Deep contrastive learning enables genome-wide virtual screening

Yinjun Jia[1,2,3,4,7,*], Bowen Gao[1,5,*], Jiaxin Tan[2,4,6,7,*], Xin Hong[1,*], Wenyu Zhu[1], Haichuan Tan[1,5], Yuan Xiao[2,4,6,7], Yanwen Huang[8], Yue Jin[2,3,4,7], Yafei Yuan[2,4,6,7], Jiekang Tian[9], Weiying Ma[1], Yaqin Zhang[1], Chuangye Yan[2,4,6,7,#], Wei Zhang[2,3,4,7,#], Yanyan Lan[1,6,10,#]

[1] Institute for AI Industry Research (AIR), Tsinghua University, Beijing, China.

[2] School of Life Sciences, Tsinghua University, Beijing, China.

[3] IDG/McGovern Institute for Brain Research, Tsinghua University, Beijing, China.

[4] Tsinghua-Peking Center for Life Sciences, Beijing, China.

[5] Department of Computer Science and Technology, Tsinghua University, Beijing, China.

[6] Beijing Frontier Research Center for Biological Structure, Tsinghua University, Beijing, China

[7] State Key Laboratory of Membrane Biology, Tsinghua University, Beijing, China

[8] State Key Laboratory of Natural and Biomimetic Drugs, School of Pharmaceutical Sciences, Peking University, Beijing 100191, China

[9] Center of Pharmaceutical Technology, School of Pharmaceutical Sciences, Tsinghua University, Beijing 100084, China

[10] Beijing Academy of Artificial Intelligence, Beijing, China.

* Contribute equally to this work.

Correspondence and requests for materials should be addressed to Y.L. (lanyanyan@air.tsinghua.edu.cn), C.Y. (yancy2019@tsinghua.edu.cn) and W.Z. (wei_zhang@mail.tsinghua.edu.cn).

Abstract

Numerous protein-coding genes are associated with human diseases, yet approximately 90% of them lack targeted therapeutic intervention. While conventional computational methods such as molecular docking have facilitated the discovery of potential hit compounds, the development of genome-wide virtual screening against the expansive chemical space remains a formidable challenge. Here we introduce DrugCLIP, a novel framework that combines contrastive learning and dense retrieval to achieve rapid and accurate virtual screening. Compared to traditional docking methods, DrugCLIP improves the speed of virtual screening by several orders of magnitude. In terms of performance, DrugCLIP not only surpasses docking and other deep learning-based methods across two standard benchmark datasets but also demonstrates high efficacy in wet-lab experiments. Specifically, DrugCLIP successfully identified agonists with < 100 nM affinities for $5HT_{2A}R$, a key target in psychiatric diseases. For another target NET, whose structure is newly solved and not included in the training set, our method achieved a hit rate of 15%, with 12 diverse molecules exhibiting affinities better than Bupropion. Additionally, two chemically novel inhibitors were validated by structure determination with Cryo-EM. Building on this foundation, we present the results of a pioneering trillion-scale genome-wide virtual screening, encompassing approximately 10,000 AlphaFold2 predicted proteins within the human genome and 500 million molecules from the ZINC and Enamine REAL database. This work provides an innovative perspective on drug discovery in the post-AlphaFold era, where comprehensive targeting of all disease-related proteins is within reach.

Introduction

The human genome comprises approximately 20,000 protein-coding genes (*1*), many of which are related to a variety of diseases. Despite this, only about 10% of these genes have been successfully targeted by FDA-approved drugs or have documented small-molecule binders in the literature (*2*). This leaves a substantial portion of the druggable genome largely unexplored, representing a promising opportunity for therapeutic innovation. The scientific community is eager to translate biological relevant targets into pharmaceutical breakthroughs. However, most researchers lack access to advanced high-throughput screening equipment or sufficient computational power to perform comprehensive virtual screenings. Additionally, proteins often function as parts of families or pathways, indicating that targeting single proteins may not always be the most effective strategy (*3, 4*). These limitations can significantly reduce the success rate of drug discovery, especially for new targets. Therefore, developing a comprehensive chemical database containing genome-wide virtual screening results would be an invaluable asset for the biomedical research community, with the potential to significantly accelerate the discovery of new drugs.

Given the impracticality of experimentally screening all human proteins, virtual screening has emerged as the only viable approach to tackle the vast number of potential targets. In classical computer-aided drug discovery (CADD), molecular docking serves as a foundational technique for target-based virtual screening. Despite advancements in simplified scoring functions, optimized algorithms, and hardware acceleration (*5-9*), molecular docking remains time-intensive, often requiring several seconds to minutes to evaluate each protein-ligand pair. For example, a recent large-scale docking campaign took two weeks to screen 1 billion molecules against a single target, even with the use of 10,000 CPU cores (*10*). As a result, the computational demands for genome-wide virtual screening are prohibitively high, rendering such efforts impractical with existing technologies.

Artificial intelligence holds great promise for drug discovery. Various deep learning methods have been developed for virtual screening, focusing on predicting ligand-receptor affinities (*11-13*). Yet, applying these methods to large-scale virtual screening still face significant challenges. A primary issue is the inconsistency of affinity values due to heterogeneous experimental conditions (*14, 15*), which may negatively impact the performance of the trained model. Moreover, a notable distribution shift between training datasets and real-world testing scenarios hinders the generalizability of AI models, as real-world virtual screenings often involve a larger proportion of inactive molecules than those represented in the curated training sets (*16*). Additionally, the computational demands of deep learning models, with millions of parameters, pose a crucial bottleneck in inference speed, especially as chemical libraries and target numbers grow. Consequently, there is an urgent need for the development of more efficient and robust AI methodologies to effectively address these challenges.

In this work, we introduce DrugCLIP, a novel contrastive learning approach for virtual screening. Contrastive learning has demonstrated significant success in various applications like image-text retrieval (*17*), enzyme function annotation (*18*), and protein homology detection (*19*). The core innovation of DrugCLIP lies in its ability to distinguish potent binders from non-binding molecules with a given protein pocket by aligning their representations. This approach effectively mitigates the impact of noisy affinity labels and chemical library imbalances that have traditionally challenged virtual screening efforts. Moreover, the inference of DrugCLIP is highly efficient, approximately 100,000 times faster than traditional docking methods and 100 times faster than current machine learning-based approaches.

Comprehensive *in silico* and wet-lab evaluations were conducted to assess the accuracy of the DrugCLIP model. Our model achieved state-of-the-art performance on two widely recognized virtual screening benchmarks, DUD-E (*20*) and LIT-PCBA (*21*), outperforming traditional docking-based screening methods and other deep neural

networks. To further validate its performance, DrugCLIP was applied to screen molecules for two real-world targets: $5HT_{2A}R$ (5-hydroxytryptamine receptor 2A) and NET (norepinephrine transporter). Remarkably, our model identified chemically diverse binders with adequate affinities, which were further validated through functional assays and structure determination. These results provide compelling evidence of the efficacy of our virtual screening method.

