Journal of Cheminformatics

RESEARCH Open Access

Evaluation of deep learning architectures for predicting ligand interactions with neurologically relevant GPCRs

Souvik Dey[1,2], Pinyi Lu[1,2], Anders Wallqvist[1] and Mohamed Diwan M. AbdulHameed[1,2*]

Abstract

G protein-coupled receptors (GPCRs) are therapeutic targets for over 30% of approved drugs, yet specific GPCR subtypes act as molecular initiating events in several neurotoxic adverse outcome pathways. Therefore, knowledge of GPCR-ligand interactions is critical for drug discovery and computational toxicology. However, accurate predictions of GPCR-ligand binding can be challenging due to receptor conformational flexibility, complex membrane environment, and lack of selectivity among ligands. Drug-target interaction (DTI) models that jointly encode protein and ligand representations offer a promising approach to predict these interactions. In this study, we identified 119 neurologically relevant GPCRs and evaluated three deep learning architectures for creating a unified GPCR-ligand DTI model: dual-projection cosine similarity networks and transformer encoders (both using pre-computed embeddings) as well as bidirectional cross-attention networks (with frozen or fine-tuned encoders). Unlike prior studies that rely primarily on random splits, we evaluated model performance across random-split, cluster-split, and novel protein scenarios to provide realistic estimates of generalization. All the models performed well, achieving area under the receiver operating characteristic curve (AUROC) values of 0.91 on random-split validation, 0.81 on cluster-split testing (structurally distinct ligands), and 0.64 on novel GPCR generalization (proteins unseen during training). However, in the unseen-proteins test set, predictions were less accurate for GPCRs from protein families not represented in the training data or those with contrasting ligand interactions. We also assessed the models' ability to detect ligand selectivity, achieving an AUROC of 0.71 on ligands with at least five known GPCR interactions, although sensitivity remained low at 0.49. Our curated neurological GPCR dataset and rigorous evaluation framework provide realistic model assessment, reveal where current models succeed or fail, and provide practical guidance for deploying GPCR-ligand predictors in drug discovery and toxicity screening.

Scientific contribution

Our work connects computational toxicology and GPCR pharmacology by systematically evaluating transformers and similarity-based proteochemometric architectures for predicting CNS-relevant GPCR-ligand interactions under rigorous cold-start conditions. Our analysis reveals that while models achieve strong performance on structurally novel ligands, generalization to novel GPCRs remains challenging, with particularly poor prediction for out-of-family receptors. These results provide a rigorous evaluation framework for predicting ligand interactions against neurologically relevant GPCRs and highlight the need for improved protein representation learning in computational toxicology frameworks.

*Correspondence:
Mohamed Diwan M. AbdulHameed
mabdulhameed@bhsai.org
Full list of author information is available at the end of the article

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

Keywords GPCR-ligand binding, Neurotoxicity, Deep learning, Transformers, DTI, Cold-start evaluation, Drug safety assessment, Proteochemometrics

Introduction

Drug discovery remains a long and costly process, with safety-related failures representing a top cause for drug attrition [1, 2]. Central nervous system toxicity alone comprises nearly 25% of these cases [3]. The financial implications of these late-stage failures highlight the critical need for earlier identification of neurotoxicity risks, particularly during the lead optimization stage. Current neurotoxicity assessment tools rely heavily on traditional animal testing despite ethical concerns, high costs, and limited throughput [4]. However, to counteract these challenges, several new approach methodologies, such as organoids, organ-on-a-chip systems, and in silico prediction models, have been developed to enable rapid screening of large chemical libraries [5–7].

G protein-coupled receptors (GPCRs) are therapeutic targets for over 30% of U.S. Food and Drug Administration (FDA)-approved drugs [8]. However, off-target GPCR binding, caused by high structural similarity, particularly in the ligand-binding transmembrane domains, has been implicated in multiple neurotoxic adverse drug reactions with severe consequences [9, 10]. For example, metoclopramide, an antiemetic commonly prescribed for gastroparesis, exerts its therapeutic effect in part through dopamine D2 receptor antagonism. However, chronic exposure to central D2 blockade is associated with tardive dyskinesia, a serious neurological movement disorder that led to the addition of an FDA Black Box Warning in 2009 [11, 12]. Similarly, oxybutynin, an antimuscarinic agent developed for overactive bladder with therapeutic activity primarily attributed to M3 receptor antagonism, can cross the blood–brain barrier and exhibits broad muscarinic receptor binding across multiple subtypes (M1–M5). Central anticholinergic activity of oxybutynin has been associated with cognitive impairment with an increased risk of dementia in large epidemiological studies [13, 14]. These examples illustrate that understanding and predicting GPCR binding is critical for neurotoxicity risk assessment.

Computational models for predicting toxicity have evolved substantially over the past four decades, progressing from simple structural alerts to sophisticated machine learning methods [15]. While traditional quantitative structure–activity relationship (QSAR) models demonstrate significant utility for specific toxicity endpoints, they require complete retraining for each new protein and perform poorly when applied to structurally diverse chemicals [16, 17]. In addition, QSAR models are inherently single-target, requiring separate models for each protein of interest, which limits their utility for multi-target prediction. These limitations motivated the development of drug-target interaction (DTI) prediction models, which combine protein and ligand descriptors within a single framework [18, 19]. This joint encoding of a protein and a ligand within a single model is formally known as proteochemometrics (PCM), a framework that combines target and chemical descriptors to enable multi-target prediction across a protein family [19]. Modern PCM approaches replace handcrafted descriptors with learned embeddings from pretrained foundation models, but the underlying paradigm of encoding both partners jointly within one model remains unchanged. While early DTI methods utilized similarity patterns between drugs and targets, deep learning enabled DTI models to learn representations directly from sequences and molecular graphs [20–24]. The recent emergence of foundation models and transformer architectures has brought a series of innovations to DTI models [25]. Protein and molecular language models, trained on large databases, have demonstrated remarkable utility in capturing general structural and functional properties [26–29]. Current state-of-the-art DTI approaches integrate these pretrained embeddings with advanced model architectures, combining the molecular understanding from foundation models with the multi-modal learning capabilities of transformer attention mechanisms [30–34].

Despite these computational advances, a critical gap still exists for predicting ligand interactions with neurologically relevant GPCRs. Current neurotoxicity prediction platforms have integrated deep learning with large, curated databases but provide minimal GPCR coverage. For example, ChemNTP screened 199 neurotoxicity targets but included only one GPCR, NTxPred2 focused on predicting neurotoxic peptides and proteins, and NeuTox 2.0 was trained on hierarchical toxicity datasets without target-specific focus [35–37]. Conversely, GPCR-specific models have advanced from simple machine learning ensemble classifiers to physics-based free-energy calculations and sophisticated deep learning architectures [38–40]. Models have also been developed for distinguishing allosteric versus orthosteric binding sites and evaluating whether focusing on transmembrane helices instead of the entire sequence enhances binding prediction [41, 42].

More recently, Brahma et al. developed AiGPro, which predicts agonist and antagonist bioactivity across 231 GPCRs [43]. However, these frameworks focus mostly on either highlighting model sophistication or drug discovery applications rather than toxicity endpoints. Additionally, most GPCR-ligand prediction studies use random data splits for validation, which often contain overlapping ligands and proteins between training and test sets, leading to overly optimistic performance estimates. More rigorous evaluation frameworks that distinguish between scenarios with novel drugs, novel proteins, or both demonstrate that model performance degrades substantially when similarity to training data decreases [20]. This combination of limited GPCR representation in neurotoxicity models, lack of focus on neurologically relevant GPCRs in existing DTI models, and inadequate validation practices motivates the development of a specialized framework for predicting ligand interactions with neurologically relevant GPCRs using rigorous out-of-distribution evaluation.

