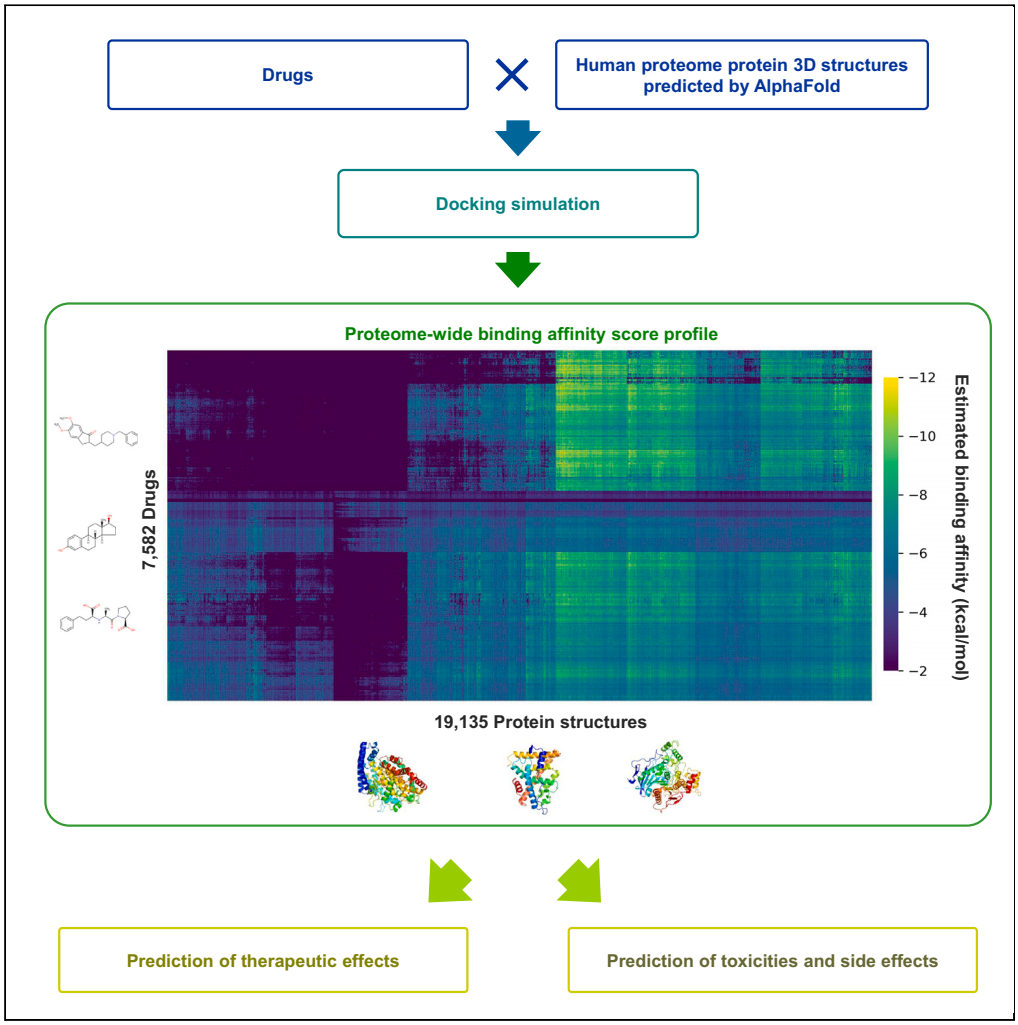
iScience

CellPress
OPEN ACCESS

Article

Predicting therapeutic and side effects from drug binding affinities to human proteome structures



Ryusuke Sawada, Yuko Sakajiri, Tomokazu Shibata, Yoshihiro Yamanishi

yamanishi@i.nagoya-u.ac.jp

Highlights

Docking simulations are performed for all pairs of drugs and protein structures

Each drug is represented by a proteome-wide binding affinity score profile

The method can predict drug indications from all related protein structures

The statistical model can extract proteins eliciting drug toxicities and side effects

Sawada et al., iScience 27, 110032
June 21, 2024
https://doi.org/10.1016/j.isci.2024.110032

iScience

CellPress
OPEN ACCESS

Article

Predicting therapeutic and side effects from drug binding affinities to human proteome structures

Ryusuke Sawada,[1,2,4] Yuko Sakajiri,[1,3,4] Tomokazu Shibata,[1] and Yoshihiro Yamanishi[1,3,5,*]

SUMMARY

Evaluation of the binding affinities of drugs to proteins is a crucial process for identifying drug pharmacological actions, but it requires three dimensional structures of proteins. Herein, we propose novel computational methods to predict the therapeutic indications and side effects of drug candidate compounds from the binding affinities to human protein structures on a proteome-wide scale. Large-scale docking simulations were performed for 7,582 drugs with 19,135 protein structures revealed by AlphaFold (including experimentally unresolved proteins), and machine learning models on the proteome-wide binding affinity score (PBAS) profiles were constructed. We demonstrated the usefulness of the method for predicting the therapeutic indications for 559 diseases and side effects for 285 toxicities. The method enabled to predict drug indications for which the related protein structures had not been experimentally determined and to successfully extract proteins eliciting the side effects. The proposed method will be useful in various applications in drug discovery.

INTRODUCTION

The identification of drug pharmacological actions such as therapeutic efficacy and side effects is a challenging issue.[1] Drugs exhibit therapeutic and side effects when they interact with target proteins in the human body and off-targets. The information on drug-protein interactions provides important clues for identifying these types of effects. Currently, computational approaches, such as machine learning and docking simulations, can give a deeper insight into drug-protein interactions.[2–5] Machine learning methods can predict drug-protein interactions with high accuracy when a training dataset of sufficient size is available, but they do not work well for target proteins with little prior information on ligands. Docking simulation can estimate the binding affinity of a drug to a target protein by calculating the binding free energy even when no prior information on interaction is available, but it requires three-dimensional (3D) structures of target proteins.

Because the experimental determination of protein 3D structures is time- and cost-consuming, 3D structural data are available for a limited number of proteins. Currently, there are approximately 200 million natural proteins with known amino acid sequences, while 3D structures have been determined experimentally only for 180,000.[6] For protein sequences encoded in the human genome, fully determined 3D structures are available for only 14% and partially determined 3D structures are available for only 21% (Figure S1A). 180,000 protein structures registered in the Protein DataBank (PDB) archive are considered to cover almost the entire fold space for monomeric proteins in the environment.[7–10] In fact, a sequence similarity search for all human proteins against PDB entries has revealed that 46% and >90% of the proteins are fully and partially homologous, respectively (Figure S1B), which suggests a possibility of predicting 3D structures for most human proteins.

Over the years, the computational prediction of protein 3D structures from amino acid sequences has been tackled using two approaches: namely homology (e.g., MODELLER[11] and SWISS-MODEL[12]) and *ab initio* (e.g., ROSETTA[13] and I-TASSER[14]) modeling, but it has been a difficult task.[15] Recently, artificial intelligence (AI) technologies including transformers have been used for predicting protein 3D structures and have been shown to outperform conventional homology modeling techniques and *ab initio* methods (RaptorX,[16] DMPFold,[17] trRosetta,[18] and AlphaFold version 1[19]). In particular, AlphaFold version 2[20] won the Critical Assessment of Structure Prediction 14[21] competition with unprecedented success and thus had a tremendous impact on the field of structural biology.[22–25] The information on protein 3D structures revealed by AlphaFold[26] is a useful resource for medical and pharmaceutical research.

In this study, we propose novel computational methods to predict the therapeutic indications and side effects of drug candidate compounds using the 3D structures of human whole proteins revealed by AlphaFold. We performed large-scale docking simulations of each drug on all human protein structures, which was previously impossible to achieve thus far. The calculated binding affinities to all the proteins (including structurally resolved and unresolved proteins) of each drug were summarized to create the proteome-wide binding affinity score

[1]Department of Bioscience and Bioinformatics, Faculty of Computer Science and Systems Engineering, Kyushu Institute of Technology, Iizuka, Japan
[2]Department of Pharmacology, Okayama University Graduate School of Medicine, Dentistry and Pharmaceutical Sciences, Okayama, Japan
[3]Graduate School of Informatics, Nagoya University, Chikusa, Nagoya, Japan
[4]These authors contributed equally
[5]Lead contact
*Correspondence: yamanishi@i.nagoya-u.ac.jp
https://doi.org/10.1016/j.isci.2024.110032

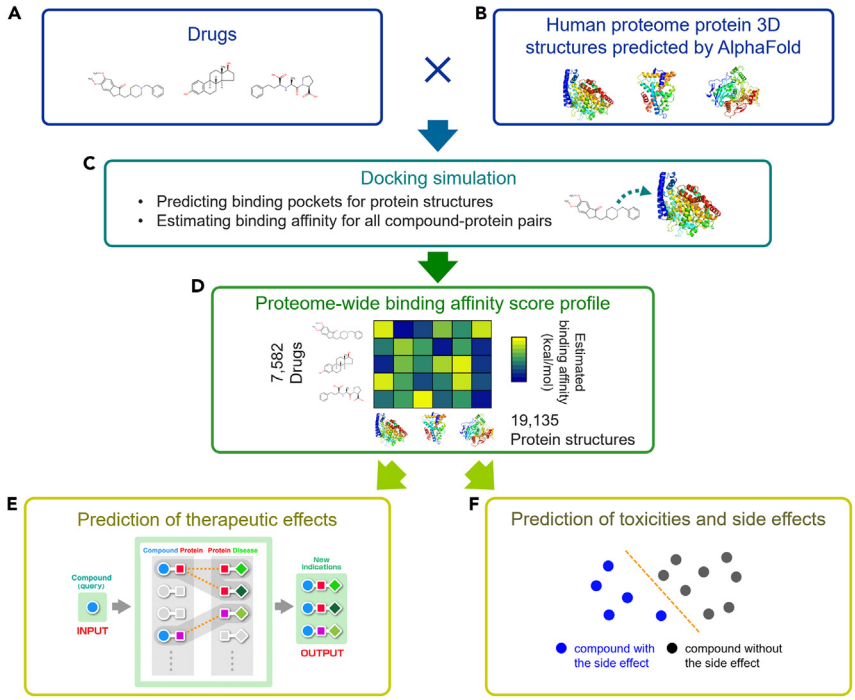


Figure 1. Workflow of the proposed drug discovery methods using AlphaFold

(A) Structures of the drugs were obtained from KEGG DRUG.

(B) 3D structural data for all human proteins were retrieved from AlphaFoldDB.

(C) Ligand-binding pockets were detected for all human protein structures, and the binding affinities for all drug-protein pairs were assessed.

(D) Estimated binding affinities were compiled to generate the proteome-wide binding affinity score (PBAS) profiles.

(E and F) Using PBAS profiles, the prediction of potential drug therapeutic indications (E) and side effects (F) was conducted.

(PBAS) profile. We show the usefulness of the method on the prediction of putative therapeutic indications and side effects of drugs from the PBAS profiles with machine learning models. The proposed methods are considered useful in various applications for drug discovery.

RESULTS

Workflow of the proposed methods for predicting therapeutic indications and side effects

The workflow of the proposed methods for predicting the therapeutic indications and side effects of drug candidate compounds using the 3D structures of human whole proteins is shown in Figure 1. First, the 3D structures of all human proteins were obtained from the AlphaFold protein structure database[26] (AlphaFoldDB) and the structures of all drugs were obtained from the Kyoto Encyclopedia of Genes and Genomes (KEGG) DRUG database.[27] Next, we identified the ligand-binding pockets in all the protein structures and performed docking simulations for all possible drug-protein pairs to determine the binding affinities of drugs to proteins. The binding free energy values calculated from the simulations were merged to create the PBAS profiles of 7,582 drugs for 19,135 human protein structures. In this study, we show the usefulness of the PBAS profiles on the prediction of therapeutic indications and side effects of drugs with machine learning models.

Docking simulations for possible pairs of drugs and all human protein structures

We performed extensive docking simulations for all drugs and human proteins using AutoDock Vina[28] for fast calculations. We evaluated the binding affinities of all drug-protein pairs using 7,582 drugs and 19,135 protein structures (18,347 different proteins). The results of the docking simulations were combined to create the PBAS profiles of all drugs, representing each of their binding affinities to all human proteins (center heatmap in Figure 2). Figure S2 shows the approximate binding affinities of the structurally unresolved and resolved proteins separately. Until now, the 3D structures of approximately 30% of all human proteins have been modeled experimentally (resolved and partially resolved on the left side of the heatmap), making it impossible to conduct docking simulations for the unresolved proteins. AlphaFold and AutoDock Vina provided the capability to predict the binding affinity of any drug for both resolved and unresolved proteins.

An example of docking poses with the PBAS profile is shown in Figure S3, where the 3D structure was predicted using AlphaFold for ABL1 (AlphaFoldDB ID: AF-P00519-F1-model_v1) and its inhibitor nilotinib. The resulting ligand-protein binding structure (PDB ID: 3cs9) is shown in gray, which indicates that the structure predicted using AlphaFold (blue) and the experimentally determined structure match well (root-mean-square deviation [RMSD] of atomic positions = 0.464 Å). Furthermore, the docking pose of nilotinib (yellow) predicted by the docking simulation was also consistent with experimentally determined structure (RMSD = 1.369 Å).

