



HAL
open science

Development of new virtual screening methods and novel unbiased benchmarking data sets

Viet Khoa Tran Nguyen

► **To cite this version:**

Viet Khoa Tran Nguyen. Development of new virtual screening methods and novel unbiased benchmarking data sets. Cheminformatics. Université de Strasbourg, 2020. English. NNT : 2020STRAF013 . tel-03504176

HAL Id: tel-03504176

<https://theses.hal.science/tel-03504176>

Submitted on 28 Dec 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ÉCOLE DOCTORALE DES SCIENCES CHIMIQUES
Laboratoire d'Innovation Thérapeutique (UMR 7200)

THÈSE présentée par :
Viet Khoa TRAN NGUYEN

soutenue le : 17 septembre 2020

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

Discipline/ Spécialité : Chimie théorique, chimie informatique

**Développement de Jeux de Données Non Biaisés
et de Nouvelles Méthodes de Criblage Virtuel**

THÈSE dirigée par :

M. ROGNAN Didier

Directeur de recherche, Université de Strasbourg

RAPPORTEURS :

M. MONTES Matthieu

Professeur, CNAM

Mme. DOUGUET Dominique

Chargée de recherche, INSERM

AUTRE MEMBRE DU JURY :

Mme. KELLENBERGER Esther

Professeure, Université de Strasbourg

UNIVERSITE DE STRASBOURG
ECOLE DOCTORALE DE SCIENCES CHIMIQUES (ED 222)
LABORATOIRE D'INNOVATION THERAPEUTIQUE (UMR 7200)

Thèse présentée par

Viet Khoa TRAN NGUYEN

soutenue le 17 septembre 2020

Titre en français

**DEVELOPPEMENT DE JEUX DE DONNEES NON BIAISES
ET DE NOUVELLES METHODES DE CRIBLAGE VIRTUEL**

Titre en anglais

**DEVELOPMENT OF NEW VIRTUAL SCREENING METHODS
AND NOVEL UNBIASED BENCHMARKING DATA SETS**

Table of Contents

Table of Contents	1
Acknowledgment	3
Thesis Summary in French	5
Introduction	25
Chapter 1. Benchmarking Data Sets from PubChem BioAssay Data: Current Scenario and Room for Improvement	29
Introduction.....	31
PubChem BioAssay Statistics: Assays and Compounds	32
What We Can Do with PubChem BioAssay Data: from the Data Set Construction Point of View	35
Note-Worthy Issues with Using Data from PubChem BioAssay for Constructing Benchmarking Data Sets.....	43
Conclusion	55
References.....	56
Supporting Information	63
Take-home Messages.....	88
Chapter 2. All in One: Cavity Detection, Druggability Estimate, Cavity-Based Pharmacophore Perception, and Virtual Screening	89
Introduction.....	91
Computational Methods.....	93
Results and Discussion	102
Conclusions.....	117
References.....	118
Supporting Information	122
Take-home Messages.....	134

Chapter 3. LIT-PCBA: An Unbiased Data Set for Machine Learning and Virtual Screening	135
Introduction.....	137
Computational Methods.....	140
Results and Discussion	144
Conclusions.....	157
References.....	157
Supporting Information	161
Take-home Messages.....	172
Chapter 4. Rescoring LIT-PCBA Docking Poses with Interaction-Based Scoring Functions	173
Introduction.....	175
Computational Methods.....	176
Results and Discussion	177
Conclusions.....	183
References.....	184
Supporting Information	186
Overall Conclusions	191

Acknowledgment

I would like to thank, first and foremost, Dr. Didier Rognan, my supervisor, for accepting me as one of his Ph.D. students, for his guidance, patience and great support. Despite being the head of a department with so many responsibilities, he still has time to discuss the projects with me on a regular basis and gives me a hand whenever I need, while granting me enough autonomy to learn how to become an independent researcher. The goal of “transforming the mind of a student into that of a scientist” that he has been adhering to will also be the principle that I embrace once I become a teacher in the future. I have learnt so much from him during the last three years, and am truly grateful for all the opportunities I have been given.

I would also like to express my sincere gratitude to Pr. Esther Kellenberger, Dr. Franck Da Silva, and Mr. Guillaume Bret for their help and support. The previous work of Franck sets the basis for the first part of my projects, and it was a pleasure cooperating with him during my first few months in the lab. Guillaume is the computer genius who I always turn to whenever I have a problem with the machines. Esther is probably the first member of the group that I met at the doctoral recruitment contest three years ago, and, along with Dr. Gilles Marcou, has shared with me lots of experiences in teaching in general, and in giving lectures in cheminformatics in particular. As a young teaching assistant who had not had many pedagogical skills, I have learnt so much from their advice and truly love working with them. Their comments to my Ph.D. work, notably at the mid-thesis defense, have greatly contributed to the quality of my final dissertation.

My appreciation also goes to all of my past and present labmates: Priscila, Florian, Célien, Merveille, Mikhail, Jenke, Joel, Xuechen, and Julia. Being on the same team with them fills me with joy, and I cherish each and every moment that we have had together, both inside and outside the lab. I thank them for the help and support they have been sending, and would love to have a chance to work with them in the near future. Also, other members of the UMR 7200, as well as those from other departments of the Strasbourg Faculty of Pharmacy, who I have the pleasure to meet and talk to, are also appreciated for their kindness and friendship.

I would also like to take this opportunity to thank my past professors, both in Vietnam and in France, for their dedication, their great support, and for the inspiration that they have been bringing. Thanks to my Vietnamese professors, I had the very first thought of pursuing a career

in academia as a lecturer-researcher at a very young age and have been so determined to keep nurturing my dream ever since. Had it not been for my former professors in Grenoble, Pr. Ahcène Boumendjel and Dr. Edwige Nicolle, I would not have won the “golden ticket” to go to France four years ago and would have never had the first great year in the lovely city by the Isère River that I always keep in my heart and will never forget. I would like to thank them for their encouragement, their advice and for the opportunities they have offered even after my departure for Strasbourg. I do hope that we will still keep in touch, and have many other chances to collaborate for many years to come.

Besides, my friends in Vietnam, in France and in many other countries around the world (Japan, the Netherlands, Sweden, Germany, the USA, Canada, etc.), either staying by my side physically or being with me emotionally, are warmly appreciated for their friendship and support. What a life that we have had! Toast to all the good memories, the smiles, the laughter, and also the sadness that have led us to where we are now. May the great moments that we have shared stay forever in our hearts, may the good times that we have had always be treasured, and may our friendship remain fulfilling throughout the coming years, despite the merciless wheel of time that keeps rolling.




Another special thank goes to the French Embassy in Vietnam and the Foundation of the Grenoble Alpes University for their financial aid during the pursuit of my Master’s degree; and to the Doctoral School of Chemical Sciences (ED 222), University of Strasbourg, for the three-year Ph.D. grant distributed to me in 2017. I would also like to thank Dr. Matthieu Montes and Dr. Dominique Douguet for having agreed to review my dissertation and to become jury members of my thesis defense.

Last but not least, I want to thank my family and my relatives in Vietnam for everything they have done, without whom I would not become the person that I am right now. Hard as it is to foretell the future, there is, still, one thing that I am certain about, it is that I will keep going the extra mile every single day to continue improving myself, and make me a better person that I can be proud of.

Love and respect.

Thesis Summary in French

The following Thesis summary in French (Résumé de thèse en français) has been validated at the thesis committee meeting organized by the Doctoral School of Chemical Sciences (Ecole Doctorale des Sciences Chimiques – ED 222), University of Strasbourg, on June 18, 2020.

		
UNIVERSITE DE STRASBOURG		
ECOLE DOCTORALE DES SCIENCES CHIMIQUES		
RESUME DE LA THESE DE DOCTORAT		
Discipline : Chimie		
Spécialité : Chimie théorique, chimie informatique		
Présentée par : TRAN NGUYEN Viet Khoa		
Titre :		
DEVELOPPEMENT DE JEUX DE DONNEES NON BIAISES ET DE NOUVELLES METHODES DE CRIBLAGE VIRTUEL		
Unité de Recherche : UMR 7200 – Laboratoire d’Innovation thérapeutique		
Directeur de Thèse : Dr. ROGNAN Didier – Directeur de recherche, CNRS		
Localisation : Faculté de Pharmacie, Université de Strasbourg		
Thèse confidentielle : <input checked="" type="checkbox"/> NON <input type="checkbox"/> OUI		

1. Introduction

Découvrir les premiers ligands pour une protéine cible, de manière rapide et économique, est un enjeu important en “drug design”. En absence d’un ligand d’une protéine dont la structure tridimensionnelle est déjà connue, le docking (amarrage moléculaire) est en général utilisé comme outil de criblage virtuel, ceci malgré un problème toujours non-résolu de prédiction quantitative des affinités de liaison des touches potentielles. Il nécessite donc de concevoir une nouvelle approche computationnelle qui peut être appliquée aux apo-protéines. En 2012, les chercheurs du Laboratoire d’Innovation Thérapeutique (Université de Strasbourg) sont arrivés à mettre en oeuvre une nouvelle méthode *in silico* qui a déjà été intégrée au logiciel IChem [1,2]. Il s’agit de la génération des pharmacophores déduits des poches de liaison potentielles à la surface d’une protéine cible. La méthode nous permet de détecter automatiquement toutes les cavités à la surface d’une protéine donnée, puis prédire la droguabilité de chaque cavité, et créer un pharmacophore pour chaque site considéré comme potentiellement droguable. Une vingtaine d’éléments pharmacophoriques “structure-based” qui représentent chaque cavité qu’on étudie sont retenus. A ce stade, il nous reste à élaborer une stratégie d’utilisation de ces pharmacophores pour faire du criblage virtuel de chimiothèques, afin de sélectionner de manière rationnelle les touches potentielles pour une protéine d’intérêt pharmaceutique, même à défaut de ligand co-cristallisé.

Lorsqu’une nouvelle méthode de criblage *in silico* est développée, il faut évaluer la performance de cette méthode pour voir si elle arrive à choisir les vraies touches d’une

cible biologique, de manière rétrospective, à partir d'une banque de molécules. Ceci est fait en utilisant les données déjà existantes, soit dans la littérature, soit dans les bases de données ouvertes au public. Cependant, de nombreux problèmes avec les jeux de données actuellement utilisés dans la communauté de chimoinformatique, tels que DUD, DUD-E, ChEMBL, ou MUV, ont été observés et avertis [3-6]. Plus précisément, il y a des biais dans la composition des actifs et des "decoys", par exemple: la puissance des "decoys" n'est pas encore connue et vérifiée par les tests biologiques, le nombre des actifs est trop élevé, et les actifs ressemblent trop à des molécules de référence. Ces jeux de données ne décrivent pas la vraie vie, car ils n'imitent pas les données utilisées au criblage à haut débit en réalité, et ils surestiment la précision des méthodes de criblage *in silico*. Il nécessite donc de concevoir un nouveau jeu de données non-biaisé qui est dédié à des méthodes de criblage virtuel "structure-based" ainsi que "ligand-based", qui a un niveau de difficulté similaire à celui des chimiothèques utilisées au criblage à haut débit, et qui est capable de capturer les différences entre les performances de différentes méthodes.

Devant les problèmes expliqués ci-dessus, mon travail de thèse se compose en deux parties principales. La première partie concerne le développement d'une procédure d'alignement de petits ligands sur les pharmacophores "structure-based" déjà générés, avant de choisir une meilleure pose pour chaque ligand, et de classer les ligands selon un certain paramètre. Une fois élaboré, ce protocole pourrait être utilisé pour prédire la pose d'un composé actif pour une protéine cible, et distinguer entre les vrais actifs et les "decoys" (qui sont chimiquement similaires à des vrais actifs), ou entre les vrais actifs et les vrais inactifs d'une cible d'intérêt pharmaceutique. La deuxième partie se

focalise sur la conception des jeux de données “PubChem BioAssay” représentant une diversité de protéines cibles dont les biais dans la composition des composés actifs et inactifs sont réduits, et l'évaluation de ces jeux de données après préparation pour voir s'il y a encore des biais ou pas.

2. Résultats et discussions

2.1. Développement de méthodes de criblage virtuel basées sur les pharmacophores déduits des poches de liaison potentielles à la surface d'une protéine cible

L'élaboration et l'évaluation de protocoles de criblage virtuel se font en utilisant les jeux de données “Astex”, “DUD-E” et “PubChem BioAssay” au long des trois challenges : un challenge de positionnement de ligand et deux challenges de criblage virtuel rétrospectif.

Le jeu de données “Astex” est utilisé pour le challenge de positionnement de ligand. Il se compose de 85 complexes, chacun est une structure d'une protéine avec un ligand co-cristallisé en 3D [7]. Une totalité de 17.555 conformères ont été créés pour toutes les entrées. Les conformères ont été ensuite alignés sur les éléments pharmacophoriques générés par le programme VolSite, puis scorés par le programme Shaper2 développé au laboratoire ; les poses ayant été préalablement optimisées en présence de la protéine avec SZYBKI 1.8.0.1 [8]. Une seule pose a été retenue pour chaque ligand, selon quelques critères. Il est observé que si l'on sélectionne la pose avec la meilleure énergie d'interaction ligand-protéine MMFF94 ou celle avec la

meilleure énergie totale MMFF94 pour chaque ligand, les valeurs moyennes d'écart quadratique moyen RMSD ("root-mean-square deviation") à la pose cristallographique sont les meilleures: 2,221 Å et 2,232 Å, respectivement, indiquant une très bonne performance qui est même meilleure que celle qu'on a eue auparavant avec le docking moléculaire en utilisant Surflex-Dock (RMSD = 2,575 Å) [9]. Le nombre des entrées qui ont donné une RMSD < 1 Å avec notre méthode est plus élevé que celui obtenu avec le docking [9] (**Figure 1**). Il est clair donc que les deux critères ci-dessus sont les meilleurs pour la sélection de pose. On a également comparé la performance de notre méthode avec celles de LigandScout et de Discovery Studio, en utilisant toujours les mêmes éléments pharmacophoriques issus d'IChem comme input. Les deux programmes ne sont pas arrivés à positionner correctement les ligands dans quasiment 90% des cas étudiés (RMSD > 4 Å). En tenant compte du fait que notre méthode d'alignement marche très bien avec les mêmes pharmacophores, il est certain que ce sont les méthodes d'alignement de LigandScout et de Discovery Studio qui échouent, et que la qualité de nos pharmacophores "structure-based" n'est pas coupable de cet échec.

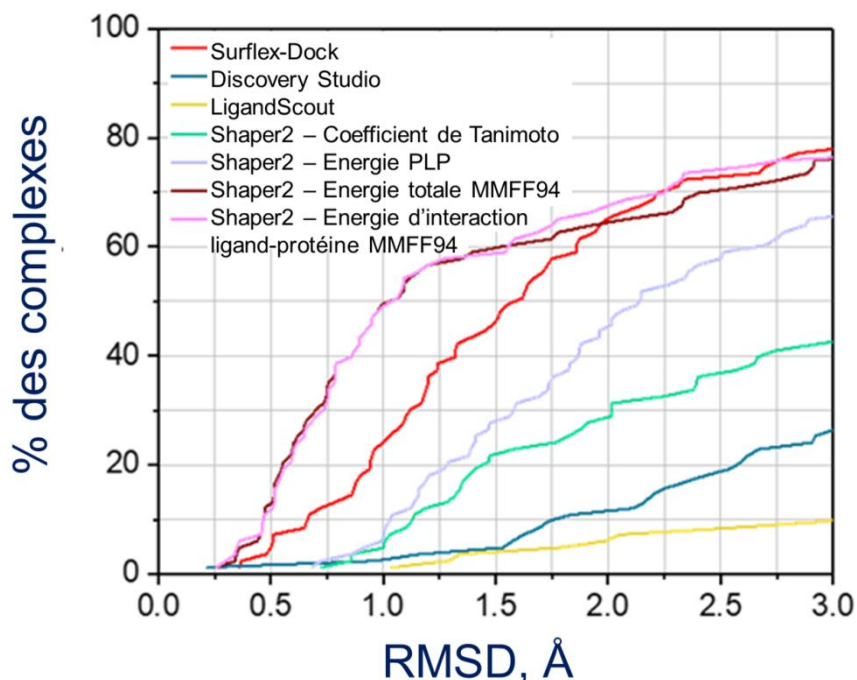


Figure 1. Performance de différentes méthodes de criblage virtuel dans la prédiction de poses des 85 ligands du jeu de données “Astex”, illustrée par le pourcentage cumulé des entrées en fonction de la valeur RMSD (en Å) de la meilleure pose de chaque ligand par rapport à la pose cristallographique correspondante.

Le jeu de données “DUD-E” est ensuite utilisé pour le premier challenge de criblage virtuel rétrospectif. Pour cette étude, on a choisi 10 entrées de protéines cibles d’intérêt pharmaceutique, y compris deux RCPGs, deux récepteurs d’hormones nucléaires, deux protéases, deux kinases, et deux autres enzymes [10]. Pour chaque entrée, il y a une protéine de structure cristallographique connue, un ligand co-cristallisé, les vrais actifs et les “decoys” qui sont chimiquement similaires à des vrais actifs. Après avoir sélectionné une meilleure pose pour chaque composé, soit avec l’énergie d’interaction ligand-protéine MMFF94, soit avec l’énergie totale MMFF94, et avoir classé les composés selon le même critère, on a observé que les valeurs moyennes de ROC AUC

qu'on a eues ont été toutes inférieures à 0,65, ce qui n'est pas suffisamment bon comme résultat. La meilleure performance a été obtenue lorsqu'on a classé les composés selon l'énergie PLP après une sélection de meilleure pose selon l'énergie totale MMFF94 : ROC AUC moyenne = 0,68, deux entrées ont donné une excellente performance (ADRB2, RENI : ROC AUC > 0,8), deux entrées ont donné une bonne performance (AKT1, FGFR1 : $0,7 < \text{ROC AUC} < 0,8$). Cette performance est assez intéressante et comparable à celle qu'on a obtenue auparavant avec le docking moléculaire en utilisant Surflex-Dock [9].

Pour le deuxième challenge de criblage virtuel rétrospectif, on a choisi les jeux de données "PubChem BioAssay", qui nous fournit les vrais actifs et les vrais inactifs de chaque protéine cible qui ont déjà été vérifiés par les essais biologiques confirmatoires. Plus précisément, les trois jeux de données suivants : ROCK2 (inhibiteurs de Rho kinase 2), ESR1 (antagonistes du récepteur alpha des œstrogènes), et OPRK1 (agonistes des récepteurs opioïdes kappa) ont été choisis. Les résultats qu'on a obtenus montrent que les nombres des vrais actifs récupérés parmi les 5% des composés les mieux classés par notre méthode (alignement de molécules sur les pharmacophores par Shaper2, sélection de pose par l'énergie totale MMFF94, classement de composés par l'énergie PLP), dans la plupart des cas, sont égaux ou supérieurs à ceux obtenus par le docking moléculaire avec Surflex-Dock, et par la recherche par similarité géométrique en 3D avec ROCS. Notre méthode est également arrivée à récupérer le plus de "chémotypes/scaffolds" des vrais actifs par rapport aux deux autres méthodes pour les deux entrées ESR1 et OPRK1 (**Figure 2**). Il est clair donc que notre approche est aussi efficiente que d'autres méthodes computationnelles

dans des challenges de criblage virtuel et tend à récupérer plus de chémotypes originaux des composés actifs.

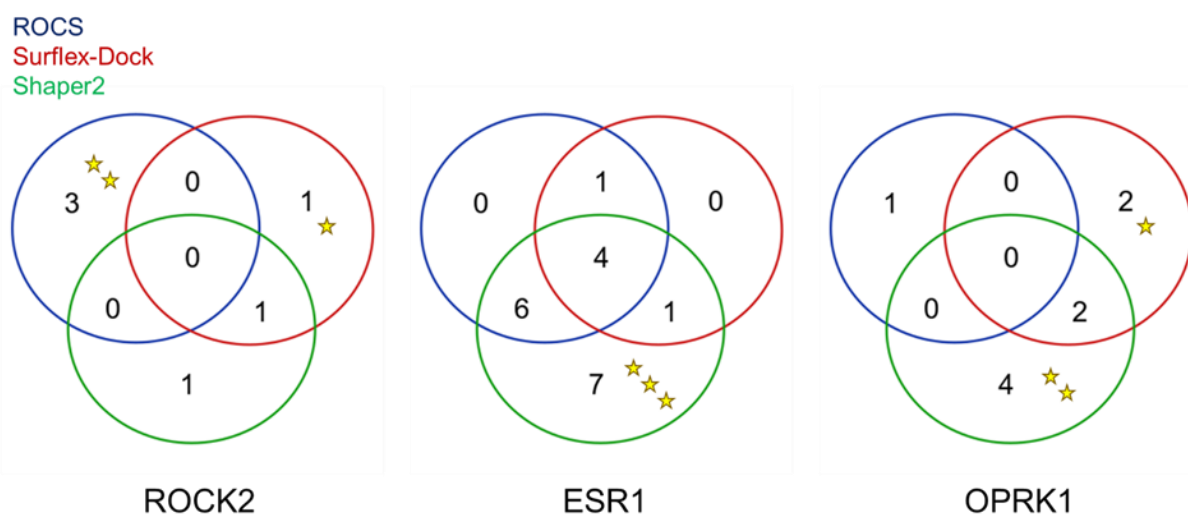


Figure 2. Performance des trois méthodes de criblage virtuel (ROCS, Surflex-Dock, Shaper2) dans le deuxième challenge de criblage virtuel avec trois jeux de données “PubChem BioAssay” (ROCK2 – AID644, ESR1 – AID743080, OPRK1 – AID1777). Les nombres de vrais actifs récupérés parmi les 5% des composés les mieux classés par les méthodes sont indiqués dans les diagrammes de Venn. Chaque étoile signifie un chémotype des actifs récupérés par une seule méthode.

2.2. Développement de jeux de données “PubChem BioAssay” non-biaisés pour les études de criblage virtuel rétrospectif

Les données de bioactivité expérimentales sont récupérées à partir du site web de “PubChem BioAssay”, où se trouvent toutes les informations relatives aux essais biologiques déjà réalisés sur une cible thérapeutique, y compris les vrais actifs et les vrais inactifs ainsi que les valeurs d'affinité (EC_{50} , IC_{50} , K_d , ou K_i) en μM ou nM. Une

première étape de pré-sélection a eu lieu pour garder seulement les jeux de données avec au moins 10.000 substances testées, dont au moins 50 ont été confirmées comme actives par une étude dose-réponse, sur une protéine cible ayant été co-cristallisée au moins une fois avec un ligand du même phénotype (inhibiteur, agoniste, antagoniste, etc.) que celui des vrais actifs validés par l'essai biologique qui correspond. Une totalité de 21 jeux de données correspondant à 21 protéines cibles d'intérêt pharmaceutique, couvrant 11 familles de protéines, ont été retenus. Plusieurs familles fortement étudiées depuis des années, telles que les RCPGs ($n = 3$), les kinases ($n = 3$), ou les récepteurs nucléaires ($n = 5$), sont choisies. 162 structures cristallographiques en 3D (protéine en complexe avec un ligand pour chacune) pour l'ensemble des 21 jeux de données sont trouvées sur la "Protein Data Bank". Tous ces résultats ont été mis à jour au 31 décembre 2018.

Chaque complexe protéine-ligand a été ensuite téléchargé directement depuis le site web de la "Protein Data Bank" en format pdb. Les hydrogènes ont été ajoutés avec Protoss. Toutes les molécules d'eau qui se trouvent dans le site de liaison qui participent à au moins trois liaisons d'hydrogène avec la protéine et/ou le ligand, dont au moins deux sont avec la protéine, ont été conservées. Les structures des protéines, des ligands et des sites de liaison ont été enregistrées séparément en format mol2.

Toutes les substances de chaque jeu de données ont été téléchargées en format sdf depuis le site web de "PubChem BioAssay". Les informations relatives à chaque substance ont été ensuite récupérées, y compris l'activité (actif/inactif), le phénotype (inhibiteur, agoniste, antagoniste), la puissance (en μM), la valeur de HillSlope, la

“fréquence de touche”, la masse moléculaire, le coefficient de partage octanol/eau (ALogP), la charge formelle, le nombre de liaisons à rotation libre, et le nombre d’accepteurs ou de donneurs de liaisons d’hydrogène. Les règles de filtrage ont été déterminées de sorte que les faux positifs ainsi que l’enrichissement artificiel soient évités. Le processus de filtrage à quatre étapes est effectué comme suit :

- Etape 1 : Filtre de substances inorganiques : les molécules qui possèdent au moins un atome autre que H, C, N, O, P, S, F, Cl, Br, et I ont été enlevées. Toutes les substances (actives et inactives) ont passé cette étape.
- Etape 2 : Filtre de faux positifs : un actif est retenu seulement si sa valeur de HillSlope est entre 0,5 et 2 (étape 2a), si la “fréquence de touche” est inférieure à 0,26 (étape 2b), s’il n’est pas considéré comme agrégateur ou inhibiteur de la luciférase et s’il n’a pas la propriété autofluorescente (étape 2c). Les substances inactives, par contre, n’ont pas passé cette étape.
- Etape 3 : Filtre de propriétés moléculaires : une substance est retenue seulement si sa masse moléculaire est entre 150 et 800 Da, si son ALogP est entre -3 et +5, s’il possède moins de 15 liaisons à rotation libre, 10 accepteurs/donneurs de liaisons d’hydrogène, et si sa charge formelle est entre -2 et +2. Toutes les substances ont passé cette étape.
- Etape 4 : Conversion en 3D et ionisation : les structures en 2D des substances restant ont été converties en 3D avec Corina, et ensuite ionisées à pH

physiologique avec Filter (OpenEye). Toutes les substances ont passé cette étape.

Presque 60% des vrais actifs ont été éliminés après toutes les étapes de filtrage. Il est observé que la sous-étape 2a a filtré le plus d'actifs (les actifs non-spécifiques ayant plusieurs sites de liaison). Par contre, seulement 10% des vrais inactifs ont été retirés, car ils n'ont pas passé l'étape 2 comme les substances actives (**Figure 3**). Ces étapes de filtrage soulignent l'importance de l'élimination des artefacts de test dans la composition des vrais actifs, car elles retirent non seulement les faux positifs qui pourraient ultérieurement impacter la performance du criblage virtuel, mais aussi font baisser le taux des actifs par rapport aux inactifs, rendant les taux de touche de nos jeux de données plus proches de ceux qui sont typiquement observés lors de criblages expérimentaux à haut débit. D'ailleurs, la puissance des composés actifs de DUD-E ou de ChEMBL est en général plus élevée que celle de nos jeux de données, c'est-à-dire que nos actifs sont plus difficiles à détecter, et permettent une meilleure discrimination entre les méthodes de criblage *in silico*, puisque la surestimation de la performance de ces méthodes est minorée.

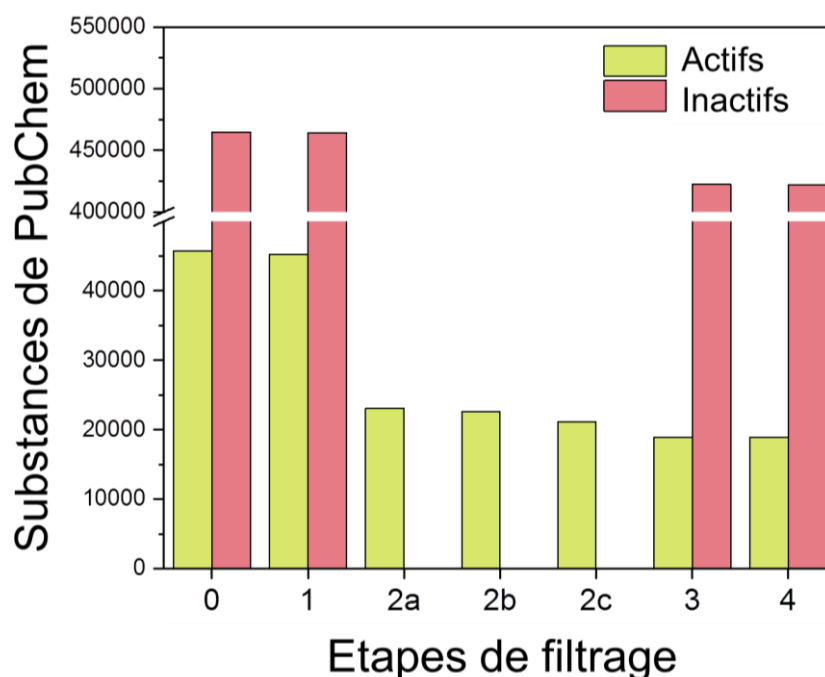


Figure 3. Les nombres de vrais actifs et de vrais inactifs de PubChem, sélectionnés après chaque étape de filtrage. Les substances inactives n'ont pas passé les sous-étapes 2a, 2b et 2c.

Les jeux de données déjà préparés ont été ensuite évalués par des méthodes de criblage virtuel "ligand-based" (recherche par similarité en 2D avec ECFP4 ou en 3D avec ROCS) ou "structure-based" (docking moléculaire avec Surflex-Dock). Le meilleur coefficient de Tanimoto (donné par les méthodes "ligand-based") et le meilleur score de docking (issu par Surflex-Dock) ont été enregistrés pour chaque substance. Chacun des 162 complexes cristallographiques trouvés dans la "Protein Data Bank" a été utilisé comme support ("template"), générant autant de listes de touches que de supports disponibles. En plus, l'approche "max-pooling" a été également utilisée, dans laquelle seulement le meilleur score donné par tous les supports a été retenu pour chaque

substance. Les valeurs d'EF1% (enrichissement en vrais actifs correspondant à un taux de faux positifs de 1%) ont été calculées pour évaluer la performance du criblage.

Il est observé que les valeurs d'EF1% de chaque entrée sont très variables dans la plupart des cas, confirmant l'influence du choix du support de référence et de méthode sur la performance du criblage. L'enrichissement comparable à ou moins bon que celui obtenu par la sélection aléatoire (EF1% = 1,0) est observé chez plusieurs entrées. Notamment, sur six jeux de données (ARO1, GLP1R, GLS, L3MBTL1, RORC, THRB), aucune méthode n'est arrivée à donner un enrichissement supérieur à 2,0 avec l'approche "max-pooling" (**Figure 4**). Pour cinq d'entre eux, aucun support n'a donné un EF1% > 2,0. Ceci signifie la difficulté remarquable de nos jeux de données, grâce à l'absence des biais structurels dans la composition des substances (les actifs, les inactifs et les ligands de référence) et la distribution de la puissance des vrais actifs qui n'est pas orientée vers les valeurs sub-micromolaires. Parmi les 21 jeux de données évalués, 15 (sauf les six mentionnés ci-dessus) ont été sélectionnés et constituent donc la nouvelle base de données intitulée LIT-PCBA. Chacun entre eux a été ensuite divisé en quatre sous-ensembles ("training actives", "validation actives", "training inactives", "validation inactives") par la méthode "asymmetric validation embedding" (AVE) qui mesure la distance dans l'espace chimique de chaque paire de molécules pour les distribuer dans les sous-ensembles de sorte que le biais total soit minimisé [6]. Pour 12 jeux de données (à part ALDH1, VDR, FEN1), une valeur de biais inférieure à 0,01 a été atteinte après seulement quelques itérations de l'algorithme génétique (pour les trois qui restent, les valeurs de biais total sont toutes inférieures à 0,10). Ceci confirme encore une fois qu'il y a très peu de biais dans la composition de nos jeux de données,

et fait preuve de la qualité de LIT-PCBA en tant qu'une base de données prête à l'emploi pour évaluer la performance de méthodes de criblage virtuel à l'avenir.

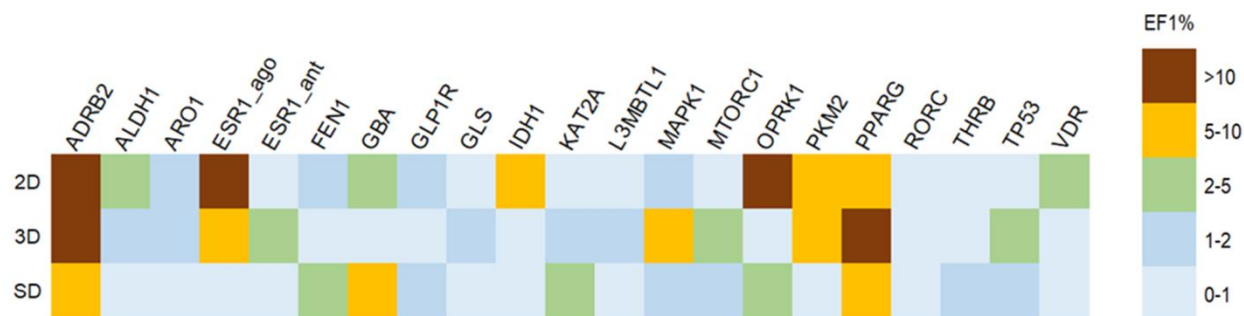


Figure 4. “Heat map” illustrant la performance des trois méthodes de criblage *in silico* : 2D – recherche par similarité en 2 dimensions avec les fingerprints ECFP4, 3D – recherche par similarité en 3 dimensions avec ROCS, et SD – docking moléculaire avec Surflex-Dock, en matière d’EF1% obtenus par l’approche “max-pooling” sur les 21 jeux de données après les quatre étapes de filtrage.

En plus, les poses de docking issues de Surflex-Dock ont été réévaluées par deux méthodes : IFP (“protein-ligand interaction fingerprints”) et GRIM (“graph-matching”) [9,11], qui ont déjà été intégrées au logiciel IChem, en utilisant les structures cristallographiques des ligands de référence et des sites de liaison. Il est observé que le classement des molécules selon la similarité des interactions protéine-ligand (IFP) ou la similarité des graphes d'interaction (GRIM) a donné les valeurs d’EF1% plus élevées que celles obtenues à partir des scores de docking de Surflex-Dock, confirmant l’importance de l’étape traitement des poses de docking d’une molécule, notamment par les approches basées sur la comparaison des modes d’interactions ligand-protéine de ces poses avec celles d’un référent dans les challenges de criblage virtuel.

3. Conclusion générale

Les éléments pharmacophoriques “structure-based” issus d’ICChem qui représentent le site actif d’une protéine (même sans ligand co-cristallisé) sont simples et assez précis pour faire du criblage virtuel. La nouvelle procédure proposée dans ce travail (alignement de molécules sur les pharmacophores par Shaper2, sélection de pose par l’énergie totale MMFF94, et classement de composés par l’énergie PLP) s’avère aussi efficiente que des méthodes computationnelles existantes dans l’identification des composés actifs et leurs chémotypes originaux, et peut donc être utilisée en parallèle avec d’autres méthodes de criblage *in silico* afin d’améliorer la performance globale du criblage. On présente également la nouvelle base de données LIT-PCBA, se composant de 15 protéines cibles, chacune avec les vrais actifs et les vrais inactifs déjà confirmés par les essais biologiques issus de “PubChem BioAssay”. Ces jeux de données, préparés par une procédure rigoureuse de plusieurs étapes, sont moins biaisés, en matière de structure des ligands et de composition des sets de molécules, que ceux qui existent déjà (DUD, DUD-E, etc.), et sont donc plus difficiles. LIT-PCBA est prête à l’emploi pour des études comparatives de nouvelles méthodes de criblage virtuel, notamment celles basées sur l’intelligence artificielle.

4. Références

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

[2] 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.

[3] Chaput, L.; Martinez-Sanz, J.; Saettel, N.; Mouawad, L. Benchmark of Four Popular Virtual Screening Programs: Construction of the Active/Decoy Dataset Remains a Major Determinant of Measured Performance. *J. Cheminformatics* **2016**, *8*, 56.

[4] Chen, L.; Cruz, A.; Ramsey, S.; Dickson, C. J.; Duca, J. S.; Hornak, V.; Koes, D. R.; Kurtzman, T. Hidden Bias in the DUD-E Dataset Leads to Misleading Performance of Deep Learning in Structure-Based Virtual Screening. *PLoS One* **2019**, *14*, No. e0220113.

[5] Sieg, J.; Flachsenberg, F.; Rarey, M. In Need of Bias Control: Evaluating Chemical Data for Machine Learning in Structure-Based Virtual Screening. *J. Chem. Inf. Model.* **2019**, *59*, 947–961.

[6] Wallach, I.; Heifets, A. Most Ligand-Based Classification Benchmarks Reward Memorization Rather Than Generalization. *J. Chem. Inf. Model.* **2018**, *58*, 916–932.

[7] Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T.; Mortenson, P. N.; Murray, C. W. Diverse, Highquality Test Set for the Validation of Protein-Ligand Docking Performance. *J. Med. Chem.* **2007**, *50*, 726–741.

[8] SZYBKI 1.8.0.1: OpenEye Scientific Software, Santa Fe, NM 87508, USA.

[9] 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.

[10] 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.

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

LISTE DES PRESENTATIONS

1. Viet-Khoa Tran-Nguyen, Didier Rognan. LIT-PCBA: An Unbiased Database for Evaluation of Virtual Screening Methods. 9e Journées de la Société Française de Chémoinformatique (SFCi). Paris, France (21-22/11/2019). Présentation en affiche.
2. Viet-Khoa Tran-Nguyen, Franck Da Silva, Guillaume Bret, Didier Rognan. Generation and Investigation of Protein-Based Cavity Pharmacophores. 7e Réunion Scientifique Publique du LabEx MEDALIS. Strasbourg, France (26/04/2019). Présentation orale.
3. Viet-Khoa Tran-Nguyen, Franck Da Silva, Guillaume Bret, Didier Rognan. Generation and Investigation of Protein-Based Cavity Pharmacophores. 21e Congrès du Groupe de Graphisme et Modélisation Moléculaire (GGMM). Nice, France (03-05/04/2019). Présentation en affiche.
4. Viet-Khoa Tran-Nguyen, Franck Da Silva, Guillaume Bret, Didier Rognan. Generation and Investigation of Protein-Based Cavity Pharmacophores. Journée du Campus d'Illkirch (JCI 2019). Illkirch-Graffenstaden, France (01-02/04/2019). Présentation orale.
5. Viet-Khoa Tran-Nguyen, Franck Da Silva, Guillaume Bret, Didier Rognan. Generation and Investigation of Protein-Based Cavity Pharmacophores. 26e Journée des Jeunes Chercheurs de la Société de Chimie Thérapeutique. Paris, France (20-22/02/2019). Présentation en affiche.
6. Viet-Khoa Tran-Nguyen, Franck Da Silva, Guillaume Bret, Didier Rognan. Generation and Virtual Screening of Protein-Based Cavity Pharmacophores: A Prospective Study. Journée des Doctorants de l'Ecole Doctorale des Sciences Chimiques. Strasbourg, France (23/11/2018). Présentation orale.
7. Viet-Khoa Tran-Nguyen, Guillaume Bret, Didier Rognan. Generation and Investigation of Protein-Based Cavity Pharmacophores: A Prospective Study. 54e Rencontres Internationales de Chimie Thérapeutique (RICT 2018). Strasbourg, France (04-06/07/2018). Présentation en affiche.

8. Viet-Khoa Tran-Nguyen, Guillaume Bret, Didier Rognan. Generation and Virtual Screening of Protein-Based Cavity Pharmacophores: A Prospective Study. 7th French-Japanese Workshop on Computational Methods in Chemistry. Strasbourg, France (02-03/07/2018). Présentation en affiche.
9. Viet-Khoa Tran-Nguyen, Guillaume Bret, Didier Rognan. Generation and Virtual Screening of Protein-Based Cavity Pharmacophores: A Prospective Study. Strasbourg Summer School in Chemoinformatics – 2018. Strasbourg, France (25-29/06/2018). Présentation en affiche.

LISTE DES PUBLICATIONS (relatives à la thèse)

1. Tran-Nguyen, V. 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. doi: 10.1021/acs.jcim.8b00684.
2. 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** (sous presse). doi: 10.1021/acs.jcim.0c00155.
3. Tran-Nguyen, V. K.; Rognan, D. Benchmarking Data Sets from PubChem BioAssay Data: Current Scenario and Room for Improvement. *Int. J. Mol. Sci.* **2020**, *21*, 4380. doi: 10.3390/ijms21124380.

Introduction

Discovering the very first ligand that exerts a desired bioactivity towards a protein target in a fast and cost-effective manner has long been a main challenge in drug design. For a particular protein whose ligands' three-dimensional structures are not yet available, the molecular docking technique is usually employed as a virtual screening tool to detect potential “hits”, despite the unresolved issues in quantitatively predicting these molecules' binding affinity. It is therefore necessary to conceive a novel computational approach that can be applied to apoproteins. In 2012, the researchers at the “Laboratoire d'Innovation Thérapeutique” (University of Strasbourg) managed to design a new pharmacophore perception method that was already integrated in the IChem software package.^{1,2} This method automatically detects all possible ligand-binding sites on the surface of any given protein target, then predicts the “druggability” of each cavity, and finally creates a set of structure-based pharmacophoric points that represent each pocket that was previously deemed potentially “druggable”. At this point, a question arises as to how we make use of these pharmacophore models to screen a chemolibrary comprising thousands, or even millions of molecules, with the aim of rationally selecting potential “hits” for a protein of pharmaceutical interest, regardless of the availability of a co-crystallized ligand.

Once a novel *in silico* screening procedure is developed, it must be evaluated in terms of discriminatory power to make sure that it manages to retrieve active molecules for a biological target among a pool of structurally diverse compounds. This has to be done with the use of existing data sets, either found in the literature, or extracted from open-access databases. However, numerous problems with the sets of ligands currently employed by the cheminformatics community have been observed and reported.³⁻⁶ Among them are:

- (i) The absence of experimental evidence confirming the impotence of presumably inactive molecules (known as “decoys”);
- (ii) The presence of too many true actives with high potency towards the target;
- (iii) The hit rates of some data sets which are too high to be deemed realistic;
- (iv) The chemical bias in the composition of ligand sets, as the actives are issued from only a few chemical series, the decoys are too different from the true hits in terms of physicochemical features, the active compounds are too structurally similar to the co-crystallized ligands used as references.

Therefore, such benchmarking data sets do not describe real life, as they fail to mimic chemolibraries used in actual high-throughput screening campaigns, and overestimate the real accuracy of virtual screening methods. As a result, there arises the need for developing a novel unbiased data collection built upon experimentally confirmed data which can be applied to validating both ligand-based and structure-based screening procedures, which has a difficulty level (in terms of distinguishing true actives from true inactives) as close as possible to that of real high-throughput screening decks, and which is able to capture the differences in the performances of different *in silico* methods.

In light of the problems explained above, the work portrayed in this Ph.D. thesis is composed of two main sections as follows:

- The first main part concerns the development of a new procedure to align small ligands on the previously generated structure-based pharmacophore models, prior to the selection of one best pose for each ligand and the creation of a hit list where all molecules are sorted according to certain scoring parameters. Once elaborated, this protocol can be employed to predict the pose of an active compound inside a “druggable” binding pocket of a protein, and to differentiate between the true actives and the “decoys” or the true inactives of a biological target of pharmaceutical interest. This part of the work is portrayed in the Chapter 2 of the dissertation.
- The second main part is focused on the construction of a new data set from experimental input deposited on PubChem BioAssay⁷ that features a wide range of protein targets, with obvious and hidden design bias already reduced. A post-preparation evaluation of this data collection using various virtual screening methods and scoring functions is also carried out to make sure that the aforementioned bias has been mitigated, confirming the advantage of employing such data to validate new *in silico* screening approaches. This part of the work is portrayed in the Chapter 3 of the dissertation.

Besides, with the aim of facilitating future high-quality benchmarking data set developments, in the Chapter 1 of this dissertation, a comprehensive review of data collections built upon PubChem BioAssay input is also provided, along with an analysis of notable issues that must not be neglected when it comes to constructing a novel database, leading to the suggestion of some good practices that should be followed to ensure the quality of data set design. Finally, the

Chapter 4 of this thesis concerns the rescoring of docking poses issued by a popular docking program (Surflex-Dock) on the ensemble of ligand sets previously presented in Chapter 3, aiming to highlight the advantage of scoring functions relying on protein-ligand interaction comparisons over energy-based empirical ones in recognizing the true hits of a biological target from a pool of chemically diverse and unbiased molecules.

References

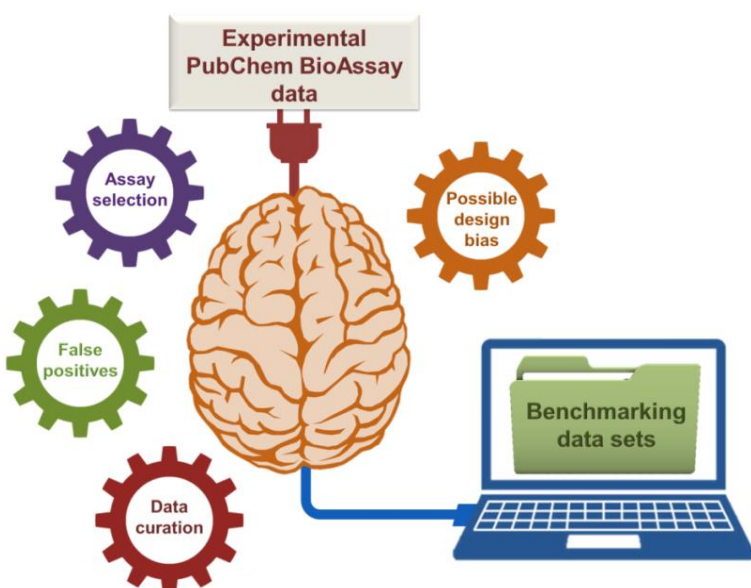
1. Da Silva, F.; Desaphy, J.; Rognan, D. IChem: A Versatile Toolkit for Detecting, Comparing, and Predicting Protein-Ligand Interactions. *ChemMedChem* **2018**, *13*, 507-510.
2. Desaphy, J.; Azdimousa, K.; Kellenberger, E.; Rognan, D. Comparison and Druggability Prediction of Protein-Ligand Binding Sites from Pharmacophore-Annotated Cavity Shapes. *J. Chem. Inf. Model.* **2012**, *52*, 2287-2299.
3. Chaput, L.; Martinez-Sanz, J.; Saettel, N.; Mouawad, L. Benchmark of Four Popular Virtual Screening Programs: Construction of the Active/Decoy Dataset Remains a Major Determinant of Measured Performance. *J. Cheminf.* **2016**, *8*, 56.
4. Chen, L.; Cruz, A.; Ramsey, S.; Dickson, C. J.; Duca, J. S.; Hornak, V.; Koes, D. R.; Kurtzman, T. Hidden Bias in the DUD-E Dataset Leads to Misleading Performance of Deep Learning in Structure-Based Virtual Screening. *PLoS One* **2019**, *14*, No. e0220113.
5. Sieg, J.; Flachsenberg, F.; Rarey, M. In Need of Bias Control: Evaluating Chemical Data for Machine Learning in Structure-Based Virtual Screening. *J. Chem. Inf. Model.* **2019**, *59*, 947-961.
6. Wallach, I.; Heifets, A. Most Ligand-Based Classification Benchmarks Reward Memorization Rather Than Generalization. *J. Chem. Inf. Model.* **2018**, *58*, 916-932.
7. PubChem BioAssay. <https://www.ncbi.nlm.nih.gov/pcassay/> (accessed December 2018).

Chapter 1

Benchmarking Data Sets from PubChem BioAssay Data: Current Scenario and Room for Improvement

Developing realistic data sets for evaluating virtual screening methods is a task that has been tackled by the cheminformatics community for many years. Numerous artificially constructed data collections were developed, but they all suffer from multiple drawbacks, one of which is the unknown potency of presumably inactive molecules, leading to possible false negatives in the ligand sets. In light of this problem, the PubChem BioAssay database, an open-access repository providing bioactivity information of compounds that were already tested on a biological target, is now a recommended source for data set construction. Nevertheless, there exist several issues with the use of such data that need to be properly addressed. In this chapter, an overview of benchmarking data collections built upon experimental PubChem BioAssay input is provided, along with a thorough discussion of note-worthy issues that one must consider during the design of new ligand sets from this database. This chapter has been published as a review article in the special issue “QSAR and Chemoinformatics in Molecular Modeling and Drug Design” of the International Journal of Molecular Sciences.

Tran-Nguyen, V. K.; Rognan, D. Benchmarking Data Sets from PubChem BioAssay Data: Current Scenario and Room for Improvement. *Int. J. Mol. Sci.* **2020**, *21*, 4380. doi: [10.3390/ijms21124380](https://doi.org/10.3390/ijms21124380).



1. Introduction

The PubChem BioAssay database (<http://pubchem.ncbi.nlm.nih.gov/bioassay>) was first introduced in 2004 as a part of the PubChem project initiated by the National Center for Biotechnology Information (NCBI), aiming to provide the scientific community with an open-access resource where experimental bioactivity high-throughput screening (HTS) data of chemical substances can be found.¹⁻⁵ Starting out with small-molecule HTS input from the National Institute of Health (NIH), the database now gathers data from over 700 different sources, including governmental organizations, world-renowned research centers, chemical vendors as well as other biochemical databases, featuring over 260 million bioactivity data points reported in both small-molecule assays and RNA interference reagents-screening projects.⁵⁻¹¹ Journal publishers are also acknowledged for a significant contribution to the growth of PubChem BioAssay, as the database has received experimental input from more than 30 million scientific publications in response to requests from over 400 peer-reviewed journals (as of April 30, 2020),¹⁰⁻¹² denoting a constant and tremendous effort from many sectors of the scientific community to support free sharing of HTS data.

Soon after its introduction, PubChem BioAssay has established itself as a reliable and highly-queried public repository where information on each biological assay, from overall descriptions to detailed screening protocols, from input data to assay results, as well as chemical features and bioactivities of all tested molecules, can be easily accessed and downloaded directly from the webpage. The two search options (limits search and advanced search) allow a systematic and thorough investigation of the assays deposited on the database, according to various parameters, e.g. assay type, target type, or quantity of featured substances, offering a practical data collection and analysis tool.¹³ Information on related targets and same-project assays enables a more complete look into the body of screening campaigns on the same or closely-related biological targets. Crosslinks to the NCBI Entrez information retrieval system,¹⁴ PubMed Central¹⁵ and the Protein Data Bank¹⁶ also facilitate research relying on the use of data extracted from the resource. Various updates have been brought to PubChem BioAssay over the years, enlarging the size of available archival data, introducing new features to the web interface and improving data sharing capability.¹⁷⁻²⁰ Several million users have been procuring data from the website and its different programmatic services each month,²¹ highlighting the importance of this public

database as a key source of chemical information for researchers, students and the general public from around the world.

In this review article, a quick summary of assays and compounds deposited on PubChem BioAssay, along with an overview of data sets built by the cheminformatics community upon the data retrieved from this repository will be provided. We also give a thorough discussion of noteworthy issues that have to be addressed prior to utilizing such data in cheminformatics-related projects, with illustrations observed in our recently introduced LIT-PCBA data collection,²² which was constructed from PubChem BioAssay data.

2. PubChem BioAssay Statistics: Assays and Compounds

As of April 30, 2020, there were 1,067,896 assays deposited on the database. The vast majority of them (99.98%) involved small-molecule screening, only 177 assays were conducted with RNA interference reagents. These assays are classified according to the number of tested substances (chemical samples provided by data contributors⁸), the number of active substances, the screening stage, and the target type, as listed in **Table S1**. It can be deduced that most PubChem assays are small-scale screening projects, with over 99% of them conducted on fewer than 100 substances, and nearly 94% giving no more than nine actives (**Figure 1**). The screening stage was, in most cases (about three quarters), not specifically annotated. Assays giving confirmatory results regarding the bioactivities of tested molecules account for a larger proportion than primary screens, though dose-response curves are not always provided. Interestingly, nearly 75% of available assays do not have a specific biological target (i.e. a protein, a gene or a nucleotide), but are rather cell-based assays identifying molecules that interfere with a certain cell function or an intracellular activity (e.g. tumor cell growth inhibitors, lipid storage modulators, HIV-1 replication inhibitors), or are pharmacokinetics studies. On the other hand, some assays take multiple macromolecules as targets (e.g. AID 1319). The utility of data extracted from these assays in cheminformatics-related research will be later discussed in the manuscript.

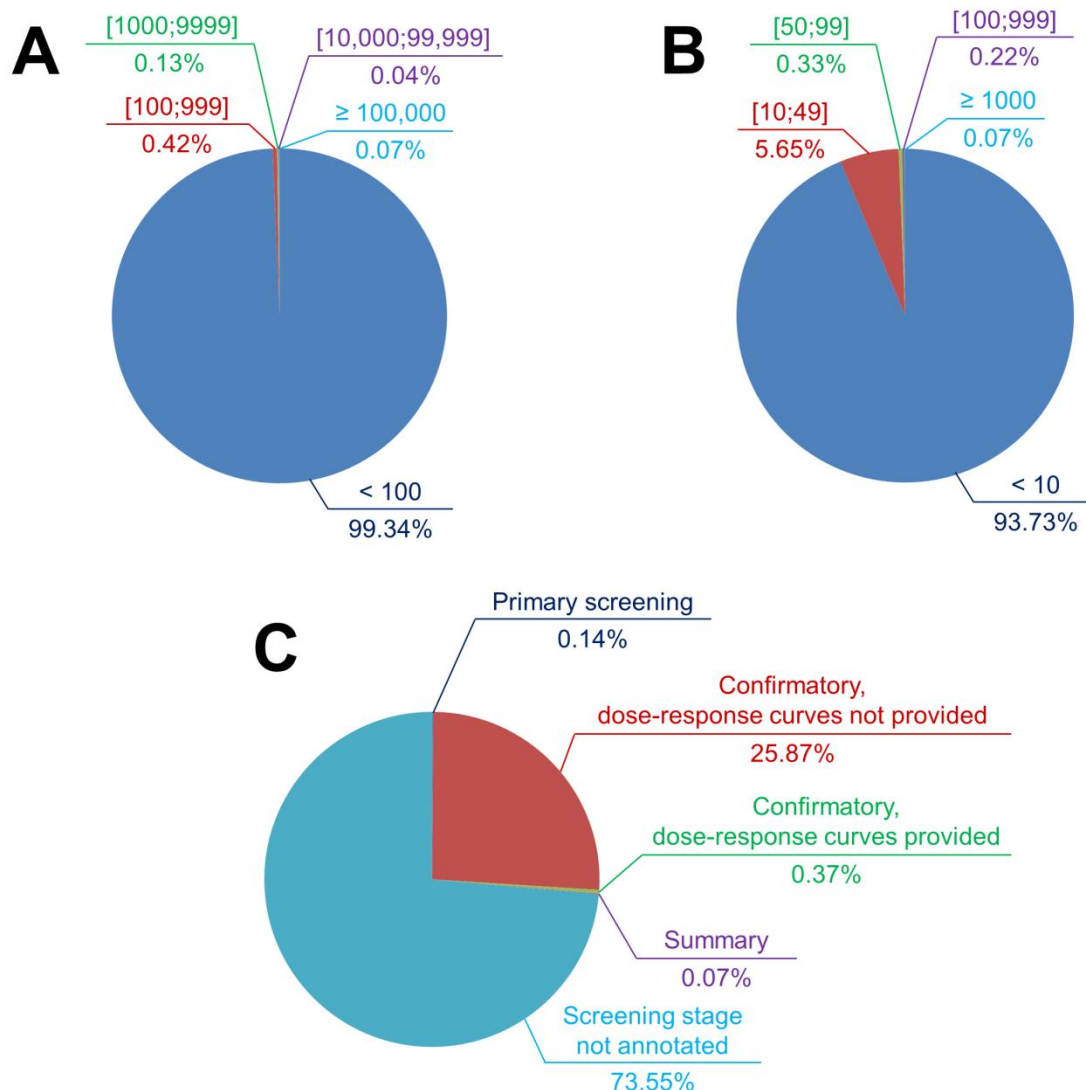


Figure 1. Partition of small-molecule PubChem bioactivity assays according to the number of tested substances (**A**), the number of active substances (**B**), and the screening stage (**C**). It is observed that most assays are small-scale screening projects in which fewer than 100 substances were tested, and no more than nine actives were identified. All statistics were updated as of April 30, 2020.

A total of 102,694,672 compounds were tested in at least one PubChem bioactivity assay (as of April 30, 2020), over 95% of which are organic molecules (i.e. molecules bearing no atom other than H, C, N, O, P, S, F, Cl, Br, and I). The term “compounds”, according to PubChem, refers to unique chemical structures that were extracted and standardized from the community-provided substances.⁸ A question always raised when it comes to drug design is whether a chemical compound is drug-like or not, or if a molecule has physicochemical properties that are deemed

favorable for oral administration in humans. Several rules of thumb have addressed this issue, giving criteria largely employed to predict a compound's drug-likeness, including the Lipinski's rule of five,^{23,24} the Ghose filter,²⁵ and the Veber's rule.²⁶ PubChem compounds are analyzed according to each criterion,²³⁻²⁷ and statistics are given in **Table S2**. Statistical results show that most compounds tested in PubChem bioactivity assays satisfy the aforementioned rules, indicating their potential to become orally active drugs (**Figure 2**). However, only 1% of them (over 1 million compounds) were deemed active in at least one screening experiment, highlighting the miniature portion of active molecules available in the database, and implying an average "hit rate" lower than those observed in artificially constructed data sets such as DUD,²⁸ DUD-E,²⁹ or DEKOIS 2.0.³⁰ The other compounds were either biologically inactive in all assays where they were tested, or were left "inconclusive" in terms of bioactivity. These "inconclusive" compounds, present in various AIDs such as 1345009, 1345010, or 743075, have to be discarded when data extracted from PubChem BioAssay are used in cheminformatics-related research. On the other hand, compounds being repeatedly inactive in HTS assays, dubbed "dark chemical matter",³¹ are in fact important to keep, notably for identifying ligands of novel targets (e.g. protein-protein interfaces).

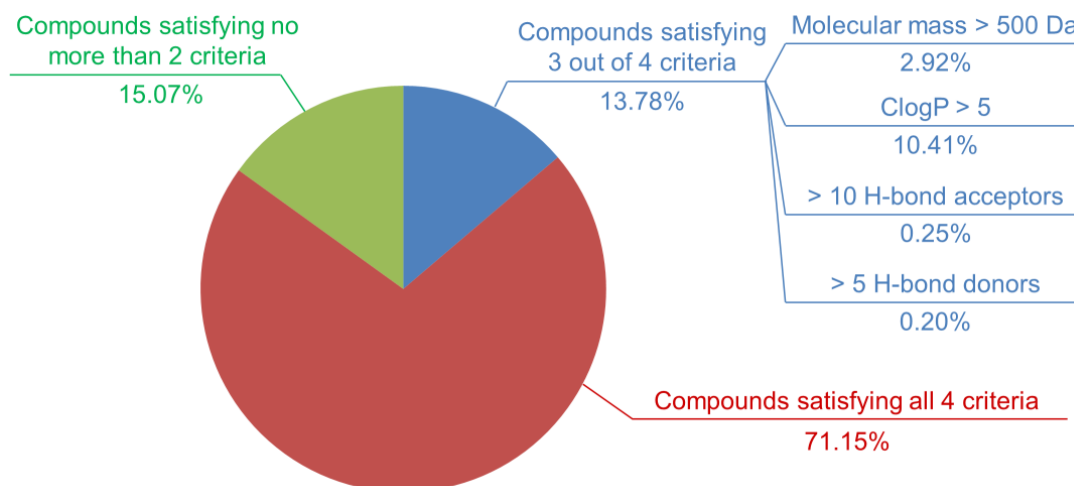


Figure 2. Partition of compounds tested in PubChem bioactivity assays according to four criteria of the Lipinski's rule of five. It is observed that most compounds (over 70%) satisfy all criteria. Nearly 85% of deposited compounds violate no more than one criterion. On the other hand, only 0.1% of all compounds (over 130,000) do not satisfy any criterion. Statistics were updated as of April 30, 2020.

