HHS Public Access

Author manuscript

Nat Chem Biol. Author manuscript; available in PMC 2024 March 01.

Published in final edited form as:

Nat Chem Biol. 2023 June ; 19(6): 712–718. doi:10.1038/s41589-022-01234-w.

Modeling the expansion of virtual screening libraries

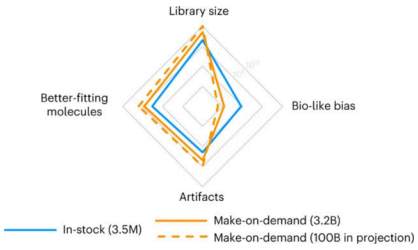
Jiankun Lyu[1], **John J. Irwin**[1,*], **Brian K. Shoichet**[1,*]

[1]Department of Pharmaceutical Chemistry, University of California, San Francisco, CA 94158, USA

Abstract

Recently, “tangible” virtual libraries have made billions of molecules readily available. Prioritizing their molecules for synthesis and testing demands computational approaches, like docking. Their success may depend on library diversity, its similarity to bio-like molecules, and how receptor-fit and artifacts change with library size. We compared a library of 3 million “in-stock” molecules with billion-plus tangible libraries. The bias toward bio-like molecules in the tangible library decreases 19,000-fold versus those “in-stock”. Similarly, thousands of high-ranking molecules, including experimental actives, from five ultra-large library docking campaigns are also dissimilar to bio-like molecules. Meanwhile, better fitting molecules are found as the library grows, with score improving log-linearly with library size. Finally, as library size increases, so too do rare molecules that rank artifactually well. Although the nature of these artifacts changes from target to target, their expectation of occurrence does not, and simple strategies can minimize their impact.

Graphical Abstract



*Corresponding authors: John Irwin: jir322@gmail.com and Brian Shoichet: bshoichet@gmail.com.

Author contributions

J.L. performed computational docking and chemoinformatic analysis, prepared figures and co-wrote the manuscript. J.J.I. developed docking libraries, edited the manuscript, and arranged funding. B.K.S. supervised the work, co-wrote the manuscript, and conceived the study with the other authors.

Code availability

DOCK3.8 is freely available for non-commercial research https://dock.compbio.ucsf.edu/DOCK3.8/. A web-based version is available at https://blaster.docking.org/. The tool to measure Tanimoto coefficient is freely accessible at https://github.com/docking-org/ChemInfTools.

Competing interests

B.K.S. is co-founder of BlueDolphin, LLC, a molecular docking contract research organization, Epiodyne, and Deep Apple Therapeutics, Inc., both drug discovery companies, has recently consulted for Umbra, Abbvie, and Dice Therapeutics, and is on the SAB of Schrodinger. J.J.I. co-founded Deep Apple Therapeutics, Inc. and BlueDolphin, LLC. J.L. declares no competing interests.

Introduction

Given over 10^{60} drug-like molecules estimated possible[1,2], screening of $10^{5.5}$ to $10^{6.5}$ random molecules, as in high-throughput screens, might in principle never work. A widely mooted explanation for why it has worked[3–5] is that screening decks are far from random, but are biased toward molecules that proteins have evolved to recognize: metabolites, natural products, and their mimicking drugs[6]—what we will call "bio-like" molecules. The implication of this idea, which we[6] and others[7–19] have promoted, is that as chemical libraries expand they should remain biased toward bio-like molecules. While popular, this idea has never been prospectively tested.

An opportunity to do so has come with the advent of ultra-large make-on-demand or "tangible" libraries[20]. These virtual libraries are composed of molecules that have not been previously made, but can readily be synthesized. Since 2016, these libraries have expanded accessible molecules from 3.5 million (available from in-stock collections) to over 29 billion[21]. While such libraries cannot be empirically screened, molecules within them can be computationally prioritized for synthesis and testing, often using molecular docking. Indeed, docking of these ultra-large tangible libraries has revealed highly potent molecules for multiple targets, with affinities often in the mid-nM and sometimes high picomolar range[22–26], results typically better than from docking the much smaller in-stock collections. If the idea that the success of compound screens reflects library bias toward bio-like molecules, one would expect these new ultra-large libraries, and the hits emerging from them, to share the biases toward metabolites, natural products and drugs observed in the "in-stock" libraries. Since we know the identity of every molecule screened in these libraries, this idea can be explicitly tested.

It also seemed interesting to consider what other factors in the tangible libraries have contributed to docking successes, and those of other library screening methods[27], and what challenges may be anticipated as the libraries continue to grow[28]. For instance, should we expect the fit of library molecules into their receptor targets to improve as the libraries grow, and if so at what rate? Do the high-ranking molecules from a library screen come to be dominated by a small number of chemotypes as libraries grow, or is diversity maintained? As the libraries grow, does the likelihood of artifacts[29] that exploit weaknesses in the docking scoring and sampling also grow?

Here we explore how similarity to bio-like molecules has changed with library growth, how goodness-of-fit and chemical diversity of high-ranking molecules changes with library growth, and how we might anticipate rare but high-ranking artifacts to change with library growth. Even at this early stage in the field, the results that emerge are strong enough to suggest strategies to maximize success in ultra-large library screening.

Results

Similarity to bio-like molecules changes with library size

An intriguing observation of HTS screening decks and of "in-stock" libraries was that they resembled bio-like molecules (metabolites, natural products, and drugs) over 1000-fold

than what would be expected at random[6]. To investigate how this similarity to bio-like molecules changed with library size, we compared the 3.5 million in-stock library and the 3.1 billion make-on-demand library to worldwide drugs, to metabolites, and to natural products ("bio-like" molecules). Using ECFP4 topological fingerprints, we calculated the Tanimoto similarity between each library molecule and each bio-like molecule. In this comparison, the Tanimoto coefficient (Tc) represents the features shared between two molecules (library and bio-like) divided by the total number of features. A Tc of 1 indicates the pair are identical while a score of 0.2 indicates that the similarity is low enough to be essentially meaningless. As seen previously[6], the in-stock set is far more similar to bio-like molecules than expected at random (Fig. 1a, blue curve), with 10,000 in-stock molecules being identical to metabolites, natural products, or drugs (Fig. 1b). Conversely, as the library grows 886-fold from 3.5 million in-stock to 3 billion tangible molecules, the number with Tc values >0.8 to bio-like molecules actually decreases 2.3-fold despite a library that is three orders-of-magnitude larger (Fig. 1a, orange curve). Most of the growth of the tangible library comes in the random similarity region compared to bio-like molecules, where the peak is around a Tc of 0.25; in this region the tangible library grows by 3,000-fold versus the in-stock. Between the two extremes of random similarity and full identity, the similarity to bio-like molecules falls much faster for the 3 billion tangible library (Fig. 1a, orange curve) than it does for the 3.5 million in-stock library (Fig. 1a, blue curve). By the time essentially full identity ($0.95 < Tc \leq 1.0$) with bio-like molecules is reached, only 0.000022% (700 molecules) of the make-on-demand library qualify, whereas 0.42% of the "in-stock" molecules do so, a 19,000-fold decrease. Thus, though docking campaigns with the new ultra-large libraries have returned potent molecules with high hit rates[22–26], the new libraries do not retain the strong bias to bio-like molecules that was a feature of both in-stock and HTS libraries[6].

Of course, it could be that the actual docking hits nevertheless resemble bio-like molecules even though the overall library does not. Accordingly, we plotted the similarity to bio-like molecules of large-library docking hits from five targets, including three G-protein couple receptors (GPCRs)[30,31] and two enzymes: the D_4 dopamine receptor[23], AmpC β-lactamase[24], the melatonin receptor[26], the σ_2 receptor[22] and the Nsp3 Macrodomain[32] from SARS-CoV-2 (Fig. 1c and Extended Data Fig. 1). In all five campaigns, the docking-prioritized molecules had Tc values < 0.6 to bio-like molecules, peaking at Tc values of 0.3 to 0.35, similarity values not much different than expected for pairs of random molecules. There was little difference in the distribution of molecules selected for synthesis and testing (orange bars, Fig. 1c and Extended Data Fig. 1) and the subset of those that were found to be active on target on experimental testing (blue bars, Fig. 1c and Extended Data Fig. 1)

While similarity to bio-like molecules confers little benefit in docking hit rate, it might improve success later in drug discovery. Several investigators[33] have noted that natural products, for instance, are more likely to be transporter substrates, improving permeability and exposure. "Transportability" is hard to calculate, but several proxies may be used to calculate cell and organ permeability, including calculated octanol-water partition coefficient (cLogP), topological polar surface area (tPSA), numbers of rotatable bonds, and formal charge. By these criteria, the "in-stock" and tangible (make-on-demand) libraries differ little (Extended Data Figs. 2a—d). Even if we only compare the 61,179 bio-like molecules

"in-stock" to the tangible library, the same conclusion emerges (Extended Data Figs. 3a—d). We can extend this analysis to violations of Lipinski's rule-of-five (Ro5)[34] and Jorgensen's rule-of-three (QPPCaco >22, LogS >-5.7, potential metabolism sites < 7)[35]. We calculated these properties for the 61,179 bio-like molecules "in-stock" and compared them to the same number drawn 30 times from the lead-like tangible molecules (Extended Data Fig. 3e and 3f). There were actually fewer violations among the tangible molecules than among the bio-like. Naturally, this partly reflects the intentional lead-like[36] character of the tangible molecules, reducing Ro5 violations, but since this is the set being docked it remains meaningful. Finally, where molecules deriving from ultra-large library docking have been tested in vivo they have had favorable plasma and brain exposure on intraperitoneal and even oral dosing[22,26,37,38]. Thus, while we cannot rule out an advantage for bio-like molecules, the physical properties of the tangible molecules put them at no obvious disadvantage to them.