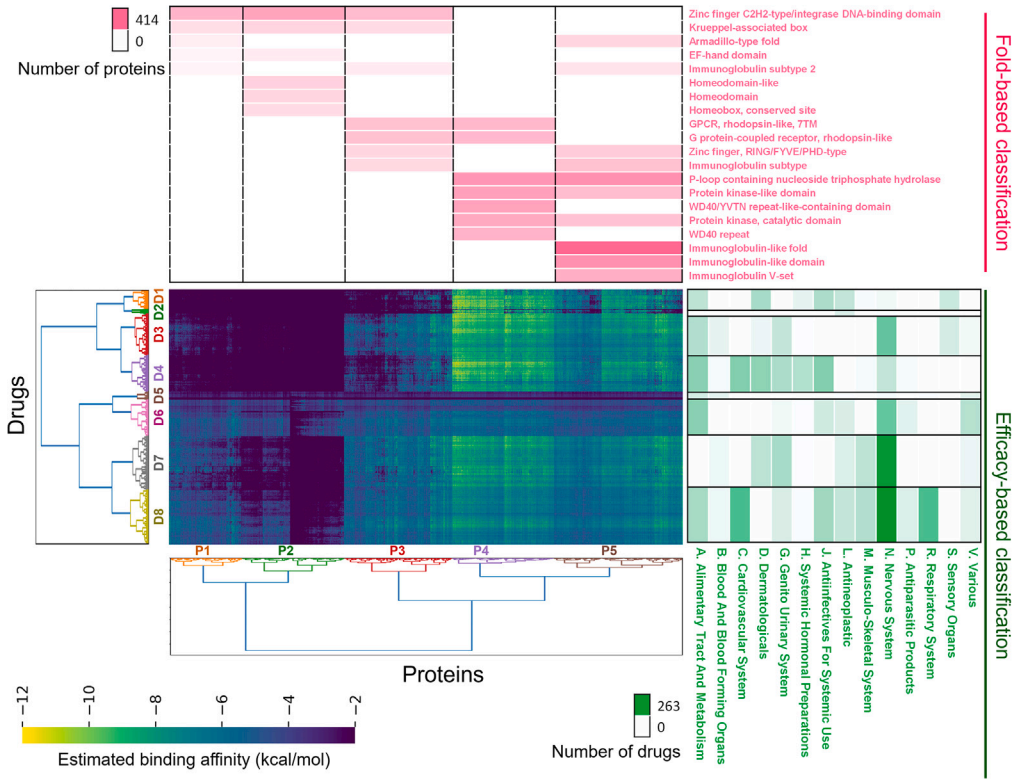


Figure 2. Proteome-wide binding affinity score (PBAS) profile and clustering analysis

The PBAS profiles created from the results of docking simulations for all drugs with all human proteins are shown in the center of the heatmap. The estimated binding affinities are expressed as a gradient, with the horizontal and vertical axes representing proteins (19,135 structures) and drugs (7,582 drugs), respectively. Proteins and drugs are clustered separately, and the labels are shown for each cluster. The dendrogram on the left side of and under the heatmap shows the clusters of drugs (D1–D8) and proteins (P1–P5), respectively. Enrichment analysis was performed for the protein and drug clusters. The table above the heatmap reflects the results of the enrichment analysis of the structural domain groups of InterPro for the protein clusters. The number of proteins in each structural domain group is shown as a gradient. The table on the right side of the heatmap contains the outcomes of the enrichment analysis for the Anatomical Therapeutic Chemical (ATC) classification system groups for the drug clusters. The number of drugs enriched in the ATC group is marked as a gradient.

Binding affinity-based protein clusters correlate with protein folds and structural domains

Hierarchical clustering was performed on proteins based on the similarities in binding affinity of the PBAS profiles (Figure 2). The horizontal and vertical dendrograms in the heatmap show the protein and drug clusters, respectively. The proteins and drugs were divided into five (P1–P5) and eight (D1–D8) clusters, respectively, according to the distance between clusters.

For proteins, we focused on their structural domains and conducted an enrichment analysis using database for annotation, visualization, and integrated discovery (DAVID).[29] The top table of the heatmap in Figure 2 shows the enrichment analysis results. Among the significantly enriched domain structural groups, the top five are indicated by the gradient of table cells with respect to the number of enriched proteins. The P1 group exhibited several motifs containing nuclear-localized proteins, such as the transcription factors "zinc finger C_2H_2-type/integrase DNA-binding domain" and "Krueppel-associated box." The P2 group, like P1, not only had many zinc finger-related transcriptional regulatory domains but was also composed of different types of transcriptional regulatory domains, including "homeodomain-like"; "homeodomain"; and "homeobox, conserved site." In the P3 group, proteins consisting of G protein-coupled receptor (GPCR)-related domains in addition to zinc finger-related domains were identified. The P4 group included many domains related to protein phosphorylation signaling, such as "P loop containing nucleoside triphosphate hydrolase"; "protein kinase-like domain"; "WD40/YVTN repeat-like-containing domain"; "protein kinase, catalytic domain"; and "WD40 repeat." The P5 group was composed of more proteins associated with immunoglobulin-related domains. These results suggest that protein clusters derived from the PBAS profiles correlate with protein folds and structural domains.

Binding affinity-based drug clusters correlate with polypharmacology and target specificity

Next, we evaluated the clustered groups of drugs. We determined which class of the Anatomical Therapeutic Chemical (ATC) classification system corresponded to the drugs of each cluster group (table on the right side of the heatmap in Figure 2). Some clusters had a correlation with the enriched ATC classification class. For example, antipsychotics (ATC group N) were abundant not only in cluster groups D7 and D8 but

also in the relatively distant D3 cluster group. These findings suggest that drugs form cluster groups based on their binding affinity, regardless of the type of therapeutic indication (ATC classification).

The ATC classification system is based on the targeted organ and the disease, for which the drug is indicated; however, even drugs for the same disease may have different binding proteins and mechanisms of action, which may be explained by the year in which the drug was developed and approved for a therapeutic use.[30] Therefore, we examined the distribution for the drug approval year of each drug cluster (Figure S4). The leftmost dendrogram displays the clusters of drug-binding affinity reflected in a gradient in the middle (Figure S4). Clusters D1–D4 and D5–D8 were divided into two main groups depending on distances in the dendrogram, and the former group showed a lower binding affinity than the latter group. Furthermore, the former group represented more recently approved drugs. It implies that in recent years, there has been a trend toward the development of drugs with lower average binding affinities to human proteins.

In recent years, drugs that bind more strongly to specific target molecules tend to be approved.[31,32] Therefore, we examined the correlation between the drug's year of approval in the PBAS profiles, its molecular weight, and logP values. The molecular weight and logP values of drugs in the PBAS profiles tended to increase from the earlier to later times (Figure S5A). For each drug, we compared the average value of the overall binding affinity with that of the top 10% with the highest value for this parameter (Figure S5B). The average binding affinity of the drug to all proteins had a decreasing trend in recent years, whereas that of the top 10% of proteins had an increasing trend in recent years. An increase in the molecular weight of each drug is associated with its structural complexity, which results in an elevated binding specificity to the protein pocket site without gaps. An increase in the logP value may also indicate that the drug is more hydrophobic, subsequently causing more selective binding to a few target proteins. Despite the decrease in binding strength to most proteins, the opposite trend for specific proteins may be explained by these increases in molecular weight and logP values and the associated binding affinity and strength to specific proteins. These results suggest that drug clusters derived from the PBAS profiles correlate with polypharmacology and target specificity.

Performance test with CASF-2016 core set

We applied our proposed method to the validation dataset of CASF-2016 core set[33] and computed the docking scores. Subsequently, we examined the correlation between these calculated docking scores and the experimentally determined binding constants, comparing them with those of other methodologies (Figure S6). Values for other methodologies were obtained from the "Table S4" of a study by Su et al.[33] Figure S7 illustrates a detailed plot of our proposed method's performance. The validation set encompasses the entire CASF-2016 core set, which includes binding structures between protein complex interaction interfaces and ligands. Notably, our proposed method does not consider protein complexes. To ensure a fair evaluation, we generated a subdataset comprising only monomers (Monomer). Furthermore, we created a human protein-only (PBAS protein-only) subdataset out of the monomer subdataset to evaluate the performance of our proposed PBAS. We evaluated the performance of these three validation sets.

Upon utilizing all validation data, we observed a correlation coefficient of 0.417. Furthermore, when utilizing the monomer set excluding complexes, a slight enhancement in performance was noted, with a correlation coefficient of 0.508. When considering only proteins within the PBAS of the monomer set, the correlation coefficient improved to 0.596. Our proposed method demonstrated improved performance compared to other methodologies, with results closely aligning with those of "AutoDock Vina," which employs the same docking simulation software.

Comprehensive prediction of drug therapeutic indications based on binding affinity to all human protein structures

We predicted the potential therapeutic indications of drugs using the PBAS profiles. The prediction was conducted using a template matching method with information on therapeutic target proteins for various diseases.[2] The previous template matching method was dependent on known target protein information, so applicability was limited to approximately 4,000 proteins with information on at least one known ligand. We proposed to use the PBAS profiles as an input in the template matching method, enabling us to take into account the interactions with a much larger number of proteins in the prediction process of drug therapeutic indications. The stronger a drug binds, the higher its predictive score for the applicable disease. The proposed method was applied to 7,582 drugs for 559 diseases.

Part of the drug-protein-disease network predicted using the proposed method is shown in Figure 3. For example, nilotinib, a tyrosine kinase inhibitor, was predicted to bind to the tyrosine kinase ABL1, one of its original therapeutic target proteins in the treatment of myeloid leukemia. Thus, it was confirmed that the proposed method was able to reproduce known drug indications based on the calculated binding affinity.

Table S1 shows the predicted therapeutic indications of the parturition-inducing drug cloprostenol. The original therapeutic indication of cloprostenol, "accelerated parturition," was predicted to have the No.1 ranking. The results of a docking simulation with the prostaglandin F receptor, the drug's original target, are shown in Figure 4A. The binding affinity used for ranking was −10.1 kcal/mol, which was the top docking score (Table S2). Other newly predicted indications were ophthalmic hypertension and glaucoma. Interestingly, the therapeutic indications of cloprostenol in lowering intraocular pressure have been reported in animals, and its analogs have been used to treat increased intraocular pressure in humans.[34–36]

As another example, Table S3 shows the results of the predicted therapeutic indications for the antidiabetic drug ertugliflozin. The original therapeutic indication of ertugliflozin, i.e., "diabetes mellitus 2," was predicted with a ranking of 11. Figure 4B shows the docking pose of ertugliflozin and SLC5A2. The estimated binding affinity was −10.426 kcal/mol, indicating strong binding and an overall high ranking (Table S4). Other predicted indications included cardiovascular disease and atherosclerotic cardiovascular disease. Ertugliflozin was reported

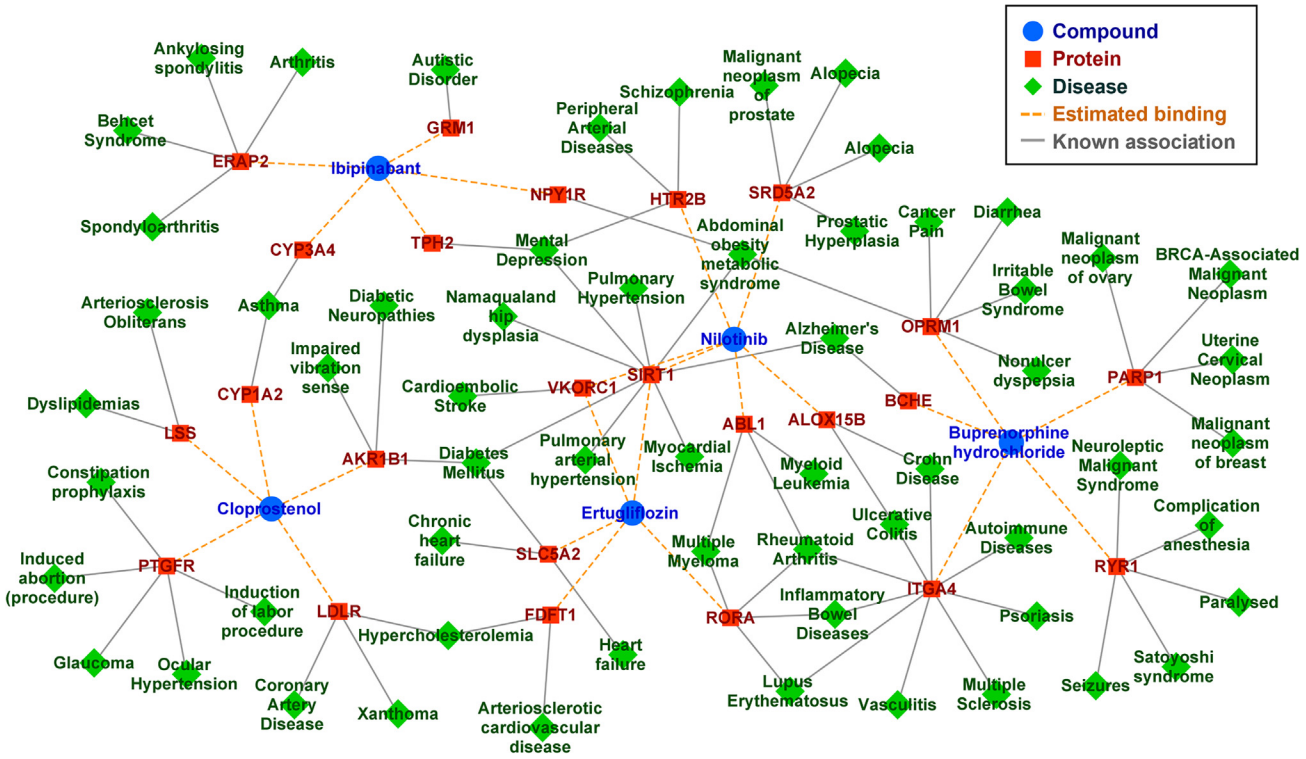


Figure 3. Small part of the drug-protein-disease network predicted by the proposed methods using proteome-wide binding affinity score profiles

Blue circles, green rhombuses, and red rectangles stand for drugs, diseases, and proteins, respectively. Orange broken and gray lines indicate the relationships between the drugs and proteins obtained from the docking simulation results and those between proteins and possible disease indications, respectively.

to be effective in treating cardiovascular disease according to a completed phase 3 study.[37] Thus, this method enabled us to predict a potential therapeutic effect of ertugliflozin.

Finally, Table S5 shows the results of predicted therapeutic indications for the analgesic drug buprenorphine. The original therapeutic indication for buprenorphine, "cancer pain," was predicted with a ranking of 15. The binding affinity to the mu-opioid receptor, which is the main target of this drug, indicates that the original therapeutic indications were correctly predicted. Figure 4C shows the docking

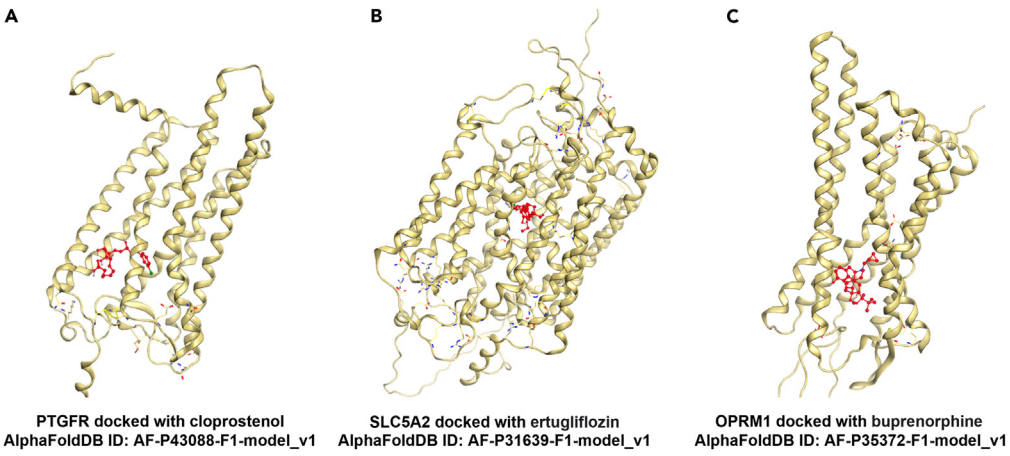


Figure 4. Protein-compound binding structures estimated by docking simulation

Compounds and proteins are shown as a red rod and an ivory ribbon model, respectively.