3. What We Can Do with PubChem BioAssay Data: from the Data Set Construction Point of View

Being a wealth of experimental bioactivity data constantly gathered from many parts of the world, PubChem BioAssay offers ample opportunities for scientists from various disciplines, e.g. biochemistry, pharmacy, or cheminformatics, to exploit this abundant resource for both teaching and researching purposes. Access to the database is facilitated by numerous online services, in both manual (via PubChem limited and advanced search engines^{32,33}) and programmatic ways (via access routes such as the Power User Gateway PUG,³⁴ PUG-SOAP,³⁵ PUG-REST,³⁶ PUG-View,³⁷ the PubChemRDF REST interface³⁸ or the Entrez Utilities¹⁴). Recently, a novel web service called ScrubChem was introduced,³⁹ gathering PubChem BioAssay data that were already reparsed, digitally curated and improved, allowing a systematic analysis of all targets, chemicals and assays featured on the database at low computational costs, after which the cleaned data can be downloaded for use in modeling applications. Upon acquiring experimental input from the resource, scientists may use it in various ways to achieve their research objectives. Several review articles have been published in this regard,^{7,40,41} summarizing a wide range of studies that were conducted on the basis of PubChem BioAssay data.⁴²⁻⁶⁰ In this section, we only place our focus on the research featuring benchmarking data collections that were constructed by the cheminformatics community from PubChem's experimental results as a means of validating *in silico* screening protocols.

Throughout the years, various artificially constructed data sets have been developed,^{28-30,61-71} including DUD, DUD-E, or DEKOIS 2.0. However, the design of these collections suffers from many drawbacks, as demonstrated in several studies.⁷²⁻⁷⁶ One of them is the unknown potency of presumably inactive molecules, also known as “decoys”, which were usually extracted from the BIOVIA Available Chemicals Directory (ACD)⁷⁷ or the ZINC database.⁷⁸ This means there is no guarantee that the “decoys” do not exert the desired bioactivity against the protein target, due to the lack of relevant experimental evidence, and it is therefore very likely that false negatives exist among the inactive molecules. Using data from PubChem BioAssay as input for database construction, on the other hand, helps alleviate this problem. A number of data collections of different sizes have been designed from PubChem data and introduced to the scientific community, offering better references for evaluating novel virtual screening methods. Not

counting non-publicly available data sets (e.g. the three small- and medium-sized ligand sets that we designed in 2019 to validate our new pharmacophore-based ligand-aligning procedure⁷⁹), in this section, we only mention open-access ones, including the MUV data sets,⁸⁰ the UCI Machine Learning Repository,⁸¹ the BCL::ChemInfo framework by Butkiewicz *et al.*,⁸² the Lindh *et al.* data collection,⁸³ and our recently introduced LIT-PCBA (**Table 1**).²²

Table 1. Overview of the main open-access benchmarking data sets developed from experimental PubChem BioAssay data.

Data sets	Year	Number of ligand sets	Number of molecules per ligand set	Active-to-inactive ratio	Assay data		Assay artifacts avoided	Chemical bias avoided	Virtual screening suitability	
					Primary	Confirmatory			Ligand-based	Structure-based
MUV ⁸⁰	2009	17	15,030	2×10^{-3}	✓	✓	✓	✓	✓ ^a	✓
UCI ⁸¹	2009	21	69 to 59,795	2×10^{-4} to 0.33	✓	✓			✓	
Butkiewicz <i>et al.</i> ⁸²	2013	9	61,849 to 344,769	5×10^{-4} to 7×10^{-3}		✓			✓	
Lindh <i>et al.</i> ⁸³	2015	7	59,462 to 338,003	7×10^{-5} to 1×10^{-3}	✓	✓	✓	✓	✓	✓
LIT-PCBA ²²	2020	15	4247 to 362,088	5×10^{-5} to 0.05		✓	✓	✓ ^b	✓	✓

^a Ligand-based approaches are preferred.

^b Unbiased training and validation sets are provided for machine learning.

3.1. The MUV Data Sets

The Maximum Unbiased Validation (MUV) data sets, built by Rohrer and Baumann in 2008 and published in early 2009,⁸⁰ are among the first benchmarking sets of compounds whose bioactivity was experimentally determined and retrieved from PubChem BioAssay, which, as a result, avoids the issue regarding unknown potency values of presumably inactive molecules (“decoys”) inherent in other data sets.⁸⁰ Based upon 18 pairs of primary HTS and corresponding confirmatory dose-response experiments, whose biological targets range from kinases, GPCRs, nuclear receptors to protein-protein interactions, 17 medium-sized ligand sets (15,030 compounds), each with an active-to-inactive ratio at 2×10^{-3} , were generated, implying smaller hit rates in comparison to those of other databases.^{76,80} Specifically designed to be maximally unbiased, the MUV data sets were prepared according to a workflow that removed assay artifacts, prevented artificial enrichment, and reduced “analogue bias” in the composition of their ligands. A series of consecutive filters was first applied to eliminate “false positives” among active molecules, including promiscuous aggregators, frequent hitters exerting off-target or cytotoxic effects, as well as chemicals which are likely to spoil the assay’s optical detection method. A subsequent “chemical space embedding filter”, encoded by vectorized descriptors related to physicochemical properties of each molecule (e.g. molecular weight, number of hydrogen bond donors/acceptors), was next employed to rule out actives that were not adequately embedded in inactive compounds, ensuring that the inactive sets did not significantly differ from the sets of actives, thus avoiding possible artificial enrichment. Finally, a refined nearest neighbor analysis was applied, based on a “nearest neighbor function” and an “empty space function”, to reduce both the level of self-similarity among the actives and the separation degree between active and inactive molecules, selecting only 30 true actives and 15,000 true inactives that were optimal as regards the criterion of spatial randomness for each ligand set. Post-design analyses on the resulting data sets showed that (i) there exist a large number of distinct molecular scaffolds presented by the ligands (1.2 compounds/scaffold class), denoting the absence of “analogue bias” and a good representation of drug-like chemical space; (ii) the correlation between the degree of data set clumping and retrospective virtual screening performance was no longer observed after MUV design, suggesting that the final ligand sets were indeed not affected by benchmarking data set bias; and (iii) the MUV data were significantly less biased than the then-standard DUD data set, as evidenced by a lower molecular

self-similarity level and a higher difficulty in distinguishing true actives from true inactives by ligand-based virtual screening simulations. The introduction of the MUV data collection therefore marks a milestone in the quest to construct realistic data sets entirely from experimental results with little design bias and applicability to evaluating both ligand-based and structure-based *in silico* methods, serving as an inspiration for future database development.

3.2. The UCI Repository

The UCI Machine Learning Repository was introduced in 2009.⁸¹ On the basis of data retrieved from 12 PubChem bioactivity assays, both primary ($n = 7$) and confirmatory ($n = 5$), a total of 21 medium- and small-sized data sets (69-59,795 compounds) were generated, either by using separately primary or confirmatory screening data, or by combining results from a primary assay and its corresponding confirmatory screen. In the latter case, compounds which were deemed as active in the primary experiments but later denounced as inactive by the confirmatory readouts were all considered inactive in the combined data sets (instead of being discarded as in the MUV collection). The active-to-inactive ratio ranges from 2×10^{-4} to 0.33. Each ligand set was then randomly split into a training-and-validation set (80% of the population) and an independent test set (the other 20%) for machine learning algorithm assessments.⁸¹ Despite being one of the earliest remarkable attempts at using experimental data from PubChem BioAssay for data set construction, the UCI database itself has several limitations. Firstly, though the author offered 21 data sets in total, only four of them, which were built by combining primary and confirmatory results, were recommended. Reasons for this lie in (i) the high portion of false positives recorded in primary experiment-based ligand sets that casts doubt on the solitary use of such data for evaluating *in silico* screening; (ii) the hit rates observed in the sets built upon confirmatory assays alone are too high (7-33%) to be deemed realistic, notably in comparison to those of real screening decks; and (iii) the size of some data sets is too tiny (tens of active molecules among fewer than 100 compounds) for virtual screening methods (especially ligand-based ones) to give any meaningful result. Secondly, due to the lack of high-quality biological target 3D structures for several bioassays (e.g. AIDs 456, 1608) and insufficient information on possible binding site(s) of the molecules, the design focus of this data collection is implied to be limitedly placed on ligand-based (machine learning) approach evaluations. Thirdly, the issue of physicochemical bias in the composition of active and inactive molecules that may lead to artificial enrichment

and an overestimation of virtual screening performance, which had been raised in the MUV paper,⁸⁰ was not addressed throughout the development of these data sets, raising questions on the real benefit of using such data for validating novel *in silico* screening procedures.

3.3. The Butkiewicz *et al.* Data Collection

Another PubChem BioAssay-based data collection was introduced in 2013 by Butkiewicz *et al.* as a part of the cheminformatics framework BCL::ChemInfo.⁸² Nine medium- and large-sized data sets (> 60,000 compounds) were constructed upon collating results from relevant confirmatory screens, thus avoiding the issue of false positives commonly observed when only primary readouts are accounted. Diverse classes of protein targets are covered in the database, including three GPCRs, three ion channels, the choline transporter, the serine/threonine kinase 33 and the tyrosyl-DNA phosphodiesterase. Active-to-inactive ratios range from 5×10^{-4} to 7×10^{-3} , implying small hit rates which are all lower than 0.8% (< 0.1% in most cases). Though the number of true actives is deemed sufficiently large (> 170 actives for each ligand set) and the hit rates are generally low, one drawback of this database is that the problems regarding assay artifacts, analogue bias, and artificial enrichment due to physicochemical differences between active and inactive molecules (which need to be properly addressed during the construction phase) were completely overlooked. These issues are even more critical when data sets intended for evaluating ligand-based virtual screening methods (which is, in fact, the design focus of this data collection) are developed. There is hence no guarantee that only a little chemical bias exists in the composition of these ligand sets, and it is likely that *in silico* screening performance could be overestimated due to such unconsidered issues.

3.4. The Lindh *et al.* Data Collection

In 2015, Lindh *et al.* introduced a novel data collection designed for evaluating both ligand-based and structure-based virtual screening methods.⁸³ A rigorous procedure of analyzing the whole PubChem BioAssay database was first carried out, after which only assays (excluding cell-based and multiplex ones) that were performed with more than 1000 compounds (at least 20 of which were identified as active) against a single protein target that had been co-crystallized with a drug-like molecule were kept. The sole protein structure chosen to represent each target had to be of the same species as that used in the corresponding high-throughput screen, must not

be bound to any DNA fragment or cofactor other than ATP (to avoid the possibility of multiple binding sites), and had the highest resolution ($< 3 \text{ \AA}$) as well as the fewest missing atoms among the available structures on the Protein Data Bank.¹⁶ Only 19 bioassays, both primary ($n = 7$) and confirmatory ($n = 12$), related to seven protein targets were retained. Molecules having been identified as active in primary assays but not validated by confirmatory screens were all discarded from the active ligand sets. The remaining active compounds were then subject to the Hill Slope filter (which takes inspiration from the MUV database) and the pan-assay interference compounds (PAINS) filter⁸⁴⁻⁸⁹ to eliminate potential false positives. In the end, seven medium- and large-sized data sets ($> 59,000$ compounds) were constructed, with active-to-inactive ratios ranging from 7×10^{-5} to 1×10^{-3} , indicating hit rates significantly lower than those commonly seen in other databases. It is observed that a large number of unique Bemis-Murcko scaffolds are present among the active molecules (1.4 compounds/scaffold), implying that there is little analogue bias and substantial structural diversity in the active set composition. Though no direct measure was taken to reduce artificial enrichment due to differences between the true actives and true inactives, retrospective virtual screening on the seven final data sets using physicochemical property similarity searches (1D approach) and molecular docking was carried out, suggesting that the docking performance was not based on artificial enrichment, as the 1D method gave much lower enrichment in true actives than the structure-based approach in most cases. The Lindh *et al.* data collection is therefore considered the next remarkable step towards employing experimental input from PubChem BioAssay to build realistic data sets suitable for both ligand-based and structure-based *in silico* screening evaluations while addressing (and avoiding, to a considerable extent) most issues inherent in many other databases, including false positives, analogue bias and artificial enrichment. However, due to the unreasonably rigorous data quality filters that were applied during the construction of this data collection, the quantity of target sets offered by the authors is relatively small (only seven), and several important protein families that have been largely investigated by biochemists, e.g. GPCRs, nuclear receptors, are neglected (only two kinases were included in the database).

3.5. The LIT-PCBA Data Collection

Five years later, we (Tran-Nguyen *et al.*) developed and introduced a novel data collection entitled LIT-PCBA.²² A rigorous systematic search was first performed on the ensemble of

PubChem bioactivity assays, keeping only confirmatory screens conducted with over 10,000 substances, giving no fewer than 50 active molecules, against a single protein target having at least one crystal PDB structure bound to a drug-like ligand of the same phenotype as that of the confirmed actives. A total of 21 assays corresponding to 21 targets covering 11 diverse protein families, including three GPCRs, three kinases and five nuclear hormone receptors, were retained. Contrary to the data sets of Lindh *et al.*, in LIT-PCBA, all relevant protein-ligand structures available on the Protein Data Bank were kept, providing 162 “templates” in total. Taking inspiration from the MUV paper, we also addressed the issues of false positives, artificial enrichment and analogue bias during the construction of the LIT-PCBA data sets. The active and inactive substances retrieved from PubChem BioAssay were subjected to a series of consecutive filters, which ruled out inorganic chemicals (bearing at least one atom other than H, C, N, O, P, S, F, Cl, Br, and I), frequent hitters, non-specific binders, promiscuous aggregators, spoilers of optical detection methods, compounds with extreme molecular properties, and ligand preparation failures. Physicochemical differences between active and inactive substances were mitigated, as all molecular properties of the remaining ligands were kept within the same range, thus avoiding the presence of molecules that are too different from others in terms of physicochemical features. Retrospective virtual screening by ligand-based methods (2D fingerprint similarity searches and 3D shape-matching) on the resulting data collection confirmed that there was indeed little chemical bias in the composition of the ligand sets, as both approaches generally gave comparable performances to random selection. Results from molecular docking were also considered along with those of the two ligand-based approaches, leading to the selection of 15 small- to large-sized target sets (4247-362,088 molecules) that finally constituted the LIT-PCBA collection. Active-to-inactive ratios span over a relatively wide range from 5×10^{-5} to 0.05, but are below 3×10^{-3} in most cases, implying smaller hit rates than those of many other databases. Moreover, active substances included in LIT-PCBA are generally less potent than those found in DUD-E and ChEMBL, which imposes a more difficult challenge for *in silico* screening. Each ligand set was then further unbiased by the asymmetric validation embedding method (AVE),⁷³ yielding validation and training subsets with minimized overall bias that are ready for benchmarking novel virtual screening procedures. To the best knowledge of the authors, LIT-PCBA is now the latest attempt at constructing realistic data sets from confirmatory PubChem BioAssay data, possessing numerous advantages. Firstly, a large variety of protein targets

(including heavily researched ones) are featured in the collection and all available PDB structures are accounted. This practice takes into consideration at the same time the entire chemical diversity of known target-bound ligands and the complete conformational space accessible to the investigated target. Secondly, assay artifacts, chemical bias as well as potency bias in the composition of ligand sets were avoided or reduced, preventing possible overestimation of *in silico* screening performances. Thirdly, the eventual data-unbiasing step based on chemical space analyses offers a rational split of every existing set of molecules (instead of the random division that was previously observed in the UCI repository design). This further ensures the absence of both obvious and hidden bias in the final data sets. And lastly, thanks to the presence of at least one high-quality 3D structure with well-defined binding site(s) that represents each protein target, and the aforementioned chemically unbiased ligand set composition, the application of LIT-PCBA is thus not intended only for evaluating ligand-based or structure-based virtual screening alone, but rather for both, and especially for the field of machine learning algorithm development. There exist, however, some limitations in the design of this data collection, such as the relatively high hit rates of some ligand sets (2-5%), or the number of remaining true actives for several targets that is quite small (tens of molecules) for *in silico* methods to give any meaningful result. The current situation, as a consequence, still leaves plenty of room for further improvement, and more data sets based on experimental bioactivity assays are encouraged to be constructed, with inspirations taken from the existing collections mentioned above, to offer more realistic sets of molecules that mimic those employed in actual high-throughput screening campaigns, and to provide better evaluation tools for novel virtual screening approaches.

4. Note-Worthy Issues with Using Data from PubChem BioAssay for Constructing Benchmarking Data Sets

As demonstrated in the literature and the previous section, data retrieved from PubChem BioAssay may be used for various purposes in cheminformatics-related research, including benchmarking data set construction. Due to the availability of a wide range of assays with diverse ligand sets that the database offers, it is important to be conscious of all the issues that may arise regarding the usage of such large data,^{22,80,83} in terms of assay selection and data curation, to properly employ these abundant resources.

4.1. Assay Selection for Evaluating Virtual Screening Methods

4.1.1. Assay Selection as Regards the Data Size and Hit Rates

One of the first questions that we have to face when using data from the PubChem BioAssay repository to build benchmarking data sets concerns which assay(s) that should be chosen. As mentioned earlier in the manuscript, as of April 30, 2020, there were over a million assays deposited on the database. However, only a few of them can be deemed suitable for method evaluation purposes. There are many factors that one should consider before deciding which assay(s) to use. We herewith propose, as primary conditions to filter out unsuitable assays, the selection of only small-molecule HTS assays yielding biologically active molecules. RNAi assays, on the other hand, were conducted on microRNA-like molecules comprising twenties of base pairs that violate most drug-likeness rules of thumb, and are therefore, not of great interest in small-molecule drug discovery. For the sake of having an acceptable amount of ligands in the data that may give meaningful retrospective evaluations of *in silico* screening methods, we recommend that only assays with no fewer than 10 actives selected among at least 300 tested substances should be kept. Data sets including only nine or fewer actives are considered too small and would be over-challenging for virtual screening, especially for machine learning algorithms to learn anything meaningful. On the other hand, assays conducted with fewer than 300 substances while yielding more than 10 actives give hit rates that are deemed too high in comparison to those typically observed in experimental screening decks,²² even higher than those of existing data sets such as DUD,²⁸ DUD-E,²⁹ or DEKOIS 2.0.³⁰ There may exist, of course, assays with high hit rates that remain after this initial check (e.g. AIDs 1, 3, 720690, 720697); however, the aforementioned conditions are proposed to demonstrate that there is only a very small portion of available PubChem assays (0.20%) whose data may be considered for evaluating virtual screening protocols (**Figure 3**). The ligand sets of the remaining assays need to be further examined, and may be filtered, to ensure that their hit rates are as close as possible to those of experimental HTS campaigns, and that they are suitable for the nature of the screening method (ligand-based or structure-based).

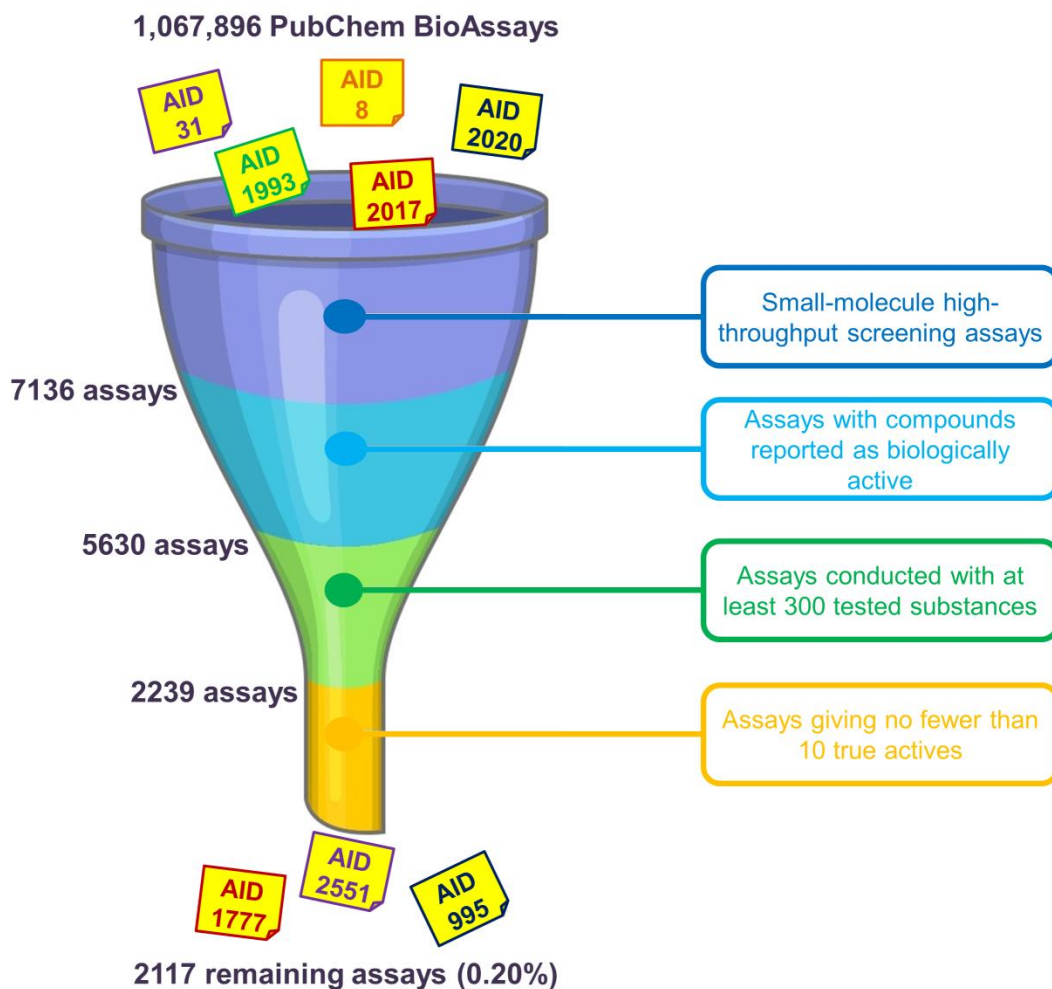


Figure 3. Primary selection of PubChem assays whose ligand sets should be further considered for evaluating virtual screening methods. We herewith recommend the use of only small-molecule HTS assays giving at least 10 biologically active molecules among no fewer than 300 tested substances. Overall, there are only 2117 assays (0.20% of 1,067,896 assays in total as of April 30, 2020) that remain, indicating a very small portion of PubChem assays that may be considered after this initial check.

4.1.2. Assay Selection as Regards the Nature of Virtual Screening

As demonstrated in various papers, a ligand set may be appropriate for evaluating only ligand-based *in silico* approaches,^{81,82} or only structure-based methods,⁷⁶ or sometimes both.^{22,80,83} This depends on the quantity and the chemical composition of all molecules that constitute the data set, the availability and the quality of 3-dimensional structures of relevant protein targets, as well as the definition of binding site(s) in which active substances exert their bioactivity. Data sets retrieved from the PubChem BioAssay database, being no exception, have to be thoroughly

examined according to the criteria mentioned above before being used to assess a certain virtual screening method. Ideally speaking, an assay whose ligands are considered for evaluating structure-based approaches needs to be conducted on a protein target whose structure has been solved at a high resolution, with no ambiguity in terms of electron density, with at least a molecule of the same phenotype (agonist, antagonist, inhibitor, etc.) as that of the active compounds. However, targets for which no crystallographic or electron-microscopic structure is deposited on the Protein Data Bank may also be considered, if high-quality homology models are available. An example of this can be seen in the assay AID 588606, featuring inhibitors of the yeast efflux pump Cdr1. Though the protein target, the ABC drug resistance protein 1 of *Candida albicans* (CaCdr1p), has not yet been available on the Protein Data Bank with a known inhibitor, a homology model of this transporter was generated using the human ABCG5/G8 crystal structure as template, and possible binding sites located in the transmembrane domain were identified and validated by means of atomic modeling and systematic mutagenesis, confirming their essential role in Cdr1p-induced multidrug resistance.⁹⁰ However, caution should be taken when one uses such artificially constructed models as input for structure-based screening approaches. On the other hand, the presence of many non-overlapping binding sites (orthosteric versus allosteric) in the 3D structures of protein targets (as observed in those of AIDs 1469, 624170, or 624417), either crystallographic or not, may ultimately become a reason for failures in screening PubChem molecules on such proteins, especially when there is no information on the exact binding site of the tested substances that can be deduced from the assay description.²² As virtual screening performances may vary quite significantly depending on the protein structure employed as input,²² one should therefore be cautious when using data of these assays for evaluating structure-based screening procedures, lest they give poorer performances than expected due to external reasons that are not related to the methods themselves. Another point that should not be overlooked concerns assays that were conducted on substances derived from only a few chemical series, as they may give rise to bias that overestimates screening performance, notably that of ligand-based approaches. If another similar assay on the same target but with a more diverse ligand set (in terms of chemical features) is available, one is recommended to make use of this assay instead. Otherwise, the “biased” data need further tuning to be deemed suitable for evaluation purposes, e.g. by filtering out “redundant” compounds (this point will be thoroughly discussed in the next section of this manuscript). However, this ligand-

filtering process should not lower the number of active substances to a value so small that ligand-based methods or machine learning algorithms cannot come up with meaningful results.

4.1.3. Assay Selection as Regards the Screening Stage

Additionally, the use of data from “primary assays” should be subject to caution, as the activity outcome was only determined at a single concentration, and has not yet been validated on the basis of a dose-response relationship with multiple tested concentrations,^{3,91} hence the potency values of active molecules are not confirmed. As a matter of fact, some substances originally deemed as active in a primary assay may be denounced as inactive by a subsequent confirmatory screen, as seen in AIDs 449 and 466, or AIDs 524 and 548. We therefore recommend that primary screening data should only be used if there exists a confirmatory assay that validates the potency of the selected active molecules. This practice was already observed in the construction of the MUV data sets by Rohrer and Baumann,⁸⁰ in which pairs of primary and corresponding confirmatory screens were employed, whose data were then combined to form the final ligand sets. In this manner, the large pool of inactive substances from the primary assay is not neglected, and the bioactivities of the confirmed hits are indeed guaranteed, affording a vast data set (usually implying a low hit rate) with fully validated active components. Otherwise, output data of primary screens alone should be used with great caution, due to the risk of assuming “false positives” that may later falsify virtual screening outcomes. An exhaustive search on the whole PubChem BioAssay database is therefore of paramount importance to select relevant data sets for retrospective assessments of *in silico* screening protocols in order to ensure the quality of such evaluations.

4.2. Detecting False Positives among Active Substances

Concerns have long been raised over the presence of chemical-induced artifacts in screening experiments, leading to false-positive findings among the molecules deemed as active.^{22,80,83-89,92}

Misinterpretation of assay results and subsequent inaccurate conclusions may stem from various reasons largely discussed in the literature. Among them are off-target effects of compounds exerting unspecific bioactivities, possible biological target precipitation by organic chemicals aggregation, inherent fluorescent properties of substances that interfere with fluorescence emission detection methods, or luciferase inhibitory activities of molecules that spoil light

emission measurement in reporter gene assays.⁸⁰ Active substances whose modes of action are subject to the aforementioned issues must therefore be removed from PubChem BioAssay ligand sets before the data can be used for retrospective virtual screening purposes. Rohrer and Baumann (2009) addressed this problem during the construction of their MUV data sets from the database, designing a so-called “assay artifacts filter” aiming to eliminate all active ligands that likely become false positives, thus prevent them from affecting subsequent screening performances. The filter is composed of three filtering “layers”, including (i) the “Hill slope filter” after which actives whose Hill slopes for the dose-response curves are lower than 0.5 or higher than 2 are eliminated, (ii) the “Frequency of hits filter” that keeps only the molecules deemed as active in no more than 26% of the bioactivity assays in which they were tested, and (iii) the “Auto-fluorescence and luciferase inhibition filter” that rules out compounds exhibiting auto-fluorescent properties along with inhibitors of luciferase.⁸⁰ All frequent hitters, unspecific binders (molecules with multiple binding sites), experimentally determined aggregators, and spoilers of optical detection methods are, as a result, removed from the PubChem data sets after these filtering steps. Such filters indeed have a profound impact on the population of active substances, as over a half of them were deleted by these “false positives filters” during the development of our recently introduced LIT-PCBA data set (**Figure 4**).²² This drastic decrease in the number of confirmed actives also helps lower the “hit rates” observed in our ligand sets (as only the actives were subjected to these filters), thus bringing them closer to those typically reported in high-throughput screening decks in reality, and lower than those of artificially constructed data sets such as DUD,²⁸ DUD-E,²⁹ or DEKOIS 2.0.³⁰ This not only denotes the particular challenge brought by our data set, but also highlights the importance of detecting, and removing false positives in assembling active substances.

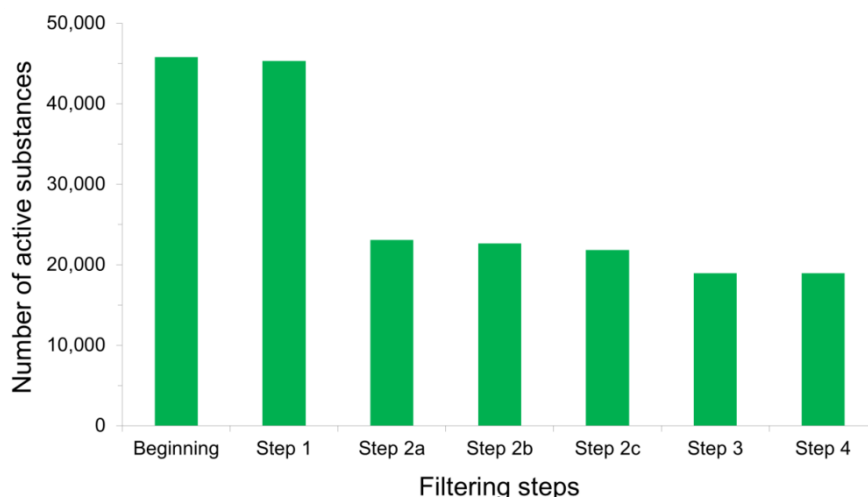


Figure 4. Total number of active substances that remained after each filtering step was applied to PubChem BioAssay ligands during the construction of the LIT-PCBA data set:²² Step 1 – inorganic molecules; Step 2a – actives with Hill slopes < 0.5 or > 2 ; Step 2b – actives with frequency of hits > 0.26 ; Step 2c – actives found among 10,892 confirmed aggregators, luciferase inhibitors or auto-fluorescent molecules; Step 3 – substances with extreme molecular properties; Step 4 – 3D conversion and ionization failures. It can be observed that the sole step 2a removed the most active molecules (over 50% of them), thus significantly reducing the population of true actives in comparison to that of true inactives.

4.3. Possible Chemical Bias in Assembling Active and Inactive Substances

As previously mentioned, a note-worthy issue of raw data published on PubChem BioAssay lies in the chemically biased composition of active and inactive substances for a particular target. More specifically, there may exist “analogue bias”⁹³ present among the molecules constituting a ligand set, which likely leads to overly good performances of virtual screening methods. This bias is generally observed in data collections whose actives (or inactives) share similar chemical features, meaning a large number of these molecules are issued from the same (or similar) scaffolds.⁷⁶ As ligand-based and structure-based screening methods tend to recognize compounds of the same chemical series, such bias may result in an overestimation of *in silico* screening performance.⁷⁶ Besides, significant differences between active and inactive molecules, in terms of physicochemical properties, such as molecular mass, octanol-water partition coefficient, or atomic formal charge, may as well be the source of artificial enrichment.⁸⁰ Raw experimental data from PubChem BioAssay therefore need to be finely tuned before further use, by filtering out most compounds representing the same scaffold while ensuring that the

physicochemical parameters of all included molecules are kept within the same range, so that chemical bias, if there were any, in the ligand set would be reduced.⁷⁶ An example of the importance of filtering input data can be seen in the MTORC1 ligand set (**Figure 5**) included in our recently introduced LIT-PCBA data collection,²² comprising the molecules tested for an inhibitory activity towards the mTORC1 signaling pathway, targeting the human serine/threonine-protein kinase mTOR.

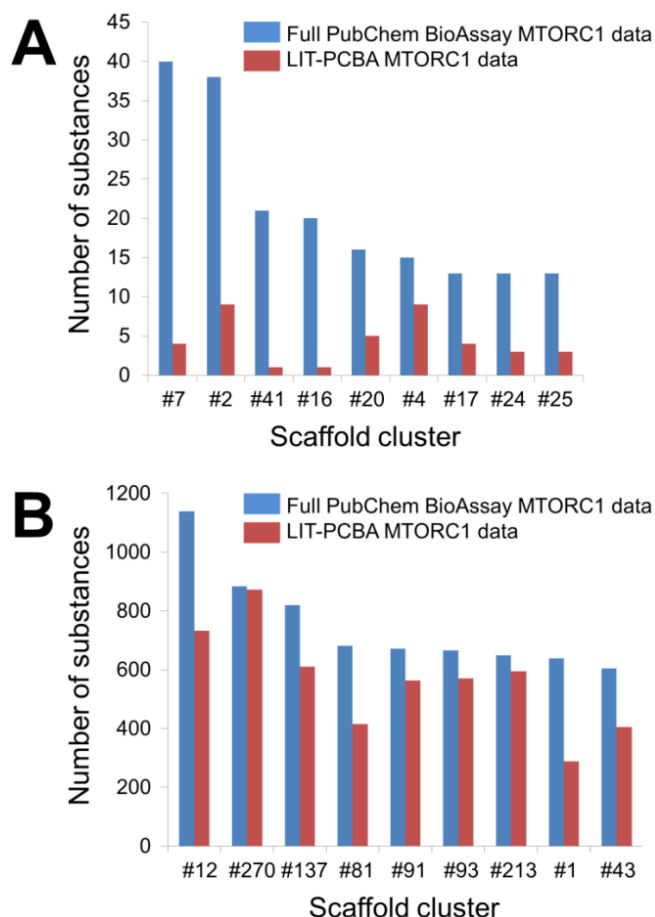


Figure 5. Number of substances falling into each scaffold cluster that includes more than 10 true active molecules (**A**) or 600 true inactive molecules (**B**). Bemis-Murcko frameworks derived from the input molecules were first created by trimming each active and each inactive separately with Pipeline Pilot 19.1.0.1964.^{94,95} A hierarchical scaffold tree consisting of canonical SMILES strings that represent the rings, linkers and double bonds in each molecule was next generated according to an iterative ring-trimming procedure described by Schuffenhauer *et al.* (2007).⁹⁶ All ligands were then clustered based on the smallest scaffold at the root of the scaffold tree for each ligand. The number that follows each hash symbol indicated in this figure refers to the ordinal number of a scaffold cluster as issued by Pipeline Pilot. Details of all clusters can be found in Supporting Information.

As to be expected, the full PubChem BioAssay data feature a larger number of scaffold clusters, with 59 clusters for the active set and 1151 clusters for the inactive set (against 41 and 1106 clusters in the LIT-PCBA active and inactive ligand sets, respectively). However, only 18 (out of 342, 5.26%) true actives possess unique scaffolds, meaning nearly 95% of all active substances in the full PubChem ligand set share chemical similarities with at least another active. Notably, nine clusters are reported to have more than 10 representatives (**Figure 5A, Table S3**). The pruned LIT-PCBA active ligand set, on the other hand, includes no cluster with over 10 members and 21 clusters (51.22%) with only one substance for each. This means nearly a quarter of LIT-PCBA active molecules (over four times the value observed in the full PubChem set) possess unique scaffolds. Moreover, the number of ligands falling into each cluster in the filtered LIT-PCBA active set is greatly reduced in comparison to that of the unfiltered data (**Figure 5A, Table S3**). On the other hand, around 25% of PubChem molecules were deemed to have extreme physicochemical properties and were therefore discarded as the MTORC1 ligand set was constructed.²² These observations suggest that (i) there is indeed significant chemical bias in the full PubChem active ligand composition; and (ii) the filtering steps that were applied to build the LIT-PCBA data collection helped reduce this bias by lowering the number of active substances sharing the same chemical features (thus avoiding the presence of too many molecules issued from the same chemotype), and by ruling out compounds that were too different from others (hence preventing artificial enrichment). A similar conclusion can be drawn from the full PubChem inactive ligand set and the corresponding LIT-PCBA data (**Figure 5B, Table S4**). The benefit of filtering PubChem ligands in reducing chemical bias is again highlighted as the data sets undergo a subsequent unbiasing procedure using the previously described asymmetric validation embedding (AVE) method,⁷³ which measures pairwise distances in chemical space between molecules belonging to four sets of compounds (training actives, training inactives, validation actives, validation inactives, training-to-validation ratio = 3) based on ECFP4.⁹⁷ A nearly zero overall bias value (0.001) was obtained from the LIT-PCBA MTORC1 ligand set after only seven iteration steps of the AVE genetic algorithm (GA),²² while 16 GA iterations were necessary to bring the overall bias of the full PubChem set down to 0.006. This denotes that the pruned LIT-PCBA ligands are much less biased in terms of chemical features than the complete PubChem molecules, and confirms the necessity of detecting chemical bias in PubChem BioAssay data and removing it so that the data set is better adapted for further use.

The impact of filtering PubChem BioAssay molecules on subsequent retrospective screening performance can also be observed with the use of two *in silico* methods: 2D similarity searches using extended-connectivity ECFP4 fingerprints⁹⁷ with Pipeline Pilot⁹⁵ (ligand-based) and molecular docking with Surflex-Dock (structure-based).⁹⁸ Both data sets (the full PubChem data and the pruned LIT-PCBA MTORC1 ligands) underwent the same screening protocols using the two aforementioned programs as described in our previous paper.²² Screening performance is evaluated according to the EF1% (enrichment in true actives at a constant 1% false positive rate over random picking) values obtained by the “max-pooling” approach, taking into account all available PDB templates of the protein target ($n = 11$), while generating only one hit list that facilitates post-screening assessments.²² It is observed that both methods performed better on the full PubChem data than on the filtered LIT-PCBA ligand set (**Table 2**). Interestingly, the true actives that were retrieved along with the top 1% false positives belong to the same scaffold clusters, or to clusters that are similar to each other. Such observations reconfirm that (i) ligand-based and structure-based screening approaches tend to recognize compounds that share chemical features, and (ii) the chemical bias present in the complete PubChem data indeed leads to over-optimistic screening performances. This, again, highlights the importance of filtering the ensemble of molecules deposited on PubChem BioAssay prior to evaluating virtual screening procedures, first to reduce chemical bias in the composition of the data, then to avoid overestimating the real discriminatory accuracy of *in silico* methods.

Table 2. Retrospective screening performance of 2D ECFP4 fingerprint similarity searches with Pipeline Pilot (ligand-based) and molecular docking with Surflex-Dock (structure-based) on the full PubChem BioAssay data and the pruned LIT-PCBA MTORC1 ligand set, demonstrated by EF1% (enrichment in true actives at a constant 1% false positive rate over random picking) values and the numbers of true actives retrieved along with the top 1% false positives by the “max-pooling” approach.

	2D ECFP4 fingerprint similarity searches		Molecular docking	
	EF1%	Number of retrieved actives	EF1%	Number of retrieved actives
Full PubChem data	0.6	2	3.2	11
LIT-PCBA MTORC1 data	0.0	0	1.0	1

4.4. Potency Bias in the Composition of Active Ligand Sets

As of April 30, 2020, there were 1,067,719 small-molecule assays deposited on the PubChem BioAssay database, but only 240,999 of them (22.6%) yielded active substances with confirmed potency values. These values are provided in different terms (EC_{50} , IC_{50} , K_d , K_i), and the threshold to distinguish true actives from true inactives varies from assay to assay, depending on the researchers who conducted the experiments. Some assays accept active substances with potency values above 100 μ M (e.g. AIDs 1030, 1490, 504847), even at millimolar level (e.g. AIDs 1045, 1047); while in some others, several substances with even sub-micromolar potency are not deemed actives (e.g. AIDs 1221, 1224, 1345010). It is therefore comprehensible that the potency range of true actives as well as its distribution is quite diverse across all assays of PubChem. As active molecules with high potency towards a biological target are easier to be picked by both ligand-based and structure-based virtual screening methods,²² ligand sets with too many actives whose potency values are in the sub-micromolar range are prone to overestimate the real accuracy of *in silico* screening. PubChem BioAssay data sets, especially those composed of highly potent true actives (potency below 1 μ M), need to be filtered so that the so-called “potency bias” in the composition of their active ligand sets is reduced before further use.

An illustration of this point can be taken from the LIT-PCBA PPARG ligand set (27 true actives and 5211 true inactives) and the corresponding full PubChem BioAssay data (AID 743094, 78 true actives, 8532 true inactives) comprising small molecules that were tested for an agonistic activity on the peroxisome proliferator-activated receptor gamma (PPAR γ) signaling pathway.²² The number of true actives with high potency ($EC_{50} < 1 \mu$ M) in the complete PubChem data is 19, nearly three times higher than that of the pruned LIT-PCBA ligand set ($n = 7$). Upon carrying out 2D similarity searches with Pipeline Pilot using ECFP4 fingerprints and ten structurally diverse crystallographic PPAR γ agonists randomly chosen from 138 available structures on the Protein Data Bank as templates, it is observed that, as expected, the screening protocol managed to retrieve more highly potent true actives from the full data set than from the filtered ligand set in 70% of the cases (**Figure 6**). Moreover, the “max-pooling” approach, when applied to the complete PubChem data, selected seven highly potent actives among the top 1%-ranked molecules, seven times higher than the amount obtained from LIT-PCBA. Among them, four even have potency values below 0.1 μ M. The same screening method, on the other hand, failed

to retrieve any true active with $EC_{50} < 0.1 \mu\text{M}$ from the pruned PPARG data. The screening performance observed on the full ligand set is, as a result, better than that obtained after ligand-filtering, as the EF1% value is nearly twice higher than that received with LIT-PCBA ligands. This reconfirms that *in silico* screening procedures tend to recognize molecules with high potency towards a protein target, and the presence of too many highly potent ligands in the data likely leads to a better screening performance. It is therefore recommended that one should filter the ensemble of PubChem BioAssay ligands to ensure that there are not too many true actives with high potency that remain, in order to avoid possible “potency bias” in the data set and the subsequent overestimation of *in silico* methods’ discriminatory power.

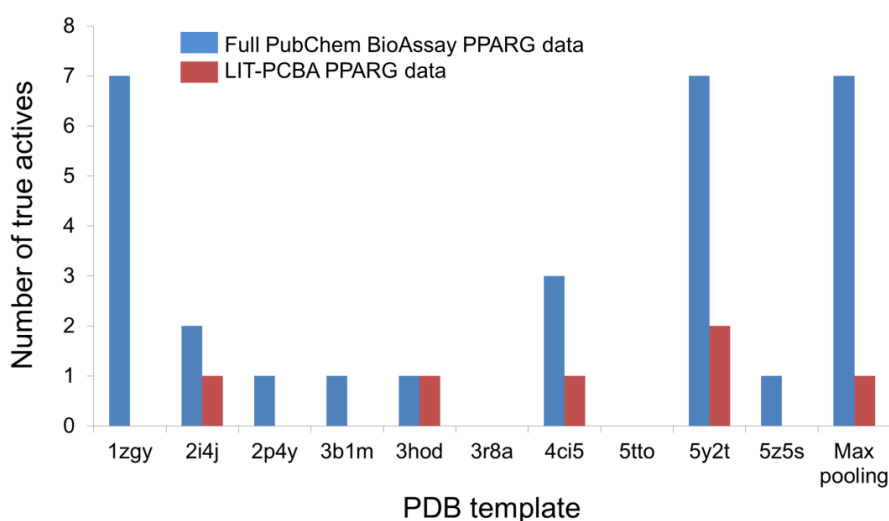


Figure 6. The number of highly potent true actives ($EC_{50} < 1 \mu\text{M}$) retrieved among the top 1%-ranked molecules by 2D ECFP4 fingerprint similarity searches from the full PubChem BioAssay data and the corresponding LIT-PCBA PPARG ligand set after ligand-filtering. Ten known crystallographic PPARG agonists were randomly chosen as templates from 138 available structures on the Protein Data Bank.

4.5. Processing Input Structures Prior to Virtual Screening

PubChem BioAssay ligands, as deposited on the database, can be downloaded either as SMILES strings,⁹⁹ or in 2D SDF format,¹⁰⁰ and are therefore, in general, not yet ready to be directly employed as input for most *in silico* screening protocols (except for 1D or 2D ligand-based approaches). A rigorous ligand-processing procedure is thus necessary to afford ready-to-use structures for virtual screening. This process concerns a wide range of aspects inherent in the 3-dimensional structural formula of a molecule, including atomic coordinates in 3D space, formal

charge assigned on each atom, the presence of different protonation states and tautomeric shifts that slightly alter the structure, the representation of undefined stereocenters or flexible rings, as well as the existence of multiple conformations and/or configurations.¹⁰¹ Various studies have concluded that database-processing has indeed an impact on screening performance, some processing stages are even indispensable to certain programs.¹⁰¹⁻¹⁰⁴ Kellenberger *et al.* (2004),¹⁰³ Perola and Charifson (2004),¹⁰⁴ and Cummings *et al.* (2007)¹⁰¹ pointed out that the initial conformation and orientation in 3D space of a molecule, which are determined based on details featured in the original SMILES string, may significantly affect the final enrichment output by a docking program. Performance of structure-based screening methods whose scoring functions rely on ligand-receptor interactions^{105,106} may be sensitive to a change in explicit hydrogen assignment or protonation states, as the positions of hydrogen-bonding groups and proton-carrying atoms are crucial to properly detecting intermolecular hydrogen bonds and ionic interactions, respectively.^{101,107} While a generation of correct multiple conformers for a molecule is not imperative when it comes to carrying out docking with GOLD¹⁰⁸ or Surflex-Dock,⁹⁸ this step has in fact a pivotal role in 3D shape similarity searches using ROCS (OpenEye).¹⁰⁹ The examples mentioned above denote that good *in silico* screening outcomes do require careful treatment of input ligand sets, and a thorough investigation of different data-processing procedures with commonly used programs (e.g. Protoss,¹¹⁰ Corina,¹¹¹ MOE,¹¹² Sybyl,¹¹³ Daylight¹¹⁴) is thus recommended. If it is possible (if the data size is not too large), one should check each output structure by hand to ensure that the assigned atom types, bond types, stereochemical properties and protonation states are correct before further use. This also applies to protein structure preparation prior to screening, as structural features of the protein target, especially those of the binding site, are of indisputable importance to structure-based virtual screening performance.

5. Conclusion

Retrieving experimental PubChem BioAssay data to construct novel data sets for virtual screening evaluations helps avoid assuming false negatives among inactive ligands, which is a problem inherent in artificially developed data collections. However, there remain several issues regarding assay selection, false active molecules, chemical bias and potency bias, as well as data curation that are worth noticing prior to employing PubChem input for database-designing

purposes. To the best of our knowledge, there have been several publicly available data sets that were constructed from the data deposited on this repository, but the quantity is not yet considerable, and there still exist some limitations in the design of these data collections. More effort in this regard is recommended, with the points raised in this manuscript taken into account, in order to offer more realistic data sets suitable for validating both ligand-based and structure-based *in silico* screening procedures in the future. Of course, the herein proposed good practices should also be applied to proprietary bioactivity data.

References

1. Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Bryant, S. H. PubChem: A Public Information System for Analyzing Bioactivities of Small Molecules. *Nucleic Acids Res.* **2009**, *37*, W623-W633.
2. Li, Q.; Cheng, T.; Wang, Y.; Bryant, S. H. PubChem as a Public Resource for Drug Discovery. *Drug Discov. Today* **2010**, *15*, 1052-1057.
3. Wang, Y.; Bolton, E.; Dracheva, S.; Karapetyan, K.; Shoemaker, B. A.; Suzek, T. O.; Wang, J.; Xiao, J.; Zhang, J.; Bryant, S. H. An Overview of the PubChem BioAssay Resource. *Nucleic Acids Res.* **2010**, *38*, D255-266.
4. Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Zhou, Z.; Han, L.; Karapetyan, K.; Dracheva, S.; Shoemaker, B. A.; Bolton, E.; Gindulyte, A.; Bryant, S. H. PubChem's BioAssay Database. *Nucleic Acids Res.* **2012**, *40*, D400-412.
5. Wang, Y.; Cheng, T.; Bryant, S. H. PubChem BioAssay: A Decade's Development toward Open High-Throughput Screening Data Sharing. *SLAS Discov.* **2017**, *22*, 655-666.
6. Austin, C. P.; Brady, L. S.; Insel, T. R.; Collins, F. S. NIH Molecular Libraries Initiative. *Science* **2004**, *306*, 1138-1139.
7. Cheng, T.; Pan, Y.; Hao, M.; Wang, Y.; Bryant, S. H. PubChem Applications in Drug Discovery: A Bibliometric Analysis. *Drug Discov. Today* **2014**, *19*, 1751-1756.
8. Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. PubChem Substance and Compound databases. *Nucleic Acids Res.* **2016**, *44*, D1202-D1213.
9. PubChem Data Sources. <https://pubchem.ncbi.nlm.nih.gov/sources/> (accessed April 2020).
10. PubChem Classification Browser. <https://pubchem.ncbi.nlm.nih.gov/classification/#hid=80/> (accessed April 2020).
11. PubChem Data Counts. <https://pubchemdocs.ncbi.nlm.nih.gov/statistics/> (accessed April 2020).
12. Kim, S.; Thiessen, P. A.; Cheng, T.; Yu, B.; Shoemaker, B. A.; Wang, J.; Bolton, E. E.; Wang, Y.; Bryant, S. H. Literature Information in PubChem: Associations between PubChem Records and Scientific Articles. *J. Cheminform.* **2016**, *8*, 32.
13. PubChem BioAssay. <https://www.ncbi.nlm.nih.gov/pcassay/> (accessed April 2020).
14. Entrez Programming Utilities Help. <https://www.ncbi.nlm.nih.gov/books/NBK25501/> (accessed April 2020).

15. PubMed Central. <https://www.ncbi.nlm.nih.gov/pmc/> (accessed May 2020).
16. 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.
17. Wang, Y.; Suzek, T.; Zhang, J.; Wang, J.; He, S.; Cheng, T.; Shoemaker, B. A.; Gindulyte, A.; Bryant, S. H. PubChem BioAssay: 2014 Update. *Nucleic Acids Res.* **2014**, *42*, D1075-D1082.
18. Wang, Y.; Bryant, S. H.; Cheng, T.; Wang, J.; Gindulyte, A.; Shoemaker, B. A.; Thiessen, P. A.; He, S.; Zhang, J. PubChem BioAssay: 2017 Update. *Nucleic Acids Res.* **2017**, *45*, D955-D963.
19. Kim, S.; Thiessen, P. A.; Cheng, T.; Yu, B.; Bolton, E. E. An Update on PUG-REST: RESTful Interface for Programmatic Access to PubChem. *Nucleic Acids Res.* **2018**, *46*, W563-570.
20. Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem 2019 Update: Improved Access to Chemical Data. *Nucleic Acids Res.* **2019**, *47*, D1102-1109.
21. About PubChem. <https://pubchemdocs.ncbi.nlm.nih.gov/about/> (accessed April 2020).
22. 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** (in press).
23. Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Deliv. Rev.* **1997**, *23*, 3-25.
24. Lipinski, C. A. Lead- and Drug-Like Compounds: the Rule-of-Five Revolution. *Drug Discov. Today Technol.* **2004**, *1*, 337-341.
25. Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. A Knowledge-Based Approach in Designing Combinatorial or Medicinal Chemistry Libraries for Drug Discovery. 1. A Qualitative and Quantitative Characterization of Known Drug Databases. *J. Comb. Chem.* **1999**, *1*, 55-68.
26. Veber, D. F.; Johnson, S. R.; Cheng, H. Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular Properties that Influence the Oral Bioavailability of Drug Candidates. *J. Med. Chem.* **2002**, *45*, 2615-2623.
27. Pyka, A.; Babuska, M.; Zachariasz, M. A Comparison of Theoretical Methods of Calculation of Partition Coefficients for Selected Drugs. *Acta. Pol. Pharm.* **2006**, *63*, 159-167.
28. Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, *49*, 6789-6801.
29. 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.
30. Bauer, M. R.; Ibrahim, T. M.; Vogel, S. M.; Boeckler, F. M. Evaluation and Optimization of Virtual Screening Workflows with DEKOIS 2.0 – A Public Library of Challenging Docking Benchmark Sets. *J. Chem. Inf. Model.* **2013**, *53*, 1447-1462.
31. Wassermann, A. M.; Lounkine, E.; Hoepfner, D.; Le Goff, G.; King, F. J.; Studer, C.; Peltier, J. M.; Grippo, M. L.; Prindle, V.; Tao, J.; Schuffenhauer, A.; Wallace, I. M.; Chen, S.; Krastel, P.; Cobos-Correa, A.; Parker, C. N.; Davies, J. W.; Glick, M. Dark chemical matter as a promising starting point for drug lead discovery. *Nat. Chem. Biol.* **2015**, *11*, 958-966.

32. PubChem BioAssay “Limits” Search. <https://www.ncbi.nlm.nih.gov/pcassay/limits> (accessed April 2020).
33. PubChem BioAssay “Advanced” Search. <https://www.ncbi.nlm.nih.gov/pcassay/advanced> (accessed April 2020).
34. PubChem Power User Gateway (PUG) Help. <https://pubchemdocs.ncbi.nlm.nih.gov/power-user-gateway> (accessed April 2020).
35. PubChem PUG SOAP. <https://pubchemdocs.ncbi.nlm.nih.gov/pug-soap> (accessed April 2020).
36. PubChem PUG REST. <https://pubchemdocs.ncbi.nlm.nih.gov/pug-rest> (accessed April 2020).
37. PubChem PUG View. <https://pubchemdocs.ncbi.nlm.nih.gov/pug-view> (accessed April 2020).
38. PubChemRDF. <https://pubchemdocs.ncbi.nlm.nih.gov/rdf> (accessed April 2020).
39. ScrubChem by Jason Bret Harris. <http://scrubchem.org/> (accessed May 2020).
40. Kim, S. Getting the Most out of PubChem for Virtual Screening. *Expert Opin. Drug Discov.* **2016**, *11*, 843-855.
41. Kim, S.; Shoemaker, B. A.; Bolton, E. E.; Bryant, S. H. Finding Potential Multitarget Ligands Using PubChem. *Methods Mol. Biol.* **2018**, *1825*, 63-91.
42. Li, Q. Y.; Jorgensen, F. S.; Oprea, T.; Brunak, S.; Taboureau, O. hERG classification model based on a combination of support vector machine method and GRIND descriptors. *Mol. Pharm.* **2008**, *5*, 117-127.
43. Su, B. H.; Shen, M. Y.; Esposito, E. X.; Hopfinger, A. J.; Tseng, Y. J. In Silico Binary Classification QSAR Models Based on 4D-fingerprints and MOE Descriptors for Prediction of hERG Blockage. *J. Chem. Inf. Model.* **2010**, *50*, 1304-1318.
44. Wang, S.; Li, Y.; Wang, J.; Chen, L.; Zhang, L.; Yu, H.; Hou, T. ADMET Evaluation in Drug Discovery. 12. Development of Binary Classification Models for Prediction of hERG Potassium Channel Blockage. *Mol. Pharm.* **2012**, *9*, 996-1010.
45. Shen, M. Y.; Su, B. H.; Esposito, E. X.; Hopfinger, A. J.; Tseng, Y. J. A Comprehensive Support Vector Machine Binary hERG Classification Model Based on Extensive but Biased End Point hERG Data Sets. *Chem. Res. Toxicol.* **2011**, *24*, 934-949.
46. Cheng, F.; Yu, Y.; Shen, J.; Yang, L.; Li, W.; Liu, G.; Lee, P. W.; Tang, Y. Classification of Cytochrome P450 Inhibitors and Noninhibitors Using Combined Classifiers. *J. Chem. Inf. Model.* **2011**, *51*, 996-1011.
47. Su, B. H.; Tu, Y. S.; Lin, C.; Shao, C. Y.; Lin, O. A.; Tseng, Y. J. Rule-Based Prediction Models of Cytochrome P450 Inhibition. *J. Chem. Inf. Model.* **2015**, *55*, 1426-1434.
48. Didziapetris, R.; Dapkunas, J.; Sazonovas, A.; Japertas, P. Trainable Structure-Activity Relationship Model for Virtual Screening of CYP3A4 Inhibition. *J. Comput. Aided Mol. Des.* **2010**, *24*, 891-906.
49. Novotarskyi, S.; Sushko, I.; Korner, R.; Pandey, A. K.; Tetko, I. V. A Comparison of Different QSAR Approaches to Modeling CYP450 1A2 Inhibition. *J. Chem. Inf. Model.* **2011**, *51*, 1271-1280.
50. Buchwald, P. Activity-Limiting Role of Molecular Size: Size-Dependency of Maximum Activity for P450 Inhibition as Revealed by qHTS Data. *Drug Metab. Dispos.* **2014**, *42*, 1785-1790.
51. Zhu, H.; Rusyn, I.; Richard, A.; Tropsha, A. Use of Cell Viability Assay Data Improves the Prediction Accuracy of Conventional Quantitative Structure-Activity Relationship Models of Animal Carcinogenicity. *Environ. Health Perspect.* **2008**, *116*, 506-513.

52. Guha, R.; Schurer, S. C. Utilizing High Throughput Screening Data for Predictive Toxicology Models: Protocols and Application to MLSCN Assays. *J. Comput. Aided Mol. Des.* **2008**, *22*, 367-384.
53. Zhang, J.; Hsieh, J. H.; Zhu, H. Profiling Animal Toxicants by Automatically Mining Public Bioassay Data: A Big Data Approach for Computational Toxicology. *PLoS One* **2014**, *9*, 11.
54. Sedykh, A.; Zhu, H.; Tang, H.; Zhang, L.; Richard, A.; Rusyn, I.; Tropsha, A. Use of in Vitro HTS-derived Concentration-Response Data as Biological Descriptors Improves the Accuracy of QSAR Models of in Vivo Toxicity. *Environ. Health Perspect.* **2011**, *119*, 364-370.
55. Kim, M. T.; Huang, R.; Sedykh, A.; Wang, W.; Xia, M.; Zhu, H. Mechanism Profiling of Hepatotoxicity Caused by Oxidative Stress Using Antioxidant Response Element Reporter Gene Assay Models and Big Data. *Environ. Health Perspect.* **2016**, *124*, 634-641.
56. Zhu, H.; Zhang, J.; Kim, M. T.; Boison, A.; Sedykh, A.; Moran, K. Big Data in Chemical Toxicity Research: The Use of High-Throughput Screening Assays to Identify Potential Toxicants. *Chem. Res. Toxicol.* **2014**, *27*, 1643-1651.
57. Chen, B.; Wild, D.; Guha, R. PubChem as a Source of Polypharmacology. *J. Chem. Inf. Model.* **2009**, *49*, 2044-2055.
58. Zhang, J.; Han, B.; Wei, X.; Tan, C.; Chen, Y.; Jiang, Y. A Two-Step Target Binding and Selectivity Support Vector Machines Approach for Virtual Screening of Dopamine Receptor Subtype-Selective Ligands. *PLoS One* **2012**, *7*, e39076.
59. Swamidass, S. J.; Schillebeeckx, C. N.; Matlock, M.; Hurle, M. R.; Agarwal, P. Combined Analysis of Phenotypic and Target-Based Screening in Assay Networks. *J. Biomol. Screen.* **2014**, *19*, 782-790.
60. Lounkine, E.; Nigsch, F.; Jenkins, J. L.; Glick, M. Activity-Aware Clustering of High Throughput Screening Data and Elucidation of Orthogonal Structure-Activity Relationships. *J. Chem. Inf. Model.* **2011**, *51*, 3158-3168.
61. Bissantz, C.; Folkers, G.; Rognan, D. Protein-Based Virtual Screening of Chemical Databases. 1. Evaluation of Different Docking/Scoring Combinations. *J. Med. Chem.* **2000**, *43*, 4759-4767.
62. McGovern, S. L.; Shoichet, B. K. Information Decay in Molecular Docking Screens against Holo, Apo, and Modeled Conformations of Enzymes. *J. Med. Chem.* **2003**, *46*, 2895-2907.
63. Diller, D. J.; Li, R. Kinases, Homology Models, and High Throughput Docking. *J. Med. Chem.* **2003**, *46*, 4638-4647.
64. Lorber, D. M.; Shoichet, B. K. Hierarchical Docking of Databases of Multiple Ligand Conformations. *Curr. Top. Med. Chem.* **2005**, *5*, 739-749.
65. Irwin, J. J.; Raushel, F. M.; Shoichet, B. K. Virtual Screening against Metalloenzymes for Inhibitors and Substrates. *Biochemistry* **2005**, *44*, 12316-12328.
66. Miteva, M. A.; Lee, W. H.; Montes, M. O.; Villoutreix, B. O. Fast Structure-Based Virtual Ligand Screening Combining FRED, DOCK, and Surflex. *J. Med. Chem.* **2005**, *48*, 6012-6022.
67. Pham, T. A.; Jain, A. N. Parameter Estimation for Scoring Protein-Ligand Interactions Using Negative Training Data. *J. Med. Chem.* **2006**, *49*, 5856-5868.

