Protein structure-based organic chemistry-driven ligand design from ultra-large chemical spaces

Didier Rognan (rognan@unistra.fr)
Laboratoire d'innovation therapeutique https://orcid.org/0000-0002-0577-641X
François Sindt
Laboratoire d'Innovation Thérapeutique, UMR7200 CNRS/Université de Strasbourg
Anthony SEYLLER
Laboratoire d'Innovation Thérapeutique, UMR7200 CNRS/Université de Strasbourg
Merveille Eguida
Laboratoire d'Innovation Thérapeutique, UMR7200 CNRS/Université de Strasbourg

Article

Keywords:

Posted Date: January 9th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-3687338/v1

Additional Declarations: Yes there is potential Competing Interest. D.R. is co-founder and shareholder of BIODOL Therapeutics. M.E. is employee of Amgen.

Abstract

Ultra-large chemical spaces describing several billion compounds are revolutionizing hit identification in early drug discovery. Because of their size, such chemical spaces cannot be fully enumerated and requires ad-hoc computational tools to navigate them and pick potentially interesting hits. We here propose a structure-based approach to ultra-large chemical space screening in which commercial chemical reagents are first docked to the target of interest and then directly connected according to organic chemistry and topological rules, to enumerate drug-like compounds under three-dimensional constraints of the target. When applied to bespoke chemical spaces of different sizes and chemical complexity targeting two receptors of pharmaceutical interest, the computational method was able to quickly enumerate hits that were either known ligands (or very close analogs) of targeted receptors as well as chemically novel candidates that could be experimentally confirmed by *in vitro* binding assays. The proposed approach is generic, can be applied to any docking algorithm and requires few computational resources to prioritize easily synthesizable hits from billion-sized chemical spaces.

Introduction

Identifying the first hit compounds able to target a macromolecule of interest is often achieved by screening experimentally or computationally a library of drug-like compounds,[1] thereby enabling a hit to lead follow-up using classical medicinal chemistry strategies.[2] Until recently, the commercially-available chemical space describing drug-like compounds amenable to screening has been restricted to 10-15 million compounds with a yearly growth of ca. half a million compounds.[3] On-demand compound libraries[4, 5] have completely changed this situation by proposing billions of compounds not yet available but easily synthesizable in a few steps and reproducible parallel synthesis. Early approaches to virtually screen subsets of ultra large chemical spaces led to spectacular successes,[6-9] notably unexpected high hit rates, very high potencies and fine selectivity.[10, 11] Today, ca. 70 billion compounds are accessible on-demand with fast delivery (6-8 weeks) and high-purity grade (> 95%).[12] Due to their huge size, compounds describing these ultra-large chemical spaces cannot be fully enumerated and requires dedicated computational tools for registration, storage and navigation.[13] Usually, large chemical spaces are described in a combinatorial manner from the building blocks and organic chemistry reactions required to synthesize them.[5] If ligand-based approaches are now available to efficiently query these large chemical spaces[14-16], structure-based approaches including macromolecular target information (e.g. topology of a binding site) still need to be developed to exhaustively mine multibillion chemical spaces. Several computational methods have indeed been described for such a task[17-23], albeit with moderate to severe restrictions. One the one hand, brute force docking of 1.4 billion compounds[23] has been successfully described with the help of costly dedicated platforms,[23, 24] but will soon reach its limits with next-to-come trillion-sized chemical spaces,[25] since full atomistic docking just scales linearly with the number of compounds to be screened. A workaround consists in the proper selection of seed fragments/scaffolds to screen a representative subset of the entire space. The seed fragment may

originate from the early docking of fragment-based representative synthons,[20] X-ray diffraction screening data[21] or medicinal chemistry knowledge.[22] Once a seed fragment has been identified, scaffold-focused two-dimensional (2D) libraries, exploring the corresponding chemical space via a set of organic chemistry reactions,[26] can be enumerated, converted in three-dimensional (3D) atomic coordinates and physically docked to propose novel hits. This approach has been applied with success to a few targets[20-22, 27] but still requires hardware settings enabling docking a significant subset (a few million) of the entire chemical space. Last, fast machine learning approaches may be first trained on a set of representative ligand-annotated docking poses to simply predict docking scores, [17-19, 28, 29] and next be applied to predict docking scores for the remaining space. Even if only a small fraction of the full space (1-5%) has to be docked at the atomic level, this strategy cannot be further applied to trillion- sized chemical spaces since it would require gathering first billion of docking scores on a single target. Moreover, this approach has led to very mitigated results with respect to hit rate and hit potencies,[30] and deserves further experimental validations.

Herein, we present a simple and fast computational approach (SpaceDock) avoiding the above-cited drawbacks. It first requires docking commercially available chemical reagents in the target of interest, in order to couple them according to standard organic chemistry reactions to propose multibillion compound libraries in one or two synthetic steps. When applied to two targets of pharmaceutical interest, the method was able to quickly retrieve hits that are chemically identical (or very close) to existing ligands, but also to propose chemically novel and potent ligands.

Results

Since the SpaceDock method heavily relies on the possibility to accurately dock chemical reagents, we first investigate the best docking prototols for the latter task by setting-up a dedicated benchmarking study. We then describe how chemical reagents are annotated by reactive groups and organic chemistry reactions, to define a chemical space of 53.5 billion synthesizable compounds. Last we present two concrete application of the SpaceDock workflow to two receptors of pharmaceutical interest.

Setting up the conditions for accurate docking of chemical reagents

To evaluate the feasibility of the SpaceDock approach, we first needed to set-up an archive of reference 3D structures for protein-bound chemical reagents. Since experimental data for such a dataset are missing, we fragmented in 3D space drug-like ligands from known protein-ligand X-ray structures (sc-PDB dataset)[31] using a set of 12 common organic chemistry reactions, then added the 3D atomic coordinates of the missing reactive moieties (e.g. boronic acid, halide; **Supplementary Fig. 1**) and last created on-the-fly "surrogate X-ray poses" for the corresponding reagents expected to yield the parent ligands with the above-described reactions. The final archive of 5,845 reagents was selected after appropriate filtering **(Supplementary Table S1**) and exhibit 13 chemical functions with a prevalence of reactive groups (e.g. amines, aryl halides, boronic acids) reflecting the frequent usage of simple organic chemistry reactions in drug discovery.[32] With a set of reference reagents in hand, we could next verify whether state-of-the-art

docking algorithms were able to reproduce the surrogate X-ray poses. Five algorithms relying on different principles (FlexX:[33] incremental construction, GOLD:[34] genetic algorithm, PLANTS:[35] ant colony optimization, RDPSOVina:[36] random drift particle swarm optimization, Surflex:[37] surface-based molecular similarity) were used for that purpose. Since the SpaceDock strategy just need a single pair of complementary reagents to be properly docked to reconstitute a full ligand, the docking performance was measured by computing the root-mean square deviation (rmsd) of the pose found to be the closest (best pose) to that of the surrogate X-ray structure (**Fig. 1**). All docking tools exhibit an excellent docking performance with 70-80% of chemical reagents being docked within 2 Å rmsd accuracy (**Fig. 1A**). Up to 70% of very high-quality poses (rmsd < 1 Å) could be generated by the apparently best docking/scoring scheme (GOLD docking, PLP scoring; **Fig. 1A**). The observed docking accuracy is therefore independent on the chosen docking algorithm, and remains in agreement with docking benchmarks on low molecular weight fragments.[38, 39] Since rmsd is a global measure that does not take into account whether key protein-reagent interactions are verified or not, we additionally computed the similarity of protein-reagent interaction fingerprints (IFPs)[40] between docked and surrogate X-ray poses. Again, an excellent performance could be noticed using this orthogonal quality descriptor, with 75-85 % of chemical reagents for which the IFP similarity to the X-ray pose is deemed acceptable (Tc-IFP > 0.60;[40] **Fig. 1B**). To ascertain that all chemical functions are equally suitable for docking, the same analysis was repeated for each of the 13 chemical groups (**Fig. 1C**) present in our library, focusing on the best docking strategy (GOLD docking, PLP scoring). Reassuringly, the docking performance appears to be relatively independent on the chemical function of the reagent (**Fig. 1C**) as well as on the target protein family (**Fig. 1D**).

Defining a readily-accessible ultra-large chemical space from simple organic chemistry reactions

Starting from the pioneering work of Hartenfeller et al.,[26] we selected 36 robust, stereo- and regioselective organic chemistry reactions to define a chemical space of 5.5 billion compounds readily accessible in one or two synthesis steps (**Supplementary Table S2, Supplementary Fig. 2**). Contrarily to previous similar approaches[26, 41, 42], chemical reagents were here carefully chosen from specific SMARTS strings in a list of 145,705 commercial chemical reagents contributing to Enamine's REAL space[43] of 36 billion compounds. Moreover, possible side reactions affecting synthesis yields were minored by selecting reagents that are monofunctional for a particular chemical function (e.g. monocarboxylic acid), and lacking additional chemical functions (e.g. nucleophilic groups for an electrophilic reactant) that would decrease the reaction yield (**Supplementary Table S2**). Altogether, 134,331 commercial reactants could be unambiguously annotated by reaction type, reactant role and reactive atoms yielding a total of 713,155 atomic tags (**Fig. 2**). Conversion in 3D atomic coordinates provided a total of 176,824 ready-to-dock unique reagents, ionized at pH 7, including stereoisomers for reactants bearing up to two undefined chiral centers.