In this work, we systematically developed and evaluated deep learning architectures for predicting GPCR-ligand binding interactions focusing on neurologically relevant GPCRs. While general DTI methods are trained on diverse proteins to broaden their applicability, this may not be relevant for toxicity screening, where we often have a fixed set of toxicity-associated targets. To this end, we curated a dataset of 152,831 GPCR-ligand interactions spanning 119 neurologically relevant GPCRs. We implemented multiple DTI approaches ranging from simple cosine similarity of concatenated embeddings to sophisticated cross-attention mechanisms with fine-tuned GPCR and ligand embeddings. We created two different test sets to assess model extrapolation to structurally novel ligands and to novel GPCRs from underrepresented families, corresponding to the cold drug (S2) and cold protein (S3) scenarios from an earlier study [20]. We conducted multiple independent runs with statistical significance testing to understand current model capabilities. Our findings demonstrate that the ligand-centric learning pattern observed earlier for kinases extends to GPCRs [44], that fine-tuning large pretrained protein encoders can sometimes be detrimental, and that models achieve useful but incomplete generalization, with performance strongly dependent on the similarity of test queries to training data. These findings have important implications for deploying DTI models in practical GPCR-ligand screening workflows for toxicity assessment and highlight clear directions for developing next-generation methods.

Materials and methods

Data extraction and preprocessing

We accessed the Gene Ontology (GO) database to identify genes involved in neurological processes using the following biological processes: (1) axon guidance, (2) neuroinflammatory response, (3) nervous system process, (4) nervous system development, (5) neuron apoptotic process, (6) regulation of neuron apoptotic process, (7) chemical synaptic transmission, (8) neurotransmitter transport, (9) learning or memory, and (10) presynaptic endocytosis. We also included generic toxicity-relevant GO biological processes response to toxic substance and response to oxidative stress. We further filtered these gene symbols using GO cellular component terms to retain the genes localized only to neurons, axons, or dendrites. We then mapped these filtered genes to their corresponding UniProt IDs to identify 3,652 unique protein targets, 639 of which were GPCRs. Furthermore, we examined ToxCast assay gene targets and Organisation for Economic Co-operation and Development-approved adverse outcome pathways for neurotoxicity, but we did not discover additional targets [45, 46]. Figure 1a summarizes our workflow.

We queried the ChEMBL and BindingDB databases and obtained GPCR-ligand activity data (IC_{50}, K_i, K_d, and EC_{50}) for 153 GPCRs; the remaining 486 GPCRs did not have any associated ligands. We removed all entries without absolute activity values (e.g., those containing < or > symbols) or with activity units other than nanomolar. We processed ligand SMILES using the ChEMBL structure pipeline, which performed validation, desalting and standardization, followed by stereochemistry removal [47]. Given the heterogeneous nature of the data sources, we opted for a classification model and used a 0.1 μM threshold to distinguish active from inactive interactions, consistent with the Pharos database recommendation for GPCRs [48]. We removed duplicate entries when a GPCR-ligand pair belonged to the same class and removed all entries when the pair belonged to opposite classes. Our final dataset consisted of 152,831 GPCR-ligand pairs, representing 119 unique GPCRs and 97,955 unique ligands, with 68,030 active and 84,801 inactive interactions (Fig. 1b).

We split the dataset into training, validation, and test sets using a 70:10:20 ratio (Fig. 1c). We generated training and validation sets via random splitting, allowing both GPCR and ligand overlap between these sets. We created two distinct test sets of similar sizes to evaluate model performance under different generalization conditions. For the unseen-proteins test set (S3, cold target scenario), we randomly selected nine GPCRs (Supplementary Table S1) and removed all of their associated GPCR-ligand pairs from the training and validation sets. For

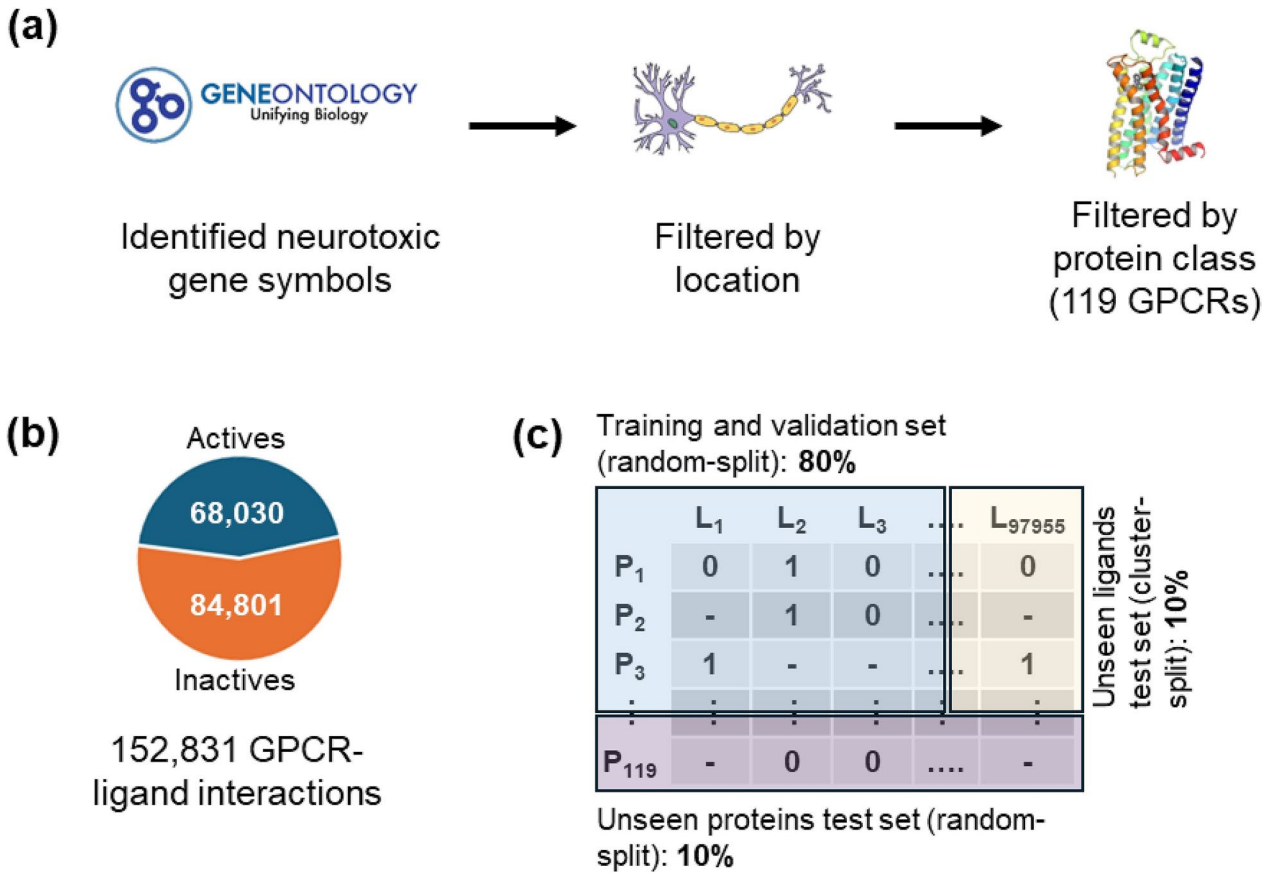


Fig. 1 Data curation procedure for creating the neurotoxic G protein-coupled receptor (GPCR)-ligand interaction profiler. **a** Method for selecting the 119 neurotoxic GPCRs using the biological processes and cellular components from the Gene Ontology database. **b** Distribution of active and inactive interactions in the GPCR-ligand dataset. **c** Illustration describing the selection of the training, validation, and two test sets

the unseen-ligands test set (S2, cold drug scenario), we performed Butina clustering on all 97,955 ligands using Tanimoto similarity (calculated using 2,048-bit Morgan fingerprints) and assigned ligands to clusters using a 0.65 similarity threshold [49]. We then ensured that ligands in this test set came from clusters that had no representatives in the training set, ensuring structural dissimilarity.

GPCR and ligand representation

We represented GPCRs using their primary amino acid sequences and encoded them using ProtBERT, a pretrained protein language model (PLM) [27]. ProtBERT is based on the BERT (Bidirectional Encoder Representations from Transformers) architecture and was trained on 217 million protein sequences from UniRef100 database using masked language modeling [50]. This pretraining enabled the model to learn contextualized amino acid representations capturing both local sequence patterns and higher-order structural and functional properties. We tokenized protein sequences with spaces between amino acids for compatibility and extracted sequence-level embeddings from ProtBERT. We also evaluated ESM2 (facebook/esm2-t30-150 M-UR50D), another state-of-the-art PLM trained on 65 million protein sequences using masked language modeling, but this model did not yield significant performance improvements over ProtBERT and is therefore not presented [26].