68. Vogel, S. M.; Bauer, M. R.; Boeckler, F. M. DEKOIS: Demanding Evaluation Kits for Objective In Silico Screening – A Versatile Tool for Benchmarking Docking Programs and Scoring Functions. *J. Chem. Inf. Model.* **2011**, *51*, 2650-2665.
69. Gatica, E. A.; Cavasotto, C. N. Ligand and Decoy Sets for Docking to G Protein-Coupled Receptors. *J. Chem. Inf. Model.* **2012**, *52*, 1-6.
70. Lagarde, N.; Ben Nasr, N.; Jeremie, A.; Guillemain, H.; Laville, V.; Labib, T.; Zagury, J. F.; Montes, M. NRLiSt BDB, the Manually Curated Nuclear Receptors Ligands and Structures Benchmarking Database. *J. Med. Chem.* **2014**, *57*, 3117-3125.
71. Xia, J.; Tilahun, E. L.; Kebede, E. H.; Reid, T. E.; Zhang, L.; Wang, X. S. Comparative Modeling and Benchmarking Data Sets for Human Histone Deacetylases and Sirtuin Families. *J. Chem. Inf. Model.* **2015**, *55*, 374-388.
72. Chaput, L.; Martinez-Sanz, J.; Saettel, N.; Mouawad, L. Benchmark of Four Popular Virtual Screening Programs: Construction of the Active/Decoy Dataset Remains a Major Determinant of Measured Performance. *J. Cheminf.* **2016**, *8*, 56.
73. Wallach, I.; Heifets, A. Most Ligand-Based Classification Benchmarks Reward Memorization Rather Than Generalization. *J. Chem. Inf. Model.* **2018**, *58*, 916-932.
74. Chen, L.; Cruz, A.; Ramsey, S.; Dickson, C. J.; Duca, J. S.; Hornak, V.; Koes, D. R.; Kurtzman, T. Hidden Bias in the DUD-E Dataset Leads to Misleading Performance of Deep Learning in Structure-Based Virtual Screening. *PLoS One* **2019**, *14*, No. e0220113.
75. Sieg, J.; Flachsenberg, F.; Rarey, M. In Need of Bias Control: Evaluating Chemical Data for Machine Learning in Structure-Based Virtual Screening. *J. Chem. Inf. Model.* **2019**, *59*, 947-961.
76. Lagarde, N.; Zagury, J. F.; Montes, M. Benchmarking Data Sets for the Evaluation of Virtual Ligand Screening Methods: Review and Perspectives. *J. Chem. Inf. Model.* **2015**, *55*, 1297-1307.
77. BIOVIA Available Chemicals Directory (ACD). <https://www.3dsbiovia.com/products/collaborative-science/databases/sourcing-databases/biovia-available-chemicals-directory.html> (accessed May 2020).
78. Irwin, J. J.; Shoichet, B. K. ZINC – A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 177-182.
79. Tran-Nguyen, V. 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.
80. Rohrer, S. G.; Baumann, K. Maximum Unbiased Validation (MUV) Data Sets for Virtual Screening Based on PubChem Bioactivity Data. *J. Chem. Inf. Model.* **2009**, *49*, 169-184.
81. Schierz, A. C. Virtual Screening of Bioassay Data. *J. Cheminform.* **2009**, *1*, 21.
82. Butkiewicz, M.; Lowe, E. W.; Mueller, R.; Mendenhall, J. L.; Teixeira, P. L.; Weaver, C. D.; Meiler, J. Benchmarking Ligand-Based Virtual High-Throughput Screening with the PubChem Database. *Molecules* **2013**, *18*, 735-756.
83. Lindh, M.; Svensson, F.; Schaal, W.; Zhang, J.; Sköld, C.; Brandt, P.; Karlen, A. Toward a Benchmarking Data Set Able to Evaluate Ligand- and Structure-Based Virtual Screening Using Public HTS Data. *J. Chem. Inf. Model.* **2015**, *55*, 343-353.

84. Baell, J. B.; Holloway, G. A. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) From Screening Libraries and for Their Exclusion in Bioassays. *J. Med. Chem.* **2010**, *53*, 2719-2740.
85. Gilberg, E.; Jasial, S.; Stumpfe, D.; Dimova, D.; Bajorath, J. Highly Promiscuous Small Molecules From Biological Screening Assays Include Many Pan-Assay Interference Compounds but Also Candidates for Polypharmacology. *J. Med. Chem.* **2016**, *59*, 10285-10290.
86. Baell, J. B. Feeling Nature's PAINS: Natural Products, Natural Product Drugs, and Pan Assay Interference Compounds (PAINS). *J. Nat. Prod.* **2016**, *79*, 616-628.
87. Capuzzi, S. J.; Muratov, E. N.; Tropsha, A. Phantom PAINS: Problems with the Utility of Alerts for Pan-Assay Interference Compounds. *J. Chem. Inf. Model.* **2017**, *57*, 417-427.
88. Kenny, P. W. Comment on the Ecstasy and Agony of Assay Interference Compounds. *J. Chem. Inf. Model.* **2017**, *57*, 2640-2645.
89. Baell, J. B.; Nissink, J. W. Seven Year Itch: Pan-Assay Interference Compounds (PAINS) in 2017 – Utility and Limitations. *ACS Chem. Biol.* **2018**, *13*, 36-44.
90. Nim, S.; Lobato, L. G.; Moreno, A.; Chaptal, V.; Rawal, M. K.; Falson, P.; Prasad, R. Atomic Modelling and Systematic Mutagenesis Identify Residues in Multiple Drug Binding Sites That Are Essential for Drug Resistance in the Major Candida Transporter Cdr1. *Biochim. Biophys. Acta* **2016**, *1858*, 2858-2870.
91. Hughes, J. P.; Rees, S.; Kalindjian, S. B.; Philpott, K. L. Principles of Early Drug Discovery. *Br. J. Pharmacol.* **2011**, *162*, 1239-1249.
92. Hsieh, J. H. Accounting Artifacts in High-Throughput Toxicity Assays. *Methods Mol. Biol.* **2016**, *1473*, 143-152.
93. Good, A. C.; Oprea, T. I. Optimization of Camd Techniques 3. Virtual Screening Enrichment Studies: A Help or Hindrance in Tool Selection? *J. Comput. Aided Mol. Des.* **2008**, *22*, 169-178.
94. Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887-2893.
95. Dassault Systèmes, Biovia Corp. <https://www.3dsbiovia.com/> (accessed April 2020).
96. Schuffenhauer, A.; Ertl, P.; Roggo, S.; Wetzel, S.; Koch, M. A.; Waldmann, H. The Scaffold Tree, Visualization of the Scaffold Universe by Hierarchical Scaffold Classification. *J. Chem. Inf. Model.* **2007**, *47*, 47-58.
97. Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742-754.
98. Jain, A. N. Surflex-Dock 2.1: Robust Performance from Ligand Energetic Modeling, Ring Flexibility, and Knowledge-Based Search. *J. Comput. Aided Mol. Des.* **2007**, *21*, 281-306.
99. Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31-36.
100. Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K.; Grier, D. L.; Leland, B. A.; Laufer, J. Description of Several Chemical Structure File Formats Used by Computer Programs Developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 244-255.
101. Cummings, M. D.; Gibbs, A. C.; DesJarlais, R. L. Processing of Small Molecule Databases for Automated Docking. *Med. Chem.* **2007**, *3*, 107-113.

102. Knox, A. J.; Meegan, M. J.; Carta, G.; Lloyd, D. G. Considerations in Compound Database Preparation – “Hidden” Impact on Virtual Screening Results. *J. Chem. Inf. Model.* **2005**, *45*, 6, 1908-1919.
103. Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D. Comparative Evaluation of Eight Docking Tools for Docking and Virtual Screening Accuracy. *Proteins* **2004**, *57*, 225-242.
104. Perola, E.; Charifson, P. S. Conformational Analysis of Drug-Like Molecules Bound to Proteins: An Extensive Study of Ligand Reorganization upon Binding. *J. Med. Chem.* **2004**, *47*, 2499-2510.
105. Marcou, G.; Rognan, D. Optimizing Fragment and Scaffold Docking by Use of Molecular Interaction Fingerprints. *J. Chem. Inf. Model.* **2007**, *47*, 195-207.
106. 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.
107. Polgar, T.; Keserue, G. M. Ensemble Docking into Flexible Active Sites. Critical Evaluation of FlexE against JNK-3 and β -Secretase. *J. Chem. Inf. Model.* **2006**, *46*, 1795-1805.
108. Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and Validation of a Genetic Algorithm for Flexible Docking. *J. Mol. Biol.* **1997**, *267*, 727-748.
109. Hawkins, P. C.; Skillman, A. G.; Nicholls, A. Comparison of Shape-Matching and Docking as Virtual Screening Tools. *J. Med. Chem.* **2007**, *50*, 74–82.
110. Bietz, S.; Urbaczek, S.; Schulz, B.; Rarey, M. Protoss: A Holistic Approach to Predict Tautomers and Protonation States in ProteinLigand Complexes. *J. Cheminf.* **2014**, *6*, 12.
111. Molecular Networks Gmbh. <https://www.mn-am.com/> (accessed April 2020).
112. *Molecular Operating Environment (MOE), 2018.01*; Chemical Computing Group, Inc.: Montreal, QC, Canada, 2015.
113. *Sybyl-X Molecular Modeling Software Packages, 2.1.1*; TRIPOS Associates, Inc.: St. Louis, MO, USA, 2013.
114. Daylight Chemical Information Systems. <https://www.daylight.com/> (accessed May 2020).

Supporting Information**Table S1.** Number of PubChem bioactivity assays according to the number of tested substances, the number of active substances, the screening stage, and the target type. Statistics were updated as of April 30, 2020.

Criteria	Assay type	
	Small-molecule assay	RNA interference assay
1. Number of tested substances (N_t):		
• $N_t < 100$	1,060,707	22
• $100 \leq N_t < 1000$	4530	92
• $1000 \leq N_t < 10,000$	1359	14
• $10,000 \leq N_t < 100,000$	422	48
• $N_t \geq 100,000$	701	1
2. Number of active substances (N_a):		
• $N_a < 10$	1,000,714	28
• $10 \leq N_a < 50$	60,328	57
• $50 \leq N_a < 100$	3562	18
• $100 \leq N_a < 1000$	2399	60
• $N_a \geq 1000$	716	14
3. Screening stage:		
• Primary screening	1416	113
• Confirmatory, dose-response curves not provided	276,216	0
• Confirmatory, dose-response curves provided	3904	0
• Summary	701	10
• Screening stage not annotated	785,482	54
4. Target type:		
• Single protein	238,096	0
• Single gene	17	0
• Single nucleotide	95,325	0
• Multiple proteins	25,649	0
• Multiple genes	3	0
• Multiple nucleotides	8646	0
• Protein-protein interaction	210	0
• None	795,301	177
All	1,067,719	177

Table S2. Number of compounds featured in PubChem bioactivity assays that satisfy each criterion of the Lipinski's rule of five, the Ghose filter, and the Veber's rule. Statistics were updated as of April 30, 2020.

Criteria	Number of PubChem compounds
1. Lipinski's rule of five:	
• Molecular mass \leq 500 Da	88,667,112
• ClogP \leq 5	78,183,471
• Number of hydrogen bond donors \leq 5	101,211,514
• Number of hydrogen bond acceptors \leq 10	99,344,677
• Compounds satisfying all criteria	73,062,126
2. Ghose filter:*	
• Molecular mass from 180 Da to 480 Da	82,926,795
• AlogP from -0.4 to +5.6	79,473,661
• Number of atoms from 20 to 70	71,554,127
3. Veber's rule:	
• Number of rotatable bonds not exceeding 10	93,857,861
• Polar surface area not exceeding 140 \AA^2	96,031,201
All	102,694,672

* The criterion regarding molar refractivity of the Ghose filter is not addressed in this table, as no relevant search option is available on PubChem Compound.

Table S3. Scaffold clusters of PubChem BioAssay active ligands (AID 493208) and the number of their representatives before and after LIT-PCBA filters.

Scaffold cluster	Scaffold structure	Number of substances falling into each scaffold cluster	
		Full data from AID 493208	Data from the LIT-PCBA MTORC1 ligand set
1	(O=C1CNCCO1)	3	1
2	(c1ccncc1)	38	9
3	(o1cccc1)	9	5
4	(O=C1CC=CN1)	15	9
5	(N=C1NC=CN1)	8	5
6	(c1nnn[nH]1)	4	1
7	(c1cnnc1)	40	4
8	(c1ccsc1)	3	0
9	(C1COCO1)	1	1
10	(O=C1NC=CC1=O)	5	5
11	(c1nc[nH]n1)	4	1
12	(O=C1NC=CSC=C1)	1	1
13	(C1CN=CN1)	1	0
14	(c1cccc1)	9	2
15	(C1CC=CCN1)	5	2
16	(c1cn[nH]c1)	20	1
17	(c1cc[nH]c1)	13	4
18	(C1CCNCC1)	6	1
19	(c1nnc[nH]1)	5	2
20	(c1csen1)	16	5
21	(C1OC=CO1)	5	3
22	(C1CNC=CC1)	2	0
23	(C1CCC\C=C/CC1)	2	1
24	(c1c[nH]cn1)	13	3
25	(C1CNCCN1)	13	3
26	(o1cccn1)	3	2
27	(C1COC=CC1)	1	1
28	(c1c[nH]nn1)	2	1
29	(C1C=CNC=N1)	1	0
30	(O=C1NC=CC=C1)	7	4
31	(C1COC=CO1)	3	1
32	(C1CC=CN1)	3	2
33	(C1CCC=NCC1)	1	1
34	(O=C1NN=CC=C1)	8	3
35	(c1ccnnc1)	4	1
36	(O=C1NC=CN=C1)	2	1
37	(C1NC=CN=C1)	2	0
38	(C1COC=CN1)	1	0
39	(O=C1NN=CN=C1)	3	0
40	(O=S1(=O)NC=CC=C1)	1	1
41	(o1ccnc1)	21	1
42	(c1nncs1)	4	2
43	(O=C1NC=CC(=O)N1)	5	1
44	(C1CC=CO1)	5	2
45	(C1OC=CC=C1)	1	0
46	(O=C1C=CNC=C1)	1	0
47	(O=C1NC=CS(=O)(=O)C=C1)	1	0
48	(C1CN=CO1)	3	0
49	(O=C1C=COC=C1)	1	1

50	(O=C1NC=NC=C1)	2	1
51	(c1ncncn1)	5	0
52	(c1cnncn1)	1	1
53	(C1CCC=CCC1)	2	0
54	(c1cn[nH]n1)	1	0
55	(C1CCCCC1)	1	0
56	(O=C1NCC=C1)	1	0
57	(C1CN=CC=C1)	2	0
58	(S1C=CC=NC=C1)	1	0
59	(o1nccn1)	1	1
All		342	97

Table S4. Scaffold clusters of PubChem BioAssay inactive ligands (AID 493208) and the number of their representatives before and after LIT-PCBA filters.

Scaffold cluster	Scaffold structure	Number of substances falling into each scaffold cluster	
		Full data from AID 493208	Data from the LIT-PCBA MTORC1 ligand set
1	(c1ccc2ncccc2c1)	638	288
2	(O=S(=O)(NC1=NCNCN1)c2ccccc2)	5	5
3	(O=C1NC=Cc2ccccc12)	196	195
4	(C1OC=Cn2ccnc12)	12	12
5	(O=C1NC=Ne2ccccc12)	256	150
6	(C1OC=Cc2ncccc12)	28	5
7	(O=C1NC(=O)c2ccnc2N1)	160	160
8	(O=C1NNe2ncccc12)	49	48
9	(O=C1OC=Cc2ccccc12)	34	25
10	(c1cnc2ccnn2c1)	439	305
11	(O=C1COc2ccccc2N1)	402	393
12	(c1ccc2[nH]ccc2c1)	1139	733
13	(O=C(CN1CCCNCC1)Nc2ccccc2)	8	8
14	(O=C1C=CNc2ccnn12)	35	31
15	(C1CCc2sccc2C1)	319	180
16	(C1COc2ccccc2O1)	281	250
17	(c1ccc2ncnc2c1)	98	16
18	(O=C(CC1NCCNC1=O)Nc2ccccc2)	19	19
19	(C1N=CNc2ncn12)	10	10
20	(c1nnc2cc[nH]c2n1)	33	27
21	(O=C(NC1C=CNC1=O)c2ccccc2)	13	13
22	(O=C1CC2=C(N1)NC(=O)NC2=O)	42	42
23	(o1cnc2ncccc12)	72	65
24	(O=C(CSC1=NC(=O)C=CN1)Nc2ccccc2)	15	15
25	(O=C1CNC(=O)N1c2ccccc2)	7	7
26	(O=C1NC(NS(=O)(=O)c2ccccc2)C(=O)N1)	15	15
27	(c1cnc2ncnn2c1)	274	205
28	(c1cn2ncccc2nn1)	6	6
29	(O=C1NN=C2COC=CN12)	6	6
30	(O=C1CN=CN1C2CCCCC2)	5	0
31	(O=C(CC1NCCOC1=O)Nc2ccccc2)	7	7
32	(N=C1Nc2ccccc2S1)	8	8
33	(C1C=NNC1c2cn[nH]c2)	5	3
34	(O=C(CNS(=O)(=O)c1ccccc1)NCc2ccccc2)	14	14
35	(O=C1C=CN2C=CC=CC2=N1)	12	12
36	(O=C(CNS(=O)(=O)c1ccccc1)Nc2ccccc2)	46	39
37	(c1ccc(cc1)c2nn[nH]n2)	50	42
38	(O=C(CNc1ccccc1)Nc2ccccc2)	31	31
39	(c1ccc(cc1)n2cnnn2)	23	23
40	(O=C(CSc1nncc1)Nc2ccccc2)	34	34
41	(O=C1C2CN3CC1CN(C2)C3c4ccccc4)	26	23
42	(O=C1NCNc2ncccc12)	5	5
43	(o1ccc2ccccc12)	604	405
44	(C(N1CCNCC1)c2ccccc2)	21	21
45	(O=C1CCC2=C(O1)C=CNC2=O)	14	14
46	(c1ccc(cc1)c2ccc[nH]2)	139	100
47	(O=S(=O)(c1ccccc1)n2ccnc2)	14	14
48	(c1ccc2nsnc2c1)	175	170

49	(o1nc2ccccc2n1)	68	68
50	(O=C1NC(=NC=C1)SCc2oncn2)	3	3
51	(O=C(CCS(=O)(=O)c1ccccc1)Nc2ccccc2)	12	12
52	(C(Nc1ncc[nH]1)c2ccccc2)	9	8
53	(O=C1NC=C(C(=O)N1)S(=O)(=O)Nc2ccccc2)	5	5
54	(N=C1NC=Nc2[nH]ncc12)	27	27
55	(O=C(CCN1cnnn1)Nc2ccccc2)	13	13
56	(C1CN(CCN1)c2ccccc2)	166	151
57	(O=S(=O)(Nc1ncccn1)c2ccccc2)	27	27
58	(O=C1Nc2nccn2C=C1)	12	12
59	(O=C1NC(=O)c2[nH]cnc2N1)	190	175
60	(O=C(CNCCc1ccccc1)Nc2ccccc2)	14	14
61	(O=C(NCCc1ccccc1)C(=O)Nc2ccccc2)	14	14
62	(C(c1noen1)n2cccn2)	6	6
63	(N(c1ccccc1)c2nccs2)	17	14
64	(O=C(Nc1ccccc1)C2CNC(=O)C2)	11	11
65	(O=C(CSc1ocnn1)c2ccccc2)	8	8
66	(O=C1C=CN=C2C=CC=CN12)	438	424
67	(O=C(CSCc1ocn1)N2CCNCC2)	13	13
68	(O=C(Oc1ccn[nH]1)c2ccccc2)	42	15
69	(O=C1NC=CS(=O)c2ccccc12)	123	72
70	(O=C(CS(=O)Cc1coen1)NCCc2ccccc2)	13	13
71	(O=C(CNc1ccccc1)NCCSCc2ccccc2)	14	14
72	(O=C(CSCc1oenn1)NCCc2ccccc2)	13	13
73	(O=C1Oc2ccccc2C=C1)	438	336
74	(C(N1CCCC1)c2coen2)	50	44
75	(O=C(Nc1ccccc1)c2occc2)	30	26
76	(O=C(NC1CCNCC1)Nc2ccccc2)	12	12
77	(O=C1CSc2ccccc2N1)	138	134
78	(o1enc2ccccc12)	528	287
79	(O=C1CCCC2=C1CC=CN2)	6	6
80	(c1cnn2cnnc2c1)	82	75
81	(c1ccc2[nH]cnc2c1)	682	415
82	(O=C(Nc1ccccc1)C2CCCN2)	5	5
83	(c1ccc2enncc2c1)	88	40
84	(O=C1Cc2ccccc2N1)	126	124
85	(O=C1Nc2ccccc2O1)	359	354
86	(O=C1NN=C(C=C1)c2ccccc2)	70	67
87	(C1NC=Cc2ccnn12)	13	13
88	(O=S(=O)(NCC1CCCC1)c2ccccc2)	18	12
89	(O=C1Nc2ccccc2N1)	128	128
90	(O=C1CN=Cc2ccccc2N1)	32	21
91	(O=C1C=CNc2ccccc12)	671	563
92	(O=C1NCC2=C1OC=CC2=O)	110	97
93	(C1Oc2ccccc2O1)	666	570
94	(O=C(NCc1cccs1)Nc2ccccc2)	17	15
95	(O=C(NCCc1ccccc1)c2occc2)	32	31
96	(O=S(=O)(c1ccccc1)c2c[nH]nn2)	15	15
97	(c1ccc2n[nH]nc2c1)	76	30
98	(O=C(Nc1ccccc1)\C=C\c2ccccc2)	21	19
99	(O=S(=O)(N1CCCCC1)c2ccccc2)	11	11
100	(c1ccc(cc1)c2ccn[nH]2)	11	0
101	(N=C1NC(=O)CC(S1)C(=O)Nc2ccccc2)	8	8
102	(O=C1N=CNc2sccc12)	2	2
103	(O=C1CCC(=NN1)c2ccccc2)	48	48
104	(C1Cc2senc2C=C1)	11	11

105	(C1C=Cc2nsc12)	13	13
106	(O=C1NCc2ccccc2N1)	44	43
107	(O=C1NCc2ccccc2N1)	70	63
108	(O=C1CCC2=C(COC2=O)N1)	12	12
109	(O=C1CCC2=C(O1)C=COC2=O)	7	7
110	(O=C(CCOc1ccccc1)Nc2ccccc2)	8	8
111	(c1ccn2ccnc2c1)	143	64
112	(O=C1CN(C2CCCCC2)C(=O)CN1)	50	41
113	(c1ccc(cc1)c2ccnnc2)	11	11
114	(O=C(CSCc1ccccc1)NCc2ccccc2)	7	6
115	(C1CCc2[nH]ncc2C1)	13	13
116	(O=C1NC2=CC=CCC2=C1)	16	16
117	(O=C(NC1ccccc1)c2occc2)	37	37
118	(o1ccnc1c2ccccc2)	32	32
119	(O=S(=O)(NCc1ccccc1)c2ccccc2)	12	5
120	(O=C(COc1ccccc1)Nc2cn[nH]c2)	6	6
121	(O=C(COc1ccccc1)Nc2cccs2)	14	14
122	(o1cnnnc1c2ccccc2)	36	22
123	(O=C(Nc1ccccc1)c2c[nH]cn2)	5	5
124	(O=S(=O)(N1CCCCC1)c2ccccc2)	27	27
125	(O=S(=O)(N1CCOCC1)c2ccccc2)	27	27
126	(O=C(Nc1ccccc1)\C=C\c2occc2)	26	21
127	(O=C1NC(=O)c2c[nH]cc2N1)	59	42
128	(C1Cc2ccccc2CN1)	142	120
129	(C(Oc1ccccc1)c2occc2)	10	10
130	(C1CC2(CCN1)OCCO2)	34	31
131	(O=C(Nc1ccccc1)c2ccccc2)	133	76
132	(O=C(COc1ccccc1)NC2CCS(=O)(=O)C2)	12	12
133	(O=C1NC=CS(=O)(=O)c2ccccc12)	170	104
134	(O=C(NC1CCS(=O)(=O)C1)c2ccccc2)	32	32
135	(c1ccc(cc1)n2ccnn2)	20	20
136	(O=C1NN=Nc2sccc12)	80	34
137	(c1ccc2scnc2c1)	819	610
138	(O=C1NCc2cn[nH]c12)	118	91
139	(o1cnc(n1)c2ccccc2)	75	72
140	(O=C(NCC12CC3CC(CC(C3)C1)C2)c4cc[nH]n4)	6	0
141	(o1cnc(n1)c2ccncc2)	20	20
142	(c1ncc2cn[nH]c2n1)	225	175
143	(O=C(CSc1ocnn1)Nc2ccccc2)	36	36
144	(C1C=CNC2nncn12)	273	164
145	(C(Oc1ccccc1)c2oncn2)	13	13
146	(c1ncc2cc[nH]c2n1)	76	3
147	(O=C1NN=Cc2ccccc12)	75	70
148	(O=C(CN1CCNCC1)Nc2ccccc2)	26	26
149	(c1cnc2sccc2c1)	168	111
150	(o1cnc(n1)c2occc2)	21	21
151	(O=C1NC(=O)c2ccsc2N1)	305	200
152	(c1ccn2cncn2c1)	33	24
153	(O=C1NC=Nc2ccsc12)	114	46
154	(O=C(Nc1ccccc1)C2CCNCC2)	40	40
155	(c1cc2nncn2cn1)	167	83
156	(O=C1Nc2ccccc2S(=O)(=O)N1)	58	37
157	(O=C1NC=Nc2sccc12)	203	151
158	(C1CN2CCC1c3ncccc23)	17	16
159	(O=C1NCC2=C(CNC2=O)N1)	13	13
160	(O=C1NC(=O)c2ccccc2N1)	571	501

161	(O=C1C=CC2=C1CC=CN2)	11	11
162	(O=C(CSc1ncc[nH]1)Nc2ccccc2)	89	72
163	(C1CNc2ccccc2C1)	299	264
164	(O=C(CSC1=NC(=O)NC=C1)Nc2ccccc2)	13	13
165	(O=S(=O)(Cc1occc1)c2cccs2)	39	39
166	(C1C=CNC2nnnn12)	24	24
167	(O=C(NCCc1occc1)C(=O)NCCc2ccccc2)	12	12
168	(O=C(NCCc1occc1)C(=O)NCc2ccccc2)	11	11
169	(C(N1CCNCC1)c2occc2)	18	18
170	(C1CC(CCO1)c2cccs2)	24	24
171	(O=C1Nc2ncn2C=C1)	12	12
172	(O=C1NNC=C1NS(=O)(=O)c2ccccc2)	6	6
173	(O=C1NC=COc2ccccc12)	4	4
174	(O=C1CC(=O)Nc2ccccc2N1)	25	25
175	(O=C(NCCc1ccccc1)C(=O)NCCc2cccs2)	12	12
176	(O=C(NCCc1cccs1)C(=O)NCc2ccccc2)	13	13
177	(c1enc2[nH]ccc2n1)	17	17
178	(C(N1CCNCC1)c2cccnc2)	63	63
179	(O=C(CSc1cccn1)Nc2ccccc2)	56	19
180	(O=C(CNS(=O)(=O)c1conc1)Nc2ccccc2)	13	13
181	(C(c1occc1)n2ccnc2)	13	13
182	(C1NC=Cc2nnnc2O1)	33	18
183	(O=C1NC=Ne2cc[nH]c12)	175	74
184	(O=C1CCCC2=C1C=CC(=O)N2)	62	42
185	(O=C1NC=CC(N1)c2en[nH]c2)	13	13
186	(c1enc2ncn2c1)	89	65
187	(O=C(CSC1=NC=CC(=O)N1)Nc2ccccc2)	49	45
188	(O=C1C2CN3CC1CN(C2)C3c4cc[nH]c4)	14	14
189	(O=C(CNc1ccccc1)NCc2ccccc2)	14	14
190	(C1CC2CNC=CN2C1)	26	26
191	(O=C1CCn2nccc2N1)	86	86
192	(O=C1NN=Cc2c[nH]nc12)	105	105
193	(O=C1CCc2ccccc2N1)	107	107
194	(c1enc2n[nH]cc2c1)	25	6
195	(O=S1(=O)NC=Cc2nncnc12)	106	75
196	(O=C(CNS(=O)(=O)c1cc[nH]c1)N2CCNCC2)	13	13
197	(c1ccc(cc1)n2cnnc2)	9	9
198	(O=C1CNC(=O)C2CCCCN12)	19	19
199	(C1Cn2cnnc2S1)	37	37
200	(C1Cc2ccsc2CN1)	86	53
201	(C(C1CCNCC1)N2CCCCC2)	13	13
202	(O=C(NCCc1occc1)c2ccccc2)	13	13
203	(C(C1CCNCC1)N2CCCC2)	13	13
204	(c1cnn2ccnc2c1)	13	13
205	(C(C1CCNCC1)N2CCOCC2)	13	13
206	(O=S(=O)(Cc1cccs1)c2cccs2)	26	26
207	(C(C1CCNCC1)N2CCCCC2)	18	18
208	(C1COc2ccccc2N1)	214	212
209	(O=C(CNS(=O)(=O)c1cc[nH]c1)NCc2ccccc2)	13	13
210	(O=C(CNS(=O)(=O)c1cc[nH]c1)Nc2ccccc2)	31	31
211	(O=S(=O)(N1CCCCC1)c2cc[nH]c2)	99	99
212	(O=C(NCc1ccccc1)c2ccn[nH]2)	13	13
213	(O=C1C=CN=C2SC=NN12)	649	595
214	(O=C1CCCc2ccccc2N1)	93	93
215	(C1CSc2ccccc2N1)	137	99
216	(c1cn2ncsc2n1)	69	69

217	(C1CCc2nnen2CC1)	24	24
218	(c1en2ccsc2n1)	99	87
219	(O=C1OC2(CCNCC2)C=C1)	25	25
220	(o1cccc1c2oncc2)	12	12
221	(O=C1NC=CC=C1CN2CCCCC2)	25	25
222	(o1nccc1c2cccs2)	36	36
223	(o1nccc1c2cccc2)	112	110
224	(O=C1NCc2ccccc12)	206	205
225	(O=S(=O)(N1CCNCC1)c2ccccc2)	80	54
226	(C(C1CCCCC1)n2cccc2)	12	12
227	(O=S1(=O)C=CC(=N1)NCc2ccccc2)	12	12
228	(O=C1C=CNC2ncccc12)	75	66
229	(O=C1OC=Cc2[nH]cnc12)	22	22
230	(O=C1NC=CC=C1CN2CCNCC2)	43	43
231	(O=C(COCc1cccn1)Nc2ccccc2)	12	12
232	(O=C(NCc1nnn[nH]1)c2ccccc2)	13	13
233	(O=S(=O)(N1CCCCC1)c2cn[nH]c2)	20	20
234	(O=S(=O)(N1CCCCC1)c2cn[nH]c2)	35	35
235	(O=C(NCCS(=O)(=O)N1CCNCC1)C2CNC(=O)C2)	11	11
236	(O=C1NCC=C(CN2CCNCC2)N1)	17	17
237	(O=C(NC1=NCCC(=O)N1)c2ccccc2)	13	13
238	(o1nce2cnenc12)	164	127
239	(C(Oc1ccccc1)c2coen2)	23	18
240	(C1CNc2ccnn2C1)	13	13
241	(C1Nc2ccenc2OC=C1)	47	38
242	(c1enc2[nH]ncc2c1)	164	99
243	(c1enc2[nH]cnc2c1)	388	296
244	(C1NCc2ccccc2O1)	43	23
245	(O=C(N1CCCCC1)c2cn[nH]c2)	45	42
246	(O=C1CCSc2ccccc2N1)	67	62
247	(c1ccn(c1)c2ccn[nH]2)	79	59
248	(C1CCc2cc[nH]c2CC1)	23	18
249	(C(Sc1ncc[nH]1)c2ccccc2)	13	13
250	(C1CCN(C1)c2ccenn2)	23	22
251	(C1CN(CCO1)c2ccenn2)	13	13
252	(C1CCN(CC1)c2ccenn2)	42	20
253	(O=C1Nc2nnen2C=C1)	241	230
254	(c1ccc(cc1)c2ccenn2)	30	30
255	(O=C1CN(Cc2occc2)C(=O)N1)	13	13
256	(O=S(=O)(N1CCCC1)c2cc[nH]c2)	25	25
257	(O=C1CN(Cc2cccs2)C(=O)N1)	13	13
258	(O=C(CC1NC(=O)NC1=O)Nc2ccccc2)	69	68
259	(C1CCN(CC1)c2nncs2)	111	87
260	(C1C=CNC2ccnn12)	39	33
261	(c1ccc(cc1)n2ccnn2)	26	21
262	(C1CC(CCN1)c2ccn[nH]2)	129	79
263	(o1nenc1c2c[nH]nn2)	65	65
264	(O=C1C=CSc2ccccc12)	39	39
265	(O=C1NC=CSc2ccccc12)	202	108
266	(c1en2nenc2s1)	220	198
267	(O=S(=O)(NCCc1csen1)c2ccccc2)	13	13
268	(O=C1Nc2ccccc2C=C1)	244	237
269	(O=C(NCCc1csen1)C(=O)Nc2ccccc2)	31	31
270	(C1Cc2ccccc2N1)	884	872
271	(O=C1NC(=O)c2[nH]ccc2N1)	158	73
272	(O=C1NC=Nc2occc12)	69	69

273	(o1ccc2cnenc12)	151	151
274	(c1en2nnnc2cn1)	52	52
275	(O=S(=O)(N1CCOCC1)c2cc[nH]c2)	13	13
276	(O=C(Cc1ccccc1)NCCc2csen2)	12	12
277	(O=C(NCCc1csen1)c2ccccc2)	31	27
278	(O=S(=O)(NCCc1cnes1)c2ccccc2)	17	17
279	(O=C(NCCc1cnes1)C(=O)Nc2ccccc2)	29	29
280	(O=C(NCCc1cnes1)c2ccccc2)	31	28
281	(O=C(NCCc1cccs1)c2ccccc2)	13	10
282	(O=S1(=O)CCCCN1c2ccccc2)	42	42
283	(O=C1Nc2ccccc2NC1=O)	130	130
284	(O=C1CC(=O)c2ccccc12)	21	13
285	(O=C(NCc1ccccc1)c2enenc2)	20	20
286	(O=C(Nc1ccccc1)c2enenc2)	18	17
287	(O=C1NN=Cc2cc[nH]c12)	353	246
288	(C1CCc2cnoc2C1)	18	18
289	(O=C(NCCc1ccccc1)c2c[nH]nn2)	13	13
290	(O=C(NCc1cccs1)c2c[nH]nn2)	13	13
291	(O=C(NCc1ccccc1)c2c[nH]nn2)	19	19
292	(C(N1CCOCC1)c2ccccc2)	39	39
293	(C(N1CCCC1)c2ccccc2)	16	16
294	(O=C(CN1C=CC=CC1=O)Nc2ccccc2)	13	13
295	(O=C(Nc1ccccc1)c2ccon2)	29	25
296	(o1nccc1c2ccccc2)	45	45
297	(O=C(NCc1ccc[nH]1)Nc2ccccc2)	12	12
298	(O=C(Nc1enon1)c2occc2)	6	6
299	(O=C(COc1ccccc1)Nc2enon2)	12	12
300	(O=C(Nc1ccccc1)c2c[nH]nn2)	69	69
301	(c1ncc2nc[nH]c2n1)	32	32
302	(O=S(=O)(N1CCNCC1)c2cc[nH]c2)	17	17
303	(C1CCc2ncn2CC1)	87	87
304	(O=C1NC=Nc2oncc12)	65	65
305	(c1ccc2ncnc2c1)	242	88
306	(O=C1NC=CSc2ncccc12)	24	24
307	(O=C1NC=Cc2ncccc12)	99	99
308	(O=S1(=O)N=Cc2ccccc12)	9	9
309	(O=S(=O)(NCc1oncn1)c2ccccc2)	10	10
310	(O=C(Cc1ncccn1)Nc2ccccc2)	11	11
311	(O=C(Cc1ncccn1)Nc2ccccc2)	29	29
312	(C(Sc1nnc[nH]1)c2oncn2)	63	27
313	(O=C(CCCc1oncn1)Nc2ccccc2)	13	13
314	(O=S1(=O)C=CNc2ccccc12)	22	19
315	(c1ccc2sccc2c1)	50	46
316	(O=C1CSC2(N1)C=CNC2=O)	102	84
317	(O=C1OC2(CCCCCC2)C=C1)	18	18
318	(O=C1OC2(CCCCCC2)C=C1)	35	35
319	(O=C(Cn1ccccc1)Nc2ccccc2)	13	13
320	(O=C(CSc1ncccn1)Nc2ccccc2)	10	10
321	(c1cc2nncn2cn1)	37	33
322	(O=C(CNS(=O)(=O)c1ccsc1)Nc2ccccc2)	12	12
323	(O=S(=O)(Nc1ccccc1)c2ccsc2)	28	20
324	(O=S(=O)(N1CCNCC1)c2ccsc2)	24	22
325	(O=C(CNS(=O)(=O)c1cccs1)Nc2ccccc2)	39	37
326	(O=C1Nc2[nH]ncc2C=C1)	176	166
327	(O=C(Nc1nncs1)c2ccccc2)	54	54
328	(O=S1(=O)NCCCN1Cc2ccccc2)	9	9

329	(O=C(CN1CCNS1(=O)=O)Nc2ccccc2)	10	10
330	(c1cn2cnnc2cn1)	64	55
331	(C1Cn2cnnc2C=N1)	23	7
332	(C1CSc2ncccc2N1)	63	60
333	(C1CCc2[nH]ccc2C1)	43	16
334	(C1CCc2ccccc2NC1)	51	35
335	(O=C1NC=Nc2[nH]ncc12)	68	44
336	(O=C(NC1CCCCC1)C2CNCC(=O)N2)	12	12
337	(O=C1NC=COc2ncccc12)	58	58
338	(c1nncc2n[nH]cc12)	122	86
339	(C(NC1CCNCC1)c2ccccc2)	6	6
340	(C1CCN2CCNCC2C1)	13	13
341	(C(NC1CCNCC1)c2occc2)	13	13
342	(O=S1(=O)C=CC(=N1)N2CCCCC2)	49	49
343	(O=C1NN2C=NC=NC2=C1)	72	50
344	(o1ncc2ccccc12)	61	60
345	(O=C1C=CN=C2CCCCCN12)	77	76
346	(O=C1NC=CN=C1NCCc2ccccc2)	13	13
347	(O=C1NC=CN=C1NCc2ccccc2)	13	13
348	(c1nc(ns1)c2c[nH]nn2)	34	23
349	(O=C(NC1C1oncn1)c2ccccc2)	30	27
350	(C1Nc2secc2C=N1)	13	11
351	(c1esc(c1)c2ccnnc2)	13	4
352	(O=C1NC=Cc2secc12)	90	83
353	(O=C(NC1c1cccc1)C2CCCCC2)	7	7
354	(O=C(NCCc1c1cccc1)C2CCCCC2)	11	9
355	(O=C1OC=Cc2secc12)	65	39
356	(O=S1(=O)NC=Cc2ccccc12)	56	53
357	(N1C=CS/C1=N\c2ccccc2)	5	5
358	(O=C1CCCc2ncncc12)	6	6
359	(c1ccc(cc1)c2cc[nH]n2)	14	10
360	(O=S1(=O)CCC(C1)NCc2cccs2)	23	23
361	(o1cccc1c2ccccc2)	11	4
362	(O=C1C=CNc2nncn12)	19	18
363	(C(CSc1nnn[nH]1)NCc2ccccc2)	36	36
364	(O=C1CCC2=C(N1)NC(=O)NC2=O)	6	6
365	(C1CCNCC1)	13	13
366	(c1c[nH]c(c1)c2cccs2)	13	7
367	(O=C(CSc1nnc[nH]1)Nc2ccccc2)	42	42
368	(C1Cc2ncccc2CN1)	26	26
369	(O=S(=O)(Nc1c1cccc1)c2cccs2)	13	13
370	(O=C1NN=Nc2ccsc12)	13	5
371	(O=C(NC1c1cccc1)C2CCNCC2)	16	16
372	(c1cnc2nncn2c1)	48	47
373	(O=S(=O)(N1CCCCC1)c2conc2)	120	108
374	(O=C(Nc1cccs1)\C=C\c2ccccc2)	11	8
375	(O=C1C=COc2ccccc12)	204	170
376	(C(=C\c1nccs1)/c2ccccc2)	10	1
377	(O=C(COc1c1cccc1)Nc2nnc[nH]2)	6	6
378	(N=C1Nc2ccccc2N1)	5	5
379	(O=C1NC(=O)C2=CC=CNC2=N1)	23	23
380	(O=C1OCCc2ccccc12)	27	19
381	(c1nncc2c[nH]cc12)	1	1
382	(O=S(=O)(N1CCCCC1)c2ccccc2)	12	12
383	(C(Sc1ocnn1)c2ccccc2)	18	18
384	(O=C(NCC1CCCO1)\C=C\c2ccccc2)	6	6

385	(O=C(CSc1nnc[nH]1)N2CCCCC2)	12	12
386	(O=C1NCC=C(CN2CCCCC2)N1)	6	6
387	(C(Oc1ccccc1)c2ocnn2)	17	17
388	(C(Cc1ccccc1)Cc2ocnn2)	11	11
389	(N(c1ccccc1)c2ncccn2)	2	2
390	(O=C(CSc1nccnn1)Nc2ccccc2)	22	12
391	(c1ncc2ccsc2n1)	63	22
392	(O=C(Nc1ccccc1)c2conc2)	11	11
393	(O=C(Cc1nccccc1)NC2CCCCC2)	8	8
394	(O=C(NS(=O)(=O)c1ccccc1)c2ccccc2)	13	5
395	(O=C(COc1ccccc1)NCc2oncn2)	13	13
396	(O=C(Nc1nccccc1)c2ccccc2)	13	13
397	(c1ccc(cc1)n2ccccc2)	9	0
398	(c1cc(cen1)c2nnc[nH]2)	7	7
399	(O=C(CCC(=O)c1ccccc1)Nc2ccccc2)	11	11
400	(O=S(=O)(N1CCNCC1)c2ccccc2)	17	15
401	(O=C1NC=C(C(=O)N1)S(=O)(=O)N2CCNCC2)	5	5
402	(O=C(Cn1ccccc1)NC2CCCCC2)	5	5
403	(O=C(Cn1ccccc1)NC2CCCCC2)	5	3
404	(O=S(=O)(NCC1CCCCC1)c2ccccc2)	12	12
405	(O=C(C1CCCCC1)N2CCNCC2)	8	8
406	(O=S(=O)(N1CCCCC1)c2ccccc2)	20	20
407	(C1C=CNe2nccn12)	31	13
408	(C1CCc2ccsc2CC1)	92	35
409	(O=C(CS(=O)Cc1coen1)N2CCNCC2)	9	9
410	(O=C(NC1CCCCC1)c2occcc2)	6	5
411	(O=C(CN1CCNC1=O)NCc2ccccc2)	12	12
412	(O=C(NCc1ccccc1)c2ccccc2)	22	21
413	(O=C(NCCc1ccccc1)c2ccccc2)	5	5
414	(C(SCc1ccccc1)c2occcc2)	23	0
415	(C1Cc2ccsc2C1)	62	56
416	(O=C1Nc2ccccc2SC=C1)	5	0
417	(O=C1Nc2ccccc2S(=O)C=C1)	5	5
418	(S1C=CC=Ne2ccccc12)	28	8
419	(O=C(C1CCNCC1)N2CCNCC2)	62	57
420	(C1CC(CN1)c2ccccc2)	8	8
421	(O=C(CS(=O)Cc1coen1)NCc2ccccc2)	12	12
422	(O=C(CCSCCc1ccccc1)Nc2ccccc2)	7	7
423	(O=C(CSCc1coen1)NCCc2ccccc2)	12	12
424	(O=C(CCSCc1ccccc1)NCc2ccccc2)	11	11
425	(O=C(CCSCc1ccccc1)NCCc2ccccc2)	10	10
426	(O=S(=O)(N1CCCCC1)N2CCCCC2)	8	8
427	(O=C(NCCCNc1ccccc1)c2ccccc2)	5	1
428	(O=C(NCCc1ccccc1)C2CNC(=O)C2)	6	6
429	(O=S(=O)(N1CCCCC1)N2CCNCC2)	5	5
430	(O=C1CN(C2CCCCC2)C(=O)CN1)	37	27
431	(C(NCc1ccccc1)C2CCCCC2)	11	1
432	(O=C(CSc1nccnn1)NCc2ccccc2)	8	0
433	(O=C(CSc1nccnn1)NC2CCCCC2)	7	0
434	(O=C(CSc1nccnn1)NCCc2ccccc2)	7	0
435	(c1ccc(cc1)c2cnenn2)	20	3
436	(O=C(CNC(=O))C=C/C(=O)Nc1ccccc1)NCc2occcc2)	6	0
437	(c1ncc2sccc2n1)	14	1
438	(N=C1NC=Ne2n[nH]cc12)	1	1
439	(C(Sc1ccccc1)c2ccccc2)	5	3
440	(O=S(=O)(NCCC1=CCCCC1)c2ccccc2)	5	3

441	(O=S(=O)(NC1CCCCC1)c2ccccc2)	8	8
442	(O=C(Oc1ccccc1)c2cnc2)	8	4
443	(o1ccc2[nH]ccc12)	73	46
444	(c1cc2sccc2[nH]1)	111	70
445	(O=S(=O)(NC1CCCCC1)c2ccccc2)	14	14
446	(c1ccc2ccccc2c1)	42	25
447	(C1Cc2cc[nH]c2CN1)	49	3
448	(O=S1(=O)Cc2c[nH]nc2C1)	14	14
449	(O=C1C=CN=C2SCC=NN12)	9	0
450	(O=C1CN=C2C=CN=CN12)	191	129
451	(O=S1(=O)N=CNCc2ccccc12)	17	17
452	(O=S1(=O)NC=NCc2ccccc12)	11	11
453	(O=C(c1ccccc1)c2cccs2)	29	0
454	(N(c1ccccc1)c2cccs2)	33	12
455	(O=C1Nc2cnnc2C=C1)	12	12
456	(O=C(Nc1ccccc1)C2CCNC2)	8	8
457	(O=C(C1CCNC1)N2CCNCC2)	22	18
458	(c1cn2nc[nH]c2n1)	11	11
459	(o1ccc2ncnc12)	11	11
460	(O=C1CNC(=O)c2ccccc2N1)	7	7
461	(O=C(Cc1ccc[nH]1)Nc2ccccc12)	12	12
462	(O=C(Nc1ccccc1)C2CCNCC2)	6	6
463	(O=C(Cc1ccc[nH]1)NCc2ccccc12)	9	9
464	(O=C(Cc1ccc[nH]1)N2CCNCC2)	11	11
465	(O=C1Nc2ccccc2N=C1)	66	64
466	(C1CC(CCN1)c2cnc2)	17	17
467	(O=C1NCCN(C2CCNCC2)C1=O)	47	47
468	(O=C1NN=C(N=C1)c2ccccc2)	5	5
469	(C1CN(CCN1)c2ccccc2)	3	3
470	(C1CN(CC=C1)c2ccccc2)	1	0
471	(C1CCN(CC1)C2CCNCC2)	14	14
472	(O=C(CN1CCNCC1)c2ccccc2)	3	3
473	(C1CC(=CCN1)c2ccccc2)	1	0
474	(C(CN1CCOCC1)C2CCNCC2)	1	1
475	(O=C1C=CNC(=C1)CN2CCCCC2)	9	9
476	(C1CN(CCO1)c2ccccc2)	2	2
477	(O=C(Nc1ccccc1)C2=NNC(=O)C=C2)	7	7
478	(C(Nc1ccccc1)c2ccccc2)	20	17
479	(C1CCN(CC1)c2ccccc2)	1	1
480	(O=C(CNc1ccccc1)NCCc2ccccc2)	8	8
481	(O=C(CSc1nnc[nH]1)c2ccccc2)	20	11
482	(O=C1NC=Cc2cnnc12)	44	36
483	(O=C(Nc1ccccc1)C2=CC=CNC2=O)	28	28
484	(O=C(Nc1nncs1)C2=CC=CNC2=O)	10	10
485	(O=C(Nc1con1)C2=CC=CNC2=O)	8	8
486	(O=S(=O)(Cc1cccs1)c2ccccc2)	7	7
487	(C1Cn2ccccc2CN1)	18	12
488	(O=S(=O)(Cc1occc1)c2ccccc2)	9	9
489	(O=C1NC2=C(CCC2)C=N1)	82	82
490	(N1C=CSc2nnc12)	30	30
491	(C1SCc2n[nH]cc12)	23	23
492	(O=C(Nc1ccccc1)C2=CNC(=O)C=C2)	9	9
493	(C1CCc2ccccc2C1)	11	5
494	(O=C1NC(=NC=C1)n2cccn2)	9	9
495	(O=C1NC2=C(CCCC2)C=N1)	19	19
496	(O=C(CSC1=NC(=O)NC=C1)NCc2ccccc2)	8	8

497	(C1CN(CCN1)c2cccn2)	7	7
498	(O=C1NC(=O)c2nc[nH]c2N1)	6	6
499	(O=C(NC1CNC(=O)C1)c2cccs2)	10	10
500	(O=C(NC1CNC(=O)C1)c2occc2)	12	12
501	(O=C(NC1CNC(=O)C1)c2ccccc2)	21	21
502	(O=C(CCc1ccccc1)NC2CNC(=O)C2)	12	12
503	(O=C(Cc1cccs1)NC2CNC(=O)C2)	10	10
504	(O=C(Cc1ccccc1)NC2CNC(=O)C2)	18	18
505	(O=C(COc1ccccc1)NC2CNC(=O)C2)	8	8
506	(O=C(Nc1ccccc1)Nc2nncs2)	10	8
507	(O=C(NC1CNC(=O)C1)\C=C/c2ccccc2)	8	8
508	(O=C(NC1CNC(=O)C1)C2CCCCC2)	10	10
509	(O=C(CSc1nncs1)N2CCCCC2)	4	4
510	(O=C1CCCN1c2ccccc2)	17	17
511	(O=C(NC1CCCCC1)Nc2nncs2)	6	6
512	(O=C(CCCc1ccccc1)NC2CNC(=O)C2)	6	6
513	(c1ncc2nn[nH]c2n1)	73	66
514	(C(N1CCOCC1)c2occc2)	9	9
515	(O=C(COc1ccccc1)NCCc2occc2)	5	5
516	(O=C(CSC1=NC=CC(=O)N1)Nc2ccon2)	5	5
517	(O=C(CSC1=NC=CC(=O)N1)N2CCOCC2)	6	6
518	(O=C(Cc1ccccc1)NCCc2cccs2)	9	8
519	(O=C(COc1ccccc1)NCCc2cccs2)	5	5
520	(O=C(CCc1ccccc1)NCCc2escn2)	6	6
521	(c1encc(c1)c2nncs2)	10	10
522	(c1ennc(c1)c2ncs2)	31	19
523	(C(c1ccccc1)n2ccnc2)	5	5
524	(O=C(CSc1ncc[nH]1)Nc2nncs2)	13	13
525	(O=C1NC(=O)C=C(N1)N2CCNCC2)	5	5
526	(O=C(CSc1ncc[nH]1)Nc2occc2)	5	5
527	(O=C(CSc1ncc[nH]1)Nc2ccccc2)	11	11
528	(O=C(COc1ccccc1)NCCc2escn2)	7	7
529	(O=C(NCCc1ncs1)c2occc2)	5	5
530	(O=C(Cc1ccccc1)NCCc2ncs2)	7	7
531	(O=C(NCCc1ncs1)c2cccs2)	5	5
532	(O=C(NCCc1scn1)c2occc2)	8	8
533	(C1CCC(NC1)c2ccnc2)	5	5
534	(c1en2cnnc2s1)	16	14
535	(O=S(=O)(c1ccccc1)c2coen2)	2	2
536	(C1CN(CCO1)c2ocnc2)	5	5
537	(C1CCN(CC1)c2ocnc2)	39	30
538	(O=C(Nc1ccccc1)N2CCCC2)	7	7
539	(O=S(=O)(NCc1ccccc1)c2cccs2)	7	7
540	(O=C(CNS(=O)(=O)c1cccs1)N2CCNCC2)	8	8
541	(O=C(CNS(=O)(=O)c1cccs1)NCc2ccccc2)	7	7
542	(o1ncc(c1)c2ccccc2)	21	21
543	(O=C1NN=C(C=C1)c2ccncc2)	5	5
544	(C1CCCc2[nH]ccc2CC1)	8	4
545	(O=C1NN=C2CCCCC2=C1)	16	16
546	(c1ccc2enccc2c1)	75	59
547	(O=S(=O)(NCc1ccccc1)c2cnnc2)	12	12
548	(O=S(=O)(Nc1ccccc1)c2cnnc2)	8	8
549	(O=S(=O)(N1CCNCC1)c2cnnc2)	2	2
550	(o1ccc2ccncc2)	27	24
551	(C(Sc1ncccn1)c2oncn2)	20	11
552	(c1ccc2sncc2c1)	92	88

553	(O=C(CCc1cc[nH]c1)NCc2ccccc2)	12	12
554	(O=C(CCc1cc[nH]c1)Nc2ccccc2)	11	11
555	(O=C(CCc1cc[nH]c1)N2CCNCC2)	7	7
556	(C1CN(CCN1)c2oncc2)	8	6
557	(O=C(NCc1occc1)c2concc2)	5	5
558	(O=C1C=CN=C2SC=CN12)	311	311
559	(C1CC(=CCN1)c2nocn2)	5	5
560	(O=C(COc1ccccc1)Nc2ccn[nH]2)	12	1
561	(O=C(Nc1ccn[nH]1)c2ccccc2)	11	2
562	(O=C1NC(=O)c2cncnc2N1)	45	45
563	(c1enc(nc1)n2cccn2)	16	9
564	(C1CCN(CC1)c2ccn[nH]2)	14	14
565	(C1CCN(CC1)c2oncn2)	13	11
566	(C1COc2ccccc2C1)	60	60
567	(C1CC2(CCN1)NC=CN=C2)	7	6
568	(O=C(CSc1nnc[nH]1)N2CCNCC2)	9	9
569	(O=C(CSc1nnc[nH]1)NCc2ccccc2)	9	9
570	(O=C1N=CNc2ncccc12)	13	9
571	(C1CCN2CCN=C2CC1)	26	26
572	(C1CCC2=NCCN2C1)	17	17
573	(O=C(CSc1nnc[nH]1)NCCCc2ccccc2)	7	7
574	(O=C(NCc1nnn[nH]1)C2CNC(=O)C2)	5	5
575	(O=C(CSc1nnc[nH]1)NCCCc2ccccc2)	7	7
576	(O=C1NN=Cc2n[nH]cc12)	24	23
577	(O=C(CCc1ccc[nH]1)NCc2ccccc2)	10	10
578	(O=C(CCc1ccc[nH]1)NCCCN2CCNCC2)	3	3
579	(O=C(CCc1ccc[nH]1)Nc2ccccc2)	9	9
580	(C1CN2CCN=C2CO1)	21	21
581	(O=S(=O)(N1CCNCC1)c2cc[nH]n2)	12	12
582	(O=S(=O)(Nc1ccccc1)c2cc[nH]n2)	11	11
583	(O=C(Cc1ccccc1)NCCS(=O)(=O)N2CCNCC2)	5	5
584	(o1cnnclc2c[nH]nn2)	5	5
585	(C1C=Nc2ccccc2N=C1)	24	9
586	(O=S(=O)(N1CCOCC1)c2cn[nH]c2)	12	12
587	(O=S(=O)(NCc1ccccc1)c2cc[nH]n2)	4	4
588	(O=C(Nc1ccccc1)C2=NNC(=O)NC2=O)	12	12
589	(O=C(N1CCNCC1)C2=NNC(=O)NC2=O)	7	7
590	(O=C(N1CCCCC1)C2=NNC(=O)NC2=O)	5	5
591	(O=C1NC(=O)N(N=C1)c2ccccc2)	8	8
592	(O=C1Nc2ccccc2S1)	27	24
593	(O=C1NC=CN2CCCCC12)	28	28
594	(C1Oc2ccccc2C=C1)	10	10
595	(O=C1NC=Nc2ncccc12)	28	28
596	(O=C1Nc2cncnc2N1)	44	44
597	(O=C(N1CCNCC1)c2ccc[nH]2)	18	18
598	(O=S(=O)(Nc1ccccc1)c2c[nH]cn2)	6	6
599	(O=S(=O)(N1CCNCC1)c2c[nH]cn2)	7	7
600	(C1CCCN(CC1)c2ccenn2)	21	13
601	(O=S(=O)(N1CCCCC1)c2cc[nH]c2)	27	27
602	(C1CCCc2senc2CC1)	5	4
603	(C1CN=C2SC=CC2=C1)	5	4
604	(O=C(NCCc1csn1)c2cccs2)	3	3
605	(O=C(CSC1=NC=CC(=O)N1)NCc2ccccc2)	9	9
606	(O=C1NCCCN2ncccc12)	29	29
607	(C1SC=Cc2[nH]ncc12)	42	39
608	(o1cccc1c2ocnc2)	12	12