Docking score improves with library size

Whether more and more favorable molecules are found as the library grows will govern how far we should expand the tangible libraries. Ideally we would like to know how the affinities and hit-rates of docked molecules improves with library size, but determining this would be an expensive undertaking. As a proxy, we can ask how docking score improves with library size. While docking score, with its errors and approximations, may be a weak link to likelihood of binding, we have found that it correlates with hit rate in two systems, the D_4 dopamine[23] and the σ_2 receptors[22], and it is the primary criterion by which molecules are selected in docking screens.

We docked ever larger libraries against the D_4, σ_2 and 5HT2A receptors, looking for how docking score change with library size. We first docked 344 million, 1.4 billion, and 1.7 billion molecules against the three receptors, respectively—and from this largest set picked ever larger subsets of the library at random 30 times, with subset size increasing by half-logs from 10^5 to over 10^9 molecules. For each subset, the scores and scaffolds of the top 5000 ranking molecules, divided into quartiles, and the number of molecules with scores better than a certain threshold were measured.

As the subsets grow from 10^5 to over 10^9 molecules, the scores of the top ranking 5000 molecules monotonically improved for all three targets (Fig. 2a). This improvement was roughly log-linear for all quartiles among the 5000, excluding the very top scoring molecule where it increases faster (but see below), and does not seem to saturate with library size. While the curves appear to have some negative curvature, this mostly reflects larger improvements in score from the smallest docking libraries; above 1 million molecules the rate of change appears steady for each log increase in library size. In short, as the library enlarges, the fit of the top-ranking docked molecules steadily improves without signs of saturation, at least on the log scale.

The improvement of the docking scores could reflect new scaffolds appearing in the library as it grows, or it could reflect the optimization of analogs of molecules already present. To investigate this, we analyzed the top 5000 molecules in each library subset for Bemis-Murcko scaffolds[39] (Fig. 2b). The scaffolds can be divided into two categories: singletons

without analogs, and scaffolds for which analogs exist. Plotting the score variations among singleton scaffolds, analogs in a group scaffold, and all top-ranking 5000 molecules, we observe that both singletons and analog clusters both contribute to the improvement of docking scores as library grow (Extended Data Fig. 4). While the proportion of analogs in the top 5000 increases with library size, molecules from both categories contribute to score improvements up to the billion molecule range (Extended Data Fig. 5).

One can also ask how the number of molecules with scores favorable enough that they are likely to bind experimentally changes with library size. Ordinarily this is difficult owing to the approximations and errors in docking but, at least for the D_4 and σ_2 receptors, the variation of hit rate with docking score has been measured experimentally by testing about 500 molecules from across the docking scoring range[22,23] (this has not been done for the 5HT2A receptor, which was thus excluded from this analysis). For both targets, this revealed a sigmoidal curve with a high-hit rate plateau; molecules that score in this plateau have a high likelihood of binding. For the D_4 and σ_2 receptor, the plateaus are defined by scores of ≤ -60 and ≤ -55 DOCK scoring units, respectively[22,23]. Both the number of molecules and the number of scaffolds in this favorable scoring region increases with library size (Fig. 2c and Fig. 2d), indicating not only molecules that better fit the site are found, but also more types of such molecules are found with library growth (Fig. 3).

Artifacts increase with library size.

An exception to the log-linear improvement of docking scores may be observed for the very best molecules from the screens (Fig. 2a, blue curves). The score of these molecules shows positive curvature with library growth, and especially in the larger libraries diverges from the other top 5000 ranking molecules. On inspection, these are not molecules that fit the receptor uniquely well, but rather molecules that cheat the scoring function by exploiting its holes and approximations. For instance, for the D_4 receptor these are molecules are conformationally strained[40], for the σ_2 receptor they are molecules have artifactually low desolvation penalties (and so too favorable scores)[22], and for the σ_2 and 5HT2A receptor they are molecules with artifactual atomic partial charges and with wrong tautomers. As the libraries grow, so too does the number of these artifactual hits, and by the time we dock 1.3 billion molecules against σ_2, over 98% of the top 100 ranking molecules have incorrect tautomerization. Meanwhile, beyond the top 100,000 docking hits these artifacts drops almost disappear—their biggest impact is in a thin slice of the top-ranking docking molecules (we distinguish between these artifacts that exploit a hole in the scoring function, and are rare, from molecules whose scores are too favorable owing to scoring function approximations, and are within some error range of what their true scores should be. A key feature of the rare artifacts is that they crowd the top of the docking scoring list; the more common decoys are more evenly spread throughout). Still, if one picked molecules exclusively from among the very top-ranking molecules, and was limited to a fixed number of them, it could easily be true that the prioritized molecules could come to be dominated by artifacts.

Naturally, one solution to these artifactual “cheating” molecules is to fix the holes in the docking scoring function. Certainly, once one finds a particular artifact one can address it.

However, two characteristics of these artifacts may make this difficult in general. First, they are rare events; if they were more common, they would be discovered by the retrospective control screens that are commonly conducted before a large prospective screen[41]. Second, they can change from target to target. For instance, in the campaign against the dopamine D4 receptor it was conformationally strained molecules that contributed most to these artifacts, for the σ_2 receptor it was molecules with artifactually low desolvation penalties[22], wrong tautomers, and artifactual partial atomic charges, the latter two of which also characterized the top-ranking molecules for 5HT2A. Other targets may reveal still other artifacts. As rare molecules in a multi-billion molecule library, these may be hard to anticipate.

One may, however, imagine a general strategy, free of any particular aspect of the docking scoring function, to treat the problem of rare artifacts. In doing so, it is important to consider two of their features: first, they are rare events that rise to the top, and second, especially for large library screens, there can be hundreds-of-thousands of molecules that score within the plateau region where molecules may be likely to bind. For instance, in a simplifying example, assume that these rare-event artifact occur at a rate of 0.001% of the molecules docked. In this case, the number of “cheating” artifacts will increase from 10 to 10,000 as the library grows from one million to one billion molecules. Usually, one can only afford to synthesize and test a fixed number of top-ranking compounds. If that number is 100 molecules, picked from the very top-ranking docked molecules, then in docking a million molecule library the cheating artifacts will only account for 10% of the molecules tested, but docking a billion molecule library they will amount to 10,000 molecules, completely dominating the top 100 ranked molecules.

More generally, we can model how the number of rare-event, cheating molecules will grow with library size, using a statistical distribution of these molecules versus the rest of the library, and considering different rates of occurrence. We simulate the distribution of these artifacts using both an extreme value distribution and a uniform distribution, while using a normal distribution for other library molecules. From these two distributions, we can estimate the effect of varying the artifact-to-library-molecule ratio with growing library size. Performance is evaluated by the percentage of artifacts in the top N-ranked molecules. With either distribution, artifacts begin to dominate the top-ranking list as the library grows for a given artifact-to-library-molecule ratio (Fig. 4a). If we cannot afford to synthesize and test more than a few hundred top-ranking molecules, the campaign will inevitably begin to falter as libraries rise toward 1 billion molecules.

A general solution to this problem is simply not to prioritize the several hundred molecules to be synthesized and tested exclusively from the very top-ranked molecules. Recall that a broad range of high-ranking docked molecules—a range that grows with library size—may have roughly equal likelihood of binding, and in a docking screen of a billion molecules, the top million might have scores that differ little from each other. To explore the impact of such a rank-spreading strategy on rare-event artifacts, we defined as five rank ranges the top 1–100 molecules, the top 101–1,000 molecules, the top 1,001–10,000 molecules, the top 10,001–100,000 molecules, and the top 100,001–1,000,000 molecules, picking 20 molecules from each. We plotted the percentage of rare-event artifacts among the 100 molecules picked in this rank-spreading strategy versus the same percentage among simply the top-100

molecules, as a function of library size (Fig 4b). For a given library size, the percentage of artifacts in the rank-spreading strategy was always lower than picking them exclusively from the top 100 molecules; for larger libraries, this strategy rank-spreading decreased the number of artifacts from 100% to between 25 to 50%. Naturally, there may be other strategies that will achieve the same goal, including rescoring the top-ranked molecules with another scoring function that, while it may also suffer from rare artifacts, may not suffer from the same ones. Even here, a strategy of picking from across the high-ranking ranges may have benefit.

