nature chemical biology

Article

https://doi.org/10.1038/s41589-024-01797-w

The impact of library size and scale of testing on virtual screening

Received: 13 June 2024

Accepted: 13 November 2024

Published online: 03 January 2025

Check for updates

Fangyu Liu[1], Olivier Mailhot[1], Isabella S. Glenn[1], Seth F. Vigneron[1], Violla Bassim[2], Xinyu Xu[1], Karla Fonseca-Valencia[1], Matthew S. Smith[1], Dmytro S. Radchenko[3], James S. Fraser[2], Yurii S. Moroz[3,4,5] ✉, John J. Irwin[1] ✉ & Brian K. Shoichet[1] ✉

Virtual ligand libraries for ligand discovery have recently increased 10,000-fold. Whether this has improved hit rates and potencies has not been directly tested. Meanwhile, typically only dozens of docking hits are assayed, clouding hit-rate interpretation. Here we docked a 1.7 billion-molecule virtual library against β-lactamase, testing 1,521 new molecules and comparing the results to a 99 million-molecule screen where 44 molecules were tested. In a larger screen, hit rates improved twofold, more scaffolds were discovered and potency improved. Fifty-fold more inhibitors were found, supporting the idea that the large libraries harbor many more ligands than are being tested. In sampling smaller sets from the 1,521, hit rates only converged when several hundred molecules were tested. Hit rates and affinities improved steadily with docking score. It may be that as the scale of docking libraries and their testing grows, both ligands and our ability to rank them will improve.

With the advent of ultra-large, make-on-demand ('tangible') libraries, available chemical space has increased from about 3.5 million to over 38 billion (https://enamine.net/compound-collections/real-compounds). Recent studies suggest that structure-based docking prioritizes potent ligands from within such libraries, with affinities often in the mid-nanomolar and sometimes high-picomolar range[1–11]. Docking the new libraries seems to improve hit rates, affinities and chemotype novelty versus smaller libraries[12,13], suggesting that bigger libraries are better for virtual screening. This is supported by simulations that show that as libraries grow, the best molecules fit ever better to protein binding sites[14]. Still, exactly how large libraries may improve docking screens versus smaller libraries, if in fact they do so, remains to be tested experimentally in side-by-side studies.

Further clouding the issue is the scale of testing of molecules prioritized from the docking campaigns. Irrespective of whether million- or billion-scale libraries are screened, rarely are more than several dozen molecules synthesized and tested[3,6–8]. From the hit rates of these screens (number active divided by number tested), it has been inferred that there are likely hundreds of thousands or even millions of potential ligands in the libraries that remain untested, but this has not been probed experimentally[1]. As important, the few molecules tested make the results subject to the statistics of small numbers. It is unclear that we can have full confidence in hit rates, affinities and the likelihood of discovering new chemotypes—all key outcomes—when testing only a few dozen compounds.

Here we begin to investigate these questions quantitatively. First, to explore the impact of library size on docking outcome, we screened over 1.7 billion molecules for inhibitors of the model enzyme AmpC β-lactamase[1,15–20] and compared the results to a previous screen on the same enzyme using essentially the same method where only 99 million molecules were docked[1]. These smaller and larger screens were compared by hit rates, affinities and the number of new chemotypes discovered. Second, we synthesized and tested 1,521 compounds for AmpC inhibition, rather than the 44 tested in the smaller library campaign[1],

[1]Department of Pharmaceutical Chemistry, University of California, San Francisco, San Francisco, CA, USA. [2]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, CA, USA. [3]Enamine Ltd, Kyïv, Ukraine. [4]Chemspace LLC, Kyïv, Ukraine. [5]Department of Chemistry, Taras Shevchenko National University of Kyïv, Kyïv, Ukraine. ✉e-mail: ysmoroz@gmail.com; jir322@gmail.com; bshoichet@gmail.com

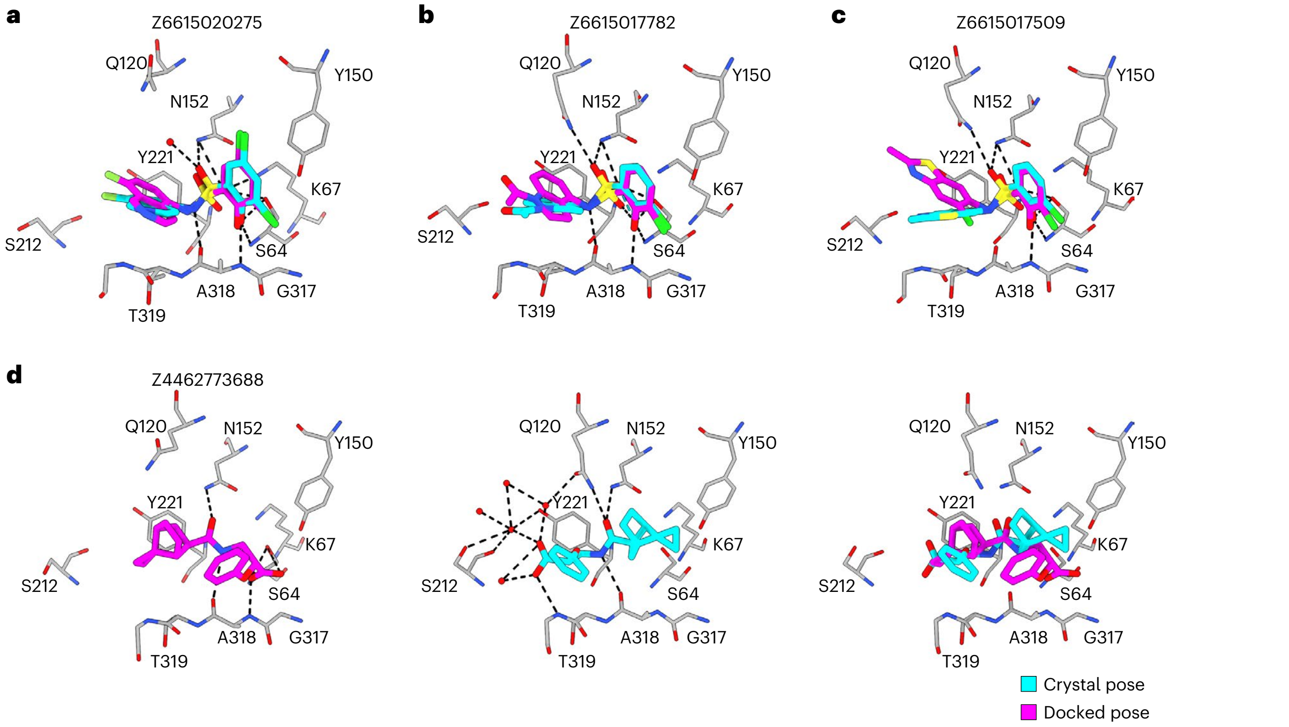


Fig. 1 | Superposition of the crystallographic and docking poses of the new AmpC inhibitors. Crystal structures (carbons in cyan) and docked poses (carbons in magenta) of the inhibitors. AmpC carbon atoms are in gray, oxygens in red, nitrogens in blue, sulfurs in yellow, chlorides in green and fluorides in light green. Hydrogen bonds are shown as black dashed lines. **a**–**c**, AmpC in complex with Z6615020275 (**a**) (**1**) (r.m.s.d. to crystal structure 0.79 Å, 1.3 μM), Z6615017782 (**b**) (**2**) (r.m.s.d. = 0.97 Å, 0.95 μM) and Z6615017509 (**c**) (**3**) (r.m.s.d. = 3.14 Å, 0.86 μM). The overlay of the crystal and docked poses are shown. **d**, AmpC in complex Z4462773688 (**4**) (r.m.s.d. = 5.61 Å, 323 μM). The docked poses (left panel), crystal poses (middle panel) and the overlay of the docked and crystal poses are shown (right panel).

and asked whether the number of inhibitors found scaled with number of top-ranking molecules investigated, something that has until now simply been an implication of large library docking. Third, with these observations in hand, we examined the sensitivity of docking hit rates and affinities to the scale of experimental testing by subsampling smaller sets from the larger one; this has implications for how we should understand docking hit rates and affinities, and how we should scale these experiments in the future. Fourth, we investigate how hit rate is predicted by docking score, and whether we might expect better molecules to be found as libraries expand into the tens of billions of molecules and beyond[5,21]. Finally, the scale of the experimental testing here allows us to investigate potential correlations between docking rank and affinity category (high, mediocre, poor). We will argue that the answers emerging from this study support further expansion of docking libraries into the trillions of compounds range, and a re-investment in docking scoring functions to optimize what is now a loose correlation between docking rank and affinity category.

Results

Selection, synthesis and testing of 1,521 docking hits

In a previous docking screen of 99 million molecules against AmpC β-lactamase, 44 high-ranking molecules were prioritized for synthesis and testing. This revealed five new inhibitors with affinities ranging from 1.3 to 400 μM, a hit rate of 11% using this range of activity[1]. Using essentially the same docking method, here we screened a 1.7 billion-molecule library against AmpC. In addition to selecting high-ranking molecules for testing, as in the smaller library screen, here molecules from across the docking scoring range were also tested, allowing us to also investigate how hit rate varied with docking score. Overall, 838,672,414 molecules ranking from −117.35 kcal mol^{-1} (best scores) to −28 kcal mol^{-1} (worst scores), were considered as candidates for testing. These were organized into bins of resolution ranging from 2 to 4 kcal mol^{-1} among the lower (better) scores to 8 kcal mol^{-1} among the higher (worse) scores. Up to 25,000 molecules were selected per bin, by rank order (for the lower and better energy bins, this amounted to all the molecules in the bin). Molecules topologically similar to known inhibitors, with ECFP4-based Tanimoto coefficient (T_c) > 0.5, were excluded, as were those with more than one unsatisfied hydrogen bond donor and more than six hydrogen bond acceptors: such molecules exploit known gaps in the DOCK3.8 scoring function[22]. The remaining 184,317 molecules were clustered by $T_c = 0.32$ based on the interaction fingerprinting[23], resulting in 80,767 clusters. In previous simulations[14] and experiments[4] we had found molecules with artifactually favorable scores concentrated among the top-ranking docked molecules. Here too, we observed molecules that achieved scores much higher than one would expect from the overall distribution; this problem became more acute as the library grew (Extended Data Fig. 1). We chose to ignore these molecules for experimental testing. The origins of these molecules, and their experimental confirmation as docking artifacts, is explored in a separate study[24].

Overall, 2,089 cluster heads, all topologically dissimilar to one another and to known inhibitors, were chosen for synthesis and testing. Of these, 1,521 were successfully synthesized (a fulfillment rate of 73%). Manual inspection ('manual-picked') from among the better scoring bins (−95.96 to −60 kcal mol^{-1}) accounted for 734 of these, and another 1,336 molecules were chosen based on rank alone ('auto-picked'): 560 molecules occurred in both sets (Supplementary Table 1).