Finally, a genome-wide virtual screening was conducted using DrugCLIP on all human proteins predicted by AlphaFold2 (*22, 23*). In this process, we first define pockets for AlphFold predictions with structure alignment (*24*), pocket detection software (*25*), and generative AI models. Next, we screened over 500 million drug-like molecules from the ZINC (*26, 27*) and Enamine REAL (*28*) databases against identified pockets. Notably, this unprecedented large-scale virtual screening was completed in just 24 hours on a single computing node equipped with 8 A100 GPUs. Lastly, we applied a CADD cluster-docking pipeline to select chemically diverse and physically proper molecules for each pocket. These result a dataset containing over 2 million potential hits targeting more than 20,000 pockets from around 10,000 human proteins. To the best of our knowledge, this is the first virtual screening campaign to perform more than 10 trillion of scoring operations on protein-ligand pairs, covering nearly half of the human genome. All molecules, scores, and poses have been made freely accessible at https://drug-the-whole-genome.yanyanlan.com, facilitating further research in drug discovery on a genome-wide scale.

Results

The design and *in-silico* evaluation of the DrugCLIP model

Unlike previous machine learning models that relied on regression to directly predict protein-ligand affinity values, DrugCLIP (Fig. 1) redefines virtual screening as a dense retrieval task. The key innovation lies in its training objective, which aims to learn an aligned embedding space for protein pockets and molecules, encoded by separated neural networks. Vector similarity metrics can then be employed to reflect their binding probability. Using contrastive loss during training, the similarity between protein pockets and their binders (positive protein-ligand pair) is maximized, whereas the similarity between protein pockets and molecules binding to other targets (negative protein-ligand pairs) is minimized.

The training process of DrugCLIP includes two stages: pretraining and fine-tuning. The molecule and pocket encoders are pretrained with large-scale synthetic data and are further refined using experimentally determined protein-ligand complex structures during fine-tuning.

In the pretraining stage, the molecule encoder is initialized with Uni-Mol, a well-established molecule encoder. With the molecule encoder frozen, the pocket encoder is randomly initialized and trained to align with the molecule encoder using contrastive learning (Fig. 1B). we developed a Protein Fragment-Surrounding Alignment (ProFSA) framework (Fig. 1A) to generate large-scale synthetic data specifically tailored for contrastive pretraining. In this approach, short peptide fragments are extracted from protein-only structures to serve as pseudo-ligands, while their surrounding regions are designated as pseudo-pockets. Appling ProFSA framework to PDB (*29*) data yielded 5.5 million pseudo-pocket and ligand pairs to facilitate the pretraining. The trained pocket encoder has been evaluated across various downstream tasks such as pocket matching (Supplementary Table 1), pocket property prediction (Supplementary Table

2) and protein-ligand affinity prediction (Supplementary Table 3). Experimental results demonstrate that our pretrained pocket encoder exhibits strong performance, even in a zero-shot setting, outperforming many supervised learning-based models as well as physical and knowledge-based models. These results underscore the success of the pretraining stage to obtaining meaningful molecule and pocket representations.

After pretraining, the molecule and pocket encoders are further fine-tuned (Fig. 1D) using 40,000 experimentally determined protein-ligand complex structures collected by BioLip2 database (*30*). Given that the binding conformations of molecules are unknown and only their topologies are provided in virtual screening, we implemented a random conformation sampling strategy for data augmentation by using RDKit (*31*) for conformation generation. This augmentation allows DrugCLIP to train on data that more accurately reflect the variability of real-world screenings, thereby enhancing the model's performance and generalization ability.

In the screening process (Fig. 1E), we first use our trained encoders to represent molecules and pockets to vectors. Cosine similarities between the pocket and molecule embeddings are then computed, and candidate molecules are ranked according to these similarity scores. Since the molecule representations can be computed offline, DrugCLIP screening is highly efficient, requiring only the calculation of a simple cosine similarity and subsequent ranking. Compared with traditional computational method like docking, DrugCLIP is more than 100,000 times faster.