Discussion

Since 2016, readily accessible molecular libraries for virtual screening have increased from 3.5 million to over 29 billion compounds. Our ability to prioritize from this vast chemical space depends on the molecules it explores and the ability of computational methods, often docking, to prioritize true ligands from an ocean of decoys. Three main observations from this study begin to illuminate the molecules that the new libraries explore, and how docking prioritizes them. First, the billion-plus tangible library is 19,000-fold less biased toward bio-like molecules than is the 3.5 million in-stock library. Second, as the libraries grow, better fitting molecules are found. The improvement in docking score is log-linear with library size and does not yet appear to saturate. Third, as the libraries grow so too do rare-event artifacts. While these are inconsequential for smaller, million molecule libraries, by the time the libraries grow to a billion molecules they can dominate hit lists. A general strategy of spreading docking picks from the docking rank curve suggests a way to overcome what might be a general problem.

Not only is the 3.1 billion make-on-demand library 19,000-fold less biased toward bio-like molecules than is the 3.5 million in-stock library (Fig. 1a), but thousands of experimentally tested high-ranking molecules from five docking campaigns are also dissimilar to bio-like molecules (Fig. 1b). This contradicts the idea, which we[6] and others[7–19] have advocated, that biasing a library toward metabolites, natural products, and drugs increases the chance of success in screening. Instead, the tangible library is little more similar to these bio-like molecules than one would expect at random, and diverges further from them as it grows. The “in-stock”, bio-like and tangible molecules have similar distributions of cLogP, tPSA and rotatable bonds (Extended Data Figs. 2a—c and 3a—c); and the tangible molecules are, if anything, more Ro5 and ADMET rules than is the bio-like set. To the extent that these physical properties contribute to success in subsequent compound optimization, they differ little between the two sets of molecules. Rather than biasing toward bio-like molecules —which may be simply a historical feature of the “in-stock” libraries and HTS decks[6]— the tangible libraries are defined by the over 200,000 intentionally diverse and stereogenic building blocks from which they are synthesized. The emphasis on the exploration of a wide range of chemotypes with high three-dimensionality ensures a diverse collection of functionalities and shape, and it may be this feature, rather than similarity to precedented molecules, that drives the better receptor fits of molecules from these libraries.

The exploration of stereogenic, functionally congested molecules ensures that as the library grows[42] more and more molecules are sampled that well-complement receptor sites.

Docking diverse libraries leads to a long-tail of high-scoring molecules, separated from the more normal distribution of docking scores from the library. By docking a library that is 1000-fold larger than the “in-stock” libraries that until recently dominated the field, we are essentially extending and filling in this long tail, such that it is populated with statistically relevant sampling of chemotypes (Fig. 3). As the libraries grow, docking scores improve log-linearly, and show no sign of saturation. Interestingly, a similar trend using ligand-based virtual screening has been previously reported; here too, best scores, in this case measuring 3D similarity to known ligands, improves log-linearly up to 10 billion molecules[27]. For the docked molecules, improved scores derive from both new chemotypes fortuitously appearing in the libraries, and from analogs of previously explored scaffolds that optimize fit (Extended Data Fig. 4). An inference from these trends is that, at least for now, screening larger and larger libraries will continue to improve docking results—with one caveat—and the hits that emerge will often have analogs to support early optimization.

Counterbalancing the improvement in docking fits with library expansion is the growth in the raw number of rare-event artifacts. If the number of molecules one could synthesize and test scaled with library size this wouldn’t be a problem. But with resources to only do so for an essentially fixed number of molecules, these rare, high-ranking artifacts will eventually overwhelm the true positives (Fig. 4a). This outcome can be alleviated by a strategy that not only tests the very top-ranking molecules, but also selects ones from slightly lower ranks that remain high scoring (Fig. 4b). We suspect that such rare-event artifacts will occur in most types of library screens, including HTS, DELs, or even genetic screens, and will become more pernicious with library size. Variations of this strategy may also be useful in these other areas.

Certain caveats bear airing. While bio-like molecules confer no advantage in docking hit-rate, nor in physical properties, they may have advantages not directly assessed here, including being transporter substrates[33]. Thus far, where molecules from ultra-large library docking have been tested in vivo they have had favorable plasma and brain exposure on intraperitoneal and even oral dosing[22,26,37,38], but this remains a small set of experiments. Mechanically, the divergence of the tangible libraries from bio-like molecules has only been measured by one type of topological similarity, other metrics may show different levels of divergence. We suspect that while this may affect the results quantitatively, qualitatively the story will remain; certainly, the number of molecules that are identical to metabolites, natural products, and drugs will not change. Apropos of the cheating artifacts, we would reemphasize that these are rare molecules that score well by finding holes in the scoring function. They are not the general run of molecules that are evaluated properly but with enough error that they rank too highly—these docking continues to struggle with, and the strategies suggested here will not address them. Finally, other, more quantitative approaches can be considered to solving the problem of rare artifacts, including rescoring to identify molecules exploiting holes in one scoring function that another function does not share.

The key observations of this work should not be obscured by these caveats. Virtual libraries are growing into a chemical space that is far less similar to bio-like molecules than are in-stock libraries. Despite this, multiple screens of these billion-molecule libraries have returned potent ligands with high success rate[22–26], suggesting that bias toward

precedented molecules might never have explained the success of large library screens. Indeed, simulation of docking performance with library size shows that we are still in the domain where ever-larger, diverse, stereogenic libraries will continue to fortuitously explore molecules with better and better fit for a target binding site. Strategies to avoid rare-event artifacts will help ensure that docking and related techniques can continue to prioritize from this growing chemical space, finding ever-more interesting molecules.

Methods

Bio-like libraries.

We used two sets of molecules from the ZINC15 database (https://zinc15.docking.org) to approximate the chemical space of bio-like molecules: worldwide drug set (https://zinc15.docking.org/substances/subsets/world/) and biogenic set (https://zinc15.docking.org/substances/subsets/biogenic/). The worldwide drug set contains 5,900 compounds and the biogenic set contains 168,185.

Screening libraries.

The in-stock and make-on-demand libraries were used in the analysis of bio-like bias. Molecules from both libraries are within the lead-like range: cLogP ≤ 3.5 and heavy atom count ≤ 25. The in-stock library contains 3,539,537 molecules. For bio-like bias quantification and physical property calculations, make-on-demand libraries contain 3,164,844,749 and 4,941,080,527 molecules at that time, respectively.

Quantifying bio-like bias.

Each molecule of the in-stock and make-on-demand libraries was in turn compared to each molecule of the bio-like library. Compounds were represented by their ChemAxon ECFP4 fingerprints (https://chemaxon.com/). The length of the fingerprints is 1,024 bits. The similarity was calculated by comparing their respective ECFP4 fingerprints with the Tanimoto coefficient. The tool to measure this similarity was deposited at https://github.com/docking-org/ChemInfTools. Related figures were made by GraphPad Prism 9.4.

Physical property calculations.

cLogP, tPSA and number of rotatable bonds for each molecule from the in-stock bio-like, in-stock and make-on-demand libraries were calculated by the RDKit v2020.09.1.0 package (https://www.rdkit.org). The net charge of 61,179 in-stock bio-likes were predicted at pH 7.4 by the majormicrospecies modular in ChemAxon Jchem 21.13 (https://chemaxon.com/). The net charge of in-stock and make-on-demand molecules were pre-calculated in ZINC22. Details can be found at https://cartblanche22.docking.org/tranches/3d. To evaluate number of violations on Lipinski's rule of five and Jorgensen's rule of three, 61,179 molecules were randomly picked 30 times from the lead-like make-on-demand library (https://cartblanche22.docking.org/search/random). These two metrics were calculated by QikProp from the 2022–1 released Schrödinger suite. Related figures were made by GraphPad Prism 9.4.

Molecular docking.

The docking setups of D_4 and σ_2 campaigns were reported previously[22,23]. The 5HT2A receptor with a ligand (unpublished) was used in the docking calculation. This unpublished structure is in an active state and is similar to the published 5HT2A active structure (PDB ID: 6WHA) with a low Cα RMSD of 0.8Å. The atoms of the Lisuride were used as the matching sphere calculation in the orthosteric site. The spheres were labeled[43] according to the charge-charge interaction and hydrogen-bond patterns of the Lisuride ligand in the cryo-EM structure. Forty-five spheres were used in total and were grouped into 7 clusters based on their spatial locations in the binding site by k-means clustering method. These labelled and clustered spheres were used to improve searching efficiency for speeding up docking calculations. The complex structure was protonated by Epik and PROPKA at pH 7.0 in Maestro (2021 release). Partial charges of residue atoms were assigned based on AMBER united atom types. The volume of the low dielectric and the desolvation volume was extended out from the surface of the receptor by 1.1 Å and 0.5 Å, respectively. Docking energy grids were pre-calculated with QNIFFT[44] for Poisson-Boltzmann-based electrostatic potentials, AMBER force fields using CHEMGRID for van der Waals potentials[45] and SOLVMAP[46] for ligand desolvation.