All molecules were initially tested at 200, 100 and 40 μM for AmpC inhibition[1,16,20]. Of the 1,447 experimentally well-behaved molecules, 1,296 were among the top scoring 1% of the docked molecules, the same cutoff used in the 99 million-molecule screen (the rest were spread out among lower ranks and were selected to test hit rate versus score dependence). Of these 1,296 compounds, 168 had an apparent inhibitory constant (K_i) < 166 μM, based on the three-point inhibition numbers and assuming competitive inhibition (below), while another 122 had apparent K_i values between 166 and 400 µM. Concentration–response curves were measured for 17 compounds across this potency range. The half-maximum inhibitory concentration (IC_{50}) values from these full curves corresponded well to those predicted by the three-point inhibition numbers (Extended Data Fig. 2 and Supplementary Table 2). For seven of the new inhibitors, each in a different chemotype family, we determined full K_i values and mechanisms by Lineweaver–Burk analysis (Extended Data Fig. 3). All seven were competitive inhibitors, consistent with docking to the AmpC active site, with K_i values ranging from 0.7 to 121 µM (Extended Data Fig. 3). Accordingly, we modeled all the new inhibitors as competitive, consistent with the X-ray crystal structures determined for four of them, which all bound in the β-lactamase active site (Fig. 1). With this assumption, K_i values ranged from 464 to 0.46 µM (ref. 25) (Fig. 2). All assays included 0.01% Triton X-100, diminishing the likelihood of artifacts from colloidal aggregation[18,26]. For further confidence, 140 of the inhibitors were checked for particle formation by dynamic light scattering[26–28]; no signs of colloid-like particle formation were detected at relevant concentrations (Supplementary Table 3).

Docked versus crystallographic geometries

To investigate how docking poses corresponded to experimentally determined geometries, the structures of four of the new inhibitors were determined by X-ray crystallography, with resolutions ranging from 1.66 to 1.88 Å (Supplementary Table 4). Unambiguous electron density allowed us to confidently model the positions of the new inhibitors in the enzymes' active site (Extended Data Fig. 4). For Z6615020275 (**1**) (1.3 µM; Fig. 1a), Z6615017782 (**2**) (0.95 µM; Fig. 1b) and Z6615017509 (**3**) (0.86 µM; Fig. 1c), the docked and experimental structures superimposed with a 0.79, 0.97 and 3.14 Å root mean square deviation (r.m.s.d.) respectively, with differences in position stemming from deviations of nonwarhead groups binding distally in the site. For a weaker inhibitor Z4462773688 (**4**) (323 µM), the crystal structure showed larger deviations from the docking prediction. An unprecedented bicyclo-alkyl carboxylate bound in a geometry flipped from that anticipated by docking, leading to an r.m.s.d. of 5.61 Å (Fig. 1d). Z4462773688 is an example of the 44 inhibitors found in this campaign that sample not only new topologies, but also new warheads for AmpC.

Hit rates are higher from the larger library screen

The overall hit rate (number experimentally active/number tested) from the 1.7 billion-molecule campaign was 22.4% (290 actives/1,296 high-ranking tested). We defined a hit as having an apparent K_i value < 400 µM, based on previous literature. This hit rate is significantly higher than that from 99 million-molecule docking screen, which was 11.4% ($P = 0.021$ by Z test) (Fig. 2a). With a more stringent definition of hits, the hit rates drop for both screens: to 8.3 and 2.3% ($K_i < 100$ µM) and to 2.5 and 2.3% ($K_i < 30$ µM) for the larger and smaller library campaigns, respectively (Extended Data Fig. 5a,b). Unlike the 44 molecules from the smaller library that were both high ranking and manually selected, the 1,296 molecules from the larger library include both manually selected compounds and those picked by score alone (both sets also selected for diversity and dissimilarity to knowns). Focusing only on the high-ranking, manually selected molecules from the larger screen (662 molecules), the hit rate is significantly higher than from the smaller library campaign: 21.4 versus 11.4% ($P = 0.032$, Extended Data Fig. 5c). Considering the top 44 manually selected molecules from the larger screen—that is, the same number picked from the smaller library campaign—the hit-rate difference is even more pronounced: 47.7 versus 11.4% ($P = 0.00005$) (Extended Data Fig. 5d,e). This hit-rate difference is supported by differences across affinity ranges. Most of the actives from the 99 million-molecule screen had apparent K_i values over 100 µM (Fig. 2b), with one inhibitor found in the 1 to 3.2 µM range and none found in the intermediate ranges. Conversely, from the 1.7 billion library each half-log affinity bin is well-populated by new inhibitors. The higher hit rate from the larger library is consistent with the idea that as the virtual libraries grow, ever more plausible molecules are fortuitously sampled and prioritized by molecular docking.

Hit-rate variability and ligand affinity ranges

While hit rate is a fair way to compare the two screens, the raw number of hits was naturally far greater from the larger library (Fig. 2c), where 29-fold more high-ranking molecules were tested. Qualitatively, this explains why all half-log affinity bins were well-populated from the larger library, whereas this was more hit-and-miss when we only tested 44 molecules (Fig. 2b). To quantify how hit rate varies with the number tested, we pulled sets of 44, 139 and 439 molecules randomly 30 times from the 1,296 and asked how hit rate was affected. When only selecting 44 molecules hit rates varied from 11% for one unlucky draw to 36% for a lucky one. Pulling sets of 439 molecules 30 times, the hit rate only varied from 20 to 27%. The standard deviation in hit rates decreased from 6.1 to 3.5 to 1.7%, respectively (Fig. 2d). This variability was mirrored in ligand affinities; for instance, it was not until set size rose to 439 molecules tested that the highest affinity molecules were reliably sampled (Fig. 2e). Re-analyzing previous campaigns against the σ_2 and dopamine D4 receptor[1,4], where around 500 molecules were experimentally tested, similar variability was seen in both hit rates and in sampling of the high-affinity ligands, which for σ_2 were in the low nanomolar range (Extended Data Fig. 6).

These results indicate that both hit rates and affinities in docking screens may be unreliable when only dozens of molecules are tested, as is common in the field. To quantify how many molecules should be tested to report stable hit rates and affinities, we drew on the observation that when large numbers of molecules are tested for the three targets, there is an exponential relationship between affinity and hit rate, something also seen in high-throughput screens[29]. For the top-ranking 1% of docked molecules from each campaign, we modeled hit rates (y) and hit affinities (x) with an exponential plateau function $y = b(1 - e^{-cx})$ for each of target (Fig. 3a). This fit the distribution of affinities for the 1,296 molecules tested for AmpC, 327 for σ_2 and 371 for D4 (all top 1% ranking molecules) with R^2 values of 0.998, 0.999 and 0.985, respectively. As smaller sets are drawn from the full sets, variability rises (Fig. 2d,e). Beginning with 1,296, sampling was stepwise reduced by 20 molecules in a bootstrapping manner, repeating this 1,000 times to evaluate divergence (Fig. 3b). By ~495 molecules, the average R^2 of D4 curves falls to 0.95, a point on all three curves where we began to see the meaningful divergence from the fit to the full range of compounds plotted. This same R^2 occurs at 215 and 135 molecules for AmpC and σ_2, respectively, perhaps reflecting an inverse relationship to hit rates for each target among the top 1% of docked molecules (22.4% for AmpC, 38.7% for σ_2 and 20.8% for D4). In these targets, testing fewer than these several hundred compounds degrades the correlation of affinity with hit rate. For targets with relatively high hit rates, this suggests that over a hundred molecules should be experimentally tested to infer confident docking hit rates and affinity ranges. For targets with lower hit rates, even more compounds would need to be tested for confident results.

To explore this further with a focus on hit-rate variability, we simulated random draws using the AmpC, σ_2 and D4 experimental hit rates from their high-ranking compounds. One hundred thousand bootstrap iterations were performed for sample sizes ranging from 10 to 1,250 compounds in increments of 10 and we considered the mean and lower bound for a single-sided 95% confidence interval at different numbers of compounds tested (Fig. 3c). The solid curves reflect the 95% likelihood that the hit rate will be at a certain level or higher. While the average hit rate over all simulations remains unchanged, the variability increases as the number of molecules tested drops and so does one's confidence that the observed hit rate reflects the true hit rate based on the overall docking rankings. This again suggests more than 100 molecules may be a sensible minimum for experimental testing

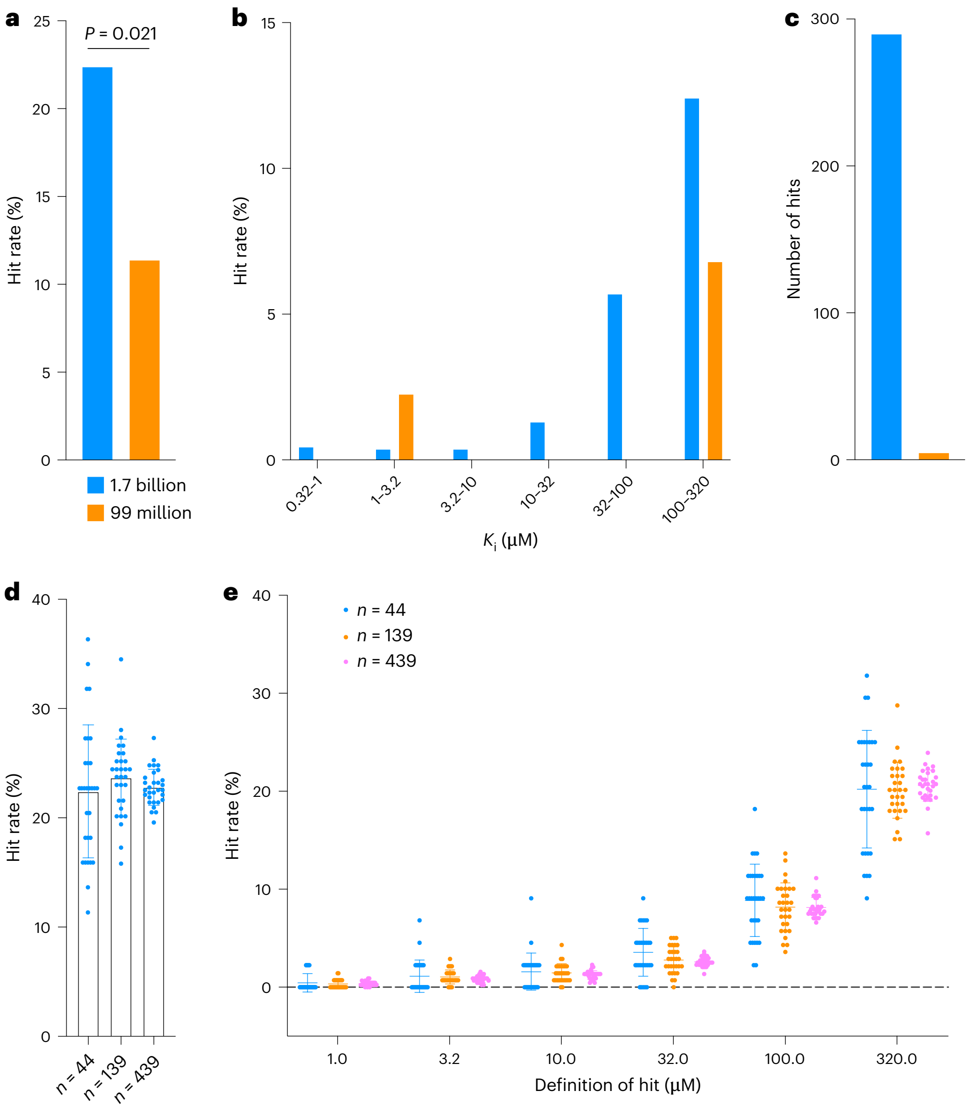


Fig. 2 | Larger-scale docking and testing increases hit rates and reduces uncertainty. a, The hit rates (number of actives/total tested) of the 1.7 billion screen (blue bar; 22.4%) versus the 99 million screen (orange bar; 11.4%). A two-sided *Z*-test was used to compare the hit rates of the two screens, under the assumption that the data followed a normal distribution. **b**, Hit rates by different affinity bins in the 2022 screen and 2019 screen. **c**, Number of hits (number of actives) of the 1.7 billion screen (blue bar) versus the 99 million screen (orange bar). **d**, The impact of randomly purchasing 44, 139 and 439 molecules out of 1,296 molecules for testing on hit rates. Each sample size is randomly drawn 30 times and the resulting hit rates were plotted. The error bars represent s.d.s of the hit rates. The hit rates are 22.42 ± 6.08% (*n* = 44), 23.67 ± 3.54% (*n* = 139) and 22.80 ± 1.65% (*n* = 439). **e**, The impact of randomly purchasing 44, 139 and 439 molecules out of 1,296 molecules for testing on hit rates with different affinity cutoffs. Each sample size is drawn 30 times and the resulting hit rates were plotted. Data represent mean ± s.d.s of the hit rates.

in large library virtual screens. While both the affinity ranges and the hit rates for the screens against AmpC, σ_2 and D4 differ substantially, the functional form relating hit affinity and hit number was the same and led to similar predictions for the minimum number of molecules to test for all three targets. This may help predict how many molecules would be found in different affinity ranges should one choose to test more molecules, a point to which we will return.