(A) Overall structure of PTGFR docked with cloprostenol.

(B) Overall structure of SLC5A2 conjugated with ertugliflozin.

(C) Overall structure of OPRM1 merged with buprenorphine. All protein structures were estimated using AlphaFold.

pose of buprenorphine and OPRM1. The binding affinity was high (Table S6). Although PDB only registered the 3D structures of similar sequences for the mu-opioid receptor, this confirms that the therapeutic indication prediction can be made even if the experimentally determined 3D structure of the target protein is unavailable. Additional therapeutic indications of buprenorphine were predicted, including inflammatory diseases in the intestinal tract. Buprenorphine was previously shown to reduce intestinal inflammation in a mouse model.[38]

Experimental validation of binding affinities of PBAS for drug therapeutic indications

In order to substantiate the *in silico* findings, we conducted experimental validation through competitive binding assays for a drug across 30 human proteins (Table S7) associated with pharmacological actions, including therapeutic effects and side effects. Metergoline, known for its vasodilatory properties to alleviate headache symptoms, acts as both a serotonin antagonist and dopamine agonist.[39] A binding assay was conducted to confirm metergoline's interaction with 30 proteins (Table S8). The proposed method predicted strong binding of metergoline to serotonin receptor and dopamine receptor groups, which are known interacting proteins, and this was corroborated by verification experiments. For instance, the inhibition rate for 5-HT1B receptor (HTR1B) was 100.4 (Inh %). The docking score stood at −10.3 (kcal/mol), validating our docking and assay findings. Despite a weak correlation coefficient of −0.191 between overall experimental values and the proposed method's binding affinity, the highest predicted binding affinity was observed for Acetylcholinesterase (ACHE), which was confirmed in experiments. Consequently, metergoline's efficacy against central nervous system degenerative diseases like Alzheimer's disease is anticipated. Experimental protocols for proteins were provided in "in vitro binding and enzymatic assays" section in the supplementary information.

Comprehensive prediction of drug side effects based on binding affinities to all human protein structures

Next, we predicted drug side effects using the PBAS profiles. The binding affinities of each drug to all human proteins were used as features in the L1-regularized logistic regression model, a sparsity-induced classifier. There have been previous studies on the side effect prediction based on the interactions of drugs with multiple proteins using machine learning models,[4,5] but previous methods used only known drug-protein interactions involving proteins with at least one known ligand, which was a serious limitation. An advantage of our proposed method was the use of information on the interactions with all human proteins.

We compared the side effect prediction accuracy of the proposed method with that of existing techniques. The proposed PBAS profile was compared with the following three types: (1) chemical substructure profile based on molecular structure descriptors, which is referred to as a Fingerprint profile[40]; (2) target protein profile based on drug-protein interactions estimated using a structural similarity search, which is referred to as the target estimation with similarity search (TESS) profile[2]; and (3) target protein profile based on drug-protein interactions estimated by supervised learning, which is referred to as the target estimation with logistic regression (TELR) profile.[5] Furthermore, we compared the performance when combining the PBAS profile and the three profiles.

We evaluated the performance of each profile-based prediction with 5-fold cross-validation experiments using known drug side effects obtained from the SIDER database[41] as gold standard data. The 5-fold cross-validation experiment was repeated 30 times, and the area under the ROC curve (AUC) and area under the Precision-Recall (AUPR) curve scores were calculated. Figure 5 shows the distribution of the AUC and AUPR scores in repeated experiments. Table S9 shows the mean value and standard deviation of the distribution of the AUC and AUPR scores. The proposed PBAS profile outperformed the other three profiles at the significant level in terms of both AUC and AUPR. One explanation about this observation would be that previous methods focused on a limited number of target proteins, whereas our approach allows for side effect predictions by considering interactions with all human proteins. In the case of combining PBAS and the other three profiles, it was found that AUC and AUPR remained almost unchanged. Only the combination of TESS profile and PBAS significantly improved the performance.

Next, we made a large-scale side effect prediction of 7,582 drugs in KEGG DRUG and 285 side effects in SIDER. The side effect prediction model was trained with all the drugs in SIDER, and the model was then applied to predicting the presence or absence of 285 side effects of 7,582 drugs. To evaluate the validity of the predicted side effects by our models, we elaborated the newly predicted side effects using independent resources. For example, we used the US Food and Drug Administration (FDA) Adverse Event Reporting System (FAERS), a dataset that is independent of the side effect resources in Side Effect Resource (SIDER). SIDER, which was used for training, is a side effect dataset collected primarily from drug inserts, whereas FAERS is a dataset collected from adverse drug events that occur after drugs are on the market. The number of overlapping drugs between 429 from SIDER and 1,673 from FAERS was 282. We examined whether the newly predicted side effects using the SIDER-based model were actually reported in FAERS (Figure 5 and Table S9). The FAERS-based validation also showed that the proposed PBAS profile exhibited higher AUC and AUPR values than the Fingerprint, TESS, and TELR profiles. These results suggest that it is useful to consider the binding affinities of drugs to all human proteins in predicting drug side effects.

Drug-protein-side effect network reveals proteins eliciting side effects

The side effect prediction model has a feature extraction ability owing to the sparse modeling, so highly weighted proteins in the predictive model can be considered important for the side effect and they are thought to be the proteins involved in the expression of side effects. Actual weight values in the side effect predictive models are shown as a heatmap in Figure S8. It was observed that most weight values were zeros and a small number of proteins were selected as important proteins for each side effect. Figure 6 shows a small part of drugs, proteins, and side effects in the network diagram, where drug-protein links correspond to the drug-binding affinities to all human protein structures, whereas protein side effect links correspond to the estimated weights in the predictive model. For example, the drug disulfiram is predicted to cause nausea because it binds to opioid receptors. It is known that drug interactions with opioid receptors cause nausea as a side effect.[42] This implies that the predictive model successfully reproduced known drug-protein-side effect associations.

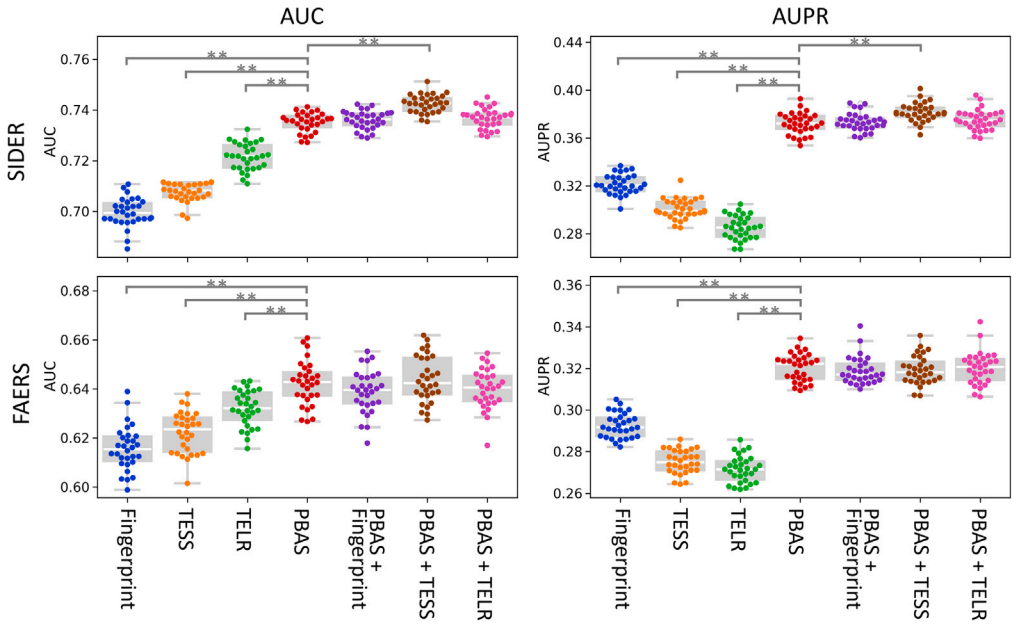


Figure 5. Performance evaluation on the side effect prediction

Asterisks indicate level of statistical significance: $^{**}p < 0.01$. Upper two panels show the AUC (left) and AUPR (right) scores for the SIDER dataset, and lower two panels show the scores for the FARES dataset. The results of replicates ($n = 30$) of cross-validation experiments for each of the seven profiles are shown. PBAS + Fingerprint, PBAS + TESS, and PBAS + TELR represent the integrated profiles that combine PBAS with the other three profiles.

Table S10 illustrates the examples of proteins that were regarded as important factors in causing side effects, and their validity was confirmed through independent resources. The proteins for which our model extracted an association with side effects from the PBAS profile are depicted further. For example, the insulin-inducible gene INSIG2 was identified as a protein responsible for weight gain. Consistently, it has been previously reported that INSIG2 is a molecule associated with weight gain during drug use.[43] A comprehensive evaluation of off-target interactions is important for predicting the potential side effects of drugs in the human body.[44,45] It is possible to elucidate the cause of side effects from the viewpoint of binding proteins including off-targets, even if the association between side effects and proteins has not been known. These results suggest that the proposed method is also useful for the biological interpretation of the mechanism of side effects in terms of binding proteins.

DISCUSSION

Summary

In this study, we characterized the mode-of-action of drugs as per their binding affinities to the entire human protein structures revealed by AlphaFold. Docking simulations for all possible pairs of drugs and human proteins were conducted, and the results were summarized as PBAS profiles. We demonstrated the usefulness of the PBAS profile for predicting the therapeutic indications and side effects of drugs. We identified proteins that bind strongly to the drug from PBAS profiles and constructed a three-way relationship between drug, protein, and disease. For predicting drug side effects, the PBAS profiles were used as the feature vectors of the drugs in machine learning models. Even for proteins whose 3D structures had not been experimentally determined, we successfully made predictions of drug therapeutic indications and side effects based on the binding affinities of the drugs with the protein 3D structures.

We used the proteome-wide structures revealed by AlphaFold for all human proteins in order to construct the PBAS profiles representing the binding affinities of drugs to all human proteins. Hierarchical clustering analysis of the PBAS profiles uncovers the link between the protein clusters and structural domain families. Moreover, the hierarchical clustering analysis revealed that drug clusters correlated with target specificity and drug approval year; the more recent the drug tends to have the higher binding affinity to specific target proteins and the lower binding affinity to the other off-target proteins. The PBAS profiles may successfully reflect the fact that recent drugs are designed to bind to specific target proteins and have fewer side effects.[31,32]

Interpretation and importance

We evaluated the effectiveness of our proposed PBAS profile using the CASF-2016 core set. We examined the correlation between these calculated docking scores and the experimentally determined binding constants, comparing them with values obtained from other methodologies. The correlation coefficients for all pairs demonstrated moderate performance, comparable to other tests utilizing AutoDock Vina

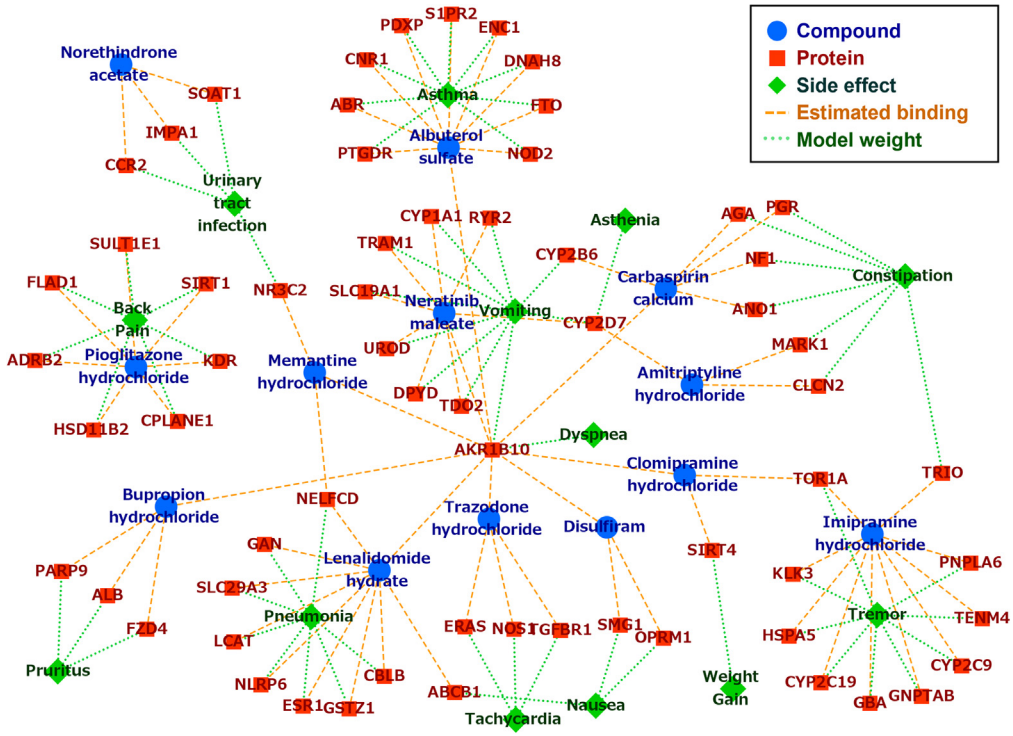


Figure 6. Small part of the drug-protein-side effect network predicted using our method

Blue circles, red rectangles, and green rhombuses indicate drugs, proteins, and side effects, respectively. Orange-dashed and green-dotted lines show the drug-protein relationships estimated from the docking simulation and the relationships between side effects and proteins with high weights in the predictive model, respectively.

(Figures S6 and S7). There is potential to enhance performance by incorporating alternative docking simulation techniques. Notably, performance showed improvement when restricting the validation dataset to proteins utilized in PBAS. Additionally, in docking simulations employing structures predicted by AlphaFold, the utilization of human proteins tended to yield better results compared to non-human proteins.