Retrospective chemical space docking of 97 million compounds for human estrogen receptor beta agonists.

For a first proof-of-concept, we selected as a target the activated form of the human estrogen receptor beta (ERβ) for the following two reasons: (i) the ligand-binding cavity is nicely druggable with a good hydrophobicity/hydrophilicity balance, (ii) the receptor has been co-crystallized with many high-affinity low molecular-weight agonists, notably compounds sharing a 2-aryl-benzoxazole scaffold[44] whose one-step synthesis from 2-aminophenols and benzaldehydes is one of the 36 reactions that we have encoded. To avoid a possible chemotype bias, we selected an X-ray receptor structure co-crystallized with genistein (PDB 1QKM), a non-benzoxazole high-affinity agonist used from hereon as "reference ligand" (**Fig. 3A**) and asked whether we could recover a "ground truth" benzoxazole agonist (WAY-338, **Fig. 3A**) or any close analog, by first docking the necessary reactants (2-aminophenols, benzaldehydes) and then enabling the benzoxazole ring formation within the protein binding site. To this end, 145 commercial 2-aminophenols and 3,874 benzaldehydes were generated in 3D and docked into the 1QKM structure, in order to explore a combinatorial space of 561,730 possible benzoxazoles. Since the later space is small, we additionally considered a much larger space of 97 million sulfonamide decoys synthesizable from 1,275 sulfonyl chlorides and 76,758 amines, thereby strongly minoring the benzoxazole space (0.57%) in the full chemical space to scan. After docking all reagents necessary to mine both chemical spaces according to the previously found best protocol (GOLD docking, PLP scoring), a series of filters of increasing complexity (**Table 1**) was iteratively passed to a decreasing number of possible solutions, first starting with pairs of potentially reacting reagent poses, then with successfully enumerated ligand poses, and last with quality checked redocking poses.

Table 1 | Incremental series of filters applied to prioritize SpaceDock hits

Filter	Type	Criteria	Applies to	Software used
1	Geometry	Distances, angles, clashes	Pair of reactant poses	this work
2	Interaction	Interaction fingerprint similarity to reference	Pair of reactant poses	IChem[45]
3	Energy, geometry	Rmsd of refined pose to non-refined pose	Fully enumerated ligand	Szybki[46] Surflex-Dock[37]
4	Interaction, structure	Interaction fingerprint similarity (IFP) to reference Number of stereocenters Number of rotatable bonds Drug-likeness	Fully enumerated ligand	IChem[45] Filter[46]
5	Redocking	Rmsd to energy-minimized SpaceDock pose IFP similarity to energy-minimized SpaceDock pose	Docking poses	GOLD[34] Surflex-Dock[37] IChem[45]
6	Quality check	Number of strained torsions Local and global strain energy Number of unsatisfied H-bond donors and acceptors, number of unsatisfied ionic bonds	Docking poses	Torsion_analyzer[47] Freeform[46] this work
7	Final selection	Duplicates removal HYDEscore	Docking poses	This work Hydescorer[48]

The SpaceDock flowchart is displayed **Fig. 3**. In a first step, pure chemical and topological filters (**Supplementary Figs. 3, 4**) are passed to all docking poses of possible reactant pairs to quickly remove impossible reactions (filter #1). To stay on a safe side, we only considered pairs of bound reactants exhibiting a total interaction fingerprint (IFP) similarity[40] to the genistein X-ray pose above an acceptable threshold[40] (IFP $\geq$ 0.60 considering all non-bonded interactions, IFP $\geq$ 0.50 considering polar interactions only; filter #2). The 821,702 remaining pairs of reactants were then converted, in the protein 3D space, into the corresponding benzoxazoles and sulfonamides, respectively and the fully enumerated ligands were quickly minimized in the protein binding site. Only 539,906 poses deviated less than 1.0 Å rmsd from the non-refined poses after energy refinement (filter #3). The remaining minimized poses were filtered again according to IFP similarity of the genistein X-ray pose (IFP $\geq$ 0.60 considering all non-bonded interactions, IFP $\geq$ 0.60 considering polar interactions only; filter #4). Compounds with more than

2 stereocenters and 8 rotatable bonds were removed at this stage, leaving 49,569 poses for further processing. To ensure that the selected SpaceDock poses might be recovered by classical docking, all remaining hits were redocked to the ERβ structure, as previously done for the reagents. Only 121,470 poses close to the corresponding energy-minimized SpaceDock poses (rmsd ≤2.0 Å; IFP ≥ 0.60 considering all non-bonded interactions, IFP ≥ 0.60 considering polar interactions only) were retained (filter #5). A quality check of remaining poses (filter #6) was next applied to remove unlikely solutions (≥ 1 strained torsion, local strain energy > 4 kcal/mol, global strain energy > 8 kcal/mol, no unsatisfied ionic bond , >2 unsatisfied h-bond donors, > 4 unsatisfied h-bond acceptors).[22, 49] The number of plausible solutions (7,712) being still important, a custom filter was finally applied to keep only poses anchored at both sides of the binding pocket (H-bond either Glu305 or Arg346, and to His475), as seen for all potent ERβ agonists (recall genistein X-ray pose, **Fig. 3A**). The final hit list comprises 102 poses from 64 unique ligands (filter #7), including 54 benzoxazoles and 10 sulfonamides (**Fig. 3B, Supplementary Table S3)** ranked by decreasing full IFP similarity to the reference ligand, then by decreasing polar IFP similarity, and last by increasing absolute binding free energy predicted by the HYDE scoring function.[48]

Despite in minority in the initial space (0.57%), it is reassuring that the ground truth chemotype was considerably enriched (84 %) in the final hit list. Inspecting the structures and binding poses of the hits, we observed that SpaceDock was indeed able to recover, among the top-ranked hits, the ground truth ligand (rank #9), a known ERβ agonist ChEMBL187673[50] (IC_{50} = 50 nM, rank #25) and 52 other 2-arylbenzoxazoles, with almost perfect binding modes (rmsd =1.15 Å for the ground-truth ligand, **Fig. 3C**). About half of the hits (30 out 64; all from the benzoxazole space), were considered chemically similar to existing ERβ ligands (**Supplementary Fig. 5**), evidencing that SpaceDock can propose both known ligands (or very close analogs thereof) and new chemical entities. However, only a lower number of compounds (17, out of which 10 share the sulfonamide space) strictly intersected Enamine REAL space (**Supplementary Fig. 5**). This observation does not preclude for their synthesizability but just illustrates that these hits, despite the commercial availability of their starting building blocks, cannot be obtained within the scope of 167 parallel synthesis protocols defining REAL space.

From this preliminary proof-of-concept, it appears that the herein presented method is able to perform a complex organic chemistry reaction (ring cyclisation) from suitably posed and chemically compatible chemical reagents, under the 3D constraints of the target's structure, to generate and prioritize fully enumerated ligands for meaningful reasons. We therefore decided to apply SpaceDock to a prospective screening of a much larger chemical space.

Prospective chemical space docking of 670 million compounds for human dopamine D3 receptor antagonists.

We next applied the method to a much larger chemical space of 670 million carboxamides targeting the human dopamine D3 receptor (DRD3). Since the only available high-resolution DRD3 receptor structure (PDB 3PBL) has been obtained in complex with the antagonist eticlopride (**Fig. 4A**),[51] the latter orthomethoxybenzamide (OMB) ligand was used as both reference and ground-truth ligand to recover.