Multiple new chemotypes discovered

Only molecules topologically dissimilar to known AmpC inhibitors, and topologically diverse from each other, were selected for synthesis and testing. Since topological diversity can emerge from changes that leave core pharmacophores intact, we also visually inspected inhibitors for novelty. We prioritized molecules by two criteria: those that sampled new scaffolds, and those that explored a new anionic warhead (Extended Data Fig. 7). For instance, Z6615021877 (**5**) and Z6722203632 (**6**) introduce tetrazolone and tetrazole anionic warheads, respectively, both of which were previously unknown for AmpC. Z2607647274 (**7**) and Z4173922012 (**8**) use cycloalkyl carboxylate and tricyclo-heptane carboxylate as their warheads. Meanwhile, Z2610488449 (**9**), which uses a new urea linker scaffold, achieves a high affinity of 12 µM. The affinity of this scaffold was readily optimized to 4 µM, marking it among the most effective noncovalent AmpC inhibitors that does not rely on a sulfonamide linker.

Docking score predicts hit rate

In earlier studies against the D4 dopamine and σ_2 receptors, we had found that docking score correlated to experimental hit rate, generating a well-behaved sigmoidal curve that plateaued at a maximum hit rate[1,4]. While these curves suggested an unexpected ability to predict binders, both receptors have well-formed, buried binding sites, making them unusually suitable for this technique. Meanwhile, the

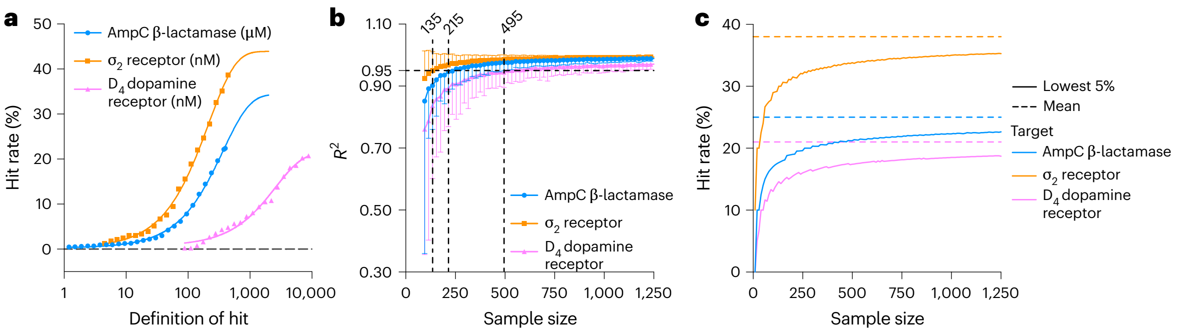


Fig. 3 | Several hundred compounds should be tested in large library docking. **a**, For the top-ranking 1% of the docked molecules, the relationship between hit affinity and hit rates can be fit with an exponential plateau model $y = b(1 - e^{-cx})$ where y is hit rate, x is the minimum affinity for a hit (for AmpC the unit is μM and for σ_2 and D4 it is nM) and b is the maximal hit rate. The fit maximal hit rates are 34.5% for AmpC with an R^2 of 0.998, 43% for an σ_2 receptor with an R^2 of 0.999, and 20.8% for D4 with an R^2 of 0.985. **b**, The impact of subsampling on the R^2. From among the top-ranking 1% of the docked molecules, starting from sample size 1,296 (AmpC, blue; σ_2, orange; D4, pink), each subsample size is bootstrapped 1,000 times and fit with the parameters derived from the entire dataset. The R^2 values are plotted with the symbols indicating the average and the error bars indicating the standard deviations. A dashed line of $R^2 = 0.95$ is labeled. The sample sizes at which the average R^2 reaches 0.95 are labeled. For σ_2, that size is 135, for AmpC, it is 215 and for D4, it is 495 molecules. **c**, Mean and 95% confidence interval for hit rate in relation to sample size for AmpC, σ_2 and D4. The dashed lines show the mean hit rate from the compounds in the top 1% by docking score, and the solid line shows the boundary of a single-sided 95% confidence interval from 100,000 bootstrap iterations. Hits are defined as 400 μM affinity or better for AmpC, 678 nM or better for σ_2 and 10 μM or better for D4.

plateauing of the score versus hit-rate curve suggests a limitation in even our ability to identify, far less rank ligands. To investigate how docking might predict binding in a more solvent-exposed, historically more difficult binding site, we reexplored this relationship for AmpC. Docked molecules were not only selected from among the best docking energies (some poses shown in Extended Data Fig. 8), as is typical in virtual screening, but also from mediocre and unfavorable scoring ranges. Molecules were picked from among 16 scoring bins, beginning at the most favorable DOCK3.8 scores (−100.58 kcal mol^{-1} for AmpC) down to −28 kcal mol^{-1}. The top 1% of the docking-ranked library extends down to scores of −72 kcal mol^{-1}, but by −28 kcal mol^{-1} 49% of the 1.7 billion-molecule library has been sampled. More than 50 molecules per bin were selected from the −100.58 to the −60 kcal mol^{-1} bin, and for scores worse (less negative) than −60 more than 20 molecules were tested per bin. Molecules were selected strictly by numerical rank at the beginning of each bin. They were tested for AmpC inhibition as above.

Hit rates fell monotonically as scores worsened (Fig. 4a, blue curve). This resembles what we had observed for the σ_2 and dopamine D4 receptors[1,4], except that here we do not observe a hit rate plateau; hit rates begin at a maximum at the best docking scores and fall steadily as scores worsen. A difference between the AmpC curve and the plateaus observed previously is that here from the beginning we excluded a small fraction of likely artifacts that concentrate among the very top scoring molecules[14] (Extended Data Fig. 1). The scale of docking in this study allows one to recognize these cheating artifacts by how they diverge from the rest of the library; in another study we find that they may be also recognized by rescoring with an orthogonal scoring function[24]. Both their differential scoring and explicit rescoring may help recognize these molecules in future studies.

To investigate how the affinities of the new inhibitors tracked, we plotted score versus hit rate in the 400, 127, 40 and 13 μM ranges (Fig. 4a, blue, orange, pink and green curves). Here too, the hit rates in each affinity-range rose steadily as score improved. The more potent inhibitors appear at better scores than the less potent ones, with those in the 127 μM or better tranche beginning to appear at scores of −64, those in the 40 μM or better tranche appearing only past −76 and the most potent inhibitors only appearing at the −85 bin. This hints at docking score correlating with gross categorical ranking of affinity, something that was not apparent from smaller studies, nor expected[30,31]. The trend observed in the auto-picked molecules is conserved when considering all molecules (both by rank and manually selected) as well as those that were manual-picked. (Extended Data Fig. 9). We undertook the same analysis with the docking campaigns against the σ_2 receptor and dopamine receptor, where hundreds of molecules were tested across docking ranks that ranged from high to mediocre to poor, as in this study. While the σ_2 and D4 receptor docking hits were more potent than the AmpC hits, typically in the nM range, the same patterns emerged; the most potent ligands appear at better (more negative) docking scores than did the mid-potency ligands, which appear at better scores than the most potent ones (Fig. 4b,c). Admittedly, the relationship between docking score and affinity is mostly categorical, but it appears to rank molecules better than simple binary classification as binders or nonbinders, with more potent ligands more concentrated in better scoring regions. As loose as these correlations are, they may support a predictive relationship between docking score and affinity category (high, medium or low), at least when at scale. This would warrant a renewed emphasis on improving the field's scoring functions and offer a metric against which they might be tested.

To compare the hit-rate curves for the three targets, we plotted the negative logarithm of the rank percentage ('pProp') for the D4 and σ_2 receptors, and for AmpC (Fig. 4d). A pProp of three denotes a compound occupying the top 0.1% scoring region, a pProp of 4 the top 0.01% and so on; plotting rank avoids scoring offsets among the targets. The hit-rate curve of the most permissive hit definition for each target is plotted against the pProp. The D4 and σ_2 curves align well, peaking around a pProp of 5, with the plateau occupying the region from 4 to 6 (top in 10,000 to top in 1,000,000), while the AmpC curve is slightly right shifted, peaking above 6 and not suffering from a plateau. These curves allow one to quantify the parts of the docking scoring range where most hits are likely to be found. For the D4 and σ_2 receptors, it also alerts one to the danger of over-emphasizing the very best ranked molecules where those that cheat the scoring function concentrate, absent controls for them[14]. As docking and virtual screening libraries climb into the tens of billions of molecules[5,21], this concern will become more pressing[24]. This may be addressed by recognizing their divergence from other molecules in the library, and by explicit rescoring with an orthogonal scoring function.

Discussion

In the last 5 years, the number of molecules accessible for ligand discovery has expanded 10,000-fold. Anecdotally, this has revealed molecules with improved activity from library docking. Here we seek to quantify this in apples-to-apples comparisons of a smaller versus larger library; five key observations emerge. First, comparing a docking screen of

https://doi.org/10.1038/s41589-024-01797-w

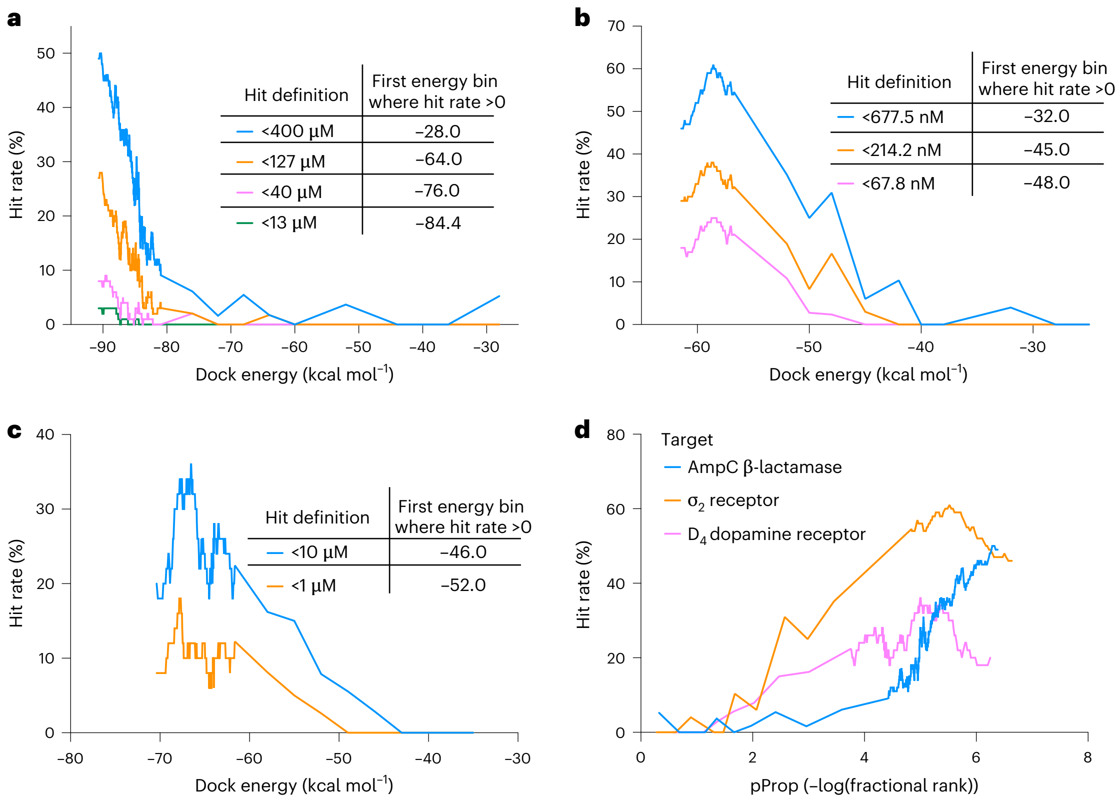


Fig. 4 | Hit rate of tested compounds plotted against DOCK scores with different affinity cutoffs. a, The AmpC hit rates of 1,293 well-behaved auto-picked compounds using four different affinity cutoffs, <400, 127, 40 and 13 μM, are plotted against DOCK scores. **b**, σ_2 receptor hit rates of 484 compounds plotted against DOCK scores with three different affinity cutoffs: <677.5, 241.2 and 67.8 nM. **c**, Dopamine D4 hit rates of 549 compounds plotted against DOCK scores with two different affinity cutoffs: <10 and <1 μM. **d**, Rescaling the hit-rate curves of the three targets by the $\log_{10}$ of fractional rank in the library. For each target, the most permissive hit definition is used.