609	(O=C1Nc2ccccc2C=N1)	61	42
610	(C1CCN(C1)c2ccccc2)	33	25
611	(O=C(NCCc1cn[nH]c1)Nc2ccccc2)	29	29
612	(O=S(=O)(NCc1ccc[nH]1)c2ccccc2)	12	12
613	(O=C1Nc2[nH]ncc2N=C1)	4	4
614	(C1Cc2ncc2CN1)	78	75
615	(O=C1NC=Cn2nccc12)	22	22
616	(C1CCN(C1)c2ncccn2)	19	19
617	(O=C(N1CCNCC1)c2cnccn2)	14	14
618	(O=C1Oc2cnccc2C=C1)	11	11
619	(O=S(=O)(N1CCCCC1)c2c[nH]cn2)	43	43
620	(O=C(Cc1cccn1)N2CCNCC2)	5	5
621	(O=C1NCCN(C2CCCCC2)C1=O)	7	7
622	(O=C(N1CCOCC1)c2cccs2)	6	6
623	(O=C1NCCN(Cc2ccccc2)C1=O)	4	4
624	(c1ccc2[nH]nnc2c1)	46	44
625	(O=C(NC1c1ccn1)NC2CCCCC2)	7	7
626	(O=C(NC1c1ccn1)Nc2ccccc2)	7	7
627	(O=C(N1CCCCC1)c2cc[nH]c2)	21	21
628	(O=S(=O)(NC1c1ccn1)c2ccccc2)	4	4
629	(O=C(CN1CCCN1(=O)=O)Nc2ccccc2)	27	27
630	(O=C(CN1CCCN1(=O)=O)NCc2ccccc2)	4	4
631	(c1cc2[nH]ncc2cn1)	79	19
632	(O=C1CC=Ne2ccccc2N1)	17	15
633	(O=C1NC(=O)c2sccc2N1)	255	192
634	(C1CN2C=NC=CC2=N1)	12	12
635	(C(N1CCCC1)c2oncc2)	12	12
636	(O=C1CSc2[nH]ncc2N1)	27	27
637	(C(N1CCOCC1)c2oncc2)	12	12
638	(O=S(=O)(N1CCCC1)c2c[nH]cn2)	8	8
639	(c1cnnc(c1)n2cccn2)	70	68
640	(O=C(Cc1cn[nH]c1)N2CCNCC2)	6	5
641	(O=C1CC(CN1)c2oncn2)	10	10
642	(O=C(N1CCCCC1)c2cc[nH]c2)	17	17
643	(O=C(CNS(=O)(=O)c1ccsc1)NCc2ccccc2)	10	10
644	(C(Nc1ccccc1)c2ocnn2)	29	3
645	(C(N1CCNCC1)c2ocnn2)	29	17
646	(O=C1C=CS(=O)(=O)N1Cc2ccccc2)	23	21
647	(O=C1NCc2cccn12)	45	40
648	(o1nnc1c2ccn[nH]2)	22	22
649	(O=S(=O)(c1ccccc1)c2cn[nH]c2)	23	23
650	(c1ccn2ccccc2c1)	33	25
651	(O=C1NC=CN=C1NCc2cccs2)	5	5
652	(O=C1NC=CN2CCCC12)	8	8
653	(O=S(=O)(NCC1CCNCC1)c2cc[nH]c2)	8	8
654	(O=S1(=O)CCCN1c2ccccc2)	19	19
655	(o1ccc(n1)c2cccs2)	9	9
656	(O=C1NC=Ne2c[nH]nc12)	30	8
657	(c1ccn(c1)c2nncs2)	15	9
658	(C1CN(CCN1)c2nncs2)	4	4
659	(C1CCN(C1)c2nncs2)	18	18
660	(O=S(=O)(NCC1CCCCC1)c2ccsc2)	6	6
661	(O=S(=O)(NCc1ccccc1)c2ccsc2)	11	11
662	(C1CCC2=C(CC1)NCC2)	22	15
663	(O=C(N1CCNCC1)c2cn[nH]c2)	86	53
664	(O=C(NC1CCNCC1)c2cn[nH]c2)	6	6

665	(O=S(=O)(N1CCCCC1)c2ccsc2)	6	6
666	(O=C1CNc2ccccc12)	12	12
667	(c1ccc2nnccc2c1)	10	6
668	(c1ccn2nccc2c1)	29	29
669	(O=C1NN=C(C=C1)N2CCNCC2)	42	42
670	(O=C(CSc1oncn1)N2CCNCC2)	8	8
671	(O=C(CSc1oncn1)Nc2ccccc2)	9	8
672	(O=C(CSc1oncn1)N2CCCCC2)	6	6
673	(O=C(NC1=CNC(=O)NC1=O)Nc2ccccc2)	12	12
674	(O=C(CNS(=O)(=O)c1cn[nH]c1)Nc2ccccc2)	10	10
675	(O=C(NCc1ccccc1)C2CCCN2)	12	12
676	(C1CC2(CCN1)N=CC=N2)	11	2
677	(C1Nc2ccccc2C=N1)	5	5
678	(O=C(NCCCN1CCNCC1)c2cn[nH]c2)	23	19
679	(O=C1CS(=O)(=O)C2(N1)C=CNC2=O)	40	40
680	(C1CN(CCN1)c2ncccn2)	8	8
681	(O=C(NCCN1CCNCC1)c2cn[nH]c2)	8	8
682	(O=C(N1CCOCC1)c2cn[nH]c2)	5	5
683	(O=C(NCCN1CCOCC1)c2cn[nH]c2)	8	8
684	(O=C1CCSC2(N1)C=CNC2=O)	9	8
685	(C1CCCN(CC1)c2nncs2)	14	12
686	(C1CN(CCO1)c2nncs2)	5	5
687	(C1CCc2oncc2CC1)	9	9
688	(O=C(NCc1occc1)c2c[nH]nn2)	7	7
689	(O=C(Cn1ccccc1)N2CCNCC2)	8	8
690	(O=S(=O)(NCc1ccccc1)c2nncs2)	8	8
691	(O=C1NCc2cc[nH]c12)	16	10
692	(C(N1CCNCC1)c2nocn2)	7	7
693	(O=C(NCCc1ccccc1)c2ccn[nH]2)	8	8
694	(O=C(Nc1ccccc1)c2ccn[nH]2)	10	10
695	(O=C1NN=C2C=CNC=C12)	11	11
696	(O=S(=O)(Cc1ocn1)c2ccccc2)	10	10
697	(C(Sc1ccccc1)c2coen2)	29	10
698	(O=C(N1CCNCC1)c2ccccc2)	7	1
699	(O=C1CN(CCc2ccncc2)C(=O)N1)	7	7
700	(O=C1CN(Cc2ccncc2)C(=O)N1)	9	9
701	(O=C(CSc1oncn1)Nc2ccccc2)	5	5
702	(C1CN(CCN1)c2oncn2)	10	10
703	(C1CCc2cnen2CC1)	30	11
704	(c1enc2nnnn2c1)	225	128
705	(O=C(NC1CCCCC1)C2CNCC(=O)N2)	11	11
706	(O=C1CNCCN1c2ccccc2)	4	4
707	(O=C(NCc1ccccc1)c2ccccc2)	1	1
708	(O=C(N1CCNCC1)c2ccccc2)	1	1
709	(O=C(N1CCNCC1)c2ccn[nH]2)	13	13
710	(c1cc2ccsc2[nH]1)	11	1
711	(O=C(COCc1cccon1)N2CCNCC2)	3	3
712	(O=C1NC=CC(N1)c2ccccc2)	19	19
713	(c1cc2cn[nH]c2s1)	62	38
714	(C1SC=CN=C2C=CC=C12)	8	2
715	(O=C1CCC=C2SCNCCN12)	8	7
716	(C1CCc2nccccc2C1)	39	6
717	(O=C1NC=Cc2[nH]ccc12)	24	24
718	(O=S(=O)(Nc1ccccc1)c2conc2)	5	5
719	(O=C1OC=Cc2[nH]ncc12)	39	39
720	(O=C(CNS(=O)(=O)c1cn[nH]c1)NCc2ccccc2)	12	12

721	(O=C(CNS(=O)(=O)c1conc1)N2CCNCC2)	12	12
722	(O=C(CNS(=O)(=O)c1conc1)NCc2ccccc2)	12	12
723	(O=C(CNS(=O)(=O)c1conc1)NCCc2ccccc2)	7	7
724	(O=C(Cc1oncn1)Nc2ccccc2)	10	10
725	(c1ccc(nc1)c2cccn2)	3	3
726	(c1cnnc(c1)c2cccs2)	17	13
727	(O=C1NC(=O)c2occc2N1)	83	65
728	(C1C=COc2ncnc12)	98	25
729	(O=C(NCCCN1CCNCC1)c2occc2)	1	1
730	(O=C(N1CCNCC1)c2occc2)	13	10
731	(O=C(NC1CCCCC1)c2occc2)	5	5
732	(C(N1CCNCC1)c2ccsc2)	15	8
733	(O=C1Nc2ncccc2N=C1)	110	110
734	(O=C1Nc2ccenc2SC=C1)	30	30
735	(c1cc2cnnc2c1)	86	63
736	(O=C(NCCc1cccc1)C2CCCNC2)	5	5
737	(O=C1COc2ccnc2N1)	11	11
738	(O=C(NC1cccc1)C2CCCNC2)	12	12
739	(O=C(Nc1cccc1)C2CCCNC2)	11	11
740	(O=C1CCOc2cccc12)	18	17
741	(C1Cc2c[nH]nc2C=C1)	63	37
742	(O=C(NCC12CC3CC(CC(C3)C1)C2)c4cccc4)	4	0
743	(C1Cn2ccnc2S1)	101	55
744	(O=C1NC=Ne2ccoc12)	77	47
745	(O=C(Nc1cccc1)c2oncc2)	8	8
746	(C1OC=Cc2oncc12)	13	12
747	(O=C1NC=Cn2cccc12)	103	103
748	(c1cc2cnnc2c1)	6	5
749	(O=C1C=CN=C2CCCN12)	29	29
750	(O=C1NC=CN(Cc2ccccc2)C1=O)	9	9
751	(O=C(CN1C=CNC(=O)C1=O)Nc2ccccc2)	39	39
752	(C(c1cccc1)c2ocnn2)	15	15
753	(O=C1NC=CN(C1=O)c2ccccc2)	12	12
754	(O=C(NCCCN1CCNCC1)c2ocnn2)	6	6
755	(O=C(CN1C=CNC(=O)C1=O)NCc2ccccc2)	8	8
756	(O=C(CN1C=CNC(=O)C1=O)c2ccccc2)	5	5
757	(O=C(CN1C=CNC(=O)C1=O)NCCc2ccccc2)	6	6
758	(O=S1(=O)Cc2cn[nH]c2C=C1)	40	40
759	(O=C1C=CN=C2CCCN12)	27	27
760	(O=C(NC1cccc1)c2cc[nH]c2)	7	7
761	(O=C(NCCc1cccc1)c2cc[nH]c2)	5	5
762	(O=C1NC=Ne2ncnc12)	17	14
763	(O=C(Nc1nncs1)C2CCCCC2)	12	12
764	(O=S(=O)(Nc1cccc1)c2nncs2)	11	11
765	(O=S(=O)(N1CCCCC1)c2nncs2)	24	24
766	(O=C(CO1cccc1)Nc2nncs2)	19	19
767	(O=S(=O)(N1CCOCC1)c2nncs2)	8	8
768	(O=C(Nc1nncs1)c2occc2)	12	12
769	(O=S(=O)(NC1occc1)c2nncs2)	6	6
770	(O=S(=O)(N1CCCCC1)c2nncs2)	9	9
771	(C(N1CCOCC1)c2ccnc2)	9	9
772	(O=C(Nc1nncs1)c2cccs2)	8	8
773	(O=S(=O)(N1CCNCC1)c2nncs2)	12	12
774	(O=S(=O)(N1CCCCC1)c2nncs2)	10	10
775	(c1cnc2snc2c1)	31	29
776	(O=C(Nc1cccc1)C(=O)c2cc[nH]c2)	12	12

777	(O=C(Nc1nncs1)C(=O)c2cc[nH]c2)	11	11
778	(O=C(Nc1cccc1)c2cn[nH]c2)	12	4
779	(O=C1NN=Cn2cccc12)	270	261
780	(O=S1(=O)NC=CC(=N1)c2cccs2)	10	10
781	(O=S1(=O)NC=CC(=N1)c2occc2)	7	7
782	(O=C1NN=Cc2sccc12)	69	66
783	(O=C1NC=NC2=C1CNCC2)	25	25
784	(O=C(Nc1cccc1)C2=CC=NS(=O)(=O)N2)	12	12
785	(O=C(NCc1cccc1)C2=CC=NS(=O)(=O)N2)	5	5
786	(C1OC=Cc2ncc12)	9	0
787	(C1OC=Cc2[nH]ncc12)	34	34
788	(C1SC=Cc2oncc12)	8	8
789	(c1ccc2[nH]ncc2c1)	17	17
790	(O=S1(=O)C=CC(=N1)N2CCNCC2)	12	12
791	(O=S1(=O)C=CC(=N1)NCCCN2CCNCC2)	7	7
792	(N(c1cccc1)c2nncs2)	9	9
793	(O=C1CNCc2ccsc2N1)	5	3
794	(C1Cc2cn[nH]c2C=C1)	12	12
795	(C1CCC(CC1)Nc2nncs2)	12	12
796	(O=C1NC2(CCNC2)OC=C1)	12	5
797	(O=C1NN=Cn2nccc12)	10	10
798	(c1cc2[nH]ccn2n1)	20	20
799	(O=C(NCc1cn[nH]c1)Nc2cccc2)	6	0
800	(C(Nc1nncs1)c2cccc2)	11	11
801	(o1nenc1c2cccc2)	142	138
802	(O=C1CCc2cn[nH]c2N1)	28	24
803	(C1NC=Cn2cccc12)	16	10
804	(O=C(Nc1cccc1)Nc2cccs2)	8	8
805	(O=C(CSc1ocnn1)Oc2cccc2)	15	15
806	(O=C1C=CNC2=C1CCCC2)	7	7
807	(O=C(Nc1cnon1)c2cccc2)	12	12
808	(O=C1NC=NC2=C1CCCC2)	17	17
809	(O=C(NCCSCc1cccc1)C2CCNCC2)	6	6
810	(o1ennc1c2cccs2)	24	24
811	(O=C1NC=NC2=C1CCCC2)	17	16
812	(O=C(NCc1occc1)c2ccon2)	6	1
813	(C1CNc2nccn2C1)	7	5
814	(C1CC=Nc2cccc2S1)	5	0
815	(O=C(Nc1cccs1)c2ccon2)	26	18
816	(O=C1CCNC(C1)c2cccs2)	5	5
817	(O=C(N1CCNCC1)c2ccon2)	12	12
818	(O=C1CSCc2cn[nH]c2N1)	11	11
819	(O=C(CSc1ocnn1)NCc2cccc2)	6	6
820	(C(CSc1ocnn1)Oc2cccc2)	10	10
821	(O=C(CSc1ocnn1)NC2CCCC2)	8	8
822	(O=C(Nc1cccc1)c2nc[nH]n2)	24	24
823	(o1enc(n1)c2cccn2)	18	8
824	(O=C(CCc1oncn1)Nc2cccc2)	29	29
825	(o1nenc1c2cccn2)	17	11
826	(o1enc(n1)c2cccn2)	14	14
827	(C1CCc2senc2C1)	35	35
828	(O=C1NC(=O)c2cc[nH]c2N1)	29	29
829	(C(Cc1ccnc1)N2CCNCC2)	23	23
830	(O=C(COc1cccc1)Nc2cccc2)	15	2
831	(C(Nc1cccn1)c2cccc2)	31	30
832	(O=C(NCc1cccs1)c2cccc2)	11	4

833	(O=C(NCc1cccc1)c2occc2)	5	5
834	(O=C(COc1cccc1)NCc2cccs2)	12	6
835	(C(Nc1cccc1)c2occc2)	23	23
836	(C(Cc1cccc1)N2CCNCC2)	20	20
837	(O=C(Oc1cccc1)c2ccon2)	12	5
838	(C1CCC(CC1)N2CCOCC2)	21	21
839	(O=C(NCc1cccc1)c2con2)	17	17
840	(O=C(COc1cccc1)NCC2CCCCC2)	10	10
841	(C1CCCCC1)	9	7
842	(O=C(NCC1CCCCC1)c2ccccc2)	10	10
843	(C1CC(CCN1)c2ocnn2)	24	24
844	(C(Oc1cccc1)c2nccs2)	12	12
845	(o1cccc1c2nc[nH]n2)	43	38
846	(C1CCC(CC1)N2CCNCC2)	3	3
847	(C1CCC2(CC1)CCNC=N2)	37	25
848	(c1enn(c1)c2nccs2)	30	5
849	(O=C1C=CC=Cc2cocc12)	24	21
850	(O=S1(=O)CCC(C1)NCc2occc2)	22	22
851	(o1cnc(n1)c2cccs2)	5	5
852	(O=C(Oc1cccc1)c2cnns2)	9	9
853	(c1esc(c1)c2ccsc2)	15	13
854	(C(NCc1cccc1)C2CCCO2)	22	22
855	(O=C(Nc1cccc1)c2csnn2)	11	11
856	(O=S(=O)(c1cccc1)c2c[nH]cn2)	7	6
857	(c1esc(c1)c2ncsn2)	6	6
858	(O=C(COc1cccc1)Nc2nens2)	11	11
859	(C(N1CCCCC1)c2ccccc2)	17	16
860	(O=C(CSc1ccnnc1)Nc2ccccc2)	12	12
861	(o1cccc1c2occn2)	56	33
862	(C(Nc1ocnc1)c2cccnc2)	5	5
863	(C1CN(CCN1)c2ocnc2)	42	25
864	(C1Cc2cnnc2C1)	1	1
865	(O=C(COc1cccc1)Nc2oncc2)	12	12
866	(O=C1CC(CN1)c2nncs2)	7	7
867	(N(c1cccc1)c2ccnnc2)	37	28
868	(O=C1NC=NC(=C1)c2ccccc2)	1	1
869	(C(CNCc1occc1)CSc2nnn[nH]2)	5	5
870	(C(CNCc1cccc1)CSc2nnn[nH]2)	23	23
871	(C(Nc1cnnc1)c2occc2)	41	24
872	(o1ncc(n1)c2nnc[nH]2)	36	36
873	(O=C(CSc1cccn1)c2ccccc2)	5	1
874	(C(CSc1nnn[nH]1)NCc2occc2)	6	5
875	(O=C1CCC2=C(NC=NC2=O)N1)	11	11
876	(C1CCN(CC1)c2cnccn2)	21	21
877	(O=C1CCC2=C(N1)N=CNC2=O)	7	7
878	(O=C(Nc1oncc1)c2ccccc2)	11	11
879	(O=C1NN=Nc2ccccc12)	12	12
880	(O=C(NCC1CCNC1)Nc2ccccc2)	2	2
881	(O=C(NCCN1CCNCC1)Nc2ccccc2)	6	6
882	(O=C(NCc1onen1)c2occc2)	9	9
883	(o1nc2ncnc2n1)	2	1
884	(O=C(NCC1CCNCC1)Nc2ccccc2)	6	5
885	(C1CCN(CC1)c2ncccn2)	8	8
886	(O=C(Nc1cccc1)C2=CC=NNC2=O)	12	12
887	(O=C(Nc1nccs1)C2=CC=NNC2=O)	5	5
888	(O=C1NN=CC2=C1CCCC2)	12	5

889	(O=C(Cc1csn1)Nc2ccccc2)	12	12
890	(O=C(Nc1ccccc1)c2ccnnc2)	33	33
891	(O=C(Nc1nncs1)c2ccnnc2)	14	14
892	(C(Sc1ncccn1)c2ccccc2)	38	15
893	(C(NCc1ccccc1)c2occc2)	41	22
894	(C(NCc1occc1)c2occc2)	24	24
895	(O=C(NC1CCS(=O)(=O)C1)c2ccnnc2)	12	12
896	(O=C(Nc1cccs1)c2ccnnc2)	10	7
897	(C(NCc1cccs1)c2occc2)	4	4
898	(O=C(NCc1ccccc1)c2cn[nH]n2)	12	12
899	(O=C(NCc1cccs1)c2cn[nH]n2)	7	7
900	(c1ccc(cc1)n2nccn2)	12	12
901	(O=C(NCCCCc1ccccc1)c2cn[nH]n2)	8	8
902	(O=C(NCCCCc1ccccc1)c2cn[nH]n2)	6	6
903	(O=C(NCc1ccccc1)c2cn[nH]n2)	6	6
904	(O=C(NCc1ccccc1)c2cn[nH]n2)	7	7
905	(O=C(NCCc1ccccc1)c2cn[nH]n2)	13	13
906	(O=C(NCc1ccccc1)c2ccnnc2)	9	6
907	(C1CCC(C1)c2ccccc2)	12	12
908	(C1CC(CCN1)c2nocn2)	5	5
909	(C1CC(CCN1)c2oncn2)	9	9
910	(O=C1NC=C(C2N3CC4CN2CC(C3)C4=O)C(=O)N1)	5	5
911	(C(CN1CCOCC1)C2CCCCN2)	9	9
912	(c1nncc2[nH]nnc12)	13	8
913	(c1ccc(cc1)c2nnn[nH]2)	8	8
914	(C(CN1CCCCC1)C2CCCCN2)	11	11
915	(O=S(=O)(c1ccccc1)c2cscn2)	15	13
916	(O(c1cccnn1)c2nccn2)	18	12
917	(O=C1C2CN3CC1CN(C2)C3c4ccccc4)	5	5
918	(c1csc(c1)c2c[nH]cn2)	3	0
919	(c1ccc(cc1)c2c[nH]cn2)	6	0
920	(O=C1NN=C(Oc2nccn2)C=C1)	11	11
921	(C1CC(CCN1)c2nnn[nH]2)	10	10
922	(N1C=NC=Nc2ccccc12)	43	23
923	(C1CC(CCN1)Oc2ccccc2)	4	4
924	(C1CCC(CC1)Nc2ocnc2)	4	4
925	(C(Nc1ocnc1)c2ccccc2)	24	13
926	(C1CCCN(CC1)c2ocnc2)	14	4
927	(N(c1ocnc1)c2ccccc2)	10	10
928	(O=C1NCNc2ccsc12)	7	5
929	(o1ccnc1\C=C\c2ccccc2)	9	9
930	(C(Cc1ccccc1)Nc2ocnc2)	13	13
931	(N(c1ccccc1)c2cn[nH]n2)	10	0
932	(O(c1ccccc1)c2cccnn2)	6	0
933	(O=C1Cc2ccccc2CN1)	10	8
934	(c1cncc(c1)c2nc[nH]n2)	17	17
935	(O=C(CCCc1oncn1)NCc2occc2)	3	3
936	(C1CCc2nocc2C1)	11	11
937	(O=C(Cc1ccccc1)NCc2oncn2)	5	5
938	(C(NC1CCCCC1)c2oncn2)	9	7
939	(o1ccnc1c2cccs2)	29	5
940	(O=C(CSC1=NC=CC(=O)N1)N2CCNCC2)	6	6
941	(O=C1NC(=NC=C1)SCCOc2ccccc2)	6	3
942	(O=C1NC(=NC=C1)Nc2ccccc2)	11	11
943	(O=C(NCc1occc1)c2ccnnc2)	4	3
944	(O=C1NC2(C=CNC2=O)C=C1)	11	11

945	(O=C(N1CCNCC1)c2ccnnc2)	4	4
946	(O=C(COc1ccccc1)NCc2occcc2)	3	1
947	(O=C(NCc1occcc1)c2ccccc2)	2	0
948	(O=C1Nc2ccccc2C1=O)	4	4
949	(C1CC2OCOC2CO1)	11	11
950	(c1ccc2snnc2c1)	22	22
951	(O=C(Nc1ccsc1)c2ccccc2)	6	5
952	(O=C1NN=Cc2c[nH]cc12)	85	85
953	(O=C(Nc1ccccc1)Nc2ccsc2)	7	4
954	(O=S(=O)(NCc1csn1)c2ccccc2)	9	9
955	(C1SC=Cc2sccc12)	20	3
956	(O=C1NC2(CCCCC2)N=C1)	12	12
957	(O=C1NCN=C1c2ccccc2)	12	12
958	(O=C1NC2(CCCCC2)N=C1)	12	12
959	(O=C(Nc1ccccc1)c2cccs2)	17	13
960	(C1CCc2c[nH]nc2CC1)	3	3
961	(C(NCc1cc[nH]c1)c2ccccc2)	30	30
962	(C(CNCc1cc[nH]c1)Cc2ccccc2)	7	7
963	(C(Cc1ccccc1)NCc2cc[nH]c2)	12	12
964	(C(NC1CCCCC1)c2cc[nH]c2)	10	10
965	(C(NC1CCCCC1)c2cc[nH]c2)	5	5
966	(C1CCc2n[nH]cc2C1)	12	12
967	(c1cc2nc[nH]e2cn1)	111	38
968	(C(Nc1cnsc1)c2ccccc2)	11	4
969	(O=C(CNS(=O)(=O)c1c[nH]cn1)Nc2ccccc2)	6	6
970	(O=C(c1ccccc1)c2ccc[nH]2)	9	9
971	(C(NCc1ccsc1)c2ccccc2)	13	13
972	(O=C1CSc2nccccc2N1)	5	5
973	(O=C(CS(=O)(=O)Cc1coen1)NCc2ccccc2)	12	12
974	(O=C1NCCc2ccccc12)	12	10
975	(O=C(N1CCCCC1)c2cccs2)	8	8
976	(c1ccn(c1)c2cccs2)	12	10
977	(O=C(NCc1ccccc1)c2cn[nH]c2)	6	6
978	(O=C1CCCN1c2nncs2)	7	3
979	(C(Cn1ccccc1)c2ccccc2)	12	12
980	(C1OC=Cc2sccc12)	68	36
981	(O=C1NN=Cc2ccsc12)	6	6
982	(C1CNc2nccccc2N1)	6	6
983	(O=C(CS(=O)(=O)Cc1coen1)N2CCNCC2)	9	9
984	(O=C(CS(=O)(=O)Cc1coen1)NCCc2ccccc2)	12	12
985	(O=C(CS(=O)(=O)Cc1coen1)NCCCc2ccccc2)	4	4
986	(O=C(Nc1ccc1)C23CC4CC(CC(C4)C2)C3)	7	3
987	(C(N1CCNCC1)c2oncn2)	5	5
988	(N(c1oncn1)c2oncn2)	4	1
989	(C1SC=Cc2n[nH]cc12)	1	1
990	(O=C(CNS(=O)(=O)c1cn[nH]c1)NCCc2ccccc2)	8	8
991	(O=C(CNS(=O)(=O)c1cn[nH]c1)N2CCNCC2)	8	8
992	(O=C1CN(CCC2=CCCC2)C(=O)N1)	12	12
993	(O=C1CN(C2CCCC2)C(=O)N1)	12	12
994	(o1cc2ccccc2n1)	11	4
995	(O=C1CN(Cc2ccccc2)C(=O)N1)	5	5
996	(O=C1CN(CCc2cccs2)C(=O)N1)	12	12
997	(c1cc2sccc2cn1)	38	10
998	(O=C(CCCc1cccn1)NCCc2ccccc2)	5	2
999	(O=C1NC(=O)N2CSCC12)	7	7
1000	(O=C(NCc1nnc[nH]1)C23CC4CC(CC(C4)C2)C3)	8	8

1001	(O=C1CCCc2[nH]ccc12)	60	7
1002	(O=C(CNS(=O)(=O)C1=CNC(=O)NC1=O)Nc2ccccc2)	12	12
1003	(O=C(CNS(=O)(=O)C1=CNC(=O)NC1=O)NCc2ccccc2)	9	9
1004	(O=C(CS(=O)(=O)Cc1cocn1)NCCCN2CCNCC2)	4	4
1005	(O=C(CS(=O)(=O)Cc1cocn1)Nc2ccccc2)	12	12
1006	(O=C(NCCCN1CCNCC1)C2CCNCC2)	3	3
1007	(O=C(Cn1ccnn1)Nc2ccccc2)	12	12
1008	(C(=C\c1nc[nH]n1)/c2ccccc2)	7	2
1009	(O=C(CSc1nnn[nH]1)NCCc2ccccc2)	6	6
1010	(O=C(OCc1ccccc1)c2cn[nH]n2)	7	7
1011	(O=C1NC(C=C1)c2ccccc2)	9	3
1012	(O=C(Nc1ccccc1)C2=CC(=O)NC(=O)N2)	1	1
1013	(C1NC=Nc2ncn12)	3	3
1014	(O=C(Nc1nncn1)NS(=O)(=O)Oc2ccccc2)	5	5
1015	(O=C(NC1NC(=O)NC1=O)c2ccccc2)	7	7
1016	(O=C(Nc1ccccc1)\C=C\c2cc[nH]c2)	5	2
1017	(N=C1NC(=CS1)c2ccccc2)	3	3
1018	(O=S(=O)(Nc1ccccc1)c2ccccc2)	12	7
1019	(O=C(CNS(=O)(=O)c1ccccc1)NCCc2ccccc2)	10	8
1020	(O=S(=O)(NCCSc1nnn[nH]1)c2ccccc2)	5	5
1021	(O=C1C=CN=C2NC=NN12)	1	1
1022	(O=C(CNS(=O)(=O)c1ccccc1)NCCSc2ccccc2)	10	7
1023	(c1ccc(nc1)c2cccs2)	13	0
1024	(c1ccc(cc1)c2ccccc2)	23	5
1025	(O=C(Nc1ccccc1)\C=C/c2ccccc2)	1	0
1026	(O=C(NCc1ccccc1)C(=O)Nc2ccccc2)	10	10
1027	(O=C(CCN1CCNCC1)Nc2ccccc2)	8	8
1028	(C(=C/c1nccs1)/c2ccccc2)	7	0
1029	(o1enc(\C=C\c2ccccc2)n1)	5	5
1030	(C1CCc2ccoc2C1)	12	6
1031	(O=C(CNc1ccccc1)NCCSc2ccccc2)	8	8
1032	(O=C(NCCNCc1ccccc1)c2cnon2)	6	6
1033	(O=C1C=CC=C2OC=CN12)	6	6
1034	(O=C(CSc1ncccn1)NCCc2ccccc2)	5	0
1035	(O=C(CSc1ncccn1)N2CCCCC2)	8	4
1036	(o1ccccc1c2cnncn2)	22	17
1037	(O=C1CNC=C2CC=CC=C2N1)	4	4
1038	(O=C1CNC2=CC=CCC2=CN1)	2	2
1039	(O=C(CSc1ncccn1)Nc2cccs2)	5	0
1040	(O=C1NC(=CC(=O)N1)Nc2ccccc2)	2	2
1041	(C1Cc2ccncc2CO1)	5	5
1042	(C(Nc1ncn[nH]1)c2occc2)	4	4
1043	(C(Nc1ncn[nH]1)c2ccccc2)	12	2
1044	(O=C(c1occc1)n2cnncn2)	5	4
1045	(O=S(=O)(Nc1cccn1)c2ccccc2)	6	6
1046	(O=C1OC=Nc2ccccc12)	6	5
1047	(O=C(CSc1ncccn1)NCCc2ccccc2)	12	0
1048	(O=C(CSc1nc[nH]n1)Nc2ccccc2)	3	3
1049	(C1Cc2cn[nH]c2C1)	8	8
1050	(O=S(=O)(Nc1cccn1)c2ccccc2)	15	15
1051	(O=C(CCCSc1ncccn1)Nc2ccccc2)	10	1
1052	(O=C1CCSc2nncn12)	10	10
1053	(C1CCc2nccncc2C1)	8	0
1054	(C1C=COc2ccccc12)	8	5
1055	(O=C1OC=Nc2sccc12)	6	5
1056	(O=C(Nc1cccs1)c2ccccc2)	9	3

1057	(C1Cc2ccccc2CO1)	4	4
1058	(O=C1CCSC2=NC=CCN12)	6	4
1059	(O=S(=O)(Nc1oncc1)c2ccccc2)	10	10
1060	(O=C(OCc1ccccc1)C2=CNC(=O)CC2)	8	8
1061	(O=C1NC(=O)C(Nc2ccccc2)S1)	9	9
1062	(C(c1ccccc1)c2ccccc2)	6	3
1063	(S(c1ccccc1)c2cn[nH]c2)	24	3
1064	(O=C1NNC=C1CC2=CNCC2=O)	3	3
1065	(O=C1NC(=O)N2C=CSC2=N1)	5	0
1066	(O=C(CCCC(=O)OCC(=O)c1ccccc1)Nc2ccccc2)	5	3
1067	(O=C(COc1ccccc1)N2CCNCC2)	10	10
1068	(O=C1OC=CC2=C1NCCC2)	4	4
1069	(O=C1CCCc2occc12)	6	5
1070	(O=C(CNC(=O)C1CCCC1)OCC(=O)c2ccccc2)	5	4
1071	(C(c1cn[nH]c1)c2cn[nH]c2)	9	6
1072	(O=C1NCC=C1Nc2ccccc2)	8	8
1073	(C1CCCc2ncccc2CC1)	3	3
1074	(O=C(Nc1ccccc1)C2=CNC(=O)NC2)	7	7
1075	(O=C1NC(=O)C2=C(NC(=O)N2)N1)	1	1
1076	(O=C(NCCCN1CCOCC1)C(=O)Nc2ccccc2)	6	6
1077	(O=C(CCNc1cccnc1)Nc2ccccc2)	2	2
1078	(O=C(Nc1ccccc1)C2CC(=O)N=CS2)	3	3
1079	(N(c1ccccc1)c2cccnn2)	10	0
1080	(N1C=CSc2ccccc12)	7	5
1081	(O=C(CNS(=O)(=O)c1ccccc1)N2CCNCC2)	5	5
1082	(C(COc1ccccc1)CN2CCNCC2)	11	11
1083	(O=C(CNS(=O)(=O)c1ccccc1)N2CCCCC2)	5	5
1084	(O=C(CNS(=O)(=O)c1ccccc1)NCCSCc2ccccc2)	6	4
1085	(O(c1ccccc1)c2ccccc2)	6	3
1086	(O=C(COC(=O)c1ccccc1)c2ccccc2)	2	2
1087	(O=C(CNS(=O)(=O)c1ccccc1)Nc2cccnc2)	5	5
1088	(O=C(CCN1CCCCC1)Nc2ccccc2)	10	10
1089	(O=C(CNCCc1ccccc1)Nc2ccccc2)	5	5
1090	(O=C(COc1ccccc1)NS(=O)(=O)c2ccccc2)	5	1
1091	(O=C1C=CC2=C1CSC=CN2)	1	0
1092	(O=S(=O)(CCCS1c1cccnc1)Cc2ccccc2)	5	0
1093	(O=C(NCc1occc1)C(=O)Nc2ccccc2)	6	6
1094	(O=C1NCNc2secc12)	5	3
1095	(C=C(NC(=O)c1ccccc1)C(=O)NC2CCCCC2)	3	1
1096	(O=C(OC1CCCCC1)c2ccccc2)	6	1
1097	(O=C(Nc1ccccc1)Nc2cccnc2)	6	6
1098	(O=C(Nc1ccccc1)C=C/c2occc2)	1	0
1099	(O=C(CSc1n[nH]1)Nc2ccccc2)	1	1
1100	(O=C(COc1ccccc1)Nc2nc[nH]n2)	1	1
1101	(C1CN=C(O1)c2ccc[nH]2)	2	1
1102	(C1Cc2[nH]ncc2C=N1)	1	0
1103	(c1ccc(cc1)c2cnc[nH]2)	22	16
1104	(O=C(CSc1c[nH]en1)NCC2CCCO2)	11	9
1105	(O=C(CCc1oncn1)Nc2ccccc2)	5	5
1106	(O=C1NC=Cc2cc[nH]c12)	34	24
1107	(O=C1OC=Cc2ncccc12)	6	6
1108	(O=C(CSc1c[nH]en1)Nc2ccccc2)	12	0
1109	(O=C(CN1CCNC1=O)NCCc2ccccc2)	6	6
1110	(O=C1C=Cc2ccccc12)	15	0
1111	(C(Cc1ccccc1)Nc2ocnn2)	10	0
1112	(O=C1CC(CN1)c2ocnn2)	10	10

1113	(O=C(CSc1ocnn1)N2CCCCC2)	12	12
1114	(O=C(OC1CCCCC1)C2=CNC(=O)CC2)	9	9
1115	(O=C1CC(C=CN1)c2ccccc2)	7	7
1116	(O=C(OCC1CCCCC1)C2=CNC(=O)CC2)	6	6
1117	(O=C1NC=C(C(=O)N1)S(=O)(=O)N2CCCCC2)	8	8
1118	(O=C(NC1=CNNC1=O)Nc2ccccc2)	12	12
1119	(O=C(Nc1ccccc1)C2CC2)	9	9
1120	(O=C1NC(=O)C2=NCCN=C2N1)	6	6
1121	(O=C1NC(=O)c2nccnc2N1)	3	3
1122	(C1Cc2cnoc2C=C1)	22	13
1123	(O=C(Nc1ccnc1)c2ccon2)	7	7
1124	(O=C(NC1=CNNC1=O)c2ccon2)	10	10
1125	(O=C(N1CCCCC1)c2ccon2)	8	8
1126	(O=C(NCCc1ccccc1)c2ccon2)	7	7
1127	(O=C(Nc1cc[nH]n1)c2ccon2)	6	1
1128	(O=C(Nc1cn[nH]1)c2ccon2)	5	4
1129	(O=C(Nc1cn[nH]c1)c2ccon2)	19	19
1130	(O=C(CCS(=O)(=O)c1cccs1)Nc2ccccc2)	19	19
1131	(O=C(CCNS(=O)(=O)c1cccs1)Nc2ccccc2)	4	4
1132	(C(c1coen1)n2ccnn2)	8	8
1133	(c1cc2cnccn2n1)	4	0
1134	(c1esc(n1)c2nccs2)	7	0
1135	(O=C1NC=Cn2cnnc12)	5	5
1136	(O=S(=O)(Nc1ccccc1)c2ccn[nH]2)	1	1
1137	(C(c1ccccc1)c2cnnc2)	1	1
1138	(O=C1NC(=O)c2[nH]ncc2N1)	15	15
1139	(O=C(Cn1ccnn1)NC(=O)Nc2ccccc2)	7	7
1140	(O=C(CS(=O)(=O)c1ncc[nH]1)Nc2ccccc2)	10	3
1141	(O=C(CNC(=O)C1CCCCC1)Nc2ccccc2)	12	12
1142	(c1cn2ccnc2cn1)	10	8
1143	(O=C(CSc1ocnc1)Nc2ccccc2)	9	1
1144	(O=C(CSc1ocnc1)NC2CCCCC2)	7	0
1145	(O=C(CSc1ocnc1)c2ccccc2)	7	0
1146	(C(Cc1ccccc1)Nc2ncccn2)	8	0
1147	(C(Nc1ncccn1)c2ccccc2)	11	0
1148	(O=C(Cc1cccs1)NCC(=O)Nc2ccccc2)	4	3
1149	(O=C(CCCS(=O)(=O)c1ncccn1)Nc2ccccc2)	4	2
1150	(O=C(CNCc1ccccc1)NC2CCCCC2)	4	1
1151	(O=C(NC1N=CNC1=O)c2ccccc2)	8	1
All		41,294	32,972

Take-home Messages

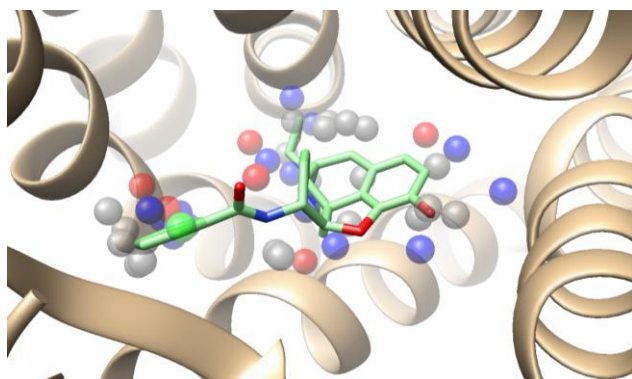
This first chapter provides a comprehensive overview of data collections developed from the experimental PubChem BioAssay database and a thorough discussion on the issues that one should take notice of when using PubChem input for data set construction purposes. We did not only review the history, or point out the challenges, but also proposed possible solutions to address the issues and provided our vision for future directions. This is potentially informative for both the cheminformatics community (including ligand-/structure-based method-developing groups) and the medicinal chemists who are working on rational drug design/drug discovery. At the time when scientists are struggling to find a good standardized data set to test their novel *in silico* screening approaches, we believe that the information provided in this chapter can answer most of the concerns we might have. This review has received rave comments from all three reviewers of the International Journal of Molecular Sciences, and was accepted over two weeks after the first submission, only with several minor modifications.

Chapter 2

All in One: Cavity Detection, Druggability Estimate, Cavity-Based Pharmacophore Perception and Virtual Screening

As explained earlier in the manuscript, well-known issues in predicting the strength of binding interactions between a small ligand and a macromolecule cast doubt on the use of scoring functions employed by docking programs, thus hampering the identification of potential “hits” for a protein target, especially in the case where structural information on neither endogenous nor synthetic ligands is available. A novel computational method tailored to ligand-free protein structures was proposed in 2012, which automatically detects ligand-binding cavities, then predicts their structural “druggability” before creating a structure-based pharmacophore model for the “druggable” binding sites. In this chapter, the design of a new accompanying tool namely Shaper2 is described, aligning small ligands to the aforementioned cavity-derived pharmacophoric features with the use of a smooth Gaussian function. The selection and validation process of scoring parameters to screen the previously aligned ligands is next reported, with the aim of selecting as many active molecules as possible among the top-ranked compounds. The work portrayed in this chapter has been published as an original research paper in the *Journal of Chemical Information and Modeling*, and was presented at various conferences, both as a poster presentation and as an oral presentation.

Tran-Nguyen, V. 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. doi: [10.1021/acs.jcim.8b00684](https://doi.org/10.1021/acs.jcim.8b00684).



1. Introduction

Computer-aided drug design¹ has become a standard tool to assist medicinal chemists in identifying and/or optimizing hits for targets of pharmaceutical interest. Corresponding computational methods are classically divided into ligand-based² or structure-based approaches³ as to whether preexisting knowledge of ligands or target structures is taken into account. Among the ligand-centric methods, pharmacophore searches⁴ are extremely popular for many reasons: (i) the concept of pharmacophore is very intuitive and easily understood for both computational and medicinal chemists, (ii) it does not require the *a priori* knowledge of the target's three-dimensional (3D) structure, (iii) it does not suffer from the main drawbacks⁵ of structure-based approaches (e.g. inaccurate binding free energy estimates) since topological scoring functions⁶ are used to rank ligand adequacy (fitness) to a pharmacophore query, and (iv) aligning a ligand onto a pharmacophore model intuitively guides its further optimization in order to gain or lose additional features.

Typical ligand-based pharmacophore searches first require that the template ligands share the same functional effect, then extract common features from these aligned ligands to derive a pharmacophore hypothesis, and search for potential hits that satisfy this hypothesis in a chemolibrary. If the X-ray structures of protein-ligand complexes are available, protein-ligand-based pharmacophores⁷⁻¹⁰ may be derived as well by mapping features onto protein-interacting ligand atoms, and therefore, complement purely ligand-based pharmacophore models. However, there are still many protein structures and/or novel cavities for which not a single ligand has ever been identified. In order to avoid problems associated with structure-based approaches (e.g. target flexibility, absolute or relative ranking of compounds of interest) for such orphan targets, several methods have been proposed over the last decade to fill the gap between structure-based methods and pharmacophore searches.

Structure-based pharmacophore perception methods classically use a set of molecular probes (atoms, fragments) to locate energetically preferred probe locations. Grid-based methods (e.g. GRID,¹¹ SuperStar,¹² FTMap,¹³ VolSite,¹⁴ T2F,¹⁵ GRAIL¹⁶) locate these preferred positions on a three-dimensional lattice encompassing either the full protein or at least a user-defined binding cavity. Energy minima on the contour maps¹⁷⁻¹⁹ are then saved for every probe and used as guides to define structure-based pharmacophoric features. Fragment-based methods rely on the

prediction of hotspots from molecular dynamic simulations of the target (e.g. MCSS,²⁰ SILCS,²¹ HSRP²²) with multiple copies of fragments bearing well-defined pharmacophoric properties. Again, the most energetically favorable positions of every fragment are later converted into pharmacophores. The positions of these features can be topologically predicted by scanning the cavity-lining and accessible amino acids, in order to generate topologically ideal interaction vectors pointing at 3D space (spheres, cones) where potential ligand atoms should be located to optimally interact with the protein surface. The pioneering method LUDI²³ has inspired many structure-based pharmacophore perception methods (e.g. Virtual ligand,²⁴ SBP,²⁵ HS-Pharm,²⁶ Snooker,²⁷ Exemplar²⁸) to position ideal pharmacophoric moieties from the 3D structure of a binding cavity.

Whatever the method, the number of generated features (a few hundreds) exceeds by far the upper complexity tolerated by pharmacophore searching algorithms. The number of features must therefore be considerably lowered to an acceptable value, usually below 10. A pre-selection phase aimed at pruning pharmacophoric features can be carried out based on energetic criteria,^{15,16,20-22} buriedness criteria,^{15,19} hydration sites overlaps,²² or locations with respect to knowledge-based predicted anchoring hotspots.²⁶ Most methods finish the filtering step by hierarchical clustering based on feature properties and inter-feature distances.

Receptor-based pharmacophore searches have proven to perform at least as effectively as molecular docking, with respect to enrichment in true actives in retrospective virtual screening experiments.^{21,22,26,28} However, they suffer from, with a few exceptions,^{21,28} a lack of automation since many of the above-cited post-processing steps are tedious, thus leaving the user with subjective decisions to make as regards, for example, the nature of probes to use, the acceptable energy minima, or the number of clusters. Moreover, the true value of receptor-based pharmacophore searches in posing a ligand has rarely been examined²⁹ and compared to that of molecular docking.

To address the above limitations, we herewith modified a previously-described cavity detection method (VolSite¹⁴) in order to automatize many steps between cavity detection and workable pharmacophore query definition. VolSite has notably been embedded in the IChem³⁰ toolkit to perform the following operations: (i) on-the-fly detection of all cavities at the surface of a target of interest, (ii) prediction of their structural druggability, and (iii) perception of potential

pharmacophores from the 3D structures of cavities predicted as “druggable”. We next modified the previously reported Shaper method¹⁴ to align ligand atoms onto cavity features by shape-matching and tested several topological as well as energy-based scoring functions in posing and virtual screening challenges.

2. Computational Methods

2.1. Data Sets

Sc-PDB Diverse Set: 213 diverse protein-ligand complexes (**Table S1**) were retrieved from the sc-PDB database³¹ according to the diversity of their protein-ligand interaction patterns, measured by a previously-reported graph-matching procedure (GRIM).³² Starting from a full GRIM similarity matrix calculated on 9283 entries of the sc-PDB archive, clusters were defined using simple agglomerative clustering, a minimal pairwise similarity (GRIM score) of 0.70 between its representatives, a minimal size of 6 entries, and a single linkage criterion. For every cluster, representative X-ray structures of the bound ligand and its cognate target (cluster center) were downloaded from the sc-PDB website.³³

Astex Diverse Set: 85 entries of the Astex Diverse Set³⁴ (**Table S2**) were downloaded from the CCDC website³⁵ and processed as follows. For each entry, the protein-ligand complex was reconstructed in Sybyl-X.2.1.1³⁶ by merging the ligand (mol2 file format) into the protein (mol2 file format). Bound water molecules were imported from the corresponding RCSB Protein Data Bank (PDB)³⁷ file, all hydrogen atoms were deleted, and the fully hydrated complex (heavy atoms only) was protonated using Protoss.³⁸ Ions and cofactors having no heavy atoms located in a 4.5-Å-radius sphere centered on the ligand’s center of mass were deleted. Water molecules were kept if two conditions were satisfied: (i) the oxygen atom was located in the above-described sphere; (ii) the bound water engaged in at least two hydrogen bonds with the protein (donor-acceptor distance not exceeding 3.5 Å, donor-hydrogen-acceptor angle not narrower than 120 deg.). The ligand, as defined in the original Astex data, and the hydrated protein (including the ions and cofactors that remained) were separately saved in mol2 file format.

DUD-E subset: 10 entries (**Table S3**), selected from a previous benchmarking study³² and representing 5 important target families (G protein-coupled receptors, nuclear receptors, protein

kinases, proteases, other enzymes) were retrieved from the DUD-E database³⁹ and further processed similarly to the Astex Diverse Set.

ROCK2 screening set: 59,805 compounds tested for Rho kinase 2 (ROCK2) inhibitory activities were downloaded from the PubChem BioAssay repository in 2D sd file format. Primary screening data (% of inhibition at a single concentration of 6 μM , AID 604)⁴⁰ for all compounds and confirmatory potency values for primary hits (IC_{50} s from the dose-response assay ID 644)⁴¹ were collected directly from PubChem. Compounds with IC_{50} values equal to or lower than 10 μM ($n = 67$) were considered active, all other compounds were considered inactive. The X-ray structure of human ROCK2 kinase in complex with an inhibitor (1426382-07-1) was retrieved from the PDB (PDB ID 4WOT) and further processed similarly to the Astex Diverse Set. The starting 3D coordinates of PubChem ligands (mol2 file format) were generated with Corina v.3.4⁴² and all compounds were ionized at physiological pH with Filter v.2.5.1.4.⁴³ The fully processed data set comprises 59,781 compounds (67 actives and 59,714 inactives).

ESR1 screening set: 10,486 compounds tested for estrogen receptor α (ESR1) antagonism were downloaded from PubChem BioAssay in 2D sd file format. Dose-response inhibitory concentrations for the confirmed hits (IC_{50} values, AID 743080)⁴⁴ were also collected from PubChem. Compounds with IC_{50} values equal to or lower than 25 μM , exhibiting full inhibition curves and devoid of Sn and P atoms ($n = 59$) were kept as actives. To avoid bias in the inactive set, inactive compounds were selected among the molecules free of Sn and P atoms, with molecular weights falling in the same range (310-750 Da) as that observed for true actives. 1530 inactive compounds were finally selected. The X-ray structure of human estrogen receptor α in complex with the selective antagonist 4-hydroxytamoxifen was retrieved from the PDB (PDB ID 3ERT) and further processed similarly to the Astex Diverse Set. The starting 3D coordinates of PubChem ligands (mol2 file format) were generated with Corina v.3.4⁴² and all compounds were ionized at physiological pH with Filter v.2.5.1.4.⁴³ The fully processed data set comprises 1589 compounds (59 actives and 1530 inactives).

OPRK1 screening set: 284,220 compounds tested for kappa opioid receptor (OPRK1) agonism were downloaded from PubChem BioAssay in 2D sd file format. Dose-response activity data (EC_{50} values, AID 1777)⁴⁵ were also collected from PubChem. Compounds with EC_{50} values equal to or lower than 20 μM ($n = 35$) were considered active. All other compounds were

considered inactive, from which a randomly selected set of 34,048 compounds was retrieved. The X-ray structure of the active state-stabilized human kappa opioid receptor in complex with the full agonist MP1104 was downloaded from the PDB (PDB ID 6B73) and further processed similarly to the Astex Diverse Set. The starting 3D coordinates of PubChem ligands (mol2 file format) were generated with Corina v.3.4⁴² and all compounds were ionized at physiological pH with Filter v.2.5.1.4.⁴³ The fully processed data set comprises 34,083 compounds (35 actives and 34,048 inactives).

2.2. Cavity-Based Pharmacophore Perception (IChem)

The previously described VolSite algorithm¹⁴ was embedded in the IChem toolkit v.5.2.9³² with small modifications compared to the original description. First, hydrogen atoms were added to the input target PDB structure using Protoss,³⁸ therefore optimizing the intra- and inter-molecular hydrogen bond network for all molecules in the input PDB file. The pharmacophoric properties of protein atoms (hydrophobic features, aromatic features, hydrogen-bond donors, hydrogen-bond acceptors, positively ionizable features, negatively ionizable features, metals) were detected on the fly from their atom types (mol2 input), thereby enabling us to consider additional molecules (ions, cofactors, water molecules, prosthetic groups, nucleic acids) as parts of the protein. Second, hydrophobic protein atoms were redefined using tighter rules in comparison to those indicated in our seminal report.¹⁴ Hydrophobic atoms were restricted to carbon or sulfur atoms not bonded to heteroatoms or halogen atoms. Cavity-based pharmacophores were defined using a four-step protocol as described in **Figure 1**.

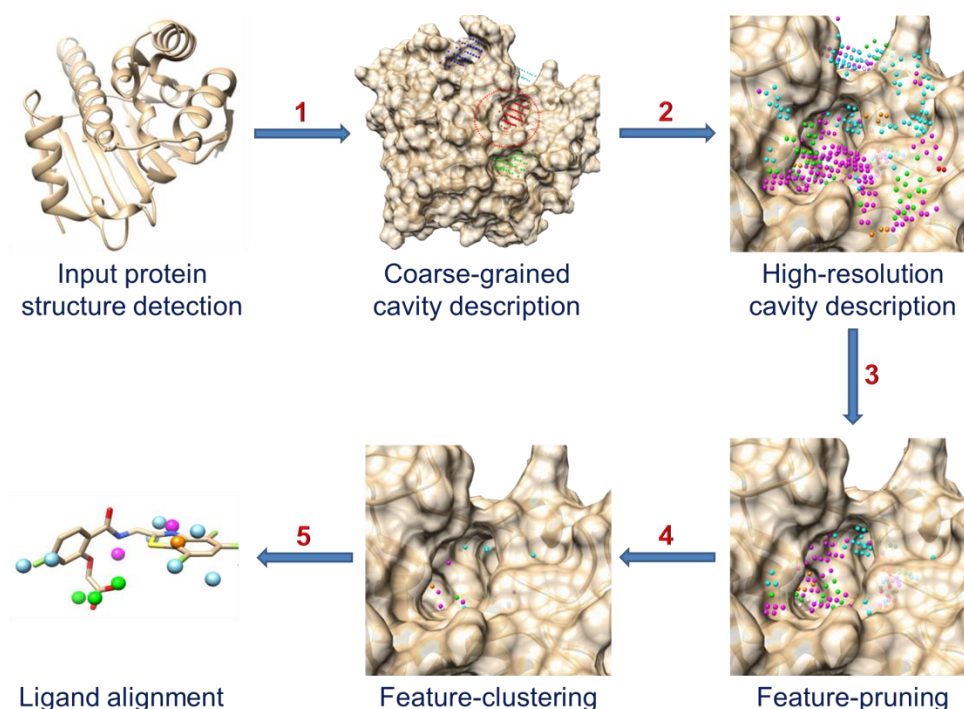


Figure 1. Overall flowchart of the method. **(1)** Starting from a hydrogens-containing protein input structure, cavities were automatically detected using standard VolSite parameters and described as a collection of pharmacophoric features (blue, cyan, red and green dots). **(2)** The cavities predicted as “druggable” (enclosed by a red circle) were submitted to a second structure-based pharmacophoric description step using a tighter grid resolution (1.0 Å). Pharmacophoric features (hydrophobic features: cyan; aromatic features: orange; hydrogen bond acceptors and negatively ionizable features: green; hydrogen bond donors and positively ionizable features: magenta) were assigned according to the pharmacophoric properties of the nearest acceptable protein atom (see “Computational methods”). **(3)** Pharmacophoric features were pruned according to knowledge-based rules (buriedness, distance to cavity center, PLP interaction energy). **(4)** Hierarchical clustering of pharmacophoric features was carried out. **(5)** Shape-based alignment of ligand atoms onto the cavity-based features (same color coding as in step 2) was done by optimizing the overlap of the corresponding molecular shapes.

Step 1 – Coarse-grained cavity detection: the general procedure for detecting cavities has already been described in a previous report¹⁴ and will just be briefly summarized here. Starting from atomic coordinates of the target protein, a three-dimensional (3D) cube was centered on the center of mass of the target and filled with a 1.5-Å-resolution grid defining voxels with a volume of 3.375 Å³ each. To every voxel was associated a site point along with a property at its center. If the corresponding voxel encompassed a protein atom or if its center was less than 2.0 Å away from any protein heavy atom, the site point would be considered inaccessible (“IN” property).

Any other point was then checked for buriedness by generating, from its coordinates, a set of 120 regularly spaced 8-Å-long rays. If the number of rays intersecting an “IN” cell (N_{ri}) was smaller than 55, the corresponding point would be deemed outside the enclosing cavity and was assigned the “OUT” property. The remaining points were claimed to encompass the cavity and checked for direct neighborhood with other cavity points. If isolated (fewer than 3 neighbors in adjacent voxels), the points were deleted. Site points closer than 4.0 Å to a protein atom were assigned one of the eight possible pharmacophoric properties (hydrophobic feature, aromatic feature, H-bond acceptor, H-bond donor, H-bond acceptor and donor, negatively ionizable feature, positively ionizable feature, metal-binding feature) complementary to that of the closest protein atom using the previously-reported interaction rules.³² Points with no neighboring protein atoms within a 4-Å distance were assigned the null property (“dummy”). For each detected cavity, a set of site points (mol2 file format) and a “druggability” score (derived from a previously-described support vector machine model)¹⁴ were given. Only cavities with positive druggability scores were further considered for the generation of cavity-based pharmacophores.

Step 2 – High-resolution cavity description: for each cavity, the previously-reported procedure (step 1) was repeated with two modifications: (i) the center of the 3D lattice was defined as the center of mass of the corresponding coarse-grained cavity, and (ii) the grid resolution was then set to 1.0 Å for a better description of cavity points. Each cavity point was assigned a pharmacophoric feature as previously reported.

Step 3 – Pruning pharmacophoric features: to describe the properties of true pharmacophoric features, “ideal pharmacophores” were deduced from 213 protein-ligand complexes of the sc-PDB Diverse Set. In an ideal pharmacophore model, a feature is assigned to any ligand atom in interaction with the target protein with a property equal to that of the corresponding interaction, but using exactly the same IChem rules (atom types, distances, angles, planes) as those used to define pharmacophoric properties of cavity points. An analysis of these ideal pharmacophoric features enables us to set threshold values for simple descriptors (buriedness, distance to the cavity center, interaction energy) in order to reduce the number of features without losing crucial information. Three pruning rules were applied in the following order: (i) buriedness N_{ri} lower than 80, (ii) distance between the feature and the cavity center shorter than 8 Å, (iii) piecewise linear potential (PLP)⁴⁶ interaction energy lower than the corresponding feature-dependent

threshold (for hydrophobic features, H-bond donors/acceptors, positively ionizable and negatively ionizable features: 0 kcal/mol; for aromatic features: -2.4 kcal/mol; for metal-binding features: -3.5 kcal/mol).

Step 4 – Refining and clustering pharmacophoric features: the remaining H-bond acceptors, aromatic features and hydrophobic features were next subjected to a refining step. As hydrogen atoms were explicitly described in the target protein, a cavity point would still be a hydrogen-bond acceptor feature only on the condition that the nearest protein atom was a hydrogen-bond donor (previous definition in steps 1 and 2) and that the donor-hydrogen-feature angle was between 120 and 180 degrees. Previously-defined acceptor features not fulfilling the new angular threshold were therefore re-assigned a novel property according to the second nearest protein atom and so on until a new property could be unambiguously assigned. If it was not possible (no clear assignment possible from any of the protein atoms closer than 4 Å from the feature), the feature was simply eliminated. The remaining aromatic features were next reconsidered from their spatial location with respect to the aromatic plane to which the closest aromatic protein atom belonged. Apart from the previously applied distance criterion (distance between the feature and the protein atom shorter than 4 Å), we herein applied a second distance threshold of 1.5 Å, corresponding to the largest possible distance between the aromatic feature and two virtual points situated 4 Å away from the closest protein aromatic ring, along a normal to the aromatic plane in both directions. Again, aromatic features not satisfying this additional filter were either reassigned a new property (starting from the second closest protein atom) or eliminated if no assignment was possible. Last, the remaining hydrophobic features were also reconsidered and kept as hydrophobic only if: (i) more than 50% of the protein atoms located within 4.5 Å from the feature were hydrophobic, and (ii) at least 50% of the neighboring protein residues (less than 4.5 Å away) were considered hydrophobic (alanine, valine, leucine, isoleucine, proline, methionine, phenylalanine, tyrosine, and tryptophan). It is note-worthy that these refinements were applied at the step 4 and not to the full set of pharmacophoric features (step 2) to speed up the overall protocol.