Commercially available carboxylic acids and primary/secondary amines (**Supplementary Table 2**) were first filtered to remove reagents that, upon amide bond formation, would lead to non-drug-like ligands (**Supplementary Table 4**), thereby keeping 19,887 acids and 33,726 amines (in 3D coordinates) to explore a chemical space of 670 million carboxamides (**Fig. 4B**). The resulting 53,613 chemical reagents were then docked to the eticlopride-free DRD3 structure using GOLD docking and PLP scoring, as previously described. Since 20 poses were saved for each reactant, a total of 268 billion (19,887*20*33,726*20) possible reactions were passed to the SpaceDock flowchart (**Fig. 4B**), removing first impossible amide bond formation according to geometrical criteria (**Supplementary Fig.6**) while keeping only amine poses exhibiting the crucial ionic bond to the key Asp110 residue[51] (filter #1, **Fig. 4B**), then retaining pair of reactant poses for which the IFP similarity to the reference ligand is higher than 0.60 for all interactions and 0.50 for polar interactions only (filter #2).[40] A total of 24,674,693 reactions were conducted *in silico* to generate the corresponding carboxamides inside the receptor pocket, that were later energy-minimized. Keeping only minimized poses that did not deviate much from the initial pose (rmsd < 1.0 Å) afforded 15,120,198 plausible solutions (filter #3, **Fig. 4B**). At this stage, hits bearing a cis-amide bond or more than 2 chiral centers or more than 9 rotatable bonds were removed to keep only drug-like compounds. The resulting number of hits being still very high, we pruned the hit list by keeping only minimized poses with a high full IFP similarity to the reference ligand (IFP similarity > 0.60) while exhibiting a perfect IFP similarity to eticlopride (IFP =1) with respect to polar interactions (H-bond and ionic bond to Asp110). This filter (filter #4, **Fig. 4B**) yielded to 518,306 SpaceDock poses (corresponding to 500,041 unique compounds) that had to be confirmed by full atomistic docking (GOLD docking, PLP scoring, 20 poses saved) of the corresponding ligands and comparison with the minimized SpaceDock poses. Only docking poses verifying the following three criteria (rmsd $\leq$2.0 Å & IFP_full $\geq$ 0.60 & IFP_polar =1) were retained, leaving 712,120 good docking poses (filter #5, **Fig. 4B**) for sanity check (no strained torsion, local strain energy $\leq$ 4 kcal/mol, global strain energy $\leq$ 8 kcal/mol, no unsatisfied ionic bond, $\leq$ 2 unsatisfied h-bond donors, $\leq$ 4 unsatisfied h-bond acceptors, filter #6, **Fig. 4B**). The number of remaining poses being still important (97,096), a custom filter (not implemented by default, **Table 1**) was added to remove poses for compounds with no aromatic ring (always present in known DRD3 antagonists),[52] exhibiting a predicted absolute binding free energy (HYDEscore) lower than 30 kJ/mol and further restricting the deviation to the original SpaceDock poses (rmsd $\leq$1.0 Å & IFP_full $\geq$ 0.75). A reasonable number of 757 docking poses from 315 unique ligands (filter #7, **Fig. 4B**) defined the final hit list. Compounds were ranked by decreasing full IFP similarity to the reference ligand, then by decreasing polar IFP similarity, and last by increasing HYDE binding free energy (**Supplementary Table 5**).

As for the first attempt on ERβ ligands, we first check whether the ground-truth ligand and its corresponding OMB scaffold were present in the list. Indeed, 15 OMBs including eticlopride (rank 30) were part of the list with binding poses very similar to that observed for the reference ligand (rmsd of eticlopride = 0.73 Å, **Fig. 4C**). Interestingly, 300 additional hits not sharing the OBM scaffold were prioritized with poses and protein-ligand interaction patterns quite close to that seen for eticlopride (**Fig. 4D**). Most ligands were scaffold hops for which the orthomethoxybenzamide has been replaced by a bicyclic heteroaryl-amide, connected by 2-3 carbon atoms to a basic amine. By comparison to the ERβ hit

list, the DRD3 hits deviate more from known ChEMBL ligands (24% considered as chemically similar) but are more easily obtainable in REAL space (53% being directly purchasable, and additional 38% being very close to REAL space compounds; **Supplementary Fig. 7**). 16 chemically diverse and representative hits were directly purchased at Enamine, out of which 15 could be synthesized in six weeks (5 mg quantity, > 90% purity) and further tested for binding to human DRD3 (**Fig. 5**).

Out of the tested 15 compounds, ten exhibited detectable binding (> 20% inhibition) to the DRD3 receptor at the single concentration of 10 μM (**Fig. 5**). The six strongest binders (#1, #25, #66, #107, #142, #161) were selected for dose-curve responses for inhibition constants (K_i) determination (**Fig. 5**, **Supplementary Fig. 8**). Three of them (#1, #66, #142) exhibited K_i values in the 300-400 nM range, the three others at 1.4-1.6 μM. The remarkable hit rates (66% at 10 μM, 20% at 500 nM) are in line with previous observations from docking ultra-large libraries,[10, 11] and suggests that SpaceDock competes rather well with much more demanding full atomistic docking when screening large chemical spaces.

Interestingly, novel heteroamatic-carboxamide scaffolds were disclosed for 4 of the strong binders (#66, #107, #142 and #161) that could not be found in any of 6,714 dopamine DRD2/DRD3 ligands from ChEMBL (**Table 2**). SpaceDock proposals should still be considered as primary hits. As such, their potency is lower than that of the closest dopamine D2/D3 antagonists from ChEMBL, albeit with a higher ligand efficiency.

Conclusion

We herein describe a novel computational method (SpaceDock) to exhaustively browse ultra-large chemical spaces under specific constraints of a target protein and known binders. When applied to two nicely druggable targets (estrogen receptor β, dopamine D3 receptor) and chemical spaces up to 670 million compounds, it enabled the fast recovery of known ligands/scaffolds (in both cases) and the identification of novel and potent new chemical entities (dopamine D3 receptor).

SpaceDock departs from existing methods[20-22] by two major differences: (i) fully unmodified chemical reagents and not synthons (scaffolds with chemistry-informed exit vectors) are used as primary sources of hits, (ii) most promising ligands are directly obtained within the protein binding site, by 3D *in silico* synthesis according to geometrical and chemical cross-compatibility of previously posed reagents pairs.

Indeed, direct docking of chemical reagents has, to the best of our knowledge, never been reported. Interestingly, our preliminary benchmark demonstrates that docking chemical reagents is as accurate as docking low-molecular weight fragments[38] with ca. 75% of chemicals properly posed with respect to their corresponding substructures in full PDB ligands. Noteworthy, the docking accuracy is independent on the docking tool used, as well on the reactive moiety of the reactants and on the target protein family; therefore opening the method to any druggable target and set of commercial building blocks. To enable an easy synthetic access to most SpaceDock hits, the method relies on chemical reagents contributing to Enamine's REAL space, and generate hits in the binding site 3D space using a set of 36 robust two-

component organic chemistry reactions. Given the 70% average docking accuracy of reactants, we therefore expect the likehood to properly couple two chemically compatible reactants into a fully enumerated and suitably posed ligand at ca. 50%. Docking the starting chemical reagents is clearly the most time-consuming step of the entire flowchart (ca 15 s/reagent), meaning that SpaceDock scales with the number of reactants and not the number of products defining the chemical space to be screened. To optimize the speed of the further processing, a series of filters of increasing complexity is applied, step to step, to a decreasing number of plausible solutions. Just checking the relative position of compatible reactants to be paired by fast distance/angles measures permits to remove 99.8% of possible solutions. Although not mandatory, we applied IFP similarity to a reference pose to remove topologically valid ligands not fulfilling expected interactions with key residues. This filter permits to reduce the number of full ligand poses to the third most time-consuming but necessary energy-minimization step (ca. 1s/ recombined pose), and remove local strains around the newly created bonds. We assume that a SpaceDock proposal is all the more interesting if it does not vary (in terms of rmsd and IFP similarity) upon energy minimization within the protein binding site, and if it can be recovered by full atomistic docking of the corresponding ligand. Although not necessary, we recommend this redocking step to ensure that SpaceDock and any state-of-the-art docking tool (we here used GOLD but other tools may be used as well) agrees on the final poses to be sent to the very important quality check. A particular importance is given to local and global strain energies (≤4 and 8 kcal/mol, respectively), as well as to the number of unsatisfied ionic bonds (none) and of unsatisfied hydrogen-bond donors/acceptors (≤ 2 and 4, respectively). In the DRD3 test case, omitting this step drastically enriched the final hit list in false positives which could not be confirmed experimentally (data not shown). The herein proposed chemical space docking approach could yield, at least for the present case of a G protein-coupled receptor, to experimentally-validated hits with a high hit rate and nanomolar potencies that agrees with tendencies already noticed upon full atomistic docking of ultra-large library virtual screens. [10, 11]

SpaceDock remains a relatively light computational procedure since browsing a chemical space of 100 million compounds can be achieved within 2 days on a 16-core Intel$^{(R)}$ Xeon$^{(R)}$ Silver 4210 processor. Mining the entire 5.5 billion chemical space has been made possible for the 4th international CACHE challenge[54] with still limited resources (1 week on 400 cores). Preliminary attempts to scan even larger chemical spaces (e.g. by adding three-component reactions) suggests that the method can be easily applied up to a trillion compounds.

Methods

Setting-up a library of chemical reagents from fragmented protein-bound ligands.