99 million molecules to one of 1.7 billion molecules against the same target, hit rates improved with library size, as did the potency of the inhibitors. Multiple new AmpC inhibitor chemotypes were discovered. Second, consistent with the idea that there are many more ligands to be discovered than are being prioritized, the number of new inhibitors scaled almost linearly with the number of top-ranking molecules tested; testing 29-fold more molecules discovered 58-fold more inhibitors. Third, to determine reliable docking statistics from a large library screen, one must also test at scale. When only a handful of molecules are tested, as is common in docking, statistics of hit rates and maximal affinities suffer from large errors. We find that at least several hundred molecules should be tested for docking statistics to be trustworthy. Fourth, in contrast to earlier studies where hit rates plateaued above a certain docking score[1,4], here hit rates continued to climb essentially linearly as score improves. This was also true for the D4 and σ_2 receptors after removing their high-ranking artifacts. This observation supports the idea that as libraries grow, hit rates and affinities will improve, as long as high-ranking docking artifacts can be removed or avoided. Finally, a loose, categorical correlation between docking score and ligand affinity was observed for AmpC, and on reanalysis also for the σ_2 and D4 receptor campaigns[1,4]. While this correlation remains loose and only by affinity category (for example, strong, mediocre, weak), it may suggest that further optimization of docking scoring functions will allow the field to distinguish not only binders from nonbinders, but also categorically rank them by activity, something we and others have long discounted[30,31].

Several caveats should be aired. The monotonic improvement of hit rate with docking score and its loose correlation with affinity have only been observed in three systems. This merits investigation in other targets, ideally using other scoring functions, at scale. Current community tests of docking methods, such as CACHE[32], may offer a forum for doing so. Methodologically, we note that for less than 10% of the molecules reported here were full IC_{50} curves determined. While these correlated well with inferred IC_{50} and K_i values based on three concentration point inhibition, such affinities must be considered approximate. An important aspect of getting well-behaved score versus hit-rate curves was our ability to recognize and exclude artifactual molecules that appeared to cheat our scoring function. Such artifacts, with different physical and chemical origins, have appeared in previous large library campaigns against the σ_2 and D4 receptors[1,4]; recognizing and removing them was important to revealing the well-behaved score versus hit-rate curves and affinity categorization that we describe here. Whether such artifacts are peculiar to DOCK3.8, the program we use in this study, or to docking more generally, is presently unknown. Finally, it is important to emphasize that docking results improve both with scale of testing and size of library. In a 1 billion-molecule library, even testing thousands of molecules will likely leave hundreds of thousands of potent ligands untested. When only dozens are tested, the statistics of small numbers ensure that not only the best but often the most representative ligands will be missed.

These caveats should not obscure the major observations of this study. Against the same target, docking a 20-fold larger library led to improved hit rates and affinities, consistent with theoretical simulations[14]. Similarly, as more high-ranking molecules are tested, more ligands are found, supporting the idea that most true ligands in the new ultra-large libraries remain to be tested (we suffer from an embarrassment of riches). Once we correct for high-ranking docking artifacts, hit rate rises monotonically with docking score. More tentatively, a correlation between affinity and score also appears at scale.

While brute force docking, of the sort described here, has been able to address a 1,000-fold increase in library size, to go up another 1,000-fold, into the trillions of molecules, seems beyond it and more guided sampling of chemical space may be required[5,11,33,34]. What this study suggests is that efforts to sample from the supra-trillion molecule space should be worthwhile. To support such efforts, we are making available the identity, docking score and experimental activities of each of the 1,521 molecules tested here (Supplementary Table 1), and extensive docking score and pose information from the full library screen (https://lsd.docking.org).

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41589-024-01797-w.

References

1. Lyu, J. et al. Ultra-large library docking for discovering new chemotypes. *Nature* **566**, 224–229 (2019).
2. Gorgulla, C. et al. An open-source drug discovery platform enables ultra-large virtual screens. *Nature* **580**, 663–668 (2020).
3. Stein, R. M. et al. Virtual discovery of melatonin receptor ligands to modulate circadian rhythms. *Nature* **579**, 609–614 (2020).
4. Alon, A. et al. Structures of the sigma(2) receptor enable docking for bioactive ligand discovery. *Nature* **600**, 759–764 (2021).
5. Sadybekov, A. A. et al. Synthon-based ligand discovery in virtual libraries of over 11 billion compounds. *Nature* **601**, 452–459 (2022).
6. Fink, E. A. et al. Structure-based discovery of nonopioid analgesics acting through the α_{2A}-adrenergic receptor. *Science* **377**, eabn7065 (2022).
7. Singh, I. et al. Structure-based discovery of conformationally selective inhibitors of the serotonin transporter. *Cell* **186**, 2160–2175.e17 (2023).
8. Gahbauer, S. et al. Docking for EP4R antagonists active against inflammatory pain. *Nat. Commun.* **14**, 8067 (2023).
9. Sadybekov, A. V. & Katritch, V. Computational approaches streamlining drug discovery. *Nature* **616**, 673–685 (2023).
10. Gorgulla, C. et al. A multi-pronged approach targeting SARS-CoV-2 proteins using ultra-large virtual screening. *iScience* **24**, 102021 (2021).
11. Klarich, K., Goldman, B., Kramer, T., Riley, P. & Walters, W. P. Thompson sampling—an efficient method for searching ultralarge synthesis on demand databases. *J. Chem. Inf. Model.* **64**, 1158–1171 (2024).
12. Walters, W. P. Virtual chemical libraries. *J. Med. Chem.* **62**, 1116–1124 (2019).
13. Gorgulla, C., Jayaraj, A., Fackeldey, K. & Arthanari, H. Emerging frontiers in virtual drug discovery: from quantum mechanical methods to deep learning approaches. *Curr. Opin. Chem. Biol.* **69**, 102156 (2022).
14. Lyu, J., Irwin, J. J. & Shoichet, B. K. Modeling the expansion of virtual screening libraries. *Nat. Chem. Biol.* **19**, 712–718 (2023).
15. Weston, G. S., Blazquez, J., Baquero, F. & Shoichet, B. K. Structure-based enhancement of boronic acid-based inhibitors of AmpC beta-lactamase. *J. Med. Chem.* **41**, 4577–4586 (1998).
16. Powers, R. A., Morandi, F. & Shoichet, B. K. Structure-based discovery of a novel, noncovalent inhibitor of AmpC beta-lactamase. *Structure* **10**, 1013–1023 (2002).
17. Feng, B. Y., Shelat, A., Doman, T. N., Guy, R. K. & Shoichet, B. K. High-throughput assays for promiscuous inhibitors. *Nat. Chem. Biol.* **1**, 146–148 (2005).
18. Feng, B. Y. et al. A high-throughput screen for aggregation-based inhibition in a large compound library. *J. Med. Chem.* **50**, 2385–2390 (2007).
19. Eidam, O. et al. Design, synthesis, crystal structures, and antimicrobial activity of sulfonamide boronic acids as beta-lactamase inhibitors. *J. Med. Chem.* **53**, 7852–7863 (2010).
20. Babaoglu, K. et al. Comprehensive mechanistic analysis of hits from high-throughput and docking screens against beta-lactamase. *J. Med. Chem.* **51**, 2502–2511 (2008).
21. Gorgulla, C. et al. VirtualFlow 2.0—the next generation drug discovery platform enabling adaptive screens of 69 billion molecules. Preprint at *bioRxiv* https://doi.org/10.1101/2023.04.25.537981 (2023).
22. Bender, B. J. et al. A practical guide to large-scale docking. *Nat. Protoc.* **16**, 4799–4832 (2021).
23. Fassio, A. V. et al. Prioritizing virtual screening with interpretable interaction fingerprints. *J. Chem. Inf. Model.* **62**, 4300–4318 (2022).
24. Wu, Y. et al. Identifying artifacts from large library docking. *J. Med. Chem.* **67**, 16796–16806 (2024).
25. Cheng, Y. & Prusoff, W. H. Relationship between the inhibition constant (K1) and the concentration of inhibitor which causes 50 per cent inhibition (I_{50}) of an enzymatic reaction. *Biochem. Pharmacol.* **22**, 3099–3108 (1973).
26. McGovern, S. L., Helfand, B. T., Feng, B. & Shoichet, B. K. A specific mechanism of nonspecific inhibition. *J. Med. Chem.* **46**, 4265–4272 (2003).
27. Feng, B. Y. & Shoichet, B. K. A detergent-based assay for the detection of promiscuous inhibitors. *Nat. Protoc.* **1**, 550–553 (2006).
28. O'Donnell, H. R., Tummino, T. A., Bardine, C., Craik, C. S. & Shoichet, B. K. Colloidal aggregators in biochemical SARS-CoV-2 repurposing screens. *J. Med. Chem.* **64**, 17530–17539 (2021).
29. Walters, W. P. & Namchuk, M. Designing screens: how to make your hits a hit. *Nat. Rev. Drug Discov.* **2**, 259–266 (2003).
30. Tirado-Rives, J. & Jorgensen, W. L. Contribution of conformer focusing to the uncertainty in predicting free energies for protein-ligand binding. *J. Med. Chem.* **49**, 5880–5884 (2006).
31. Irwin, J. J. & Shoichet, B. K. Docking screens for novel ligands conferring new biology. *J. Med. Chem.* **59**, 4103–4120 (2016).
32. Ackloo, S. et al. CACHE (Critical Assessment of Computational Hit-finding Experiments): a public–private partnership benchmarking initiative to enable the development of computational methods for hit-finding. *Nat. Rev. Chem.* **6**, 287–295 (2022).
33. Gentile, F. et al. Artificial intelligence-enabled virtual screening of ultra-large chemical libraries with deep docking. *Nat. Protoc.* **17**, 672–697 (2022).
34. Yang, Y. et al. Efficient exploration of chemical space with docking and deep learning. *J. Chem. Theory Comput.* **17**, 7106–7119 (2021).