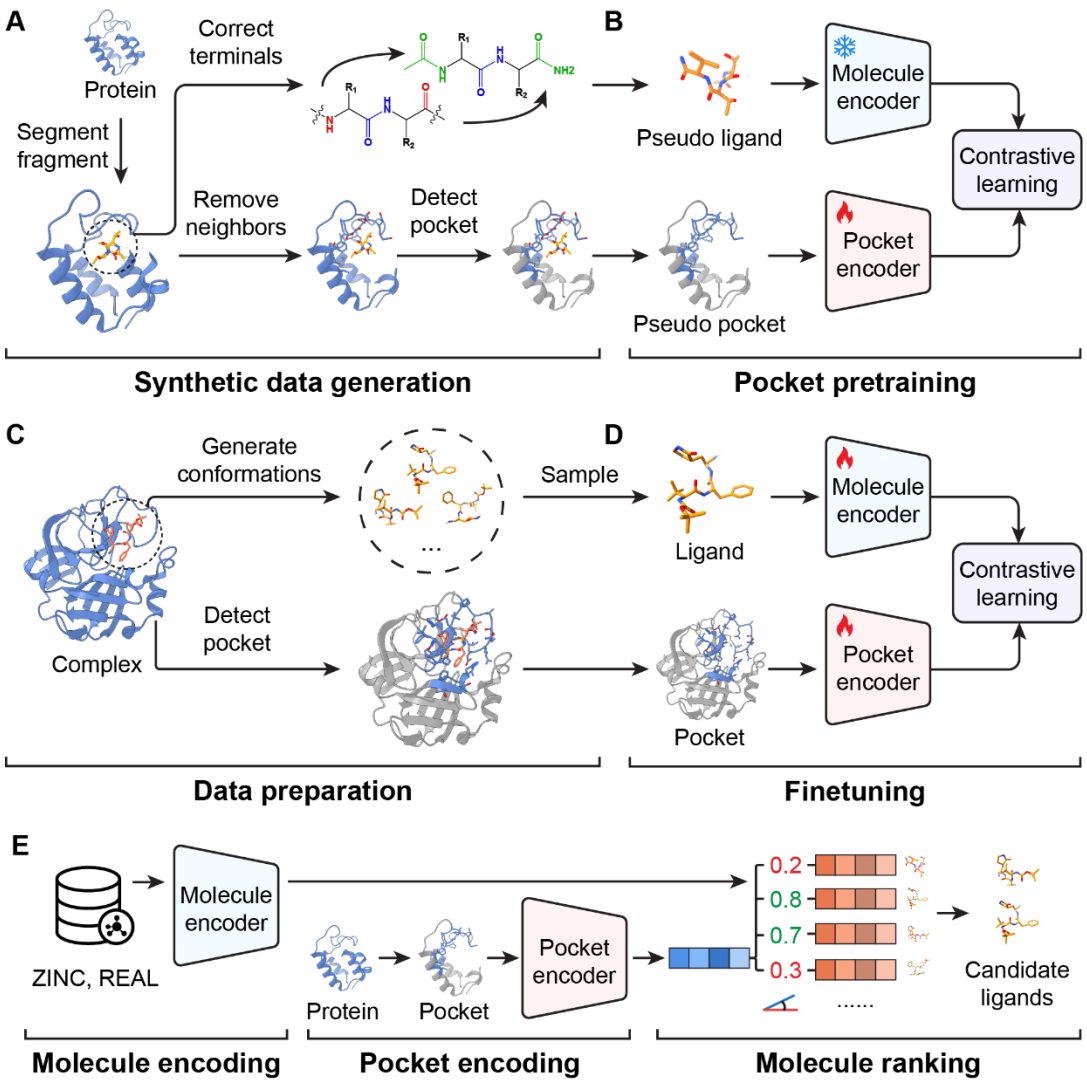

Fig.1 The framework of DrugCLIP. **(A)** In the pretraining stage, a large-scale synthetic data was created using the ProFSA strategy. Specifically, pseudo pocket-ligand pairs were constructed through a series of operations, including fragment segmentation, terminal correction, neighbor removal and pocket detection, on protein data. **(B)** The pocket encoder is pretrained with pseudo pocket-ligand pairs in a contrastive distillation manner to transfer knowledge from a well-established molecular encoder to the pocket encoder. **(C)** During the fine-tuning process, experimentally determined protein-ligand pairs were used as training data, with multiple ligand conformations generated by RDKit. **(D)** In the fine-tuning stage, both the pocket and molecule encoders were updated using a contrastive loss, which maximizes the similarity between positive pairs and minimize it between negative pairs. **(E)** The pipeline for virtual screening with DrugCLIP. The candidate molecules from the library were pre-encoded with the trained molecular encoder. For a given pocket, the trained pocket encoder converts it to a vector, and the cosine similarity is then utilized to select top ligands with the highest scores.

We benchmarked DrugCLIP on two widely used virtual screening datasets, DUD-E (*20*) and LIT-PCBA (*21*). The DUD-E dataset contains 22,886 active compounds of 102

protein targets. For each active compounds, 50 decoys with similar physical properties but different structures are generated. In contrast, LIT-PCBA comprises approximately 8,000 active and 2.64 million inactive compounds across 15 targets, derived from experimental results of the PubChem BioAssay database. DrugCLIP was compared with established physical-informed docking software including Glide-SP (*5*), Autodock Vina (*6*), Surflex (*32*), and regression-oriented machine learning models including NNscore (*13*), RFscore (*33*), Pafnucy (*34*), OnionNet (*12*), PLANET (*11*), Gnina (*35*), BigBind (*36*). In both sets of results (Fig. 2A and 2B), DrugCLIP demonstrated a substantial improvement over all baseline methods in terms of EF1%, measuring the recall capacity of virtual screening models.

We also investigated the influence of homology information and protein structure accuracy on DrugCLIP's performance. Remarkably, even when test protein families were entirely excluded from the training set, DrugCLIP still outperform one of the most popular virtual screening method AutoDock Vina (Fig. 2C), highlighting its strong generalization capability to new targets. Moreover, DrugCLIP shows exceptional robustness by outperforming AutoDock Vina even with a 3 Å RMSD error in the side chain conformations of protein pockets (Fig. 2D), indicating its robustness to structural inaccuracies.

Furthermore, DrugCLIP is exceptionally efficient (Fig. 2E), making it highly suitable for large-scale screening tasks. For instance, DrugCLIP completed the screening for LIT-PCBA in merely 38 seconds with the aid of pre-encoding acceleration, significantly faster than Glide docking (3 days), Uni-Dock (22 hours) (*8*), and another machine learning method PLANET (3 hours) (*11*). Moreover, the time consumption of DrugCLIP screening scales linearly with the simultaneously increase of target and molecule numbers (Fig. 2F), which can facilitate multi-target virtual screening.

These *in silico* results confirm that DrugCLIP possesses superior virtual screening capabilities, combining high performance, generalizability, robustness, and efficiency.

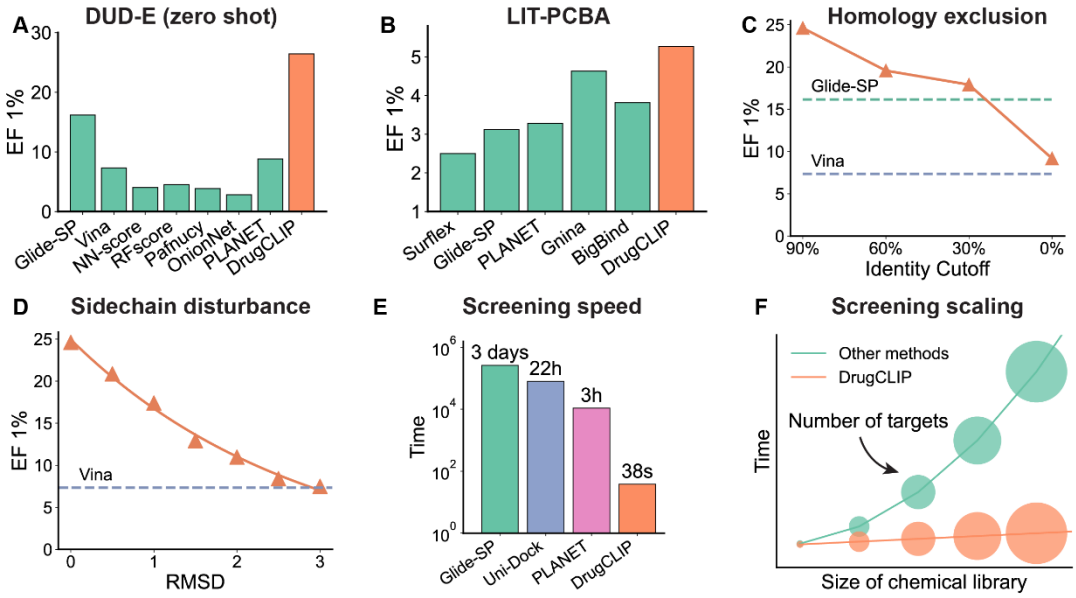

Fig.2 *In silico* benchmarking results of DrugCLIP. **(A)** The evaluation of DrugCLIP on the DUD-E dataset using the EF1% to assess model performance. The results of baseline models are token from previous studies (*11, 37*). **(B)** The evaluation of DrugCLIP on the LIT-PCBA dataset, also using the EF1% for performance measurement. The results of baseline models are token from previous studies (*11, 21, 36, 38*). **(C)** The assessment of DrugCLIP's generalization ability was conducted by varying the identity cutoffs between testing targets and training data in DUD-E, with Glide-SP and Vina represented as dashed lines. **(D)** The evaluation of DrugCLIP's robustness regarding errors in pocket side-chain conformations, was conducted by using RMSD values ranging from 0 Å to 3 Å, with Vina shown as a dashed line for reference. **(E)** The screening speed on the LIT-PCBA dataset, compared with docking methods like Glide-SP and Uni-Dock, and the machine learning model PLANET. Speeds of baseline methods are token from previous studies (*8, 11*). **(F)** An illustration of time consumption as the screening scale increases, with the x-axis representing the size of the molecule library and the size of the circles representing the numbers of targets. DrugCLIP (the orange line) has a computational complexity of O(M+N), where M is the number of targets and N is the number of compounds, whereas most existing methods (the green line) have a complexity of O(MN).