37,922 ligands from the sc-PDB database of druggable protein-ligand 3D structures,[31, 55] were fragmented using a set of 12 RECAP[56]-inspired retrosynthetic rules to yield 97,024 chemical reagents (**Supplementary Fig. 1**) with standard topologies (bond length, angle bending, torsion angles) retrieved from the TRIPOS force-field.[57] The resulting building blocks were then filtered using the following rules: (i)

IChem v.5.2. 8[45] detection of at least four non-covalent interactions (one of which being a ionic bond or an hydrogen-bond) with the original sc-PDB target protein, (ii) a total number of heavy atoms between 3 and 23, (iii) a total number of rotatable bonds inferior or equal to 6, (iv) a heteroatom to carbon ratio between 0.05 and 4.5, (v) no more than two fused cycles, (vi) a number of aromatic rings inferior to 3. The final library comprised 5,845 reagents (mol2 file format) derived from 4,656 unique sc-PDB ligands. Although the building blocks have not been explicitly crystallized with their target, the corresponding poses will be further annotated as "surrogate X-ray" pose.

Docking sc-PDB building reagents to their cognate targets

The above described reagents were docked to the sc-PDB target originally bound to the ligand they were derived of, after randomizing their initial orientation and dihedral angles with the Surflex[37] *ran_archive* routine, using 5 state-of-the-art docking tools (FlexX v.5.2.0,[33] GOLD v.2022,[34] PLANTS v1.2,[35] RDPSOVina v.2.0,[36] Surflex v.4.5.4.3[37]) with almost standard parameters (**Supplementary Tables 6-8**). Since the boron atom is not parametrized in some docking tools, it was replaced by either a dummy atom (FlexX, GOLD, PLANTS, Surflex) or a carbon (RDPSOvina) while keeping the trigonal planar geometry of the boronic acid unchanged. Up to 20 poses were preferentially saved in mol2 file format whenever possible (GOLD, PLANTS, Surflex), in sd file format (FlexX) or in pdbqt file format (RDPSOVina). For each docking pose, the root-mean-square deviation (rmsd) of heavy atoms to the corresponding surrogate X-ray pose was computed thanks to the Surflex *rms* routine when comparing mol2 files, or the ADFRsuite-1.0[58] *obrms* routine when comparing files of different formats (mol2 vs. pdbqt, mol2 vs. sd). In addition, we measured the similarity of protein-ligand interactions between docked and X-ray poses with the IFP module of the IChem v.5.2.8 package.[45]

Preparation of bespoke chemical spaces encoded by 36 robust organic chemistry reactions

The global stock of commercially available building blocks (250,355 compounds, sd file format, date: 2022-12-28) was downloaded from Enamine's website[59] and filtered by catalog identification number to retain 145,707 reagents contributing to the REAL space.[43] Building blocks were then filtered to remove unsuitable entries as previously described.[41] For each of 36 different one or two-steps organic chemistry reactions (**Supplementary Table 2**), the corresponding reactants were retrieved using SMARTS strings[41] queries in PipelinePilot v.22.1.0.2935[60] (**Supplementary Figure 9**). In order to avoid side reactions, building blocks need to be monofunctional for the reactive group of interest and free of any possible poisoning chemical function for the reaction of interest (**Supplementary Table 2**). For each retained building block and possible reaction, an annotation triplet is provided: (i) reaction type, reactant role, reactive atoms. The final annotation table comprises 713,155 annotation triplets for 134,331 REAL building blocks. Selected building blocks were finally ionized at their most likely ionization state at pH 7.4 using PipelinePilot and converted into 3D atomic coordinates with Corina v.3.40,[61] allowing to generate up to 4 diastereoisomers by entry, in a single ready-to-dock mol2 file format.

Docking of chemical reagents to human estrogen receptor beta

The X-ray structure of the human estrogen receptor beta in complex with the agonist genistein[62] was downloaded from the Protein Data Bank (PDB 1QKM). Hydrogen atoms and simultaneous optimisation of protonation states of protein, water and ligand atoms was performed with Protoss v.4.0.[63] All water molecule and genistein were removed, keeping only remaining protein atoms of chain A which were saved in mol2 file format. The commercial building blocks selected for a possible benzoxazole ring or sulfonamide bond formation (145 aminophenols and 3,874 benzaldehydes; 1,275 sulfonyl chlorides and 76,758 amines) were docked to the ERβ atomic coordinates with GOLD using previously reported parameter settings (**Supplementary Table 7**). The cavity was detected from X-ray atomic coordinates of genistein. Up to 20 poses, scored by the PLP scoring function, were retained for each building block.

Docking of chemical reagents to the human dopamine D3 receptor (DRD3)

The X-ray structure of the human dopamine D3 receptor in complex with the antagonist eticlopride[51] was downloaded from the Protein Data Bank (PDB 3PBL). Hydrogen atoms and simultaneous optimisation of protonation states of protein, water and ligand atoms was performed with Protoss v.4.0.[63] The inserted T4-lysozyme sequence (Asn1002-Tyr1161), all water molecule and eticlopride were removed, keeping only remaining protein atoms of chain A which were saved in mol2 file format. The commercial building blocks were initially filtered based on their capacity to form a drug-like molecule through an amide bond formation (**Supplementary Table 4**) and their inclusion in the pool of reagents utilized in the REAL Space. The reagents selected for a possible amide bond formation (33,726 amines and 19,887 carboxylic acids) were docked to the DRD3 atomic coordinates with GOLD using previously reported parameter settings (**Supplementary Table 7**). The cavity was detected from X-ray atomic coordinates of eticlopride. Up to 20 poses, scored by the PLP scoring function, were retained for each building block. To decrease the number of possible recombinations, only docking poses of amines exhibiting an ionic bond to the key residue Asp110, detected on the fly with IChem, were further retained for amide bond formation.

Ligand enumeration by reagents coupling

Given two poses of chemically compatible reagents, a ligand is generated within the protein binding site, according to their respective location and chemical compatibility. Reagent poses are initially loaded using an in-house mol2 parser and annotated for at least one reaction based on the tag table shown in **Fig. 2**. Atomic coordinates of reactive atoms carbon and their immediate neighbors, are extracted and stored for subsequent calculations. This process is repeated for each reaction, following a similar workflow. A subsequent set of filters is applied to pairs of reagent poses, including the distance between their center of mass to promptly eliminate distant pairs, the distance between connectable atoms, examination of certain angles of the future formed bond/ring to ensure a suitable geometry, and consideration of clashes ($\leq$4 between non-reacting atoms) to prevent overlapping substituents. If a pair satisfies all the rules, a bond is created between the connectable atoms. The hybridization of reacting atoms is then updated to reflect the newly created bonds and exit atoms (to be removed after the reaction) are deleted. The fully enumerated molecule is then saved into a single mol2 file. An optional step is available at this stage. If a reference ligand exists, the molecule is initially written to a temporary mol2 file to assess its IFP similarity

(default values are ≥ 0.60 for all non-bonded interactions and ≥ 0.50 for polar interactions) to the reference pose using IChem v.5.2.8. If the similarity threshold is reached, the molecule is transferred to the final mol2 file. Detailed rules of these filters can be found in **Supplementary Figs. 3, 4, 6**. The fully enumerated molecule, in presence of the target protein, is last energy-minimized in Szybki v2.4.0.0,[46] using standard settings and the MMFF94 force-field.[64]

Comparisons to reference ligands

Interaction fingerprint similarity search between any pose (before and after energy refinement) and a reference X-ray ligand was done using standard parameters of the IFP module implemented in the IChem v.5.2.8 package.[45] Likewise, root-mean square deviations were computed with the *rms* routine of Surflex-Dock v.4.5.4.3.[37]

Redocking of SpaceDock poses

The coupling of two reagent poses, followed by protein constraint refinement (referred to as the "SpaceDock" pose), was redocked into the target protein structure using GOLD. The scoring function employed was PLP, with 20 generated poses, and the same parameter file as described in **Supplementary Table 7**. To eliminate structural biases, input ligand structures were converted to SMILES format using OEChem Toolkit v.3.4.0.1[46] and further transformed into 3D structures with Corina v.3.40.[61] Up to four diastereoisomers were generated in a single mol2 file. The resulting full atomistic docking pose, exhibiting a rmsd (computed with Surflex rms) below 2 Å, all non-bonded interactions IFP similarity ≥ 0.60, and precisely the same polar IFP as the corresponding SpaceDock pose, was considered as confirmation and retained for subsequent investigations. If multiple docking poses satisfy these rules for each SpaceDock pose, all of them are retained.

Quality check of redocked poses

The number of torsion strains in every redocking pose was estimated with TorsionAnalyzer v.2.0.0.[47] Any pose with at least one torsion annotated as 'strained' was discarded from further analysis. Local strain (distortion of the specific conformation from the nearest local minima) and global strain (energy required to select the specific conformation from the full conformational ensemble of the corresponding compound in water) energies were then computed with standard parameter of Freeform v.2.4.0.0.[46] Any pose with local and global strain energies higher than 4 and 8 kcal/mol, respectively, were discarded.