We represented ligands as SMILES strings and encoded them using MolFormer, a chemical language model pretrained on 1.1 billion molecules from ZINC and PubChem databases [28]. Like ProtBERT, MolFormer also employs a transformer-based architecture trained via masked atom prediction, enabling it to learn molecular representations that capture both physicochemical properties and structural features. We also evaluated 2,048-bit Morgan fingerprints as an alternative molecular representation, but this model performed comparably to MolFormer and is not shown [51].

Model training

We built all the models using PyTorch and Python 3.11 on a single NVIDIA L40S GPU (48 GB VRAM) with CUDA

13.1. We developed and evaluated three architectures to predict GPCR-ligand binding: (1) dual-projection cosine similarity networks (CosSim), (2) transformer encoders applied to concatenated GPCR-ligand embeddings (transformer), and (3) bidirectional cross-attention networks with either frozen or fine-tuned GPCR and ligand encoders (CA-Base/CA-Lig/CA-Prot/CA-Full). Figure 2 shows a schematic overview of the architectures. We used XGBoost with concatenated GPCR-ligand embeddings as a baseline classifier. We trained each model five times with different random seeds to assess reproducibility, and we reported the mean and standard deviation for the performance metrics.

We adapted the CosSim approach (Fig. 2, left panel) from an earlier study [52]. We applied mean pooling to the pre-computed GPCR and ligand embeddings obtained from ProtBERT and MolFormer (dimensions of 1,024 and 768, respectively) and passed them through separate single-layer feed-forward neural networks with ReLU activation. These layers projected the embeddings into a shared latent space where we calculated cosine similarity to predict binding probability $p(G^*,L^*)$:

$$p\left(G^*, L^*\right) = \left(\frac{G^* \cdot L^*}{\|G^*\| \cdot \|L^*\|}\right)$$

where G^* and L^* represent the projected GPCR and ligand embeddings, respectively. We trained the model for 50 epochs using binary cross-entropy loss and AdamW optimizer with a learning rate of 10^{-4} and batch size of 32.

The transformer model (Fig. 2, middle panel) followed a similar preprocessing pipeline as CosSim but concatenated the mean-pooled GPCR and ligand embeddings instead of projecting them separately [53]. We passed the concatenated embeddings through two transformer encoder layers, each with four attention heads and 10% dropout [54]. We passed the output to a feed-forward classifier with one hidden layer, ReLU activation, 10%

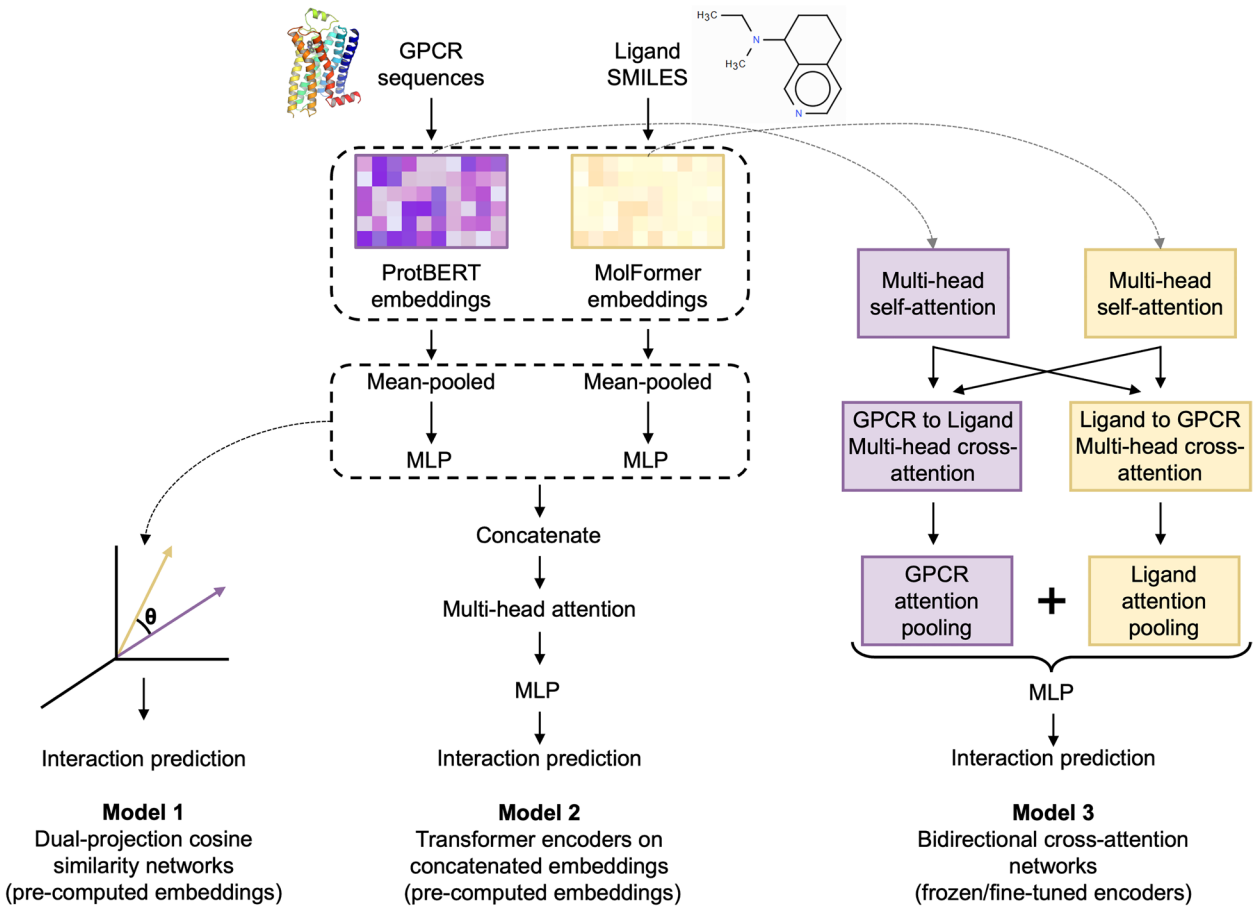


Fig. 2 Architectures for our different models. We started with ligand SMILES and G protein-coupled receptor (GPCR) sequences that were tokenized using MolFormer and ProtBERT, respectively, and used the following to create the three models: (left panel) dual-projection cosine similarity networks for pre-computed mean-pooled embeddings, (middle panel) transformer encoders on concatenated pre-computed mean-pooled embeddings, and (right panel) bidirectional cross-attention networks on either frozen or fine-tuned encoders

dropout, and a sigmoid activation function at the output layer. We trained this model for 50 epochs using binary cross-entropy loss and AdamW optimizer with a learning rate of 10^{-4} and batch size of 32.

The cross-attention models (Fig. 2, right panel) extended the transformer approach by introducing bidirectional cross-attention between GPCR and ligand representations [55]. Unlike the previous architectures, we computed embeddings on the fly rather than using pre-computed representations, enabling gradient-based fine-tuning of the pretrained encoders during backpropagation. We explored four variants of this model: CA-Base (frozen ProtBERT and MolFormer encoders), CA-Lig (fine-tuned MolFormer with frozen ProtBERT), CA-Prot (fine-tuned ProtBERT with frozen MolFormer), and CA-Full (both encoders fine-tuned). We used the padded sequence outputs from ProtBERT and MolFormer and passed them through separate single-layer feed-forward neural networks with ReLU activation. We passed the projected embeddings through individual transformer encoder layers with four attention heads and 10% dropout to compute self-attention for GPCRs and ligands independently. We applied bidirectional cross-attention using multi-head attention mechanisms with 10% dropout, where GPCR representations served as queries to address ligand representations, and vice versa. We performed mean pooling on the cross-attention outputs, concatenated the pooled GPCR and ligand representations, and passed them through a feed-forward classifier with two hidden layers, ReLU activation, 10% dropout, and a sigmoid output layer. We trained this model for 10 epochs (reduced due to computational cost) using binary cross-entropy loss and AdamW optimizer with a learning rate of 10^{-4}, weight decay of 10^{-2}, and batch size of 8. We employed a ReduceLROnPlateau scheduler that reduced the learning rate by a factor of 0.5 when validation area under the receiver operating characteristic curve (AUROC) plateaued for three consecutive epochs.