Validating the performance of DrugCLIP with wet-lab experiments

In addition to *in silico* evaluation, we tested the DrugCLIP model on real-world targets using wet-lab experiments. We focused on two well-established targets for psychiatric diseases: the serotonin receptor 2A ($5HT_{2A}R$) and the norepinephrine transporter (NET).

235 $5HT_{2A}R$ is an emerging target for antidepressant development. Psychedelics such as
236 lysergic acid diethylamide (LSD) and psilocybin act as agonists of $5HT_{2A}R$. Despite
237 their potent hallucinogenic effects, these psychedelics have demonstrated strong and
238 long-lasting antidepressant effects in both rodent models and humans (*39, 40*), making
239 $5HT_{2A}R$ a crucial target for the development of next-generation antidepressants.
240 Previous research suggests that the recruitment of β-arrestin2 following $5HT_{2A}R$
241 activation is a key biochemical mechanism underlying these antidepressant effects (*41,*
242 *42*).

243 Since the structure of $5HT_{2A}R$ was included in our training data, we trained a
244 customized DrugCLIP model more specifically for $5HT_{2A}R$ virtual screening. We
245 excluded similar experimental structures (90% homology) from the training dataset to
246 prevent potential data leakage. A library of 1,648,137 in stock molecules from
247 ChemDiv were screened, and 78 compounds from the top results were ordered,
248 selecting with criteria including drug-likeness, docking score, and chemical diversity.

249 Firstly, a calcium flux assay was used for a primary screening of these compounds in
250 an agonist model at a concentration of 10 μM. Eight of the 78 molecules were identified
251 as positive agonists, exhibiting a minimal activity of 10% compared to serotonin. The
252 affinities of these molecules to $5HT_{2A}R$ were further assessed using [^{3}H]-labeled
253 ketanserin competitive binding assays, with six showing a K_i of less than 10 μM. We
254 then evaluated the cellular function of these hit molecules using NanoBit assays for β-
255 arrestin2 recruitment, and all six molecules achieved an EC_{50} of less than 1 μM.

256 Among the six validated agonists, two chemically novel molecules are particularly
257 interesting. The molecule with the highest affinity, V008-4481, binds to the orthosteric
258 (*43*), extended (*44*) and side-extended (*41*) pocket of $5HT_{2A}R$ with its indole ring,
259 phenylpiperazine part and linear tail respectively. In previous researches, molecules that
260 bind to all three regions are rarely seen. This molecule achieves an affinity of 21.0 nM
261 and exhibits an EC_{50} of 60.3 nM with an E_{max} of 35.8% in the NanoBit assay. Another

262 molecule, V006-3328, features a unique branched topology uncommon among $5HT_{2A}R$
263 binders. It has an affinity of 3.51 μM, with an EC_{50} of 599 nM and an E_{max} of 23.4% in
264 the NanoBit assay. These two molecules present strong potential as starting points for
265 the discovery of new serials of $5HT_{2A}R$ agonists.

266 Unlike $5HT_{2A}R$, the norepinephrine transporter (NET) is a well-established drug
267 target for depression and attention deficit hyperactivity disorder (ADHD), with multiple
268 FDA-approved inhibitors (*45*). However, the structures of NET with or without its
269 inhibitors in complexes were not solved until 2024 (*46-48*). The closest protein
270 structure in our dataset is the dopamine transporter from *Drosophila* (*49*), which shares
271 less than 60% similarity with NET. Therefore, screening against NET provides a more
272 challenging test of our model's ability to generalize to relatively new targets.

273 For this target, we directly utilized the DrugCLIP model trained with all collected
274 complex structures to screen the ChemDiv in-stock library. We ultimately selected 104
275 compounds considering chemical novelty and diversity. We tested their inhibition of
276 NET protein by measuring the transport of [^{3}H]-labeled norepinephrine in NET-
277 containing liposomes. Among these compounds, 15% of them exhibited more than 60%
278 inhibition of NET, with 12 compounds demonstrating greater potency than the widely
279 used antidepressant bupropion.

280 Unlike previous NET inhibitors that typically feature aliphatic nitrogen atoms
281 capable of forming a salt bridge interaction with ASP75 of NET (*46-48*), our screening
282 identified several hits with positively charged aromatic nitrogen atoms. Notably, two
283 such molecules, 0086-0043 and Y510-9709, demonstrated better IC_{50} (with values of
284 1.14 μM and 0.31 μM, respectively) than bupropion (1.5 μM). Structural analysis of the
285 complexes between these compounds and the NET protein revealed that the aromatic
286 rings indeed form more favorable interactions with NET: the isoquinoline ring of 0086-
287 0043 engages in a T-shaped π-π interaction with PHE72, and the thiazole ring of Y510-
288 9709 likely interacts with surrounding aromatic side chains like PHE323. These

findings highlight the potential of DrugCLIP model to providing new chemical insights for drug discovery.