The remaining features were then clustered using a simple hierarchical clustering method by pharmacophoric property and inter-feature distance (< 3.1 Å). The final pharmacophoric features were saved in three possible file formats (TRIPOS mol2 format, CATALYST chm file format,⁴⁷

and LigandScout pml format⁸). The pharmacophore models describe for each feature the following items:

- Property: hydrophobic feature, aromatic feature, H-bond acceptor, H-bond donor, negatively ionizable feature, positively ionizable feature, metal-binding feature;
- Atomic coordinates of the feature (head);
- A 3-Å-long projection vector to a tail (H-bond acceptors, H-bond donors, aromatic features) directed to the complementary protein atom;
- Special attributes for aromatic features (centroid, normal, vector, plane);
- Location spheres for directional features (H-bond acceptors, H-bond donors, aromatic features) of 1.6 and 2.2 Å radius for head and tail spheres, respectively;
- Exclusion volumes placed, for each cavity-lining residue (one exclusion volume per residue), on the geometric center of the residue's heavy atoms located at a distance range of 4.1-5.0 Å from any pharmacophoric feature. The radii of exclusion spheres are dependent on the number of close heavy atoms of the protein (1 close atom: 1.15 Å; 2 atoms: 1.25 Å; 3 atoms: 1.35 Å; 4 atoms: 1.45 Å; 5 atoms: 1.55 Å; 6 atoms: 1.60; 7 atoms: 1.65; 8 atoms and above: 1.70 Å).

It is worth noting that features having the double property H-bond donor and H-bond acceptor were described by two separate properties (donor, acceptor) matched on the same point.

2.3. Ligand Alignment to IChem Pharmacophoric Features (Shaper2)

The previously-described Shaper algorithm,¹⁴ designed to align cavities, was slightly modified to align ligand atoms (mol2 file format) onto the aforementioned set of cavity points. Shaper2 relies on OpenEye python libraries⁴³ to describe molecular shapes by a smooth Gaussian function and to align two molecular objects (ligand features, cavity features) by optimizing the intersection of their corresponding volumes.⁴⁸ During the alignment, cavity features are kept rigid while a maximum of 200 pre-defined conformers of the ligand to fit (fit object, constructed in Omega2 v.2.5.1.4)^{43,49} undergo rigid body rotations and translations. Contrary to the original Shaper method, the updated version allows users to choose among different overlap methods (by default: Exact), different overlap minimization techniques (by default: Subrocs) and diverse similarity metrics (by default: TanimotoCombo). A detailed description of all options is available online.⁵⁰

A specific force field (**Table S4**) has been set up to align ligand atoms to cavity features. It consists of SMARTS (simplified molecular-input line-entry system arbitrary target specification) patterns for nine pharmacophoric feature properties (hydrophobic features, rings, H-bond donors, H-bond acceptors, H-bond donors and acceptors, cations, anions, Ca_Mg, Zn) and 56 pattern-matching rules to score the shape-based alignment by pharmacophoric similarity (**Table S4**). All aligned poses were then subjected to a two-step structure optimization process using the MMFF94 force field⁵¹ implemented in SZYBKI v.1.8.0.1.⁴³ First, each pose was minimized with the steepest descent algorithm with respect to the MMFF94 potential in full Cartesian coordinates using default settings. Then, a single point calculation was done with the Poisson-Boltzmann (PB) protein-ligand electrostatics,⁵² calculating protein-ligand interaction energy including solvent effects. All possible ligand-cavity matches were scored according to the four following metrics:

- The TanimotoCombo similarity score:

$$\text{TanimotoCombo} = \text{ShapeTanimoto} + \text{ColorTanimoto} = \frac{\text{OS}_{C,L}}{\text{IS}_C + \text{IS}_L + \text{OS}_{C,L}} + \frac{\text{OC}_{C,L}}{\text{IC}_C + \text{IC}_L + \text{OC}_{C,L}}$$

- $\text{OS}_{C,L}$ is the overlap between the shapes of cavity and ligand features
 - IS_C and IS_L are the non-overlapping shapes of each entity
 - $\text{OC}_{C,L}$ is the overlap between the colors of cavity and ligand features
 - IC_C and IC_L are the non-overlapping colors of each entity
 - The score is asymmetric and varies between 0 and 2.
- The PLP interaction of each feature with the protein, as implemented in the original publication.⁴⁶
 - The MMFF94 protein-ligand interaction energy IntE:

$$\text{IntE} = E_{\text{vdW-PL}} + E_{\text{Coulomb-PL}} + E_{\text{Protein_desolv_PB-PL}} + E_{\text{Ligand_desolv_PB-PL}} + E_{\text{Solvent_screening_PB-PL}}$$

- The MMFF94 total energy TotE = TotIE + IntE:

$$\text{TotIE (ligand MMFF94 intramolecular energy)} = E_{\text{vdW}} + E_{\text{Coulomb}} + E_{\text{Bond}} + E_{\text{Bend}} + E_{\text{StretchBend}} + E_{\text{Torsion}} + E_{\text{Improper_Torsion}}$$

$$\text{IntE} = E_{\text{vdW-PL}} + E_{\text{Coulomb-PL}} + E_{\text{Protein_desolv_PB-PL}} + E_{\text{Ligand_desolv_PB-PL}} + E_{\text{Solvent_screening_PB-PL}}$$

For more details, the reader is directed to the SZYBKI document on the OpenEye website, describing the MMFF94 force field implementation.⁵³

2.4. Ligand Alignment to IChem Pharmacophores (Discovery Studio)

The input ligand 3D structure was converted from mol2 to sd file format using Corina v.3.4⁴² and employed as input to generate 3D conformers using the “Generate Conformations” protocol of Discovery Studio v.2017.⁵⁴ The conformer generation method was set as “FAST”, a maximum of 200 conformers were generated within an energy threshold of 20 kcal/mol (as regards the global minimum). Ligand conformers were next aligned to IChem pharmacophoric features (chm format) using the “citest” command of Discovery Studio. A maximum of 2000 pharmacophore models including from 2 to 6 features were generated to map ligand conformers in the rigid mode. The best mapping conformer (highest fit value) was finally saved in sd file format.

2.5. Ligand Alignment to IChem Pharmacophoric Features (LigandScout)

Ligands (sd file format) were converted to the LigandScout⁵⁵ v.4.1.10 ldb database format with the “idbgen” script that saved up to 200 conformations for each ligand using high-quality settings of the “iCon” conformer generator (“icon-best” option).⁵⁶ The conformations were next aligned, with standard settings of the “iscreen” routine, to the IChem-generated pharmacophores (pml format). The best mapping conformer (highest fit value) was saved in sd file format.

2.6. Docking (Surflex-Dock)

Surflex-Dock v.4.227 was used as prototypical docking engine.⁵⁷ A protomol was first generated from the list of residues, ions, cofactors and water molecules lining the ligand-binding site (any molecule with a heavy atom in a 4.5-Å-radius sphere centered on the ligand’s center of mass) using default settings.⁵⁷ The protomol was further used to dock a randomly generated conformation of the ligand using the “-pgeom” option. Only the best-ranked pose (scored by pK_d values) was saved.

2.7. ROCS Shape Overlap

A maximal number of 200 conformers (sd file format) were generated for every PubChem ligand using standard settings of Omega2 v.2.5.1.4.^{43,49} All conformers were then compared to the query (protein-bound ligand X-ray pose, mol2 file format) with ROCS v.3.2.0.4^{43,58} and scored by TanimotoCombo values, after which the best matching one (highest Tc) was determined.

3. Results and Discussion

The pharmacophore concept is more than one century old⁵⁹ and has been widely used in ligand-based⁴ and, more recently, protein-ligand-based^{7,8} virtual screening. When only structures of ligand-free proteins are available, defining simple and workable pharmacophore queries is more difficult for the simple reason that cavity structure-based pharmacophore perception is a complex and multi-step procedure. Cavities first need to be detected at the protein surface, and then evaluated for their potential “druggability”. The positions of pharmacophoric features mimicking a perfect ligand must then be inferred from the coordinates of cavity-lining protein residues. Very often, the number of ideal features exceeds by far the upper complexity tolerated by standard 3D pharmacophore searches. Therefore, they need to be rationally pruned, usually from interaction energy maps, to downsize the population and to enable the definition of a workable pharmacophore model (usually comprising fewer than 10 features). Moreover, there exist many methods²⁰⁻²² that rely on lengthy molecular dynamic simulations to locate the energetically preferred positions of probes, which prohibits their usage even at a low throughput. Although recent efforts have been reported to simplify the steps described above,^{21,28} it is still necessary to design a tool that is able to quickly and reliably automatize the entire process from early cavity detection to late final pharmacophore definition.

3.1. Cavity-Based Pharmacophore Perception

The herein proposed cavity-based pharmacophore perception workflow is made of four consecutive steps (**Figure 1**). First, potentially druggable cavities were detected on the fly from the input protein structure using standard parameters of our in-house developed VolSite algorithm.¹⁴ The method centers the protein in a 1.5-Å-resolution lattice and assigns a pharmacophoric feature (hydrophobic/aromatic/positively-charged/negatively-charged/metal-binding feature, H-bond donor and/or acceptor) to every accessible voxel, depending on the pharmacophoric property of the nearest accessible protein atom. The structural “druggability” of every detected cavity was predicted with the use of a support vector machine model¹⁴ that showed a very good accuracy level in comparison to state-of-the-art methods. For each cavity, the detection procedure was repeated using a higher-resolution grid (1.0 Å) that was centered on the cavity’s center of mass, after which the obtained features were pruned in order to decrease their population.

The previously published VolSite algorithm¹⁴ was modified to take into account the positions of explicit hydrogen atoms, added by the Protoss knowledge-based method.³⁸ The main advantage of using hydrogen coordinates of the target protein is that hydrogen acceptor features can be better assigned from the corresponding vectors (donor-hydrogen-voxel center) than using the previous protocol that just relied on distances. Along the same spirit, we have also refined the definition of cavity aromatic features by taking into account additional topological measurements for detecting face-to-face aromatic interactions (see “Computational methods”). Last, the assignment of hydrophobic features is stricter and now requires that the closest protein atom be also annotated as hydrophobic and located in a global hydrophobic environment. The consequence of these changes is that the pharmacophoric assignment of cavity features may require several steps. For example, a hydrophobic protein atom (e.g. CB atom of an alanine) cannot be used to assign a hydrophobic property to a cavity voxel if the latter does not satisfy the above-described proximity conditions, even if it is the closest protein atom of that particular voxel. In that case, a second assignment step is done by considering the second closest protein atom to the voxel, and so on until one protein atom perfectly suits all the required conditions. Therefore, contrary to the original VolSite implementation,¹⁴ in this updated version, some cavity voxels may not be assigned a pharmacophoric property.

A key issue in the current work is the implementation of knowledge-based rules to limit the number of pharmacophoric features to the lowest possible number. To reach this objective, we carefully analyzed the position of “ideal” pharmacophoric features derived from a training set of 213 diverse protein-ligand structures. By “ideal”, we mean that pharmacophoric features are directly mapped onto protein-bound ligand atoms if the corresponding atom is in direct interaction, according to IChem rules, with the protein. To define a set of ideal features, 213 high-resolution protein-ligand X-ray structures were extracted from the sc-PDB archive of druggable protein-ligand complexes.³¹ These structures present a maximal diversity of protein-ligand interaction patterns, as assessed by our previously described GRIM methodology³² that directly computes the pairwise similarity of protein-ligand interaction patterns. Out of the 213 most diverse complexes, we could identify 4871 ideal features for which three properties were inspected: buriedness, distance to the cavity center, and PLP interaction energy (**Figure 2**).

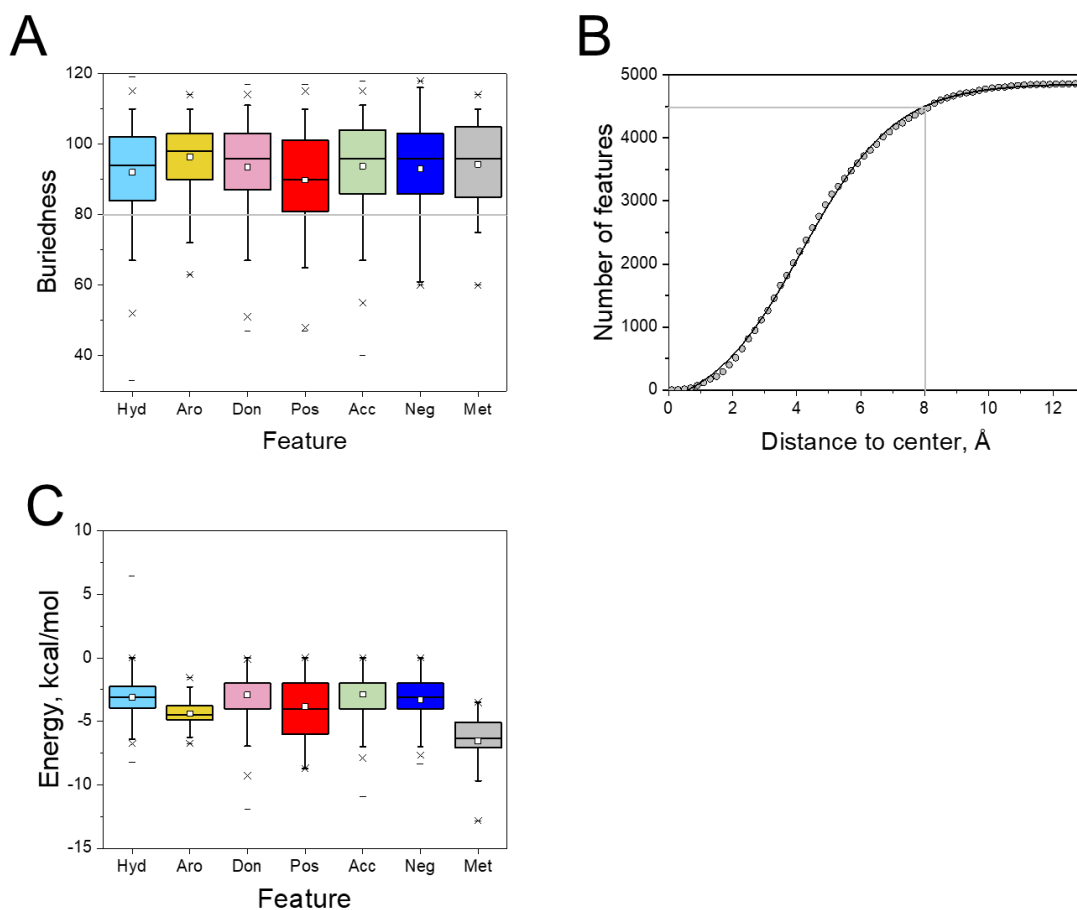


Figure 2. Properties of 4871 ideal pharmacophoric features generated from the sc-PDB Diverse Set (213 complexes). **(A)** Box-and-whisker plot of the distribution of pharmacophoric features' buriedness (Hyd: hydrophobic features; Aro: aromatic features; Don: H-bond donors; Pos: positively ionizable features; Acc: H-bond acceptors; Neg: negatively ionizable features; Met: metal-binding features) expressed by the number of 8-Å-long rays (out of 120 in total) originating from the feature center and the intersecting protein atoms. The boxes delimit the 25th and the 75th percentiles, the whiskers delimit the 5th and the 95th percentiles. The median and mean values are indicated by a horizontal line and an empty square in the box, respectively. The crosses delimit the 1st and the 99th percentiles. The minimum and maximum values are indicated by the dashes. **(B)** Distance of the feature (in Å) to the cavity center, expressed by the cumulative number of features. The cumulative distribution follows a Boltzmann sigmoidal function ($R^2 = 0.999$). **(C)** Box-and-whisker plot of the distribution of inter-feature PLP⁴⁶ interaction energy (Hyd: hydrophobic features; Aro: aromatic features; Don: H-bond donors; Pos: positively ionizable features; Acc: H-bond acceptors; Neg: negatively ionizable features; Met: metal-binding features) and their protein environment. The boxes delimit the 25th and the 75th percentiles, the whiskers delimit the 5th and the 95th percentiles. The median and mean values are indicated by a horizontal line and an empty square in the box, respectively. The crosses delimit the 1st and the 99th percentiles. The minimum and maximum values are indicated by the dashes.

Whatever the feature type, more than 75% of the ideal features had buriedness values higher than 80 (**Figure 2A**). Likewise, over 90% of them were closer than 8 Å from the corresponding cavity center (**Figure 2B**). As expected, the recorded PLP interaction energy values of these features with their protein environment clearly show that they are negative and feature type-dependent (**Figure 2C**). Applying feature-dependent cut-off thresholds (for hydrophobic/positively ionizable/negatively ionizable features, H-bond donors and/or acceptors: 0 kcal/mol; for aromatic features: -2.4 kcal/mol; for metal-binding features: -3.5 kcal/mol) ensured that at least 95% of these ideal features would be selected.

The application of the above-described pruning rules all along the flowchart (**Figure 3A**) indeed limited the number of output features from 326 ± 90 at the beginning of the process (fine-grained cavity description) to 259 ± 95 after buriedness evaluations, 253 ± 88 after cavity center-feature distance calculations, 37 ± 7 after clustering, and finally 27 ± 7 after PLP interaction energy calculations (**Figure 3B**). The chronological order in applying these three filters does not affect the obtained results. To avoid repeating the PLP interaction energy evaluation before and after clustering, we decided to place this step at the end of the protocol. Here again, we verified that this choice did not bias the obtained results.

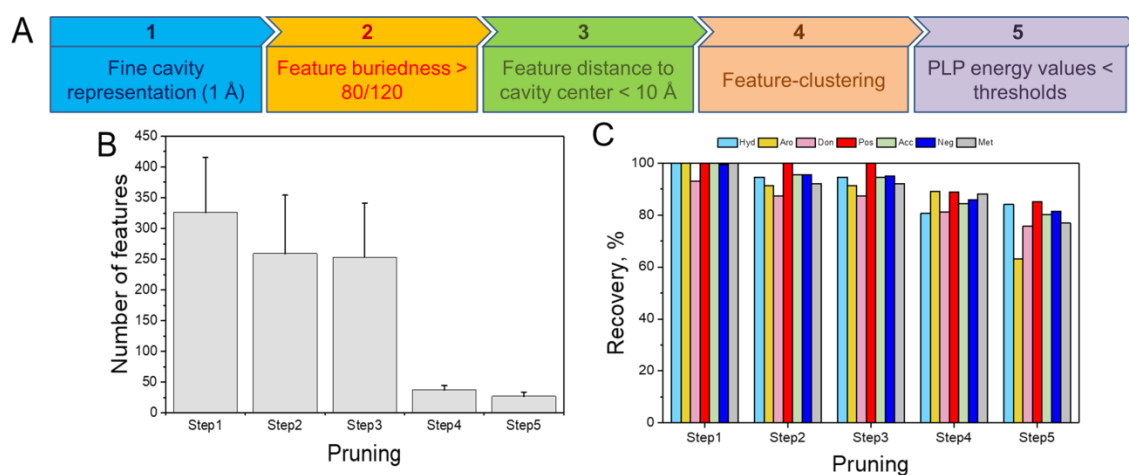


Figure 3. The five-step protocol to prune cavity-based pharmacophoric features in IChem. Features were defined from the IChem-detected ligand-binding sites of 213 entries of the sc-PDB Diverse Set. (A) The flowchart. (B) The decreasing number of pharmacophoric features that remained all along the protocol. (C) The percentage of ideal features recovery all along the protocol. An ideal feature is deemed “recovered” if it is located closer than 2.0 Å from a predicted feature of the same type, generated for the same test set according to identical topological rules by matching pharmacophoric properties to protein-interacting ligand atoms.

We also verified that the observed drastic reduction in the number of features did not lead to a global loss of information. For that purpose, we estimated the percentage of ideal features recovery, by computing the closest distance between every IChem-predicted element and an ideal feature of a compatible pharmacophoric type. If the distance is smaller than 2.0 Å, the predicted feature is deemed close enough to the ideal one and the latter is recovered. Estimating the percentage of ideal features recovery at every step of the pruning stage (**Figure 3C**), we conclude that the filtering process did not discard a significant proportion of key elements. After the last step, about 80% of all features belonging to every feature type (except aromatic ones, for which the recovery rate was about 70%) were within a radius of 2 Å from a predicted element of the same type. We thus assume that our feature selection process is accurate enough to simplify the final cavity-based pharmacophore model without any major loss of information.

3.2. Ligand Posing Accuracy

Ligands were aligned onto the above-described cavity-based pharmacophoric features using a modified version (Shaper2) of our Shaper algorithm,¹⁴ employing a smooth Gaussian function to maximize the shape overlap of ligand atoms and cavity features, and score the alignment by both shape and color (feature type) similarity. In comparison to the previous Shaper version that had been designed for pairwise cavity comparisons, the force field was modified in this updated one (**Table S4**) to enable ligand alignment to cavity features. A test set of 85 high-quality protein-ligand complexes (Astex Diverse Set),³⁴ specifically designed to assess docking performance, was used for that purpose. To estimate the posing quality, we compared the results obtained with Shaper2 alignment on IChem features (this work) to those of a state-of-the-art docking tool (Surflex-Dock).⁵⁷ Moreover, we also compared the alignment accuracy of Shaper2 to that of two standard pharmacophore search methods (Discovery Studio, LigandScout), using the same set of IChem-derived features. Four scoring functions were evaluated to analyze Shaper2 matching poses to IChem pharmacophores. The first one (Tc) just computes the TanimotoCombo similarity (shape + color) between the aligned poses and the protein-bound ligand X-ray coordinates. The second one (PLP) computes the PLP interaction energy of the feature with its protein environment. The third and fourth ones (TotE, IntE) register the MMFF94 total interaction energy and MMFF94 protein-ligand interaction energy using a Poisson-Boltzmann treatment of desolvation effects.

Plotting, for each Astex Diverse Set entry, the root-mean square deviation (RMSD) of the best Surflex-Dock pose (heavy atoms only) to the true X-ray pose, defines the base line for applying a structure-based docking tool to this data set (**Figure 4**).

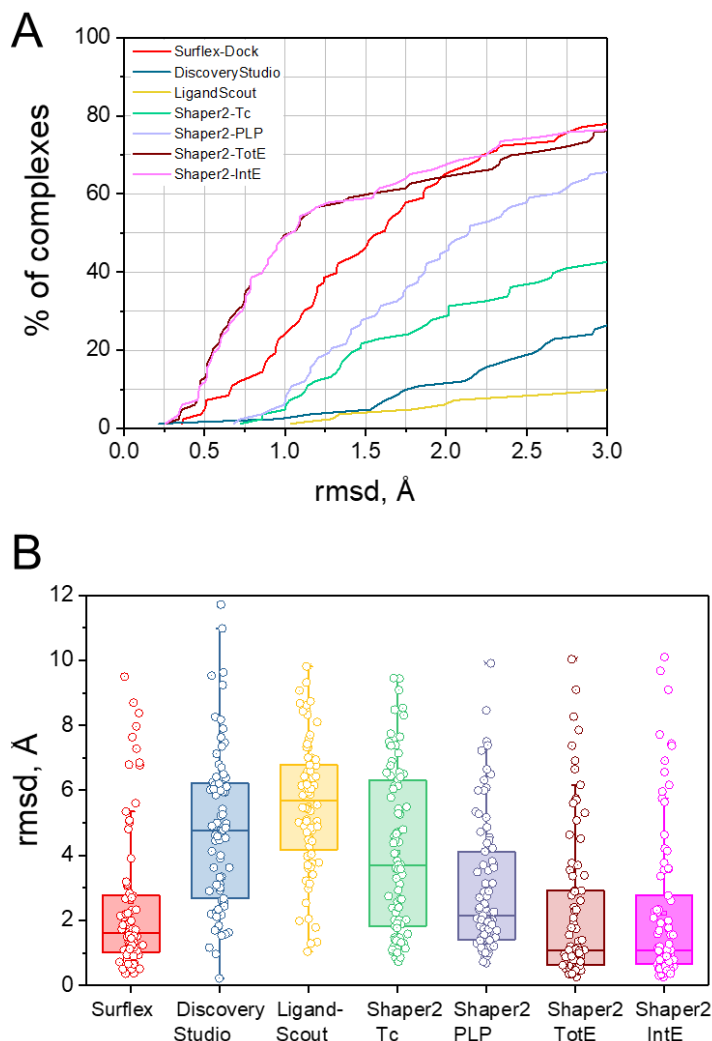


Figure 4. Performance of different methods in predicting the poses of 85 ligands from the Astex Diverse Set. Posing was done using docking (Surflex-Dock), ligand-based pharmacophore searches (Discovery Studio, LigandScout), and cavity-based pharmacophore searches (IChem). IChem alignment was scored by four different functions: TanimotoCombo similarity (Tc), PLP interaction energy (PLP), total MMFF94 energy (TotE), MMFF94 protein-ligand interaction energy (IntE). **(A)** Cumulative percentage of entries from the Astex set for which the top-ranked pose of the cognate ligand is within a certain RMSD to the X-ray pose. **(B)** Distribution of RMSD values to the X-ray pose. The boxes delimit the 25th and the 75th percentiles, the whiskers delimit the 5th and the 95th percentiles. The median and mean values are indicated by a horizontal line and an empty square in the box, respectively. The crosses delimit the 1st and the 99th percentiles. The minimum and maximum values are indicated by the dashes.

Surflex-Dock indeed posed quite accurately the Astex ligands with a median RMSD of 1.62 Å. 65% of all ligands were docked with RMSD values to the X-ray pose below 2 Å (**Table 1**). This docking performance is quite similar to previous results obtained on this peculiar data set⁶⁰ and on other sets by us⁶¹ and other groups.^{5,62} We can therefore assess that no particular bias is present in both the data set and the manner we set the input files. In our hands, the two ligand-based pharmacophore tools (Discovery Studio, LigandScout) failed to predict the correct pose (RMSD < 2.0 Å) in approximately 90% of the cases (**Figure 4, Table 1**). In other words, the complexity of IChem cavity-based features (27 features on average for the Astex Diverse Set) is still too important for hard sphere-based alignment tools. The quality of IChem cavity-based pharmacophores is not responsible for this observation since Shaper2 alignment to the same pharmacophores produced much better results, albeit with significant differences as regards the chosen scoring function (**Figure 4, Table 1**). Just relying on the similarity of shapes and colors (Tc metric) was not sufficient to yield high-quality poses (average RMSD = 4.10 Å) although the obtained results were already better than those received from Discovery Studio and LigandScout. Rescoring Shaper2 poses according to the PLP energy significantly improved the alignment (median RMSD = 2.95 Å, **Table 1**). However, this scoring method remains inferior to Surflex-Dock in producing high-quality poses (**Figure 4**).

We therefore minimized the pose (ligand in its protein environment) with the MMFF94 force field that includes an explicit Poisson-Boltzmann treatment of desolvation effects.⁵² Using either the total MMFF94 energy (TotE: ligand strain energy + protein-ligand interaction energy) or just the protein-ligand interaction energy term (IntE) yielded very accurate poses (identical median RMSD to the X-ray pose of 1.06 Å). Interestingly, although the fraction of high-quality poses (RMSD < 2.0 Å) was almost identical to that obtained with Surflex-Dock (approximately 65%), these two scoring functions were much more effective in producing very high-quality poses (RMSD to the X-ray pose < 1.0 Å; **Table 1**).

Altogether, Shaper2 alignment on IChem cavity-based pharmacophore models is therefore competitive with a standard docking tool as regards posing accuracy. The competitive advantage of a Gaussian function (Shaper2) in comparison to either the Kabsch algorithm⁶³ (Discovery Studio) or the Hungarian matcher⁶⁴ (LigandScout) appears quite significant, when it comes to considering the complexity of pharmacophore queries (27 features on average) produced by our

method. It is also note-worthy that the scoring function employed to rank Shaper2 poses is very important. Energy-based scoring functions are preferred to accelerate shape/color overlap estimations. Moreover, an explicit treatment of desolvation effects yields a very accurate pose ranking, albeit at the cost of an extra computational demand (approximately 5 seconds per pose).

Table 1. Posing accuracy of molecular docking (Surflex-Dock), ligand-based pharmacophore searches (Discovery Studio, LigandScout), and receptor-based pharmacophore searches (ICChem), applied to 85 protein-ligand complexes from the Astex Diverse Set.

Program	Average RMSD, Å ^a	Median RMSD, Å ^b	% of entries with RMSD < 1 Å	% of entries with RMSD < 2 Å
Surflex-Dock ^c	2.57	1.62	24	65
Discovery Studio ^d	4.80	4.77	3	12
LigandScout ^e	5.53	5.70	0	6
Shaper2-Tc ^f	4.10	3.70	4	28
Shaper2-PLP ^g	2.95	2.14	6	45
Shaper2-TotE ^h	2.23	1.06	49	64
Shaper2-IntE ⁱ	2.22	1.06	48	67

^a Average root-mean-square deviation (heavy atoms) to the ligand X-ray pose

^b Median root-mean-square deviation (heavy atoms) to the ligand X-ray pose

^c Surflex-Dock pose with the lowest internal score (pK_d)

^d Discovery Studio pose with the highest fit score

^e LigandScout pose with the highest fit score

^f Shaper2 pose with the highest TanimotoCombo score

^g Shaper2 pose with the lowest PLP interaction energy

^h Shaper2 pose with the lowest MMFF94 total energy

ⁱ Shaper2 pose with the lowest MMFF94 ligand-protein interaction energy

3.3. Virtual Screening Accuracy (DUD-E Set)

In the next challenge, we probed the accuracy of Shaper2 alignment to ICChem cavity-based pharmacophores to discriminate between true actives and chemically similar decoys for a set of ten DUD-E targets (**Table S3**).^{32,39} Although results obtained on such benchmarks are not fully predictive of real-life prospective virtual screening studies,⁶⁵ we still wanted to compare our approach to Surflex-Dock in this exercise. Ten targets were selected to span major target families (G protein-coupled receptors, kinases, nuclear hormone receptors, proteases, other enzymes) and caution was given to discard easy test cases (targets leading to areas under the ROC curves above 0.85) as suggested by the seminal paper.³⁹ The chosen subset is believed to be rather difficult for

docking (DUD-E authors used the Dock3.6 docking program as screening engine) with an average AUC value of 0.66, well below the mean AUC value (0.76) observed for the entire DUD-E database.³⁹ Results obtained with Surflex-Dock generally confirmed the previous report with a mean AUC value of 0.73 (**Table 2**). For two targets (GCR, FGFR1), the observed ROC AUCs were statistically better than random selection but still below 0.70, therefore indicating just a fair performance. Shaper2 alignment to IChem pharmacophores scored by the PLP potential led to a poor performance in this challenge (mean AUC value of 0.57; **Table 2**). Conversely to the above-described challenge, scoring matching poses by either MMFF94 protein-ligand interaction energy or MMFF94 total energy marginally enhanced the virtual screening accuracy of the method (mean AUC values of 0.62 and 0.65, respectively; **Table 2**) despite significant ameliorations ($AUC \geq 0.70$) for five out of the ten targets (ADRB2, GCR, ACE, FGFR1, AKT1), using the MMFF94 total energy as a scoring function. Given that the MMFF94 total energy led to the best performance, we tried to decouple the scoring function used to select the best poses from that utilized to sort compounds. The best combination was obtained by selecting the poses by MMFF94 total energy and sorting the compounds (actives and decoys) by PLP energy (**Table 2**). Using this approach, a mean AUC value of 0.68, comparable to that observed with the docking program Dock3.6, was obtained. The performance was excellent for two targets (ADRB2, RENI: ROC AUC > 0.80), good for two other entries (FGFR1, AKT1: $0.70 < \text{ROC AUC} < 0.80$), fair for four targets (AA2AR, GCR, ADA, ACE: ROC AUC ≥ 0.57) and remained poor but still better than random picking for two entries (ANDR, PGH2). Despite the small sample size, the distribution of ROC values observed from the three Shaper2 protocols with MMFF94 refinement (IntE, TotE, TotE + PLP) is statistically different from that seen when only PLP energy was taken into account in a two-sample t-test assuming either equal or unequal variance at a confidence interval of 95% ($p < 0.05$). The differences observed with respect to each pair of the refinement protocols are however statistically not significant in the same test. Compared to Surflex-Dock, the mixed approach gave a better performance for three targets (ADRB2, ANDR, FGFR1), a rather similar accuracy level for three entries (GCR, RENI, AKT1), but gave a poorer performance for the other four entries (AA2AR, ADA, PGH2, ACE; **Table 2, Figure 5**).

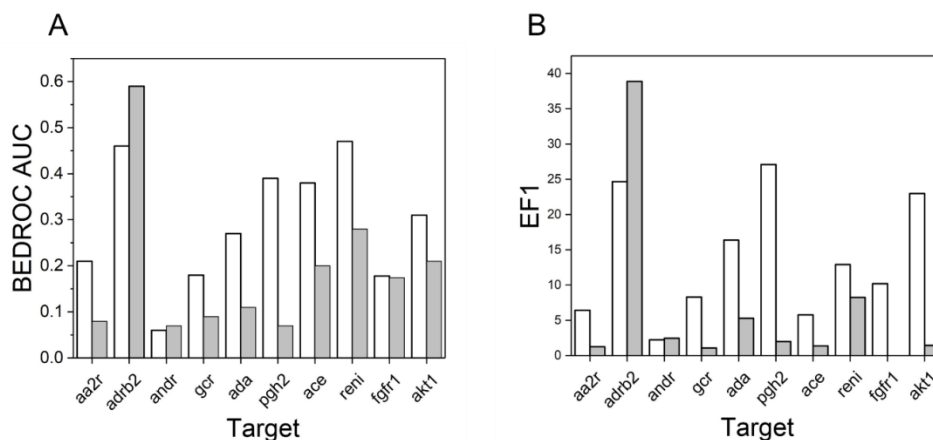


Figure 5. Virtual screening performance of Surflex-Dock (white bars) and Shaper2 (gray bars) on 10 entries of the DUD-E set.³² Shaper2 alignment to IChem cavity-based pharmacophores was scored by MMFF94 total energy, whereas DUD-E compounds were ranked by increasing PLP interaction energy. **(A)** Area under the BEDROC curve ($\alpha = 20$). **(B)** Enrichment in true actives at a constant 1% false positive rate.

We must acknowledge that we have no clear explanation on the positive role of PLP rescoring on the poses selected by MMFF94 total energy. We could not either explain the successes and failures of the approach with respect to target and/or ligand properties. To account for early enrichment in true actives, the areas under the Boltzmann-enhanced discrimination of the ROC (BEDROC) curves, as well as the enrichment in true actives at 1% decoys retrieval, were also computed for each of the entries (**Figure 5**). Disappointingly, BEDROC curves clearly show that our method was inferior to Surflex-Dock in early enrichment in true actives for seven out of the ten cases.

Table 2. Area under the ROC plot of a binary classification (actives, decoys) of DUD-E ligand poses to the X-ray structures of 10 representative targets.³²

Posing	Dock3.6^a	SF-Dock	Shaper2	Shaper2	Shaper2	Shaper2
Conformer selection	Dock3.6	SF-Dock	PLP	IntE	TotE	TotE
Scoring	Dock3.6	SF-Dock	PLP	IntE	TotE	PLP
<i>G protein-coupled receptors</i>						
Adenosine A2A receptor (AA2AR)	0.83	0.74	0.57	0.61	0.56	0.58
Beta2 adrenergic receptor (ADRB2)	0.76	0.85	0.51	0.61	0.71	0.96
<i>Nuclear hormone receptors</i>						
Androgen receptor (ANDR)	0.51	0.47	0.56	0.52	0.59	0.54
Glucocorticoid receptor (GCR)	0.44	0.56	0.56	0.64	0.73	0.57
<i>Other enzymes</i>						
Adenosine deaminase (ADA)	0.76	0.83	0.60	0.56	0.53	0.63
Prostaglandin G/H synthase 2 (PGH2)	0.62	0.76	0.57	0.62	0.54	0.55
<i>Proteases</i>						
Angiotensin-converting enzyme (ACE)	0.72	0.84	0.58	0.60	0.75	0.64
Renin (RENI)	0.66	0.88	0.56	0.68	0.66	0.82
<i>Protein kinases</i>						
Fibroblast growth factor receptor 1 (FGFR1)	0.73	0.67	0.60	0.78	0.70	0.76
RAC-alpha protein kinase (AKT1)	0.72	0.76	0.57	0.61	0.72	0.74
Mean ROC area under the curve	0.67	0.73	0.57	0.62	0.65	0.68

^a Report from the original paper describing the DUD-E database³⁹

3.4. Virtual Screening Accuracy (PubChem BioAssay)

The real value of DUD-E ligands in evaluating virtual screening performance is currently under debate because of severe ligand- and target-based drawbacks in selecting decoys.⁶⁵ The discriminatory power of most docking tools was reported to be overestimated, with the use of this data collection, for the simple reason that DUD-E actives tend to be chemically similar to the co-crystallized ligand in the 3D target structure that is selected for docking.⁶⁵ We therefore challenged our method with true experimental screening data from the PubChem BioAssay repository,⁶⁶ in which both true active and true inactive compounds have been explicitly defined according to *in vitro* assays. Three targets of pharmaceutical importance (one kinase, one nuclear hormone receptor, one G protein-coupled receptor) for which both high-quality screening data (primary assay, confirmatory dose-response assay) and 3D structural information (ligand-bound high-resolution X-ray structure) are available were selected as test cases (**Table 3**).

Virtual screening was carried out using one ligand-based method (3D shape-matching with ROCS),⁵⁸ and two structure-based approaches (molecular docking with Surflex-Dock, pharmacophore-based ligand-aligning with Shaper2). The virtual screening accuracy was simply estimated from the number of true actives ranked among the top 1% and the top 5% scorers. The experimentally determined hit rate is low (approximately 0.1%) for two screens (ROCK2, OPRK1) and much higher (3.71%) for the ESR1 challenge. Activity data range from low nanomolar to two-digit micromolar values. The ESR1 ligand set is the most enriched in molecules of very high potency (**Table 3**), and should, therefore, be easier to predict. This assumption was confirmed by an analysis of screening results given by 3D shape-matching using ROCS, as spectacular enrichment over random picking was observed when the top 1%-ranked ESR1 ligands were considered (**Table 3**). This means that the true actives in this set are similar in both shape and pharmacophoric properties to the reference ligand (4-hydroxytamoxifen) that was co-crystallized in the protein structure used for the structure-based approaches.

Table 3. Virtual screening of PubChem BioAssay data.

Target	Rho kinase 2		Estrogen receptor α		Kappa opioid receptor	
Encoding gene	ROCK2		ESR1		OPRK1	
PubChem BioAssay AID	604, 644		743080		1777	
Number of actives	67		59		35	
Number of inactives	59,714		1530		34,048	
Activity range, μ M	0.03-9.78		0.03-9.69		0.06-18.10	
Hit rate, %	0.11		3.71		0.10	
Virtual screening ^a	<i>Top 1%</i>	<i>Top 5%</i>	<i>Top 1%</i>	<i>Top 5%</i>	<i>Top 1%</i>	<i>Top 5%</i>
• ROCS ^b	2 (3.0)	3 (0.9)	11 (18.5)	11 (3.7)	1 (2.9)	1 (0.6)
• Surflex-Dock ^c	1 (1.5)	2 (0.6)	1 (1.7)	6 (2.0)	3 (8.8)	4 (2.3)
• Shaper2 ^d	1 (1.5)	2 (0.6)	2 (3.4)	18 (6.1)	1 (2.9)	6 (3.5)

^a Number of true actives among the top 1% and the top 5% scoring molecules. Numbers in brackets indicate the observed enrichment over random picking.

^b Ligands ranked by TanimotoCombo similarity scores to the template ROCK2-bound inhibitor (ligand ID 3SG, PDB ID 4WOT), ESR1-bound antagonist (ligand ID OHT, PDB ID 3ERT), and OPRK1-bound agonist (ligand ID CVV, PDB ID 6B73).

^c Ligands ranked by pK_d (Surflex-Dock score).

^d Ligands ranked by PLP energy after MMFF94 energy minimization.

For the two targets ROCK2 and OPRK1, ROCS screening performed three times better than random selection when the top 1% scorers were considered, the enrichment logically decreased when selecting more compounds from the screen with a performance equal or even inferior to random picking when the top 5% scoring compounds were accounted (**Table 3**). In other words, two screening sets (ROCK2, OPRK1) were deemed difficult for structure-based approaches, whereas the third one (ESR1) was much easier.

Surflex-Dock and Shaper2 gave identical results when the top 1% scorers of the ROCK2 screen were considered, although their performances were inferior to that of ROCS (**Table 3**). Accounting a higher percentage of top scoring compounds (5%) allowed us to retrieve one additional active, but at the cost of a lower hit rate. For the easier ESR1 test case, Shaper2 gave much better results than Surflex-Dock, whatever the fraction that was considered to qualify virtual hits. Enrichment factors over random picking of 3.4 and 6.1 were observed for the top 1% and the top 5% scoring molecules, respectively (**Table 3**). It is note-worthy that Shaper2 continued to retrieve novel actives as the number of selected virtual hits was increased, and even outperformed ROCS when the top 5% scoring hits were accounted. For the last data set (OPRK1), both Surflex-Dock and Shaper2 gave statistically good enrichment over random picking (8.8 and 2.9 at the top 1% scorers, 2.3 and 3.5 at the top 5% scorers). Docking performed better than cavity-based pharmacophore searches in the initial enrichment, but Shaper2 retrieved more actives than Surflex-Dock among the top 5% scorers (**Table 3**).

In agreement with many previous studies,⁶⁷⁻⁶⁹ we observed that the three virtual screening methods used in this study tend to retrieve different true actives, and most importantly, different chemotypes (**Figure 6**). In all screens, Shaper2 was able to identify true actives (one ROCK2 inhibitor, seven ESR1 antagonists, four OPRK1 agonists, **Figure 6**) not found by any other method. If one restricts the analysis to the retrieval of unique scaffolds, Shaper2 was the method producing the highest number of uniquely retrieved chemotypes (**Figure 6**), thereby demonstrating its utility and orthogonality to other virtual screening methods.

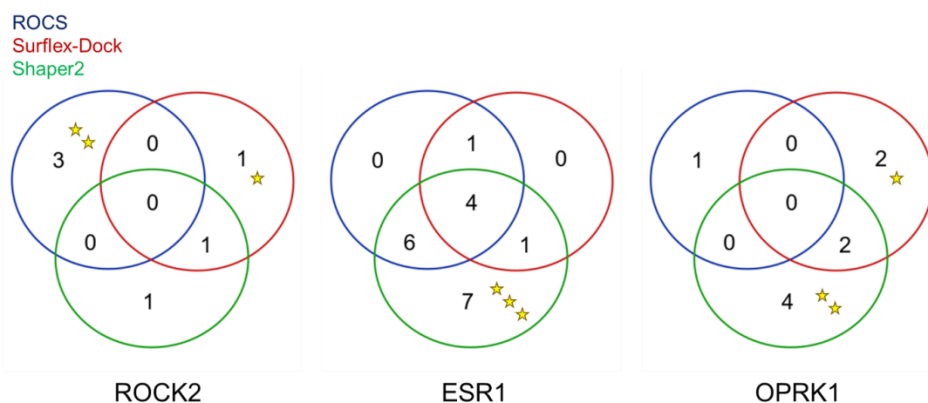


Figure 6. Orthogonality of three virtual screening methods (ROCS, Surflex-Dock, Shaper2) in retrieving true actives among the top 5% ranking hits, from three PubChem BioAssay high-throughput screens (ROCK2 inhibitors, PubChem BioAssay AID 644; ESR1 antagonists, PubChem BioAssay AID 743080; OPRK1 agonists, PubChem BioAssay AID 1777). The numbers of true actives recovered by each method are displayed by Venn diagrams,⁷⁰ highlighting molecules uniquely found by a single method or common to two or three hit lists. Each chemotype retrieved by a single method is highlighted by a star.

The motivations for retrieving the top 5% scorers were two-fold. Firstly, since we were really mining HTS data with very few high affinity ligands, the number of hits retrieved among the top 1% scorers was low (even for the ligand-based ROCS shape-matching method). We therefore increased the threshold to select the top 5% scoring molecules in order to begin to see statistically meaningful differences between the screening methods. Secondly, retrieving a higher proportion of virtual hits enabled us to cluster them by scaffolds (maximum common substructures) and pick a more representative set of hits for experimental validation (in terms of scaffold coverage) than a strategy based on a harder cut-off (say, pick the top 100 scoring compounds). Of course, no definitive conclusion can be drawn from the present benchmarking exercise focusing on three independent HTS data. However, it appears that Shaper2 alignment on IChem cavity-based pharmacophores is at least as effective as other virtual screening methods (shape alignment, docking) when applied to three test cases for which the entire screening results were known. The good performance of Shaper2 in true virtual screening benchmarks is in contradiction to the previously reported poorer performance observed in artificially constructed DUD-E training sets, for which severe target and ligand bias has been noticed.⁶⁵ We therefore recommend benchmarking virtual screening methods with true experimentally determined high-throughput screening data. Fortunately, the PubChem BioAssay repository⁶⁶ proposes an

increasing number of high-quality screening sets with both primary and confirmatory dose-response data to guide computational method development and validation.

3.5. Comparison to Other Cavity-Based Pharmacophore Perception Methods

In comparison to current structure-based pharmacophore perception methods,¹¹⁻²⁹ the herein described approach presents five noticeable assets. First, the pharmacophore perception method is fully automated, does not rely on any third party tool, and is freely available for non-profit research. The last criterion is particularly important to enable fair benchmarking. Second, in contrast to many alternative approaches,^{11,15,16} IChem does not require user intervention in defining grid lattice coordinates. It scans the entire surface and can therefore generate as many pharmacophores as the non-overlapping binding sites. Third, IChem offers a unique opportunity to restrict pharmacophore perception to binding cavities predicted as structurally druggable. Druggability (or ligandability) is predicted on the fly thanks to a robust support vector machine model, immediately after cavity detection. Fourth, IChem rules to select the most valuable pharmacophoric features have been derived from an exhaustive training set of 213 high-resolution protein-ligand X-ray structures featuring non-redundant interaction patterns and 4871 pharmacophoric features. Fifth, the method has been extensively validated on different test sets (Astex Diverse Set, DUD-E, PubChem BioAssay) for its accuracy in ligand posing and virtual screening. We also provide herein several HTS data mimicking real life scenarios with fully validated true positives and true negatives. Such benchmarking data are, to our opinion, much more valuable than commonly used data sets in which actives (usually high affinity ligands) are mixed with chemically similar decoys of unknown affinity for the intended target.

4. Conclusion

We herewith propose an alternative computational method (IChem-Shaper2) to molecular docking to identify ligands from the single knowledge of a protein 3D structure. The concept of structure-based pharmacophores has already been exploited, but rarely led to pharmacophore queries truly adapted to virtual screening purposes. The proposed approach is fully automatized and consists of three consecutive steps, each of which can be customized if necessary: (i) detection of druggable cavities at the surface of the target of interest, (ii) generation of cavity-based pharmacophore queries, and (iii) alignment of library compounds to the structure-based

pharmacophores. The method appears to be quite robust in producing high-quality poses, distinguishing true actives from decoys, and retrieving confirmed hits from high-throughput experimental screens. It should be considered as a novel weapon to the arsenal of current virtual screening methods such as protein-ligand docking or ligand-centric similarity searches. Since virtual screening benchmarks suggest its strong orthogonality to existing methods, we recommend its usage in parallel with docking and/or ligand-based approaches to retrieve different chemotypes and optimize virtual screening hits for medicinal chemistry research.

References

1. Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E. W., Jr. Computational Methods in Drug Discovery. *Pharmacol. Rev.* **2014**, *66*, 334-395.
2. Ripphausen, P.; Nisius, B.; Bajorath, J. State-of-the-Art in Ligand-Based Virtual Screening. *Drug Discov. Today* **2011**, *16*, 372-376.
3. Spyrikis, F.; Cavasotto, C. N. Open Challenges in Structure-Based Virtual Screening: Receptor Modeling, Target Flexibility Consideration and Active Site Water Molecules Description. *Arch. Biochem. Biophys.* **2015**, *583*, 105-119.
4. Leach, A. R.; Gillet, V. J.; Lewis, R. A.; Taylor, R. Three-Dimensional Pharmacophore Methods in Drug Discovery. *J. Med. Chem.* **2010**, *53*, 539-558.
5. Plewczynski, D.; Lazniewski, M.; Augustyniak, R.; Ginalski, K. Can We Trust Docking Results? Evaluation of Seven Commonly Used Programs on PDBbind Database. *J. Comput. Chem.* **2011**, *32*, 742-755.
6. 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.
7. Meslamani, J.; Li, J.; Sutter, J.; Stevens, A.; Bertrand, H. O.; Rognan, D. Protein-Ligand-Based Pharmacophores: Generation and Utility Assessment in Computational Ligand Profiling. *J. Chem. Inf. Model.* **2012**, *52*, 943-955.
8. 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.
9. Salam, N. K.; Nuti, R.; Sherman, W. Novel Method for Generating Structure-Based Pharmacophores Using Energetic Analysis. *J. Chem. Inf. Model.* **2009**, *49*, 2356-2368.
10. Koes, D. R.; Camacho, C. J. ZINCPharmer: Pharmacophore Search of the ZINC Database. *Nucleic Acids Res.* **2012**, *40*, W409-W414.
11. Goodford, P. J. A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules. *J. Med. Chem.* **1985**, *28*, 849-857.
12. 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.

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. 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.
15. Mortier, J.; Dhakal, P.; Volkamer, A. Truly Target-Focused Pharmacophore Modeling: A Novel Tool for Mapping Intermolecular Surfaces. *Molecules* **2018**, *23*, 1959.
16. Schuetz, D. A.; Seidel, T.; Garon, A.; Martini, R.; Korbel, M.; Ecker, G. F.; Langer, T. GRAIL: GRids of pharmacophore Interaction fields. *J. Chem. Theory Comput.* **2018**, *14*, 4958-4970.
17. Ahlstrom, M. M.; Ridderstrom, M.; Luthman, K.; Zamora, I. Virtual Screening and Scaffold Hopping Based on GRID Molecular Interaction Fields. *J. Chem. Inf. Model.* **2005**, *45*, 1313-1323.
18. Ortuso, F.; Langer, T.; Alcaro, S. GBPM: GRID-Based Pharmacophore Model: Concept and Application Studies to Protein-Protein Recognition. *Bioinformatics* **2006**, *22*, 1449-1455.
19. 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.
20. Miranker, A.; Karplus, M. Functionality Maps of Binding Sites: A Multiple Copy Simultaneous Search Method. *Proteins* **1991**, *11*, 29-34.
21. Yu, W.; Lakkaraju, S. K.; Raman, E. P.; Fang, L.; MacKerell, A. D., Jr. Pharmacophore Modeling Using Site-Identification by Ligand Competitive Saturation (SILCS) with Multiple Probe Molecules. *J. Chem. Inf. Model.* **2015**, *55*, 407-420.
22. Hu, B.; Lill, M. A. Protein Pharmacophore Selection Using Hydration-Site Analysis. *J. Chem. Inf. Model.* **2012**, *52*, 1046-1060.
23. Bohm, 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.
24. Schuller, A.; Fechner, U.; Renner, S.; Franke, L.; Weber, L.; Schneider, G. A Pseudo-Ligand Approach to Virtual Screening. *Comb. Chem. High Throughput Screen.* **2006**, *9*, 359-364.
25. Kirchhoff, P. D.; Brown, R.; Kahn, S.; Waldman, M.; Venkatachalam, C. M. Application of Structure-Based Focusing to the Estrogen Receptor. *J. Comput. Chem.* **2001**, *22*, 993-1003.
26. 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.
27. Roland, W. S.; Sanders, M. P.; van Buren, L.; Gouka, R. J.; Gruppen, H.; Vincken, J. P.; Ritschel, T. Snooker Structure-Based Pharmacophore Model Explains Differences in Agonist and Blocker Binding to Bitter Receptor hTAS2R39. *PLoS One* **2015**, *10*, e0118200.
28. Johnson, D. K.; Karanicolas, J. Ultra-High-Throughput Structure-Based Virtual Screening for Small-Molecule Inhibitors of Protein-Protein Interactions. *J. Chem. Inf. Model.* **2016**, *56*, 399-411.
29. Hu, B.; Lill, M. A. Exploring the Potential of Protein-Based Pharmacophore Models in Ligand Pose Prediction and Ranking. *J. Chem. Inf. Model.* **2013**, *53*, 1179-1190.

30. Da Silva, F.; Desaphy, J.; Rognan, D. IChem: A Versatile Toolkit for Detecting, Comparing, and Predicting Protein-Ligand Interactions. *ChemMedChem* **2018**, *13*, 507-510.
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-D404.
32. 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.
33. <http://bioinfo-pharma.u-strasbg.fr/scPDB> (accessed September 2018).
34. Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T.; Mortenson, P. N.; Murray, C. W. Diverse, High-Quality Test Set for the Validation of Protein-Ligand Docking Performance. *J. Med. Chem.* **2007**, *50*, 726-741.
35. <https://www.ccdc.cam.ac.uk/support-and-resources/Downloads/> (accessed September 2018).
36. *Sybyl-X Molecular Modeling Software Packages, 2.1.1*; TRIPOS Associates, Inc.: St. Louis, MO, USA, 2013.
37. 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.
38. 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.
39. 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.
40. National Center for Biotechnology Information. PubChem BioAssay Database; AID = 604. <https://pubchem.ncbi.nlm.nih.gov/bioassay/604> (accessed September 2018).
41. National Center for Biotechnology Information. PubChem BioAssay Database; AID = 644. <https://pubchem.ncbi.nlm.nih.gov/bioassay/644> (accessed September 2018).
42. Molecular Networks GmbH, Erlangen, Germany.
43. OpenEye Scientific Software, Santa Fe, NM 87508, USA.
44. National Center for Biotechnology Information. PubChem BioAssay Database; AID = 743080. <https://pubchem.ncbi.nlm.nih.gov/bioassay/743080> (accessed September 2018).
45. National Center for Biotechnology Information. PubChem BioAssay Database; AID = 1777. <https://pubchem.ncbi.nlm.nih.gov/bioassay/1777> (accessed September 2018).
46. Gehlhaar, D. K.; Verkhivker, G. M.; Rejto, P. A.; Sherman, C. J.; Fogel, D. B.; Fogel, L. J.; Freer, S. T. Molecular Recognition of the Inhibitor AG-1343 by HIV-1 Protease: Conformationally Flexible Docking by Evolutionary Programming. *Chem. Biol.* **1995**, *2*, 317-324.
47. Kurogi, Y.; Guner, O. F. Pharmacophore Modeling and Three-Dimensional Database Searching for Drug Design Using Catalyst. *Curr. Med. Chem.* **2001**, *8*, 1035-1055.
48. Grant, J. A.; Gallardo, M.; Pickup, B. A Fast Method of Molecular Shape Comparison: A Simple Application of a Gaussian Description of Molecular Shape. *J. Comput. Chem.* **1996**, *17*, 1653-1666.
49. Hawkins, P. C.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T. Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database. *J. Chem. Inf. Model.* **2010**, *50*, 572-584.

50. Shape Toolkit 2.0.1. <https://docs.eyesopen.com/toolkits/python/shapetk/index.html> (accessed August 2018).
51. Halgren, T. A. Merck Molecular Force Field. I. Basis, Form, Scope, Parameterization, and Performance of MMFF94. *J. Comp. Chem.* **1996**, *17*, 490-519.
52. Nicholls, A.; Wlodek, S.; Grant, J. A. SAMPL2 and Continuum Modeling. *J. Comput. Aided Mol. Des.* **2010**, *24*, 293-306.
53. SZYBKI theory. <https://docs.eyesopen.com/szybki/szybkitheory.html> (accessed August 2018).
54. Dassault Systèmes, BIOVIA Corp., San Diego, CA 92121, USA.
55. Inte:Ligand GmbH, Vienna, Austria.
56. Poli, G.; Seidel, T.; Langer, T. Conformational Sampling of Small Molecules with iCon: Performance Assessment in Comparison with OMEGA. *Front. Chem.* **2018**, *6*, 229.
57. Jain, A. N. Surflex-Dock 2.1: Robust Performance from Ligand Energetic Modeling, Ring Flexibility, and Knowledge-Based Search. *J. Comput. Aided Mol. Des.* **2007**, *21*, 281-306.
58. Hawkins, P. C.; Skillman, A. G.; Nicholls, A. Comparison of Shape-Matching and Docking as Virtual Screening Tools. *J. Med. Chem.* **2007**, *50*, 74-82.
59. Guner, O. F.; Bowen, J. P. Setting the Record Straight: The Origin of the Pharmacophore Concept. *J. Chem. Inf. Model.* **2014**, *54*, 1269-1283.
60. Spitzer, R.; Jain, A. N. Surflex-Dock: Docking Benchmarks and Real-World Application. *J. Comput. Aided Mol. Des.* **2012**, *26*, 687-699.
61. Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D. Comparative Evaluation of Eight Docking Tools for Docking and Virtual Screening Accuracy. *Proteins* **2004**, *57*, 225-242.
62. Cleves, A. E.; Jain, A. N. Knowledge-Guided Docking: Accurate Prospective Prediction of Bound Configurations of Novel Ligands Using Surflex-Dock. *J. Comput. Aided Mol. Des.* **2015**, *29*, 485-509.
63. Kabsch, W. A Solution for the Best Rotation to Relate Two Sets of Vectors. *Acta Cryst. Sect. A* **1976**, *32*, 922-923.
64. Kuhn, H. W. The Hungarian Method for the Assignment Problem. *Nav. Res. Logist.* **1955**, *2*, 83.
65. Chaput, L.; Martinez-Sanz, J.; Saettel, N.; Mouawad, L. Benchmark of Four Popular Virtual Screening Programs: Construction of the Active/Decoy Dataset Remains a Major Determinant of Measured Performance. *J. Cheminform.* **2016**, *8*, 56.
66. Wang, Y.; Bryant, S. H.; Cheng, T.; Wang, J.; Gindulyte, A.; Shoemaker, B. A.; Thiessen, P. A.; He, S.; Zhang, J. PubChem BioAssay: 2017 Update. *Nucleic Acids Res.* **2017**, *45*, D955-D963.
67. Kruger, D. M.; Evers, A. Comparison of Structure- and Ligand-Based Virtual Screening Protocols Considering Hit List Complementarity and Enrichment Factors. *ChemMedChem* **2010**, *5*, 148-158.
68. Tanrikulu, Y.; Kruger, B.; Proschak, E. The Holistic Integration of Virtual Screening in Drug Discovery. *Drug Discov. Today* **2013**, *18*, 358-364.
69. Tian, S.; Sun, H.; Li, Y.; Pan, P.; Li, D.; Hou, T. Development and Evaluation of an Integrated Virtual Screening Strategy by Combining Molecular Docking and Pharmacophore Searching Based on Multiple Protein Structures. *J. Chem. Inf. Model.* **2013**, *53*, 2743-2756.
70. <http://bioinformatics.psb.ugent.be/webtools/Venn/> (accessed September 2018).