The PBAS profiles revealed that the interaction pattern with all human proteins varied markedly among the drugs. Some drugs were bound to only a small fraction of proteins, whereas other drugs were bound to many proteins. In fact, it has been shown that pharmaceutical compounds may interact with more proteins than previously thought *in vivo*.[44–48] Differences in the interaction patterns are difficult to explain based only on drug action or efficacy. Although drug development research tends to focus on specific therapeutic proteins that are directly associated with disease treatment, this study suggests that off-target activities on other proteins are also essential factors in evaluating both efficacy and side effects.[44]

Limitations of the study

Our proposed method has several limitations. In the docking simulation of this study, only one ligand-binding pocket was considered per protein; however, some proteins have multiple binding sites. For example, some drugs exert inhibitory or active effects by binding to the allosteric site of target proteins (e.g., ivermectin and glutamate-gated chloride channel,[49] omecamtiv mecarbil, and myosin[50]). In addition, our study used only a fixed parameter for all docking simulations. For a more precise evaluation, we plan to perform docking simulations for multiple-binding pockets with refined parameters in a future study. In this study, the protein structures used in the docking simulation were those of a monomer. However, some proteins function as heteromeric or homomeric complexes, whereas some drugs actually target the protein-protein binding interface of these complexes (e.g., benzodiazepine and gamma-aminobutyric acid type A receptor[51]). In the future, we plan to use AlphaFold-Multimer[52] and other methods to predict the structure of protein complexes and use them in docking simulations for the PBAS profiles.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.110032.

ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI (20H05797 and 21H04915). The computation was carried out using the computer resource offered under the category of General Projects by Research Institute for Information Technology, Kyushu University.

AUTHOR CONTRIBUTIONS

R.S. and Y.Y. designed and supervised the study. R.S. and Y.S. carried out data the computational simulation, analyzed the data, and prepared the manuscript's first draft and figures. Y.S. and T.S. carried out the data preparation and provided and the methodological advice. T.S. and Y.Y. provided constructive suggestions for results and discussion in manuscript completion. All authors read and approved the final version of the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: December 9, 2023
Revised: April 8, 2024
Accepted: May 16, 2024
Published: May 20, 2024

REFERENCES

1. DiMasi, J.A., and Grabowski, H.G. (2007). The cost of biopharmaceutical R&D: is biotech different? MDE. Manage. Decis. Econ. *28*, 469–479. https://doi.org/10.1002/mde.1360.
2. Sawada, R., Iwata, H., Mizutani, S., and Yamanishi, Y. (2015). Target-Based Drug Repositioning Using Large-Scale Chemical-Protein Interactome Data. J. Chem. Inf. Model. *55*, 2717–2730. https://doi.org/10.1021/acs.jcim.5b00330.
3. Sawada, R., Iwata, M., Tabei, Y., Yamato, H., and Yamanishi, Y. (2018). Predicting inhibitory and activatory drug targets by chemically and genetically perturbed transcriptome signatures. Sci. Rep. *8*, 156. https://doi.org/10.1038/s41598-017-18315-9.
4. Mizutani, S., Pauwels, E., Stoven, V., Goto, S., and Yamanishi, Y. (2012). Relating drug-protein interaction network with drug side effects. Bioinformatics *28*, i522–i528. https://doi.org/10.1093/bioinformatics/bts383.
5. Amano, Y., Honda, H., Sawada, R., Nukada, Y., Yamane, M., Ikeda, N., Morita, O., and Yamanishi, Y. (2020). In silico systems for predicting chemical-induced side effects using known and potential chemical protein interactions, enabling mechanism estimation. J. Toxicol. Sci. *45*, 137–149. https://doi.org/10.2131/jts.45.137.
6. Service, R.F. (2020). The game has changed.' AI triumphs at protein folding. Science *370*, 1144–1145. https://doi.org/10.1126/science.370.6521.1144.
7. Tonddast-Navaei, S., and Skolnick, J. (2015). Are protein-protein interfaces special regions on a protein's surface? J. Chem. Phys. *143*, 243149. https://doi.org/10.1063/1.4937428.

8. Kihara, D., and Skolnick, J. (2003). The PDB is a covering set of small protein structures. J. Mol. Biol. *334*, 793–802. https://doi.org/10.1016/j.jmb.2003.10.027.
9. Zhang, Y., Hubner, I.A., Arakaki, A.K., Shakhnovich, E., and Skolnick, J. (2006). On the origin and highly likely completeness of single-domain protein structures. Proc. Natl. Acad. Sci. USA *103*, 2605–2610. https://doi.org/10.1073/pnas.0509379103.
10. Skolnick, J., Arakaki, A.K., Lee, S.Y., and Brylinski, M. (2009). The continuity of protein structure space is an intrinsic property of proteins. Proc. Natl. Acad. Sci. USA *106*, 15690–15695. https://doi.org/10.1073/pnas.0907683106.
11. Sali, A., and Blundell, T.L. (1993). Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. *234*, 779–815. https://doi.org/10.1006/jmbi.1993.1626.
12. Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F.T., de Beer, T.A.P., Rempfer, C., Bordoli, L., et al. (2018). SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res. *46*, W296–W303. https://doi.org/10.1093/nar/gky427.
13. Simons, K.T., Bonneau, R., Ruczinski, I., and Baker, D. (1999). Ab initio protein structure prediction of CASP III targets using ROSETTA. Proteins Suppl. *3*, 171–176. https://doi.org/10.1002/(sici)1097-0134(1999)37:3+<171::aid-prot21>3.3.co;2-q.
14. Roy, A., Kucukural, A., and Zhang, Y. (2010). I-TASSER: a unified platform for automated protein structure and function prediction. Nat. Protoc. *5*, 725–738. https://doi.org/10.1038/nprot.2010.5.
15. Kuhlman, B., and Bradley, P. (2019). Advances in protein structure prediction and design. Nat. Rev. Mol. Cell Biol. *20*, 681–697. https://doi.org/10.1038/s41580-019-0163-x.
16. Peng, J., and Xu, J. (2011). RaptorX: exploiting structure information for protein alignment by statistical inference. Proteins *79*, 161–171. https://doi.org/10.1002/prot.23175.
17. Greener, J.G., Kandathil, S.M., and Jones, D.T. (2019). Deep learning extends de novo protein modelling coverage of genomes using iteratively predicted structural constraints. Nat. Commun. *10*, 3977. https://doi.org/10.1038/s41467-019-11994-0.
18. Anishchenko, I., Baek, M., Park, H., Hiranuma, N., Kim, D.E., Dauparas, J., Mansoor, S., Humphreys, I.R., and Baker, D. (2021). Protein tertiary structure prediction and refinement using deep learning and Rosetta in CASP14. Proteins *89*, 1722–1733. https://doi.org/10.1002/prot.26194.
19. Senior, A.W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A.W.R., Bridgland, A., et al. (2020). Improved protein structure prediction using potentials from deep learning. Nature *577*, 706–710. https://doi.org/10.1038/s41586-019-1923-7.
20. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589. https://doi.org/10.1038/s41586-021-03819-2.
21. Pereira, J., Simpkin, A.J., Hartmann, M.D., Rigden, D.J., Keegan, R.M., and Lupas, A.N. (2021). High-accuracy protein structure prediction in CASP14. Proteins *89*, 1687–1699. https://doi.org/10.1002/prot.26171.
22. Callaway, E. (2022). What's next for AlphaFold and the AI protein-folding revolution. Nature *604*, 234–238. https://doi.org/10.1038/d41586-022-00997-5.
23. Jones, D.T., and Thornton, J.M. (2022). The impact of AlphaFold2 one year on. Nat. Methods *19*, 15–20. https://doi.org/10.1038/s41592-021-01365-3.
24. Porta-Pardo, E., Ruiz-Serra, V., Valentini, S., and Valencia, A. (2022). The structural coverage of the human proteome before and after AlphaFold. PLoS Comput. Biol. *18*, e1009818. https://doi.org/10.1371/journal.pcbi.1009818.
25. Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G.R., Wang, J., Cong, Q., Kinch, L.N., Schaeffer, R.D., et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. Science *373*, 871–876. https://doi.org/10.1126/science.abj8754.
26. Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., et al. (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. *50*, D439–D444. https://doi.org/10.1093/nar/gkab1061.
27. Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., and Hirakawa, M. (2010). KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res. *38*, D355–D360. https://doi.org/10.1093/nar/gkp896.
28. Trott, O., and Olson, A.J. (2010). AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J. Comput. Chem. *31*, 455–461. https://doi.org/10.1002/jcc.21334.
29. Dennis, G., Jr., Sherman, B.T., Hosack, D.A., Yang, J., Gao, W., Lane, H.C., and Lempicki, R.A. (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol. *4*, P3.
30. Dahlen, A.D., Dashi, G., Maslov, I., Attwood, M.M., Jonsson, J., Trukhan, V., and Schioth, H.B. (2021). Trends in Antidiabetic Drug Discovery: FDA Approved Drugs, New Drugs in Clinical Trials and Global Sales. Front. Pharmacol. *12*, 807548. https://doi.org/10.3389/fphar.2021.807548.
31. Fagerholm, U. (2022). Investigation of Molecular Weights and Pharmacokinetic Characteristics of Older and Modern Small Drugs. Preprint at bioRxiv. https://doi.org/10.1101/2022.09.21.508888.
32. Radhakrishnan, M.L., and Tidor, B. (2007). Specificity in Molecular Design: A Physical Framework for Probing the Determinants of Binding Specificity and Promiscuity in a Biological Environment. J. Phys. Chem. B *111*, 13419–13435. https://doi.org/10.1021/jp074285e.
33. Su, M., Yang, Q., Du, Y., Feng, G., Liu, Z., Li, Y., and Wang, R. (2019). Comparative Assessment of Scoring Functions: The CASF-2016 Update. J. Chem. Inf. Model. *59*, 895–913. https://doi.org/10.1021/acs.jcim.8b00545.
34. Apostol, S., Gafencu, O., Carstocea, B., Selaru, D., Filip, M., and Cocu, F.G. (1995). [The use of prostaglandin compounds in treating glaucoma]. Oftalmologia *39*, 214–220.
35. Hirata, T., and Narumiya, S. (2011). Prostanoid receptors. Chem. Rev. *111*, 6209–6230. https://doi.org/10.1021/cr200010h.
36. Sava, A., Motoc, A.G.M., and Stan, C.I. (2015). Electron microscopic aspects of the effects of certain prostaglandin analogs on mouse testes. Rom. J. Morphol. Embryol. *56*, 771–775.
37. von Lewinski, D., Tripolt, N.J., Sourij, H., Pferschy, P.N., Oulhaj, A., Alber, H., Gwechenberger, M., Martinek, M., Seidl, S., Moertl, D., et al. (2022). Ertugliflozin to reduce arrhythmic burden in ICD/CRT patients (ERASe-trial) - A phase III study. Am. Heart J. *246*, 152–160. https://doi.org/10.1016/j.ahj.2022.01.008.
38. Blennerhassett, M.G., Lourenssen, S.R., Parlow, L.R.G., Ghasemlou, N., and Winterborn, A.N. (2017). Analgesia and mouse strain influence neuromuscular plasticity in inflamed intestine. Neuro Gastroenterol. Motil. *29*, 1–12. https://doi.org/10.1111/nmo.13097.
39. Hušák, M., Jegorov, A., Brus, J., van Beek, W., Pattison, P., Christensen, M., Favre-Nicolin, V., and Maixner, J. (2008). Metergoline II: structure solution from powder diffraction data with preferred orientation and from microcrystal. Struct. Chem. *19*, 517–525. https://doi.org/10.1007/s11224-008-9312-0.
40. Morgan, H.L. (1965). The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. J. Chem. Doc. *5*, 107–113. https://doi.org/10.1021/c160017a018.
41. Kuhn, M., Letunic, I., Jensen, L.J., and Bork, P. (2016). The SIDER database of drugs and side effects. Nucleic Acids Res. *44*, D1075–D1079. https://doi.org/10.1093/nar/gkv1075.
42. Smith, H.S., Smith, J.M., and Seidner, P. (2012). Opioid-induced nausea and vomiting. Ann. Palliat. Med. *1*, 121–129. https://doi.org/10.3978/j.issn.2224-5820.2012.07.08.
43. Cai, H.L., Tan, Q.Y., Jiang, P., Dang, R.L., Xue, Y., Tang, M.M., Xu, P., Deng, Y., Li, H.D., and Yao, J.K. (2015). A potential mechanism underlying atypical antipsychotics-induced lipid disturbances. Transl. Psychiatry *5*, e661. https://doi.org/10.1038/tp.2015.161.
44. Nobeli, I., Favia, A.D., and Thornton, J.M. (2009). Protein promiscuity and its implications for biotechnology. Nat. Biotechnol. *27*, 157–167. https://doi.org/10.1038/nbt1519.
45. Keiser, M.J., Setola, V., Irwin, J.J., Laggner, C., Abbas, A.I., Hufeisen, S.J., Jensen, N.H., Kuijer, M.B., Matos, R.C., Tran, T.B., et al. (2009). Predicting new molecular targets for known drugs. Nature *462*, 175–181. https://doi.org/10.1038/nature08506.
46. Ericson, E., Gebbia, M., Heisler, L.E., Wildenhain, J., Tyers, M., Giaever, G., and Nislow, C. (2008). Off-target effects of psychoactive drugs revealed by genome-wide assays in yeast. PLoS Genet. *4*, e1000151. https://doi.org/10.1371/journal.pgen.1000151.