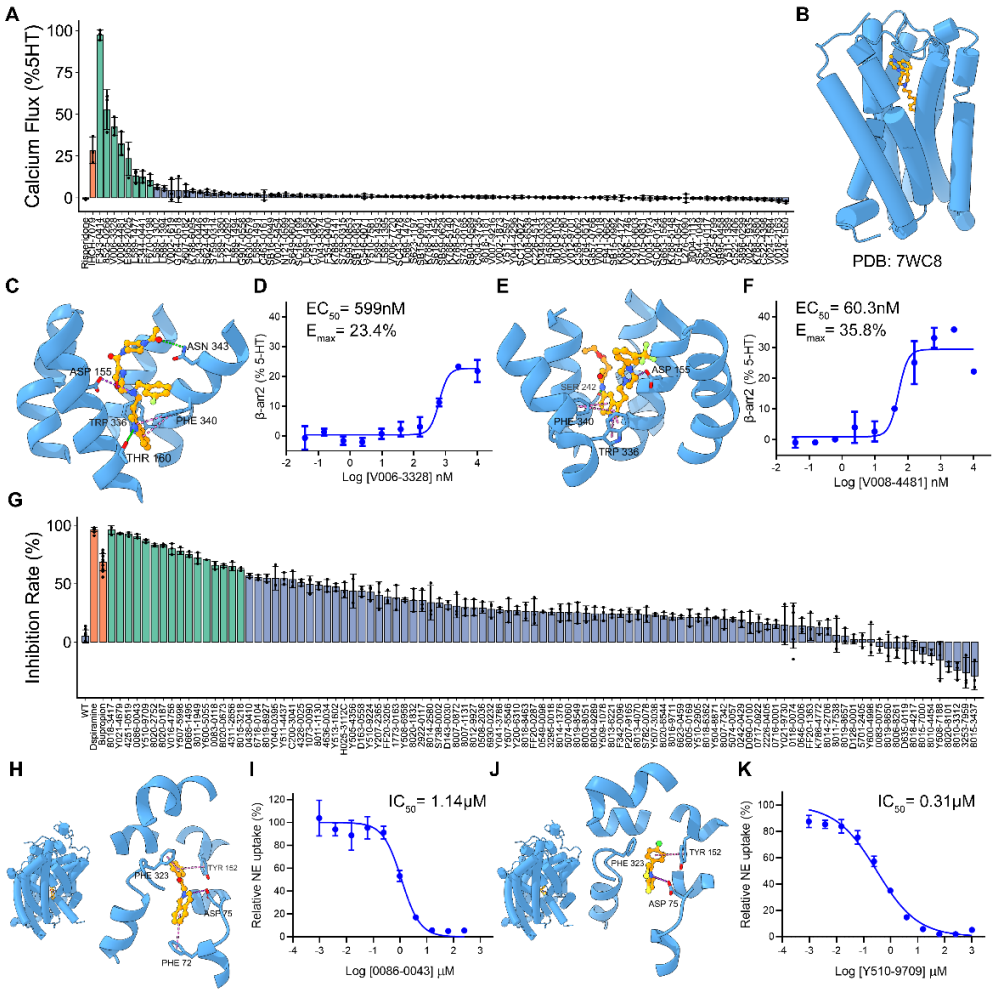

Fig.3 Wet-lab validations of DrugCLIP. **(A)** The screening results of 78 DrugCLIP identified molecules using calcium flux assays for $5HT_{2A}R$ agonist at a concentration of 10 μM. Eight molecules showed signals larger than 10%. **(B)** The 7WC8 structure was used as the receptor for docking analysis. **(C)** The docking pose of V006-3328 to $5HT_{2A}R$. **(D)** The dose response curve of V006-3328 in the NanoBit assay measuring the β-arrestin2 recruitment. **(E)** The docking pose of V008-4481 to $5HT_{2A}R$. **(F)** The dose response curve of V008-4481 in the NanoBit assay measuring the β-arrestin2 recruitment. **(G)** The evaluation of 104 DrugCLIP identified molecules with radio-ligand transportation assays for NET inhibitor at a concentration of 10 μM, and 16 molecules showed inhibition larger than 60% **(H)** The complex structure of 0086-0043 and NET determined with Cryo-EM. **(I)** The dose response curve of 0086-0043 in the radio-ligand transportation assay. **(J)** The complex structure of Y510-9709 and NET determined with Cryo-EM. **(K)** The dose response curve of Y510-9709 in the radio-ligand transportation assay.

306 **Genome-wide virtual screening with DrugCLIP**

307 After validating the DrugCLIP model through both *in silico* and wet-lab experiments,
308 we apply it to a genome-wide virtual screening campaign. The effort is grounded on
309 significant advancements in protein structure prediction, specifically targeting all
310 human protein structures predicted by AlphaFold2 (*22, 23*).

311 Virtual screening using AlphaFold-predicted structures remains a topic of debate.
312 The primary concern is that these predicted structures may lack the accuracy needed to
313 replicate experimental conformations and effectively filter out inactive molecules (*50,*
314 *51*). Despite this, some studies have shown that virtual screening with AlphaFold-
315 predicted structures can still yield reasonable results for certain targets (*52, 53*). Despite
316 of the robustness of DrugCLIP to sidechain inaccuracies (Fig. 2D), we further assess
317 the influence of predicted structure by replacing the structures in the DUD-E dataset
318 with AlphaFold predictions and using Fpocket (*25*) to identify the nearest pockets. This
319 substitution resulted in a significant performance drop for DrugCLIP, with the EF1%
320 value decreasing from 27.4% to 13.3% (Fig.4C), reflecting similar challenges observed
321 with docking methods in both virtual screening (*50*) and conformation prediction
322 (Fig.4B).

323 To improve the utilities of AlphaFold-predicted structures, we developed a strategy
324 called GenPack (Generation-Packing, Fig.4A). This strategy involves training
325 molecular generative models conditioned on the backbone structures of protein pockets.
326 While the generated molecules may not always be synthesizable, they help to localize
327 pockets more precisely and induce the pocket conformation into a more suitable state.
328 After this generation step, side chains are reintroduced, and the overall conformation is
329 refined using physical force fields. With the GenPack strategy, we significantly
330 enhanced the screening power of AlphaFold-predicted structures, increasing EF1%
331 value on the DUD-E benchmark from 13.3% to 19.5% (Fig.4C) and improving the
332 docking success rate from 19.1% to 38.7% (Fig.4B).

Finally, we introduced a genome-wide virtual screening pipeline to facilitate future drug discovery. We began with splitting AlphaFold predictions into high-confidence regions based on plDDT and PAE scores. For each region, we used homology alignment and Fpocket (*25*) along with GenPack to detect potential pockets. The DrugCLIP model was then employed to screen over 500 million drug-like molecules from the ZINC (*26, 27*) and Enamine REAL (*28*) databases. The screening process, which involved more than 10 trillion of scoring operations on protein-ligand pairs, was completed in about 24 hours on a single computing node equipped with 8 A100 GPUs. The top-ranked molecules were then clustered and further evaluated using molecular docking, filtering out poor poses with Glide score < -6 kcal/mol. The final database contains over 2 million potential hits molecules for more than 20,000 pockets from 10,000 human targets. All molecules, docking scores and poses have been made freely accessible at https://drug-the-whole-genome.yanyanlan.com, facilitating further research and drug discovery processes.

Our genome-wide screening results cover a more extensive range of targets than ChEMBL (*54*), one of the most comprehensive databases for bioactive molecules. While UniProt (*1*) contains 20,436 reviewed human proteins, the latest ChEMBL release (ChEMBL 34) covers 4,810 of them. Moreover, not all targets in the ChEMBL database have high-affinity small molecule binders; some targets only have peptide or antibody binders, or merely vague results from low-quality assays. In contrast, our database spans 9,908 targets, more than twice the number in ChEMBL and covers nearly half of the human genome. To visualize the difference between the two protein spaces, we encoded all protein sequences using the ESM1b model (*55*). The t-SNE plot shows that our space encompasses a broader range of proteins, including many that are not closely related to those in ChEMBL.

Our database includes a diverse range of targets, from well-studied proteins to less-explored members of well-known families, as well as proteins with limited

360 pharmacological understanding. For example, the c-Jun N-terminal kinase 3 (JNK3) is
361 a classical kinase target with many ligand-bound crystal structures (*56, 57*). DrugCLIP
362 identified molecules binds to the ATP-binding pockets, forming H-bonds with
363 backbone atoms of MET149 in the hinge region. SLC45A2 belongs to the solute carrier
364 (SLC) superfamily, many of whose members are important drug targets. Nevertheless,
365 SLC45A2 has limited pharmacological studies. This gene plays a crucial role in
366 pigmentation (*58*) and is widely expressed in cutaneous melanomas (*59*), with evidence
367 suggesting its oncogenic potential (*60*). All molecules in the database could bind near
368 L374, which is an important site for protein stability (*58*), thus having potential
369 modulatory effects. Another interesting example OR6A2 belongs to the olfactory
370 receptor family, whose members are mainly found to be expressed in olfactory receptor
371 neurons, yet many of them are expressed in various other tissues with unexplored
372 pharmaceutical potentials (*61*). OR6A2 is expressed in macrophages, sensing blood
373 octanal and promoting the formation of atherosclerotic plaques (*62*). Our predicted
374 molecules fit the orthosteric pocket of OR6A2 and can serve as potential inhibitors for
375 treating atherosclerosis. The final example Sestrin-2 can sense leucine (*63*) and promote
376 drug resistance of cancer cells (*64*), which belongs to a unique highly-conserved stress-
377 inducible protein family (PF04636 or IPR006730) with only three members in the
378 human genome. Our database contains predicted molecules bind to the same pocket of
379 leucine (*65*) that may serve as good starting points for anti-cancer therapies. These
380 examples highlight the potential of our database as a valuable resource for exploring
381 the undrugged genome and facilitate future drug discovery.