We also implemented an XGBoost classifier using the concatenated projected embeddings (protein + ligand) as features [56]. We applied the same projection layers as in the CosSim model, concatenated them, and used these features to train an XGBoost classifier. To select optimal hyperparameters, we used Optuna's Tree-Structured Parzen Estimator for 50 iterations, optimizing validation AUROC [57, 58]. The tuned hyperparameters were max_depth = 10, learning_rate = 0.19, n_estimators = 335, subsample = 0.80, colsample_bytree = 0.59, min_child_weight = 5, gamma = 0.29, $\alpha = 3.13$, and $\lambda = 0.004$, consistent with ranges recommended by Sheridan et al. [59].

We trained models for the full number of epochs without early stopping but saved checkpoints after each epoch based on validation AUROC, retaining only the best performing checkpoint.

Model evaluation

We evaluated all the models on the validation set and two test sets: an unseen-ligands test set created via Butina clustering to ensure structural dissimilarity from training compounds and an unseen-proteins test set containing nine randomly selected GPCRs. We computed accuracy, sensitivity (recall), specificity, Matthews correlation coefficient (MCC), and AUROC for all the evaluations. To check whether one method performed significantly better than another, we calculated p-values between methods using the pairwise Conover-Friedman post-hoc test using scikit-posthocs package (referred to as the Conover-Friedman test henceforth).

Representation-level analysis

To investigate why fine-tuning the protein encoder degraded performance, we compared the fine-tuned ProtBERT encoders from the CA-Prot and CA-Full checkpoints against the pretrained ProtBERT baseline using four complementary analyses. For each analysis, we extracted mean-pooled, residue-level embeddings for all unique GPCR sequences and analyzed the five seeds per variant. First, for each receptor i we computed the cosine distance between its pretrained embedding e_i^{pre} and its fine-tuned embedding e_i^{ft}

$$d_i = 1 - \frac{e_i^{pre} * e_i^{ft}}{\left\| e_i^{pre} \right\| \left\| e_i^{ft} \right\|}$$

Second, we projected the stacked pretrained and fine-tuned embeddings of the training receptors into two dimensions using a single joint Uniform Manifold Approximation and Projection (UMAP) fit (n_neighbors = 15, min_dist = 0.1). Third, to detect representation collapse, we computed the effective rank that measures the number of dimensions the embeddings effectively span. If $s_1, s_2, \ldots, s_i$ are the singular values of the mean-centered embedding matrix and the fraction of total spread along the direction i, $p_i = \frac{s_i}{\sum_j s_j}$, then effective rank (ER) is the exponential of the Shannon entropy of that distribution, as follows:

$$\mathrm{ER} = \exp\left(-\sum_{k=1}^{r} p_i \log p_i \right)$$

Fourth, for each parameter tensor W in the 30 ProtBERT encoder layers, we computed the relative Frobenius drift (δ_F), as follows:

$$\delta_F(W) = \frac{\left\| W^{ft} - W^{pre} \right\|_F}{\| W^{pre} \|_F}$$

which normalizes for tensor size so that layers and sub-blocks are directly comparable. We averaged the drift within each layer and across the five seeds.

Calibration analysis

We assessed the calibration of the predicted binding probabilities on the ligand-selectivity dataset. We partitioned N predictions into 10 equal-width probability bins and computed the expected calibration error (ECE), as follows:

$$\text{ECE} = \sum_{b=1}^{10} \frac{n_b}{N} \left| act\,(b) - prob\,(b) \right|$$

where n_b represents the number of predictions, $act(b)$ denotes the observed active rate, and $prob(b)$ represents the mean predicted probability within each bin b. We also computed the Brier score for predictions p_i and binary labels y_i, as follows:

$$\text{Brier score} = \frac{1}{N} \sum_{i=1}^{N} \left(p_i - y_i \right)^2$$

and compared it against the no-skill baseline, $\bar{y}(1-\bar{y})$, where $\bar{y}$ represents the active rate.

Per-ligand selectivity analysis

To evaluate each model's ability to rank a single ligand's preferred GPCRs, we performed per-ligand analyses on the ligand-selectivity subset. For each ligand in this subset, we computed the AUROC and the area under the precision-recall curve (AUPRC) over its GPCR targets. Ligands whose GPCR targets shared a single label (actives or inactives) were excluded, as their AUROC was undefined. We also computed the top-k sensitivity, the fraction of a ligand's true active receptors that appear among its k highest-ranked predictions, for k = 1 and 3 and the probability difference between active and inactive receptors.

Phylogenetic tree

We used the data mapper tool from GPCRdb to create a phylogenetic tree for all the GPCRs in our dataset [60].

Results

Data exploration and model training

Our final curated dataset consisted of 152,831 GPCR-ligand pairs representing interactions between 119 neurologically relevant GPCRs and 97,955 unique ligands, with 68,030 active (45%) and 84,801 inactive (55%) interactions. The 119 GPCRs represented diverse families: Class A (rhodopsin-like, $n = 80$), Class B (secretin-like, $n = 3$), Class C (glutamate-like, $n = 7$), Class F (frizzled, $n = 4$), and Class T2 (taste, $n = 25$). The dataset exhibited sparse coverage, with the maximum number of associated ligands for a GPCR being 8,594 (median: 574), whereas the maximum number of associated GPCRs for a single ligand was 31.

We split the dataset into training (108,419 pairs), validation (15,489 pairs), and test sets (28,923 pairs). The unseen-proteins test set contained 14,419 GPCR-ligand pairs involving nine GPCRs with no representatives in the training set (Supplementary Table S1), while the unseen-ligands test set contained 14,504 pairs with ligands from clusters lacking any representatives in the training set. To quantify structural dissimilarity across evaluation scenarios, we calculated the Tanimoto similarity (using 2,048-bit Morgan fingerprints) of ligands in the validation and unseen-ligands test set against their five nearest neighbors in the training set. Supplementary Figure S1 shows the distribution and reveals the stark difference between the two sets, with mean Tanimoto similarities of 0.74 and 0.48 for the validation set and unseen-ligands test set, respectively. Previous models developed for general DTI prediction have focused mostly on random-split test sets, which are much easier to predict, and we wanted to create more challenging datasets to assess how the models adapt to out-of-distribution chemical structures. We designed the unseen-proteins test set to assess whether our model can capture novel GPCR-ligand interactions not encountered in the training set.

To further characterize the composition of our dataset and the effect of the cluster-based split on label balance, we classified each GPCR from the training and unseen-ligands test sets into their subfamilies (Aminergic, Peptide, Lipid, Nucleotide, Class A [others], Class B, Class C, Class F, or Class T2). ChEMBL and BindingDB are known to over-represent aminergic GPCRs, and our dataset inherited that bias, with aminergic receptors accounting for 50.6% of the training pairs across 27 receptors (Supplementary Table S2). We found that the cluster-based split shifted this distribution slightly, with the test set containing 34.1% aminergic pairs across 28 receptors. Cluster split also asymmetrically affected label balance, with aminergic %active dropping modestly from 46.4% in the training set to 39.2% in the test set, while non-aminergic %active rose from 42.1 to 52.9%. We also note that 17 bitter taste receptors (TAS2R family) had extremely sparse data, and their entire set of pairs (1 pair each) was assigned to the unseen-ligands test set. These GPCRs contribute negligibly to the overall test metrics but inflate the apparent Class T2 active rate.

We trained seven models (CosSim, Transformer, XGBoost, CA-Base, CA-Lig, CA-Prot, and CA-Full) across five independent runs with different random seeds. Training curves demonstrated stable convergence for all architectures, with validation AUROC typically plateauing within 10–15 epochs for CosSim and transformer models and within 7–8 epochs for cross-attention variants (Supplementary Figure S2). The cross-attention models required a substantially longer training time (approximately 2 h per epoch on Nvidia L40S GPU for CA-Full) compared to the CosSim and transformer models (approximately 0.01 h per epoch) due to on-the-fly embedding computation and fine-tuning of the ProtBERT and MolFormer encoders.