47. Caldara, M., Graziano, S., Gullì, M., Cadonici, S., and Marmiroli, N. (2017). Editor's Highlight: Off-Target Effects of Neuroleptics and Antidepressants on Saccharomyces cerevisiae. Toxicol. Sci. *156*, 538–548. https://doi.org/10.1093/toxsci/kfx007.
48. Ietswaart, R., Arat, S., Chen, A.X., Farahmand, S., Kim, B., DuMouchel, W., Armstrong, D., Fekete, A., Sutherland, J.J., and Urban, L. (2020). Machine learning guided association of adverse drug reactions with in vitro target-based pharmacology. EBioMedicine *57*, 102837. https://doi.org/10.1016/j.ebiom.2020.102837.
49. Hibbs, R.E., and Gouaux, E. (2011). Principles of activation and permeation in an anion-selective Cys-loop receptor. Nature *474*, 54–60. https://doi.org/10.1038/nature10139.
50. Planelles-Herrero, V.J., Hartman, J.J., Robert-Paganin, J., Malik, F.I., and Houdusse, A. (2017). Mechanistic and structural basis for activation of cardiac myosin force production by omecamtiv mecarbil. Nat. Commun. *8*, 190. https://doi.org/10.1038/s41467-017-00176-5.
51. Kim, J.J., and Hibbs, R.E. (2021). Direct Structural Insights into GABA(A) Receptor Pharmacology. Trends Biochem. Sci. *46*, 502–517. https://doi.org/10.1016/j.tibs.2021.01.011.
52. Humphreys, I.R., Pei, J., Baek, M., Krishnakumar, A., Anishchenko, I., Ovchinnikov, S., Zhang, J., Ness, T.J., Banjade, S., Bagde, S.R., et al. (2021). Computed structures of core eukaryotic protein complexes. Science *374*, eabm4805. https://doi.org/10.1126/science.abm4805.
53. Bowes, J., Brown, A.J., Hamon, J., Jarolimek, W., Sridhar, A., Waldron, G., and Whitebread, S. (2012). Reducing safety-related drug attrition: the use of in vitro pharmacological profiling. Nat. Rev. Drug Discov. *11*, 909–922. https://doi.org/10.1038/nrd3845.
54. Varani, K., Gessi, S., Dalpiaz, A., and Borea, P.A. (1996). Pharmacological and biochemical characterization of purified A2a adenosine receptors in human platelet membranes by [3H]-CGS 21680 binding. Br. J. Pharmacol. *117*, 1693–1701.
55. Ford, A.P., Daniels, D.V., Chang, D.J., Gever, J.R., Jasper, J.R., Lesnick, J.D., and Clarke, D.E. (1997). Pharmacological pleiotropism of the human recombinant alpha1A-adrenoceptor: implications for alpha1-adrenoceptor classification. Br. J. Pharmacol. *121*, 1127–1135.
56. Sato, S., Hatanaka, T., Yuyama, H., Ukai, M., Noguchi, Y., Ohtake, A., Taguchi, K., Sasamata, M., and Miyata, K. (2012). Tamsulosin potently and selectively antagonizes human recombinant α1A/1D-adrenoceptors: Slow dissociation from the α1A-adrenoceptor may account for selectivity for α1A-adrenoceptor over α1B-adrenoceptor subtype. Biol. Pharm. Bull. *35*, 72–77.
57. Gleason, M.M., and Hieble, J.P. (1991). Ability of SK&F 104078 and SK&F 104856 to identify alpha-2 adrenoceptor subtypes in NCB20 cells and guinea pig lung. J. Pharmacol. Exp. Ther. *259*, 1124–1132.
58. Lalchandani, S.G., Lei, L., Zheng, W., Suni, M.M., Moore, B.M., Liggett, S.B., Miller, D.D., and Feller, D.R. (2002). Yohimbine dimers exhibiting selectivity for the human α2c-adrenoceptor subtype. J. Pharmacol. Exp. Ther. *303*, 979–984.
59. Feve, B., Elhadri, K., Quignard-Boulange, A., and Pairault, J. (1994). Transcriptional down-regulation by insulin of the beta 3-adrenergic receptor expression in 3T3-F442A adipocytes: a mechanism for repressing the cAMP signaling pathway. Proc. Natl. Acad. Sci. USA *91*, 5677–5681.
60. McCrea, K.E., and Hill, S.J. (1993). Salmeterol, a long-acting β2-adrenoceptor agonist mediating cyclic AMP accumulation in a neuronal cell line. Br. J. Pharmacol. *110*, 619–626.
61. Jung, M., Calassi, R., Rinaldi-Carmona, M., Chardenot, P., Le Fur, G., Soubrié, P., and Oury-Donat, F. (1997). Characterization of CB1 receptors on rat neuronal cell cultures: binding and functional studies using the selective receptor antagonist SR 141716A. J. Neurochem. *68*, 402–409.
62. Melck, D., De Petrocellis, L., Orlando, P., Bisogno, T., Laezza, C., Bifulco, M., and Di Marzo, V. (2000). Suppression of nerve growth factor Trk receptors and prolactin receptors by endocannabinoids leads to inhibition of human breast and prostate cancer cell proliferation. Endocrinology *141*, 118–126.
63. Dearry, A., Gingrich, J.A., Falardeau, P., Fremeau, R.T., Bates, M.D., and Caron, M.G. (1990). Molecular cloning and expression of the gene for a human D1 dopamine receptor. Nature *347*, 72–76.
64. Zhou, Q.-Y., Grandy, D.K., Thambi, L., Kushner, J.A., Van Tol, H.H., Cone, R., Pribnow, D., Salon, J., Bunzow, J.R., and Civelli, O. (1990). Cloning and expression of human and rat D1 dopamine receptors. Nature *347*, 76–80.
65. Grandy, D.K., Marchionni, M.A., Makam, H., Stofko, R.E., Alfano, M., Frothingham, L., Fischer, J.B., Burke-Howie, K.J., Bunzow, J.R., and Server, A.C. (1989). Cloning of the cDNA and gene for a human D2 dopamine receptor. Proc. Natl. Acad. Sci. USA *86*, 9762–9766.
66. Hayes, G., Biden, T.J., Selbie, L.A., and Shine, J. (1992). Structural subtypes of the dopamine D2 receptor are functionally distinct: expression of the cloned D2A and D2B subtypes in a heterologous cell line. Mol. Endocrinol. *6*, 920–926.
67. De Backer, M.D., Gommeren, W., Moereels, H., Nobels, G., Vangompel, P., Leysen, J.E., and Luyten, W.H. (1993). Genomic cloning, heterologous expression and pharmacological characterization of a human histamine H1 receptor. Biochem. Biophys. Res. Commun. *197*, 1601–1608.
68. Ruat, M., Traiffort, E., Bouthenet, M.L., Schwartz, J.-C., Hirschfeld, J., Buschauer, A., and Schunack, W. (1990). Reversible and irreversible labeling and autoradiographic localization of the cerebral histamine H2 receptor using [125I] iodinated probes. Proc. Natl. Acad. Sci. USA *87*, 1658–1662.
69. Buckley, N.J., Bonner, T.I., Buckley, C.M., and Brann, M.R. (1989). Antagonist binding properties of five cloned muscarinic receptors expressed in CHO-K1 cells. Mol. Pharmacol. *35*, 469–476.
70. Luthin, G.R., and Wolfe, B.B. (1984). Comparison of [3H] pirenzepine and [3H] quinuclidinylbenzilate binding to muscarinic cholinergic receptors in rat brain. J. Pharmacol. Exp. Ther. *228*, 648–655.
71. Clark, M.J., Emmerson, P.J., Mansour, A., Akil, H., Woods, J.H., Portoghese, P.S., Remmers, A.E., and Medzihradsky, F. (1997). Opioid efficacy in a C6 glioma cell line stably expressing the delta opioid receptor. J. Pharmacol. Exp. Ther. *283*, 501–510.
72. Martin, N.A., and Prather, P.L. (2001). Interaction of co-expressed μ-and δ-opioid receptors in transfected rat pituitary GH3 cells. Mol. Pharmacol. *59*, 774–783.
73. Maguire, P., Tsai, N., Kamal, J., Cometta-Morini, C., Upton, C., and Loew, G. (1992). Pharmacological profiles of fentanyl analogs at mu, delta and kappa opiate receptors. Eur. J. Pharmacol. *213*, 219–225.
74. Simonin, F., Gavériaux-Ruff, C., Befort, K., Matthes, H., Lannes, B., Micheletti, G., Mattéi, M.G., Charron, G., Bloch, B., and Kieffer, B. (1995). kappa-Opioid receptor in humans: cDNA and genomic cloning, chromosomal assignment, functional expression, pharmacology, and expression pattern in the central nervous system. Proc. Natl. Acad. Sci. USA *92*, 7006–7010.
75. Martin, G.R., and Humphrey, P.P. (1994). Receptors for 5-hydroxytryptamine: current perspectives on classification and nomenclature. Neuropharmacology *33*, 261–273.
76. May, J.A., McLaughlin, M.A., Sharif, N.A., Hellberg, M.R., and Dean, T.R. (2003). Evaluation of the ocular hypotensive response of serotonin 5-HT1A and 5-HT2 receptor ligands in conscious ocular hypertensive cynomolgus monkeys. J. Pharmacol. Exp. Ther. *306*, 301–309.
77. Maier, D.L., Sobotka-Briner, C., Ding, M., Powell, M.E., Jiang, Q., Hill, G., Heys, J.R., Elmore, C.S., Pierson, M.E., and Mrzljak, L. (2009). [N-methyl-3H3] AZ10419369 binding to the 5-HT1B receptor: in vitro characterization and in vivo receptor occupancy. J. Pharmacol. Exp. Ther. *330*, 342–351.
78. Xie, Z., Lee, S.P., O'Dowd, B.F., and George, S.R. (1999). Serotonin 5-HT1B and 5-HT1D receptors form homodimers when expressed alone and heterodimers when co-expressed. FEBS Lett. *456*, 63–67.
79. Bonhaus, D.W., Bach, C., DeSouza, A., Salazar, F.H., Matsuoka, B.D., Zuppan, P., Chan, H.W., and Eglen, R.M. (1995). The pharmacology and distribution of human 5-hydroxytryptamine2B (5-HT2B) receptor gene products: comparison with 5-HT2A and 5-HT2C receptors. Br. J. Pharmacol. *115*, 622–628.
80. Saucier, C., and Albert, P.R. (1997). Identification of an endogenous 5-hydroxytryptamine2A receptor in NIH-3T3 cells: agonist-induced down-regulation involves decreases in receptor RNA and number. J. Neurochem. *68*, 1998–2011.
81. Ehlert, F.J., Roeske, W.R., Itoga, E., and Yamamura, H.I. (1982). The binding of [3H] nitrendipine to receptors for calcium channel antagonists in the heart, cerebral cortex, and ileum of rats. Life Sci. *30*, 2191–2202.
82. Gould, R.J., Murphy, K.M., and Snyder, S.H. (1982). [3H] nitrendipine-labeled calcium channels discriminate inorganic calcium agonists and antagonists. Proc. Natl. Acad. Sci. USA *79*, 3656–3660.
83. Huang, X.-P., Mangano, T., Hufeisen, S., Setola, V., and Roth, B.L. (2010). Identification of human Ether-à-go-go related gene modulators by three screening platforms in an academic drug-discovery setting. Assay Drug Dev. Technol. *8*, 727–742.

