Blodow, N.; Beetz, M. Fast Point Feature Histograms (FPFH) for 3D Registration. *2009 IEEE Int. Conf. Robot. Autom.* **2009**, 3212–3217.

35. Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D. Comparative Evaluation of Eight Docking Tools for Docking and Virtual Screening Accuracy. *Proteins Struct. Funct. Bioinforma.* **2004**, *57*, 225–242.

36. Cross, J. B.; Thompson, D. C.; Rai, B. K.; Baber, J. C.; Fan, K. Y.; Hu, Y.; Humblet, C. Comparison of Several Molecular Docking Programs: Pose Prediction and Virtual Screening Accuracy. *J. Chem. Inf. Model.* **2009**, *49*, 1455–1474.

# GENERAL CONCLUSIONS

# General conclusions

This thesis has proposed novel computational approaches for molecular design, by exploiting available protein cavities represented as clouds of points. Starting from the idea to investigate the application of image recognition approaches to compare protein cavities represented as point clouds, the projects were progressively built to tackle several problems: (1) estimation of protein cavities similarity at the structural proteome scale and their prospective applications to (2) secondary target prediction and (3) target-focused library design, (4) comparison of ligands to protein cavities, (5) prediction of *interacting* cavity points (**Figure 6.1**).
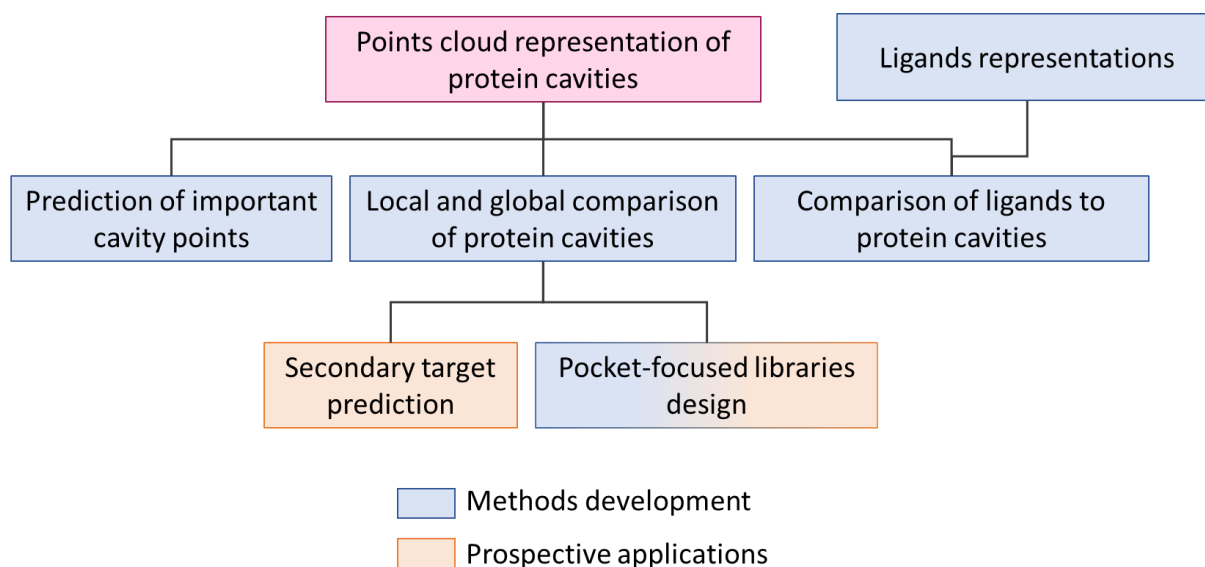


**Figure 6.1.** Computational strategies based on point cloud processing implemented in this thesis and their prospective applications.

Literature review of state-of-the-art methods revealed the intricacies of estimating the similarity between protein cavities and the need for methods enabling subpockets comparison. By developing ProCare to this end, we showed that sampling-based point cloud registration, originally applied to other computer vision tasks can identify common motifs between subpockets of unrelated proteins. From the initial retrospective validations, we went on to evaluating our method by confronting the computational predictions to experimental validations. As a result, the similarity between the binding sites of two functionally and structurally unrelated targets, the cytokine tumor necrosis factor-alpha (TNF-α) and the HIV-1 reverse transcriptase (RT) could be identified for the first time. Direct *in vitro* binding measurement showed two HIV-1 non-nucleoside inhibitors interacting with TNF-α trimer with an affinity comparable to a high-throughput screening hit. Moreover, we developed a workflow, POEM, to design a focused library of small molecules based on subpocket similarity prediction. Cognate fragments