Since the 5HT2A receptor structure used in this study is in active state, the docking setup was evaluated for its ability to enrich known 5HT2A agonists over property-matched decoys. Decoys are molecules with dissimilar topology but share similar physical properties to known ligands, so they are unlikely to bind to the receptor. Forty-seven known 5HT2A agonists were extracted from the IUPHAR[47] and ChEMBL[48] databases and 2,050 property-matched decoys were generated by the DUD-E pipeline[49]. Docking performance was judged by the ability to enrich the 5HT2A known agonists over the decoys based on docking rank, using logAUC values. The docking setup described above achieved a logAUC value of 5. An ‘extrema’ set[49] of 146,620 were constructed through the DUDE-Z web server (http://tldr.docking.org) to make sure that molecules with extreme charge properties were not prioritized. This docking setup enriched over 97% mono-cations among the top1000 ranking molecules with a high logAUC value of 27. A small ‘goldilocks’ set< ($2 < cLogP \leq 3$ and $300\ Da < molecular\ weight \leq 350\ Da$) of 1,161,497 were also downloaded from the DUDE-Z web server (http://tldr.docking.org) to check if 5HT2A known agonists remain among the highest scored compounds. The docking setup achieved a decent logAUC value of 25 in this control experiment.

Using DOCK3.8, over 344 million, over 1.3 billion and over 1.6 billion library molecules from ZINC20/ZINC22 (http://zinc20.docking.org and https://cartblanche22.docking.org) were docked against the D_4, σ_2 and 5HT2A receptor, respectively. Each library molecule was sampled in about 2761, 3409 and 713 orientations and, on average, 174, 235 and 350 conformations for the D_4, σ_2 and 5HT2A receptor, respectively. The total calculation time was 70,705, 740,030 and 680,653 hours for the D_4, σ_2 and 5HT2A receptor, respectively.

Simulating docking performance with library size.

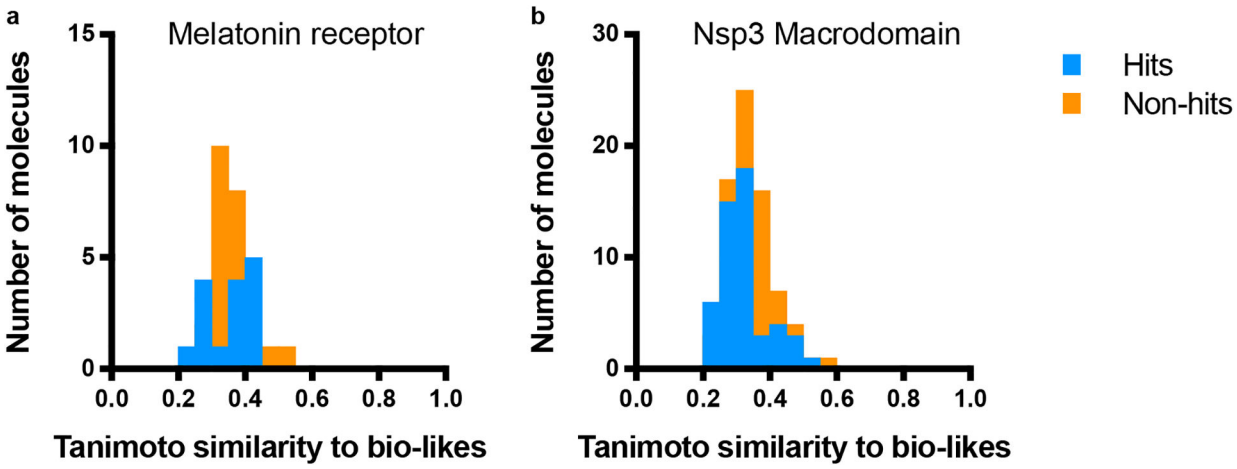
To investigate how dock scores of top 5000, number of chemotypes in top5000 and number of molecules below given dock score cutoff change with library size, we docked the full

make-on-demand library against the D_4, σ_2 and 5HT2A receptors with docking setups described above. To evaluate the effects of library size on the three metrics mentioned above, 10^5, 3×10^5, 10^6, 3×10^6, 10^7, 3×10^7, 10^8, 3×10^8, 10^9 (if possible) sets of molecules were randomly picked from the full docking-ranked list and the three metrics above were measured. Each set was selected thirty times with random selection from the full library. Chemotypes here are defined by Bemis-Murcho scaffold analysis[39]. The program mitools version 2020.04.4 (https://www.molinspiration.com/) was used to calculate scaffolds for this analysis. Related figures were made by GraphPad Prism 9.4.

Toy model for artifacts with library size.

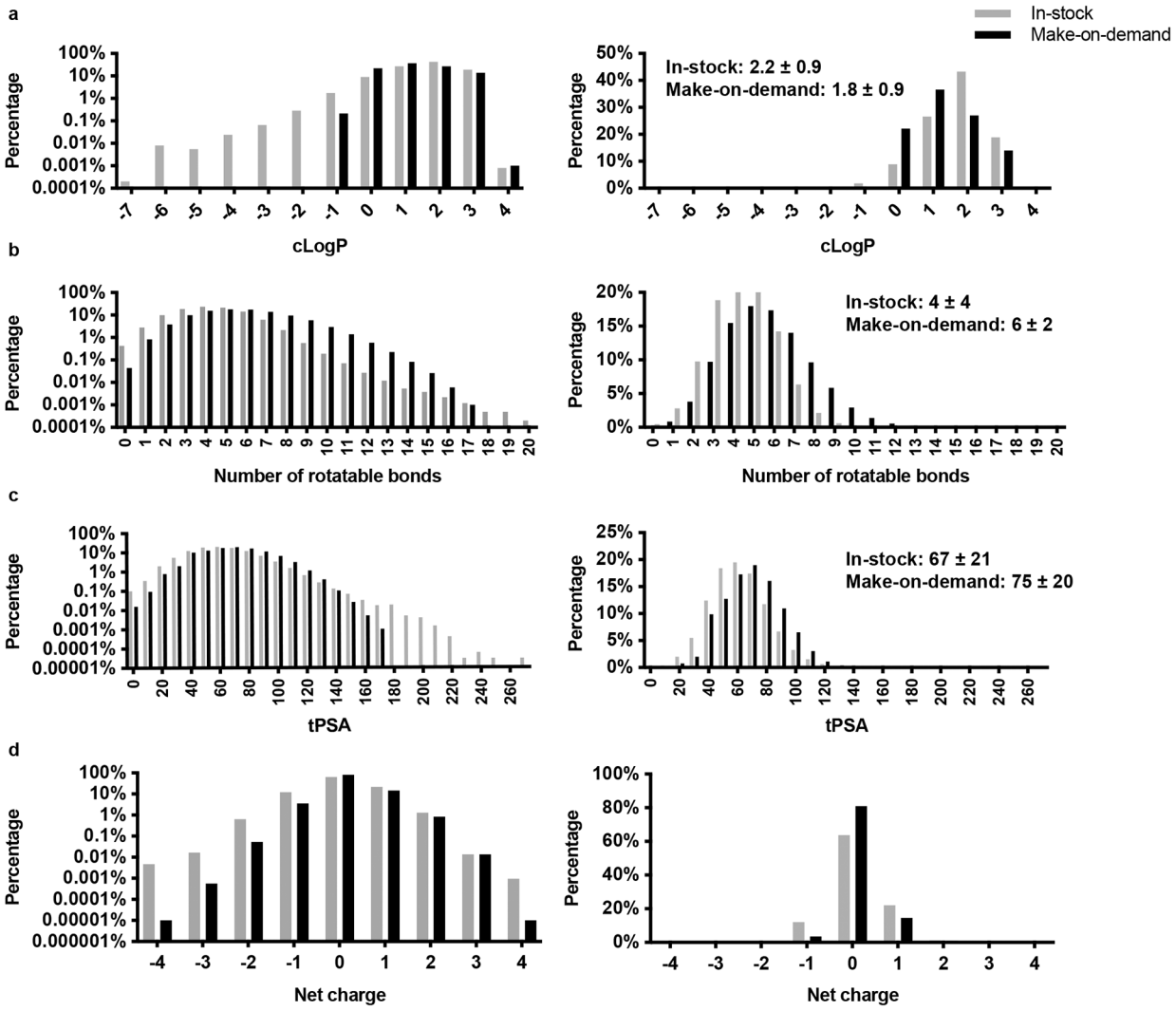
We constructed a model of how artifacts change with library size. Inputs to the model are the artifact-to-library-molecule ratio and the library size. The distributions are sampled from these two parameters 20 times independently for given artifact-to-library-molecule ratio and library size. We use the extreme value distribution (the shape parameter c = −0.1) or uniform distribution (the parameters loc = −10 and scale = −20) to sample artifacts while normal distribution (the parameter loc = 5 and 0, respectively) for library molecules. The percentage of artifactual molecules were calculated for the top 100, top 1,000, top 10,000 and top 100,000. Related figures were made by GraphPad Prism 9.4.