Last, remaining poses were inspected, in their protein-bound state, for counting the number of unsatisfied ionic bonds, hydrogen-bond donors and acceptors. First, protein-ligand ionic and hydrogen-bonds were registered with IChem. Any charged atom or hydrogen-bond donor/acceptor atom of the ligand (according to IChem definitions)[40] not present in the above list was annotated as "unsatisfied" atom. Unsatisfied heavy atoms being both donors and acceptors (e.g. hydroxyl oxygen atom) were only counted once. Ligand atoms participating to intra-molecular hydrogen bonds were considered as satisfied.

Altogether, ligand poses with more than 2 unsatisfied donors and 4 unsatisfied acceptors were removed from the final hit list.

Similarity to ChEMBL and REAL Space ligands

Known ligands of the human estrogen receptor beta (CHEMBL242) and human dopamine D2 (CHEMBL217) and D3 (CHEMBL234) receptors were retrieved from the ChEMBL database (release 33)[50] as SMILES strings for ligand entries fulfilling the following criteria: K_i < 1 μM, Assay_type = B). Pairwise chemical similarity between SpaceDock hits and ChEMBL ligands was computed with PipelinePilot v.22.1.0.2935[60] from ECFP4 circular fingerprints and scored by the value of the Tanimoto coefficient.

Maximum common substructure (MCS) similarity of SpaceDock hits (converted from mol2 to SMILES strings, thanks to Open Babel v.3.1.0)[65] to 36 billion REAL space ligands (version REALSpace_36bn_2023-03.space[12]) was computed with SpaceMACS v.0.9.2,[15] to save the top 15 REAL space compounds ranked by decreasing MCS-Tanimito similarity value.

Declarations

Data availability

List of reactants to build benzoxazole, sulfonamide and amide chemical spaces, docked poses of test reactants (ERβ, DRD3 test cases), annotation table of Enamine REAL reactants, IChem configuration files for IFP filtering.

All data and SpaceDock processing scripts are available at https://github.com/litfsindt/LIT-SpaceDock

Code availability

Filter v.4.2.1.1, Szbyki v2.5.1.1, OEChem Toolkit v.3.4.0.1; Freeform v.2.5.1.1: OpenEye Scientific, Santa Fe, N.M., USA, https://www.eyesopen.com/

FlexX v.5.2.0, Hyde v.1.5.0, SpaceMACS v.0.9.2, REAL space in fragment space format: BioSolveIT GmbH, Sankt Augustin, Germany, www.biosolveit.de

GOLD v.2022: CCDC Software Ltd., Cambridge CB2 1EZ, United Kingdom, www.ccdc.cam.ac.uk

Open Babel v.3.1.0, https://github.com/openbabel/openbabel

PLANTS v1.2: University of Konstanz, Germany, http://www.tcd.uni-konstanz.de/research/plants.php

RDPSOVina v2.0: Jiangnan University, Jiangsu, China, https://github.com/li-jin-xing/RDPSOVina

SpaceDock v.1.0.0: https://github.com/litfsindt/LIT-SpaceDock

Surflex-Dock v4.5.4.3: BioPharmics LLC, https://www.biopharmics.com

Acknowledgements

We thank Guillaume Bret (Laboratoire d'innovation thérapeutique) for technical assistance, Michael Bossert and Yurii Moroz (Enamine Ltd.) for sharing the list of REAL space reagents, and the CC-IN2P3 calculation center (Villeurbanne, France) for allocation of computing time and excellent support.

Contributions

D.R conceived the study. A.S. performed the initial benchmarking study on chemical reagents. M.E. designed the rules to filter sc-PDB building blocks. F.S. encoded organic chemistry reactions in 3D space, performed the whole docking of Enamine chemical reagents and wrote the SpaceDock source code to enumerate full ligands. F.S. and D.R analyzed the data. All of the authors contributed to writing and editing the manuscript.

Competing interests

D.R. is co-founder and shareholder of BIODOL Therapeutics. M.E. is employee of Amgen

References

1. Bleicher KH, Bohm HJ, Muller K, Alanine AI. Hit and lead generation: beyond high-throughput screening. *Nat Rev Drug Discov* **2**, 369-378 (2003).
2. Hughes JP, Rees S, Kalindjian SB, Philpott KL. Principles of early drug discovery. *Br J Pharmacol* **162**, 1239-1249 (2011).
3. Lucas X, Gruning BA, Bleher S, Gunther S. The purchasable chemical space: a detailed picture. *J Chem Inf Model* **55**, 915-924 (2015).
4. Tingle BI*, et al.* ZINC-22 horizontal line A Free Multi-Billion-Scale Database of Tangible Compounds for Ligand Discovery. *J Chem Inf Model* **63**, 1666-1776 (2023).
5. Grygorenko OO, Radchenko DS, Dziuba I, Chuprina A, Gubina KE, Moroz YS. Generating Multibillion Chemical Space of Readily Accessible Screening Compounds. *iScience* **23**, 101681 (2020).
6. Lyu J*, et al.* Ultra-large library docking for discovering new chemotypes. *Nature* **566**, 224-229 (2019).
7. Sadybekov AA*, et al.* Structure-Based Virtual Screening of Ultra-Large Library Yields Potent Antagonists for a Lipid GPCR. *Biomolecules* **10**, (2020).
8. Stein RM*, et al.* Virtual discovery of melatonin receptor ligands to modulate circadian rhythms. *Nature* **579**, 609-614 (2020).
9. Alon A*, et al.* Structures of the sigma2 receptor enable docking for bioactive ligand discovery. *Nature* **600**, 759-764 (2021).
10. Lyu J, Irwin JJ, Shoichet BK. Modeling the expansion of virtual screening libraries. *Nat Chem Biol* **19**, 712-718 (2023).

11. Sadybekov AV, Katritch V. Computational approaches streamlining drug discovery. *Nature* **616**, 673-685 (2023).
12. Readily-accessible on-demand chemical spaces, https://www.biosolveit.de/infiniSee (accessed 11-16-2023)
13. Warr WA, Nicklaus MC, Nicolaou CA, Rarey M. Exploration of Ultralarge Compound Collections for Drug Discovery. *J Chem Inf Model* **62**, 2021-2034 (2022).
14. Bellmann L, Penner P, Rarey M. Topological Similarity Search in Large Combinatorial Fragment Spaces. *J Chem Inf Model* **61**, 238-251 (2021).
15. Schmidt R, Klein R, Rarey M. Maximum Common Substructure Searching in Combinatorial Make-on-Demand Compound Spaces. *J Chem Inf Model* **62**, 2133-2150 (2022).
16. Meyenburg C, Dolfus U, Briem H, Rarey M. Galileo: Three-dimensional searching in large combinatorial fragment spaces on the example of pharmacophores. *J Comput Aided Mol Des* **37**, 1-16 (2023).
17. Gentile F*, et al.* Deep Docking: A Deep Learning Platform for Augmentation of Structure Based Drug Discovery. *ACS Cent Sci* **6**, 939-949 (2020).
18. Berenger F, Kumar A, Zhang KYJ, Yamanishi Y. Lean-Docking: Exploiting Ligands' Predicted Docking Scores to Accelerate Molecular Docking. *J Chem Inf Model* **61**, 2341-2352 (2021).
19. Graff DE, Aldeghi M, Morrone JA, Jordan KE, Pyzer-Knapp EO, Coley CW. Self-Focusing Virtual Screening with Active Design Space Pruning. *J Chem Inf Model* **62**, 3854-3862 (2022).
20. Sadybekov AA*, et al.* Synthon-based ligand discovery in virtual libraries of over 11 billion compounds. *Nature* **601**, 452-459 (2022).
21. Muller J*, et al.* Magnet for the Needle in Haystack: "Crystal Structure First" Fragment Hits Unlock Active Chemical Matter Using Targeted Exploration of Vast Chemical Spaces. *J Med Chem* **65**, 15663-15678 (2022).
22. Beroza P*, et al.* Chemical space docking enables large-scale structure-based virtual screening to discover ROCK1 kinase inhibitors. *Nat Commun* **13**, 6447 (2022).
23. Gorgulla C*, et al.* An open-source drug discovery platform enables ultra-large virtual screens. *Nature* **580**, 663-668 (2020).
24. Gadioli D*, et al.* EXSCALATE: An extreme-scale in-silico virtual screening platform to evaluate 1 trillion compounds in 60 hours on 81 PFLOPS supercomputers. *arXiv:211011644v1*, (2021).
25. Neumann A, Marrison L, Klein R. Relevance of the Trillion-Sized Chemical Space "eXplore" as a Source for Drug Discovery. *ACS Med Chem Lett* **14**, 466-472 (2023).
26. Hartenfeller M*, et al.* A collection of robust organic synthesis reactions for in silico molecule design. *J Chem Inf Model* **51**, 3093-3098 (2011).
27. Penner P*, et al.* FastGrow: on-the-fly growing and its application to DYRK1A. *J Comput Aided Mol Des* **36**, 639-651 (2022).