Methods

Large-scale docking

The campaign used the structure in the Protein Data Bank (PDB) 1L2S (ref. 16). Three Q120 conformations were modeled based on the X-ray density of PDB 3FKW (ref. 35) using qFit-3.0, with an occupancy of 0.49, 0.34 and 0.17 (ref. 36). The occupancy of the alternative conformations was converted into an additional energy term and incorporated in the DOCK scoring function as described previously[37]. The protein structure was protonated using Reduce[38]. Energy grids for the different energy terms of the scoring function were pregenerated van der Waals terms based on the AMBER force fields using CHEMGRID[39]; Poisson–Boltzmann-based electrostatic potentials using QNIFFT[40,41]; context-dependent ligand desolvation was calculated using SOLVMAP[42]. The volume of the low dielectric and the desolvation volume was extended out 2.0 and 0.25 Å. The thiophene carboxylate inhibitor solved in PDB 1L2S was used to generate matching spheres, which are later used by the docking software to fit pregenerated ligands' conformations into the small molecule binding sites[43]. The resulting docking setups were evaluated for its ability to enrich known AmpC ligands over property-matched decoys. Decoys are theoretical nonbinders to the receptor as they are topologically dissimilar to known ligands but retain similar physical properties. We curated 31 AmpC ligands based on their dissimilarity among themselves. 2,480 decoys were generated by using the DUDE-Z pipeline[44]. The docking set-up can rank ligands over decoys with a logAUC of 28.5 with most of the ligands recapitulating their experimental poses. For docking against 1.7 billion molecules, each molecule from the ZINC22 database[45] was sampled in about 3,822 orientations and 875 conformations by using DOCK3.8 (ref. 43). Overall, over 1,841 trillion complexes were sampled and scored, spending 2,129,230 core hours or about 1 month on a 3,000 core cluster, using DOCK3.8 (ref. 43).

Hit-picking strategy

To increase novelty, high-ranking molecules with scores down to −79.25 (99,277 molecules), and molecules from different energy bins (~25,000 from −76, −72, −68, −64 and −60 bins and 5,000 from −52, −44, −36 and −28 bins), summed to 244,217 molecules, were filtered to exclude those similar to 237 previously known ligands. A T_c cutoff of 0.5 was used; no molecule more similar than this value was allowed, removing 9,561 molecules. We also filtered out molecules that buried too many uncompensated polar groups: while DOCK3.8 penalizes for desolvation, we find that these artifacts can nevertheless occur. Using LUNA 1,024-length binary fingerprints[23], molecules that had more than one hydrogen bond donor and more than six hydrogen bond acceptors that were not compensated with polar interactions to the protein were removed; 50,339 molecules were filtered out at this step. This left 184,317 for further processing. For autopicking, these molecules were clustered for self-similarity using an ECFP4 T_c = 0.32, resulting in 80,767 cluster heads.

Most of the molecules tested were auto-picked based on docking rank. With almost all the high-ranking molecules being negatively charged, we wanted to ensure that their representation as anions at pH 7.4 was likely. We used JChem to calculate the distribution of protonation states of the high-ranking cluster heads and compared this to the dominant state represented in their docked poses (multiple protonation states of a molecule can be docked). Only when the calculated dominant charge state matched with that of the docked pose, and the species is calculated to be more than 80% anionic, was the molecule accepted for autopicking, which left 56,814 molecules. Molecules were picked based on their docking ranks across different affinity bins, selecting 1,336 molecules for synthesis and testing.

For manual picking from the different energy bins, all cluster heads were again filtered for interactions using LUNA, seeking molecules that formed hydrogen bonds with backbone of A318, that made pi–pi interactions with Y221, and that made at least two more interactions with the binding pocket (that is, hydrogen bonds with N152, N346, G320, S212, R204, Q120, cation-pi with K315, K67 or pi–pi interaction with Y150). The molecules that passed these filters were reclustered at a T_c = 0.32; cluster heads were visually inspected and prioritized. The rest of the high-scoring cluster heads were also manually inspected seeking new interesting chemotypes. A total of 734 were prioritized manually, slightly less than half of the molecules that were synthesized and tested.

Synthesis of the molecules

Compounds were sourced from the Enamine REAL database (https://enamine.net/compound-collections/real-compounds). The purities of active molecules were at least 90% and typically above 95%. The detailed chemical synthesis can be found in the Supplementary Information.

AmpC enzymology

All candidate inhibitors were dissolved in dimethylsulfoxide (DMSO) at 20 mM, and more dilute DMSO stocks were prepared as necessary so that the concentration of DMSO was held constant at 1% v/v in 50 mM sodium cacodylate buffer, pH 6.5. AmpC activity and inhibition was monitored spectrophotometrically using either CENTA or nitrocefin as substrates. All assays included 0.01% Triton X-100 to reduce compound aggregation artifacts. Active compounds were further investigated for aggregation by dynamic light scattering and by detergent-dependent inhibition of the counter-screening enzyme malate dehydrogenase.

For initial screening, the docking hits were diluted such that final concentrations in the reaction buffer was 200, 100 and 40 µM. In these assays, two widely studied AmpC substrates were used, depending on availability, CENTA[46] and nitrocefin[16]. The first was tested at an $[S]/(K_m)$ ratio of 1.81 (K_m CENTA 27.6 µM; [S] = 50 µM, where K_m is the Michaelis constant) and the second was tested at $[S]/K_m$ ratios of 0.556 (K_m nitrocefin 180 µM; [S] = 100 µM) and 0.156 ([S] = 28 µM). The colorimetric assay was converted to a medium throughput manner using a BMG Labtech CLARIOstar. Substrate (CENTA (EC_{50} K_m = 27.6 µM) or nitrocefin (EC_{50} = 180 µM)) and protein were injected into buffer containing the putative inhibitor, followed by rate measurement for 50 s in 96-well format. IC_{50} values reflect the percentage inhibition fit to a dose–response equation in GraphPad Prism with a Hill coefficient set to one ($f(x) = \max - \frac{\max - \min}{1 + \frac{x}{IC_{50}}}$). The K_i was calculated using the Cheng–Prusoff equation ($K_i = \frac{IC_{50}}{1 + \frac{[S]}{Km}}$). For 18 of the more potent compounds, based on the initial three concentration point results, full dose–response curves were measured and for another eight full K_i values were measured and calculated using Lineweaver–Burk plots. Data were analyzed using GraphPad Prism software v.9.

AmpC crystallization, data collection and structure determination

AmpC crystallization was carried out as previously described[16]. Briefly, cocrystals of AmpC and inhibitors were grown by vapor diffusion in hanging drops equilibrated over 1.7 M potassium phosphate buffer (pH 8.7) using microseeding. The initial concentration of protein in the drop was 6 mg ml^{-1} and the concentration of the inhibitor was 0.5 mM. The inhibitor was added to the crystallization drop in a 4% DMSO, 1.7 M potassium phosphate buffer (pH 8.7) solution. Crystals appeared within 3–5 days after equilibration at 23 °C.

Data were measured from a single crystal per complex on the Beamline 8.3.1 of the Berkeley Advanced Light Source, with wavelength 1.11583 Å at 100 K. Before data collection, cocrystals of AmpC were immersed in a cryoprotectant solution of 20% sucrose and 1.7 M potassium phosphate (pH 8.7) for about 20 s and then flash-cooled in liquid nitrogen. The structures were solved by molecular replacement with PHENIX[47] using PDB 1L2S as the search model. Structure refinement was carried out with PHENIX and COOT[48]. MolProbity[49] was used for validation (Extended Data Fig. 3); structural figures were prepared using ChimeraX[50]. To test model bias, polder omit maps were calculated after perturbing the model by omitting the selected ligands: the ligands were first omitted, and the resulting model was subjected to three cycles of phenix.refine in both real and reciprocal space with simulated annealing. The ligands were then inserted to calculate the polder omit maps[51].

Hit-rate curves

To obtain hit rate curves, the experimentally tested molecules for each target (AmpC, the σ_2 and dopamine D4 receptors) were ordered by increasing DOCK score. A rolling window was passed over the list, calculating the hit rate as the percentage of molecules with experimentally determined affinity equal to or better than the hit definition, and the DOCK score as the average for the window. A window size of 100 was used for AmpC and σ_2, and a window of 50 for D4 receptor. For all three targets, molecules were picked from both within and outside what would typically be considered high-ranking regions. The rolling window was stopped for those scores outside the high-ranking region since discrete score bins were used in the hit-picking of these likely nonbinders. The scores at which the rolling window was stopped are −78 for AmpC, −52.5 for σ_2 and −60 for D4. For the pProp rescaling, the same strategy was used, but the DOCK scores were transformed to fractional rank based on the observed score distribution. The negative base 10 logarithm of the fractional rank is then reported, termed ‘pProp’.

Hit-rate modeling

For sampling hit-rate variability in relation to sample size, we used sample sizes for 10 to 1,250 in jumps of 10. For each sample size, we picked 100,000 random samples of the uniform distribution [0, 1] using Python. The hit rate of the sample was then defined as the number of observations with equal to or lower than the observed experimental hit rate for that target. A single-sided 95% confidence interval is built by taking the boundary value between the top 95% observed hit rates and the bottom 5%.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The compounds docked in this study are freely available from the ZINC20 and ZINC22 databases, https://zinc20.docking.org and https://cartblanche22.docking.org. All compounds tested can be purchased from Enamine. Compound information including their ZINC ID, catalog ID, SMILES, DOCK score, ranking and affinity can be found in Supplementary Table 1. The synthetic procedures and purity information for the hits can be found in the Supplementary Note. Extensive docking-related files can be found at https://lsd.docking.org. DOCK3.8 is freely available for noncommercial research at https://dock.compbio.ucsf.edu/DOCK3.8/. A web-based version is available without restriction at https://blaster.docking.org/. X-ray structures and maps are available in the PDB under accession numbers 9C81 (Z4462773688), 9C6P (Z6615017509), 9C84 (Z6615020275) and 9DHL (Z6615017782), respectively. Source data are provided with this paper.

References

35. Chen, Y., McReynolds, A. & Shoichet, B. K. Re-examining the role of Lys67 in class C beta-lactamase catalysis. *Protein Sci.* **18**, 662–669 (2009).
36. Riley, B. T. et al. qFit 3: protein and ligand multiconformer modeling for X-ray crystallographic and single-particle cryo-EM density maps. *Protein Sci.* **30**, 270–285 (2021).
37. Fischer, M., Coleman, R. G., Fraser, J. S. & Shoichet, B. K. Incorporation of protein flexibility and conformational energy penalties in docking screens to improve ligand discovery. *Nat. Chem.* **6**, 575–583 (2014).
38. Word, J. M., Lovell, S. C., Richardson, J. S. & Richardson, D. C. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol.* **285**, 1735–1747 (1999).
39. Meng, E. C., Shoichet, B. K. & Kuntz, I. D. Automated docking with grid-based energy evaluation. *J. Comput. Chem.* **13**, 505–524 (1992).
40. Gallagher, K. & Sharp, K. Electrostatic contributions to heat capacity changes of DNA-ligand binding. *Biophys. J.* **75**, 769–776 (1998).
41. Sharp, K. A. Polyelectrolyte electrostatics: salt dependence, entropic, and enthalpic contributions to free energy in the nonlinear Poisson–Boltzmann model. *Biopolymers* **36**, 227–243 (1995).
42. Mysinger, M. M. & Shoichet, B. K. Rapid context-dependent ligand desolvation in molecular docking. *J. Chem. Inf. Model.* **50**, 1561–1573 (2010).
43. Coleman, R. G., Carchia, M., Sterling, T., Irwin, J. J. & Shoichet, B. K. Ligand pose and orientational sampling in molecular docking. *PLoS ONE* **8**, e75992 (2013).
44. Stein, R. M. et al. Property-unmatched decoys in docking benchmarks. *J. Chem. Inf. Model.* **61**, 699–714 (2021).
45. Tingle, B. I. et al. ZINC-22—a free multi-billion-scale database of tangible compounds for ligand discovery. *J. Chem. Inf. Model.* **63**, 1166–1176 (2023).
46. Eidam, O. et al. Fragment-guided design of subnanomolar beta-lactamase inhibitors active in vivo. *Proc. Natl Acad. Sci. USA* **109**, 17448–17453 (2012).
47. Liebschner, D. et al. Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallogr. D* **75**, 861–877 (2019).
48. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
49. Chen, V. B. et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66**, 12–21 (2010).
50. Pettersen, E. F. et al. UCSF ChimeraX: structure visualization for researchers, educators, and developers. *Protein Sci.* **30**, 70–82 (2021).
51. Liebschner, D. et al. Polder maps: improving OMIT maps by excluding bulk solvent. *Acta Crystallogr. D* **73**, 148–157 (2017).