Extended Data



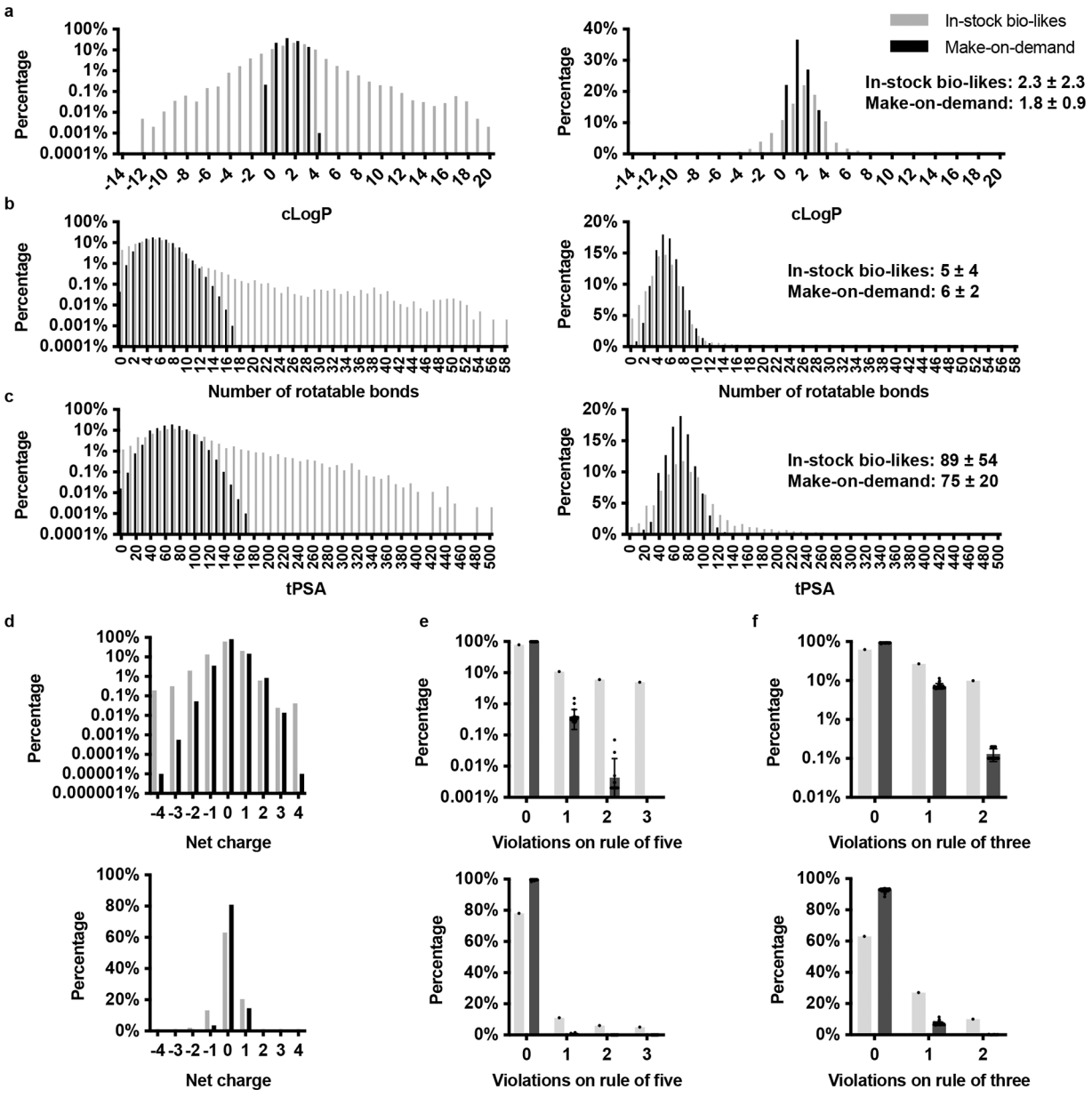
Extended Data Fig. 1.

The distribution of docking-prioritized and experimentally active (blue) and non-active (orange) molecules from two different docking campaigns as a function of the Tanimoto similarity to their nearest neighbor in the bio-like molecule set.



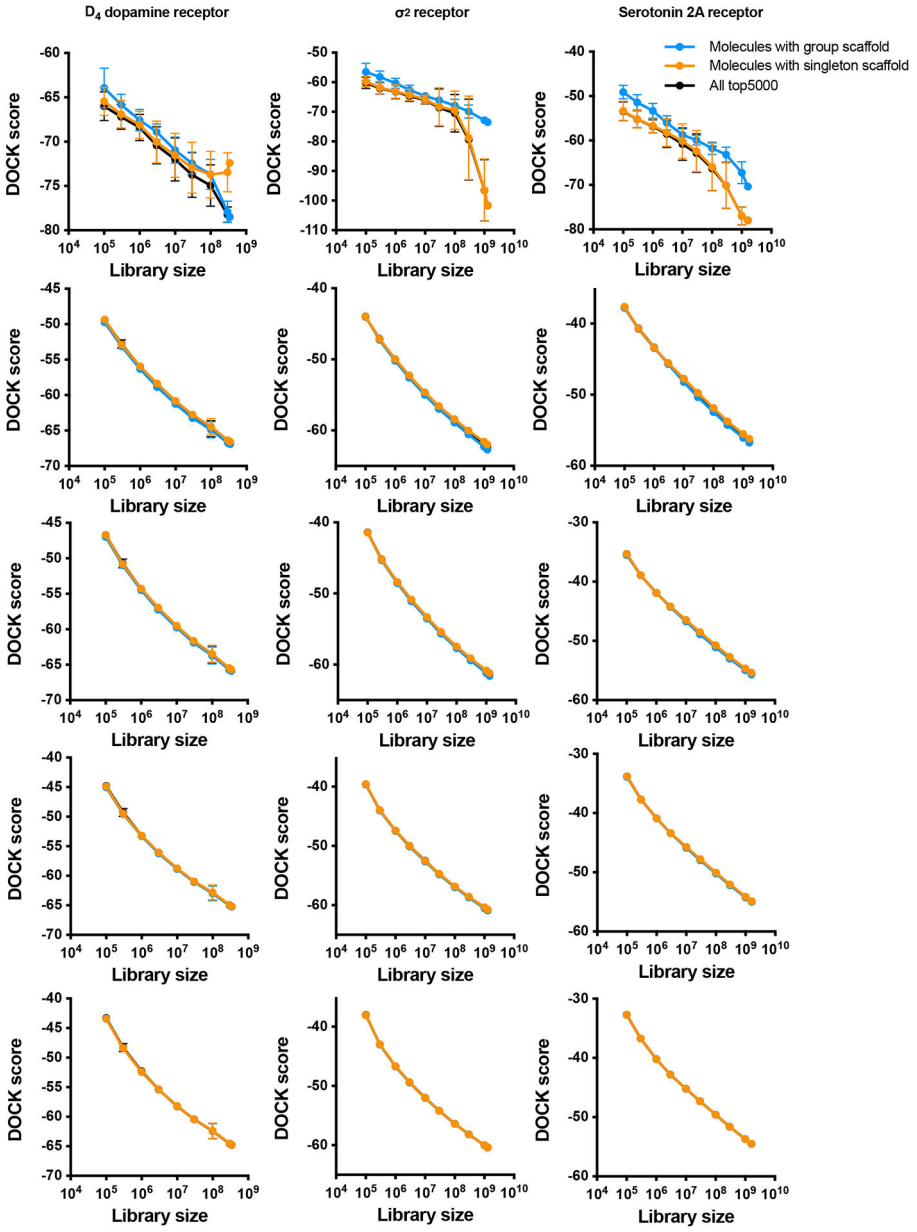
Extended Data Fig. 2.

The distribution of in-stock (grey) and make-on-demand (black) libraries as a function of physical properties.



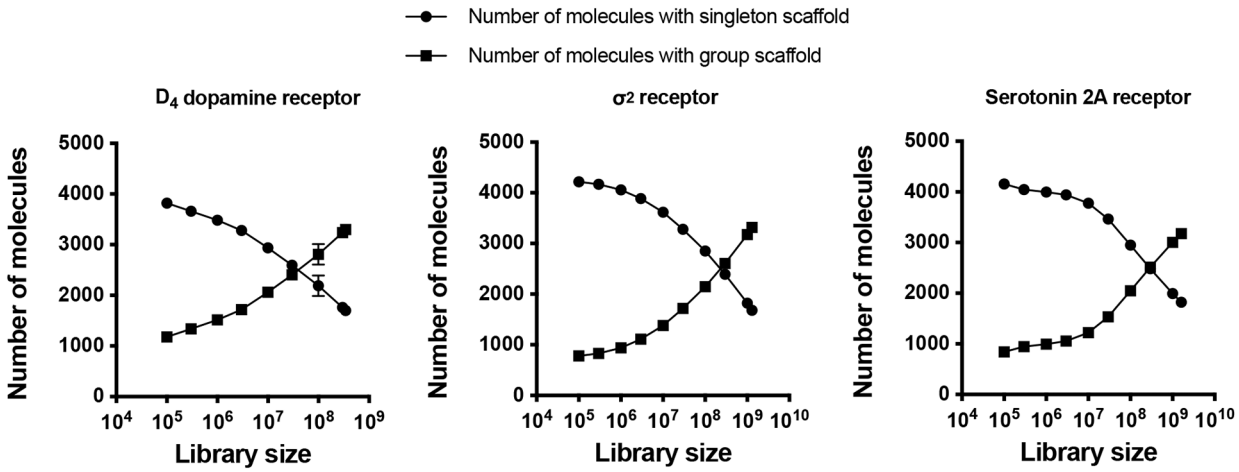
Extended Data Fig. 3.