Overall model performance

Tables 1, 2, 3 show a comprehensive performance comparison across all models and evaluation sets. On the

Table 1 Performance metrics for the seven models on the validation set

Model	Accuracy	Sensitivity	Specificity	MCC	AUROC
CosSim	0.807 ± 0.004	0.753 ± 0.013	0.849 ± 0.008	0.607 ± 0.007	0.884 ± 0.001
Transformer	0.809 ± 0.002	0.798 ± 0.011	0.819 ± 0.010	0.615 ± 0.003	0.881 ± 0.001
CA-Base	0.813 ± 0.004	0.795 ± 0.029	0.827 ± 0.028	0.622 ± 0.006	0.891 ± 0.002
CA-Lig	0.831 ± 0.002	0.803 ± 0.026	0.852 ± 0.017	0.656 ± 0.005	0.907 ± 0.001
CA-Prot	0.747 ± 0.005	0.711 ± 0.020	0.776 ± 0.014	0.487 ± 0.009	0.774 ± 0.018
CA-Full	0.772 ± 0.003	0.715 ± 0.015	0.818 ± 0.015	0.537 ± 0.006	0.786 ± 0.023
XGB	0.824 ± 0.002	0.786 ± 0.004	0.855 ± 0.002	0.643 ± 0.005	0.904 ± 0.000

MCC Matthews correlation coefficient, *AUROC* area under the receiver operating characteristic curve, *CosSim* dual-projection cosine similarity networks, *CA-Base* bidirectional cross-attention with frozen protein and ligand encoders, *CA-Lig* fine-tuned ligand encoders, *CA-Prot* fine-tuned protein encoders, *CA-Full* fine-tuned ligand and protein encoders, *XGB* XGBoost

Table 2 Performance metrics for the seven models on the test set (unseen ligands)

Model	Accuracy	Sensitivity	Specificity	MCC	AUROC
CosSim	0.715 ± 0.005	0.604 ± 0.015	0.819 ± 0.023	0.435 ± 0.013	0.786 ± 0.002
Transformer	0.706 ± 0.006	0.652 ± 0.030	0.756 ± 0.018	0.410 ± 0.012	0.763 ± 0.008
CA-Base	0.741 ± 0.010	0.679 ± 0.042	0.798 ± 0.030	0.483 ± 0.017	0.809 ± 0.007
CA-Lig	0.737 ± 0.009	0.656 ± 0.037	0.813 ± 0.026	0.476 ± 0.018	0.811 ± 0.007
CA-Prot	0.670 ± 0.003	0.576 ± 0.011	0.758 ± 0.011	0.340 ± 0.007	0.687 ± 0.013
CA-Full	0.693 ± 0.007	0.586 ± 0.028	0.792 ± 0.017	0.388 ± 0.014	0.700 ± 0.024
XGB	0.719 ± 0.001	0.587 ± 0.008	0.842 ± 0.007	0.445 ± 0.003	0.797 ± 0.002

MCC Matthews correlation coefficient, *AUROC* area under the receiver operating characteristic curve, *CosSim* dual-projection cosine similarity networks, *CA-Base* bidirectional cross-attention with frozen protein and ligand encoders, *CA-Lig* fine-tuned ligand encoders, *CA-Prot* fine-tuned protein encoders, *CA-Full* fine-tuned ligand and protein encoders, *XGB* XGBoost

Table 3 Performance metrics for the seven models on the test set (unseen proteins)

Model	Accuracy	Sensitivity	Specificity	MCC	AUROC
CosSim	0.582 ± 0.009	0.414 ± 0.058	0.709 ± 0.054	0.128 ± 0.015	0.586 ± 0.003
Transformer	0.603 ± 0.006	0.503 ± 0.015	0.678 ± 0.021	0.183 ± 0.010	0.595 ± 0.009
CA-Base	0.615 ± 0.007	0.420 ± 0.092	0.760 ± 0.071	0.194 ± 0.018	0.594 ± 0.029
CA-Lig	0.597 ± 0.009	0.404 ± 0.052	0.742 ± 0.042	0.155 ± 0.019	0.585 ± 0.011
CA-Prot	0.595 ± 0.006	0.479 ± 0.026	0.682 ± 0.017	0.164 ± 0.014	0.584 ± 0.033
CA-Full	0.603 ± 0.005	0.469 ± 0.034	0.703 ± 0.024	0.178 ± 0.012	0.585 ± 0.011
XGB	0.608 ± 0.008	0.318 ± 0.028	0.824 ± 0.017	0.166 ± 0.019	0.637 ± 0.011

MCC Matthews correlation coefficient, *AUROC* area under the receiver operating characteristic curve, *CosSim* dual-projection cosine similarity networks, *CA-Base* bidirectional cross-attention with frozen protein and ligand encoders, *CA-Lig* fine-tuned ligand encoders, *CA-Prot* fine-tuned protein encoders, *CA-Full* fine-tuned ligand and protein encoders, *XGB* XGBoost

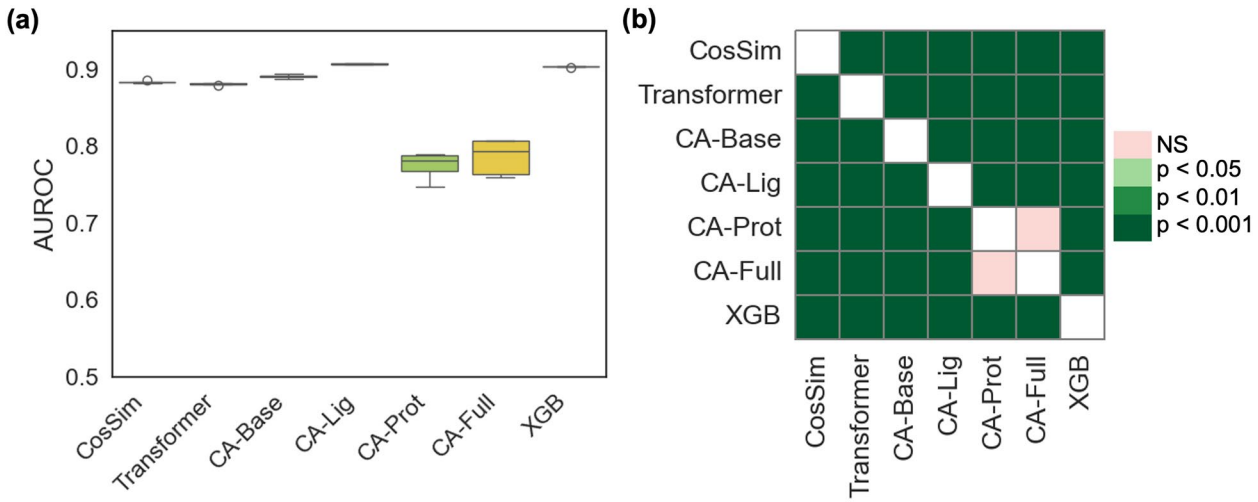


Fig. 3 Comparison of model performance on the validation set. **a** Boxplots comparing the area under the receiver operating characteristic curve (AUROC) for our seven models. **b** Sign plots showing the pairwise significance matrices obtained using a Conover-Friedman test. Cell color represents the corrected *p*-value for the difference between the two methods in the rows and columns. *CosSim* dual-projection cosine similarity networks, *CA-Base* bidirectional cross-attention with frozen protein and ligand encoders, *CA-Lig* fine-tuned ligand encoders, *CA-Prot* fine-tuned protein encoders, *CA-Full* fine-tuned ligand and protein encoders, *XGB* XGBoost, *NS* not significant

validation set, CA-Lig achieved the best performance with an AUROC of 0.907, followed closely by the simpler XGBoost baseline (AUROC: 0.904), (Fig. 3). The CosSim and transformer models achieved intermediate performance (AUROC: 0.884 and 0.881, respectively). However, CA-Prot and CA-Full displayed substantially degraded performance (AUROC: 0.774 and 0.786, respectively), with CA-Prot performing even worse than the XGBoost baseline, indicating issues with fine-tuning the ProtBERT encoders.