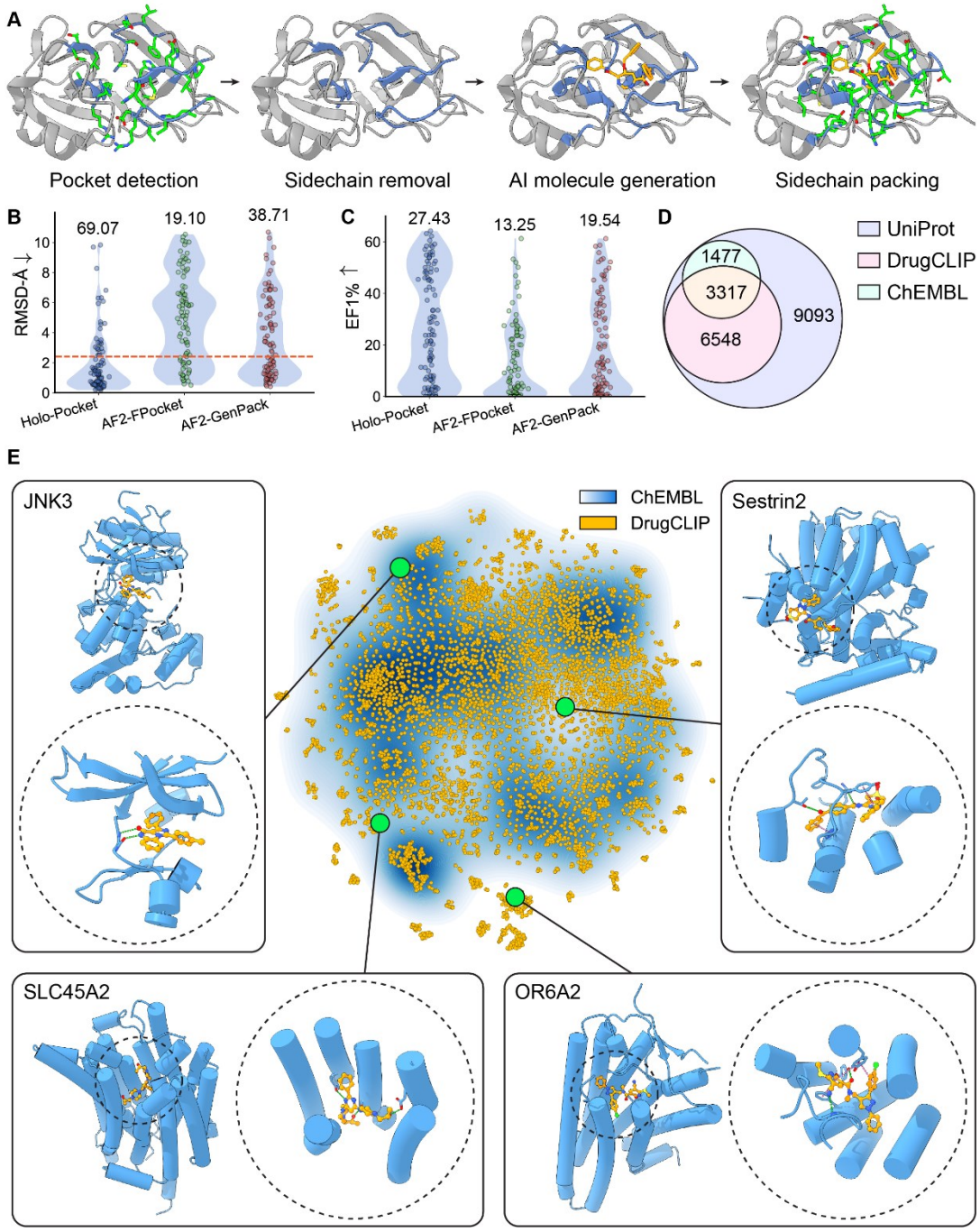

Fig.4 DrugCLIP enables genome-wide virtual screening. **(A)** The GenPack (Generation-Packing) process for extracting pockets from AlphaFold2-predicted structures involves using Fpocket to detect initial pockets, removing sidechains, applying an AI-generative model to create molecules based on the backbone structure, and then performing sidechain packing with the generated molecules. **(B)** The redocking RMSD comparisons for different pocket definitions: holo-pocket, pockets detected by Fpocket on AF2-predicted structures, and pockets generated by GenPack. The red dash line indicates the RMSD threshold of 2 Å, and corresponding docking success rates are labeled above each column. **(C)** The EF1% comparisons for virtual screening on the DUD-E dataset using different pocket definitions:

holo-pocket, pockets detected by Fpocket on AlphaFold2-predicted structures, and pockets generated by GenPack on AlphaFold2 predictions, and averaged EF1% values are labeled above each column. **(D)** The Venn diagram of target numbers in different database, with UniProt, DrugCLIP, and ChEMBL shown as different circles. **(E)** The t-SNE visualization and examples for the genome-wide virtual screening results. Yellow dots indicate targets in our database while the blue-white gradient represents targets in the ChEMBL database, with density ranging from high (blue) to low (white).

Conclusions and Discussions

With the rapid advancement of protein structure prediction methods and the availability of a comprehensive atlas of predicted protein structures for human and disease related species (*23, 66*), we have entered a new era where effective drug discovery for all disease-related targets is within reach. In this paper, we introduce DrugCLIP, a groundbreaking contrastive learning based virtual screening approach that aims to achieve genome-wide drug discovery. The efficacy of DrugCLIP has been rigorously validated through both *in silico* benchmarks and wet-lab experiments. In well-established benchmarks, DrugCLIP consistently outperformed traditional docking software and contemporary machine learning models. Notably, for the $5HT_{2A}R$ and NET targets, DrugCLIP identified diverse high-affinity binders and novel chemical entities. These findings underscore the potential of DrugCLIP model as a reliable tool for virtual screening in real-world drug development. We demonstrate its application through a genome-wide virtual screening campaign, encompassing more than 20,000 pockets across approximately 10,000 human proteins, using a chemical library of 500 million molecules from ZINC and Enamine REAL. Remarkably, DrugCLIP completes this trillion-level virtual screening campaign in just 24 hours using just a single computational node with 8 GPU accelerators. Beyond the screening results, we have generated over 2 million high-confidence protein-ligand complex structures accompanied with their docking score. By making this extensive database freely accessible, we aim to make a substantial contribution to the research community, accelerating drug discovery and fostering innovation in therapeutic development.