28. Sivula T, Yetukuri L, Kalliokoski T, Kasnanen H, Poso A, Pohner I. Machine Learning-Boosted Docking Enables the Efficient Structure-Based Virtual Screening of Giga-Scale Enumerated Chemical Libraries. *J Chem Inf Model* **63**, 5773-5783 (2023).
29. Roggia M, Natale B, Amendola G, Di Maro S, Cosconati S. Streamlining Large Chemical Library Docking with Artificial Intelligence: the PyRMD2Dock Approach. *J Chem Inf Model*, https://doi.org/10.1021/acs.jcim.1023c00647 (2023).
30. Gentile F*, et al.* Automated discovery of noncovalent inhibitors of SARS-CoV-2 main protease by consensus Deep Docking of 40 billion small molecules. *Chem Sci* **12**, 15960-15974 (2021).
31. Desaphy J, Bret G, Rognan D, Kellenberger E. sc-PDB: a 3D-database of ligandable binding sites–10 years on. *Nucleic Acids Res* **43**, D399-404 (2015).
32. Bostrom J, Brown DG, Young RJ, Keseru GM. Expanding the medicinal chemistry synthetic toolbox. *Nat Rev Drug Discov* **17**, 709-727 (2018).
33. Rarey M, Kramer B, Lengauer T, Klebe G. A fast flexible docking method using an incremental construction algorithm. *J Mol Biol* **261**, 470-489 (1996).
34. Jones G, Willett P, Glen RC, Leach AR, Taylor R. Development and validation of a genetic algorithm for flexible docking. *J Mol Biol* **267**, 727-748 (1997).
35. Korb O, Stutzle T, Exner TE. Empirical scoring functions for advanced protein-ligand docking with PLANTS. *J Chem Inf Model* **49**, 84-96 (2009).
36. Li J, Li C, Sun J, Palade V. RDPSOVina: the random drift particle swarm optimization for protein-ligand docking. *J Comput Aided Mol Des* **36**, 415-425 (2022).
37. Jain AN. Surflex-Dock 2.1: robust performance from ligand energetic modeling, ring flexibility, and knowledge-based search. *J Comput Aided Mol Des* **21**, 281-306 (2007).
38. Chachulski L, Windshugel B. LEADS-FRAG: A Benchmark Data Set for Assessment of Fragment Docking Performance. *J Chem Inf Model* **60**, 6544-6554 (2020).
39. Verdonk ML, Giangreco I, Hall RJ, Korb O, Mortenson PN, Murray CW. Docking performance of fragments and druglike compounds. *J Med Chem* **54**, 5422-5431 (2011).
40. Marcou G, Rognan D. Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. *J Chem Inf Model* **47**, 195-207 (2007).
41. Hartenfeller M*, et al.* DOGS: reaction-driven de novo design of bioactive compounds. *PLoS Comput Biol* **8**, e1002380 (2012).
42. Sommer K, Flachsenberg F, Rarey M. NAOMInext - Synthetically feasible fragment growing in a structure-based design context. *Eur J Med Chem* **163**, 747-762 (2019).
43. Moroz YS. 2022q3-4 REAL database reagents, personal communication. (ed^(eds) (2023).
44. Malamas MS*, et al.* Design and synthesis of aryl diphenolic azoles as potent and selective estrogen receptor-beta ligands. *J Med Chem* **47**, 5021-5040 (2004).
45. Da Silva F, Desaphy J, Rognan D. IChem: A Versatile Toolkit for Detecting, Comparing, and Predicting Protein-Ligand Interactions. *ChemMedChem* **13**, 507-510 (2018).

46. OpenEye Scientific Software, Sante Fe, NM, U.S.A. (ed^(eds).
47. Penner P, Guba W, Schmidt R, Meyder A, Stahl M, Rarey M. The Torsion Library: Semiautomated Improvement of Torsion Rules with SMARTScompare. *J Chem Inf Model* **62**, 1644-1653 (2022).
48. Schneider N, Lange G, Hindle S, Klein R, Rarey M. A consistent description of HYdrogen bond and DEhydration energies in protein-ligand complexes: methods behind the HYDE scoring function. *J Comput Aided Mol Des* **27**, 15-29 (2013).
49. Fischer A, Smiesko M, Sellner M, Lill MA. Decision Making in Structure-Based Drug Discovery: Visual Inspection of Docking Results. *J Med Chem* **64**, 2489-2500 (2021).
50. https://www.ebi.ac.uk/chembl/ (accessed 11-16-2023).
51. Chien EY*, et al.* Structure of the human dopamine D3 receptor in complex with a D2/D3 selective antagonist. *Science* **330**, 1091-1095 (2010).
52. Maramai S*, et al.* Dopamine D3 Receptor Antagonists as Potential Therapeutics for the Treatment of Neurological Diseases. *Front Neurosci-Switz* **10**, 451 (2016).
53. Hopkins AL, Groom CR, Alex A. Ligand efficiency: a useful metric for lead selection. *Drug Discov Today* **9**, 430-431 (2004).
54. CACHE challenge: Critical assessment of computational hit-finding experiments. https://cache-challenge.org/ , (accessed 11-16-2023)
55. sc-PDB: An Annotated Database of Druggable Binding Sites from the Protein DataBank. http://bioinfo-pharma.u-strasbg.fr/scPDB/, (accessed 11-16-2023)
56. Lewell XQ, Judd DB, Watson SP, Hann MM. RECAP–retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J Chem Inf Comput Sci* **38**, 511-522 (1998).
57. Clark M, Cramer RD, III., Van Opdenbosch N. Validation of the general purpose tripos 5.2 force field. *J Comput Chem* **10**, 982-1012 (1989).
58. ADFR software suite downloads, https://ccsb.scripps.edu/adfr/downloads/ (accessed 11-16-2023).
59. Enamine building blocks catalog. https://enamine.net/building-blocks/building-blocks-catalog (accessed 03-25-2023)
60. Dassault Systèmes Biovia Corp., San Diego, CA. https://www.3ds.com/products-services/biovia/products/data-science/pipeline-pilot/
61. Molecular Networks GmbH, Nürnberg, Germany. https://mn-am.com/products/corina/ (accessed 03-25-2023)
62. Pike AC*, et al.* Structure of the ligand-binding domain of oestrogen receptor beta in the presence of a partial agonist and a full antagonist. *EMBO J* **18**, 4608-4618 (1999).
63. Bietz S, Urbaczek S, Schulz B, Rarey M. Protoss: a holistic approach to predict tautomers and protonation states in protein-ligand complexes. *J Cheminform* **6**, 12 (2014).
64. Halgren TA. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J Comput Chem* **17**, 490-519 (1996).

65. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: An open chemical toolbox. *J Cheminform* **3**, 33 (2011).

Table

Table 2 is available in the Supplementary Files section.

Supplementary Information

Supplementary Figures and Supplementary Tables are not available with this version.