84. Finlayson, K., Turnbull, L., January, C.T., Sharkey, J., and Kelly, J.S. (2001). [3H] dofetilide binding to HERG transfected membranes: a potential high throughput preclinical screen. Eur. J. Pharmacol. *430*, 147–148.
85. Zhou, Z., Gong, Q., Ye, B., Fan, Z., Makielski, J.C., Robertson, G.A., and January, C.T. (1998). Properties of HERG channels stably expressed in HEK 293 cells studied at physiological temperature. Biophys. J. *74*, 230–241.
86. Kanda, H., Mimura, T., Hamasaki, K., Yamamoto, K., Yazaki, Y., Hirai, H., and Nojima, Y. (1999). Fyn and Lck tyrosine kinases regulate tyrosine phosphorylation of p105CasL, a member of the p130Cas docking protein family, in T-cell receptor-mediated signalling. Immunology *97*, 56–61.
87. Chan, C.-C., Boyce, S., Brideau, C., Charleson, S., Cromlish, W., Ethier, D., Evans, J., Ford-Hutchinson, A.W., Forrest, M.J., Gauthier, J.Y., et al. (1999). Rofecoxib [Vioxx, MK-0966; 4-(4′-methylsulfonylphenyl)-3-phenyl-2-(5H)-furanone]: a potent and orally active cyclooxygenase-2 inhibitor. Pharmacological and biochemical profiles. J. Pharmacol. Exp. Ther. *290*, 551–560.
88. Swinney, D.C., Mak, A.Y., Barnett, J., and Ramesha, C.S. (1997). Differential allosteric regulation of prostaglandin H synthase 1 and 2 by arachidonic acid. J. Biol. Chem. *272*, 12393–12398.
89. Riendeau, D., Charleson, S., Cromlish, W., Mancini, J.A., Wong, E., and Guay, J. (1997). Comparison of the cyclooxygenase-1 inhibitory properties of nonsteroidal anti-inflammatory drugs (NSAIDs) and selective COX-2 inhibitors, using sensitive microsomal and platelet assays. Can. J. Physiol. Pharmacol. *75*, 1088–1095.
90. Warner, T.D., Giuliano, F., Vojnovic, I., Bukasa, A., Mitchell, J.A., and Vane, J.R. (1999). Nonsteroid drug selectivities for cyclo-oxygenase-1 rather than cyclo-oxygenase-2 are associated with human gastrointestinal toxicity: a full in vitro analysis. Proc. Natl. Acad. Sci. USA *96*, 7563–7568.
91. Ellman, G.L., Courtney, K.D., ANDRES, V., Jr., and Featherstone, R.M. (1961). A new and rapid colorimetric determination of acetylcholinesterase activity. Biochem. Pharmacol. *7*, 88–95.
92. Nadarajah, B. (1992). The effect of pralidoxime chloride in the assay of acetylcholinesterase using 5, 5′-dithio-bis (2-nitrobenzoic acid)(Ellman's reagent). J. Anal. Toxicol. *16*, 192–193.
93. Hambleton, R., Krall, J., Tikishvili, E., Honeggar, M., Ahmad, F., Manganiello, V.C., and Movsesian, M.A. (2005). Isoforms of cyclic nucleotide phosphodiesterase PDE3 and their contribution to cAMP hydrolytic activity in subcellular fractions of human myocardium. J. Biol. Chem. *280*, 39168–39174.
94. Hung, S.-H., Zhang, W., Pixley, R.A., Jameson, B.A., Huang, Y.C., Colman, R.F., and Colman, R.W. (2006). New insights from the structure-function analysis of the catalytic region of human platelet phosphodiesterase 3A: a role for the unique 44-amino acid insert. J. Biol. Chem. *281*, 29236–29244.
95. Houslay, M.D. (2005). The long and short of vascular smooth muscle phosphodiesterase-4 as a putative therapeutic target. Mol. Pharmacol. *68*, 563–567.
96. MACKENZIE, S.J., and HOUSLAY, M.D. (2000). Action of rolipram on specific PDE4 cAMP phosphodiesterase isoforms and on the phosphorylation of cAMP-response-element-binding protein (CREB) and p38 mitogen-activated protein (MAP) kinase in U937 monocytic cells. Biochem. J. *347*, 571–578.
97. Galli, A., DeFelice, L.J., Duke, B.-J., Moore, K.R., and Blakely, R.D. (1995). Sodium-dependent norepinephrine-induced currents in norepinephrine-transporter-transfected HEK-293 cells blocked by cocaine and antidepressants. J. Exp. Biol. *198*, 2197–2212.
98. Giros, B., and Caron, M.G. (1993). Molecular characterization of the dopamine transporter. Trends Pharmacol. Sci. *14*, 43–49.
99. Gu, H., Wall, S.C., and Rudnick, G. (1994). Stable expression of biogenic amine transporters reveals differences in inhibitor sensitivity, kinetics, and ion dependence. J. Biol. Chem. *269*, 7124–7130.
100. Shearman, L.P., McReynolds, A.M., Zhou, F.C., and Meyer, J.S. (1998). Relationship between [125I] RTI-55-labeled cocaine binding sites and the serotonin transporter in rat placenta. Am. J. Physiol. *275*, C1621–C1629.
101. Wolf, W.A., and Kuhn, D.M. (1992). Role of essential sulfhydryl groups in drug interactions at the neuronal 5-HT transporter. Differences between amphetamines and 5-HT uptake inhibitors. J. Biol. Chem. *267*, 20820–20825.
102. Yin, R., Feng, B.Y., Varshney, A., and Pierce, B.G. (2022). Benchmarking AlphaFold for protein complex modeling reveals accuracy determinants. Protein Sci. *31*, e4379. https://doi.org/10.1002/pro.4379.
103. AlQuraishi, M. (2021). Protein-structure prediction revolutionized. Nature *596*, 487–488. https://doi.org/10.1038/d41586-021-02265-4.
104. Akdel, M., Pires, D.E.V., Pardo, E.P., Jänes, J., Zalevsky, A.O., Mészáros, B., Bryant, P., Good, L.L., Laskowski, R.A., Pozzati, G., et al. (2022). A structural biology community assessment of AlphaFold2 applications. Nat. Struct. Mol. Biol. *29*, 1056–1067. https://doi.org/10.1038/s41594-022-00849-w.
105. Scardino, V., Di Filippo, J.I., and Cavasotto, C.N. (2023). How good are AlphaFold models for docking-based virtual screening? iScience *26*, 105920. https://doi.org/10.1016/j.isci.2022.105920.
106. Holcomb, M., Chang, Y.T., Goodsell, D.S., and Forli, S. (2023). Evaluation of AlphaFold2 structures as docking targets. Protein Sci. *32*, e4530. https://doi.org/10.1002/pro.4530.
107. Bodenreider, O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. *32*, D267–D270. https://doi.org/10.1093/nar/gkh061.
108. Bento, A.P., Gaulton, A., Hersey, A., Bellis, L.J., Chambers, J., Davies, M., Krüger, F.A., Light, Y., Mak, L., McGlinchey, S., et al. (2014). The ChEMBL bioactivity database: an update. Nucleic Acids Res. *42*, D1083–D1090. https://doi.org/10.1093/nar/gkt1031.
109. Liu, T., Lin, Y., Wen, X., Jorissen, R.N., and Gilson, M.K. (2007). BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. Nucleic Acids Res. *35*, D198–D201. https://doi.org/10.1093/nar/gkl999.
110. Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., Pon, A., Banco, K., Mak, C., Neveu, V., et al. (2011). DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. Nucleic Acids Res. *39*, D1035–D1041. https://doi.org/10.1093/nar/gkq1126.
111. Ravindranath, P.A., and Sanner, M.F. (2016). AutoSite: an automated approach for pseudo-ligands prediction-from ligand-binding sites identification to predicting key ligand atoms. Bioinformatics *32*, 3142–3149. https://doi.org/10.1093/bioinformatics/btw367.
112. O'Boyle, N.M., Banck, M., James, C.A., Morley, C., Vandermeersch, T., and Hutchison, G.R. (2011). Open Babel: An open chemical toolbox. J. Cheminform. *3*, 33. https://doi.org/10.1186/1758-2946-3-33.
113. Gasteiger, J., and Marsili, M. (1980). Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. Tetrahedron *36*, 3219–3228. https://doi.org/10.1016/0040-4020(80)80168-2.
114. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2.
115. Paysan-Lafosse, T., Blum, M., Chuguransky, S., Grego, T., Pinto, B.L., Salazar, G.A., Bileschi, M.L., Bork, P., Bridge, A., Colwell, L., et al. (2023). InterPro in 2022. Nucleic Acids Res. *51*, D418–D427. https://doi.org/10.1093/nar/gkac993.
116. Kotera, M., Tabei, Y., Yamanishi, Y., Moriya, Y., Tokimatsu, T., Kanehisa, M., and Goto, S. (2013). KCF-S: KEGG Chemical Function and Substructure for improved interpretability and prediction in chemical bioinformatics. BMC Syst. Biol. *7*, S2. https://doi.org/10.1186/1752-0509-7-S6-S2.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
AutoDock Vina	Version 1.1	https://vina.scripps.edu/
AutoSite	Version 1.0	https://ccsb.scripps.edu/autosite/
Open Babel	Version 3.1.0	https://openbabel.org/index.html
Open-Source PyMOL	Version 2.5.0	https://github.com/schrodinger/pymol-open-source/
Python	Version 3.9.11	https://www.python.org/downloads/
RDKit	Version 2020.09.1.0	https://www.rdkit.org
Other		
PBAS profiles	This study	https://yamanishi.cs.i.nagoya-u.ac.jp/pbas/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to the lead contact, Yoshihiro Yamanishi (yamanishi@i.nagoya-u.ac.jp).

Materials availability

This study did not generate new unique materials.

Data and code availability

- PBAS profiles, ligand-binding pocket information for docking simulation and four types of compound profiles for side effect predictive models were available at https://yamanishi.cs.i.nagoya-u.ac.jp/pbas/.
- All data reported in this paper can be made available by lead contact upon request.
- This paper does not report the original code.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

In vitro binding and enzymatic assays

The validity of the results of Compound–protein interactions (CPIs) predicted by docking simulations was verified by binding studies. The molecular targets for the binding studies were 30 proteins involved in therapeutic and side effects.[53] The targets included 19 G-protein-coupled receptors, 2 ion channels, 6 enzymes, and 3 transporters expressed in the central nervous, cardiovascular, respiratory, digestive, and renal systems. The 30 proteins are listed below: A2A Human Adenosine GPCR (A2A), α1A Human Adrenoceptor GPCR (α1A), α2A Human Adrenoceptor GPCR (α2A), β1 Human Adrenoceptor GPCR (β1), β2 Human Adrenoceptor GPCR (β2), CB1 Human Cannabinoid GPCR (CB1), D1 Human Dopamine GPCR (D1), D2S Human Dopamine GPCR (D2S), H1 Human Histamine GPCR (H1), H2 Human Histamine GPCR (H2), M1 Human Acetylcholine (Muscarinic) GPCR (M1), M3 Human Acetylcholine (Muscarinic) GPCR (M3), δ(DOP) Human Opioid GPCR (δ (DOP)), κ (KOP) Human Opioid GPCR (κ (KOP)), μ (MOP) Human Opioid GPCR (μ(MOP)), 5-HT1A Human Serotonin GPCR (5-HT1A), 5HT1B Human Serotonin GPCR (5-HT1B), 5-HT2A Human Serotonin GPCR(5-HT2A), 5-HT2B Human Serotonin GPCR (5-HT2B), Cav1.2 (L-type) Rat Calcium Ion Channel (Dihydropyridine Site)(Cav1.2 (L-type)), hERG Human Potassium Ion Channel [3H] Dofetilide (hERG), Lck Human TK kinase (Lck), COX-1 Human Cyclooxygenase (COX-1), COX-2 Human Cyclooxygenase (COX-2), Acetylcholinesterase, PDE3A Human Phosphodiesterase (PDE3A), PDE4D2 Human Phosphodiesterase (PDE4D), NET Human Norepinephrine Transporter (NET), DAT Human Dopamine Transporter (DAT), Serotonin Transporter (SET) Human SET.

For the Binding Assay, one of Radioligand binding, spectrofluorimetry and spectrophotometry was used. The inhibition rate (%Inh.) was calculated by measuring the extent to which the additional compound inhibited the original ligand's specific binding to the protein. Details of the panel target list, which includes 30 target proteins, are given in Table S7.

Details of experimental methods for each protein type

Assay no. 1: A2A Human Adenosine GPCR (A2A)

Human recombinant adenosine A_{2A} receptors[54] expressed in human HEK-293 cells are used in modified Tris-HCl buffer pH 7.4. A15 μg aliquot is incubated with 50 nM [^{3}H] CGS-21680 for 90 min at 25°C. Non-specific binding is estimated in the presence of 50 μM NECA. Receptors are filtered and washed, the filters are then counted to determine [^{3}H] CGS- 21680 specifically bound. Compounds are screened at 10 μM.

Assay no. 2:α1A Human Adrenoceptor GPCR (α1A)

Human recombinant adrenergic α_{1A} receptors[55,56] expressed in human Chem-1 cells are used in modified HEPES buffer pH 7.4. A 2 μg aliquot is incubated with 0.6 nM [^{3}H] Prazosin for 60 min at 25°C. Non-specific binding is estimated in the presence of 10 μM phentolamine. Receptors are filtered and washed, the filters are then counted to determine [^{3}H] Prazosin specifically bound. Compounds are screened at 10 μM.

Assay no. 3:α2A Human Adrenoceptor GPCR (α2A)

Human recombinant adrenergic α_{2A} receptors[57,58] expressed in CHO-K1 cells are used in modified Tris-HCl buffer pH 7.4. A 2 μg aliquot is incubated with 1.5 nM [^{3}H] Rauwolscine for 60 min at 25°C. Non-specific binding is estimated in the presence of 10 μM WB-4101. Receptors are filtered and washed, the filters are then counted to determine [^{3}H] Rauwolscine specifically bound. Compounds are screened at 10 μM.

Assay no. 4:β1 Human Adrenoceptor GPCR (β1)

Human recombinant adrenergic β_1 receptors[59] expressed in CHO-K1 cells are used in modified Tris-HCl buffer pH 7.4. A 25 μg aliquot is incubated with 0.03 nM [^{125}I] Cyanopindolol for 120 min at 25°C. Non-specific binding is estimated in the presence of 100 μM S (−)-Propranolol. Receptors are filtered and washed, the filters are then counted to determine [^{125}I] Cyanopindolol specifically bound. Compounds are screened at 10 μM.

Assay no. 5:β2 Human Adrenoceptor GPCR (β2)

Human recombinant adrenergic β_2 receptors[60] expressed in CHO cells are used in modified Tris-HCl buffer pH 7.4. A 50 μg aliquot is incubated with 0.2 nM [^{3}H] CGP-12177 for 60 min at 25°C. Non-specific binding is estimated in the presence of 10 μM ICI-118551. Receptors are filtered and washed, the filters are then counted to determine [^{3}H] CGP-12177 specifically bound. Compounds are screened at 10 μM.

Assay no. 6: CB1 Human Cannabinoid GPCR (CB1)

Human recombinant cannabinoid CB_1 receptors[61,62] expressed in rat hematopoietic Chem-1 cells are used in modified HEPES buffer pH 7.4. A 5 μg aliquot of membrane is incubated with 2 nM [^{3}H]SR141716A for 60 min at 37°C. Non-specific binding is estimated in the presence of 10 μM CP 55,940. Membranes are filtered and washed 4 times and the filters are counted to determine [^{3}H] SR141716A specifically bound. Compounds are screened at 10 μM.

Assay no. 7: D1 Human Dopamine GPCR (D1)

Human recombinant dopamine D_1 receptors[63,64] expressed in CHO cells are used in modified Tris-HCl buffer pH 7.4. A 20 μg aliquot is incubated with 1.4 nM [^{3}H] SCH-23390 for 120 min at 37°C. Non-specific binding is estimated in the presence of 10 μM (+)-butaclamol. Receptors are filtered and washed, the filters are then counted to determine [^{3}H] SCH-23390 specifically bound. Compounds are screened at 10 μM.

Assay no. 8: D2S Human Dopamine GPCR (D2S)

Human recombinant dopamine D_{2S} receptors[65,66] expressed in CHO cells are used in modified Tris-HCl buffer pH 7.4. A 15 μg aliquot is incubated with 0.16 nM [^{3}H] Spiperone for 120 min at 25°C. Non-specific binding is estimated in the presence of 10 μM haloperidol. Receptors are filtered and washed, the filters are then counted to determine [^{3}H] Spiperone specifically bound. Compounds are screened at 10 μM.

Assay no. 9: H1 Human Histamine GPCR (H1)

Human recombinant histamine H_1 receptors[67] expressed in CHO-K1 cells are used in modified Tris-HCl buffer pH 7.4. A 10 μg aliquot is incubated with 1.2 nM [^{3}H] Pyrilamine for 180 min at 25°C. Non-specific binding is estimated in the presence of 1 μM pyrilamine. Receptors are filtered and washed, the filters are then counted to determine [^{3}H] Pyrilamine specifically bound. Compounds are screened at 10 μM.

Assay no. 10: H2 Human Histamine GPCR (H2)

Human recombinant histamine H_2 receptors[68] expressed in CHO-K1 cells are prepared in K-Na phosphate buffer pH 7.4. A 2 μg aliquot is incubated with 0.1 nM [^{125}I] Aminopotentidine for 120 min at 25°C. Non-specific binding is estimated in the presence of 3 μM Tiotidine. Receptors are filtered and washed, the filters are then counted to determine [^{125}I] Aminopotentidine specifically bound. Compounds are screened at 10 μM.

Assay no. 11: M1 Human Acetylcholine (muscarinic) GPCR (M1)

Human recombinant muscarinic M_1 receptors[69,70] expressed in CHO-K1 cells are used in modified Tris-HCl buffer pH 7.4. A 16 μg aliquot is incubated with 0.8 nM [^{3}H] N-Methylscopolamine for 120 min at 25°C. Non-specific binding is estimated in the presence of 1 μM Atropine. Receptors are filtered and washed, the filters are then counted to determine [^{3}H] N-Methylscopolamine specifically bound. Compounds are screened at 10 μM.

Assay no. 12: M3 Human Acetylcholine (muscarinic) GPCR (M3)

Human recombinant muscarinic M_3 receptors[69,70] expressed in CHO-K1 cells are used in modified Tris-HCl buffer pH 7.4. A 12 μg aliquot is incubated with 0.8 nM [^{3}H] N-Methylscopolamine for 120 min at 25°C. Non-specific binding is estimated in the presence of 1 μM atropine. Receptors are filtered and washed, the filters are then counted to determine [^{3}H] N- Methylscopolamine specifically bound. Compounds are screened at 10 μM.