DrugCLIP is more than just a new tool. It represents a transformative shift in the development of new therapeutics, heralding a new paradigm in drug discovery. Its genome-wide virtual screening capability opens the door to truly end-to-end drug discovery on a genomic scale, allowing researchers to screen all relevant targets simultaneously, rather than focusing on a few promising targets. This expansive approach facilitates the creation of customized chemical libraries for advanced phenotypic screening with high-fidelity models such as organoids (*67-69*) or humanized mice (*70-72*), potentially reducing failure rates in drug development.

DrugCLIP paves the way for new advancements in AI driven drug discovery. Its outstanding efficiency allows the screening scale to the largest ultra-large chemical library available today, e.g., 48 billion-compound Enamine REAL Space library. This effort pushes the boundaries of what virtual screening can achieve in drug discovery. Moreover, the release of these genome-wide virtual screening results could serve as a valuable resource for molecular generation, particularly through a retrieval-augmented generation approach (*73, 74*), enhancing our capacity for drug discovery and design.

Acknowledgement

We would like to take this opportunity to thank Prof. Jingjing Liu, Dr. Xiaobing Deng, Prof. Zhenming Liu and Dr. Qiwei Ye for their valuable discussions, insightful comments, and assistance with manuscript revisions. This work is supported by the National Key R&D Program of China No.2021YFF1201600, Beijing Academy of Artificial Intelligence, and Beijing Frontier Research Center for Biological Structure. Fundings.

References

1. The-UniProt-Consortium, UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research* **51**, D523-D531 (2022).

2. T. I. Oprea *et al.*, Unexplored therapeutic opportunities in the human genome. *Nature Reviews Drug Discovery* **17**, 317-332 (2018).

3. P. R. Caron *et al.*, Chemogenomic approaches to drug discovery. *Current Opinion in Chemical Biology* **5**, 464-470 (2001).

4. A.-L. Barabási, N. Gulbahce, J. Loscalzo, Network medicine: a network-based approach to human disease. *Nature Reviews Genetics* **12**, 56-68 (2011).

5. T. A. Halgren *et al.*, Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J Med Chem* **47**, 1750-1759 (2004).

6. O. Trott, A. J. Olson, AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem* **31**, 455-461 (2010).

7. O. Korb, T. Stützle, T. E. Exner, Empirical scoring functions for advanced protein-ligand docking with PLANTS. *J Chem Inf Model* **49**, 84-96 (2009).

8. Y. Yu *et al.*, Uni-Dock: GPU-Accelerated Docking Enables Ultralarge Virtual Screening. *Journal of Chemical Theory and Computation* **19**, 3336-3345 (2023).

9. M. Ling *et al.*, Vina-FPGA: A Hardware-Accelerated Molecular Docking Tool With Fixed-Point Quantization and Low-Level Parallelism. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* **31**, 484-497 (2023).

10. C. Gorgulla *et al.*, An open-source drug discovery platform enables ultra-large virtual screens. *Nature* **580**, 663-668 (2020).

11. X. Zhang *et al.*, PLANET: A multi-objective graph neural network model for protein–ligand binding affinity prediction. *Journal of Chemical Information and Modeling* **64**, 2205-2220 (2024).

12. L. Zheng, J. Fan, Y. Mu, OnionNet: A multiple-layer intermolecular-contact-based convolutional neural network for protein–ligand binding affinity prediction. *ACS Omega* **4**, 15956-15965 (2019).

13. J. D. Durrant, J. A. McCammon, NNScore 2.0: a neural-network receptor-ligand scoring function. *J Chem Inf Model* **51**, 2897-2903 (2011).

14. B. T. Burlingham, T. S. Widlanski, An intuitive look at the relationship of K_i

and IC_{50}: A more general use for the Dixon plot. *Journal of Chemical Education* **80**, 214 (2003).

15. G. A. Landrum, S. Riniker, Combining IC50 or Ki values from different sources is a source of significant noise. *Journal of Chemical Information and Modeling* **64**, 1560-1567 (2024).

16. R. Rodríguez-Pérez, M. Vogt, J. Bajorath, Influence of Varying Training Set Composition and Size on Support Vector Machine-Based Prediction of Active Compounds. *J Chem Inf Model* **57**, 710-716 (2017).

17. A. Radford *et al.*, Learning transferable visual models from natural language supervision. *International Conference on Machine Learning*, (2021).

18. T. Yu *et al.*, Enzyme function prediction using contrastive learning. *Science* **379**, 1358-1363 (2023).

19. L. Hong *et al.*, Fast, sensitive detection of protein homologs using deep dense retrieval. *Nature Biotechnology*, (2024).

20. M. M. Mysinger, M. Carchia, J. J. Irwin, B. K. Shoichet, Directory of useful decoys, enhanced (DUD-E): Better ligands and decoys for better benchmarking. *Journal of Medicinal Chemistry* **55**, 6582-6594 (2012).

21. V.-K. Tran-Nguyen, C. Jacquemard, D. Rognan, LIT-PCBA: An Unbiased Data Set for Machine Learning and Virtual Screening. *Journal of Chemical Information and Modeling* **60**, 4263-4273 (2020).

22. J. Jumper *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021).

23. M. Varadi *et al.*, AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research* **50**, D439-D444 (2021).

24. Y. Zhang, J. Skolnick, TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res* **33**, 2302-2309 (2005).

25. V. Le Guilloux, P. Schmidtke, P. Tuffery, Fpocket: An open source platform for ligand pocket detection. *BMC Bioinformatics* **10**, 168 (2009).

26. J. J. Irwin *et al.*, ZINC20—A Free Ultralarge-Scale Chemical Database for Ligand Discovery. *Journal of Chemical Information and Modeling* **60**, 6065-6073 (2020).

27. J. J. Irwin, B. K. Shoichet, ZINC--a free database of commercially available compounds for virtual screening. *J Chem Inf Model* **45**, 177-182 (2005).

28. O. O. Grygorenko *et al.*, Generating Multibillion Chemical Space of Readily Accessible Screening Compounds. *iScience* **23**, 101681 (2020).

29. H. M. Berman *et al.*, The Protein Data Bank. *Nucleic Acids Research* **28**, 235-242 (2000).

30. C. Zhang, X. Zhang, Peter L. Freddolino, Y. Zhang, BioLiP2: an updated structure database for biologically relevant ligand–protein interactions. *Nucleic Acids Research* **52**, D404-D412 (2023).

31. G. Landrum *et al.*, RDKit: Open-source cheminformatics. *Zenodo*, 13469390 (2024).

32. A. N. Jain, Surflex: Fully Automatic Flexible Molecular Docking Using a Molecular Similarity-Based Search Engine. *Journal of Medicinal Chemistry* **46**, 499-511 (2003).

33. M. Wójcikowski, P. J. Ballester, P. Siedlecki, Performance of machine-learning scoring functions in structure-based virtual screening. *Scientific Reports* **7**, 46710 (2017).

34. M. M. Stepniewska-Dziubinska, P. Zielenkiewicz, P. Siedlecki, Development and evaluation of a deep learning model for protein–ligand binding affinity prediction. *Bioinformatics* **34**, 3666-3674 (2018).

35. A. T. McNutt *et al.*, GNINA 1.0: molecular docking with deep learning. *Journal of Cheminformatics* **13**, 43 (2021).