Performance trends differed across the two test sets. On the unseen-ligands test set, we saw a relatively consistent performance drop of approximately 0.10 AUROC units across all models compared to the validation set

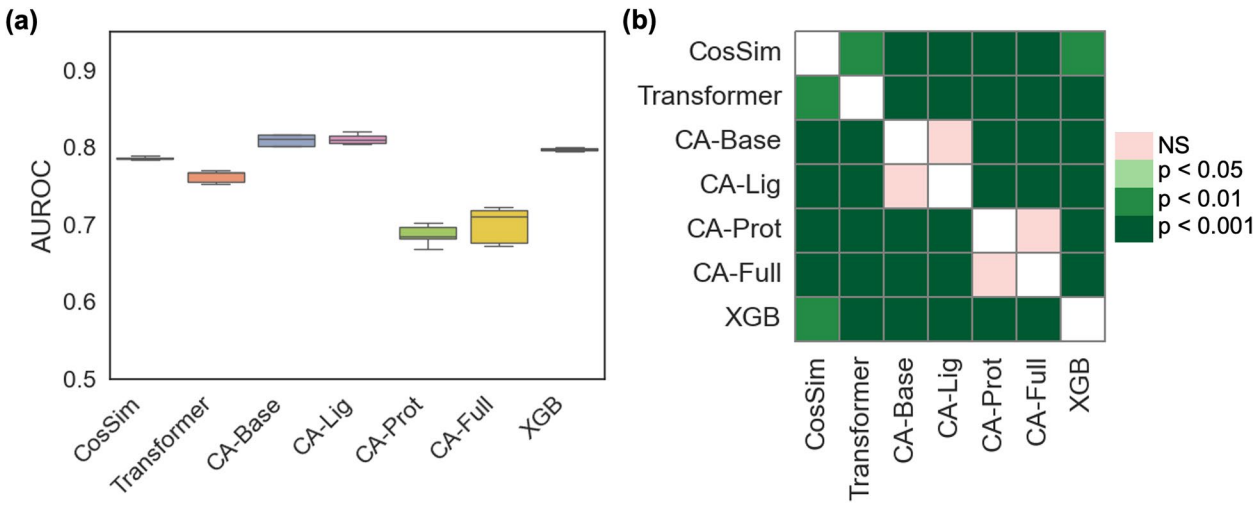


Fig. 4 Comparison of model performance on the test set (unseen ligands). **a** Boxplots comparing the area under the receiver operating characteristic curve (AUROC) for our seven models. **b** Sign plots showing the pairwise significance matrices obtained using a pairwise Conover-Friedman test. Cell color represents the corrected *p*-value for the difference between the two methods in the rows and columns. *CosSim* dual-projection cosine similarity networks, *CA-Base* bidirectional cross-attention with frozen protein and ligand encoders, *CA-Lig* fine-tuned ligand encoders, *CA-Prot* fine-tuned protein encoders, *CA-Full* fine-tuned ligand and protein encoders, *XGB* XGBoost, *NS* not significant

(Fig. 4), demonstrating reasonable generalization to novel chemical scaffolds despite their structural dissimilarity (mean Tanimoto similarity of 0.48 with the training set). CA-Lig was the most successful model here, achieving an AUROC of 0.811, followed by CA-Base (AUROC: 0.809) and XGB (AUROC: 0.797). Again, CA-Prot and CA-Full displayed poor performance (AUROC: 0.687 and 0.700, respectively).

In contrast, the unseen-proteins test set presented a substantially greater challenge, with AUROC drops ranging from 0.19 to 0.32 units compared to the validation set (Fig. 5), indicating that predicting binding to novel GPCRs was more difficult than extrapolating to novel ligand scaffolds. Interestingly, the performance hierarchy shifted for this scenario, with all models achieving similar AUROC values and large standard deviations across the five runs. Conover-Friedman test revealed that none of the five models was statistically significantly different from the others. The XGBoost baseline had the highest mean AUROC (0.637), followed closely by the transformer model and CA-Base with nearly identical AUROCs (0.595 and 0.594, respectively), while CA-Lig had a slightly lower AUROC (0.585). Additionally, the differences between the models became less pronounced (Supplementary Figure S3), with only 0.053 AUROC units separating the best (XGBoost) and worst (CA-Prot) performers, compared to spreads of 0.133 and 0.124 AUROC units on the validation and unseen-ligands test sets, respectively. This highlights that all the models struggled to predict novel GPCR-ligand interactions that were absent in the training set.

We also examined performance metrics beyond AUROC and observed consistent patterns across evaluation scenarios. We found that sensitivity consistently fell below specificity across all models and test sets (Tables 1, 2, 3), indicating greater difficulty in correctly identifying active interactions. We observed that this sensitivity–specificity gap became more pronounced on the unseen-proteins test set, where the highest sensitivity was only 0.420 for the CA-Base model, but specificity remained strong at 0.760. In contrast, for the same CA-Base model, sensitivity was 0.795 in the validation set, while specificity was 0.827. We also calculated MCC values and found that they strongly correlated with AUROC, confirming that AUROC appropriately captured overall performance despite the modest class imbalance.

Additionally, to directly test the relative contribution of protein and ligand information, we retrained all the models after shuffling the ProtBERT embeddings, MolFormer embeddings, and labels, one at a time. Permuting the MolFormer embeddings or the labels reduced AUROC to near-random levels for every model on the unseen-ligands test set, whereas permuting the ProtBERT embeddings produced only a modest decrease across all the models (Supplementary Table S3), as also seen in an earlier study [61]. These results provide additional support for ligand-centric learning, where the models retain most of their predictive power without correct protein

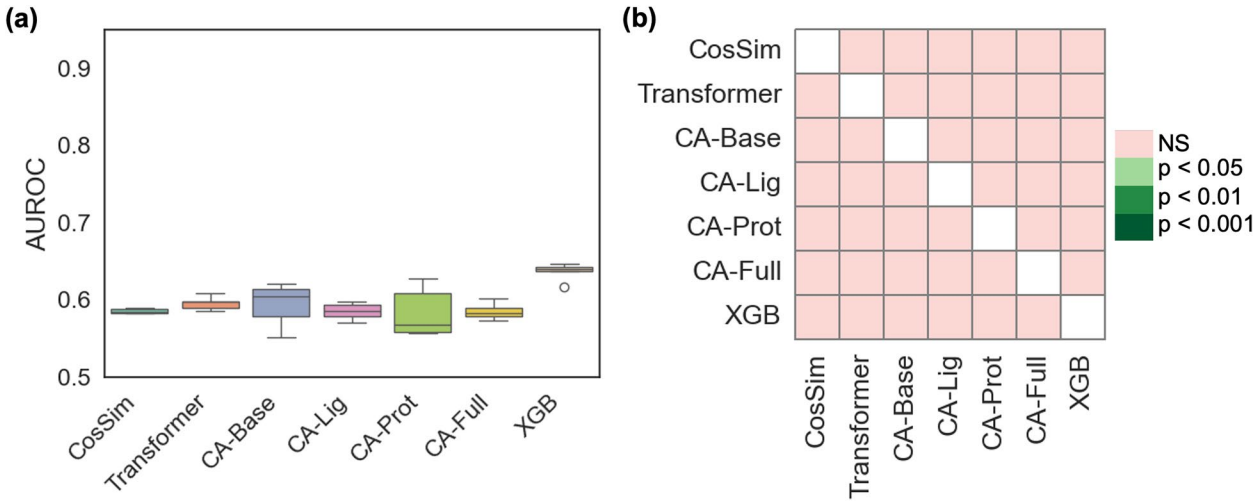


Fig. 5 Comparison of model performance on the test set (unseen proteins). **a** Boxplots comparing the area under the receiver operating characteristic curve (AUROC) for our seven models. **b** Sign plots showing the pairwise significance matrices obtained using a Conover-Friedman test. Cell color represents the corrected *p*-value for the difference between the two methods in the rows and columns. *CosSim* dual-projection cosine similarity networks, *CA-Base* bidirectional cross-attention with frozen protein and ligand encoders, *CA-Lig* fine-tuned ligand encoders, *CA-Prot* fine-tuned protein encoders, *CA-Full* fine-tuned ligand and protein encoders, *XGB* XGBoost, *NS* not significant

information but lose nearly all of it without correct ligand information.