Figures

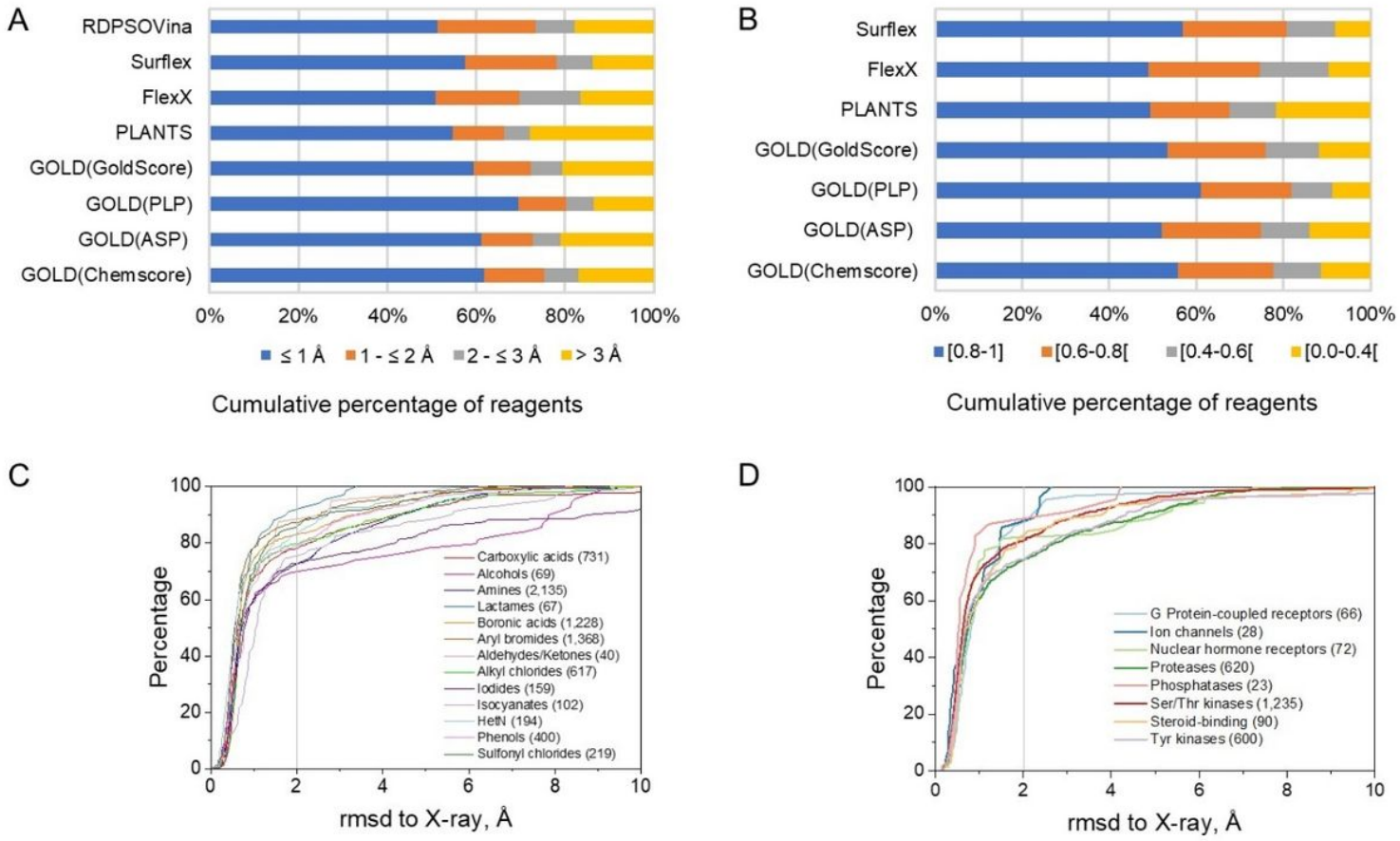
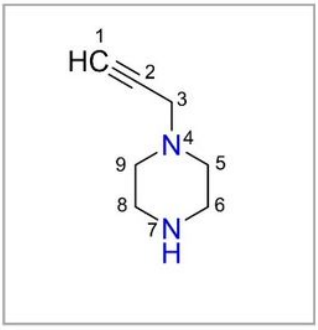


Figure 1

Accuracy of state-of-the-art docking tools to dock 5,845 sc-PDB reagents in their cognate targets. (A) Root-mean square deviation (rmsd) of the best pose (lowest rmsd, heavy atoms only) to the surrogate X-ray structure, **(B)** Similarity of protein-reagent interaction fingerprints between the best pose (highest interaction fingerprint similarity) and surrogate X-ray structures, measured by a Tanimoto coefficient. Fingerprints could not be measured for RDPSOVina poses in pdbqt format, **(C)**Cumulative rmsd of the

best pose (GOLD-PLP docking) for each of the 13 chemical functions. Numbers in brackets indicate the absolute number of each chemical function, **(D)** Cumulative rmsd of the best pose (GOLD-PLP docking) according to protein class. Numbers in brackets indicate the absolute number of samples from each protein family.



Reagent_ID	Tag_atom	Reaction	Reactant
EN300-27346	7	Amide	Amine
EN300-27346	1 2	Benzofuran	Alkyne
EN300-27346	1 2	Benzothiophene	Alkyne
EN300-27346	7	Buchwald	Amine
EN300-27346	2 1	Huisgen	Alkyne
EN300-27346	1 2	Indole	Alkyne
EN300-27346	7	SNAr	Amine
EN300-27346	1	Sonogashira	Alkyne
EN300-27346	7	Sulfonamide	Amine
EN300-27346	7	Thiourea	Amine
EN300-27346	7	Urea	Amine

Figure 2

Annotation of chemical reagents by reaction type, reactant role and reactive atoms.

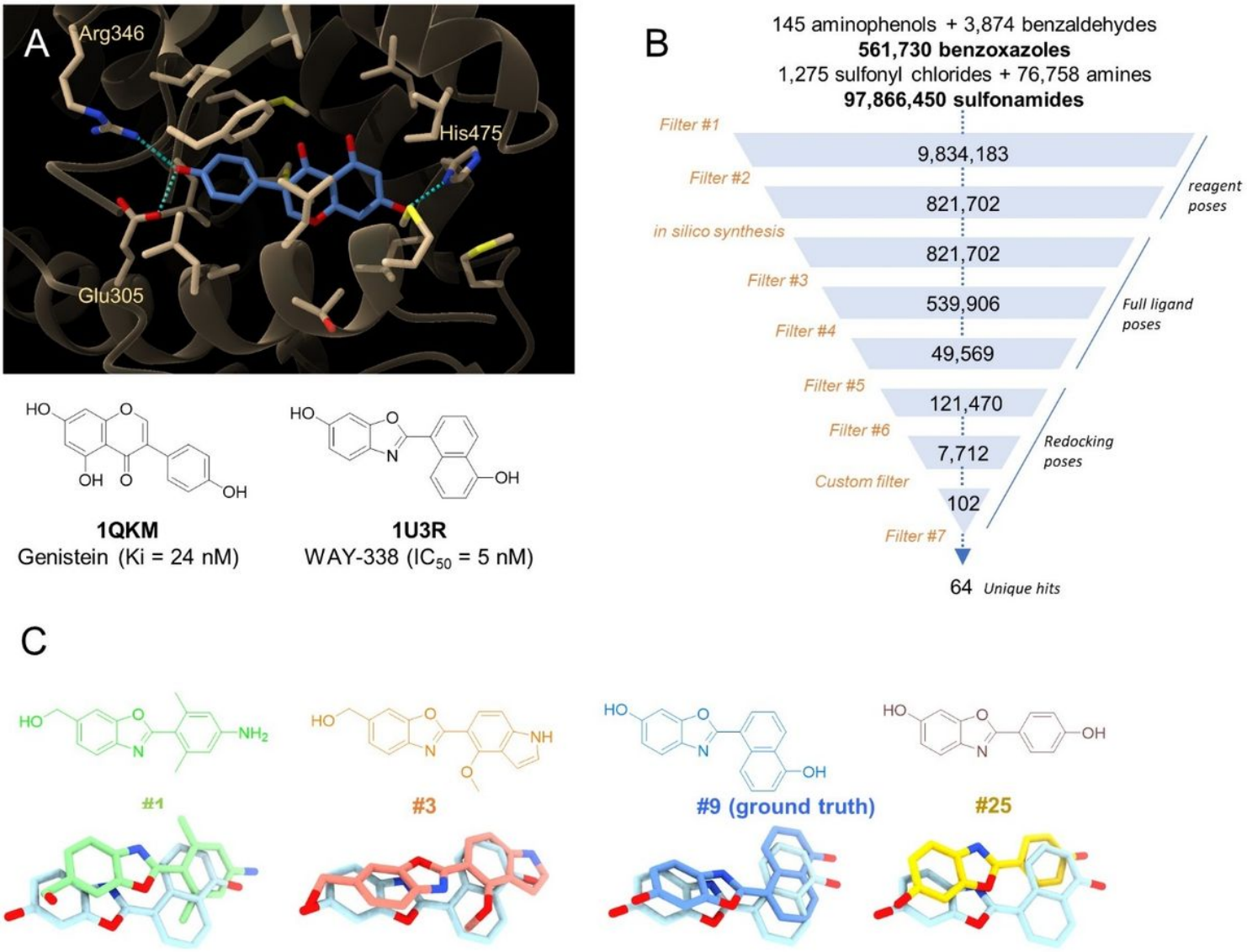


Figure 3

Space docking of benzoxazole and sulfonamide chemical spaces to the human estrogen receptor beta (ERβ). (A) X-ray structure of human ERβ (tan ribbons, PDB 1QKM) in complex with the agonist genistein (blue sticks). The genistein binding site is delimited by ERβ residues displayed as tan sticks with main receptor-ligand hydrogen-bonds indicated by cyan broken lines. The known benzoxazole agonist (WAY-338) is taken as the ground truth ligand to recover. **(B)** SpaceDock flowchart affording 64 potential ERβ agonists according to a series of filters (**Table 1**). The custom filter (H-bond either Glu305 or Arg346, and to His475) is target-specific. **(C)** Structures and rank (#) of 4 representative benzoxazoles. The proposed binding poses are overlayed to the X-ray pose of the ground truth ligand (WAY-338, cyan), the protein being masked for sake of clarity.