36. M. Brocidiacono *et al.*, BigBind: Learning from Nonstructural Data for Structure-Based Virtual Screening. *Journal of Chemical Information and Modeling* **64**, 2488-2495 (2024).

37. C. Shen *et al.*, Beware of the generic machine learning-based scoring functions in structure-based virtual screening. *Briefings in Bioinformatics* **22**, (2020).

38. J. Sunseri, D. R. Koes, Virtual Screening with Gnina 1.0. *Molecules* **26**, (2021).

39. C. L. Raison *et al.*, Single-dose psilocybin treatment for major depressive disorder: A randomized clinical trial. *JAMA* **330**, 843-853 (2023).

40. W. Duan, D. Cao, S. Wang, J. Cheng, Serotonin 2A Receptor (5-HT2AR) Agonists: Psychedelics and Non-Hallucinogenic Analogues as Emerging Antidepressants. *Chemical Reviews* **124**, 124-163 (2024).

41. D. Cao *et al.*, Structure-based discovery of nonhallucinogenic psychedelic analogs. *Science* **375**, 403-411 (2022).

42. J. Wallach *et al.*, Identification of 5-HT2A receptor signaling pathways associated with psychedelic potential. *Nature Communications* **14**, 8221 (2023).

43. K. Kim *et al.*, Structure of a Hallucinogen-Activated Gq-Coupled 5-HT(2A) Serotonin Receptor. *Cell* **182**, 1574-1588.e1519 (2020).

44. D. Wacker *et al.*, Crystal Structure of an LSD-Bound Human Serotonin Receptor. *Cell* **168**, 377-389.e312 (2017).

45. J. Zhou, Norepinephrine transporter inhibitors and their therapeutic potential. *Drugs Future* **29**, 1235-1244 (2004).

46. J. Tan *et al.*, Molecular basis of human noradrenaline transporter reuptake and inhibition. *Nature* **632**, 921-929 (2024).

47. H. Zhang *et al.*, Dimerization and antidepressant recognition at noradrenaline transporter. *Nature* **630**, 247-254 (2024).

48. T. Hu *et al.*, Transport and inhibition mechanisms of the human noradrenaline transporter. *Nature* **632**, 930-937 (2024).

49. S. Pidathala, A. K. Mallela, D. Joseph, A. Penmatsa, Structural basis of norepinephrine recognition and transport inhibition in neurotransmitter transporters. *Nature Communications* **12**, 2199 (2021).

50. Y. Zhang *et al.*, Benchmarking Refined and Unrefined AlphaFold2 Structures for Hit Discovery. *J Chem Inf Model* **63**, 1656-1667 (2023).

51. M. Karelina, J. J. Noh, R. O. Dror, How accurately can one predict drug binding modes using AlphaFold models? *eLife*, (2023).

52. J. Lyu *et al.*, AlphaFold2 structures guide prospective ligand discovery. *Science* **384**, eadn6354 (2024).

53. A. Díaz-Holguín *et al.*, AlphaFold accelerated discovery of psychotropic agonists targeting the trace amine–associated receptor 1. *Science Advances* **10**, eadn1524 (2024).

54. A. Gaulton *et al.*, ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res* **40**, D1100-1107 (2012).

55. A. Rives *et al.*, Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci U S A* **118**, (2021).

56. T. Kamenecka *et al.*, Structure-Activity Relationships and X-ray Structures Describing the Selectivity of Aminopyrazole Inhibitors for c-Jun N-terminal

580 Kinase 3 (JNK3) over p38*. *Journal of Biological Chemistry* **284**, 12853-
581 12861 (2009).

582 57. K. Zheng *et al.*, Design and Synthesis of Highly Potent and Isoform Selective
583 JNK3 Inhibitors: SAR Studies on Aminopyrazole Derivatives. *Journal of*
584 *Medicinal Chemistry* **57**, 10013-10030 (2014).

585 58. L. Le *et al.*, SLC45A2 protein stability and regulation of melanosome pH
586 determine melanocyte pigmentation. *Mol Biol Cell* **31**, 2687-2702 (2020).

587 59. J. Park *et al.*, SLC45A2: A Melanoma Antigen with High Tumor Selectivity
588 and Reduced Potential for Autoimmune Toxicity. *Cancer Immunol Res* **5**, 618-
589 629 (2017).

590 60. Z. H. Zuo *et al.*, Oncogenic Activity of Solute Carrier Family 45 Member 2
591 and Alpha-Methylacyl-Coenzyme A Racemase Gene Fusion Is Mediated by
592 Mitogen-Activated Protein Kinase. *Hepatol Commun* **6**, 209-222 (2022).

593 61. R. G. Naressi, D. Schechtman, B. Malnic, Odorant receptors as potential drug
594 targets. *Trends in Pharmacological Sciences* **44**, 11-14 (2023).

595 62. M. Orecchioni *et al.*, Olfactory receptor 2 in vascular macrophages drives
596 atherosclerosis by NLRP3-dependent IL-1 production. *Science* **375**, 214-221
597 (2022).

598 63. R. L. Wolfson *et al.*, Sestrin2 is a leucine sensor for the mTORC1 pathway.
599 *Science* **351**, 43-48 (2016).

600 64. J. Qu *et al.*, A paradoxical role for sestrin 2 protein in tumor suppression and
601 tumorigenesis. *Cancer Cell International* **21**, 606 (2021).

602 65. R. A. Saxton *et al.*, Structural basis for leucine sensing by the Sestrin2-
603 mTORC1 pathway. *Science* **351**, 53-58 (2016).

604 66. Z. Lin *et al.*, Evolutionary-scale prediction of atomic-level protein structure
605 with a language model. *Science* **379**, 1123-1130 (2023).

606 67. L. Thorel *et al.*, Patient-derived tumor organoids: a new avenue for preclinical
607 research and precision medicine in oncology. *Experimental & Molecular*
608 *Medicine* **56**, 1531-1551 (2024).

609 68. F. Weeber, S. N. Ooft, K. K. Dijkstra, E. E. Voest, Tumor Organoids as a Pre-
610 clinical Cancer Model for Drug Discovery. *Cell Chemical Biology* **24**, 1092-
611 1100 (2017).

612 69. J. J. Vandana, C. Manrique, L. A. Lacko, S. Chen, Human pluripotent-stem-
613 cell-derived organoids for drug discovery and evaluation. *Cell Stem Cell* **30**,
614 571-591 (2023).

70. J. Chuprin *et al.*, Humanized mouse models for immuno-oncology research. *Nature Reviews Clinical Oncology* **20**, 192-206 (2023).

71. G. Peltz, Can humanized mice improve drug development in the 21st century? *Trends in Pharmacological Sciences* **34**, 255-260 (2013).

72. N. Legrand *et al.*, Humanized Mice for Modeling Human Infectious Disease: Challenges, Progress, and Outlook. *Cell Host & Microbe* **6**, 5-9 (2009).

73. P. Lewis *et al.*, Retrieval-augmented generation for knowledge-intensive NLP tasks. *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Article 793 (2020).

74. Z. Wang *et al.*, Retrieval-based Controllable Molecule Generation. *ArXiv* **abs/2208.11126**, (2022).