The distribution of in-stock bio-like molecules (grey) and lead-like make-on-demand molecules (black) as a function of physical properties.



Extended Data Fig. 4.

Variation of score of top 5000 molecules with library size against the D_4 (left), σ_2 (middle) and 5HT2A (right) receptor.



Extended Data Fig. 5.

Number of top 5000 molecules in a singleton or group scaffold changes with library size against the D_4 (left), σ_2 (middle) and 5HT2A (right) receptor.

Acknowledgements

Funding was provided by US NIH grants R35GM122481 (to B.K.S.) and by GM133836 (to J.J.I.). We thank OpenEye Software for the use of Omega and Schrodinger LLC for the use of prepwizard, ligprep and qikprop in Maestro. We thank Khanh Tang, Benjamin Tingle and Jose Castanon for helping with calculations. We thank Tia Tummino and Stefan Gahbauer for reading this work.

Data availability

The compounds docked in this study are freely available from our ZINC20 and ZINC22 databases, https://zinc20.docking.org and https://cartblanche22.docking.org. Bio-like molecules for similarity comparison are freely available from ZINC15 database: https://zinc15.docking.org/substances/subsets/world/ for the worldwide drug set and https://zinc15.docking.org/substances/subsets/biogenic/ for the biogenic set. PDB codes associated with this study are: 5WIU (the D_4 receptor), 7MFI (the σ_2 receptor) and 6WHA (the 5HT2A receptor). Source data are available for all figures.

References

1. Bohacek RS, McMartin C & Guida WC The art and practice of structure-based drug design: A molecular modeling perspective. Medicinal research reviews 16, 3–50 (1996). [PubMed: 8788213]
2. Fink T, Bruggesser H & Reymond JL Virtual exploration of the small-molecule chemical universe below 160 daltons. Angewandte Chemie International Edition 44, 1504–1508 (2005). [PubMed: 15674983]
3. Wilhelm S et al. Discovery and development of sorafenib: a multikinase inhibitor for treating cancer. Nature reviews Drug discovery 5, 835–844 (2006). [PubMed: 17016424]
4. Macarron R et al. Impact of high-throughput screening in biomedical research. Nature reviews Drug discovery 10, 188–195 (2011). [PubMed: 21358738]
5. Brown DG & Boström J Where do recent small molecule clinical development candidates come from? Journal of medicinal chemistry 61, 9442–9468 (2018). [PubMed: 29920198]
6. Hert J, Irwin JJ, Laggner C, Keiser MJ & Shoichet BK Quantifying biogenic bias in screening libraries. Nature chemical biology 5, 479–483 (2009). [PubMed: 19483698]

7. Martin YC Diverse viewpoints on computational aspects of molecular diversity. Journal of Combinatorial Chemistry 3, 231–250 (2001). [PubMed: 11350246]
8. Breinbauer R, Vetter IR & Waldmann H From protein domains to drug candidates—natural products as guiding principles in the design and synthesis of compound libraries. Angewandte Chemie International Edition 41, 2878–2890 (2002).
9. Koehn FE & Carter GT The evolving role of natural products in drug discovery. Nature reviews Drug discovery 4, 206–220 (2005). [PubMed: 15729362]
10. Arve L, Voigt T & Waldmann H Charting biological and chemical space: PSSC and SCONP as guiding principles for the development of compound collections based on natural product scaffolds. QSAR & Combinatorial Science 25, 449–456 (2006).
11. Ertl P, Roggo S & Schuffenhauer A Natural product-likeness score and its application for prioritization of compound libraries. Journal of chemical information and modeling 48, 68–74 (2008). [PubMed: 18034468]
12. Gupta S & Aires-de-Sousa J Comparing the chemical spaces of metabolites and available chemicals: models of metabolite-likeness. Molecular Diversity 11, 23–36 (2007). [PubMed: 17447158]
13. Bon RS & Waldmann H Bioactivity-guided navigation of chemical space. Accounts of chemical research 43, 1103–1114 (2010). [PubMed: 20481515]
14. Lenci E & Trabocchi A Diversity-Oriented Synthesis and Chemoinformatics: A Fruitful Synergy towards Better Chemical Libraries. European Journal of Organic Chemistry.
15. Grigalunas M, Brakmann S & Waldmann H Chemical evolution of natural product structure. Journal of the American Chemical Society 144, 3314–3329 (2022). [PubMed: 35188375]
16. Rodrigues T, Reker D, Schneider P & Schneider G Counting on natural products for drug design. Nature chemistry 8, 531–541 (2016).
17. Chen Y, de Bruyn Kops C & Kirchmair J Data resources for the computer-guided discovery of bioactive natural products. Journal of chemical information and modeling 57, 2099–2111 (2017). [PubMed: 28853576]
18. Petrone PM et al. Biodiversity of small molecules–a new perspective in screening set selection. Drug discovery today 18, 674–680 (2013). [PubMed: 23454345]
19. Oprea TI Property distribution of drug-related chemical databases. Journal of computer-aided molecular design 14, 251–264 (2000). [PubMed: 10756480]
20. Warr WA, Nicklaus MC, Nicolaou CA & Rarey M Exploration of ultralarge compound collections for drug discovery. Journal of Chemical Information and Modeling 62, 2021–2034 (2022). [PubMed: 35421301]
21. <https://enamine.net/compound-collections/real-compounds> (
22. Alon A et al. Structures of the σ2 receptor enable docking for bioactive ligand discovery. Nature 600, 759–764 (2021). [PubMed: 34880501]
23. Lyu J et al. Ultra-large library docking for discovering new chemotypes. Nature 566, 224–229 (2019). [PubMed: 30728502]
24. Gorgulla C et al. An open-source drug discovery platform enables ultra-large virtual screens. Nature 580, 663–668 (2020). [PubMed: 32152607]
25. Sadybekov AA et al. Synthon-based ligand discovery in virtual libraries of over 11 billion compounds. Nature 601, 452–459 (2022). [PubMed: 34912117]
26. Stein RM et al. Virtual discovery of melatonin receptor ligands to modulate circadian rhythms. Nature 579, 609–614 (2020). [PubMed: 32040955]
27. Grebner C et al. Virtual screening in the cloud: how big is big enough? Journal of Chemical Information and Modeling 60, 4274–4282 (2019). [PubMed: 31682421]
28. Walters WP Virtual chemical libraries: miniperspective. Journal of medicinal chemistry 62, 1116–1124 (2018). [PubMed: 30148631]
29. Irwin JJ et al. An aggregation advisor for ligand discovery. Journal of medicinal chemistry 58, 7076–7087 (2015). [PubMed: 26295373]
30. Venkatakrishnan A et al. Molecular signatures of G-protein-coupled receptors. Nature 494, 185–194 (2013). [PubMed: 23407534]