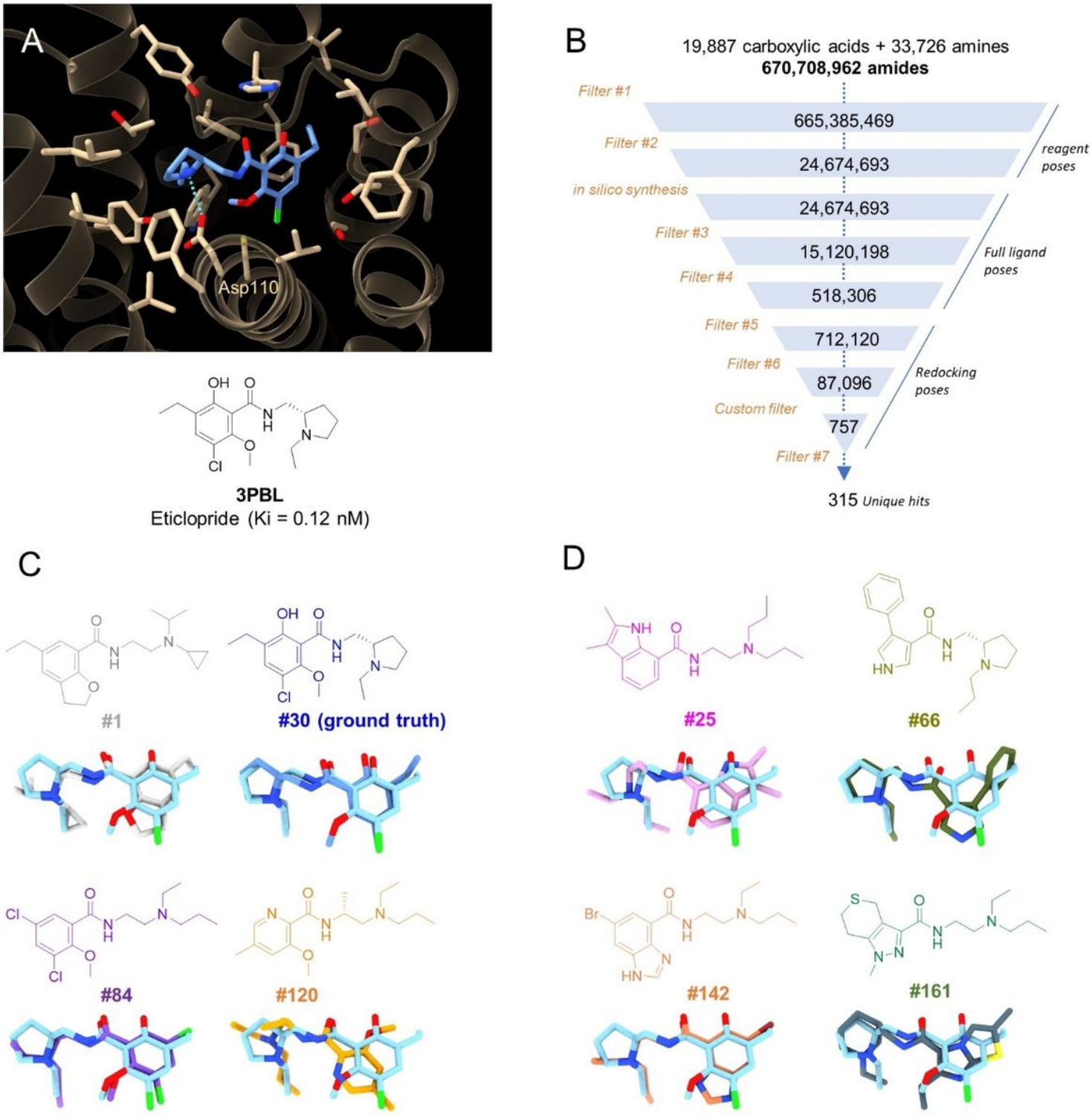


Figure 4

Space docking of an amides chemical space to the human dopamine D3 receptor (DRD3). (A) X-ray structure of human DRD3 (tan ribbons, PDB 3PBL) in complex with the antagonist eticlopride (blue sticks). The eticlopride binding site is delimited by DRD3 residues displayed as tan sticks with the main receptor-ligand ionic bond indicated by cyan broken lines. Eticlopride is taken as both the reference and the ground truth ligand to recover. **(B)**SpaceDock flowchart affording 315 potential DRD3 antagonists

according to a series of filters (**Table 1**). The custom filter (IFP similarity to eticlopride X-ray pose) is target-specific. **(C)** Structures and rank of 4 representative orthomethoxybenzamides. The proposed binding poses are overlayed to the X-ray pose of the ground truth ligand (eticlopride, cyan), the protein being masked for sake of clarity. **(D)** Structure and binding poses of other hits, aligned to the X-ray pose of eticlopride.

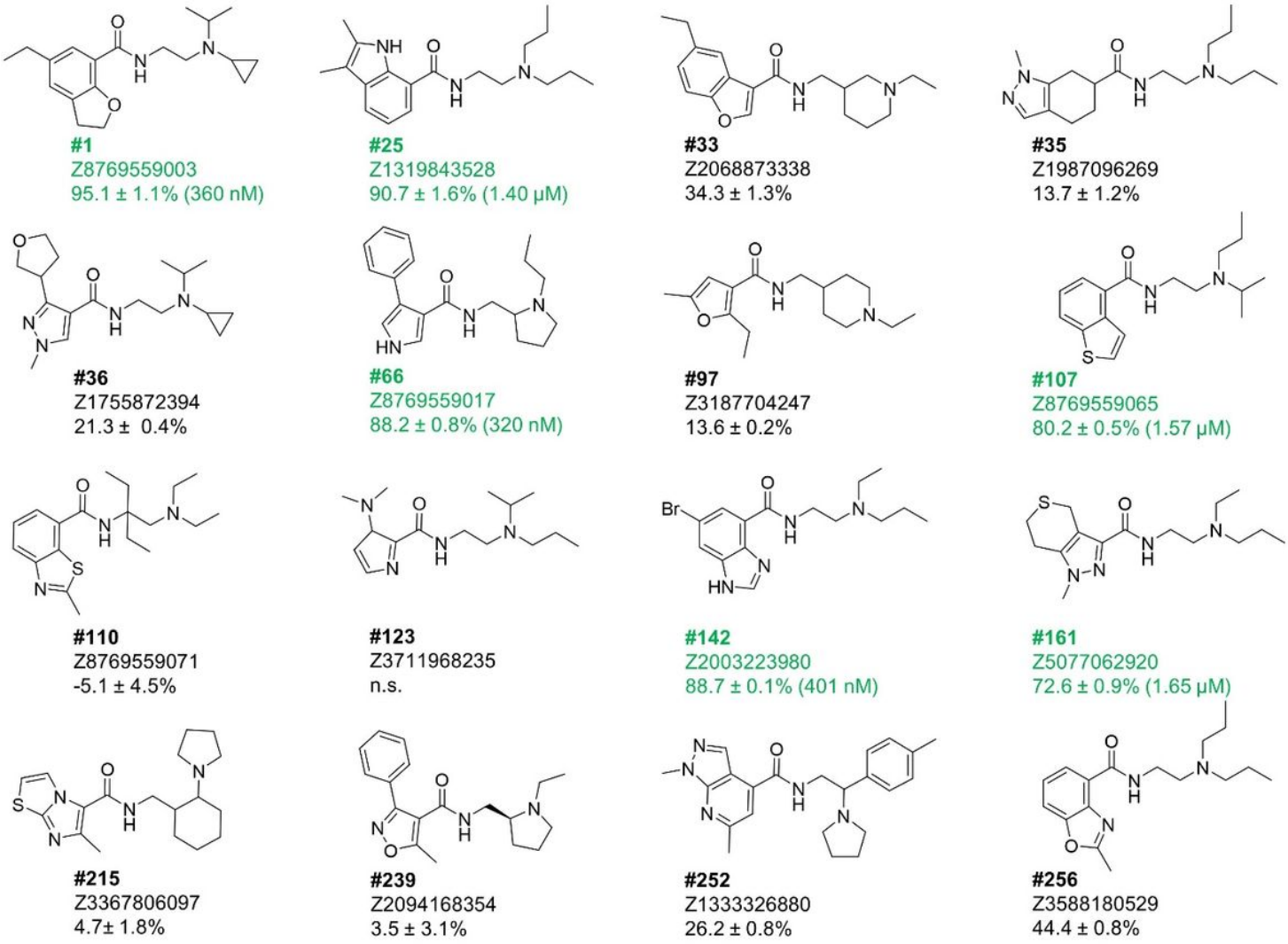


Figure 5

Structure and binding to human DRD3 of 15 SpaceDock hits from the amide space. Hits are labelled according to their SpaceDock rank, Enamine's catalog identifiers and purchased as racemates, unless specified. Binding affinities to human DRD3 are expressed as the percentage of inhibition of [^{3}H]-methylspiperone binding to human recombinant DRD3 expressed in CHO cells (Eurofins Discovery assay #48) at a single concentration of 10 µM competitor (mean of two independent experiments). The inhibition constant (K_i) was determined from dose-response curves for six strong binders (in green). Compound #123 could not be synthesized (n.s.)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Table2.docx