Acknowledgements

This work is supported by US National Institutes of Health (NIH) grant nos. R35GM122481 (to B.K.S.), GM71896 (to J.J.I.) and GM145238 (to J.S.F.) and a Damon Runyon Postdoctoral Research Fellowship (F.L.). We thank ChemAxon for JChem, OpenEye Scientific software for Omega and Schrödinger LLC for the Maestro suite. We thank G. Meigs and J. Holton for their assistance at Beamline 8.3.1 at the Advanced Light Source, operated by UCSF with NIH grant nos. R01 GM124149 for technology development and P30 GM124169 for beamline operations, and the Integrated Diffraction Analysis Technologies program of the US Department of Energy Office of Biological and Environmental Research. We thank K. Srinivasan for his assistance on data collection. The Advanced Light Source (Berkeley, CA, USA) is a national user facility operated by Lawrence Berkeley National Laboratory on behalf of the US Department of Energy under contract number DE-AC02-05CH11231, Office of Basic Energy Sciences.

Author contributions

F.L. conducted the docking screens and the ligand optimization assisted by S.F.V. and advised by B.K.S. F.L. and I.S.G. conducted the in vitro enzymatic assays, with early assistance from S.F.V. F.L.

determined the structures by X-ray crystallography, with assistance from V.B. and X.X., advised by J.S.F. F.L. and O.M. did the analysis with advice from M.S.S. Aggregation studies were conducted by K.F.-V. and I.S.G. J.J.I. developed and prepared the make-on-demand library assisted with large library docking strategies. D.S.R. and Y.S.M. supervised compound synthesis of Enamine compounds purchased from the ZINC22 database and the 46 billion catalog library.

Competing interests

B.K.S. is a founder of Epiodyne, Inc.; BlueDolphin, LLC; and Deep Apple Therapeutics, Inc., and serves on the SAB of Schrodinger LLC and of Vilya Therapeutics, and on the SRB of Genentech. J.J.I. cofounded Deep Apple Therapeutics, Inc., and BlueDolphin, LLC. J.S.F. is a consultant for, has equity in and receives research support from Relay Therapeutics. The other authors declare no competing interests.

Additional information

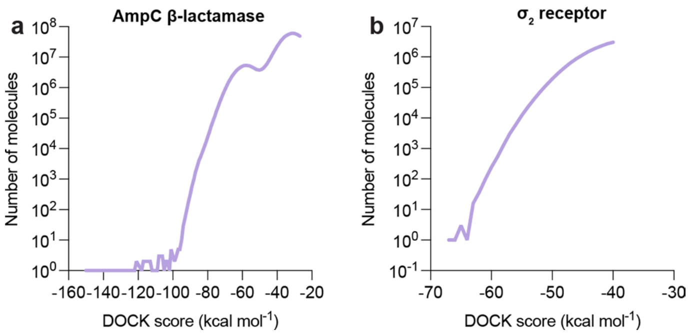
Extended data is available for this paper at https://doi.org/10.1038/s41589-024-01797-w.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41589-024-01797-w.

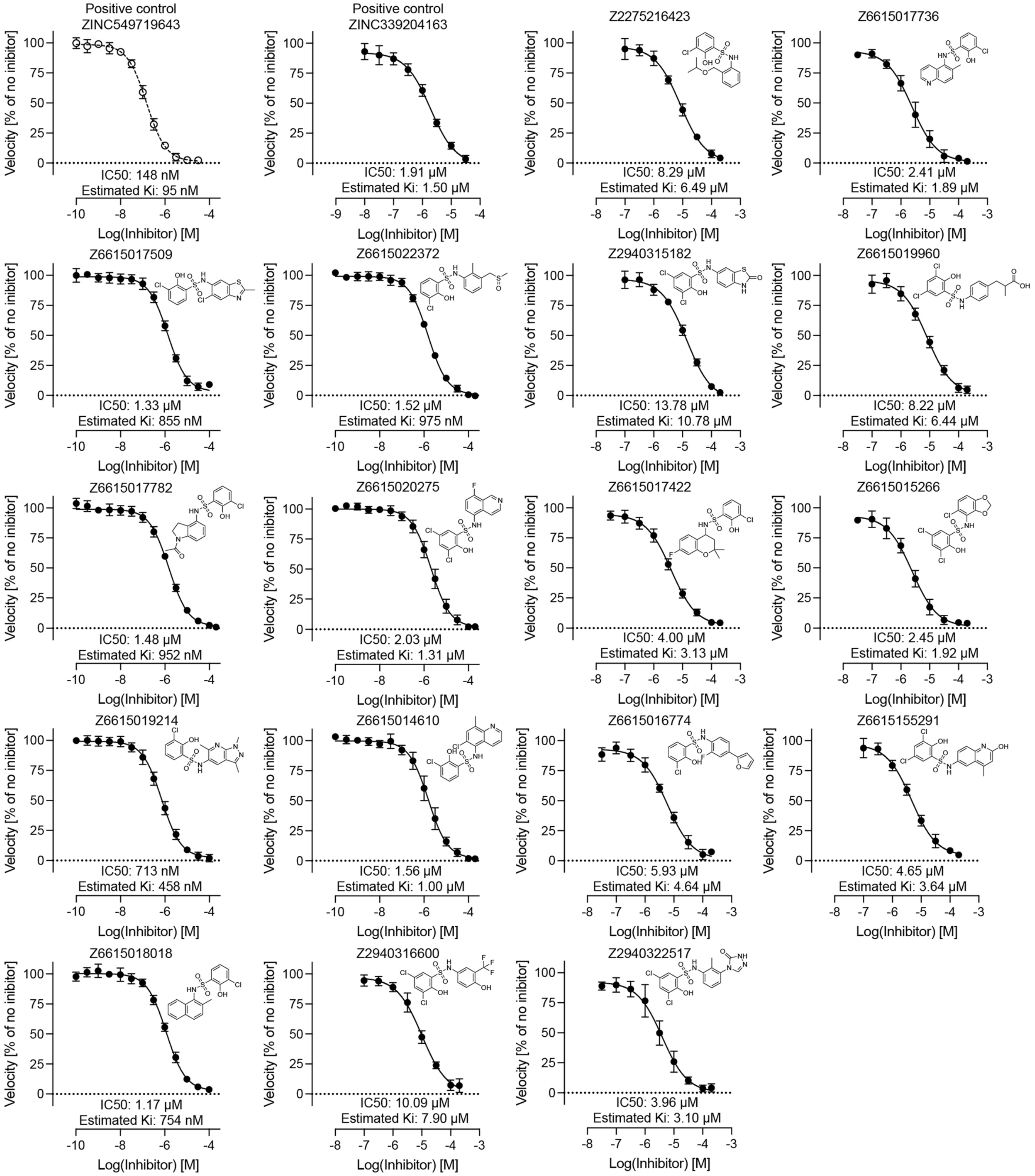
Correspondence and requests for materials should be addressed to Yurii S. Moroz, John J. Irwin or Brian K. Shoichet.

Peer review information *Nature Chemical Biology* thanks Artem Cherkasov, Tyuji Hoshino and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

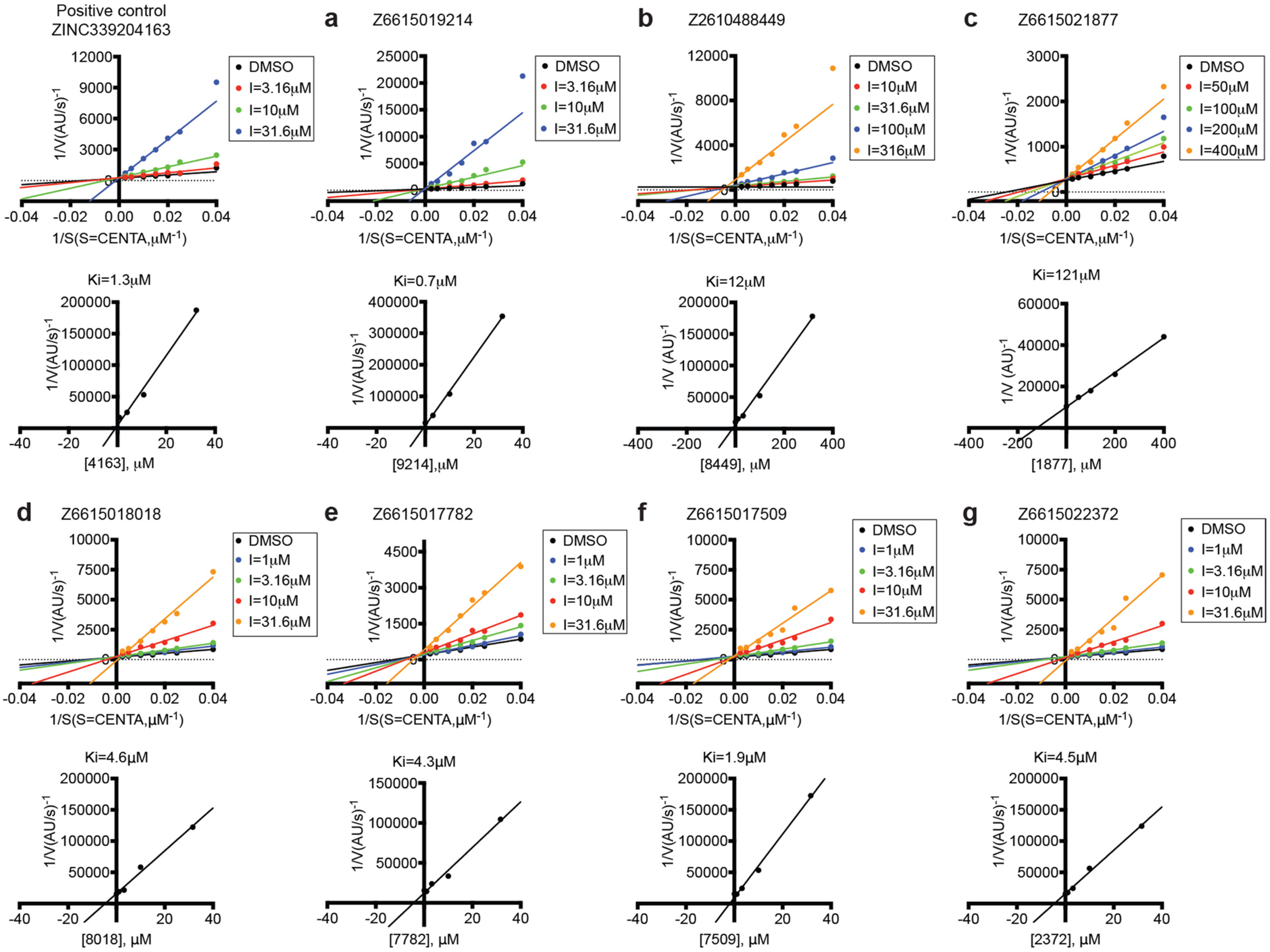
Reprints and permissions information is available at www.nature.com/reprints.