of the most similar subpockets were used as building blocks and linked to generate fully connected molecules. By applying POEM to the cyclin-dependent kinase 8 (CDK8), we successfully designed a new nanomolar ligand in just two rounds. Finally, the application of POEM to orphan targets (quinolinate synthase, WD40 repeats domain of leucine-rich repeat kinase 2), for which no pharmacological ligand is known to this date enables to improve the workflow while providing a fully blind challenge to delineate limitations regarding the fragments' selection. The biological assays of the predicted compounds are ongoing. The representation of the protein cavities as clouds of points occupying the entire ligand space can be explored to develop computational methods for small molecules screening. In this perspective, we studied point cloud registration and graph matching of ligands to protein cavities. Although ligands pharmacophoric points alignment to protein cavities is a difficult task since structurally different objects are being compared, the information contained in the cavity clouds proved to be rich for comparison to small molecules and supported the investigation of machine learning models to predict important cavity points corresponding to pharmacophores in the ligands. Some of these preliminary results were encouraging and have suggested further analyses to investigate these research questions, and have opened the perspective for other target classes.

The volumetric point-cloud representation of the protein pockets presented advantages and drawbacks. By working around the latter, we showed a variety of applications of subpocket clouds comparison in drug design under the constraint of experimental and collaborative resources available. We would have liked to pursue some questions that arose from the results presented herein, even if they fall out of the scope of this thesis. Finally, feedback from more prospective applications would be beneficial to improve the implementations according to and beyond what has been already discussed in this thesis.

To conclude on the scientific level, we hope that the novel contributions of this thesis to the state-of-the-art have provided useful insights as part of the general pursuit of computational drug design. The different evaluations provided herein have suggested improvements and new research ideas, that will be investigated by future work in our lab.

To conclude on the personal level, this thesis allowed me to learn at different levels: the process of scientific research, from the identification of questions to the investigation and communication of results in different formats, the flexibility to adjust to mishaps, collaborative multidisciplinary work, teaching, supervision, gaining knowledge of concepts in related fields (computer science, geometry, medicinal chemistry), while I was venturing out of my comfort zone as a dominantly trained biologist. The exchange of scientific reflections with colleagues and my advisor have always filled me with wonder. This experience came with its ups and downs; even so, I found that science is exciting and applies to everyday life. I also had the chance to be involved in non-research activities such as representing my fellow PhD students in our Doctoral School and lab committees, and volunteering in the ADDAL PhD association, while helping with solving problems and developing important transversal skills.

x

**Université de Strasbourg**

# Kossiwa Ikafui Merveille EGUIDA

**EDSC**
École Doctorale des
Sciences Chimiques

# Comparaison de cavités protéiques par traitement numérique de nuages de points : principes et applications en drug design

## Résumé

Les cavités de protéines sont au cœur d'interactions moléculaires nécessaires aux fonctions biologiques du vivant. Grâce à l'augmentation incessante des données structurales, les méthodes de comparaison de cavités protéiques offrent diverses applications en conception de molécules bioactives mais doivent relever plusieurs défis.

Cette thèse propose de nouveaux algorithmes basés sur le traitement d'images tridimensionnelles pour comparer les motifs globaux et locaux de (sous-) cavités protéiques, représentées en nuages de points. Leurs applications concrètes, validées par des essais biologiques *in vitro*, illustrent leurs utilisations pour prédire des cibles secondaires à l'échelle du protéome structural et pour générer des chimiothèques focalisées permettant d'augmenter le taux de touches en criblage virtuel. A partir de la caractérisation des cavités, l'élaboration de pharmacophores et le développement de méthodes de criblage virtuel ont été investigués.

Mots-Clés : comparaison de sites de protéines, nuage de points, alignement 3D, prédiction de cible secondaire, chimiothèque focalisée, criblage virtuel, pharmacophore, alignement de graphe, intelligence artificielle, conception de molécules bioactives, structure, Chémoinformatique.

## Résumé en anglais

Protein cavities are the heart of molecular interactions that trigger and regulate biological processes in living organisms. Supported by the constant augmentation of characterized pockets in three-dimensional protein structures, methods to assess the similarity between protein cavities have multiple applications in drug design but face many challenges.

This thesis proposes new algorithms based on three-dimensional (3D) image processing to compare global and subtle patterns in different protein (sub-) pockets represented by point clouds. Through prospective applications validated by *in vitro* biological experiments, we showed how these methods can predict a secondary target at the proteome scale and design a target-focused library for faster small molecule hit identification. In the next stages, better characterization of the cavities for pharmacophore elaboration and the development of virtual screening methods were investigated.

Keywords: protein subpocket comparison, point cloud, 3D alignment, secondary target prediction, focused library, virtual screening, pharmacophore, graph matching, machine learning, drug design, structure-based, Cheminformatics.