Supporting Information**Table S1.** Sc- PDB Diverse Set of 213 protein- ligand complexes.

Cluster	PDB ID	Ligand ID	Protein name	sc-PDB entries in cluster	DPI value, Å
0	10GS	VWW	Glutathione S-transferase P	18	0.350
1	1KJX	IMP	Adenylosuccinate synthetase	9	0.386
2	2R3A	SAM	Histone-lysine N-methyltransferase SUV39H2	25	0.160
3	2R3F	SC8	Cyclin-dependent kinase 2	6	0.094
4	3E5H	GNP	Ras-related protein Rab-28	171	0.062
5	13PK	ADP	Phosphoglycerate kinase, glycosomal	6	0.451
6	2FDE	385	Protease	81	0.442
7	1V3S	ATP	Signaling protein	31	0.142
8	2FDP	FRP	Beta-secretase 1	66	0.398
10	1V45	3DG	Purine nucleoside phosphorylase	11	0.408
12	3ORF	NAD	Dihydropteridine reductase	58	0.208
14	3ORN	3OR	Dual specificity mitogen-activated protein kinase kinase 1	11	0.473
15	3ORO	AGS	Serine/threonine protein kinase	65	0.353
16	2R4B	GW7	Receptor tyrosine-protein kinase erbB-4	9	0.558
18	3E65	XXZ	Nitric oxide synthase, inducible	8	0.171
20	3ORZ	BI4	3-phosphoinositide-dependent protein kinase 1	10	0.255
21	1KLK	PMD	Dihydrofolate reductase	9	N/A
22	1A28	STR	Progesterone receptor	19	0.155
24	2R4F	RIE	3-hydroxy-3-methylglutaryl-coenzyme A reductase	17	0.118
28	1A2N	TET	UDP-N-acetylglucosamine 1-carboxyvinyltransferase	9	0.266
32	1V79	FR7	Adenosine deaminase	6	0.756
34	2R4T	ADP	Glycogen synthase	13	0.117
36	2FEQ	34P	Prothrombin	76	0.519
37	2R59	PH0	Leukotriene A-4 hydrolase	7	0.136
38	1A42	BZU	Carbonic anhydrase 2	9	N/A
39	2R5C	C6P	Kynurenine aminotransferase	23	0.224
42	3OTF	CMP	Potassium/sodium hyperpolarization-activated cyclic nucleotide-gated channel 4	18	0.509
45	1KNR	FAD	L-aspartate oxidase	125	0.392
46	1KNU	YPA	Peroxisome proliferator-activated receptor gamma	7	0.470
52	3OU2	SAH	SAM-dependent methyltransferase	174	0.077
53	1KOL	NAD	Glutathione-independent formaldehyde dehydrogenase	75	0.096
54	2R6H	FAD	NADH:ubiquinone oxidoreductase, Na translocating, F subunit	32	0.435
55	2R6J	NDP	Eugenol synthase 1	6	0.098
58	1KOR	ANP	Argininosuccinate synthase	21	0.125
59	2R6W	LLB	Estrogen receptor	17	0.219
68	2R7M	AMP	5-formaminoimidazole-4-carboxamide-1-(beta)-D-ribofuranosyl 5'-monophosphate synthetase	28	0.213
70	1A4Z	NAD	Aldehyde dehydrogenase, mitochondrial	31	0.529
73	1V9N	NDP	Malate dehydrogenase	10	0.206
79	3OW3	SMY	cAMP-dependent protein kinase catalytic subunit alpha	8	0.135

80	3E7X	AMP	D-alanine–poly(phosphoribitol) ligase subunit 1	14	0.315
83	3E87	G95	RAC-beta serine/threonine-protein kinase	8	0.427
85	1VBM	YSA	Tyrosine–tRNA ligase	9	0.418
87	1VC2	NAD	Glyceraldehyde 3-phosphate dehydrogenase	52	0.877
88	3OWA	FAD	Acyl-CoA dehydrogenase	22	0.132
89	3OWB	BSM	Heat shock protein HSP 90-alpha	21	0.140
90	1VCF	FMN	Isopentenyl-diphosphate delta-isomerase	8	0.507
92	2R8O	T5X	Transketolase 1	32	0.059
96	1KP8	ATP	60 kDa chaperonin	21	0.225
100	3E8X	NAP	BH1520 protein	17	0.115
102	1KPG	SAH	Cyclopropane mycolic acid synthase 1	33	0.239
104	1A80	NDP	2,5-diketo-D-gluconic acid reductase A	30	0.328
105	3E92	G6A	Mitogen-activated protein kinase 14	21	0.152
107	1VDC	FAD	Thioredoxin reductase 1	28	0.213
111	3E9H	KAAs	Lysine–tRNA ligase	10	0.278
116	4C4F	7CE	Dual specificity protein kinase TTK	8	0.179
118	3OX4	NAD	Alcohol dehydrogenase 2	7	0.252
119	1KQB	FMN	Oxygen-insensitive NAD(P)H nitroreductase	18	N/A
120	2R97	FMN	NAD(P)H dehydrogenase (quinone)	12	0.243
122	4C58	824	Cyclin-G-associated kinase	7	0.201
125	4C5O	FAD	Putative monooxygenase	30	0.818
126	4C61	LMM	Tyrosine-protein kinase JAK2	7	0.228
128	3OY1	589	Mitogen-activated protein kinase 10	7	0.198
130	3OY3	XY3	Tyrosine-protein kinase ABL1	28	0.196
133	1KQM	ANP	Myosin heavy chain, striated muscle	10	0.581
135	2R9R	NAP	Voltage-gated potassium channel subunit beta-2	10	0.239
136	1KQN	NAD	Nicotinamide mononucleotide adenyltransferase 1	6	0.199
151	4C8G	C5P	2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase	9	0.190
156	2FKY	N2T	Kinesin-like protein KIF11	17	0.319
162	4CA6	3EF	Angiotensin-converting enzyme	21	0.529
173	3P0N	BPU	Tankyrase-2	12	0.121
180	1VHN	FMN	tRNA-dihydrouridine synthase	28	0.081
181	2RD2	QSI	Glutamine–tRNA ligase	6	0.376
183	3EBH	BES	M1 family aminopeptidase	12	0.238
184	1VHW	ADN	Purine nucleoside phosphorylase DeoD-type 1	18	0.077
185	3P19	NAP	Putative blue fluorescent protein	91	0.177
189	4CCB	OFG	ALK tyrosine kinase receptor	8	0.203
192	3P23	ADP	Serine/threonine-protein kinase	39	0.400
195	2FOI	JPA	Enoyl-acyl carrier reductase	22	0.475
199	1ADC	PAD	Alcohol dehydrogenase E chain	17	N/A
200	4CDG	ADP	Bloom syndrome protein	16	0.359
202	4CDQ	7VR	Polyprotein	6	0.063
203	3P3C	3P3	UDP-3-O-[3-hydroxymyristoyl] N- acetylglucosamine deacetylase	8	0.035
212	2FPT	ILB	Dihydroorotate dehydrogenase (quinone), mitochondrial	14	0.226
223	3EEI	MTM	5'-methylthioadenosine/S- adenosylhomocysteine nucleosidase	8	0.125
225	3EEJ	53R	Strain CBS138 chromosome J complete sequence	13	0.254
228	3P5S	AVU	CD38 molecule	7	0.186
232	3EFQ	714	Farnesyl pyrophosphate synthase	6	0.215
234	2RH1	CAU	Beta-2 adrenergic receptor	6	0.273
238	3P7N	FMN	Sensor histidine kinase	8	0.316

241	1AJ2	2PH	Dihydropteroate synthase	6	N/A
242	2FSN	ADP	Archaeal actin homolog	15	0.550
246	2FSV	NAP	NAD(P) transhydrogenase subunit beta	9	0.221
247	1KYI	ATP	ATP-dependent protease ATPase subunit HslU	6	0.472
249	3P88	P88	Bile acid receptor	8	0.447
252	2FTO	TMP	Thymidylate synthase	10	0.170
255	3P8X	ZYD	Vitamin D3 receptor	66	0.099
256	1AKW	FMN	Flavodoxin	9	0.043
258	3EHG	ATP	Sensor histidine kinase DesK	26	0.120
259	3P8Z	36A	RNA-directed RNA polymerase NS5	12	0.114
261	1KYX	CRM	6,7-dimethyl-8-ribityllumazine synthase	20	0.358
263	1AM1	ADP	ATP-dependent molecular chaperone HSP82	14	0.148
264	3EHX	BDL	Macrophage metalloelastase	13	0.165
267	2FV9	002	Disintegrin and metalloproteinase domain-containing protein 17	6	0.248
269	2FVC	888	Genome polyprotein	12	0.207
274	4D86	ADP	Poly [ADP-ribose] polymerase 14	6	0.159
276	2RKG	AB1	Pol protein	11	0.144
278	3P9J	P9J	Aurora kinase A	11	0.961
279	2RKU	R78	Serine/threonine-protein kinase PLK1	13	0.142
281	2RL5	2RL	Vascular endothelial growth factor receptor 2	13	0.342
295	4D9T	0JG	Ribosomal protein S6 kinase alpha-3	7	0.273
296	4D9W	X32	Thermolysin	13	0.043
303	1L2T	ATP	Uncharacterized ABC transporter ATP-binding protein MJ0796	12	0.156
307	1L4E	RBZ	Nicotinate-nucleotide–dimethylbenzimidazole phosphoribosyltransferase	8	0.187
314	2G1N	1IG	Renin	8	N/A
319	2UDP	UPP	UDP-glucose 4-epimerase	11	0.129
322	1AQB	RTL	Retinol-binding protein 4	6	0.106
323	3ELJ	GS7	Mitogen-activated protein kinase 8	10	0.121
325	3ELM	24F	Collagenase 3	6	0.145
333	4DC3	2FA	Putative adenosine kinase	14	0.242
336	3EN4	KS1	Proto-oncogene tyrosine-protein kinase Src	22	0.407
351	2UUO	LK3	UDP-N-acetylmuramoylalanine–D-glutamate ligase	8	0.428
354	1AUX	AGS	Synapsin-1	7	0.374
355	3EOS	PK2	Queuine tRNA-ribosyltransferase	12	0.109
356	3EPP	SFG	mRNA cap guanine-N7 methyltransferase	14	0.445
360	3EPT	FDA	Putative FAD-monoxygenase	10	0.519
364	4DFP	0L7	DNA polymerase I, thermostable	9	0.158
369	4DGM	AGI	Casein kinase II subunit alpha	8	0.143
378	1B0H	LYS_LYS_ALN	Periplasmic oligopeptide-binding protein	10	0.140
379	3PD3	A3T	Threonine–tRNA ligase	6	0.183
383	3EQP	T95	Activated CDC42 kinase 1	7	0.482
388	1B0P	TPP	Pyruvate-flavodoxin oxidoreductase	7	0.372
389	3ERK	SB4	Mitogen-activated protein kinase 1	7	0.239
400	1VRW	NAD	Enoyl-ACP reductase	15	0.297
401	4DK5	OKO	Phosphatidylinositol 4,5-bisphosphate 3-kinase catalytic subunit gamma isoform	7	0.476
407	1VSO	AT1	Glutamate receptor ionotropic, kainate 1	9	0.125
411	4DKO	OLM	Envelope glycoprotein gp160	6	0.163
412	1B3D	S27	Stromelysin-1	16	0.643
419	3PEH	IBD	Endoplasmin homolog, putative	7	0.738
427	2GA2	A19	Methionine aminopeptidase 2	6	0.209

431	2UYY	NA7	Putative oxidoreductase GLYR1	13	0.277
432	4DLK	ATP	Phosphoribosylaminoimidazole carboxylase, ATPase subunit	15	0.199
434	1VTK	TMP	Thymidine kinase	9	0.553
444	1LHN	AON	Sex hormone-binding globulin	8	N/A
453	1LIK	ADN	Adenosine kinase	6	0.241
454	2V0I	UD1	Bifunctional protein GlmU	10	0.105
459	1B9I	PXG	Putative UDP-kanosamine synthase aminotransferase subunit	7	0.155
475	3PJG	UGA	UDP-glucose 6-dehydrogenase	6	0.265
478	2V1U	ADP	ORC1-type DNA replication protein 1	11	0.521
488	3EWR	APR	Non-structural protein 3	6	0.219
501	4DQW	ATP	Inosine-5'-monophosphate dehydrogenase	13	0.318
502	3PLQ	RP2	cAMP-dependent protein kinase type I-alpha regulatory subunit	6	0.390
503	4DR9	BB2	Peptide deformylase	16	0.148
511	1W05	W05	Isopenicillin N synthase	20	0.199
512	4DRX	GTP	Tubulin alpha chain	10	0.170
521	3EXH	TPP	Pyruvate dehydrogenase E1 component subunit alpha, somatic form, mitochondrial	19	0.319
532	1BIF	AGS	6-phosphofructo-2-kinase	13	0.201
535	1BJY	CTC	Tetracycline repressor protein class D	7	0.368
538	1LVG	5GP	Guanylate kinase	7	0.184
540	2GLX	NDP	1,5-anhydro-D-fructose reductase	7	0.223
551	3EYG	MI1	Tyrosine-protein kinase JAK1	16	0.142
567	1BOO	SAH	Modification methylase PvuII	17	0.334
586	2GQT	FAD	UDP-N-acetylenolpyruvoylglucosamine reductase	14	0.078
598	2V6G	NAP	3-oxo-Delta(4,5)-steroid 5-beta-reductase	6	0.179
609	3PTQ	NFG	OSIGBa0135C13.7 protein	8	0.438
616	3F3Y	4OA	Bile salt sulfotransferase	6	0.355
621	2GTB	AZP	Orf1ab polyprotein	6	0.229
627	2V95	HCY	Corticosteroid-binding globulin	6	0.166
632	4DYA	0MF	Nucleocapsid protein	6	1.304
640	1W7K	ADP	Dihydrofolate synthase	10	0.185
649	4E0I	FAD	Mitochondrial FAD-linked sulfhydryl oxidase ERV1	10	0.503
650	3F82	353	Hepatocyte growth factor receptor	8	0.475
665	1C1C	612	Reverse transcriptase/ribonuclease H	6	0.910
677	3PZB	NAP	Aspartate-semialdehyde dehydrogenase	9	0.157
685	1C30	ADP	Carbamoyl-phosphate synthase large chain	6	0.141
689	3FBU	COA	Acetyltransferase, GNAT family	7	0.144
704	3Q0U	LL3	HTH-type transcriptional regulator EthR	6	0.100
759	1CBF	SAH	Cobalt-precorrin-4 C(11)-methyltransferase	6	0.224
765	2HA8	SAH	Probable methyltransferase TARBP1	11	0.088
768	2VFZ	UPF	N-acetyllactosaminide alpha-1,3-galactosyltransferase	9	0.583
773	4E7Z	ADP	Unconventional myosin-VI	10	0.258
785	1WKG	POI	Acetylmornithine/acetyl-lysine aminotransferase	11	0.312
801	4EAW	0NQ	RNA-directed RNA polymerase	9	0.277
802	4EB3	0O3	4-hydroxy-3-methylbut-2-enyl diphosphate reductase	7	0.141
826	3FLK	NAD	D-malate dehydrogenase [decarboxylating]	8	0.168
885	2VNA	NAP	Prostaglandin reductase 2	6	0.179
889	3QCF	NXY	Receptor-type tyrosine-protein phosphatase γ	7	0.461
903	1MP3	TTP	Glucose-1-phosphate thymidyltransferase	6	N/A

969	3FY4	FAD	(6-4)DNA photolyase	11	0.858
978	2HSD	NAD	3-alpha-(or 20-beta)-hydroxysteroid dehydrogenase	6	N/A
984	3QGZ	ADN	Histidine triad nucleotide-binding protein 1	7	0.025
1031	2VWW	7X2	Ephrin type-B receptor 4	7	0.178
1061	3G5E	Q74	Aldose reductase	7	0.103
1073	3QOV	ADP	Phenylacetate-coenzyme A ligase	8	0.180
1099	2W0J	ZAT	Serine/threonine-protein kinase Chk2	7	0.165
1163	1XOI	288	Glycogen phosphorylase, liver form	6	0.214
1202	3R04	UNQ	Serine/threonine-protein kinase pim-1	7	0.099
1260	3GJQ	TRP_GLU_ HIS_ASP_ ACE	Caspase-3	7	N/A
1265	2WE3	DUT	Deoxyuridine 5'-triphosphate nucleotidohydrolase	6	0.211
1271	1XWK	GDN	Glutathione S-transferase Mu 1	7	0.682
1310	4FHH	OU3	Vitamin D3 receptor A	9	0.282
1418	2WQO	VGK	Serine/threonine-protein kinase Nek2	6	0.204
1440	3RLL	RLL	Androgen receptor	10	0.133
1453	4FSM	HK1	Serine/threonine-protein kinase Chk1	12	0.178
1505	1O6H	W37	Squalene-hopene cyclase	7	0.374
1717	4GFD	0YB	Thymidylate kinase	6	0.120
1719	4GFN	SUY	DNA gyrase subunit B	10	0.136
1801	4GPJ	0Q1	Bromodomain-containing protein 4	6	0.116
1845	4GV2	5ME	Poly [ADP-ribose] polymerase 3	19	1.312
2170	3IUB	FG2	Pantothenate synthetase	6	0.071
2615	4JD4	JDM	Dihydroorotate dehydrogenase (fumarate)	10	0.059
2716	1SQB	AZO	Cytochrome b	6	0.549
2898	4KFN	1QR	Nicotinamide phosphoribosyltransferase	8	0.109
3197	3ZCM	PX3	Integrase	13	0.088

The Diffraction Precision Index (DPI) is calculated according to Kumar, K. S. *et al.* Online_DPI: a web server to calculate the diffraction precision index for a protein structure. *J. Appl. Crystallogr.* **2015**, *48*, 939-942.

N/A: not available. The Diffraction precision index (DPI) cannot be calculated due to insufficient parameters.

Table S2. Astex Diverse Set of 85 protein-ligand complexes.

PDB ID	Ligand ID	Protein Name	DPI, Å
1G9V	RQ3	Hemoglobin alpha chain	0.146
1GKC	NFH	92 kDa type IV collagenase	0.316
1GM8	SOX	Penicillin G acylase beta subunit	0.181
1GPK	HUP	Acetylcholinesterase	0.140
1HNN	SKF	Phenylethanolamine N-methyltransferase	0.292
1HP0	AD3	Inosine-adenosine-guanosine-preferring nucleoside hydrolase	0.273
1HQ2	PH2	6-hydroxymethyl-7,8-dihydropterin pyrophosphokinase	0.047
1HVY	D16	Thymidylate synthase	0.212
1HWI	115	HMG-CoA reductase	0.291
1HWW	SWA	Alpha-mannosidase II	0.150
1IA1	TQ3	Dihydrofolate reductase	0.117
1IG3	VIB	Thiamin pyrophosphokinase	0.144
1J3J	CP6	Bifunctional dihydrofolate reductase-thymidylate synthase	0.318
1JD0	AZM	Carbonic anhydrase XII	0.077
1JJE	BYS	IMP-1 metallo beta-lactamase	0.158
1JLA	TNK	HIV-1 RT A-chain	0.502
1K3U	IAD	Tryptophan synthase alpha chain	0.088
1KE5	LS1	Cell division protein kinase 2	0.338
1KZK	JE2	Protease	0.029
1L2S	STC	Beta-lactamase	0.153
1L7F	BCZ	Neuraminidase	0.099
1LPZ	CMB	Blood coagulation factor Xa	0.497
1LRH	NLA	Auxin-binding protein 1	0.173
1M2Z	DEX	Glucocorticoid receptor	0.936
1MEH	MOA	Inosine-5'-monophosphate dehydrogenase	0.142
1MMV	3AR	Nitric-oxide synthase, brain	0.212
1MZC	BNE	Protein farnesyltransferase beta subunit	0.133
1N1M	A3M	Dipeptidyl peptidase IV soluble form	0.812
1N2J	PAF	Pantothenate synthetase	0.137
1N2V	BDI	Queuine tRNA-ribosyltransferase	0.215
1N46	PFA	Thyroid hormone receptor beta-1	0.329
1NAV	IH5	Hormone receptor alpha 1, THRA1	0.297
1OF1	SCT	Thymidine kinase	0.144
1OF6	DTY	Phospho-2-dehydro-3-deoxyheptonate aldolase, tyrosine-inhibited	0.238
1OPK	P16	Proto-oncogene tyrosine-protein kinase ABL1	0.122
1OQ5	CEL	Carbonic anhydrase II	0.085
1OWE	675	Urokinase-type plasminogen activator	0.133
1OYT	FSN	Thrombin heavy chain	0.094
1P2Y	NCT	Cytochrome p450cam	0.369
1P62	GEO	Deoxycytidine kinase	0.129
1PMN	984	Mitogen-activated protein kinase 10	0.291
1Q1G	MTI	Uridine phosphorylase putative	0.191
1Q41	IXM	Glycogen synthase kinase-3 beta	0.156
1Q4G	BFL	Prostaglandin G/H synthase 1	0.139
1R1H	BIR	Nepriylsin	0.195
1R55	097	Adam 33	0.112
1R58	AO5	Methionine aminopeptidase 2	0.198
1R9O	FLP	Cytochrome p450 2C9	0.172
1S19	MC9	Vitamin D3 receptor	0.181
1S3V	TQD	Dihydrofolate reductase	0.156
1SG0	STL	NRH dehydrogenase [quinone] 2	0.086
1SJ0	E4D	Estrogen receptor	0.206
1SQ5	PAU	Pantothenate kinase	0.283

1SQN	NDR	Progesterone receptor	0.082
1T40	ID5	Aldose reductase	0.118
1T46	STI	Homo sapiens V-kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog	0.083
1T9B	1CS	Acetolactate synthase, mitochondrial	0.157
1TOW	CRZ	Fatty acid-binding protein, adipocyte	0.249
1TT1	KAI	Glutamate receptor, ionotropic kainate 2	0.161
1TZ8	DES	Transthyretin	0.125
1U1C	BAU	Uridine phosphorylase	0.334
1U4D	DBQ	Activated Cdc42 kinase 1	0.239
1UML	FR4	Adenosine deaminase	0.896
1UNL	RRC	Cyclin-dependent kinase 5	0.250
1UOU	CMU	Thymidine phosphorylase	0.604
1V0P	PVB	Cell division control protein 2 homolog	0.193
1V48	HA1	Purine nucleoside phosphorylase	0.282
1V4S	MRK	Glucokinase isoform 2	0.298
1VCJ	IBA	Neuraminidase	0.406
1W1P	GIO	Chitinase B	0.275
1W2G	THM	Thymidylate kinase TMK	0.199
1X8X	SO4	Tyrosyl-tRNA synthetase	0.170
1XM6	5RM	cAMP-specific 3',5'-cyclic phosphodiesterase 4B	0.139
1XOQ	ROF	cAMP-specific 3',5'-cyclic phosphodiesterase 4D	0.122
1XOZ	CIA	cGMP-specific 3',5'-cyclic phosphodiesterase	0.063
1Y6B	AAX	Vascular endothelial growth factor receptor 2	0.188
1YGC	905	Coagulation factor VII	0.126
1YQY	915	Lethal factor	0.323
1YV3	BIT	Myosin II heavy chain	0.138
1YVF	PH7	HCV NS5B polymerase	0.390
1YWR	LI9	Mitogen-activated protein kinase 14	0.185
1Z95	198	Androgen receptor	0.147
2BM2	PM2	Human beta2 tryptase	0.297
2BR1	PFP	Serine/threonine-protein kinase CHK1	0.143
2BSM	BSM	Heat shock protein HSP90-alpha	0.173

The Diffraction Precision Index (DPI) is calculated according to Kumar, K. S. *et al.* Online_DPI: A Web Server to Calculate the Diffraction Precision Index for a Protein Structure. *J. Appl. Crystallogr.* **2015**, *48*, 939-942.

Table S3. 10 DUD-E entries.

Protein Name	PDB ID	Ligand ID	Number of pharmacophoric features	Number of actives	Number of decoys
<i>G protein-coupled receptors</i>					
Adenosine A2A receptor (AA2AR)	3PWH	ZMA	35	482	31,500
Beta2 adrenergic receptor (ADRB2)	3NY8	JRZ	42	231	15,000
<i>Nuclear hormone receptors</i>					
Androgen receptor (ANDR)	2AM9	TES	33	269	14,350
Glucocorticoid receptor (GCR)	1P93	DEX	36	258	15,000
<i>Other enzymes</i>					
Adenosine deaminase (ADA)	1A41	DCF	19	282	16,900
Prostaglandin G/H synthase 2 (PGH2)	3LN1	CEL	44	104	6958
<i>Proteases</i>					
Angiotensin-converting enzyme (ACE)	3ZQZ	SLC	28	139	8700
Renin (RENI)	3SFC	S53	43	293	16,450
<i>Protein kinases</i>					
Fibroblast growth factor receptor 1 (FGFR1)	3TT0	07J	23	93	5450
RAC-alpha protein kinase (AKT1)	4EKL	0RF	41	435	123,150

Table S4. Shaper2 force-field for aligning ligand atoms to cavity features.

```
#####
#      DEFINE      #
#####
##### define degree (independent of explicit/implicit)
DEFINE hd1 [X1H0,X2H1,X3H2,X4H3,X5H4,X6H5]
DEFINE hd2 [X2H0,X3H1,X4H2,X5H3,X6H4]
DEFINE hd3 [X3H0,X4H1,X5H2,X6H3]
DEFINE hd4 [X4H0,X5H1,X6H2]
##### hydrophobic
DEFINE php [#6,#16&$hd2&!$(S=*) ,#35,#53;R0;!$(~[!#1;!#6;!$(#16;$hd2)])]
DEFINE thp [$php;$hd1]
DEFINE hp [$php;!$hd1]
DEFINE ehp [$hp;!$(*)($hp)]($hp)]
##### acceptors
DEFINE ACamine [N;!$(N*=[!#6]);!$(N~[!#6;!#1]);!$(Na;!$(N#*);!$(N=*))
DEFINE ACphosphate [O;$hd1;$O~P(~O)~O]
DEFINE ACcarboxylate [O;$hd1;$O[C;!$(N)=O],$(O=[C;!$(N)]O;$hd1)]
DEFINE ACwater [OH2]
DEFINE AChet6N [nH0;X2;$n1aaaa1]
DEFINE ACphosphinyl [O;$O=P);!$(O=P~O)]
DEFINE ACSulphoxide [O;$O=[S;!$(S(~O)~O);$(S([#6])[#6])]
#DEFINE ACprimaryAmine [$ACamine;$hd1;!X4] #leave off for implicit charge
DEFINE AChet5N [nH0;X2;$n1aaaa1]
DEFINE ACthiocarbonyl [S;X1;$S=[#6]]
DEFINE AChydroxyl [O;$hd1;$O-[C;!$(C=*)]
DEFINE ACSulphate [O;$hd1;$O~S(~O)~O]
#DEFINE ACtertiaryAmine [$ACamine;$hd3] #leave off for implicit charge
DEFINE ACamide [O;$O=[#6][#7];!$(O=[#6]([#7])[#7,#8,#16])]
DEFINE ACCarbamate [O;$O=[#6]([#7])[#8]]
DEFINE ACurea [O;$O=[#6]([#7])[#7]]
DEFINE ACester [O;$O=[#6][#8]*;!$(O=[#6]([#7,#8,#16])[#8]*)]
DEFINE ACnitrile [N;$hd1;$N#C]
DEFINE ACimine [N;!$hd3;$N(=C)C,$N=[#6];!$(N=[#6][#7,#8;!$(S=O)])]
DEFINE ACketone [O;$hd1;$O=[#6;$([H2]),$([H1]-[#6]),$(*)($[6])[#6])]
#DEFINE ACsecondaryAmine [$ACamine;$hd2;!X4] #leave off for implicit charge
DEFINE ACphenol [O;$hd1;$Oa]
DEFINE ACether [O;$*(#[6];!$(=[O,S,N]))[#6;!$(=[O,S,N])]
DEFINE ACprimaryAniline [N;$Na;$hd1]
DEFINE ACnitro [O;$hd1;$O~N~[O;$hd1]]
DEFINE AChet5O [o;X2;$o1cccc1),$o1cccc1);!$(#[6]=O)
DEFINE ACSulphone [O;$O=[S;$S(~O)~O]([#6,#7])[#6])]
### strong acceptors
DEFINE strongAcceptor [$ACphosphate,$ACcarboxylate,$ACwater,$AChet6N,$ACphosphinyl]
### moderate acceptors
#DEFINE moderateAcceptor
[$ACSulphoxide,$ACprimaryAmine,$AChet5N,$ACthiocarbonyl,$AChydroxyl,$ACSulphate,$ACtertiaryAmine,$
ACamide,$ACcarbamate,$ACurea]
DEFINE moderateAcceptor
[$ACSulphoxide,$AChet5N,$ACthiocarbonyl,$AChydroxyl,$ACSulphate,$ACamide,$ACcarbamate,$ACurea]
### weak acceptors
#DEFINE weakAcceptor
[$ACnitrile,$ACimine,$ACketone,$ACsecondaryAmine,$ACester,$ACphenol,$ACether,$ACprimaryAniline,$ACn
itro,$AChet5O,$ACSulphone]
```

```

DEFINE weakAcceptor
[$ACnitrile,$ACimine,$ACketone,$ACester,$ACphenol,$ACether,$ACprimaryAniline,$ACnitro,$AChet5O,$ACsulphone]
##### donors
DEFINE Damine [N;!$(N*=[!#6]);!$(N~[!#6;!#1]);!(Na);!(N#*);!(N=*)]
DEFINE Dhet5NH [nH;$(n1aaaa1),$(n1aaaaa1)]
DEFINE DNpH [NH,H2,H3;+]
#DEFINE DacidOH [OH1;$(O-[C,S,P]=[O,S])] #leave off for implicit protonation
DEFINE Dhydroxyl [OH1;$hd1;$(O-C);!(OC=[O,N,S])]
DEFINE Dwater [OH2]
DEFINE DprimaryAmide [N;$hd1;$(NC=O),$(NS=O)]
DEFINE DanilineNH [NH1,NH2;$hd2;$(Nc);!(NS(=O)=O)]
DEFINE DamidineNH [NH1,NH2;$(N~C~N),$(N~C(~N)~N)]
DEFINE DsecondaryAmide [#7;$hd2;$(*[#6,#16]=O);!(N(a)S=O)]
DEFINE DanilineNH2 [N;$hd1;$(Nc)]
DEFINE DhydraN [NH1,NH2,NH3;$hd1&$(NN[#6]),$hd2&$(N(N)[#6])]
DEFINE DimineNH [NH1;$(N=C)]
DEFINE DphenylOH [OH1;$(Oc)]
DEFINE DprimaryAmine [$Damine;$hd1]
DEFINE DsecondaryAmine [$Damine;$hd2]
### strong donors
#DEFINE strongDonor [$Dhet5NH,$DNpH,$DacidOH,$Dhydroxyl]
DEFINE strongDonor [$Dhet5NH,$DNpH,$Dhydroxyl]
### moderate donors
DEFINE moderateDonor
[$Dwater,$DprimaryAmide,$DanilineNH,$DamidineNH,$DsecondaryAmide,$DanilineNH2]
### weak donors
DEFINE weakDonor [$DhydraN,$DimineNH,$DphenylOH,$DprimaryAmine,$DsecondaryAmine]
##### anion intermediate
DEFINE negHet [#8,#16;$hd1]
DEFINE terminalHet [#7,#8,#16;$hd1]
DEFINE ANarylsulfonamide [N;$(N(a)S(=O)(=O)*)]
DEFINE ANmalonic [C;!$hd4;$(C(C=[O,S])C=[O,S])]
DEFINE ANarylthiol [S;$hd1;$(Sa)]
DEFINE ANhalideion [I,Br,Cl,F;!H0,-]
DEFINE ANhydroxylamine [O;$hd1;$(ON~C),$(O[n+]),$(O=n);!(ONC=[S,O,N])]
##### cation intermediate
DEFINE CATnonewN [#7;!$(NC=O);!(NS(=O)=O)]
DEFINE CATguanidine [CATnonewN]!:[#6](!:[CATnonewN])!:[CATnonewN]
DEFINE CATguanidineC [#6]~[CATguanidine]
DEFINE CATamine [N;!$(N*=[!#6]);!$(N~[!#6;!#1]);!(Na);!(N=*);!(N#*);!(#[7;X0])]
##### Zn intermediates
DEFINE hydroxamate O=[CX3]N[O-]
DEFINE reverseHydrox O=[CH][NX3][O-]
#
#
#####
#          TYPES          #
#####
TYPE donor
TYPE acceptor
TYPE cation
TYPE anion
TYPE rings
TYPE hydrophobe
TYPE metal

```

```

TYPE donac
#
#
#####
#      PATTERNS      #
#####
##### rings
PATTERN rings [R]~1~[R]~[R]~[R]1
PATTERN rings [R]~1~[R]~[R]~[R]~[R]1
PATTERN rings [R]~1~[R]~[R]~[R]~[R]1
PATTERN rings [R]~1~[R]~[R]~[R]~[R]~[R]1
### hydrophobic
# terminal hp
PATTERN hydrophobe [$thp]~*~([$thp])~[$thp]          #triple
PATTERN hydrophobe [$thp][!$(~[$thp])(~[$thp])~[$thp]);!$(*=[N,S,O])[thp] #double
PATTERN hydrophobe [$thp;!$(*~*~[$thp]);$(*~[$php])]    #single
PATTERN hydrophobe [$thp;#35,#53]                        #large
# non-terminal hp
PATTERN hydrophobe [$ehp][$hp][$hp][$ehp]
PATTERN hydrophobe [$ehp]([$ehp])[$hp][$ehp]
PATTERN hydrophobe [$hp]([$ehp])([$ehp])[$ehp]
PATTERN hydrophobe [$ehp][$hp][$hp][$ehp]
PATTERN hydrophobe [$ehp][$hp][$hp;$(*[$hp])[$hp][$ehp])]
### donor/acceptor patterns
PATTERN acceptor [$strongAcceptor,$moderateAcceptor,$weakAcceptor]
PATTERN donor [$strongDonor,$moderateDonor,$weakDonor]
### anion/cation patterns
# cations
PATTERN cation [$CATnonewN]!:[#6;!$(C(N)(N)N)](!:[$CATnonewN])!:[$CATnonewN] #guanidine
PATTERN cation [$CATnonewN]!:[#6;!$(CATguanidineC)]!$(C(N)N)!:[$CATnonewN] #amidine
PATTERN cation n:1cnce1          #azole
PATTERN cation [$CATamine]
# anions
PATTERN anion [$negHet][#6X3]~[$terminalHet]          #carboxylate
PATTERN anion [$negHet][#16X4](~[$terminalHet])~[$terminalHet] #sulfonate
PATTERN anion [$negHet][#15X4](=O)[$negHet,$terminalHet] #phosphonate
PATTERN anion [n;$hd2]1[n;$hd2][n;$hd2][n;$hd2]c1    #tetrazole
PATTERN anion [$ANarylSulfonamide,$ANmalonic,$ANarylthiol,$ANhalideion,$ANhydroxylamine]
#####
#      Type Patterns      #
#####
##### metal binders #####
#### Pattern Ca_Mg & ZN
PATTERN metal [#8;-]
PATTERN metal [#16;-]
#### Pattern Ca_Mg
# PATTERN metal [nh0]1aaaa1
##### Pattern Zn
PATTERN metal [nh0;-]
PATTERN metal [#7;-]
PATTERN metal [#8;-;!$([$hydroxamate,$reverseHydrox])]
PATTERN metal O=[CX3]N[O-]
PATTERN metal O=[CH][NX3][O-]
PATTERN metal [S]([N,-1])(=[O])(=[O])
PATTERN metal O=C[ND2][O-]

```



```
#####  
#   CAVITY Type Patterns   #  
#####  
PATTERN donor [14#7]  
PATTERN acceptor [14#8]  
PATTERN cation [15#7]  
PATTERN anion [17#8]  
PATTERN rings [15#6]  
PATTERN hydrophobe [13#6]  
#PATTERN donac [15#8]  
PATTERN metal [54#30]  
#  
#  
#####  
#   INTERACTIONS   #  
#####  
INTERACTION rings rings attractive gaussian weight=1.0 radius=1.0  
INTERACTION hydrophobe hydrophobe attractive gaussian weight=1.0 radius=1.0  
INTERACTION donor donor attractive gaussian weight=1.0 radius=1.0  
INTERACTION donac donac attractive gaussian weight=1.0 radius=1.0  
INTERACTION acceptor acceptor attractive gaussian weight=1.0 radius=1.0  
INTERACTION cation cation attractive gaussian weight=1.0 radius=1.0  
INTERACTION anion anion attractive gaussian weight=1.0 radius=1.0  
INTERACTION metal metal attractive gaussian weight=10.0 radius=1.0  
#  
INTERACTION rings hydrophobe attractive gaussian weight=1.0 radius=1.0  
INTERACTION donor cation attractive gaussian weight=1.0 radius=1.0  
INTERACTION acceptor anion attractive gaussian weight=1.0 radius=1.0  
INTERACTION donac donor attractive gaussian weight=1.0 radius=1.0  
INTERACTION donac acceptor attractive gaussian weight=1.0 radius=1.0  
INTERACTION metal anion attractive gaussian weight=10.0 radius=1.0  
INTERACTION metal acceptor attractive gaussian weight=10.0 radius=1.0  
INTERACTION metal donac attractive gaussian weight=10.0 radius=1.0  
#  
INTERACTION anion cation repulsive gaussian weight=1.0 radius=1.0  
INTERACTION metal hydrophobe repulsive gaussian weight=10.0 radius=1.0  
INTERACTION metal rings repulsive gaussian weight=10.0 radius=1.0  
INTERACTION metal cation repulsive gaussian weight=10.0 radius=1.0  
INTERACTION metal donor repulsive gaussian weight=10.0 radius=1.0  
#  
INTERACTION hydrophobe donor repulsive gaussian weight=1.0 radius=1.0  
INTERACTION hydrophobe acceptor repulsive gaussian weight=1.0 radius=1.0  
INTERACTION hydrophobe donac repulsive gaussian weight=1.0 radius=1.0  
INTERACTION hydrophobe cation repulsive gaussian weight=1.0 radius=1.0  
INTERACTION hydrophobe anion repulsive gaussian weight=1.0 radius=1.0
```

Take-home Messages

The work portrayed in this chapter presents a novel structure-based virtual screening tool whose performance is comparable to that of other *in silico* approaches. However, the total amount of time required to fully process the ligands (notably during the treatment of desolvation effects: 5 seconds per pose x hundreds of poses per ligand) is larger than that consumed by most molecular docking tools or ligand-based screening programs, which, as a result, impedes the application of this novel approach to virtual screening campaigns using ultra-large chemolibraries that comprise millions (or even billions) of ligands. A possible resolution is to start such screens with faster approaches, and subject only the top scorers issued from these methods (0.01-0.1% of the total population) to our protocol in order to rescore the ligands. Other approaches may be used simultaneously, yielding as many hit lists as the employed methods, after which all hit lists (including that given by our procedure) are fused to select, for example, the top-ranked compounds that the lists have in common. The ligand-aligning script can also be modified to allow a faster calculation on multiple cores, in hopes of reducing the computation time.

Another remarkable point of this work is that we used experimentally confirmed data from a highly-queried public repository (PubChem BioAssay) to validate our method. This rules out the issue regarding unknown potency values of presumably inactive molecules (“decoys”) inherent in artificially constructed data sets (DUD, DUD-E). Though a few data-processing steps were carried out before the employment of these data in retrospective virtual screening, the question as to whether the resulting ligand sets are still biased was not fully addressed. Starting from our first attempts described in this study, and taking inspiration from other publications reporting data set construction based on PubChem BioAssay data (reviewed in the Chapter 1 of this manuscript), we developed a novel unbiased data collection entitled LIT-PCBA from fully validated components in terms of bioactivity towards a macromolecular target, which can be applied to validating both ligand-based and structure-based *in silico* screening approaches. More details concerning the preparation and the evaluation of this new data set will be given in the next chapter of this thesis.

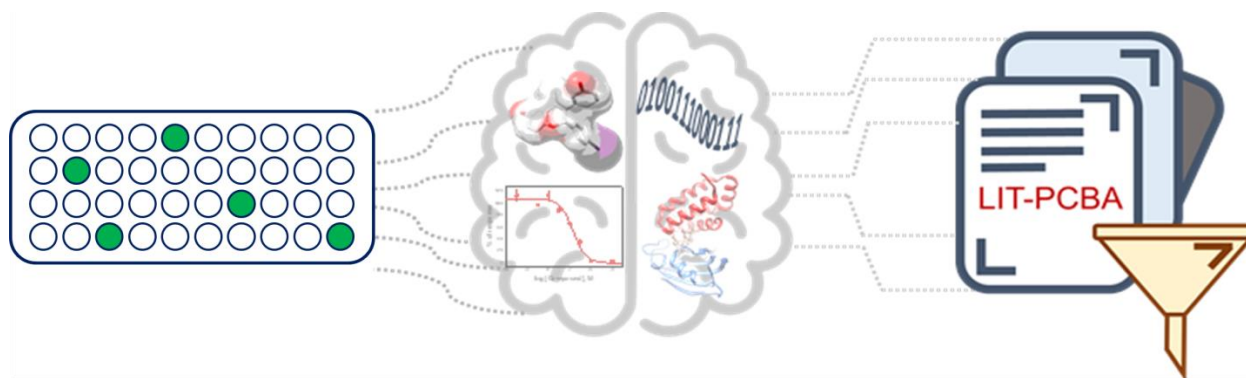
Chapter 3

LIT-PCBA: An Unbiased Data Set for Machine Learning and Virtual Screening

Chapter 3. LIT-PCBA: An Unbiased Data Set for Machine Learning and Virtual Screening

As previously mentioned in the manuscript, artificially constructed ligand sets classically used by the cheminformatics community (DUD, DUD-E, DEKOIS 2.0) suffer from multiple drawbacks ranging from the presence of possible false negatives/positives to obvious and hidden design bias, therefore overestimating the true accuracy of virtual screening methods. In this chapter, we present a novel data set entitled LIT-PCBA that was specifically designed for virtual screening and machine learning, relying on data from dose-response PubChem bioactivity assays that were additionally processed to avoid the issues inherent in other databases. The resulting ligand sets were finally unbiased by the recently described asymmetric validation embedding procedure to afford the final data collection that mimics experimental screening decks in terms of hit rate (ratio of active to inactive compounds) and potency distribution, and is ready for benchmarking novel virtual screening methods (both ligand-based and structure-based), notably those relying on machine learning. The work portrayed in this chapter has been published as an original research paper in the *Journal of Chemical Information and Modeling*, and was presented at the 9th Meeting of the French Cheminformatics Society in November 2019 in Paris.

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** (in press). doi: [10.1021/acs.jcim.0c00155](https://doi.org/10.1021/acs.jcim.0c00155).



1. Introduction

Virtual screening (VS) of compound libraries has established itself, notably in academic settings, as a fast and cost-efficient alternative to high-throughput screening (HTS) for identifying preliminary hits of pharmaceutically interesting targets.¹⁻³ Because of the availability of hundreds of virtual screening tools,⁴ choosing the right method for a specific project often relies on benchmarking studies designed to delineate the context-specific advantages and drawbacks of each method. Many target-specific ligand sets⁵⁻¹⁰ and statistical evaluation protocols¹¹⁻¹³ have been reported during the last decade to pinpoint the ability of a VS method to prioritize, for purchase and validation, the shortest possible hit list with an optimal enrichment factor in true actives. In the early 2000s, such data sets were limited in size due to the paucity of available experimental data. Inactive compounds were randomly chosen from databases of drug-like chemicals.^{5,14,15} Very soon, it appeared that the random selection of presumably inactive molecules (“decoys”) led to artificially high enrichment values, because of the bias in molecular property ranges (e.g. molecular weight) that often differed between active and inactive sets.¹⁶ One of the first attempts at designing a docking-dedicated benchmarking database led to the introduction of the DUD data collection,⁶ gathering 2950 ligands of 40 different targets from the literature, seeded among property-matched decoys (36 decoys for each active) from the ZINC archive of commercially available ligands.¹⁷ In DUD, decoys were specifically designed to share physicochemical properties with actives but with a different chemical topology. Despite the caution given to the selection of decoys, independent groups rapidly noticed three major issues for both DUD active and decoy sets: (i) actives tend to spread over a few dominant scaffolds (so-called “analog bias”),¹⁸ (ii) decoys exhibited molecular net charges different from those of actives,¹⁹ and (iii) decoys were too similar to true actives and were likely false negatives.⁸ The DUD set was upgraded to a revised version (DUD-E)¹⁰ describing an enhanced and more diverse target space (102 targets), containing 22,886 clustered true actives with known experimental data from the ChEMBL database,²⁰ enhancing the proportion of decoys in the ligand sets (50 decoys for each active). The debate on the best protocol to select decoys has led to many contributions^{8,9,21} to design novel decoy sets. As an alternative to DUD-E, other sources of active compounds (e.g. PubChem BioAssay²²) have also been utilized. Note-worthy is the MUV database⁷ that provides many advantages: (i) the data sets (targets, ligands, assay conditions) are publicly available, (ii) the included compounds are drug-like, (iii) many experimental data were

utilized to remove false positives and assay artifacts, and (iv) ligands were selected by a nearest neighbor analysis to permit a spatially unbiased distribution of actives and decoys. Consequently, the MUV data collection is considered more challenging than DUD-E.²³

For many years, the DUD-E has been considered the gold standard for benchmarking VS and machine learning methods, until recent reports²⁴⁻²⁷ warned the community about both obvious and hidden bias in its design. First, Chaput *et al.*²⁴ noticed that differences in key molecular properties (polar surface area, H-bond donor count, embranchment count) remain between DUD-E actives and decoys. Moreover, chemical bias is still present in actives that tend to resemble target-bound PDB ligands, thereby overestimating the real discriminatory power of standard docking methods.²⁴ In 2018, Wallach and Heifets described the asymmetric validation embedding (AVE) method²⁵ to quantify data set bias and optimally design training/validation ligand sets. When applied to ligand-based VS methods, all standard benchmarking data collections (DUD, DUD-E, MUV) were shown to be massively biased, rewarding memorization rather than learning.²⁵ The latter danger is even higher for currently popular artificial intelligence methods (e.g. machine learning, deep neural networks)²⁸ that are hardly interpretable and tightly dependent on the quality of the input data and the way they are split to train and test a model. Two different groups^{26,27} just reported hidden bias in the DUD-E data set when applying deep neural networks (DNNs) to either predict binding affinities or classify complexes as active/inactive from X-ray structures or docking poses. Intriguingly, DNNs trained with rigorous cross-validation procedures on simple ligand descriptors were almost as accurate as those trained on protein-ligand attributes, suggesting that deep learning did not learn anything about the physics of protein-ligand interactions. Strikingly, the literature is full of overoptimistic reports describing machine learning models²⁹⁻³¹ with near perfect performances on the above-described data sets, although true VS practitioners have known for long that such an accuracy level does not mirror the proportion of experimentally confirmed hits in real prospective VS experiments.

There is more than ever an urgent need to design an unbiased and realistic data set specifically dedicated to virtual screening and machine learning.²⁷ We herewith present our contribution based on the following eight principles:

- (i) The data set should mimic “real-life” screening decks and guide VS methods to discriminate moderately potent actives (primary hits) from inactive compounds;

- (ii) The potency of all compounds (actives, inactives) for a particular target should have been determined experimentally in homogeneous conditions;
- (iii) The ratio of actives to inactives should reflect hit rates typically observed in HTS campaigns against targets of pharmaceutical interest;³²
- (iv) Actives should be filtered to remove false positives, frequent hitters, assay artifacts and truly “undruggable” compounds; besides, dose-response curves should be available for all actives;
- (v) Active and inactive compounds should span common molecular property ranges;
- (vi) Potency distribution of confirmed actives should not be biased towards too high affinities and should ideally mimic that observed in HTS decks;
- (vii) The data set should be applicable to both ligand-based and structure-based virtual screening;
- (viii) Unbiased training and validation sets should be available for machine learning.

We therefore decided to choose the PubChem BioAssay database (PCBA)²² as the source of experimental bioactivity data. PCBA is an open-access archive hosted by the National Center for Biotechnology Information (NCBI), the National Library of Medicine (NLM) and the National Institute of Health (NIH). At the time this manuscript was written, the database stores over 1 million assay records, 134,000 of which are annotated by an activity type (IC₅₀, EC₅₀, K_d, K_i). It covers about 7200 HTS projects from 80 sources (pharmaceutical companies, academic sources, governmental sources) on a chemical repository of 2.2 million compounds. The database can be easily queried according to numerous filters and is a first-class source of bioactivity data for computer-aided drug discovery.³³

We hereby describe a workflow for retrieving assays of interest and filtering compounds and targets for bioactivity data acquisition. The retrieved target sets were then subjected to state-of-the-art virtual screening experiments in order to ascertain their suitability. The final data collection entitled LIT-PCBA contains 15 targets, 7844 true active and 407,381 true inactive compounds in total; with ready-to-use input files (ligands, targets) that have been unbiased for machine learning applications. It is available for download at <http://drugdesign.unistra.fr/LIT-PCBA>.

2. Computational Methods

2.1. Data Selection

Bioactivity data were retrieved from the PubChem BioAssay database,²² where all information on true active and true inactive substances for a protein target is provided based on experimental results from confirmatory dose-response bioactivity assays, whose related details including assay principles, general protocols and other remarks are also given. All data were updated as of December 31, 2018. The “limits” search engine (<https://www.ncbi.nlm.nih.gov/pcassay/limits>) was used to filter the PubChem BioAssay resource by various options, with “Activity Outcome” set as “Active”, “Substance Type” set as “Chemical”, and “Screening Stage” defined as “Confirmatory, Dose-Response”. 149 assays, each targeting a single protein target, operated on at least 10,000 substances, and giving no fewer than 50 confirmed actives were first retained. The experimental screening data were kept if the target was characterized by at least one Protein Data Bank (PDB)³⁴ entry, in complex with a ligand of the same phenotype (i.e. inhibitor, agonist, or antagonist) as that of the tested active substances of the corresponding bioactivity assay. Altogether, 21 raw HTS data tables were directly retrieved as csv files from the PubChem BioAssay website along with actives and inactives in separate sd files. The PDB resource was then browsed by Uniprot identifiers (Uniprot IDs)³⁵ to retrieve the corresponding PDB entries in the suitable ligand-bound form.

2.2. Template Structure Preparations for Each Target Set

Protein-ligand complexes (in pdb file format) corresponding to the chosen target sets were processed as follows. For each PDB entry, explicit hydrogen atoms were added with Protoss³⁶ to any molecule (protein, cofactor, prosthetic group, ion, ligand, water). The output pdb file was then visualized in Sybyl-X 2.1.1.³⁷ A water molecule was kept under two conditions: (i) it was found at the binding site of the ligand, i.e., the distance between the oxygen atom of the water molecule and at least one heavy atom of the co-crystallized ligand was not greater than 5 Å; and (ii) it engaged in no fewer than three hydrogen bonds with the protein and/or the ligand, at least two of which were with the protein. Hydrogen bonds must satisfy the following criteria: the donor-acceptor distance must not exceed 3.5 Å; the angle formed by the donor, the hydrogen atom and the acceptor (with the vertex of the angle positioned at the hydrogen atom) must be

larger than 120 degrees. The protonated ligand and protein (including all remaining bound water molecules, cofactors, prosthetic groups and ions) were saved separately in mol2 file format with Sybyl-X 2.1.1.³⁷

In case more than 20 ligand-bound protein entries were available for each target, all protein-ligand structures were clustered according to the diversity of protein-ligand interaction patterns. These patterns were computed as graphs with IChem³⁸ as previously described,³⁹ and target-specific interaction pattern similarity matrices were computed using the GRIM score metric.³⁹ Each matrix was then used as input for agglomerative nesting clustering using the “agnes” function in R v.3.5.2, the Ward clustering method, a Euclidean distance matrix and a total number of clusters fixed to 15. For each cluster, the PDB entry with the highest resolution was chosen as the protein-ligand PDB template for the corresponding target set.

2.3. Determination of Filtering Rules for True Active and True Inactive Substances of Each Target Set

Metadata on every substance (true active and true inactive) constituting each selected target set were collected directly from the website of PubChem BioAssay, including: the substance identifier (SID), the activity label (active or inactive), the phenotype (inhibitor, agonist, or antagonist), the potency (EC₅₀ or IC₅₀, in μM), and the Hill slope for the dose-response curve of each true active. The frequency of hits (*FoH*) for a confirmed active molecule was computed as the ratio of the number of PubChem bioactivity assays in which the substance was identified as true active to the number of assays in which it was tested. Additional molecular properties (molecular weight, AlogP, total formal charge, number of rotatable bonds, number of hydrogen bond donors and acceptors) were computed in Pipeline Pilot v.19.1.0.1964.⁴⁰

For each target set, all true actives and true inactives were then filtered according to four steps:

- Step 1: Inorganic compounds filter. Molecules bearing at least one atom other than H, C, N, O, P, S, F, Cl, Br, and I were removed.
- Step 2: False positives filter (this particular step was applied only to true active substances).
 - Step 2a: Actives with Hill slope $h < 0.5$ or > 2.0 were discarded;⁷
 - Step 2b: Actives with frequency of hits $FoH > 0.26$ were removed;⁷

- Step 2c: Aggregation – auto-fluorescence – luciferase inhibition filter:⁷ actives identified as promiscuous aggregators (actives in AID 585 or AID 485341 but not in AID 584 and AID 485294), luciferase inhibitors (actives in AID 411), or compounds having auto-fluorescent properties (actives in AID 587, AID 588, AID 590, AID 591, AID 592, AID 593, AID 594) were eliminated.
- Step 3: Molecular property range filter. The remaining actives and inactives were kept if:
 - $150 < \text{molecular weight} < 800 \text{ Da}$;
 - $-3.0 < \text{AlogP} < 5.0$;
 - Number of rotatable bonds < 15 ;
 - H-bond acceptor count < 10 ;
 - H-bond donor count < 10 ;
 - $-2.0 < \text{total formal charge} < +2.0$.
- Step 4: 3D conversion and normalization filter. The two-dimensional (2D) sd files of the remaining compounds (actives, inactives) were converted into 3D sd file format using the default settings of Corina v.3.4.⁴¹ Last, compounds were standardized and ionized at physiological pH with Filter v.2.5.1.4.⁴² All preparation failures were discarded.

2.4. 2D Similarity Searches

Extended-connectivity circular ECFP4 fingerprints⁴³ were computed for PubChem compounds and PDB ligands in Pipeline Pilot v.19.1.0.1964.⁴⁰ Pairwise similarity of PubChem compounds to PDB ligands was estimated by the Tanimoto coefficient (Tc), thereby leading to a PDB ligand-specific hit list sorted by decreasing Tc values. The areas under the ROC (receiver operating characteristic)¹¹ and BEDROC (Boltzmann-enhanced discrimination of ROC)¹² curves ($\alpha = 20$) along with the enrichment in true actives at a constant 1% false positive rate over random picking (EF1%) were calculated for each separate hit list. The same procedure was applied by fusing all lists and keeping the maximal Tc value for each compound (the “max-pooling” approach).

2.5. 3D Similarity Searches

For each target set, a maximal number of 200 conformers were generated for every PubChem compound with the standard settings of Omega2 v.2.5.1.4.⁴⁴ All conformers were then compared

to the query (PDB ligand) with ROCS v.3.2.0.4.⁴⁵ The best matching conformer was selected for every ligand according to the TanimotoCombo similarity score,¹³ and all molecules of each target set were sorted based on this same value in descending order. Retrospective virtual screening performance was evaluated by ROC AUC, BEDROC AUC and EF1% values calculated as described above.

2.6. Molecular Docking

Starting from the mol2 structure of a fully processed template protein (including remaining bound water molecules after preparation) and that of its co-crystallized ligand, a protomol representing the ligand-binding site was generated from protein-bound ligand atomic coordinates using the default settings of Surflex-Dock v.3066.⁴⁶ All molecules in the relevant target set were docked into the protomol with the “-pgeom” option of the docking engine. The best-ranked pose according to docking scores (pK_d values) was retained for each molecule, and all ligands of the set were then sorted based on this value in descending order. Retrospective virtual screening performance was evaluated by ROC AUC, BEDROC AUC and EF1% values calculated as described above.

2.7. Target Set Unbiasing

For each target set, the unbiasing of the training and validation sets was done using the previously described asymmetric validation embedding (AVE) method,²⁵ which systematically measures pairwise distance in chemical space between molecules belonging to four sets of compounds (training actives, training inactives, validation actives, validation inactives). Using circular ECFP4 fingerprints⁴³ as chemical descriptors and a training-to-validation ratio of 3, a maximal number of 300 iteration steps of the AVE genetic algorithm (GA) were run to select training and validation molecules while minimizing the overall bias B ($B \in [0;1]$) of the target set. Convergence was reached when the bias value B was lower than 0.01, i.e., the GA was programmed to stop as soon as the total bias was below 0.01. To enable the script to process large sets of compounds (more than 100,000 molecules), the bias-removing script (*remove_AVE_bias.py*) originally proposed by Wallach and Heifets²⁵ was modified to allow a faster calculation on multiple cores.

3. Results and Discussion

The aim of the present study is to design an unbiased data set dedicated to virtual screening as well as machine learning, along four main ideas:

- (1) Experimental data should be available for all compounds, including the inactives. Each true active should have been confirmed by a full dose-response curve.
- (2) The target should be a single protein, for which a high-resolution X-ray structure is available on the PDB. Moreover, the target should have been crystallized at least once, with a ligand exhibiting a phenotype (e.g. inhibitor, full agonist, neutral antagonist) identical to that of active compounds in the corresponding bioassay.
- (3) PubChem target sets should be suitable for both ligand-based and structure-based virtual screening. The performance of three orthogonal methods (2D fingerprint similarity searches, 3D shape similarity searches, molecular docking) was evaluated to select only the target sets for which at least one of these three methods achieves an EF1% value ≥ 2.0 , or in other words, performs at least twice better than random picking (EF1% = 1.0).
- (4) The finally selected target sets should be as unbiased as possible, when it comes to comparing true actives to true inactives in chemical space, and when the data are split into training and validation sets.

To this end, we designed a computational workflow (**Figure 1**) that will be presented and discussed, step-by-step in the following sections.