Assay no. 13:δ(DOP) Human Opioid GPCR (δ (DOP))

Human recombinant opiateδ_1 receptors[71,72] expressed in HEK293 cells are used in modified Tris-HCl buffer pH 7.4. A 9 μg aliquot is incubated with 1.3 nM [^{3}H] Naltrindole for 60 min at 25°C. Nonspecific binding is estimated in the presence of 1 μM Naltrindole. Receptors are filtered and washed, the filters are then counted to determine [^{3}H] Naltrindole specifically bound. Compounds are screened at 10 μM.

Assay no. 14: κ (KOP) Human Opioid GPCR (κ (KOP))

Human recombinant opiate κ receptors[73,74] expressed in human HEK-293 cells are used in modified Tris-HCl buffer pH 7.4. A 30 μg aliquot is incubated with 0.6 nM [^{3}H] Diprenorphine for 60 min at 25°C. Nonspecific binding is estimated in the presence of 10 μM naloxone. Receptors are filtered and washed, the filters are then counted to determine [^{3}H] Diprenorphine specifically bound. Compounds are screened at 10 μM.

Assay no. 15: μ (MOP) Human Opioid GPCR (μ(MOP))

Human recombinant opiate μ receptors[74] expressed in CHO-K1 cells are used in modified Tris-HCl buffer pH 7.4. A 11μg aliquot is incubated with 0.6 nM [^{3}H] Diprenorphine for 60 min at 25°C. Nonspecific binding is estimated in the presence of 10 μM naloxone. Receptors are filtered and washed, the filters are then counted to determine [^{3}H] Diprenorphine specifically bound. Compounds are screened at 10 μM.

Assay no. 16: 5-HT1A Human Serotonin GPCR (5-HT1A)

Human recombinant serotonin 5-HT_{1A} receptors[75,76] expressed in CHO-K1 cells are used in modified Tris-HCl buffer pH 7.4. An 8 μg aliquot is incubated with 1.5 nM [^{3}H]8-OH-DPAT for 60 min at 25°C. Non-specific binding is estimated in the presence of 10 μM metergoline. Receptors are filtered and washed, the filters are then counted to determine [^{3}H]8-OH-DPAT specifically bound. Compounds are screened at 10 μM.

Assay no. 17: 5HT1B Human Serotonin GPCR (5-HT1B)

Human recombinant serotonin 5-HT_{1B} receptors[77,78] expressed in Chem-1 cells are used in modified Tris-HCl buffer pH 7.4. A 2 μg aliquot of membrane is incubated with 1 nM [^{3}H]GR125743 for 90 min at 37°C. Non-specific binding is estimated in the presence of 10 μM 5-HT. Membranes are filtered and washed, the filters are then counted to determine [^{3}H]GR125743 specifically bound. Compounds are screened at 10 μM.

Assay no. 18: 5-HT2A Human Serotonin GPCR (5-HT2A)

Human recombinant serotonin 5-HT_{2A} receptors[79,80] expressed in CHO-K1 cells are used in modified Tris-HCl buffer pH 7.4. A 30 μg aliquot is incubated with 0.5 nM [^{3}H]Ketanserin for 60 min at 25°C. Non-specific binding is estimated in the presence of 1 μM Mianserin. Receptors are filtered and washed, the filters are then counted to determine [^{3}H] Ketanserin specifically bound. Compounds are screened at 10 μM.

Assay no. 19: 5-HT2B Human Serotonin GPCR (5-HT2B)

Human recombinant serotonin 5-HT_{2B} receptor[79] expressed in CHO-K1 cells are used to prepare membranes in modified Tris-HCl buffer pH 7.4. A 30 μg aliquot of membrane protein is incubated with 1.2 nM [^{3}H] LSD for 60 min at 37°C. Non-specific binding is estimated in the presence of 10 μM serotonin. Membranes are filtered and washed, the filters are then counted to determine [^{3}H] LSD specifically bound. Compounds are screened at 10 μM.

Assay no. 20: Cav1.2 (L-type) Rat Calcium Ion Channel (Dihydropyridine Site)(Cav1.2 (L-type))

Cerebral cortices of male Wistar derived rats weighing 175 ± 25 g are used to prepare L-type dihydropyridine calcium channel[81,82] in Tris-HCl buffer pH 7.4. A 2.5 mg aliquot is incubated with 0.1 nM [^{3}H] Nitrendipine for 90 min at 25°C. Non-specific binding is estimated in the presence of 1 μM nitrendipine. Membranes are filtered and washed, the filters are then counted to determine [^{3}H] Nitrendipine specifically bound. Compounds are screened at 10 μM.

Assay no. 21: hERG Human Potassium Ion Channel [^{3}H] Dofetilide (hERG)

Human recombinant potassium channel HERG[83–85] expressed in human HEK- 293 cells are used in modified Tris-HCl buffer pH 7.4. A 7.5 μg aliquot is incubated with 3 nM [^{3}H]Dofetilide for 60 min at 25°C. Non-specific binding is estimated in the presence of 10 μM Dofetilide. Channel proteins are filtered and washed, the filters are then counted to determine [^{3}H] Dofetilide specifically bound. Compounds are screened at 10 μM.

Assay no. 22: Lck Human TK kinase (Lck)

Human recombinant protein kinase LCK[86] expressed in insect cells is used. Test compound and/or vehicle is preincubated with 0.4 μg/mL enzyme in modified HEPES buffer pH 7.4 for 15 min at 37°C. The reaction is initiated by addition of 0.2 mg/mL poly (Glu:Tyr), 10 μM ATP and 0.25 μCi [γ^{32}P]ATP for another 30 min incubation period and terminated by further addition of 3% H_3PO_4. An aliquot is removed and counted to determine the amount of [^{32}P] Poly (Glu:Tyr) formed. Compounds are screened at 10 μM.

Assay no. 23: COX-1 Human Cyclooxygenase (COX-1)

Human recombinant cyclooxygenase COX-1[87,88] expressed in baculovirus infected Sf9 cells are used. Test compound and/or vehicle is incubated with 0.44 μg/mL COX-1 in modified Tris-HCl buffer pH 8.0 for 15 min at 25°C. The reaction is initiated by addition of 3 μM arachidonic acid and 100 μM Ampliflu Red for another 3 min incubation period. Determination of the amount of Resorufin formed is read spectrofluorimetrically at 535 nm/590 nm. Compounds are screened at 10 μM.

Assay no. 24: COX-2 Human Cyclooxygenase (COX-2)

Human recombinant cyclooxygenase-2[89,90] expressed in insect Sf21 cells is used. Test compound and/or vehicle is preincubated with 34 U/mL enzyme in modified Tris-HCl buffer pH 8.0 for 15 min at 25°C. The reaction is initiated by addition of 3 μM arachidonic acid and 100 μM Ampliflu Red for another 3 min incubation period. Determination of the amount of Resorufin formed is read spectrofluorimetrically at 535 nm/590 nm. Compounds are screened at 10 μM.

Assay no. 25: Acetylcholinesterase

Human recombinant acetylcholinesterase[91,92] expressed in HEK-293 cells (Sigma, C-1682) is used. Test compound and/or vehicle is preincubated with 4.1 ng/mL of enzyme for 15 min at 25°C in phosphate buffer pH 7.4. The reaction is initiated by addition of 0.7 mM acetylthiocholine iodide and 0.5 mM 5,5-dithiobis-2-nitrobenzoic acid for another 20 min incubation period. The thiocholine generated reacts continuously with dithiobisnitrobenzoic acid to produce 5-thio-2-nitro-benzoic acid, and its spectrophotometric absorbance is read at 405 nm. Compounds are screened at 10 μM.

Assay no. 26: PDE3A Human Phosphodiesterase (PDE3A)

Human recombinant PDE3A[93,94] expressed in insect Sf9 cells are used. Test compound and/or vehicle is preincubated with 20 ng/mL enzyme in Tris- HCl buffer pH 7.2 for 15 min at 25°C. The reaction is initiated by addition of 100 nM fluorescein labeled cAMP for another 30 min incubation period and terminated by addition of IMAP binding solution. IMAP complexes with phosphate groups on nucleotide monophosphate generated from cyclic nucleotides through PDE activity. Determination of the amount of complex formed is read spectrofluorimetrically at 470 nm/525 nm. Compounds are screened at 10 μM.

Assay no. 27: PDE4D2 Human Phosphodiesterase (PDE4D)

Human recombinant PDE4D2[95,96] expressed in insect Sf9 cells are used. Test compound and/or vehicle is preincubated with 5 ng/mL enzyme in Tris-HCl buffer pH 7.2 for 15 min at 25°C. The reaction is initiated by addition of 100 nM fluorescein labeled cAMP for another 15 min incubation period and terminated by addition of IMAP binding solution. IMAP complexes with phosphate groups on nucleotide monophosphate generated from cyclic nucleotides through PDE activity. Determination of the amount of complex formed is read spectrofluorimetrically at 470 nm/525 nm. Compounds are screened at 10 μM.

Assay no. 28: NET Human Norepinephrine Transporter (NET)

Human recombinant norepinephrine transporters[97] expressed in dog kidney MDCK cells are used in modified Tris-HCl buffer pH 7.4. A 40 μg aliquot is incubated with 0.2 nM [^{125}I] RTI-55 for 3 h at 4°C. Non-specific binding is estimated in the presence of 10 μM desipramine. Transporters are filtered and washed, the filters are then counted to determine [^{125}I] RTI- 55 specifically bound. Compounds are screened at 10 μM.

Assay no. 29: DAT Human Dopamine Transporter (DAT)

Human recombinant dopamine transporters[98,99] expressed in CHO-S cells are used in modified Tris-HCl buffer pH 7.4. A 0.4 μg aliquot is incubated with 0.15 nM [^{125}I] RTI-55 for 3 h at 4°C. Non-specific binding is estimated in the presence of 10 μM nomifensine. Transporter are filtered and washed, the filters are then counted to determine [^{125}I] RTI-55 specifically bound. Compounds are screened at 10 μM.

Assay no. 30: Serotonin Transporter Human Serotonin Transporter

Human recombinant SET[100,101] expressed in human HEK-293 cell are used in modified Tris-HCl buffer pH 7.4. A 9 μg aliquot is incubated with 0.4 nM [^{3}H] Paroxetine for 60 min at 25°C. Non-specific binding is estimated in the presence of 10 μM imipramine. Transporters are filtered and washed, the filters are then counted to determine [^{3}H] Paroxetine specifically bound. Compounds are screened at 10 μM.

METHOD DETAILS

Protein 3D structures

The 3D structures of all human proteins were obtained from the AlphaFoldDB[26] (20,594 proteins and 23,391 structures, UniProt proteome ID: UP000005640, AlphaFoldDB structure version 1). The number of structures was greater than that of proteins because proteins with long amino acid sequences were divided into fragment sequences before the 3D structure prediction. Therefore, multiple structures could arise from a single protein, but these structures were treated separately.

Next, the low-confidence parts of the protein structure predicted using AlphaFold were removed. In the AlphaFold-based framework, each amino acid in the resulting structure is assigned a prediction confidence level (i.e., pLDDT, PAE, and pTMP).[102] The confidence level primarily reflects the confidence level of the training data, and many of the regions with low confidence levels are intrinsically disordered protein (IDP) domains.[103,104] In general, IDP domains do not have a fixed structure, and the rigid-body model in the docking simulations cannot correctly evaluate the binding affinity with the protein regions without fixed structure and ligand compounds. Therefore, to minimize the influence of these regions of undetermined structure on the docking simulation results, the parts with low predictive reliability were removed in advance. The criterion for removal was pLDDT >70, which is considered highly reliable in the AlphaFoldDB. After preprocessing the structures and the binding pocket search (for details, see the "detection of ligand-binding pockets for docking simulation" subsection), the final number of 3D structures available for docking simulation were found to be 19,135 structures and 18,347 protein types.

Some studies indicate that docking simulations using AlphaFold predicted structures is not good enough.[105,106] However, the aim of this study is to explore the potential applications of assessing the binding affinity of each drug to all human proteins in various crucial aspects of drug discovery, such as predicting indications and side effects. While employing a homology modeling structure could sometimes yield more precise binding affinities, we opted for the AlphaFold structure in this research to prioritize the comprehensive coverage of protein structures across the genome.

Drug chemical structures

The two-dimensional (2D) structures of 8,112 drugs were obtained from the KEGG DRUG[27] database, which stores a variety of information on drugs registered in Japan, Europe, and the United States. The KEGG DRUG database comprises approximately 8,000 drugs that have been approved in Japan, Europe, and the United States. Consequently, the drugs examined in this study encompass all those approved by the FDA. Additionally, optimized 3D structures were generated from 2D data using RDKit (RDKit: Open-source cheminformatics, https://www.rdkit.org).

For the preliminary analysis of the PBAS profiles, the information on drug approval year was also obtained from KEGG DRUG.[27] Of the drugs, for which docking simulations were performed, the year of approval for 524 drugs was obtained (Figure S9).

Fingerprint profiles representing the substructures and related features of the drugs were prepared for comparisons of side effect predictions. The fingerprint profiles were prepared according to the protocol used in previous studies.[2,5] The Morgan fingerprint,[40] an extended-connectivity fingerprints (ECFP)-like fingerprint, was calculated. RDKit was used to calculate the Morgan fingerprints, and the number of features in each fingerprint was set to 2,048.

Disease–therapeutic target protein associations

We manually constructed a set of the relationships between diseases and therapeutic target proteins for use in the prediction of drug therapeutic indications. There are already several databases that integrate the data on diseases and genes, but most of them focus on genetic diseases, and the genes correlate to disease-causing genes, which do not always correspond to the therapeutic target proteins. We collected and manually curated the information on the therapeutic target proteins of various diseases from the medical literature. Finally, 2,062 disease–target protein relationships were prepared, including 250 diseases and 462 proteins.

Drug side effects used for training dataset in the side effect prediction model

The information on drug side effects in a training set for the side effect prediction model was obtained from the SIDER database.[41] SIDER is a large database that collects information on side effects from pharmaceutical package inserts. Because it is difficult to correctly predict rare adverse drug reactions in the framework of supervised learning, we used only side effects that have been frequently reported. The criteria are as follows: (1) Side effects are defined as "frequent" or with a frequency of $\geq$1% in the frequency information defined in SIDER. (2) Additionally, the types of side effects correspond to those classified as "disorder" in the Unified Medical Language System.[107] (3) Only drugs, for which at least 10 side effects have been reported, are included. Based on these criteria, a dataset of 15,035 drug side effect pairs involving 429 drugs and 285 side effects was constructed. This dataset was also used as the gold standard data for cross-validation experiments during the performance evaluation of the prediction model.