31. Munk C et al. An online resource for GPCR structure determination and analysis. Nature methods 16, 151–162 (2019). [PubMed: 30664776]
32. Schuller M et al. Fragment binding to the Nsp3 macrodomain of SARS-CoV-2 identified through crystallographic screening and computational docking. Science advances 7, eabf8711 (2021). [PubMed: 33853786]
33. Lipinski CA in Molecular informatics: confronting complexity, proceedings of the Beilstein-Institut Workshop (Frankfurt, Germany).
34. Lipinski CA, Lombardo F, Dominy BW & Feeney PJ Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Advanced drug delivery reviews 23, 3–25 (1997).
35. QikProp, Schrödinger, LLC, New York, NY, 2021.
36. Hann MM & Oprea TI Pursuing the leadlikeness concept in pharmaceutical research. Current opinion in chemical biology 8, 255–263 (2004). [PubMed: 15183323]
37. Singh I et al. Structure-based Discovery of Conformationally Selective Inhibitors of the Serotonin Transporter. bioRxiv (2022).
38. Fink EA et al. Structure-based discovery of nonopioid analgesics acting through the α2A-adrenergic receptor. Science 377, eabn7065, doi:DOI:10.1126/science.abn7065 (2022). [PubMed: 36173843]
39. Bemis GW & Murcko MA The properties of known drugs. 1. Molecular frameworks. Journal of medicinal chemistry 39, 2887–2893 (1996). [PubMed: 8709122]
40. Gu S, Smith MS, Yang Y, Irwin JJ & Shoichet BK Ligand Strain energy in large library docking. Journal of Chemical Information and Modeling 61, 4331–4341 (2021). [PubMed: 34467754]
41. Bender BJ et al. A practical guide to large-scale docking. Nature protocols 16, 4799–4832 (2021). [PubMed: 34561691]
42. Bellmann L, Penner P, Gastreich M & Rarey M Comparison of Combinatorial Fragment Spaces and Its Application to Ultralarge Make-on-Demand Compound Catalogs. Journal of Chemical Information and Modeling 62, 553–566 (2022). [PubMed: 35050621]
43. Shoichet BK & Kuntz ID Matching chemistry and shape in molecular docking. Protein Engineering, Design and Selection 6, 723–732 (1993).
44. Gallagher K & Sharp K Electrostatic contributions to heat capacity changes of DNA-ligand binding. Biophysical journal 75, 769–776 (1998). [PubMed: 9675178]
45. Meng EC, Shoichet BK & Kuntz ID Automated docking with grid-based energy evaluation. Journal of computational chemistry 13, 505–524 (1992).
46. Mysinger MM & Shoichet BK Rapid context-dependent ligand desolvation in molecular docking. Journal of chemical information and modeling 50, 1561–1573 (2010). [PubMed: 20735049]
47. Southan C et al. The IUPHAR/BPS Guide to PHARMACOLOGY in 2016: towards curated quantitative interactions between 1300 protein targets and 6000 ligands. Nucleic acids research 44, D1054–D1068 (2016). [PubMed: 26464438]
48. Mendez D et al. ChEMBL: towards direct deposition of bioassay data. Nucleic acids research 47, D930–D940 (2019). [PubMed: 30398643]
49. Stein RM et al. Property-unmatched decoys in docking benchmarks. Journal of chemical information and modeling 61, 699–714 (2021). [PubMed: 33494610]

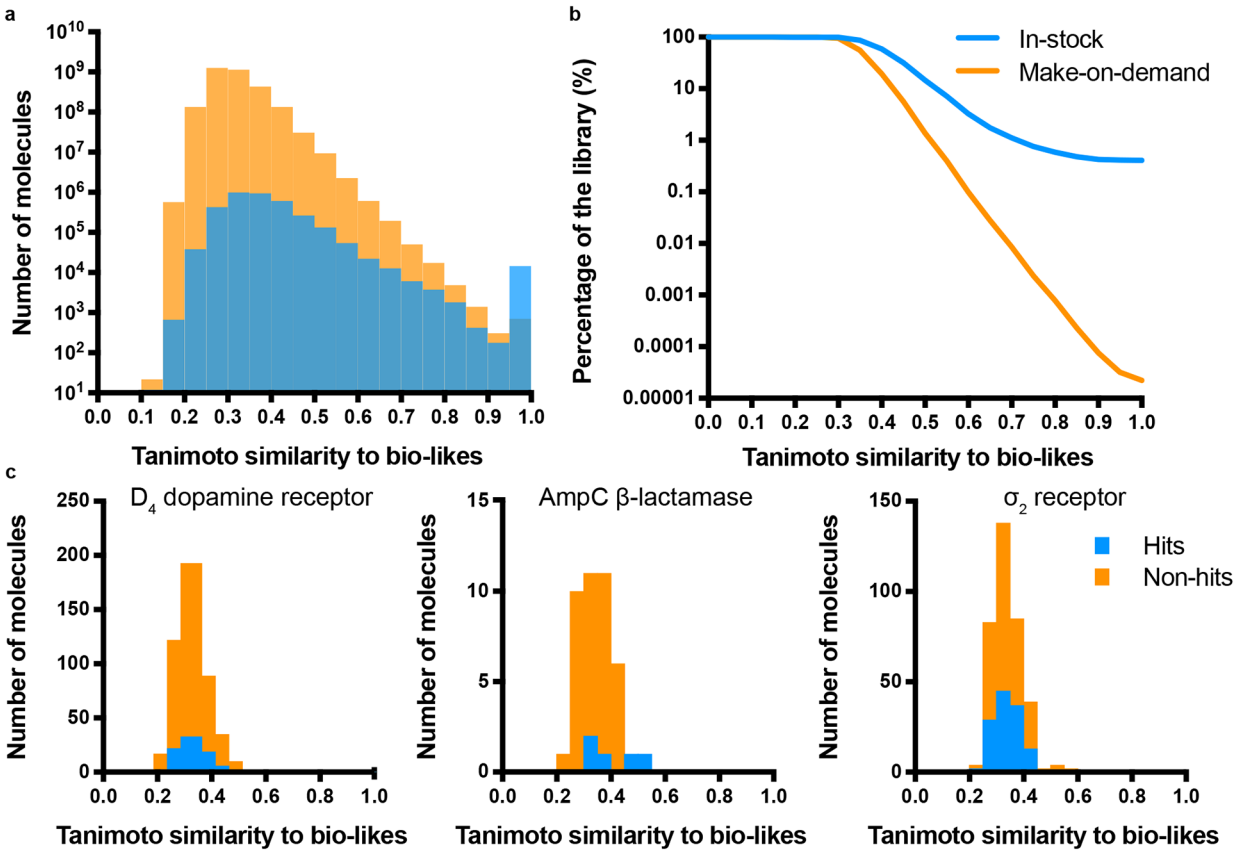


Figure 1 |. Bio-like bias decreases dramatically as the screening library grows from the in-stock library to the make-on-demand library.

a, The distribution of molecules in the in-stock (blue) and make-on-demand (orange) libraries as a function of the Tanimoto similarity to their nearest neighbor in the bio-like molecule set, which contains worldwide drugs, metabolites and natural products. **b**, Percentage of the in-stock (blue) and make-on-demand (orange) libraries as a function of the Tanimoto similarity to their nearest neighbor in the bio-like molecule set. **c**, The distribution of docking prioritized and experimentally active (blue) and non-active (orange) molecules from five different docking campaigns as a function of the Tanimoto similarity to their nearest neighbor in the bio-like molecule set. The docking campaigns from left to right are the D_4 dopamine receptor, the AmpC β-lactamase and the σ_2 receptor. The rest two docking campaigns on the melatonin receptor and the Nsp3 Macrodomain are shown in Extended Data Figure 1.

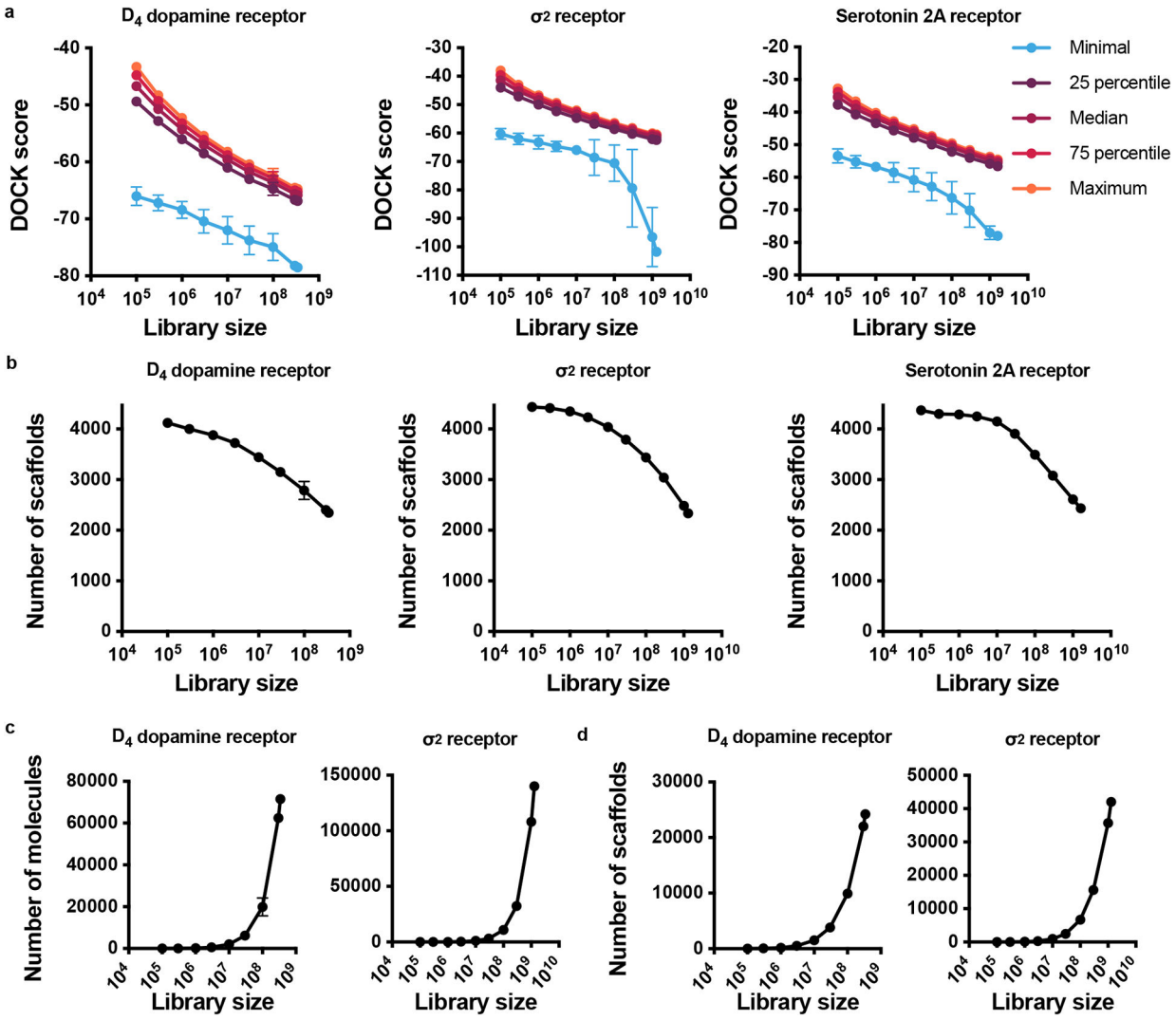


Figure 2 |. Docking performance improves with library size docking against the D_4 (left), σ_2 (middle) and 5HT2A (right) receptor.