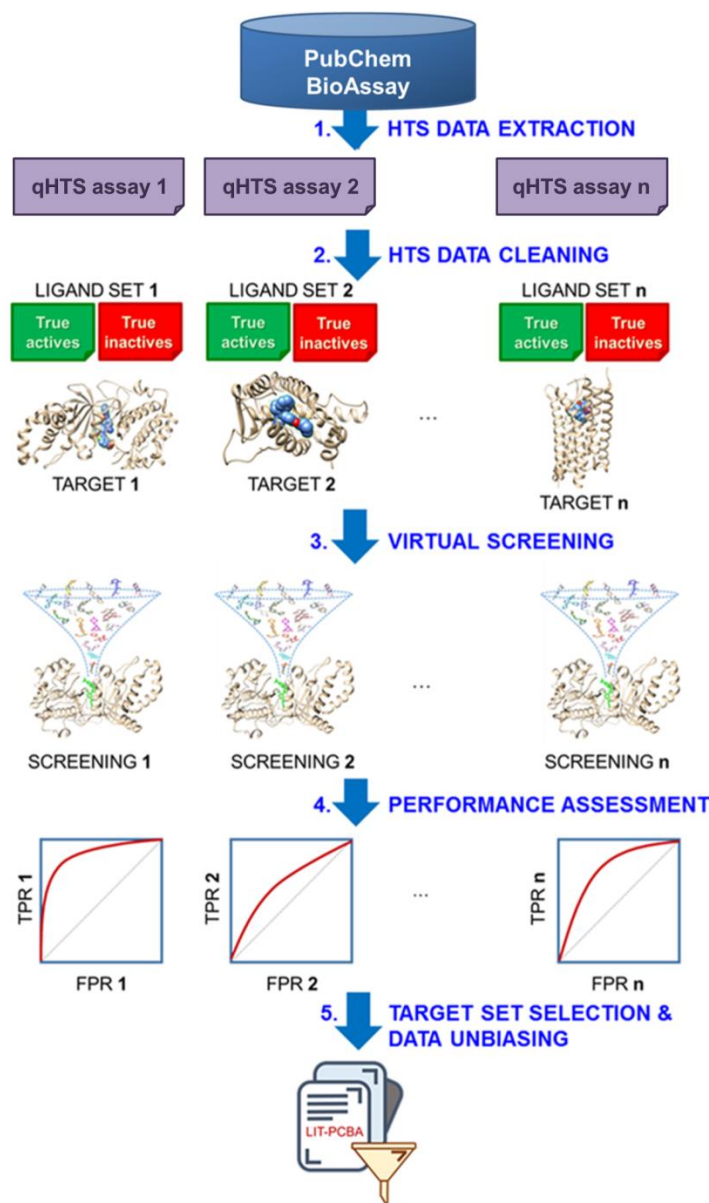


Figure 1. Workflow for LIT-PCBA data set construction. (1) Data retrieval from PubChem BioAssay according to user-defined filters (activity outcome: active; $\geq 10,000$ tested substances; ≥ 50 active substances; substance type: chemical; screening stage: confirmatory, dose-response; target: single; target type: protein target). (2) Data cleaning: removal of inorganic compounds, false positives, frequent hitters, assay artifacts and compounds with extreme molecular properties. Selection of target sets having at least a representative target structure on the Protein Data Bank co-crystallized with a ligand of the same phenotype as that of the actives in the corresponding bioassay. (3) Virtual screening of the cleaned HTS target sets with three methods (2D similarity, 3D similarity, docking). (4) Performance assessments of the methods on all cleaned target sets (ROC, BEDROC, EF1%). (5) Selection of target sets for which at least one method achieves an EF1% value ≥ 2.0 . AVE unbiassing²⁵ of the corresponding ligand sets and definition of training and validation sets for machine learning.

3.1. HTS Data Extraction

PubChem (<https://pubchem.ncbi.nlm.nih.gov>) is a public repository for information on 91 million chemical substances and 268 million biological activities, launched in 2004 as a component of the Molecular Libraries Roadmap Initiatives of the US National Institutes of Health (NIH). The PubChem BioAssay resource²² was queried to retrieve 149 assays according to multiple queries (see “Computational methods”). To ascertain that the data set will be further suitable for both ligand-based and structure-based virtual screening, we checked that each single protein target not only had a representative structure on the PDB, but was also co-crystallized with a ligand sharing the same phenotype or function with the true actives. This sanity check enables the selection of the right activation state (e.g. for G-protein coupled receptors) and the right binding site for docking. Of course, we cannot ensure at this step that all true actives share the same binding site with all PDB ligand templates. However, it serves as the first filter to avoid comparing ligands with known opposite or different functions. To control the bioactivity of each compound, only confirmatory dose-response screening assays were kept. A total of 21 assays (**Table 1**) performed on isolated enzymes ($n = 6$), soluble protein-protein interactions ($n = 4$) and target-expressing cells ($n = 11$); using four different readouts (fluorescence intensity, fluorescence polarization, luminescence, alpha screen) were finally saved. Except for five screens in which only 10,000 compounds were tested, most assays were run on a large number of compounds (from 200,000 to 400,000). Importantly, each assay was analyzed in detail, notably regarding the activity threshold qualifying a compound as active, which is target-dependent and was not further modified in this study. Compounds whose activity outcome was deemed as “inconclusive” were removed from the final ligand sets, only ligands confirmed as either actives or inactives were retained.

Corresponding targets are single proteins representing 11 families of pharmaceutical interest, including nuclear hormone receptors ($n = 5$), protein kinases ($n = 3$), and G protein-coupled receptors ($n = 3$). Most target sets describe compounds tested for an inhibitory activity against a protein target (13 target sets). Overall, 162 structures of protein-ligand complexes in PDB format were chosen as templates for the 21 target sets (**Table 1**). More information on each selected PubChem bioactivity assay (brief assay description, readout, format, PDB templates) can be found in **Table S1**.

Table 1. List of 21 selected PubChem bioactivity assays

ID	Target	Assay	Tested	Substances ^b		PDB entries
	Name	AID ^a		Actives	Phenotype	
ADRB2	Beta2 adrenergic receptor	492947	331,108	80	Agonist	8
ALDH1	Aldehyde dehydrogenase 1	1030	220,402	16,117	Inhibitor	8
ARO1	Aromatase	743083	10,486	905	Inhibitor	3
ESR1-ago	Estrogen receptor alpha	743075	10,486	589	Agonist	15
ESR1-ant	Estrogen receptor alpha	743080	10,486	477	Antagonist	15
FEN1	Flap endonuclease 1	588795	391,275	1368	Inhibitor	1
GBA	Glucocerebrosidase	2101	326,770	299	Inhibitor	6
GLP1R	Glucagon-like peptide-1 receptor	624417	408,352	6432	Inverse agonist	2
GLS	Glutaminase	624170	409,400	846	Inhibitor	11
IDH1	Isocitrate dehydrogenase	602179	390,606	365	Inhibitor	14
KAT2A	Histone acetyltransferase KAT2A	504327	387,485	817	Inhibitor	3
L3MBTL1	Lethal(3)malignant brain tumor-like protein isoform I	485360	225,505	1495	Inhibitor	1
MAPK1	Mitogen-activated protein kinase 1	995	72,004	711	Inhibitor	15
MTORC1	Mechanistic target of rapamycin	493208	43,989	342	Inhibitor	11
OPRK1	Kappa opioid receptor	1777	284,220	51	Agonist	1
PKM2	Pyruvate kinase muscle isoform 2	1631	264,516	892	Agonist	9
PPARG	Peroxisome proliferator-activated receptor gamma	743094	10,486	78	Agonist	15
RORC	Retinoic acid-related orphan receptor gamma	2551	309,031	16,824	Inhibitor	15
THRB	Thyroid hormone receptor	1469	282,587	183	Inhibitor	1
TP53	Cellular tumor antigen p53	651631	10,488	602	Agonist	6
VDR	Vitamin D receptor	504847	401,452	3735	Antagonist	2

^a Full details for each assay are available at <https://pubchem.ncbi.nlm.nih.gov/bioassay/AID>.

^b Structures deposited by individual data contributors. Unique chemical structures are called “compounds”.

3.2. HTS Data Cleaning

All active and inactive compounds were next submitted to a series of filters (see “Computational methods”) aimed at removing inorganic compounds (step 1), frequent hitters and assay artifacts (step 2),⁷ compounds exhibiting molecular properties outside pre-defined ranges (step 3), and molecules for which either 2D-to-3D conversion or ionization at pH 7.4 failed (step 4). It can be observed that nearly 60% of true active substances were removed during the filtering steps (see **Table S2, S3** for exhaustive statistics), with step 2a eliminating the most true actives (**Figure 2**). This step is aimed at ruling out actives that exhibit very strong binding cooperativity and have multiple binding sites.⁴⁷ True inactive substances, on the other hand, were not subjected to the three filtering steps 2a, 2b and 2c, thus lost much fewer members than the true actives, with over 90% of substances still remaining in the end.

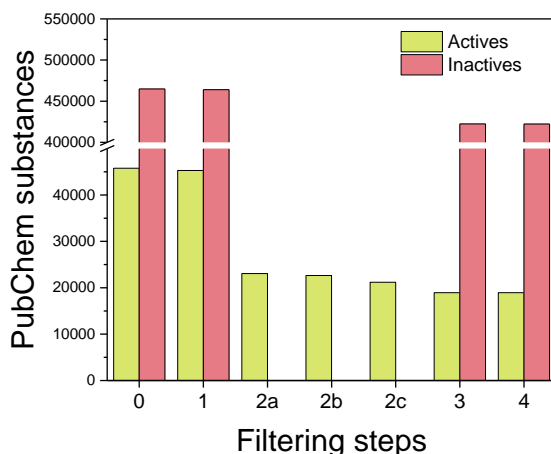


Figure 2. Total number of actives and inactives remaining after each filtering step was applied to the 21 selected target sets from PubChem BioAssay. Step 1: inorganic molecules; Step 2a: compounds with Hill slope $h < 0.5$ or > 2 ; Step 2b: compounds with frequency of hits $FoH > 0.26$; Step 2c: assay artifacts interfering with the readouts (10,892 substances classified as aggregators or auto-fluorescent molecules or luciferase inhibitors); Step 3: compounds with extreme molecular properties; Step 4: 3D conversion and ionization failures. Steps 2a, 2b and 2c were not applied to true inactives.

The filtering steps highlight the importance of removing assay artifacts in the composition of active substances. These steps not only prevented false positives that could affect subsequent screening performance, but also significantly reduced the number of true actives in comparison to that of true inactives, thus bringing hit rates closer to those typically observed in experimental

screening decks,³² but lower (in 15 out of 21 cases) than those of artificially constructed data sets commonly used by the cheminformatics community (**Figure 3A**).

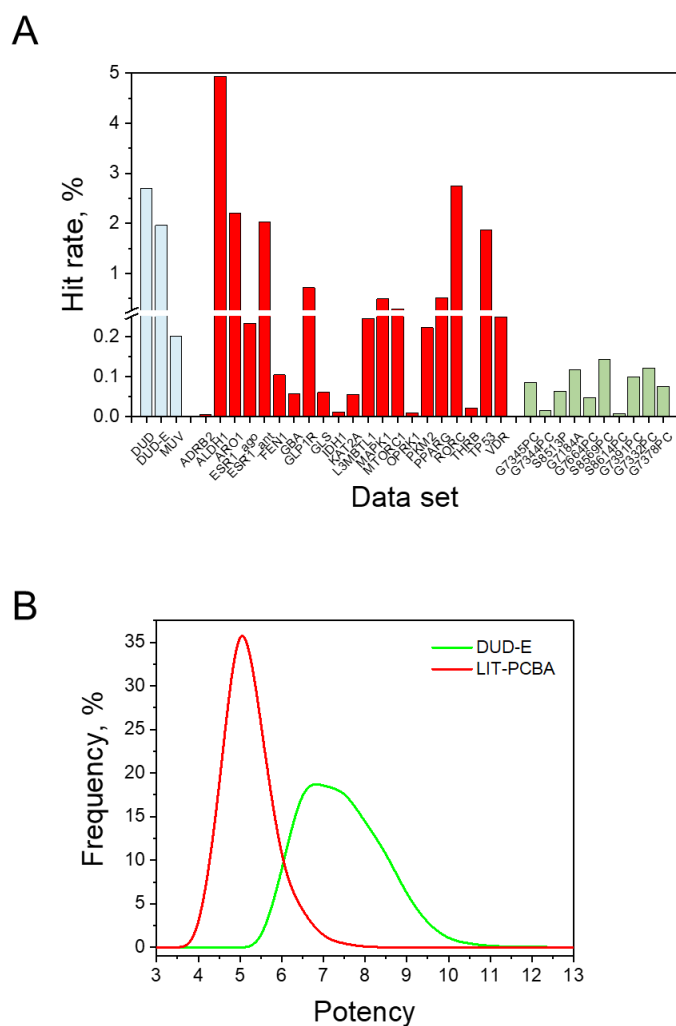


Figure 3. Properties of LIT-PCBA and standard data sets. **(A)** Confirmed hit rates for the LIT-PCBA data set (red bars), standard cheminformatics data sets (DUD,⁶ DUD-E,¹⁰ MUV;⁷ blue bars), and a representative sample of 10 high-throughput screens from a major pharmaceutical company (green).³² **(B)** Potency distribution of actives in the LIT-PCBA (red) and DUD-E (green) data sets. Potency is expressed as pIC₅₀, pEC₅₀, pK_i or pK_d.

We next looked at the potency distribution of true actives (**Figure 3B**) in our data set in comparison to that of DUD-E and ChEMBL.²⁰ We can observe different potency distribution for DUD-E actives ($n = 67,659$; median potency = 7.46 ± 0.96) and for LIT-PCBA actives ($n = 19,985$; median potency = 5.22 ± 0.54). The micromolar potency values observed for most LIT-PCBA actives reflect affinities typically observed in HTS campaigns. Conversely, DUD-E

actives tend to be much more potent (potency at the sub-micromolar level in most cases) and consequently easier to be picked, thereby overestimating the real benefit of virtual screening methods. At the individual target set level, the same trend applies when comparing the potency of LIT-PCBA and ChEMBL ligands for 19 common target sets (**Figure S1**). Importantly, we believe that the enhanced difficulty proposed by our data collection may enable a better discrimination of *in silico* screening methods.

3.3. Virtual Screening and Performance Assessments

The suitability of the 21 fully processed target sets for virtual screening was next assessed by three standard methods: 2D fingerprint similarity searches, 3D shape similarity searches, and molecular docking. The aim of the computational experiments was not to compare the virtual screening accuracy degrees of all methods but to check which of the 21 target sets may be unsuitable for *in silico* screening purposes. Hence, there is no guaranty that PubChem and PDB template ligands are strictly comparable (e.g. sharing the same binding site and molecular mechanism of action). Ligand-based screening will rapidly assess whether obvious bias is present in the ligand sets in terms of either 2D or 3D topologies. In addition, docking will ascertain if PubChem ligands share binding sites and interaction patterns with PDB templates. In each screen, all available PDB ligand/target templates were iteratively used as references, thereby generating as many hit lists as the available 162 templates. This exhaustive approach, albeit cumbersome, enables the selection of all references and takes into account the known chemical diversity of target-bound ligands (ligand-based virtual screening) or the known conformational space accessible to the target of interest (docking). In addition, a target-based “max-pooling” approach was followed by merging all screening data related to any LIT-PCBA ligand, whatever the corresponding template, and retaining the highest value (2D similarity, 3D similarity, docking score) per ligand. Statistical analyses of the data were primarily focused on enrichment factors in true actives at a constant 1% false positive rate (EF1%, **Figure 4**) as such values mirror the expectation of prospective virtual screening practices. Besides, areas under the ROC and BEDROC curves have also been calculated and are given in **Tables S4-S6**.

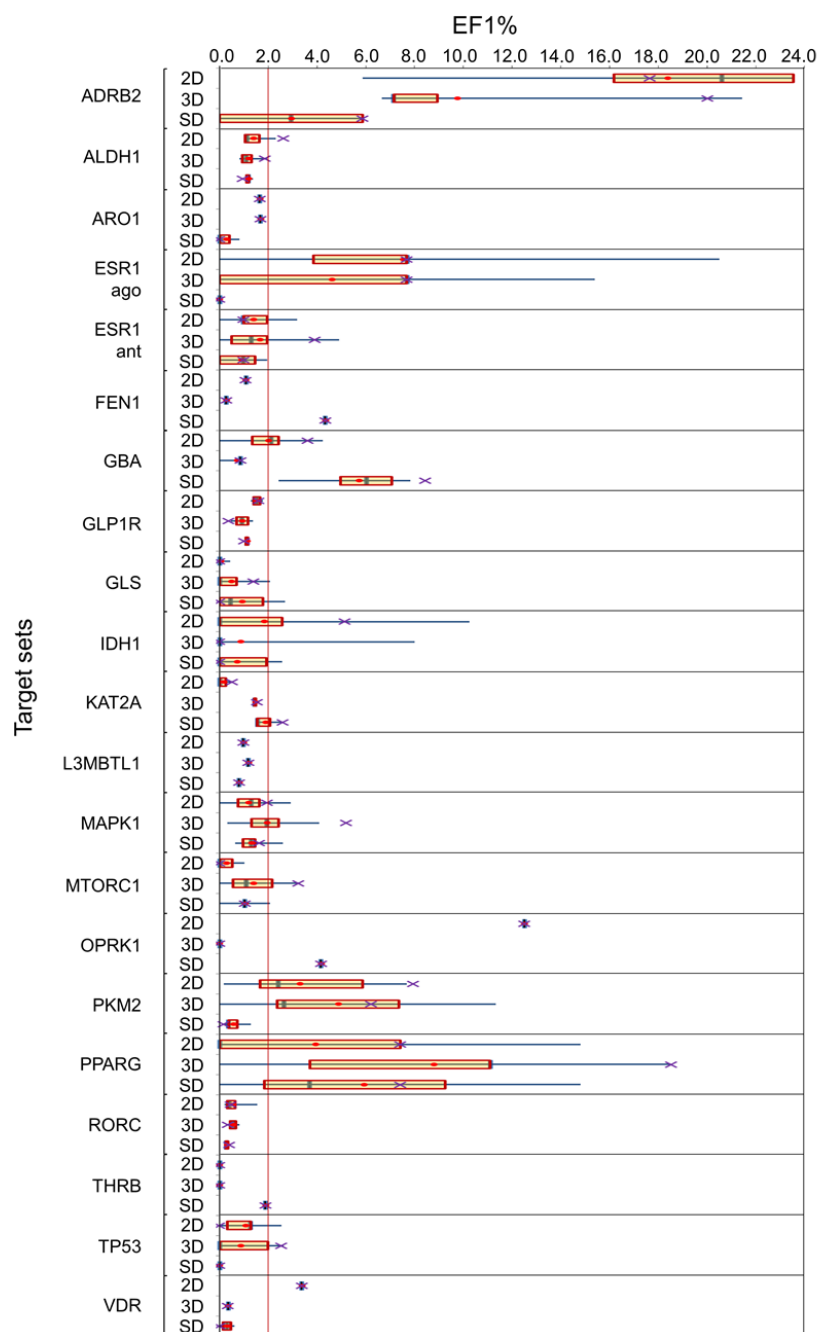


Figure 4. Performance of three different virtual screening methods (2D: ECFP4 similarity, 3D: 3D shape similarity, SD: molecular docking with Surflex-Dock) on 21 fully processed target sets. The graphs represent the distribution of EF1% values (enrichment in true actives at a constant 1% false positive rate over random picking) obtained after screening. The boxes delimit the 1st and 3rd quartiles, and the whiskers delimit the minimum and the maximum values. The median and the mean values are indicated by a green vertical line and a red dot located in each box, respectively. In cases where there is only one PDB template for a target set, or all templates gave the same EF1% value, the boxes are shrunk down into a single line. The purple crosses represent the EF1% values obtained by the “max-pooling” approach.

As expected, inspection of the observed enrichment in true actives for all 21 target sets clearly shows that the EF1% values may vary quite significantly according to the chosen template. In many instances, enrichment close to or even poorer than that obtained by random picking (EF1% = 1.0) is observed (**Figure 4**). We considered as acceptable any virtual screening protocol yielding an EF1% value ≥ 2 , or in other words, at least twice better than random picking. At this threshold, ligand-based methods clearly outperformed docking (**Figure 4**). Interestingly, only 10% of all *in silico* screening assays led to enrichment higher than 10. This result highlights the particular challenge of screening the current data set that we attribute to two main reasons: (i) the apparent absence of obvious bias in the distribution of PubChem actives in chemical space, and (ii) the potency distribution of PubChem actives not centered on sub-micromolar values.

3.4. Final Target Set Selection and Unbiasing

In order to facilitate the analysis, we will from now on discuss the results obtained by fusing, for each virtual screening method, all data across all available target-specific templates (“max-pooling” approach). This strategy was supported by two main reasons: (i) the fused approach provides enrichment values usually close to that obtained with the best possible template (**Figure 4**), and (ii) it enables the definition of a single hit list for each screening run while considering all templates. 15 out of the initial 21 target sets can be considered suitable (EF1% ≥ 2.0) for at least one of the three methods (**Figure 5**).

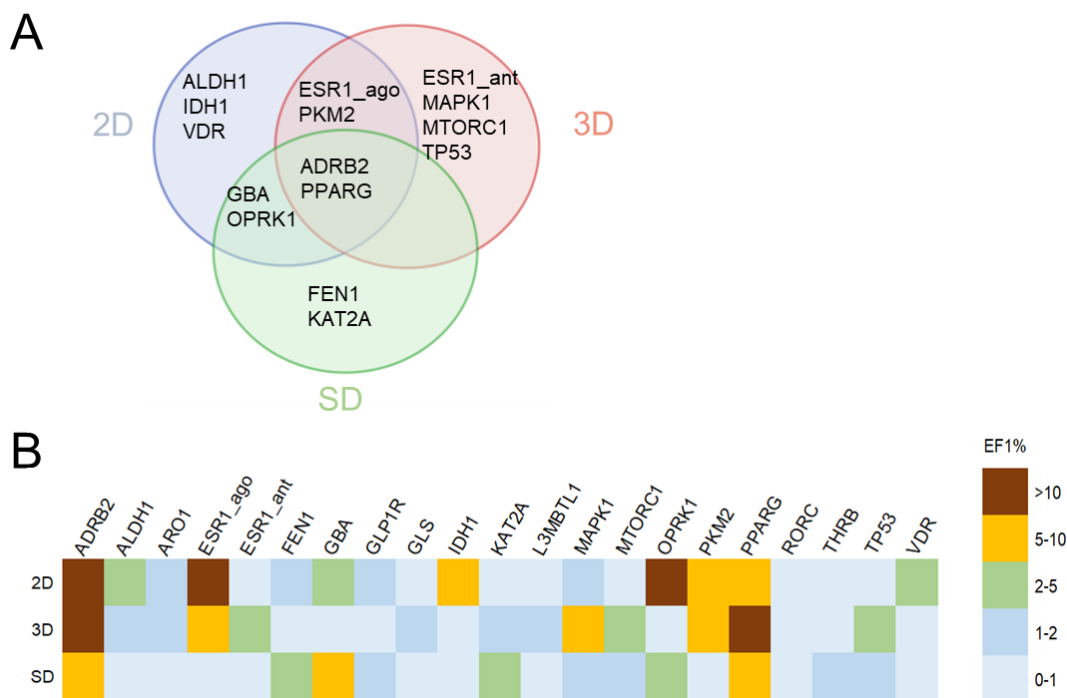


Figure 5. Comparative performance of three virtual screening protocols (2D: ECFP4 similarity searches, 3D: shape similarity searches, SD: molecular docking with Surflex-Dock) for the 21 target sets, processed by the “max-pooling” approach. **(A)** Venn diagram of target sets for which an EF1% ≥ 2.0 is observed. **(B)** Heatmap representing the fused values of EF1% obtained for each of the 21 fully processed target sets by the three *in silico* screening methods. Abbreviations of target sets are indicated above the heat map.

The current virtual screening exercise suggests that six target sets (GLS, GLP1R, ARO1, THRB, RORC, L3MBTL1) are not adequate for *in silico* screening purposes since none of the three methods was able to clearly distinguish the confirmed actives from inactive compounds when the “max-pooling” approach was applied (EF1% < 2.0) (**Figures 4-5**). Moreover, for five target sets among them (GLS as the only exception), the template-based scoring approach did not give any EF1% value above 2.0 either. Reasons for failures in screening these targets were: (i) the promiscuity of the binding site towards many low-affinity chemotypes (e.g. ARO1), (ii) the presence of non-overlapping binding sites (orthosteric versus allosteric) for PDB templates and PubChem actives (e.g. GLP1R, GLS, RORC), and (iii) the availability of a single PDB template (e.g. L3MBTL1, THRB).

Two target sets (ADRB2, PPARG) seem easier to handle since all three virtual screening methods could successfully retrieve true actives with enrichment factors higher than 5.0. In four

cases (GBA, OPRK1, PKM2, ESR1-ago), two methods succeeded. Last, only one method was able to perform correctly for 9 sets (ALDH, IDH1, VDR, MTORC1, MAPK1, ESR1-ant, TP53, FEN1, KAT2A; **Figure 5**). This result is in agreement with many previous studies⁴⁸⁻⁵⁰ suggesting that *in silico* screening methodologies are orthogonal, and is reassuring as it highlights the absence of obvious bias in both 2D molecular graphs and 3D shapes of LIT-PCBA compounds. It can therefore be implied that the remaining true actives (besides the ADRB2 and PPARG sets) do not resemble their corresponding PDB template ligands in both 2D and 3D shapes; meaning similarities between them, if there were any, did not significantly contribute to improving virtual screening performance, notably in early enrichment of true actives.

For each of the remaining 15 target sets, we ensured that the chemical diversity of PDB template ligands was not biasing our analysis. A first comparison of the number of Bemis-Murcko frameworks⁵¹ to the total number of templates indicates that a wide variety of chemotypes are indeed available among the chosen PDB template ligands (**Table S7**). A self-similarity plot of templates (Tanimoto coefficient on MDL public keys) confirms this observation and shows, for most of the target sets (MTORC1 being an exception), a large chemical diversity (**Figure S2**).

The 15 target sets were last unbiased by the AVE method²⁵ to propose optimal training and validation sets for machine learning applications. In brief, a genetic algorithm (GA) is used to select four subsets of active and inactive compounds for training and validation sets, based on pairwise distances in chemical space (ECFP4 circular fingerprints) between the above-described four ligand subsets. The objective function of the GA (bias value) gears the splitting procedure to select training and validation sets for which distances in chemical space are homogeneously distributed when training actives, validation actives, training inactives and validation inactives are compared. For 14 out of 15 target sets, just a few iterations (< 100) of the GA were necessary to unbiased the corresponding target sets with low bias values (**Table 2**). Interestingly, optimal splitting was achieved without removing a single compound from 13 out of the 15 initial PubChem compound collections, thereby suggesting that the latter input did not exhibit major bias. The final AVE-unbiased LIT-PCBA data set covers 15 target sets, 7844 unique actives and 407,381 unique inactives (**Table 2**).

For two target sets (ALDH1, VDR), the high number of true actives forced us to reduce by 25% the size of the data set in order to reach GA search completion. In both cases, care was taken to

keep the hit rates unchanged after data reduction. A one-nearest neighbor (knn1) binary classification of the 15 validation sets, still using ECFP4 fingerprints as descriptors, led to areas under the ROC curves close to random (0.50) and thereby supports the *bona fide* unbiasing of all corresponding target sets. Analyses of the three baseline virtual screening experiments for the AVE validation sets only (**Table S8**) confirm the very challenging nature of the data set as the performance dropped drastically for many target sets, notably those with low numbers of actives (e.g. ADRB2, IDH1) or few PDB template ligands (e.g. OPRK1). As previously indicated, the baseline *in silico* screening protocol was just intended to remove PubChem HTS data unsuitable for virtual screening applications, and is not indicative of the performance of modern machine learning approaches. We however recommend the application of such methods to target sets exhibiting enough true actives to train on (ALDH1, FEN1, GBA, KAT2A, MAPK1, PKM2, VDR; **Table 2**).

Table 2. Final list of 15 target sets of the LIT-PCBA data collection

Target	Target name	AVE		Actives		Inactives		Knn1 ^a ROC AUC
		Bias	Iterations	Validation	Training	Validation	Training	
ADRB2	Beta2 adrenergic receptor	0.003	2	4	13	78,120	234,363	0.500
ALDH1 ^b	Aldehyde dehydrogenase 1	0.092	195	1344	4032	25,868	77,606	0.556
ESR1-ago	Estrogen receptor alpha	0.001	1	3	10	1395	4188	0.499
ESR1-ant	Estrogen receptor alpha	0.006	9	25	77	1237	3711	0.517
FEN1	Flap endonuclease 1	0.076	39	92	277	88,850	266,552	0.499
GBA	Glucocerebrosidase	0.005	9	41	125	74,013	222,039	0.524
IDH1	Isocitrate dehydrogenase	0.001	4	9	30	90,512	271,537	0.500
KAT2A	Histone acetyltransferase KAT2A	0.001	5	48	146	87,137	261,411	0.500
MAPK1	Mitogen-activated protein kinase 1	0.000	8	77	231	15,657	46,972	0.505
MTORC1	Mechanistic target of rapamycin	0.001	7	24	73	8243	24,729	0.499
OPRK1	Kappa opioid receptor	0.000	3	6	18	67,454	202,362	0.500
PKM2	Pyruvate kinase muscle isoform 2	0.009	28	136	410	61,380	184,143	0.507
PPARG	Peroxisome proliferator-activated receptor γ	0.000	4	6	21	1302	3909	0.500
TP53	Cellular tumor antigen p53	0.008	29	19	60	1042	3126	0.491
VDR ^b	Vitamin D receptor	0.044	62	165	498	66,635	199,906	0.499

^a Area under the ROC curve for a binary classification of validation compounds (active, inactive) based on a one-nearest neighbor similarity search (ECFP4 fingerprints) model trained on target-specific training sets.

^b The size of the target set was reduced by 25% at the unbiasing stage due to the large number of remaining true actives.

4. Conclusion

A rigorous ligand set preparation process is necessary to benchmark virtual screening and/or machine learning methods. Since the body of known experimental data is continuously increasing, such benchmarking data sets need periodical revisions to remove both obvious and hidden bias inherent in human decision-making. Otherwise, errors are propagated across the literature and prevent a true comparison of novel methodological developments. Several recent reports²⁴⁻²⁷ unambiguously demonstrated that the cheminformatics community is currently facing this situation, leading notably to overoptimistic reports on the real benefit of artificial intelligence methods (e.g. deep neural networks) when applied to structure-based ligand design. We herewith present LIT-PCBA as a novel generation of virtual screening benchmarking data sets, specifically designed to reveal the true potential of computational methods in *in silico* screening exercises. The data collection has been designed from dose-response PubChem bioactivity assays for which active and inactive compounds are unambiguously defined. Importantly, a careful examination of metadata allowed the removal of assay artifacts, frequent hitters and false positives. LIT-PCBA comprises 15 target sets covering a wide diversity of ligands and target proteins. Preliminary virtual screening attempts with state-of-the-art methods (2D similarity searches, 3D shape-matching, and molecular docking) suggest that the data set is very challenging, notably because potency distribution bias among the labeled active compounds is no longer present. A recently described unbiasing procedure²⁵ was finally applied to LIT-PCBA to enable a rational and optimal distribution of training and validation sets for machine learning. We do believe that the particular challenge brought by this data collection will allow a clearer appreciation of modern artificial intelligence methods in structure-based virtual screening scenarios. The full LIT-PCBA data set is now freely accessible for download at <http://drugdesign.unistra.fr/LIT-PCBA>.

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. Wingert, B. M.; Camacho, C. J. Improving Small Molecule Virtual Screening Strategies for the Next Generation of Therapeutics. *Curr. Opin. Chem. Biol.* **2018**, *44*, 87-92.
3. Perez-Sianes, J.; Perez-Sanchez, H.; Diaz, F. Virtual Screening Meets Deep Learning. *Curr. Comput. Aided Drug Des.* **2019**, *15*, 6-28.

4. Gimeno, A.; Ojeda-Montes, M. J.; Tomas-Hernandez, S.; Cereto-Massague, A.; Beltran-Debon, R.; Mulero, M.; Pujadas, G.; Garcia-Vallve, S. The Light and Dark Sides of Virtual Screening: What Is There to Know? *Int. J. Mol. Sci.* **2019**, *20*, 1375.
5. Bissantz, C.; Folkers, G.; Rognan, D. Protein-Based Virtual Screening of Chemical Databases. 1. Evaluation of Different Docking/Scoring Combinations. *J. Med. Chem.* **2000**, *43*, 4759-4767.
6. Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, *49*, 6789-6801.
7. Rohrer, S. G.; Baumann, K. Maximum Unbiased Validation (MUV) Data Sets for Virtual Screening Based on PubChem Bioactivity Data. *J. Chem. Inf. Model.* **2009**, *49*, 169-184.
8. Vogel, S. M.; Bauer, M. R.; Boeckler, F. M. DEKOIS: Demanding Evaluation Kits for Objective In Silico Screening – A Versatile Tool for Benchmarking Docking Programs and Scoring Functions. *J. Chem. Inf. Model.* **2011**, *51*, 2650-2665.
9. Wallach, I.; Lilien, R. Virtual Decoy Sets for Molecular Docking Benchmarks. *J. Chem. Inf. Model.* **2011**, *51*, 196-202.
10. 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.
11. Triballeau, N.; Acher, F.; Brabet, I.; Pin, J. P.; Bertrand, H. O. Virtual Screening Workflow Development Guided by the “Receiver Operating Characteristic” Curve Approach. Application to High-Throughput Docking on Metabotropic Glutamate Receptor Subtype 4. *J. Med. Chem.* **2005**, *48*, 2534-2547.
12. Truchon, J. F.; Bayly, C. I. Evaluating Virtual Screening Methods: Good and Bad Metrics for the “Early Recognition” Problem. *J. Chem. Inf. Model.* **2007**, *47*, 488-508.
13. Empereur-Mot, C.; Guillemain, H.; Latouche, A.; Zagury, J. F.; Viallon, V.; Montes, M. Predictiveness Curves in Virtual Screening. *J. Cheminform.* **2015**, *7*, 52.
14. McGovern, S. L.; Shoichet, B. K. Information Decay in Molecular Docking Screens against Holo, Apo, and Modeled Conformations of Enzymes. *J. Med. Chem.* **2003**, *46*, 2895-2907.
15. Pham, T. A.; Jain, A. N. Parameter Estimation for Scoring Protein-Ligand Interactions Using Negative Training Data. *J. Med. Chem.* **2006**, *49*, 5856-5868.
16. Verdonk, M. L.; Berdini, V.; Hartshorn, M. J.; Mooij, W. T. M.; Murray, C. W.; Taylor, R. D.; Watson, P. Virtual Screening Using Protein-Ligand Docking: Avoiding Artificial Enrichment. *J. Chem. Inf. Comp. Sci.* **2004**, *44*, 793-806.
17. Irwin, J. J.; Shoichet, B. K. ZINC – A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 177-182.
18. Good, A. C.; Oprea, T. I. Optimization of CAMD Techniques 3. Virtual Screening Enrichment Studies: A Help or Hindrance in Tool Selection? *J. Comput. Aided Mol. Des.* **2008**, *22*, 169-178.
19. Hawkins, P. C.; Warren, G. L.; Skillman, A. G.; Nicholls, A. How to Do an Evaluation: Pitfalls and Traps. *J. Comput. Aided Mol. Des.* **2008**, *22*, 179-190.

20. 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-D1107.
21. Réau, M.; Langenfeld, L.; Zagury, J.-F.; Lagarde, N.; Montes, M. Decoys Selection in Benchmarking Datasets: Overview and Perspectives. *Front. Pharmacol.* **2018**, *9*, 11.
22. Wang, Y.; Bryant, S. H.; Cheng, T.; Wang, J.; Gindulyte, A.; Shoemaker, B. A.; Thiessen, P. A.; He, S.; Zhang, J. PubChem BioAssay: 2017 Update. *Nucleic Acids Res.* **2017**, *45*, D955-D963.
23. Tiikkainen, P.; Markt, P.; Wolber, G.; Kirchmair, J.; Distinto, S.; Poso, A.; Kallioniemi, O. Critical Comparison of Virtual Screening Methods against the MUV Data Set. *J. Chem. Inf. Model.* **2009**, *49*, 2168-2178.
24. Chaput, L.; Martinez-Sanz, J.; Saettel, N.; Mouawad, L. Benchmark of Four Popular Virtual Screening Programs: Construction of the Active/Decoy Dataset Remains a Major Determinant of Measured Performance. *J. Cheminform.* **2016**, *8*, 56.
25. Wallach, I.; Heifets, A. Most Ligand-Based Classification Benchmarks Reward Memorization Rather Than Generalization. *J. Chem. Inf. Model.* **2018**, *58*, 916-932.
26. Chen, L.; Cruz, A.; Ramsey, S.; Dickson, C. J.; Duca, J. S.; Hornak, V.; Koes, D. R.; Kurtzman, T. Hidden Bias in the Dud-E Dataset Leads to Misleading Performance of Deep Learning in Structure-Based Virtual Screening. *PLoS One* **2019**, *14*, e0220113.
27. Sieg, J.; Flachsenberg, F.; Rarey, M. In Need of Bias Control: Evaluating Chemical Data for Machine Learning in Structure-Based Virtual Screening. *J. Chem. Inf. Model.* **2019**, *59*, 947-961.
28. Yang, X.; Wang, Y.; Byrne, R.; Schneider, G.; Yang, S. Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery. *Chem. Rev.* **2019**, *119*, 10520-10594.
29. Wallach, I.; Dzamba, M.; Heifets, A. AtomNet: A Deep, Convolutional Neural Network for Bioactivity Prediction in Structure-Based Drug Discovery. *arXiv* **2015**, 1510.02855.
30. Pereira, J. C.; Caffarena, E. R.; Dos Santos, C. N. Boosting Docking-Based Virtual Screening with Deep Learning. *J. Chem. Inf. Model.* **2016**, *56*, 2495-2506.
31. Imrie, F.; Bradley, A. R.; van der Schaar, M.; Deane, C. M. Protein Family-Specific Models Using Deep Neural Networks and Transfer Learning Improve Virtual Screening and Highlight the Need for More Data. *J. Chem. Inf. Model.* **2018**, *58*, 2319-2330.
32. Posner, B. A.; Xi, H.; Mills, J. E. Enhanced HTS Hit Selection via a Local Hit Rate Analysis. *J. Chem. Inf. Model.* **2009**, *49*, 2202-2210.
33. Kim, S. Getting the Most out of PubChem for Virtual Screening. *Expert Opin. Drug Dis.* **2016**, *11*, 843-855.
34. 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.
35. The Uniprot Consortium. Uniprot: The Universal Protein Knowledgebase. *Nucleic Acids Res.* **2018**, *46*, 2699.
36. 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.
37. *Sybyl-X Molecular Modeling Software Packages, 2.1.1*; TRIPOS Associates, Inc.: St. Louis, MO, USA, 2013.

38. Da Silva, F.; Desaphy, J.; Rognan, D. IChem: A Versatile Toolkit for Detecting, Comparing, and Predicting Protein-Ligand Interactions. *ChemMedChem* **2018**, *13*, 507-510.
39. 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.
40. Dassault Systèmes, Biovia Corp., San Diego, CA 92121, USA.
41. Molecular Networks GmbH, Erlangen, Germany.
42. OpenEye Scientific Software, Santa Fe, NM 87508, USA.
43. Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742-754.
44. Hawkins, P. C.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T. Conformer Generation with Omega: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database. *J. Chem. Inf. Model.* **2010**, *50*, 572-584.
45. Hawkins, P. C.; Skillman, A. G.; Nicholls, A. Comparison of Shape-Matching and Docking as Virtual Screening Tools. *J. Med. Chem.* **2007**, *50*, 74-82.
46. Jain, A. N. Surflex-Dock 2.1: Robust Performance from Ligand Energetic Modeling, Ring Flexibility, and Knowledge-Based Search. *J. Comput. Aided Mol. Des.* **2007**, *21*, 281-306.
47. Prinz, H. Hill Coefficients, Dose-Response Curves and Allosteric Mechanisms. *J. Chem. Biol.* **2010**, *3*, 37-44.
48. Kruger, D. M.; Evers, A. Comparison of Structure- and Ligand-Based Virtual Screening Protocols Considering Hit List Complementarity and Enrichment Factors. *ChemMedChem* **2010**, *5*, 148-158.
49. Tian, S.; Sun, H. Y.; Li, Y. Y.; Pan, P. C.; Li, D.; Hou, T. J. Development and Evaluation of an Integrated Virtual Screening Strategy by Combining Molecular Docking and Pharmacophore Searching Based on Multiple Protein Structures. *J. Chem. Inf. Model.* **2013**, *53*, 2743-2756.
50. Tran-Nguyen, V. 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.
51. Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887-2893.

Supporting Information**Table S1.** Description of 21 selected PubChem bioactivity assays.

Target set	AID	Assay description	Readout	Format	PDB templates
ADRB2	492947	qHTS assay of beta-arrestin-biased ligands of beta2-adrenergic receptor	Lumi	CBA	3P0G, 3PDS, 3SN6, 4LDE, 4LDL, 4LDO, 4QKX, 6MXT
ALDH1	1030	qHTS assay for inhibitors of aldehyde dehydrogenase 1	Fluo	EAE	4WP7, 4WPN, 4X4L, 5AC2, 5L2M, 5L2N, 5L2O, 5TEI
ARO1	743083	qHTS assay to identify aromatase inhibitors	Fluo	CBA	3S7S, 4GL5, 4GL7
ESR1-ago	743075	qHTS assay to identify small molecule agonists of the estrogen receptor alpha (ER-alpha) signaling pathway	Fluo	CBA	1L2I, 2B1V, 2B1Z, 2P15, 2Q70, 2QR9, 2QSE, 2QZO, 4IVW, 4PPS, 5DRJ, 5DU5, 5DUE, 5DZI, 5E1C
ESR1-ant	743080	qHTS assay to identify small molecule antagonists of the estrogen receptor alpha (ER-alpha) signaling pathway using the BG1 cell line	Lumi	CBA	1XP1, 1XQC, 2YAR, 2IOG, 2IOK, 2OUZ, 2POG, 2R6W, 3DT3, 5AAU, 5FQV, 5T92, 5UFX, 6B0F, 6CHW
FEN1	588795	qHTS assay for the inhibitors of human flap endonuclease 1	Fluo	EAE	5FV7
GBA	2101	qHTS assay for inhibitors and activators of N370S glucocerebro-sidase as a potential chaperone treatment of Gaucher disease	Fluo	EAE	2V3D, 2V3E, 2XWD, 2XWE, 3RIK, 3RIL
GLP1R	624417	qHTS of GLP-1 receptor inverse agonists	Lumi	CBA	5VEW, 5VEX
GLS	624170	qHTS for inhibitors of glutaminase	Fluo ^a	EAE ^b	3UO9, 3VOZ, 3VP1, 5FI2, 5FI6, 5FI7, 5HL1, 5I94, 5JYO, 5WJ6, 5JYP
IDH1	602179	qHTS for inhibitors of mutant isocitrate dehydrogenase 1	Fluo	EAE	4I3K, 4I3L, 4UMX, 4XRX, 4XS3, 5DE1, 5L57, 5L58, 5LGE, 5SUN, 5SVF, 5TQH, 6ADG, 6B0Z, 5H84, 5H86, 5MLJ
KAT2A	504327	qHTS assay for inhibitors of GCN5L2	Fluo	PPI	3P8H
L3MBTL1	485360	qHTS assay for the inhibitors of L3MBTL1	Alpha	PPI	3P8H
MAPK1	995	qHTS assay for inhibitors of the ERK signaling pathway using a homogeneous screening assay	Alpha ^e	CBA	1PME, 2OJG, 3SA0, 3W55, 4QP3, 4QP4, 4QP9, 4QTA, 4QTE, 4WJ0, 4ZZN, 5AX3, 5BUJ, 5V62, 6G9H
MTROC1	493208	Acumen qHTS assay for inhibitors of the mTORC1 signaling pathway in MEF (Tsc2 ^{-/-} , p53 ^{-/-}) cells: Sytravon	Fluo	CBA	1FAP, 1NSG, 2FAP, 3FAP, 4DRH, 4DRI, 4DRJ, 4FAP, 4JSX, 4JT5, 5GPG
OPRK1	1777	uHTS identification of small molecule agonists of the kappa opioid receptor via a luminescent beta-arrestin assay	Lumi ^c	CBA ^d	6B73
PKM2	1631	qHTS assay for activators of human muscle isoform 2 pyruvate kinase	Lumi	EAE	3GQY, 3GR4, 3H6O, 3ME3, 3U2Z, 4G1N, 4JPG, 5X1V, 5X1W
PPARG	743094	qHTS assay to identify small molecule agonists of the peroxisome proliferator-activated receptor gamma (PPAR γ) signaling pathway	Fluo	CBA	1ZGY, 2I4J, 2P4Y, 2Q5S, 2YFE, 3B1M, 3HOD, 3R8A, 4CI5, 4FGY, 4PRG, 5TTO, 5TWO, 5Y2T, 5Z5S

RORC	2551	qHTS for inhibitors of ROR gamma transcriptional activity	Lumi	CBA	4WPQ, 4YMQ, 5APH, 5C4T, 5NTK, 5NTN, 5NTP, 5NTQ, 5NTW, 5UFR, 5VB6, 5X8Q, 6A22, 6B33, 6CVH
THRB	1469	qHTS for inhibitors of the interaction of thyroid hormone receptor and steroid receptor coregulator 2	FP ^f	PPI ^g	2PIN
TP53	651631	qHTS assay for small molecule agonists of the p53 signaling pathway	Fluo	CBA	2VUK, 3ZME, 4AGO, 4AGQ, 5G4O, 5O1I
VDR	504847	Inhibitors of the vitamin D receptor (VDR): qHTS	FP	PPI	3A2J, 3A2I

^a fluorescence intensity

^b enzyme activity assay

^c luminescence

^d cell-based assay

^e alpha screen

^f fluorescence polarization

^g soluble protein-protein interaction assay

Table S2. Number of remaining active compounds after each filtering step.

Target set	PubChem AID	Start	Filtering steps					
			Step 1	Step 2a	Step 2b	Step 2c	Step 3	Step4
ADRB2	492947	80	80	19	19	19	17	17
ALDH1	1030	16,117	16,070	8052	8023	7716	7170	7168
ARO1	743083	905	852	298	150	150	121	121
ESR1-ago	743075	105	89	20	18	18	15	13
ESR1-ant	743080	473	453	217	145	145	103	102
FEN1	588795	1368	1353	502	448	425	370	369
GBA	2101	299	298	240	236	233	166	166
GLP1R	624417	6432	6431	3000	2997	2942	2180	2180
GLS	624170	846	842	255	251	236	224	224
IDH1	602179	365	364	57	56	54	39	39
KAT2A	504327	817	794	297	268	234	194	194
L3MBTL1	485360	1495	1492	587	583	541	501	501
MAPK1	995	711	707	414	402	322	308	308
MTORC1	493208	342	342	137	136	136	97	97
OPRK1	1777	35	35	30	30	29	24	24
PKM2	1631	892	892	578	578	557	546	546
PPARG	743094	78	75	46	41	41	27	27
RORC	2551	16,824	16,805	8397	8355	8053	6874	6874
THR3	1469	183	179	92	78	64	53	53
TP53	651631	602	571	181	111	111	81	79
VDR	504847	3735	3685	1099	1067	1041	886	884
Unique compounds		45,771	45,294	23,058	22,653	21,819	18,939	18,930
% remaining		100.00	98.96	50.38	49.49	47.67	41.38	41.36

Table S3. Number of remaining inactive compounds after each filtering step.

Target set	PubChem AID	Start	Filtering steps			Final Actives/Inactives ratio
			Step 1	Step 3	Step 4	
ADRB2	492947	329,716	329,642	312,493	312,483	1/18,381
ALDH1	1030	148,322	148,166	137,980	137,965	1/19
ARO1	743083	8846	8661	5440	5381	1/44
ESR1-ago	743075	9089	8897	5640	5583	1/429
ESR1-ant	743080	8297	8121	5003	4948	1/49
FEN1	588795	382,244	382,117	355,420	355,402	1/963
GBA	2101	314,877	314,654	296,080	296,052	1/1783
GLP1R	624417	321,735	321,657	304,879	304,866	1/140
GLS	624170	401,810	401,672	371,883	371,860	1/1660
IDH1	602179	388,463	388,376	362,063	362,049	1/9283
KAT2A	504327	376,634	376,467	348,571	348,548	1/1797
L3MBTL1	485360	217,165	217,107	204,490	204,480	1/408
MAPK1	995	66,078	65,908	62,652	62,629	1/203
MTORC1	493208	41,294	41,294	32,972	32,972	1/340
OPRK1	1777	284,169	284,120	269,818	269,816	1/11,242
PKM2	1631	259,866	259,782	245,525	245,523	1/450
PPARG	743094	8532	8357	5267	5211	1/193
RORC	2551	256,777	256,580	243,311	243,284	1/35
THRB	1469	281,374	281,090	254,491	254,442	1/4801
TP53	651631	6973	6836	4215	4168	1/53
VDR	504847	384,189	383,989	355,415	355,388	1/402
Unique compounds		464,805	464,047	422,400	422,256	
% remaining		100.00	99.84	90.88	90.85	

Table S4. Virtual screening results obtained by 2D ECFP4 similarity searches on 21 fully processed selected target sets.

Target set	PubChem AID	ROC				BEDROC			
		Min	Max	Mean \pm SD	Fused	Min	Max	Mean \pm SD	Fused
ADRB2	492947	0.53	0.70	0.63 \pm 0.06	0.68	0.14	0.28	0.24 \pm 0.05	0.24
ALDH1	1030	0.49	0.52	0.51 \pm 0.01	0.52	0.07	0.11	0.09 \pm 0.01	0.11
ARO1	743083	0.50	0.52	0.51 \pm 0.01	0.52	0.06	0.06	0.06	0.06
ESR1-ago	743075	0.56	0.72	0.65 \pm 0.05	0.72	0.06	0.28	0.16 \pm 0.06	0.22
ESR1-ant	743080	0.42	0.54	0.50 \pm 0.03	0.50	0.02	0.09	0.06 \pm 0.02	0.04
FEN1	588795	0.44	0.44	0.44	0.44	0.04	0.04	0.04	0.04
GBA	2101	0.45	0.53	0.50 \pm 0.03	0.48	0.03	0.10	0.06 \pm 0.03	0.07
GLP1R	624417	0.48	0.50	0.49 \pm 0.01	0.50	0.06	0.07	0.07 \pm 0.01	0.07
GLS	624170	0.32	0.37	0.34 \pm 0.02	0.33	0.01	0.02	0.01	0.01
IDH1	602179	0.28	0.52	0.42 \pm 0.07	0.38	0.01	0.15	0.04 \pm 0.04	0.06
KAT2A	504327	0.36	0.37	0.40 \pm 0.06	0.44	0.03	0.06	0.04 \pm 0.02	0.04
L3MBTL1	485360	0.41	0.41	0.41	0.41	0.02	0.02	0.02	0.02
MAPK1	995	0.45	0.58	0.52 \pm 0.04	0.53	0.03	0.12	0.06 \pm 0.02	0.06
MTORC1	493208	0.47	0.52	0.48 \pm 0.02	0.45	0.03	0.05	0.04 \pm 0.01	0.04
OPRK1	1777	0.69	0.69	0.69	0.69	0.26	0.26	0.26	0.26
PKM2	1631	0.41	0.64	0.55 \pm 0.08	0.64	0.03	0.16	0.09 \pm 0.05	0.16
PPARG	743094	0.58	0.80	0.68 \pm 0.07	0.78	0.01	0.27	0.14 \pm 0.09	0.21
RORC	2551	0.36	0.56	0.43 \pm 0.05	0.44	0.02	0.10	0.04 \pm 0.02	0.04
THRB	1469	0.37	0.37	0.37	0.37	0.03	0.03	0.03	0.03
TP53	651631	0.38	0.56	0.47 \pm 0.06	0.42	0.03	0.06	0.05 \pm 0.01	0.03
VDR	504847	0.44	0.44	0.44	0.44	0.06	0.06	0.06	0.06
Overall		0.45	0.54	0.50 \pm 0.05	0.51	0.05	0.11	0.08 \pm 0.02	0.09

Table S5. Virtual screening results obtained by 3D shape similarity searches on 21 fully processed selected target sets.

Target set	PubChem AID	ROC				BEDROC			
		Min	Max	Mean \pm SD	Fused	Min	Max	Mean \pm SD	Fused
ADRB2	492947	0.47	0.66	0.53 \pm 0.06	0.67	0.08	0.24	0.14 \pm 0.05	0.20
ALDH1	1030	0.46	0.53	0.50 \pm 0.03	0.49	0.07	0.11	0.08 \pm 0.01	0.09
ARO1	743083	0.60	0.69	0.65 \pm 0.05	0.62	0.07	0.10	0.09 \pm 0.02	0.07
ESR1-ago	743075	0.46	0.65	0.56 \pm 0.05	0.65	0.03	0.22	0.12 \pm 0.05	0.15
ESR1-ant	743080	0.58	0.64	0.60 \pm 0.02	0.61	0.06	0.14	0.10 \pm 0.02	0.13
FEN1	588795	0.45	0.45	0.45	0.45	0.03	0.03	0.03	0.03
GBA	2101	0.33	0.40	0.38 \pm 0.03	0.34	0.02	0.05	0.03 \pm 0.01	0.03
GLP1R	624417	0.51	0.52	0.52 \pm 0.01	0.52	0.04	0.06	0.05 \pm 0.01	0.05
GLS	624170	0.37	0.45	0.40 \pm 0.03	0.44	0.01	0.04	0.02 \pm 0.01	0.04
IDH1	602179	0.35	0.50	0.41 \pm 0.05	0.39	0.00	0.09	0.03 \pm 0.02	0.02
KAT2A	504327	0.38	0.44	0.39 \pm 0.03	0.43	0.05	0.06	0.06 \pm 0.01	0.06
L3MBTL1	485360	0.50	0.50	0.50	0.50	0.04	0.04	0.04	0.04
MAPK1	995	0.45	0.62	0.53 \pm 0.05	0.55	0.03	0.13	0.08 \pm 0.03	0.11
MTORC1	493208	0.44	0.52	0.47 \pm 0.03	0.52	0.03	0.07	0.04 \pm 0.01	0.06
OPRK1	1777	0.55	0.55	0.55	0.55	0.03	0.03	0.03	0.03
PKM2	1631	0.48	0.67	0.60 \pm 0.07	0.59	0.02	0.20	0.12 \pm 0.07	0.15
PPARG	743094	0.59	0.76	0.72 \pm 0.05	0.73	0.04	0.30	0.20 \pm 0.06	0.30
RORC	2551	0.38	0.51	0.45 \pm 0.03	0.44	0.02	0.07	0.04 \pm 0.01	0.04
THRB	1469	0.57	0.57	0.57	0.57	0.07	0.07	0.07	0.07
TP53	651631	0.54	0.62	0.58 \pm 0.04	0.56	0.04	0.13	0.08 \pm 0.04	0.11
VDR	504847	0.37	0.37	0.37	0.37	0.02	0.02	0.02	0.02
Overall		0.47	0.55	0.51 \pm 0.03	0.52	0.04	0.10	0.07 \pm 0.03	0.09

Table S6. Virtual screening results obtained by molecular docking on 21 fully processed selected target sets.

Target set	PubChem AID	ROC				BEDROC			
		Min	Max	Mean \pm SD	Fused	Min	Max	Mean \pm SD	Fused
ADRB2	492947	0.41	0.52	0.46 \pm 0.04	0.44	0.03	0.08	0.06 \pm 0.02	0.09
ALDH1	1030	0.51	0.53	0.52 \pm 0.01	0.53	0.09	0.10	0.09	0.09
ARO1	743083	0.47	0.53	0.51 \pm 0.03	0.5	0.03	0.07	0.05 \pm 0.02	0.04
ESR1-ago	743075	0.26	0.51	0.36 \pm 0.07	0.48	0.00	0.05	0.01 \pm 0.02	0.03
ESR1-ant	743080	0.43	0.54	0.50 \pm 0.03	0.53	0.04	0.08	0.06 \pm 0.01	0.06
FEN1	588795	0.47	0.47	0.47	0.47	0.08	0.08	0.08	0.08
GBA	2101	0.48	0.72	0.64 \pm 0.08	0.69	0.09	0.21	0.16 \pm 0.04	0.18
GLP1R	624417	0.50	0.51	0.51 \pm 0.01	0.51	0.05	0.06	0.06 \pm 0.01	0.06
GLS	624170	0.34	0.43	0.38 \pm 0.03	0.35	0.02	0.06	0.04 \pm 0.01	0.02
IDH1	602179	0.30	0.48	0.37 \pm 0.05	0.38	0.00	0.09	0.04 \pm 0.03	0.04
KAT2A	504327	0.35	0.40	0.38 \pm 0.03	0.39	0.04	0.06	0.05 \pm 0.01	0.06
L3MBTL1	485360	0.45	0.45	0.45	0.45	0.04	0.04	0.04	0.04
MAPK1	995	0.48	0.55	0.52 \pm 0.02	0.54	0.04	0.07	0.06 \pm 0.01	0.07
MTORC1	493208	0.49	0.54	0.52 \pm 0.02	0.51	0.04	0.07	0.06 \pm 0.01	0.05
OPRK1	1777	0.58	0.58	0.58	0.58	0.09	0.09	0.09	0.09
PKM2	1631	0.49	0.56	0.54 \pm 0.02	0.62	0.04	0.05	0.05 \pm 0.01	0.05
PPARG	743094	0.65	0.74	0.69 \pm 0.02	0.71	0.08	0.24	0.16 \pm 0.04	0.18
RORC	2551	0.36	0.42	0.37 \pm 0.01	0.36	0.02	0.04	0.03 \pm 0.01	0.02
THRB	1469	0.36	0.36	0.36	0.36	0.04	0.04	0.04	0.04
TP53	651631	0.47	0.57	0.51 \pm 0.04	0.51	0.02	0.06	0.04 \pm 0.02	0.02
VDR	504847	0.34	0.36	0.35 \pm 0.01	0.34	0.01	0.01	0.01	0.01
Overall		0.44	0.51	0.48 \pm 0.02	0.49	0.04	0.08	0.06 \pm 0.02	0.06

Table S7. Chemical diversity of PDB template ligands assessed by the number of unique Bemis-Murcko frameworks.⁵¹

Target set	Number of PDB templates	Number of Bemis-Murcko scaffolds
ADRB2	8	5
ALDH1	8	8
ESR1-ago	15	14
ESR1-ant	15	15
FEN1	1	1
GBA	6	4
IDH1	14	14
KAT2A	3	2
MAPK1	15	15
MTORC1	11	5
OPRK1	1	1
PKM2	9	8
PPARG	15	15
TP53	6	5
VDR	2	1

Bemis-Murcko frameworks were computed from mol2 files, with the “Generate Fragments” component of Pipeline Pilot 2019.

Table S8. Virtual screening results (EF1%) obtained by 2D ECFP4 similarity searches, 3D shape similarity searches and molecular docking on 15 validation sets after debiasing with AVE.