Drug side effects used as an independent resource for performance evaluation

To evaluate the validity of the prediction results using external resources, adverse event reports were extracted from the FAERS database for validation. FAERS is a large database that contains reports of adverse events that occurred after the drug was marketed and used. It was independent of the SIDER database used for training the predictive model. The adverse drug reactions reported from the first quarter of 2004 to the second quarter of 2019 were downloaded from the FDA website (https://www.fda.gov). We used adverse drug reactions that were detected with a frequency of $\geq$1% for drugs with >300 reports of the drug itself. To further exclude water, metal ions, and macromolecules, the following conditions were set: drugs composed of three or more atoms with molecular weights ranging from 50 to 1,000. Only side effects that were the outputs of the SIDER learning model were selected. The final validation dataset consisted of 1,673 drugs, 280 side effects, and 83,456 drug–side effect relationships. The overlap between the 429 SIDER drugs used in the training data and the 1,673 FAERS drugs in the validation data was 282.

Compound–protein interactions for side effect prediction methods

A CPI dataset was prepared to generate the two target protein profiles, namely TESS and TELR, for the comparison of accuracy in predicting side effects. According to the previous study,[5] the CPIs were obtained from public databases including ChEMBL,[108] BindingDB,[109] DrugBank,[110] and KEGG DRUG.[27] With reference to existing reports,[45] we selected only CPIs that clearly showed pharmacological activity (e.g., half maximal inhibitory concentration of <1 μM. Finally, 1,830,624 CPIs were investigated for 1,288,343 compounds and 4,643 proteins.

Detection of ligand-binding pockets for docking simulation

Ligand-binding pockets were detected using AutoSite[111] for docking simulations. First, hydrogen atoms were added to the protein conformation data (pdb file) using Open Babel 3.1.0[112] at pH 7.4. Next, the Gasteiger charge[113] was calculated for each atom and converted to a pdbqt file. AutoSite detects multiple ligand-binding pockets ranked according to their ligand-binding site likeness. Only the top-1-ranked ligand-binding pocket for each protein was used for docking simulations. All the detected ligand binding pocket of each protein are shown in the supplementary information.

Docking simulation to construct the proteome-wide binding affinity score profiles

We conducted large-scale docking simulations for drugs on human whole-protein 3D structure data for constructing the PBAS profiles. AutoDock Vina v1.1,[28] a widely used docking program elaborated by the Scripps Research Institute, was applied. It evaluates the binding affinity of a compound to a protein with high speed and moderate scoring accuracy. Protein 3D structure data (pdbqt file format) generated during the detection of binding pocket sites were used for docking simulations on the binding pocket regions. For the parameter option for running Auto dock vina, the "energy_range" was set to 4, and default values were used for the other options. This docking setting was fixed for all ligand-receptor pairs. All estimated docking results for possible pairs of drugs and human proteins are publicly available as PBAS profiles in the supplementary information.

In this study, our objective is to conduct comprehensive docking simulations for all potential combinations of drugs and human proteins. While numerous docking simulation software and tools exist, each presents a trade-off between calculation accuracy and speed. We opted for AutoDock Vina due to its rapid calculation speed and cost-free availability, although it may yield less accurate binding affinity values compared to commercial alternatives.

Clustering and enrichment analysis of the proteome-wide binding affinity score profiles

Hierarchical clustering and enrichment analysis were performed to investigate the similarities of drugs and proteins according to their binding affinity patterns encoded in the PBAS profiles. Hierarchical clustering was conducted on both the protein and drug sides of the PBAS profile matrix. Clustering with the Ward algorithm was performed using the Python SciPy[114] library.

An enrichment analysis of the ATC classification was conducted for each drug cluster to examine the relationships between drugs and their therapeutic indications. The ATC classification of drugs was obtained from KEGG DRUG,[27] and *p*-values were calculated with Fisher's exact test using the Python SciPy[114] library. Only ATC codes that were statistically significant ($p < 0.05$) were assigned and extracted.

To examine the relationships between proteins and their structural features, DAVID[29] was used for each cluster to calculate the enrichment of the proteins classified in InterPro.[115] InterPro is a taxonomic database of protein families, domains and functional sites, and it allows functional characterization of known proteins. Only InterPro structural domain groups that were statistically significant ($p < 0.05$) assignments were identified and extracted.

Performance test with CASF-2016 core set

The CASF-2016 core set,[33] obtained from BindingDB, served as the validation dataset. The docking site used binding pocket information outlined in the CASF-2016 core set. Using the Open Source version of PyMOL (https://github.com/schrodinger/pymol-open-source), binding pocket structure data and corresponding protein AlphaFold structures were aligned. The align command facilitated the superimposition of structures, defining a box region encompassing all matched atoms as the docking site. Three validation datasets were prepared: All, Monomer, and PBAS. "All" denotes the set of pairs (254 ligand-receptor pairs) assessable using the proposed method (AlphaFold structure + AutoDock Vina) from all ligand-receptor pairs in the CASF-2016 core set. Although the CASF-2016 core set encompasses approximately

100 pairs of complex protein receptors, these complexes were excluded in the proposed method, necessitating the creation of a fair evaluation subset, Monomer, comprising only monomer pairs (167 ligand-receptor pairs). Furthermore, to evaluate the performance of the PBAS profile introduced in this study, a subset PBAS was derived, selecting only protein receptors included in the PBAS profile from the Monomer set (97 ligand-receptor pairs). AutoDock Vina's docking options mirrored those used in PBAS.

Prediction of drug therapeutic indications by template matching

For therapeutic indication prediction, we used the template matching method[2] based on the therapeutic target protein–disease associations. The previous method used only known target proteins with at least one known ligand, whereas our method used PBAS consisting of all proteins. The predictive score (referred to as *TMscore*) of the *i*-th drug against the *k*-th disease was calculated using the following formula:

$$TMscore(i,k) = min\left(\mathbf{x}_i^{(PBAS)} \odot t_k\right)$$

where $\odot$ is the Adamar product, meaning the element-by-element product of the vectors. $\mathbf{x}_i^{(PBAS)}$ is a vector of PBAS profiles of the *i*-th drug ($\mathbf{x}^{(PBAS)} = (e_1,e_2,\ldots,e_j,\ldots,e_d)\,|\,e_j \in \mathbb{R}, e_j < 0, j = 1,\ldots,d$), where element e_j indicates the binding affinity to the *j*-th protein. There are *d* proteins in total. $\mathbf{t}_k$ is a multihot vector ($\mathbf{t} = (t_1,t_2,\ldots,t_j,\ldots,t_d)\,|\,t_j \in \{0,1\}, j = 1,\ldots,d$) representing therapeutic target proteins for the *k*-th disease. If the *j*-th protein is a therapeutic target for a disease, t_j returns 1, and if it is not, t_j returns 0.

The *TMscore* is calculated for all drug–disease pairs. The final prediction score is the binding free energy obtained from the docking score. Therefore, the prediction method is designed to select drugs that bind more strongly to the therapeutic target protein for a disease. The use of the template matching method enables the provision of a drug-protein-disease network, making the biological interpretation of predicted drug therapeutic indications easier.

Construction of TESS profiles

The target protein profile for each drug was prepared by TESS, which is referred to as a TESS profile and used for comparing the side effect prediction performance. In accordance with previous studies,[2,5] the similarities of each drug with compounds in the CPI dataset were evaluated.

The predicted interaction score (*TESSscore*) between drug *i* and protein *j* was calculated as follows:

$$TESSscore(i,j) = \max\left(\mathbf{s}_i^{(sim)} \odot \mathbf{c}_j^{(CPI)}\right)$$

where $\odot$ is the Adamar product, meaning the element-wise product of the vectors. $\mathbf{s}_i^{(sim)}$ is the structural similarity profile of drug *i* to compounds in the CPI dataset ($\mathbf{s}^{(sim)} = (s_1,s_2,\ldots,s_u,\ldots,s_n)|u = 1,\ldots,n$). s_u is the structural similarity between drug *i* and CPI dataset compound *u*. $\mathbf{c}_j^{(CPI)}$ is a multihot vector showing the presence or absence of an interaction between protein *j* and compounds in the CPI dataset. For the vector ($\mathbf{c}^{(CPI)} = (c_1,c_2,\ldots,c_u,\ldots,c_n)\,|\,c_u \in \{0,1\}, u = 1,\ldots,n$), the value of element c_u is 1 if protein *j* interacts and 0 if it does not. There are *n* compounds in the database in total, and the compounds in the CPI dataset indicated by each element between vector $\mathbf{s}^{(sim)}$ and vector $\mathbf{c}^{(CPI)}$ are matched.

s_u is the structural similarity between drug *i* and compound *u* in the CPI dataset. It was calculated using the generalized Jaccard correlation coefficient (*GJ*) as follows:[2]

$$s_u = GJ\left(f_i^{(FP)}, f_u^{(FP)}\right) = \frac{\sum_h^l min(a_h,b_h)}{\sum_h^l max(a_h,b_h)}$$

where drug *i* is represented by the binary feature vector as $\mathbf{f}_i^{(FP)} = (a_1,a_2,\ldots,a_l)$, and compound *u* in the CPI dataset is represented by the binary feature vector as $\mathbf{f}_u^{(FP)} = (b_1,b_2,\ldots,b_l)$, where *l* is the number of features. KEGG chemical function and substructure[116] is used ($l = 24,401$) as a chemical descriptor, where $\min(\cdot,\cdot)$ takes the minimum of the two given values and $\max(\cdot,\cdot)$ stands for the maximum.

Construction of target estimation with logistic regression profiles

The target protein profile for each drug was prepared using TELR, which is referred to as a TELR profile and used for comparing the side effect prediction performance. Following previous studies,[2,5] a target protein profile was developed using a supervised classification framework with the CPI dataset.

The supervised learning method predicts the likelihood of interaction with individual target proteins in CPI data using the chemical structure profile $\mathbf{z}^{(chem)}$ of a query compound and L1-regularized logistic regression classifier. For a training set consisting of *n* compounds and *p* target protein labels, we constructed *p* supervised classifiers that predict the interaction of a query compound with the *j*-th target protein ($j = 1,2,\ldots,p$), where *p* is the number of target proteins with at least one known ligand. Hyperparameters of the predictive models were optimized for each target protein using a grid search based on cross-validation. KCF-S[116] was used for the chemical structure profile $\mathbf{z}^{(chem)}$.

Prediction of drug side effects based on proteome-wide binding affinity score profiles

A supervised classification machine learning model was used to predict side effects. The method is similar to that presented in previous studies.[2,5] The classifier was constructed using a linear model, where each drug is represented using a feature vector as $\mathbf{x} \in \mathbb{R}^d$, where d is the number of all proteins. The prediction score calculated by the linear function $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}^{(PBAS)}$ is then used to predict the presence or absence of each side effect. The weight vector $\mathbf{w} \in \mathbb{R}^d$ is defined during the learning process; therefore, it correctly predicts the side effect of the drug in the training data. Because each element of the feature vector $\mathbf{x}^{(PBAS)}$ corresponds to an element of its weight vector $\mathbf{w}$, it is possible to extract important features that contribute to the correct prediction by referring to the weight value $\mathbf{w}$ that corresponds to an element of $\mathbf{x}^{(PBAS)}$. Here, the predictive model was constructed using the logistical regression algorithm.

$$f\left(\mathbf{x}^{(PBAS)}\right) = \frac{1}{1+\exp(\mathbf{w}^T\mathbf{x}^{(PBAS)})}$$

Furthermore, by replacing the feature vector $\mathbf{x}^{(PBAS)}$ of the classifier by each of the other profiles (Fingerprint, TESS, and TELR), the side effect prediction model of each of the other profiles was constructed. All profile scores were binarized with thresholds. We assumed that each drug would bind to the top 10% of proteins based on binding affinities, thus setting threshold values vary according to each drug's binding affinity.

When training a model, if the number of proteins is very large compared with that of drugs in the training data, overfitting may occur. Here, L1 regularization was used to avoid it. Learning weights using L1 regularization is expected to facilitate model interpretation, because most elements of the weight vector are set to zero.

Given a training dataset $\{\mathbf{x}_i^{(PBAS)}, y_i\}_{i=1}^n, y_i \in \{+1, -1\}$ composed of n drugs and labels, the value of element y_i is 1 if the i-th drug has the side effect and 0 if it does not. The weights of the model are estimated using the following equation with the L1 regularization term (first term) and the loss function (second term):

$$\min_{\mathbf{w}} \|\mathbf{w}\|_1 + C\sum_{i=1}^{n} \log\left(1 + \exp\left(-y_i\mathbf{w}^T\mathbf{x}_i^{(PBAS)}\right)\right)$$

where $\|\mathbf{w}\|_1$ is the L1 norm, which means the sum of the absolute values of the vector $\mathbf{w}$ ($\|\mathbf{w}\|_1 = \sum_j^d |w_j|$), and C is a hyperparameter to control overfitting. To solve the problem of imbalance between positive- and negative-labeled drugs in the training data, the error term for the former was emphasized over that for the latter.

Performance evaluations of side effect prediction models

For the performance evaluation using SIDER, the gold standard set of drugs was divided into five subsets of approximately equal size, and five accuracy validation experiments were conducted. Each subset was in turn used as the test set, and a prediction model was trained on the remaining four subsets. This model was used on the test drug, and the accuracy of the test set was evaluated. Cross-validation experiments were performed for each model. The same training and test drugs were used in all experiments; therefore, the same conditions for cross-validation experiments were fulfilled.

Performance was also evaluated using FAERS, which was independent of the training data. Prediction models were trained on the SIDER dataset, and the models were adjusted to the drugs in FAERS. Prediction accuracy was evaluated based on the concordance of prediction scores with the correct labels in FAERS.

The prediction performance of the model was assessed using ROC and PR curves. The ROC curve is a plot of true-positive against false-positive ratios, and the AUPR is 1 and 0.5 for perfect inference and random prediction, respectively. The PR curve is a plot of precision (goodness of fit) against recall (recall). The ratio of the number of positive example samples to that of samples was considered the evaluation score. The hyperparameters for each method were optimized using a grid search applying the AUC and AUPR scores as objective functions.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical analyses were performed with Python. Statistical significance was assessed using unpaired, two-tailed Student's t-tests. p values are indicated in figure legends and source data. $p < 0.05$ is indicated with single asterisks, $p < 0.001$ with double asterisks, and $p < 0.0001$ with triple asterisks.