a. Change of docking score of the top-ranking 5000 molecules as library size grows. **b**. Number of Bemis-Murko scaffolds in the top-ranking 5000 molecules as library size grows. **c**. Change with library size of the number of molecules with docking scores suggesting high likelihood of binding, based on experimental correlation[22,23] for the D_4 and σ_2 receptors (the 5HT2A receptor was excluded as this experimental correlation has not been measured for it). **d**. Change with library size of the number of scaffolds with docking scores suggesting high likelihood of binding. All data are mean ± standard deviation. Each set was selected 30 times with random selection from the full library

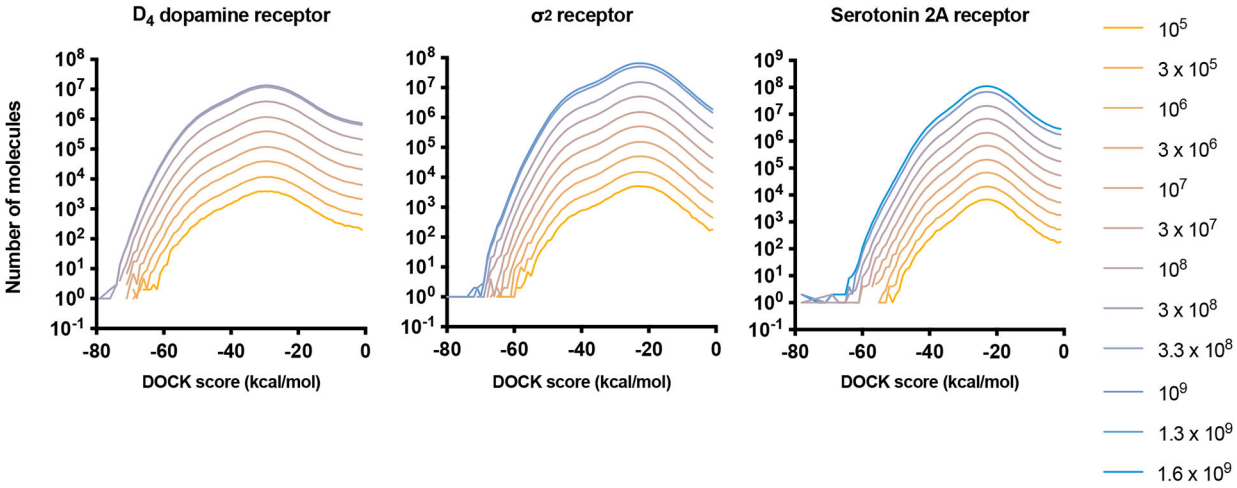


Figure 3 |. The long tail expands with library size against the D_4 (left), σ_2 (middle) and 5HT2A (right) receptor.

The X axis is in linear scale while the Y axis is in log scale. The color gradient changes from orange to marine, representing library size changes from small to large.

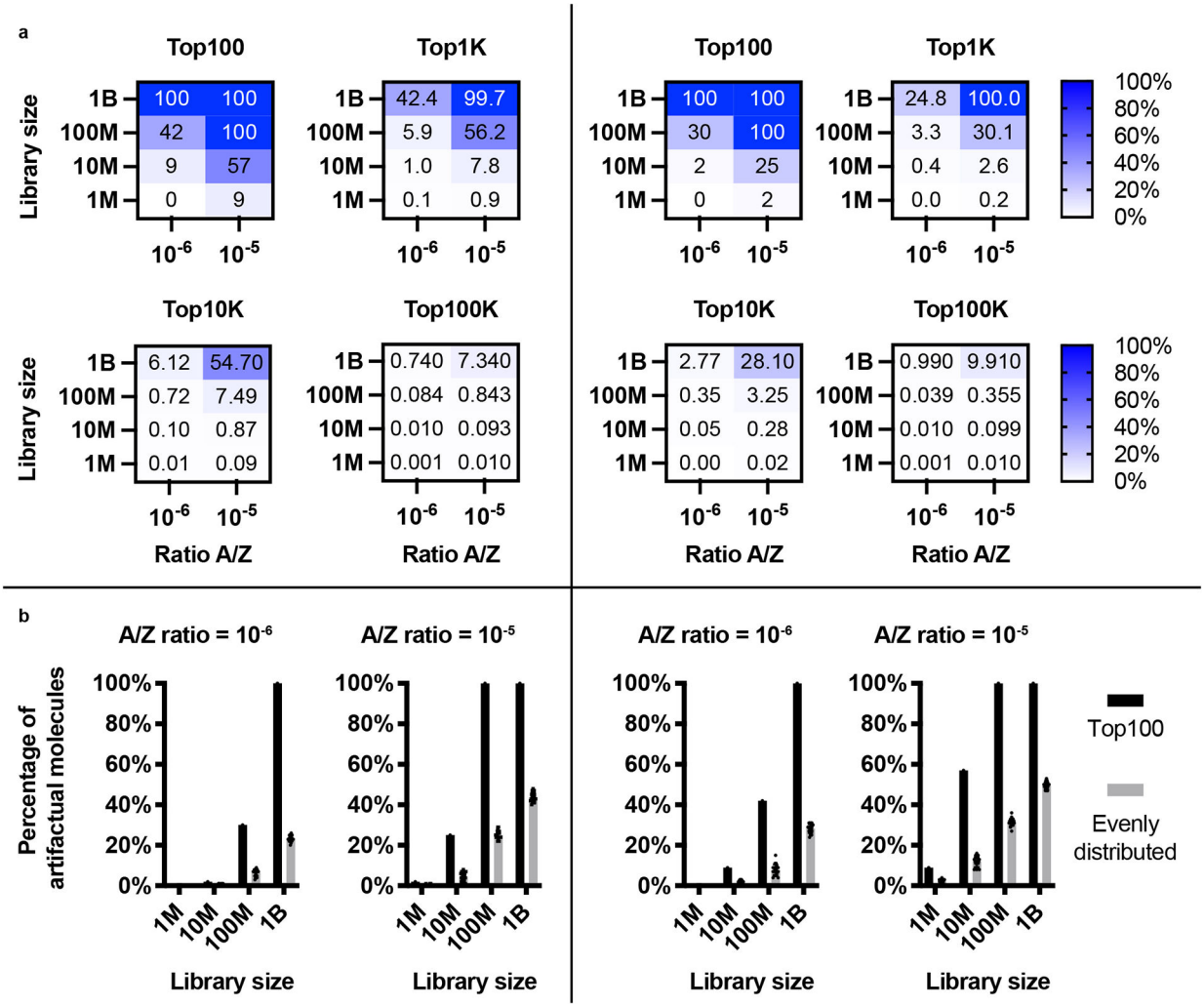


Figure 4 |. Number of artifacts increases with library size.

a, Heat maps of the percentage of artifacts in the top N (N = 100, 1,000, 10,000 and 100,000) docked molecules. The percentage of artifacts in the top N docked molecules for a given library size and the ratio between artifacts and library molecules is colored using a linear scale ranging from 0% (white) to 100% (blue). The artifactual molecules were sampled from the extreme value distribution for the left panel and the uniform distribution for the right panel. **b**, The percentage of artifactual molecules in the 100 selected molecules between the two strategies. The first strategy is just picking the top 100 molecules colored by black bars and the second strategy is selecting 100 molecules evenly distributed from 5 ranking ranges colored by grey bars. Five ranking ranges were top 1–100, top 101–1,000, top 1,001–10,000, top 10,001–100,000 and top 100,001–1,000,000. 20 molecules were drawn at random from each ranking tranche. This selection was repeated 20 times at random. The artifactual molecules were sampled from the extreme value distribution for the left panel and the uniform distribution for the right panel. Data shown here are mean ± standard deviation