To further test whether the degradation in CA-Prot was specific to ProtBERT, we repeated the CA-Base and CA-Prot experiments with ESM2 (facebook/esm2-t30-150M-UR50D) as the protein encoder. In the CA-Base variant, ESM2 did not improve performance over ProtBERT on the unseen-ligands test set (AUROC of 0.728 and 0.809 with ESM2 and ProtBERT, respectively). More importantly, fine-tuning the ESM2 protein encoder again reduced performance relative to its frozen counterpart (CA-Prot-ESM2 AUROC 0.708 vs. 0.728), reproducing the direction of the effect observed with ProtBERT, although with a smaller magnitude (Supplementary Table S4). This suggests that the performance drop from fine-tuning the protein encoder is not specific to ProtBERT but reflects a more general behavior of protein language models fine-tuned on a focused GPCR dataset of this size.

Ligand selectivity

Structural similarity among GPCRs, particularly in their seven transmembrane domain regions, often leads to off-target binding among ligands, causing interactions with unintended receptors. Understanding and predicting this selectivity is critical for both drug discovery and toxicity assessments, as off-target activity can cause adverse effects. To assess our models' ability to predict such patterns, we selected a subset of the unseen-ligands test set consisting of ligands with at least five interactions (either active or inactive) across GPCRs. This focused dataset consisted of 717 GPCR-ligand interactions between 62 GPCRs and 100 unique ligands, with 169 active and 548 inactive interactions (24% active).

We evaluated all seven models on this selectivity-focused dataset and found CA-Base to be the best performing model with an AUROC of 0.714, followed closely by CA-Lig with an AUROC of 0.698 (Supplementary Figure S4 and Supplementary Table S5), with no significant difference between them. The performance hierarchy among the models reflected the pattern seen for the larger unseen-ligands test set, with CA-Prot again occupying the lowest position (AUROC: 0.553). For the best performing CA-Base model, we observed a sensitivity of 0.494, but specificity reached 0.888, consistent with the pattern of greater difficulty in predicting active interactions.

We observed that all the models showed reduced performance on this selectivity dataset compared to the full unseen-ligands test set, with the most extensive decrease occurring in sensitivity (from 0.656 to 0.506 for CA-Lig). This suggests that predicting ligand selectivity among structurally similar GPCRs presents an additional challenge beyond predicting interactions with structurally dissimilar ligands. We attribute this difficulty to the dual challenge faced by models on this dataset: not only must they extrapolate to novel chemical scaffolds (as ligands were selected from unique clusters with a mean Tanimoto similarity of 0.48 to the training set), but they must also discriminate between closely related GPCRs that share high sequence identity (often > 70% in transmembrane regions) yet exhibit distinct ligand selectivity profiles. This difficulty likely arises from subtle differences in binding pocket residues that determine selectivity, which may not be fully captured by our sequence-based protein representations.

To better assess the reliability of model predictions, we further evaluated model calibration and per-ligand selectivity on this dataset. Calibration analysis showed that all the models tended to be over-confident in the mid-to-high probability range (Supplementary Figure S5). XGBoost was the best-calibrated model (ECE = 0.06), followed by CosSim (ECE = 0.08) and CA-Base (ECE = 0.11), while the transformer model was the least calibrated (ECE = 0.20). The Brier scores were below the no-skill baseline of 0.18 for CosSim, CA-Base, CA-Lig, and XGBoost (Supplementary Table S6). Per-ligand analyses within each ligand's GPCR list further revealed that CA-Base and CA-Lig substantially outperformed the other models on selectivity discrimination (per-ligand AUROC $\approx$ 0.86, AUPRC $\approx$ 0.85, top-1 sensitivity $\approx$ 0.48, active-inactive probability gap $\Delta \approx 0.43$), while CA-Prot and CA-Full showed near-random per-ligand discrimination (AUROC 0.59 and 0.67, respectively) with large inter-seed variability (Supplementary Table S7). These results confirm that protein encoder fine-tuning impaired not only overall binding prediction but also the ability to rank off-target GPCRs for individual ligands. Together, these two analyses show that our model set covers both aspects of ligand selectivity prediction. XGBoost and CosSim provide well-calibrated probability estimates for absolute risk assessment, while CA-Base and CA-Lig deliver excellent per-ligand rankings for identifying a compound's most likely off-target GPCRs.

Individual GPCR performance analysis

While the unseen-ligands test set evaluated model performance on novel chemical scaffolds, the unseen-proteins test set provided more insight into how models generalize to novel GPCRs and into the inner workings of the model predictions. Supplementary Figure S6 shows the diversity of the GPCRs in our dataset using a phylogenetic tree constructed from the entire sequence. The nine unseen proteins in our test set (Supplementary Table S1) represented different GPCR families. While some belonged to larger subfamilies with close homologs

in the training set (ACM4, OPRD, 5-HT1E, GRM3), others came from smaller subfamilies (P2RY1, OX1R, LSHR). Two GPCRs in this set (NTR1, NTR2) were the only members of their subfamily, with no related receptors in the training data.

We evaluated the performance of individual GPCRs using the CA-Base model, which achieved the second-highest overall AUROC on this test set, just behind the XGBoost baseline. Since some GPCRs had imbalanced datasets (with percent active interactions ranging from 5% in 5-HT1E to 76% in NTR2), we used MCC as the primary performance metric rather than AUROC (Supplementary Table S8). We overlaid the performance of individual GPCRs onto the phylogenetic tree (Supplementary Figure S6), revealing a clear pattern. GPCRs with above average MCC values (ACM4, OPRD, 5-HT1E, and NTR1) mostly belonged to larger subfamilies with multiple representatives in the training set, suggesting successful transfer of learned interaction patterns from related receptors. In contrast, GPCRs from underrepresented subfamilies showed below-average MCC values, indicating difficulty in predicting interactions with phylogenetically distant GPCRs.

We further investigated the role of ligand similarity in determining individual GPCR performance (Supplementary Figure S7a). We defined ligand similarity as the mean Tanimoto similarity of each GPCR's ligands against their top 5 nearest neighbors in the training set. We observed that GPCRs with high ligand similarity had a higher MCC value, and vice versa, with a few exceptions. For example, NTR1 and NTR2 were the only members of their subfamily and were both present in the test set, yet NTR1 performed substantially better than NTR2 (MCC difference of 0.367). We investigated this discrepancy by calculating the label difference, defined as the mean absolute difference between each test ligand's activity label (0 or 1), against the test GPCR and its average activity label across all GPCRs in the training set. NTR1 showed a lower label difference of 0.413, indicating that ligands active/inactive against NTR1 tended to show similar activity patterns in the training set, facilitating accurate predictions (Supplementary Table S8). In contrast, NTR2 exhibited a higher label difference of 0.546, suggesting that its pharmacology diverged substantially from training set patterns. Ligands that were typically active during training were often inactive against NTR2, and vice versa.

We also identified OX1R as an outlier that had a low MCC value of 0.073 despite a high ligand similarity of 0.658. We examined this case in detail and attribute the poor performance again to a high label difference (0.483). Ligands active against OX1R were frequently inactive against OX2R, its closest homolog in the training set, and vice versa (Supplementary Figure S7b). This systematic difference means that the model's learned associations from OX2R, i.e., whether specific ligands are active or inactive, did not transfer to OX1R. Therefore, despite high sequence similarity (67% identity in the entire sequence), OX1R and OX2R exhibit distinct pharmacological selectivity that our sequence-based model failed to capture, likely due to subtle differences in key binding pocket residues.