Target set	PubChem AID	2D ECFP4 similarity searches				3D shape similarity searches				Molecular docking			
		Min	Max	Mean \pm SD	Fused	Min	Max	Mean \pm SD	Fused	Min	Max	Mean \pm SD	Fused
ADRB2	492947	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ALDH1	1030	0.82	2.75	1.58 \pm 0.62	2.68	0.67	1.64	1.08 \pm 0.35	1.64	0.89	1.56	1.25 \pm 0.23	0.82
ESR1-ago	743075	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ESR1-ant	743080	0.00	12.00	2.67 \pm 3.60	0.00	0.00	4.00	1.07 \pm 1.83	4.00	0.00	4.00	1.60 \pm 2.03	4.00
FEN1	588795	1.09	1.09	1.09	1.09	0.00	0.00	0.00	0.00	3.26	3.26	3.26	3.26
GBA	2101	0.00	2.44	1.63 \pm 1.26	2.44	0.00	4.88	0.81 \pm 1.99	0.00	0.00	9.76	4.47 \pm 3.59	4.88
IDH1	602179	0.00	11.11	1.59 \pm 4.03	0.00	0.00	11.11	0.79 \pm 2.97	0.00	0.00	11.11	0.79 \pm 2.97	0.00
KAT2A	504327	0.00	2.08	0.69 \pm 1.20	0.00	0.00	2.08	0.69 \pm 1.20	0.00	2.08	6.25	4.17 \pm 2.09	2.08
MAPK1	995	0.00	5.19	0.95 \pm 1.43	1.30	0.00	5.19	1.39 \pm 1.59	2.60	0.00	5.19	1.99 \pm 1.38	1.30
MTORC1	493208	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	4.17	1.52 \pm 2.10	4.17
OPRK1	1777	16.67	16.67	16.67	16.67	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
PKM2	1631	0.00	4.41	1.31 \pm 1.79	0.74	0.00	5.15	2.13 \pm 1.93	2.21	0.00	1.47	0.90 \pm 0.61	0.74
PPARG	743094	0.00	16.67	5.56 \pm 8.13	16.67	0.00	16.67	5.56 \pm 8.13	16.67	0.00	16.67	5.56 \pm 8.13	0.00
TP53	651631	0.00	0.00	0.00	0.00	0.00	5.26	0.88 \pm 2.15	0.00	0.00	0.00	0.00	0.00
VDR	504847	3.64	3.64	3.64	3.64	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Overall		1.48	5.20	2.49 \pm 1.47	3.02	0.04	3.73	0.96 \pm 1.48	1.81	0.42	4.23	1.70 \pm 1.54	1.42

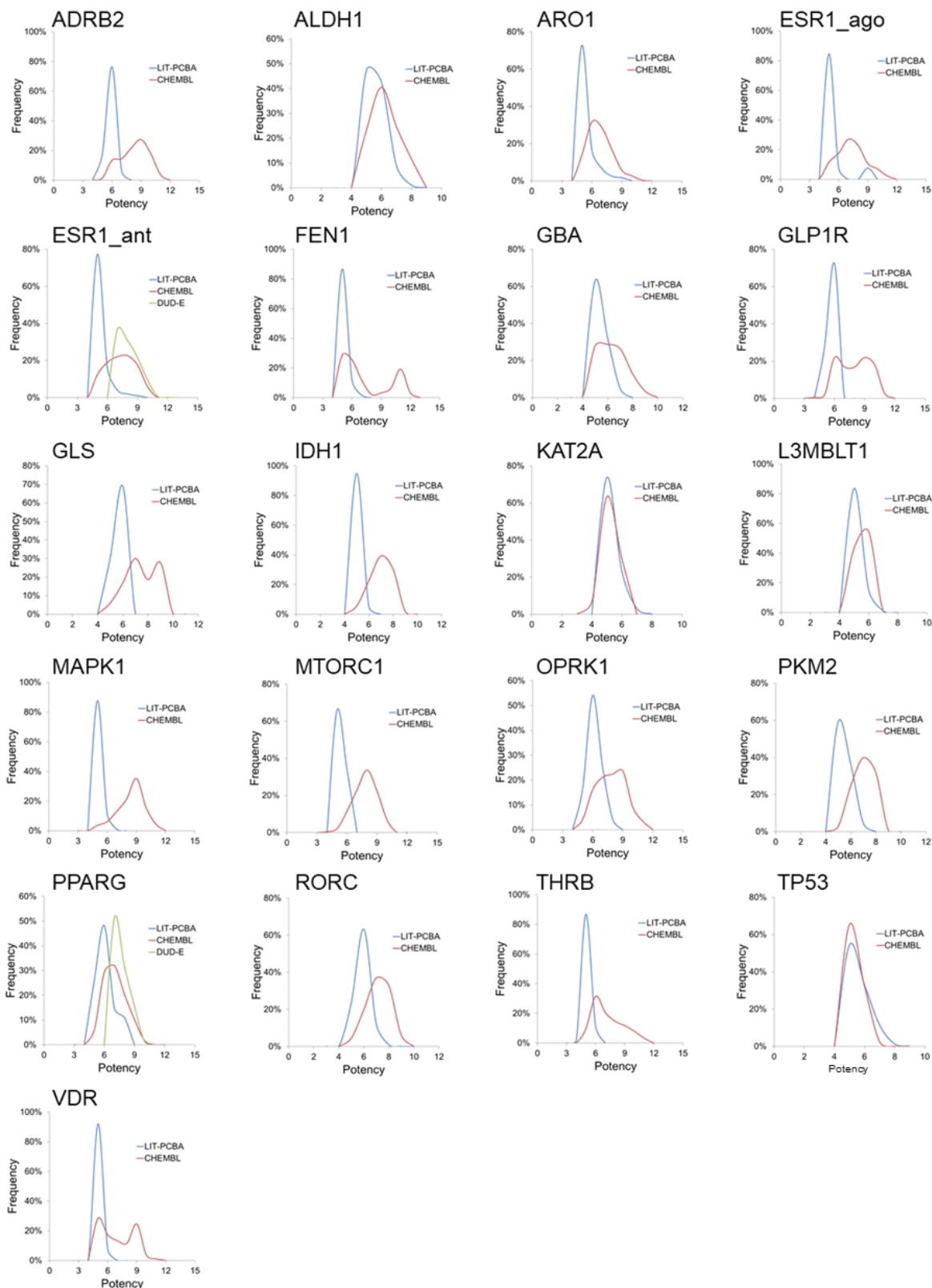


Figure S1. Comparison of potency values (in pIC₅₀, pEC₅₀, pK_i, pK_d) for confirmed actives of the LIT-PCBA, DUD-E and ChEMBL ligands.

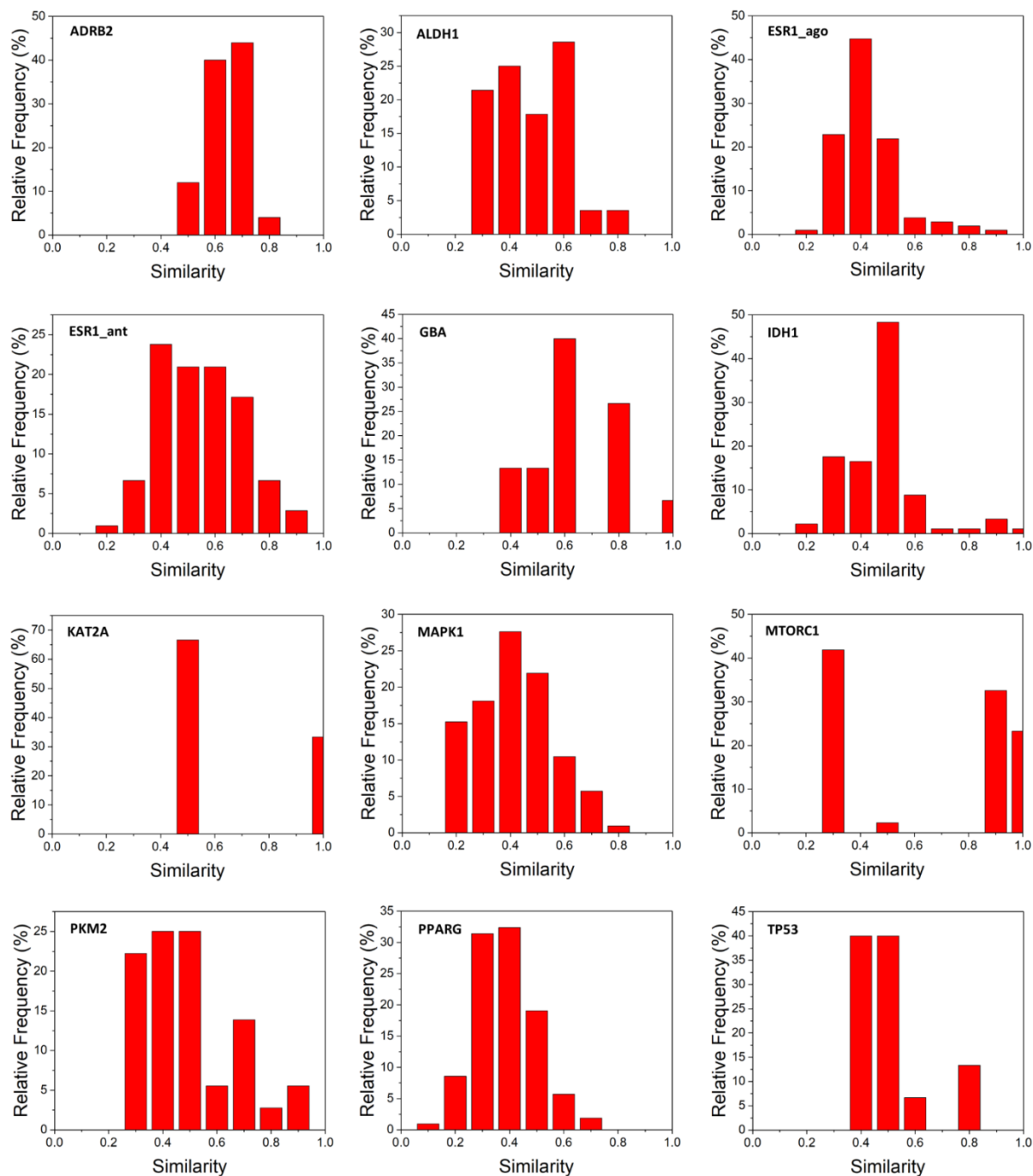


Figure S2. Self-similarity matrix of PDB template ligands. Pairwise similarity between template ligands is expressed by a Tanimoto coefficient calculated from MDL public keys implemented in Pipeline Pilot 2019.⁴⁰ No analysis is provided for three target sets (FEN1, OPRK1, VDR) for which a single PDB template ligand is available.

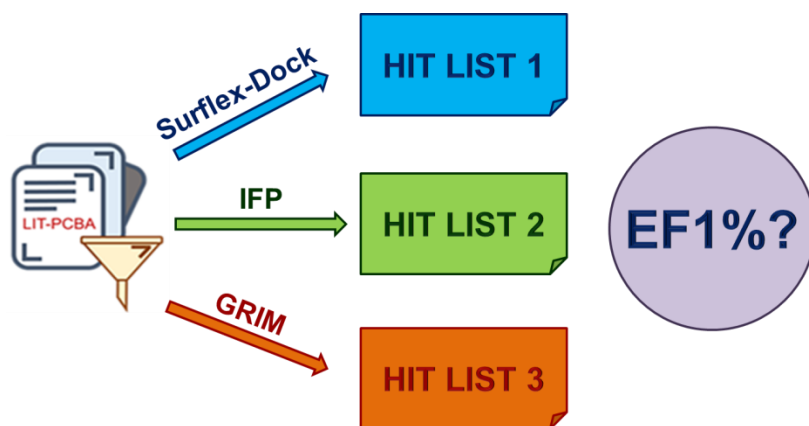
Take-home Messages

As explained in Chapter 1, LIT-PCBA marks the latest milestone in the quest to construct realistic benchmarking data sets for validating virtual screening methods entirely from experimental data. This data collection offers a pool of chemically unbiased ligands whose activity has been tested on a wide range of protein targets of pharmaceutical interest, presenting hit rates lower than those observed in most artificially constructed data sets and generally close to those of real-life high-throughput screening decks. Four subsets of ligands for each target were rationally designed, using a recently published method, to offer unbiased materials ready for evaluating both ligand-based and structure-based screening approaches, especially those relying on machine learning. Despite the existence of some limitations, e.g. the moderately high hit rates for several target sets or the relatively low number of remaining true actives in a few cases, the LIT-PCBA data collection does not suffer from serious drawbacks inherent in other benchmarking databases. More efforts in building novel data sets are recommended, with inspiration taken from the design of LIT-PCBA portrayed in this chapter and the good practices proposed in Chapter 1, in hopes of offering better evaluation tools for *in silico* screening methodologies.

Chapter 4

Rescoring LIT-PCBA Docking Poses with Interaction-Based Scoring Functions

From the results portrayed in Chapter 3, it can clearly be inferred that Surflex-Dock generally gave comparable performances to random selection on the LIT-PCBA data collection, suggesting that the energy-based empirical scoring function of this docking program was not highly effective in selecting true active molecules from a pool of chemically diverse and unbiased ligands. Several alternatives have been introduced in the literature, including two methods relying on the comparison of protein-ligand interaction fingerprints (IFP) and of interaction pattern graphs (GRIM) that were in-house developed by the researchers of our laboratory. They have both been proven more effective than popular docking programs in several virtual screening experiments, with encouraging results in terms of areas under the ROC curves and early enrichment of true actives. The questions are: will these approaches still give good performances when applied to the challenging LIT-PCBA data set, and will they once again outperform the Surflex-Dock scoring function on such a difficult data collection? This final chapter serves to answer the questions above.



1. Introduction

The scoring problem in molecular docking has been the subject of various studies aiming to select the correct pose (that matches experimentally determined output) for a ligand, and to ameliorate the screening utility as well as the scoring accuracy of a docker, i.e., to improve its ability to rank bioactive ligands above inactive ones in the hit list according to the calculated binding affinity.¹⁻⁵ Many approaches were designed to address this problem, defining a function composed of physical/chemical terms inherent in the process of protein-ligand binding, on the basis of existing complexes with known affinities and 3D structures.^{3,5-16} The energy-based scoring functions employed in several popular docking programs such as Surflex-Dock⁶ or FlexX^{11,17} rely on the empirical Bohm approach² that takes into account hydrophobic contacts and polar interactions that are formed between the involved molecules, along with the costs of entropic fixation due to torsional, translational and rotational freedom losses as the ligand and the protein are bound to each other.⁶ However, concerns have long been raised over the accuracy of such empirical methods to estimate the binding affinity of a small molecule with its macromolecular target, and the reliability of using data obtained from them for *in silico* screening purposes.^{3,18-20} Several alternatives to these scoring functions were developed, with the aim of rescoring the ensemble of poses generated by docking programs so that active molecules can be better ranked than inactive ligands, leading to an improvement in early enrichment of true actives. Among them are the two rescoring methods based on comparing ligand-protein interactions observed in a reference (e.g., a crystallographic structure found on the Protein Data Bank²¹) and those of a molecule's docking pose as issued by a docker.^{22,23}

The first method (IFP) relies on the similarity of protein-ligand interactions between a docking pose and any given template (e.g., the X-ray structure of the cognate protein with a known active molecule).²² In the first step, an interaction fingerprint for each docked ligand is generated as a fixed-length bitstring that registers the presence or the absence of non-covalent interactions between a set of user-defined protein residues (along with cofactors, ions and water molecules) and the ligand. Interaction fingerprints of the screened molecules are then compared to that of the template and are sorted by decreasing similarity as expressed by the Tanimoto coefficient.

The second method (GRIM) computes a graph whose nodes are interaction pseudoatoms which are placed on the ligand interacting atom, the protein interacting atom, and the barycenter of any

given protein-ligand interaction.²³ A clique detection algorithm is used to find the maximal common subgraph between the graph generated from the docking pose and that from the template.²³ In comparison to interaction fingerprints, interaction pattern graphs are not restricted to a fixed list of binding site atoms such that pairwise comparisons are also possible for binding cavities of different sizes.

The two aforementioned methods have been proven effective in predicting the binding modes of various ligands before the release of experimental crystallographic structures in international docking competitions, and in screening large pools of chemically diverse molecules, giving even better performances than popular docking algorithms.²²⁻²⁵ In this final chapter, these two methods are applied to the 15 target sets of the LIT-PCBA data collection, on which the energy-based scoring function of Surflex-Dock only managed to give comparable performances to random selection,²⁶ in order to assess the discriminatory power of such methods when a challenging set of different ligands from various biological targets is employed, allowing a comparison between their accuracy levels and that of Surflex-Dock.

2. Computational Methods

2.1. Rescoring LIT-PCBA Docking Poses by Protein-Ligand Interaction Fingerprint (IFP) Similarity

The IFP module²² of the IChem package²⁷ was employed to compute the similarity between the IFP recorded for each docked ligand from LIT-PCBA and that of the corresponding reference ligand, expressed by a Tanimoto coefficient (Tc) as the final output. The mol2 structures of the binding site and the reference (already prepared during the LIT-PCBA data set construction), along with the multi-mol2 files containing the docking poses issued by Surflex-Dock were used as input. The binding site refers to amino acid residues (plus water molecules, ions and cofactors) of the protein having at least one heavy atom within 5.0 Å from any heavy atom of the co-crystallized ligand (preparation was done with Sybyl-X 2.1.1²⁸). All docking poses were rescored and the pose with the highest Tc value was retained for each LIT-PCBA ligand, giving template-specific hit lists in which all ligands were sorted by decreasing Tc scores. The areas under the ROC (receiver operating characteristic)²⁹ and BEDROC (Boltzmann-enhanced discrimination of ROC)³⁰ curves (ROC AUC, BEDROC AUC, $\alpha = 20$) along with the

enrichment in true active molecules at a constant 1% false positive rate over random picking (EF1%) were calculated for each separate hit list. The same procedure was carried out by fusing all lists and keeping the maximal Tc value for each compound (“max-pooling” approach).

2.2. Rescoring LIT-PCBA Docking Poses by Interaction Graph-Matching (GRIM)

The GRIM module²³ of the IChem package²⁷ was employed to post-process the docking results obtained from Surflex-Dock. All docking poses in multi-mol2 file format were matched to the crystallographic reference ligand pose (in mol2) for rescoring based upon the similarity scores (GrScore) of interaction pattern graphs with the corresponding binding site (in mol2). The best matching pose was selected for every ligand according to the GRIM score,²³ and all molecules of each target set were sorted based on this same value in descending order. ROC AUC, BEDROC AUC and EF1% values were calculated as described above.

3. Results and Discussion

Virtual screening results on 15 ligand sets of the LIT-PCBA data collection,²⁶ demonstrated by EF1% values, using IFP and GRIM rescoring on the docking poses issued by Surflex-Dock are portrayed in **Figure 1**. It can be observed that IFP and GRIM rescoring generally gave good performances that surpassed those of native Surflex-Dock scoring, as the average values of EF1% given by both methods are higher than those received from Surflex-Dock for all 15 target sets: overall enrichment factors were recorded at 4.77 ± 2.85 , 4.78 ± 3.11 , and 2.07 ± 1.00 by IFP (**Table S1**), GRIM (**Table S2**), and Surflex-Dock,²⁶ respectively. When the “max-pooling” approach was applied, at least one of these interaction rescoring methods performed better than energy-based scoring across the whole data collection. Notably, both GRIM and IFP outperformed Surflex-Dock in nearly three quarters of the cases (including the “easy” sets ADRB2 and GBA, on which Surflex-Dock gave significantly better performances than random selection;²⁶ and several “challenging” sets where the Surflex-Dock scoring function failed, e.g. ALDH1, ESR1-ago, or PKM2). This reconfirms the conclusions drawn in earlier publications, which highlight the necessity of post-processing docking poses issued by docking programs and the benefit of using scoring functions based on ligand-protein interaction comparisons (rather than energy-based empirical docking scores, e.g. pK_d values given by Surflex-Dock) for virtual screening.²²⁻²⁵

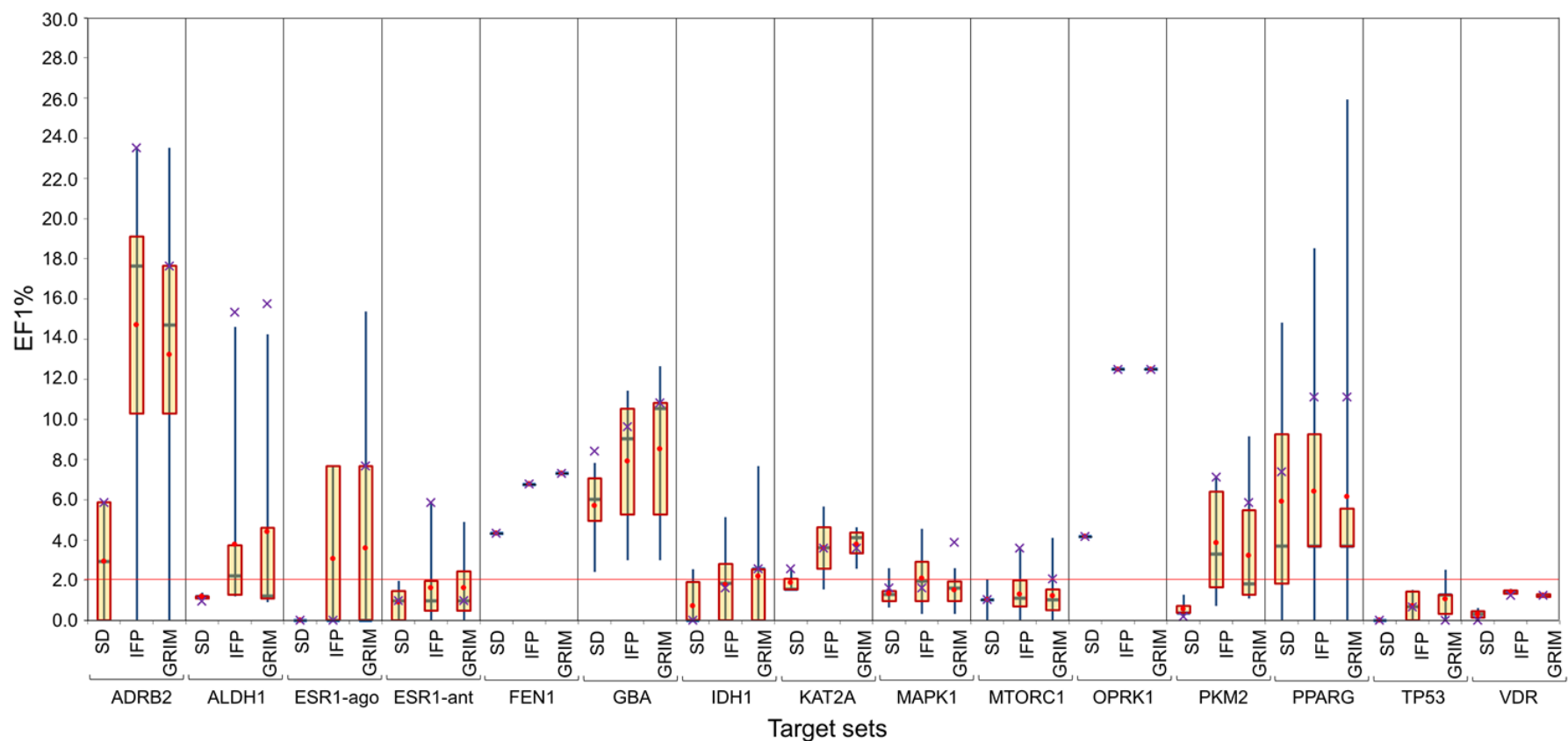


Figure 1. Retrospective virtual screening results on 15 target sets of the LIT-PCBA data collection using the native Surflex-Dock scoring function (SD), protein-ligand interaction fingerprint rescoring (IFP), and interaction graph-matching rescoring (GRIM). Scores were obtained from the same set of docking poses generated by the Surflex-Dock docking engine.

The differences in screening performances given by Surflex-Dock, IFP and GRIM can be further analyzed by examining each target set. An example can be taken from the ESR1-ago set, gathering 5596 substances tested for an agonistic activity on the estrogen receptor alpha (ER-alpha) signaling pathway. Among them, 13 have been confirmed as active, the other 5583 molecules were deemed inactive. A total of 15 protein-ligand complex structures were selected from the Protein Data Bank²¹ and used as templates. The scoring function of Surflex-Dock failed to retrieve any active compound along with the top 1% false positives (EF1% = 0.00 across all 15 templates), while IFP managed to select one true active for six templates, and GRIM successfully retrieved one active for five templates and two actives for one template (**Table S3**). Interestingly, the active substance ID 144206564 (**Figure 2**) was repeatedly selected by the two interaction-comparing scoring functions (in 75% of the cases, **Table S3**). This denotes the agreement of these methods in choosing active molecules among a pool of chemically diverse ligands in the data set. Moreover, this true active shares several key chemical features with the co-crystallized template ligands (**Figure 2**), including the presence of two hydroxyl groups linked to a series of aromatic rings, facilitating three hydrogen bonds with the residues Glu353, Arg394 and His524 of the binding pocket that can also be seen in the PDB template structures (**Figure 3**). While the pK_d scores issued by Surflex-Dock constantly failed to select this molecule, IFP and GRIM rescoring managed to recognize this compound among the top rankers multiple times, thanks to the advantage of comparing ligand-protein interactions in *in silico* screening. This, again, supports the use of this strategy rather than the energy-based empirical scoring functions of popular docking programs in identifying potential hits on the basis of known ligand structures. However, the above observations on the chemical similarities between this IFP-/GRIM-retrieved active molecule and the PDB template ligands do not imply that the screening performance of these two interaction-comparing scoring functions depends on how similar the true actives are to the references; as the Tanimoto values obtained from 2D ECFP4 fingerprint similarity searches are not correlated to those received from IFP comparisons, and also to the computed GRIM scores ($R^2 < 0.1000$ across all 15 templates of the ESR1-ago target set, **Figure S1**). This suggests that the similarity level of protein-ligand interaction fingerprints and of interaction pattern graphs is independent of the chemical similarity of the compared molecules.

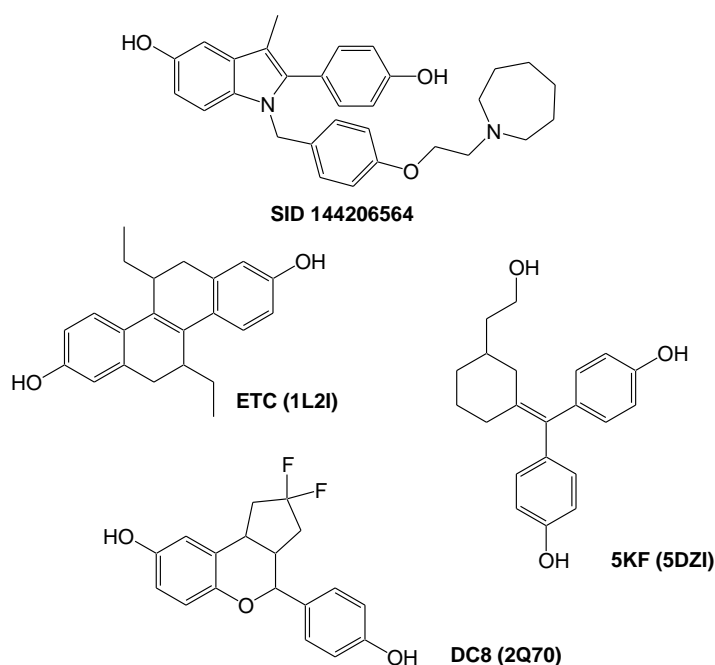


Figure 2. 2D structure of SID 144206564 from the LIT-PCBA ESR1-ago ligand set (the PubChem active substance repeatedly selected by IFP and GRIM along with the top 1% false positives), and those of several PDB template ligands. It can be observed that SID 144206564 shares several key chemical features with the known templates, including two –OH groups linked to a series of aromatic rings, forming three hydrogen bonds also observed in the template structures, which partly explains why the two ligand-protein interaction-comparing methods managed to retrieve this molecule.

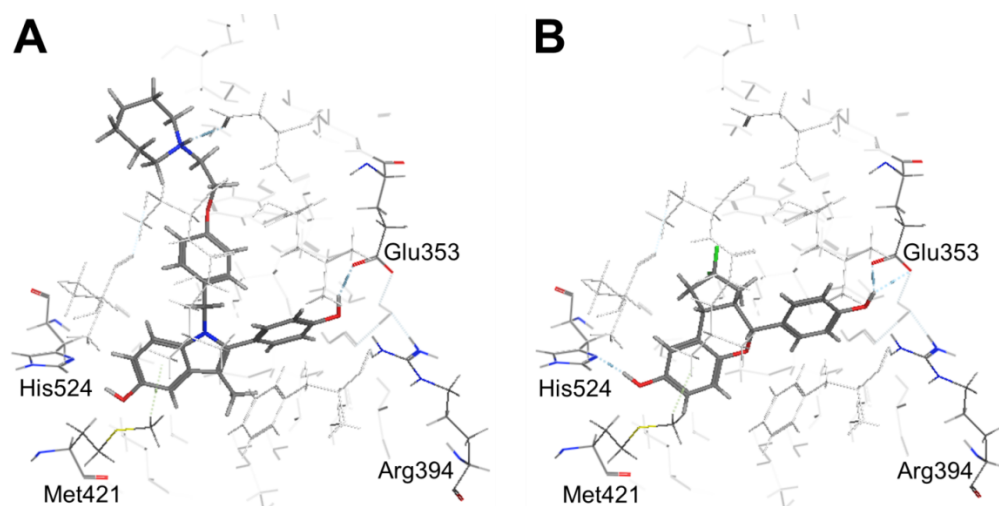


Figure 3. The best pose inside the binding pocket (PDB ID 2Q70) of the active substance ID 144206564 selected by IFP rescoring (A) and the crystallographic pose of a known ligand (HET code: DC8) retrieved from the Protein Data Bank (B) explaining why this active molecule was successfully selected by comparing protein-ligand interaction fingerprints. Identical hydrogen

bonds with the site residues are observed from both poses, including one bond with His524, one bond with Glu353, and another bond with Arg394, all involving the hydroxyl groups in the structures of both ligands (the bond acceptors and bond donors are also identical). Moreover, all hydrophobic interactions recorded in the PDB template are preserved in the IFP-selected pose, e.g. the interaction between Met421 and an aromatic ring of the ligands. This figure was prepared with MOE 2018.01.³¹ The ligands (SID 144206564 and DC8) are portrayed as sticks, while the involved protein residues are portrayed as lines and labeled.

It is observed that the pK_d docking scores issued by Surflex-Dock did not manage to select the pose with the closest interaction patterns with the binding site to those of the references in most cases (nearly 90%). In rare instances where docking selected the same pose as the ligand-protein interaction-comparing algorithms, the empirical pK_d values still failed to rank active molecules above inactive ones in the hit list. An example of this can be taken from the three inactive substances IDs 144203677, 144203979 and 144204501 (**Figure 4**) included in the ESR1-ago set of LIT-PCBA. Both IFP and Surflex-Dock chose the same best pose for each of these three molecules when the PDB template ID 2Q70 was employed. However, while the native energy-based scoring function ranked all these inactives above the confirmed hit SID 144206564, IFP rescoring successfully assigned a higher rank to this true active. An analysis of ligand-protein interactions observed from the aforementioned molecules is provided in **Table 1**.

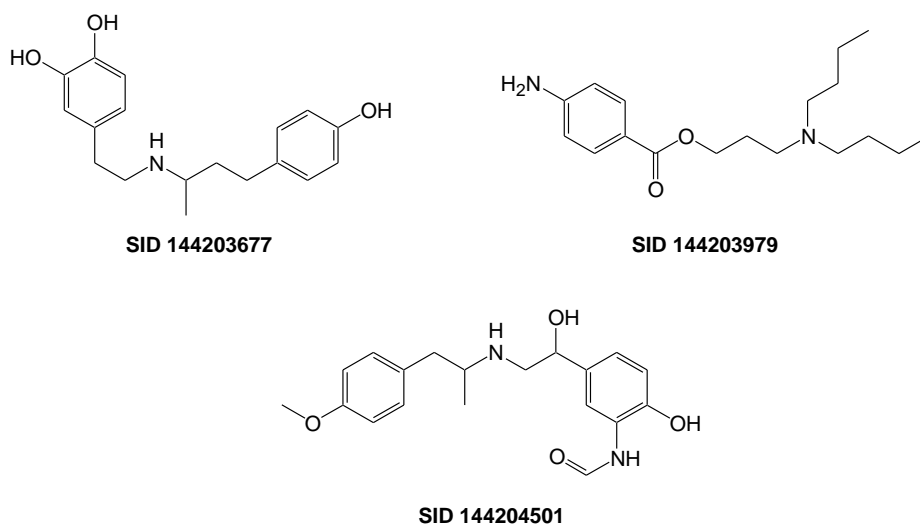


Figure 4. 2D structures of three inactive substances IDs 144203677, 144203979 and 144204501 from the LIT-PCBA ESR1-ago ligand set. IFP and Surflex-Dock agreed on the best pose for each substance, but Surflex-Dock failed to rank the confirmed hit SID 144206564 above these three inactives in the hit list (with the PDB ID 2Q70 used as template), while IFP rescoring managed to do so.

Table 1. Analysis of protein-ligand interactions observed from the best poses (selected by both IFP and Surflex-Dock) of SIDs 144203677, 144203979 and 144204501 (LIT-PCBA ESR1-ago ligand set) inside the binding pocket of the PDB ID 2Q70. A similar analysis of the true active SID 144206564 (best pose selected by IFP) is also provided for comparison.

	SID 144206564 (active)	SID 144203677 (inactive)	SID 144203979 (inactive)	SID 144204501 (inactive)
Hydrogen bonds	All hydrogen bonds observed in the PDB template were retained. No additional hydrogen bond was formed.	All hydrogen bonds observed in the PDB template were retained. However, the ligand engaged in another hydrogen bond with Met421.	The hydrogen bond with the residue Arg394 observed in the PDB template was not formed by the ligand. Besides, the ligand engaged in another hydrogen bond with Leu346.	The hydrogen bond with the residue His524 observed in the PDB template was not formed by the ligand. No additional hydrogen bond was formed.
Hydrophobic interactions	All 36 hydrophobic interactions with 14 residues in the binding site observed in the PDB template were retained. The ligand also engaged in hydrophobic interactions with one more site residue (Met388).	33 hydrophobic interactions with 15 residues in the binding site were formed. In comparison to the PDB template, this ligand did not engage in hydrophobic interactions with Trp383, but with two other residues (Met388 and His524).	33 hydrophobic interactions with 15 residues in the binding site were formed. In comparison to the PDB template, this ligand formed hydrophobic interactions with another residue (Met388).	30 hydrophobic interactions with 13 residues in the binding site were formed. In comparison to the PDB template, this ligand did not engage in hydrophobic interactions with Leu349 and Leu384, but with another residue (Met388).
pK _d by Surflex-Dock	8.0184	8.1467	10.5816	8.1929
Tc values by IFP rescoring	0.9000	0.7368	0.7500	0.7895

Based on the above analyses, it is clear that the true active SID 144206564 gave the most similar interaction patterns with the binding site to those observed in the PDB template 2Q70. The IFP rescoring, upon comparing interaction fingerprints of the PubChem molecules with those of the reference (**Table S4**), managed not only to select the right pose for the true active, but also to rank this confirmed hit above the three inactives (SIDs 144203677, 144203979 and 144204501) in the hit list, thus recognizing it among the top rankers, while Surflex-Dock gave this true active the lowest (and poorest) pK_d docking score. The observations detailed herein reconfirm that the energy-based empirical scoring functions employed by docking programs (e.g. Surflex-Dock) are not as effective as those relying on comparisons of ligand-protein interactions in selecting potential hits for a protein target among chemically diverse ligands.

On a side note, while IFP and GRIM outperformed Surflex-Dock on the 15 target sets of LIT-PCBA, their performances on this data collection are still poorer than those obtained from other databases, including DUD-E.^{22,23} This once again highlights the particular challenge brought by our newly introduced data set, thanks to the absence of both obvious and hidden bias in its design, which indeed prevents an overestimation of virtual screening performances.

4. Conclusion

Finding a scoring function to select one best pose for a ligand among those issued by a docking program and to rank these molecules in a hit list with the aim of retrieving as many potential hits as possible is a task that has long been tackled by the cheminformatics community. Various publications in the literature have raised the issue with energy-based empirical scoring functions employed by popular docking programs, as regards their inaccuracy in estimating the binding affinity of a molecule, and in screening a data set in several virtual screening challenges. The findings portrayed in this chapter reconfirm the conclusions indicated in earlier papers, pointing out that the pK_d docking scores issued by Surflex-Dock gave generally poorer performances on the LIT-PCBA data collection (in terms of early enrichment of true actives) than the two scoring functions based on measuring the similarity level of protein-ligand interaction fingerprints (IFP) and of interaction pattern graphs (GRIM). This highlights the importance of post-processing the docking poses output by docking programs, notably by the approaches relying on comparing the interaction modes inside the binding pocket of these poses with those of a high-quality reference in *in silico* screening.

References

1. Pham, T. A.; Jain, A. N. Customizing Scoring Functions for Docking. *J. Comput. Aided Mol. Des.* **2008**, *22*, 269-286.
2. Bohm, H. J. The Development of a Simple Empirical Scoring Function to Estimate the Binding Constant for a Protein-Ligand Complex of Known Three-Dimensional Structure. *J. Comput. Aided Mol. Des.* **1994**, *8*, 243-256.
3. Jain, A. N. Scoring Noncovalent Protein-Ligand Interactions: A Continuous Differentiable Function Tuned to Compute Binding Affinities. *J. Comput. Aided Mol. Des.* **1996**, *10*, 427-440.
4. Wang, R.; Liu, L.; Lai, L.; Tang, Y. SCORE: A New Empirical Method for Estimating the Binding Affinity of a Protein-Ligand Complex. *J. Mol. Model.* **1998**, *4*, 379-384.
5. Wang, R.; Lai, L.; Wang, S. Further Development and Validation of Empirical Scoring Functions for Structure-Based Binding Affinity Prediction. *J. Comput. Aided Mol. Des.* **2002**, *16*, 11-26.
6. Jain, A. N. Surflex-Dock 2.1: Robust Performance from Ligand Energetic Modeling, Ring Flexibility, and Knowledge-Based Search. *J. Comput. Aided Mol. Des.* **2007**, *21*, 281-306.
7. Welch, W.; Ruppert, J.; Jain, A. N. Hammerhead: Fast, Fully Automated Docking of Flexible Ligands to Protein Binding Sites. *Chem. Biol.* **1996**, *3*, 449-462.
8. Pham, T. A.; Jain, A. N. Parameter Estimation for Scoring Protein-Ligand Interactions Using Negative Training Data. *J. Med. Chem.* **2006**, *49*, 5856-5868.
9. Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical Scoring Functions: I. The Development of a Fast Empirical Scoring Function to Estimate the Binding Affinity of Ligands in Receptor Complexes. *J. Comput. Aided Mol. Des.* **1997**, *11*, 425-445.
10. Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and Validation of a Genetic Algorithm for Flexible Docking. *J. Mol. Biol.* **1997**, *267*, 727-748.
11. Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A Fast Flexible Docking Method Using an Incremental Construction Algorithm. *J. Mol. Biol.* **1996**, *261*, 470-489.
12. Zavodszky, M. I.; Sanschagrin, P. C.; Korde, R. S.; Kuhn, L. A. Distilling the Essential Features of a Protein Surface for Improving Protein-Ligand Docking, Scoring, and Virtual Screening. *J. Comput. Aided Mol. Des.* **2002**, *16*, 883-902.
13. Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Improved Protein-Ligand Docking Using GOLD. *Proteins* **2003**, *52*, 609-623.
14. Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004**, *47*, 1739-1749.
15. Verkhivker, G. M.; Bouzida, D.; Gehlhaar, D. K.; Rejto, P. A.; Arthurs, S.; Colson, A. B.; Freer, S. T.; Larson, V.; Luty, B. A.; Marrone, T.; Rose, P. W. Deciphering Common Failures in Molecular Docking of Ligand-Protein Complexes. *J. Comput. Aided Mol. Des.* **2000**, *14*, 731-751.
16. Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated Docking Using a Lamarckian Genetic Algorithm and an Empirical Binding Free Energy Function. *J. Comput. Chem.* **1998**, *19*, 1639-1662.

17. Rarey, M.; Kramer, B.; Lengauer, T. Multiple Automatic Base Selection: Protein- Ligand Docking Based on Incremental Construction without Manual Intervention. *J. Comput. Aided Mol. Des.* **1997**, *11*, 369-384.
18. Smith, R.; Hubbard, R. E.; Gschwend, D. A.; Leach, A. R.; Good, A. C. Analysis and Optimization of Structure-Based Virtual Screening Protocols. (3). New Methods and Old Problems in Scoring Function Design. *J. Mol. Graph. Model.* **2003**, *22*, 41-53.
19. Jain, A. N.; Dietterich, T. G.; Lathrop, R. H.; Chapman, D.; Critchlow, R. E.; Bauer, B. E.; Webster, T. A.; Lozano-Perez, T. A Shape-Based Machine Learning Tool for Drug Design. *J. Comput. Aided Mol. Des.* **1994**, *8*, 635-652.
20. Jain, A. N.; Harris, N. L.; Park, J. Y. Quantitative Binding Site Model Generation: Compass Applied to Multiple Chemotypes Targeting the 5-HT1A Receptor. *J. Med. Chem.* **1995**, *38*, 1295-1308.
21. 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.
22. Marcou, G.; Rognan, D. Optimizing Fragment and Scaffold Docking by Use of Molecular Interaction Fingerprints. *J. Chem. Inf. Model.* **2007**, *47*, 195-207.
23. 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.
24. Slynko, I.; Da Silva, F.; Bret, G.; Rognan, D. Docking Pose Selection by Interaction Pattern Graph Similarity: Application to the D3R Grand Challenge 2015. *J. Comput. Aided Mol. Des.* **2016**, *30*, 669-683.
25. Da Silva Figueiredo Celestino Gomes, P.; Da Silva, F.; Bret, G.; Rognan, D. Ranking Docking Poses by Graph Matching of Protein-Ligand Interactions: Lessons Learned from the D3R Grand Challenge 2. *J. Comput. Aided Mol. Des.* **2018**, *32*, 75-87.
26. 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** (in press).
27. Da Silva, F.; Desaphy, J.; Rognan, D. IChem: A Versatile Toolkit for Detecting, Comparing, and Predicting Protein-Ligand Interactions. *ChemMedChem* **2018**, *13*, 507-510.
28. *Sybyl-X Molecular Modeling Software Packages, 2.1.1*; TRIPOS Associates, Inc.: St. Louis, MO, USA, 2013.
29. Triballeau, N.; Acher, F.; Brabet, I.; Pin, J. P.; Bertrand, H. O. Virtual Screening Workflow Development Guided by the “Receiver Operating Characteristic” Curve Approach. Application to High-Throughput Docking on Metabotropic Glutamate Receptor Subtype 4. *J. Med. Chem.* **2005**, *48*, 2534-2547.
30. Truchon, J. F.; Bayly, C. I. Evaluating Virtual Screening Methods: Good and Bad Metrics for the “Early Recognition” Problem. *J. Chem. Inf. Model.* **2007**, *47*, 488-508.
31. *Molecular Operating Environment (MOE), 2018.01*; Chemical Computing Group, Inc.: Montreal, QC, Canada, 2015.

Supporting Information**Table S1.** Virtual screening results, in terms of EF1%, obtained by IFP rescoring on the docking poses issued by Surflex-Dock across all 15 target sets of the LIT-PCBA data collection.

Target set	PubChem AID	EF1%			
		Min	Max	Mean \pm SD	Fused
ADRB2	492947	0.00	23.53	14.71 \pm 8.32	23.53
ALDH1	1030	1.21	14.61	3.79 \pm 4.51	15.35
ESR1-ago	743075	0.00	7.69	3.08 \pm 3.90	0.00
ESR1-ant	743080	0.00	5.88	1.63 \pm 1.76	5.88
FEN1	588795	6.78	6.78	6.78	6.78
GBA	2101	3.01	11.45	7.93 \pm 3.52	9.64
IDH1	602179	0.00	5.13	1.78 \pm 1.68	1.61
KAT2A	504327	1.55	5.67	3.61 \pm 2.06	3.61
MAPK1	995	0.32	4.55	2.10 \pm 1.33	1.62
MTORC1	493208	0.00	3.61	1.30 \pm 1.06	3.61
OPRK1	1777	12.50	12.50	12.50	12.50
PKM2	1631	0.73	7.14	3.85 \pm 2.56	7.14
PPARG	743094	0.00	18.52	6.42 \pm 5.32	11.11
TP53	651631	0.00	1.51	0.73 \pm 0.80	0.66
VDR	504847	1.24	1.58	1.41 \pm 0.24	1.24
Overall		1.82	8.68	4.77 \pm 2.85	6.95

Table S2. Virtual screening results, in terms of EF1%, obtained by GRIM rescoring on the docking poses issued by Surflex-Dock across all 15 target sets of the LIT-PCBA data collection.

Target set	PubChem AID	EF1%			
		Min	Max	Mean \pm SD	Fused
ADRB2	492947	0.00	23.53	13.24 \pm 7.54	17.65
ALDH1	1030	0.92	14.23	4.42 \pm 6.06	15.76
ESR1-ago	743075	0.00	15.38	3.59 \pm 4.92	7.69
ESR1-ant	743080	0.00	4.90	1.63 \pm 1.51	0.98
FEN1	588795	7.32	7.32	7.32	7.32
GBA	2101	3.01	12.65	8.53 \pm 4.13	10.84
IDH1	602179	0.00	7.69	2.20 \pm 2.22	2.56
KAT2A	504327	2.58	4.64	3.78 \pm 1.07	3.61
MAPK1	995	0.32	2.60	1.51 \pm 0.69	3.90
MTORC1	493208	0.00	4.12	1.22 \pm 1.20	2.06
OPRK1	1777	12.50	12.50	12.50	12.50
PKM2	1631	1.10	9.16	3.23 \pm 2.84	5.86
PPARG	743094	0.00	25.93	6.17 \pm 7.09	11.11
TP53	651631	0.00	2.53	1.06 \pm 0.95	0.00
VDR	504847	1.13	1.36	1.25 \pm 0.16	1.24
Overall		1.93	9.90	4.78 \pm 3.11	6.87

Table S3. List of true active SIDs included in the ESR1-ago target set of LIT-PCBA that were retrieved along with the top 1% false positives by rescoring the docking poses issued by Surflex-Dock with the two ligand-protein interaction-comparing methods IFP and GRIM. The numbers in brackets represent the EF1% values obtained after virtual screening. Results from Surflex-Dock²⁶ are also indicated for comparison.

PDB entry	True active SIDs retrieved by Surflex-Dock scoring		True active SIDs retrieved by IFP rescoring		True active SIDs retrieved by GRIM rescoring	
1L2I	None	(0.00)	144206564	(7.69)	None	(0.00)
2B1V	None	(0.00)	144209467	(7.69)	144207138	(7.69)
2B1Z	None	(0.00)	None	(0.00)	None	(0.00)
2P15	None	(0.00)	144206564	(7.69)	144206564	(7.69)
2Q70	None	(0.00)	144206564	(7.69)	None	(0.00)
2QR9	None	(0.00)	None	(0.00)	None	(0.00)
2QSE	None	(0.00)	None	(0.00)	None	(0.00)
2QZO	None	(0.00)	None	(0.00)	None	(0.00)
4IVW	None	(0.00)	None	(0.00)	144206564	(7.69)
4PPS	None	(0.00)	None	(0.00)	None	(0.00)
5DRJ	None	(0.00)	None	(0.00)	None	(0.00)
5DU5	None	(0.00)	None	(0.00)	144207138	(7.69)
5DUE	None	(0.00)	None	(0.00)	144206564 144203706	(15.38)
5DZI	None	(0.00)	144206564	(7.69)	None	(0.00)
5E1C	None	(0.00)	144206564	(7.69)	144206564	(7.69)
Max-pooling	None	(0.00)	None	(0.00)	144206564	(7.69)
EF1%	0.00		3.08 ± 3.90		3.59 ± 4.92	

Table S4. The interaction fingerprints issued by IFP (IChem) of the true active SID 144206564 and the three true inactive SIDs 144203677, 144203979 and 144204501 included in the ESR1-ago target set of LIT-PCBA, using the PDB entry 2Q70 as template (co-crystallized ligand HET code: DC8). The bold red digits in the bit strings mark the differences between the IFP of the LIT-PCBA ligands and those of the reference. It can clearly be seen that the active SID 144206564 gave the most similar IFP to those of DC8, with only one difference; while the IFP observed in all three inactive molecules differed significantly from those of the PDB entry. Thanks to these comparisons, IFP managed to rank the true active higher than the true inactives in the hit list, thus recognizing it among the top rankers, while the energy-based empirical scoring function of Surflex-Dock failed to do so. The readers are addressed to the Table 1 of this chapter for detailed analyses.

Molecule	IFP
DC8 (PDB ID: 2Q70, reference)	A HOH3 A M343 A L346 A T347 A L349 A A350 A E353 A W383 A L384 A L387 A M388 A L391 A R394 A L402 A F404 A V418 A G420 A M421 A I424 A F425 A L428 A G521 A M522 A H524 A L525 0001000100000010000001000000100000010000000000100100000010000001000000000000 00100000000010000000000100 0000000000000000001001000000
SID 144206564 (active)	A HOH3 A M343 A L346 A T347 A L349 A A350 A E353 A W383 A L384 A L387 A M388 A L391 A R394 A L402 A F404 A V418 A G420 A M421 A I424 A F425 A L428 A G521 A M522 A H524 A L525 00010001000000100000010000001000000100000000000100100000010000001000000 1 0000 001000000000100000000001000 0000000000000000001001000000
SID 144203677 (inactive)	A HOH3 A M343 A L346 A T347 A L349 A A350 A E353 A W383 A L384 A L387 A M388 A L391 A R394 A L402 A F404 A V418 A G420 A M421 A I424 A F425 A L428 A G521 A M522 A H524 A L525 00010001000000100000010000001000000100000000000100 0 000000010000001000000 1 0000 00100000000010000000000100 00000000000 1 0001001000000
SID 144203979 (inactive)	A HOH3 A M343 A L346 A T347 A L349 A A350 A E353 A W383 A L384 A L387 A M388 A L391 A R394 A L402 A F404 A V418 A G420 A M421 A I424 A F425 A L428 A G521 A M522 A H524 A L525 000 0 00010000001000 1 0010000001000000100000000000100100000010000001000000 1 0000 001000000000 0 000000000001000 00000000000000001001000000
SID 144204501 (inactive)	A HOH3 A M343 A L346 A T347 A L349 A A350 A E353 A W383 A L384 A L387 A M388 A L391 A R394 A L402 A F404 A V418 A G420 A M421 A I424 A F425 A L428 A G521 A M522 A H524 A L525 000100010000001000000 0 00000001000000000001001000000 0 00000001000000 1 0000 00100000000010000000000100 0000000000000000 0 001000000

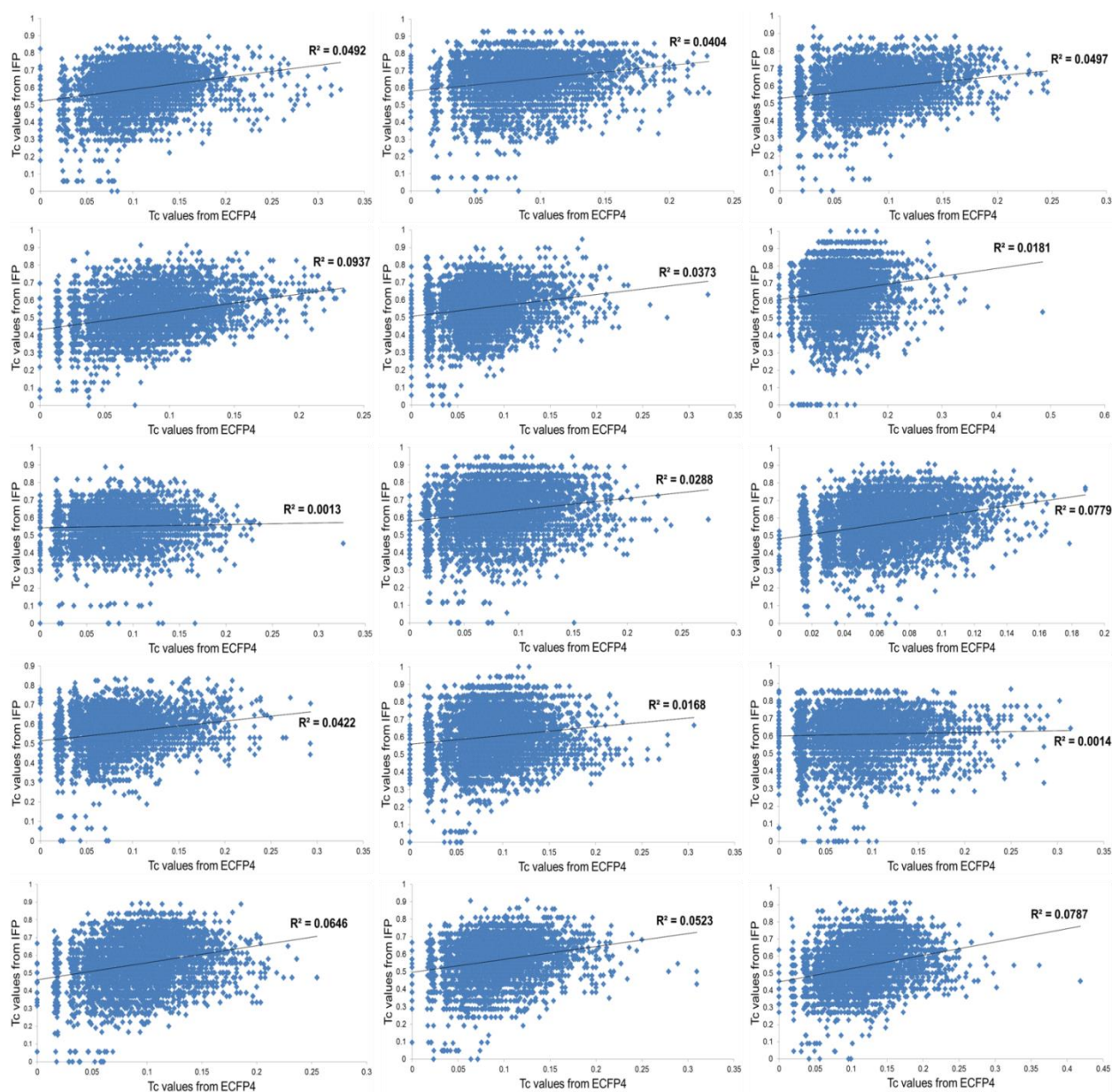


Figure S1. Scatter graphs portraying the Tanimoto (Tc) similarity values obtained from 2D ECFP4 fingerprint comparisons and those issued by IFP rescoring across all 15 templates of the ESR1-ago target set included in the LIT-PCBA data collection. From left to right and top to bottom: 1L2I, 2B1V, 2B1Z, 2P15, 2Q70, 2QR9, 2QSE, 2QZO, 4IVW, 4PPS, 5DRJ, 5DU5, 5DUE, 5DZI, 5E1C. It is observed that the R^2 values are below 0.1000 for all templates, denoting that there is almost no correlation between the Tc values received from computing 2D structural similarity and those from IFP comparisons. This suggests that the similarity level of protein-ligand interaction fingerprints is independent of the chemical similarity of the compared molecules.

Overall Conclusions

Overall, the original work portrayed in this doctoral thesis addressed the issues explained in the Introduction section, offering novel solutions that may come in useful for future *in silico* screening-related research. More specifically:

- A novel small molecule-aligning procedure based on pharmacophoric points derived from the residues constituting a potentially “druggable” cavity of any given protein target was developed. This method was proven more effective than Surflex-Dock, LigandScout and Discovery Studio in predicting the exact binding poses of various ligands inside their binding pockets, and was deemed comparable in discriminatory power to several state-of-the-art virtual screening programs in retrieving true active molecules, and recognizing their scaffolds, among different pools of chemically diverse ligands. Moreover, this method is applicable to apoprotein structures, denoting its high utility even in the absence of a co-crystallized ligand. The method is expected to contribute to virtual screening campaigns in the future, with a view to improving the overall hit rates obtained by using it in parallel with other *in silico* screening methods.
- A new unbiased benchmarking data set named LIT-PCBA based on experimentally confirmed data deposited on PubChem BioAssay was developed. Many disadvantages inherent in other data collections, especially the artificially constructed DUD, DUD-E, or DEKOIS, have been avoided or alleviated, to a certain extent, during the design of this data set, as evidenced by post-design evaluation results using various virtual screening procedures. LIT-PCBA is expected to become a new generation of realistic data sets that mimic those employed in real-life high-throughput screening campaigns, offering better validation tools for novel *in silico* screening approaches, both ligand-based and structure-based, especially those relying on machine learning.

Apart from the two main points indicated above, this Ph.D. thesis also provides a comprehensive review of data sets built upon PubChem BioAssay data, analyzes the note-worthy issues that must be addressed when it comes to constructing novel data collections, and proposes a set of good practices that should be followed in order to avoid the aforementioned problems and ensure the quality of data set design. Besides, a part of this dissertation serves to reconfirm the advantages of using ligand-protein interaction-comparing methods, e.g. those relying on interaction fingerprints and interaction pattern graphs, rather than the energy-based empirical

scoring functions of popular docking programs, in virtual screening exercises; as such methods were deemed more effective in retrieving true hits for a protein target, even when applied to a challenging data collection like LIT-PCBA.

Further improvements may be brought to the output of the work portrayed in this thesis; for example, by modifying the ligand-aligning script to allow a faster calculation on multiple cores, thus enabling the screening of a larger set of molecules while reducing the amount of time required to finish the jobs; or by applying more filtering rules on the LIT-PCBA ligands (e.g., to limit the quantity of highly potent molecules so that their population does not exceed 10% of the active data size), in order to further reduce the hit rates of several target sets, especially those at 2-5%. Moreover, other virtual screening methods, notably deep neural networks, are expected to be applied to LIT-PCBA, in hopes of delineating the true benefit of machine learning approaches in “real-life” structure-based design scenarios. Inspiration can also be taken from the points raised in the review article featured in Chapter 1, even other good practices are encouraged to be added, to give a more complete and effective guideline for developing novel realistic data sets adapted to *in silico* screening evaluation purposes in the future.

DEVELOPPEMENT DE JEUX DE DONNEES NON BIAISES ET DE NOUVELLES METHODES DE CRIBLAGE VIRTUEL

Résumé en français

Les éléments pharmacophoriques issus d'IChem qui représentent le site actif d'une protéine (même sans ligand co-cristallisé) sont simples et assez précis pour faire du criblage virtuel. La nouvelle procédure proposée dans ce travail s'avère aussi efficace que des méthodes computationnelles existantes dans l'identification des composés actifs et leurs chémotypes originaux, et peut donc être utilisée en parallèle avec d'autres méthodes de criblage *in silico* afin d'améliorer la performance globale du criblage. On présente également la nouvelle base de données LIT-PCBA, se composant de 15 protéines cibles, chacune avec les vrais actifs et les vrais inactifs déjà confirmés par les essais biologiques issus de "PubChem BioAssay". Ces jeux de données, préparés par une procédure rigoureuse de plusieurs étapes, sont moins biaisés, en matière de structure des ligands et de composition des sets de molécules, que ceux qui existent déjà (DUD, DUD-E, etc.), et sont donc plus difficiles. LIT-PCBA est prête à l'emploi pour des études comparatives de nouvelles méthodes de criblage virtuel, notamment celles basées sur l'intelligence artificielle.

Mots-clés : pharmacophore, site actif, *in silico*, criblage virtuel, alignement, jeux de données, PubChem BioAssay, biais.

Résumé en anglais

The pharmacophoric points issued by IChem that represent the active site of any given protein target (even without co-crystallized ligands) are simple and accurate enough to be employed for virtual screening. The novel ligand-aligning procedure proposed herein has been proven as effective as existing computational methods in identifying active compounds among a pool of chemically diverse molecules, and can be used in parallel with other *in silico* methods in hopes of improving the overall screening performance. Also presented in this work is the novel data collection entitled LIT-PCBA, comprising 15 target sets built upon experimentally confirmed data deposited on PubChem BioAssay. Undergoing a rigorous procedure involving multiple preparation steps, this data set is much less biased, in terms of chemical composition, than the artificially constructed DUD, DUD-E, or DEKOIS, and does not suffer from many drawbacks inherent in other databases. LIT-PCBA therefore imposes a more difficult challenge on virtual screening methods, and is now ready for benchmarking studies of novel *in silico* screening procedures, notably those relying on machine learning.

Keywords: pharmacophore, active site, *in silico*, virtual screening, alignment, data set, PubChem BioAssay, bias.