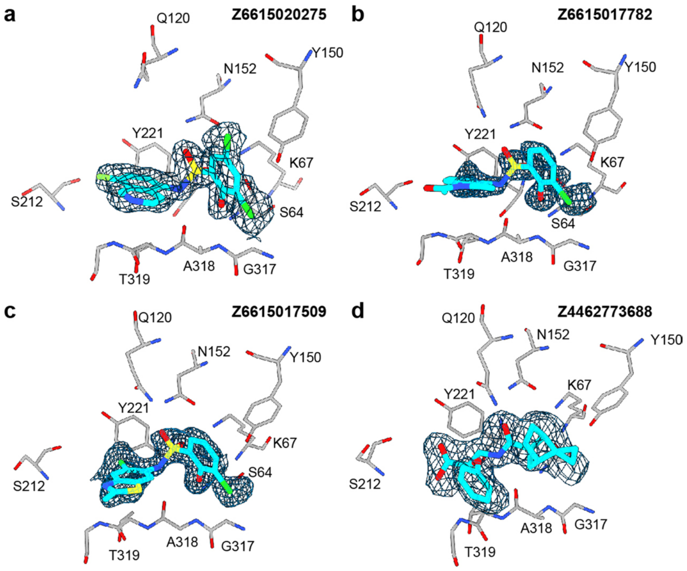
Extended Data Fig. 1 | Molecules with artifactually favorable scores disrupt the distribution of docking scores and concentrate among the top-ranking docked molecules. a, DOCK scores of molecules against AmpC. **b**, DOCK scores of molecules against σ_2 receptor[4].



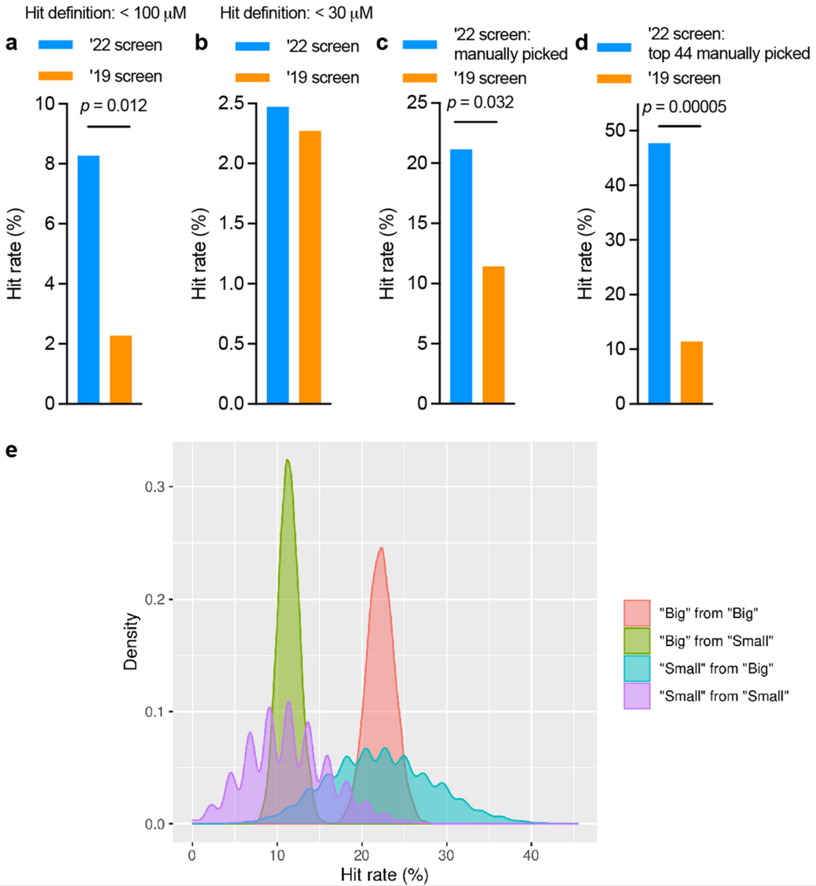
Extended Data Fig. 2 | Concentration-response curves for 17 of the new docking-derived AmpC inhibitors. Nitrocefin was kept at a constant concentration of 100 µM (for positive control ZINC549719643, new inhibitors Z6615018018, Z6615017509, Z6615022372, Z6615017782, Z6615020275, Z6615019214 and Z6615014610) or 50 µM (for positive control ZINC339304163, new inhibitors Z2275216423, Z6615017736, Z2940316600, Z2940315182, Z6615019960, Z2940322517, Z6615017422, Z6615015266, Z6615016774 and Z6615155291). The estimated K_i is calculated based on the K_d of nitrocefin (180 µM) calculated from a Lineweaver-Burk analysis. The previously reported K_i for ZINC549719643 is 77 nM[1] and for ZINC339304163 is 1.25 µM[1]. Data represent mean ± s.d.s from three biological replicates.



Extended Data Fig. 3 | Lineweaver-Burk plots of seven of the new AmpC inhibitors (a-g). ZINC339304163 is a positive control inhibitor identified in a previous docking campaign[1].

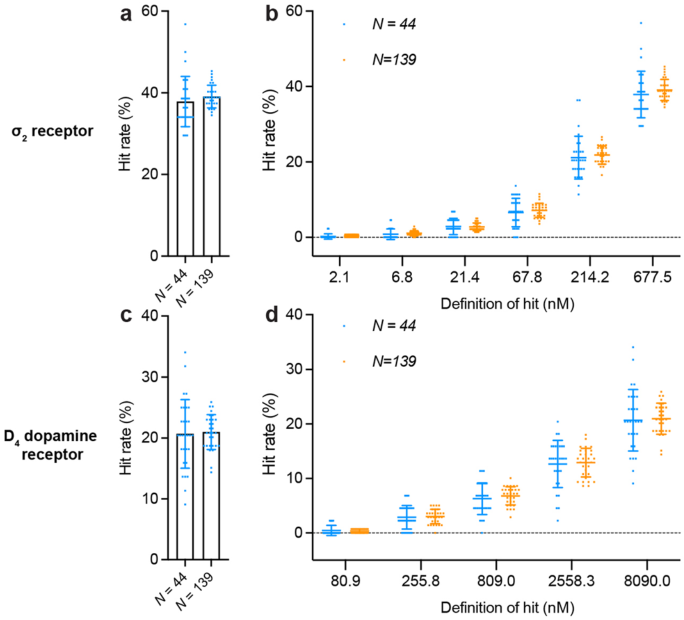


Extended Data Fig. 4 | Electron density omit maps of the AmpC inhibitors. a-d, Polder omit maps of the inhibitors (3σ).

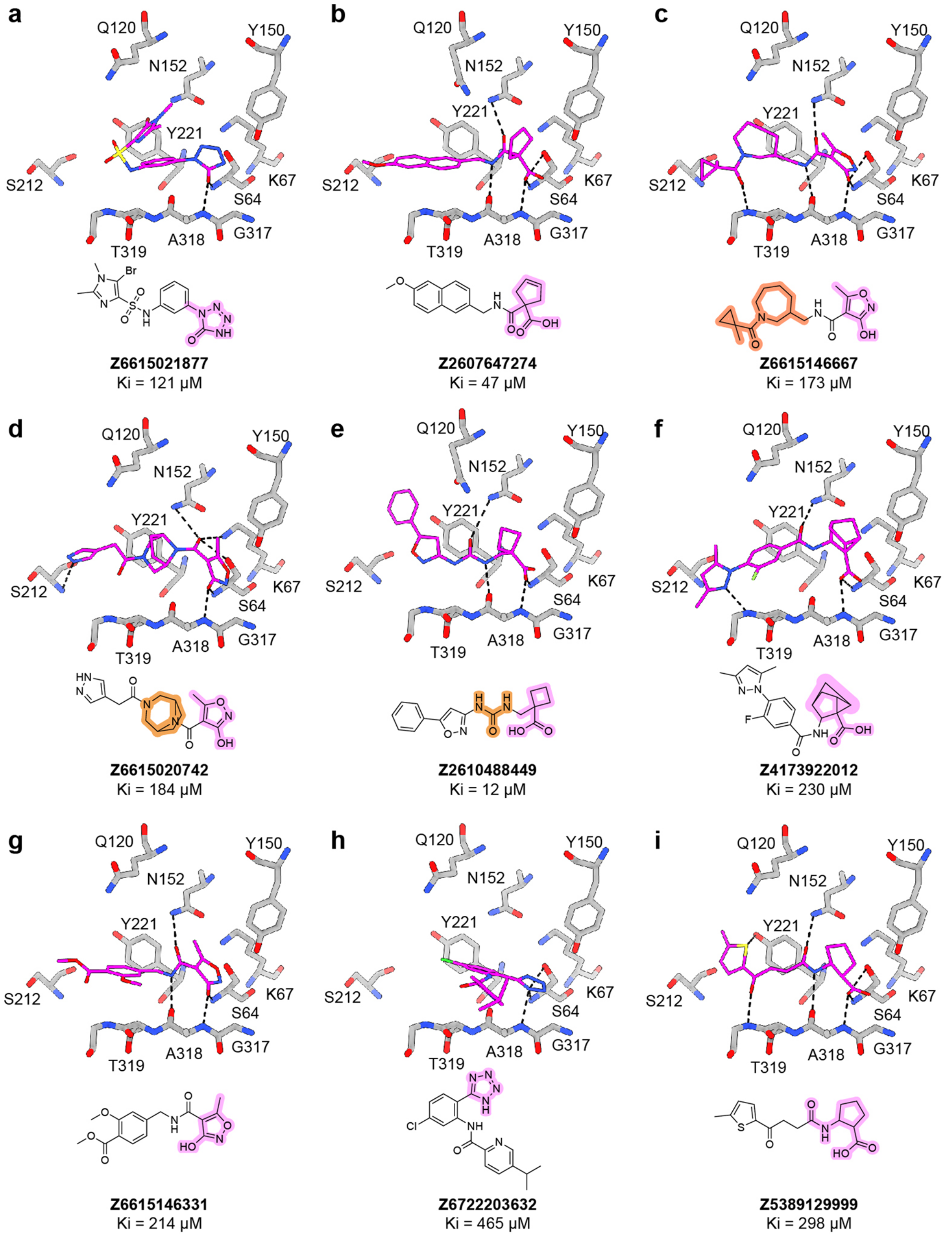


	"Big" from "Big" (22.2%)	"Big" from "Small" (11.4%)	"Small" from "Big" (22.3%)	"Small" from "Small" (11.3%)
"Big" from "Big" (22.2%)	—	p < 0.0001	p = 0.5087	p = 0.0225
"Big" from "Small" (11.4%)	—	—	p = 0.0374	p = 0.4804
"Small" from "Big" (22.3%)	—	—	—	p = 0.1069
"Small" from "Small" (11.3%)	—	—	—	—

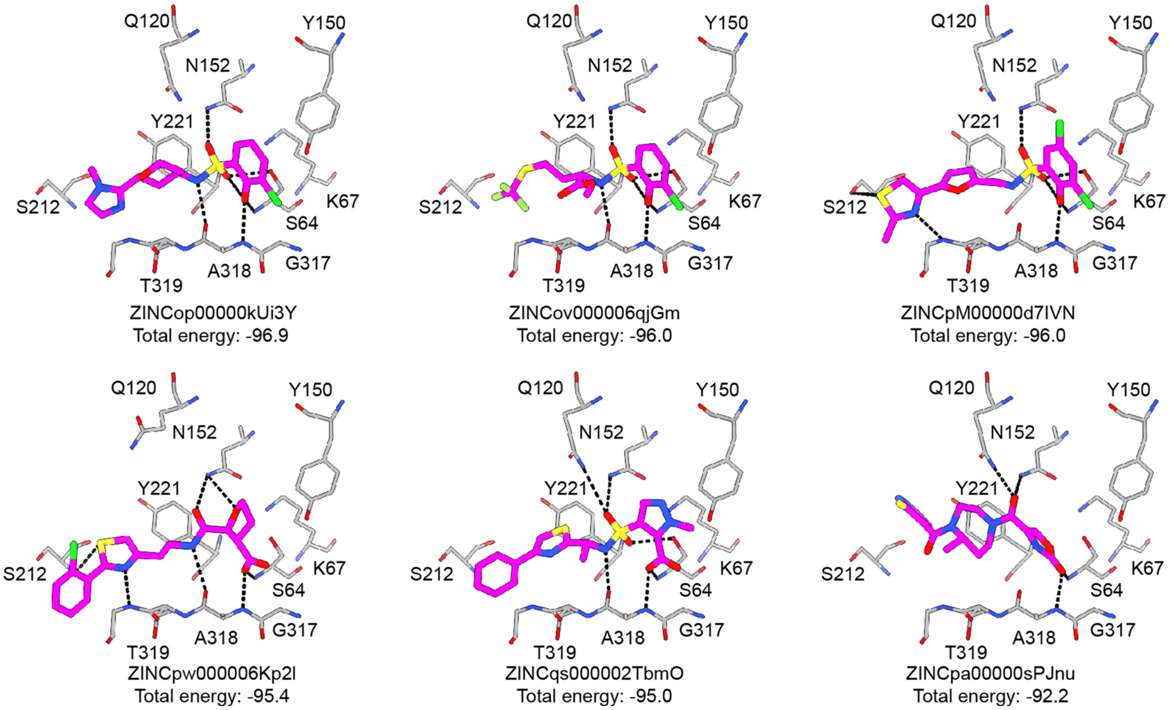
Extended Data Fig. 5 | Comparative analysis of hit rates from large-scale and small-scale AmpC screens with statistical validation. a, The hit rates (number of actives/total tested) of the 1.7 Billion screen (blue bar; 8.26%) versus the 99 Million screen (orange bar; 2.27%) with a hit defined as less than 100 µM. **b**, The hit rates (number of actives/total tested) of the 1.7 Billion screen (blue bar; 2.47%) versus the 99 Million screen (orange bar; 2.27%) with a hit defined as less than 30 µM. **c**, The hit rates of all manually picked molecules of the 1.7 Billion screen (blue bar; 21.14%) versus the 99 Million screen (orange bar; 11.4%). **d**, The hit rates of the top 44 manually picked molecules of the 1.7 Billion screen (blue bar; 47.7%) versus the 99 Million screen (orange bar; 11.4%). **e**, Hit rates from the manually picked, experimentally tested molecules of the 99 Million and 1.7 Billon screens (44 and 626 molecules, respectively), referred to as the "Small" and "Big" screens. For each set, 44 or 626 molecules were resampled for 10,000 bootstrap iterations, and the mean of the resampled hit rates is shown in parenthesis. *P*-values for the null hypothesis that the difference between two resampled distributions is zero are provided. For panels a-d, a two-sided Z-test was used to compare the hit rates of the two screens, under the assumption that the data followed a normal distribution. For panel e, *P*-values were obtained from a one-tailed non-parametric bootstrap test (10,000 iterations) comparing the means of the resampled distributions, with no assumption of normality.



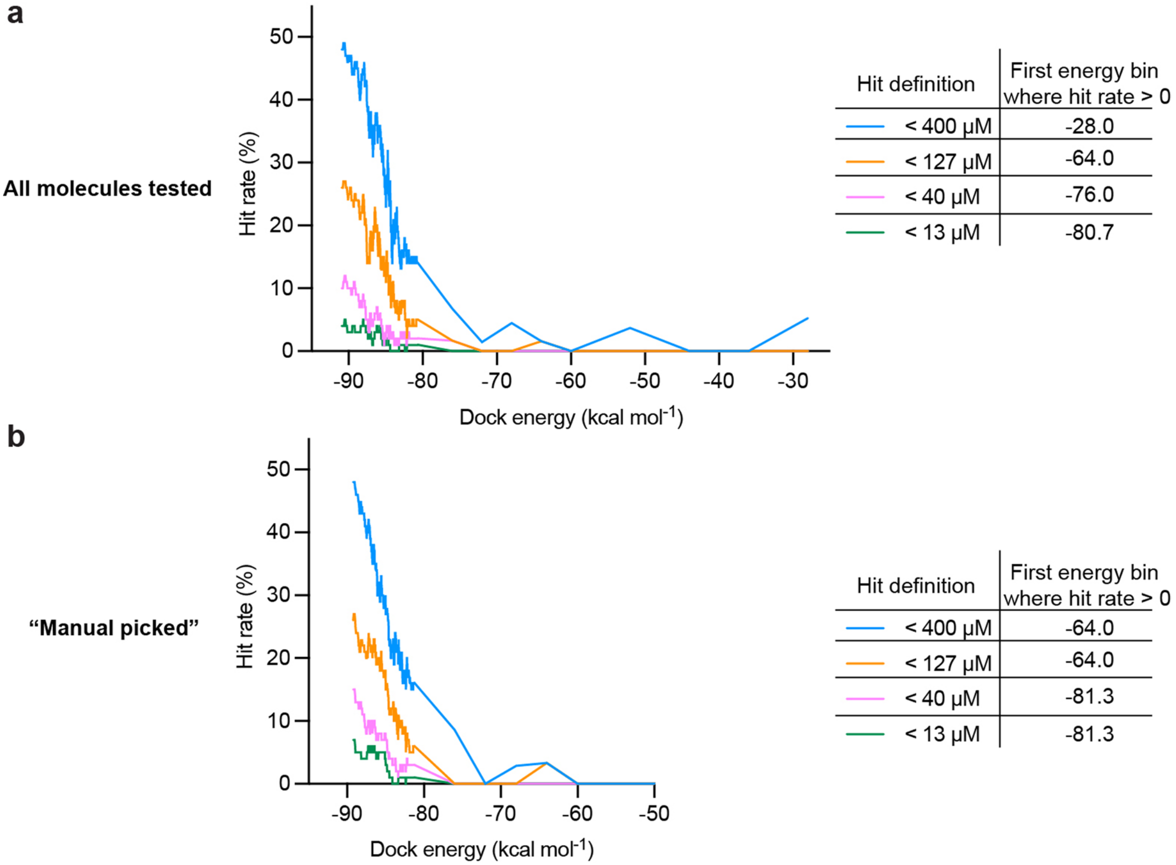
Extended Data Fig. 6 | The impact of testing fewer molecules on hit rate confidence. **a**, For 327 molecules tested against the σ2 receptor, each sample size is randomly drawn 30 times and the resulting hit rates were plotted. The error bars represent s.d.s of the hit rates. **b**, The impact of randomly purchasing 44 and 139 molecules out of 327 molecules for testing on hit rates with different affinity cutoffs. Each sample size is drawn 30 times and the resulting hit rates were plotted. The error bars represent s.d.s of the hit rates. **c**, For 371 molecules tested against the D4 receptor, each sample size is randomly drawn 30 times and the resulting hit rates were plotted. The error bars represent s.d.s of the hit rates. **d**, The impact of randomly purchasing 44 and 139 molecules out of 371 molecules for testing on hit rates with different affinity cutoffs. Each sample size is drawn 30 times and the resulting hit rates were plotted. Data represent mean ± s.d.s of the hit rates.



Extended Data Fig. 7 | Examples of the new warheads and chemotypes from the AmpC screen, in their docked poses in the enzyme active site. **a**, docked pose of Z6615021877 (K_i = 121 µM). **b**, docked pose of Z2607647274 (K_i = 47 µM). **c**, docked pose of Z6615146667 (K_i = 173 µM). **d**, docked pose of Z6615020742 (K_i = 184 µM). **e**, docked pose of Z2610488449 (K_i = 12 µM). **f**, docked pose of Z4173922012 (K_i = 230 µM). **g**, docked pose of Z6615146331 (K_i = 214 µM). **h**, docked pose of Z6722203632 (K_i = 465 µM). **i**, docked pose of Z5389129999 (K_i = 298 µM). The K_i values for Z6615021877 and Z2610488449 were calculated using Lineweaver-Burk plots, while the rest were determined based on the three-point inhibition assays.



Extended Data Fig. 8 | Docking poses of the some of the top scoring molecules. Docking poses of ZINCop00000kUi3Y, ZINCov000006qjGM, ZINCpM00000d7IVN, ZINCpw000006Kp2I, ZINCqs000002TbmO and ZINCpa00000sPJnu are shown.



Hit definition		First energy bin where hit rate > 0
—	< 400 μM	-28.0
—	< 127 μM	-64.0
—	< 40 μM	-76.0
—	< 13 μM	-80.7

Hit definition		First energy bin where hit rate > 0
—	< 400 μM	-64.0
—	< 127 μM	-64.0
—	< 40 μM	-81.3
—	< 13 μM	-81.3

Extended Data Fig. 9 | Hit rate of experimentally tested compounds plotted against DOCK scores with different affinity cutoffs. a, Hit rates of all compounds tested (1,447 well-behaved molecules among 1,521 purchased) plotted against DOCK scores with four different affinity cutoffs: < 400, <137, <40 and <13 μM. **b**, Hit rates of manually picked compounds (687 compounds) plotted against DOCK scores with four different affinity cutoffs: <400, <137, <40 and <13 μM.

nature portfolio

Corresponding author(s):	Yurii Moroz, John J. Irwin, Brian K. Shoichet
Last updated by author(s):	Oct 25, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	□	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	□	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Software used are is publicly available: DOCK 3.8
Data analysis	GraphPad Prism 9, ChimeraX 1.8, Reduce v2, AUMBER v18 (2017), Corina v3.6.0026, Omega v.2.5.1.4, QNIFFT 2.2, python 2.7, python 3.7, ChemAxon Jchem 21.13, RDKit 2020.09.1.0 package (https://www.rdkit.org), PHENIX 1.21.2-5419, Coot 0.9.8.95

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The compounds docked in this study are freely available from the ZINC20 and ZINC22 databases, https://zinc20.docking.org and https://cartblanche22.docking.org. All compounds tested can be purchased from Enamine. Compound information including their ZINC ID, catalog ID, SMILES, DOCK score, ranking, and affinity can be found in Supplementary Table 1. The synthetic procedures and purity information for the hits can be found in the Supplementary Data 5 and Supplementary Table

6. Extensive docking-related files can be found at https://lsd.docking.org. DOCK3.8 is freely available for non-commercial research at https://dock.compbio.ucsf.edu/DOCK3.8/. A web-based version is available without restriction at https://blaster.docking.org/. X-ray structures and maps are available in the Protein Data Bank under accession numbers PDBID 9C81 (Z4462773688), PDBID 9C6P (Z6615017509), PDBID 9C84 (Z6615020275), and PDBID 9DHL (Z6615017782) respectively.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	N/A
Reporting on race, ethnicity, or other socially relevant groupings	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The largest docking library size is determined by the number of available molecules at that time in ZINC22 databases
Data exclusions	No data were excluded from the analyses.
Replication	Findings were reliably replicated. The number of experimental replicates are specified in the legends of all figures.
Randomization	All samples mentioned below were allocated randomly.
Blinding	Investigators were not blinded. Blinding is not necessary for experiments in this study, because the nature of these experiments do not require subject assessment of the data that may influence the validity of the results.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*