Discussion

We developed and systematically evaluated multiple deep learning architectures for predicting GPCR-ligand binding interactions relevant to neurotoxicity assessment, creating a curated dataset of 152,831 interactions across 119 neurologically relevant GPCRs. All of our architectures fall within the proteochemometrics paradigm, differing from classical PCM only in their use of learned foundation-model embeddings in place of handcrafted descriptors. Our best-performing models (CA-Base and CA-Lig) achieved AUROC values of 0.809 and 0.811 on structurally dissimilar ligands and 0.594 and 0.585 on novel GPCRs, respectively. However, fine-tuning the ProtBERT encoder (CA-Prot, CA-Full) substantially reduced performance, while fine-tuning the MolFormer encoder (CA-Lig) provided marginal benefits over frozen encoders (CA-Base). The reduced performance when fine-tuning ProtBERT suggests that frozen embeddings may already provide sufficient family-level information for classification tasks. Fine-tuning may have disrupted useful representations faster than the development of task-specific features, particularly for the larger ProtBERT models (420M). Hence, frozen pretrained encoders coupled with well-trained task layers may be preferable for these tasks. Furthermore, the consistent AUROC drop of ~ 0.100 for structurally dissimilar ligands (mean Tanimoto of 0.48 against training set) suggests that MolFormer's chemical features transfer reasonably well and the model can utilize the shared GPCRs to understand binding patterns. However, the larger AUROC drop of ~ 0.250 for novel GPCRs indicates that protein encodings transfer poorly across GPCR families and all models struggle without learned GPCR-ligand associations. These findings are consistent with earlier conclusions that predicting interactions for novel proteins is more challenging than predicting novel ligands, particularly when the number of ligands is significantly higher than the number of proteins [20].

Our analysis revealed that sequence-based architectures learn predominantly from ligand features, with protein encodings contributing less to binding-affinity predictions. Our findings align with recent observations on two kinase-ligand datasets, KIBA and DAVIS, where protein encodings, whether sequence based or structure

based, did not substantially impact binding predictions [44, 62, 63]. We attribute this pattern to several possible causes. First, because many ligands are tested against multiple proteins from the same subfamily, models may exploit these subfamily associations rather than learn actual binding site features. Second, ligand features may be more predictive and diverse than subtle protein sequence variations within families. Third, current protein language model embeddings may not adequately encode the fine-grained binding site features that distinguish closely related GPCRs. The convergence of this pattern across protein families (kinases and GPCRs) and diverse model architectures suggests that ligand-centric learning represents a general characteristic of current sequence-based models rather than a family- or architecture-specific limitation. However, for screening drug-like compounds against neurologically relevant GPCRs, this approach provides substantial value, achieving AUROC values of 0.907 and 0.811 on structurally similar and diverse ligands, respectively, in the CA-Base model.

To further investigate the mechanism underlying the performance degradation of CA-Prot and CA-Full, we performed additional representation-level analyses. We first measured the per-GPCR cosine distance between the pretrained and fine-tuned embeddings, a value that ranges from 0 (no change) to 1 (orthogonal, i.e., completely unrelated representations). The mean drift ranged between 0.58 and 0.66 across the CA-Prot and CA-Full seeds and was nearly identical on the training and unseen-protein receptors, indicating a global shift of the protein representations rather than overfitting to training GPCRs. Joint UMAP projections of ProtBERT embeddings showed that nine of 10 GPCR subfamilies clustered together in the pretrained embedding space and shifted together but remained co-located after fine-tuning (Supplementary Figure S8). The effective rank of GPCR embeddings on the training set dropped from 50.5 for pretrained ProtBERT to 22.8 (CA-Full) and 14.2 (CA-Prot), with the most decreased seeds reaching 3.6 and 1.1, respectively, indicating a near-total representation collapse. Layer-wise parameter analysis showed that relative Frobenius drift averaged 1–2% per layer, with the attention output sub-block showing the largest shifts and peak drift concentrated in the late layers of the encoder (Supplementary Figure S9). Collectively, these findings indicate that the relatively small GPCR training set is insufficient to guide ProtBERT fine-tuning productively; instead, it distorts the pretrained representation geometry and impairs generalization. The frozen ProtBERT used in CA-Base preserves this discriminative geometry, explaining its consistent advantage over the fine-tuned variants. A similar pattern was observed when we replaced ProtBERT with ESM2, confirming that the finding was not specific to the ProtBERT architecture. These findings are consistent with prior reports of unstable fine-tuning behavior in BERT-style models on small downstream datasets, with large seed-to-seed variability in the resulting model [64].

Comparing our results with broader DTI models is difficult due to three key differences. First, our primary objective was not to develop novel DTI models but rather understand how GPCR and ligand embeddings contribute to binding predictions and provide mechanistic insights beyond model performance. Second, we trained our models on 119 neurologically relevant GPCRs, which makes it difficult to compare them with general DTI models, which are generally trained on broad kinase datasets (DAVIS, KIBA) or diverse protein classes (BindingDB). Third, we tested our models under more challenging conditions by employing both ligand clustering and GPCR holdout, whereas most DTI classification models use random data splitting for model validation and report AUROC values exceeding 0.900. If the same models were evaluated on more stringent data splits, performance would most likely be worse. Our AUROC of 0.811 on clustering-based splits represents competitive performance with state-of-the-art methods under comparable conditions.

However, our work has limitations, which suggest important future directions for predicting protein–ligand binding. For example, we used binary classification because we combined heterogeneous assay data and multiple assay types from ChEMBL and BindingDB, which introduces noise and limits performance. Binary labels also discard quantitative affinity information and may group both strong and weak binders in the same class. Access to large-scale homogeneous datasets from standardized assays would substantially improve model development. Furthermore, our focus on GPCRs leaves open questions about whether ligand-centric patterns persist across other protein classes, such as ion channels, nuclear receptors, and transporters. In addition, we did not perform experimental validation, which could prove whether our models actually work in practice.

We also offer several ideas to improve general protein–ligand binding prediction. First, incorporating three-dimensional protein structures with binding site information could enable the models to look beyond subfamily patterns and learn geometrically relevant regions. Next, we could try to improve the pretraining objectives for new foundational models that focus on binding predictions. For example, using contrastive learning to train models to recognize when different protein families bind similar compounds could provide more relevant protein representations. Finally, better evaluation strategies should explicitly test cross-family

generalization and selectivity between similar targets and be integrated with experimental validation.

Conclusions

Our deep learning models to predict ligand interactions with neurologically relevant GPCRs performed well on out-of-distribution test sets, revealing the models' capabilities. We also demonstrated that current sequence-based architectures learn predominantly from ligand features across model architectures, from simple concatenation to sophisticated cross-attention. Our results indicate that future model architectures should incorporate mechanisms to learn from protein binding site geometry, apply constraints to prevent protein family identification, and develop foundational models with joint protein–ligand binding-specific pretraining. Our models will work well for drug-like molecules against well-characterized GPCRs, the regime most relevant for neurologically relevant GPCRs-based toxicity screening but will require careful implementation for novel chemical scaffolds or underrepresented GPCR families. We also emphasize the importance of understanding the true capabilities and limitations of models beyond chasing state-of-the-art benchmarks. Without carefully analyzing what the models actually learn, we risk deploying tools that work well on validation sets but fail to predict novel compounds. Here, our curated dataset, rigorous evaluation framework, and transparent characterization provide a clear understanding of current model capabilities and a direction for developing next-generation methods balancing practical utility with mechanistic understanding.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13321-026-01258-7.

Supplementary material 1.

Acknowledgements

We acknowledge support from the Department of Defense High Performance Computing Modernization Program and The United States Air Force Research Laboratory for use of their supercomputing centers. We thank Mr. Andrew Frock and Mr. Zhenbo Yan for helpful discussions and Ms. Maria Kuhrmann for editorial assistance. The opinions and assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the views of the Defense Health Agency, the U.S. Department of Defense, the U.S. Government, or The Henry M. Jackson Foundation for the Advancement of Military Medicine, Inc. Distribution Statement A. Approved for public release: distribution is unlimited.

Author contributions

SD performed the computations and data analysis and wrote the first draft of the manuscript. MDMA conceived and supervised the study and provided overall project leadership and scientific guidance. SD, PL, AW, and MDMA contributed to study design, interpretation of the results, and critical revision of the manuscript. All authors reviewed and approved the final manuscript.

Funding

This research was funded by the U.S. Army Medical Research and Development Command under Contract Nos. W81XWH20C0031 and HT942524F0189 and by Defense Threat Reduction Agency Grant CBCall14-CBS-05-2-0007.

Data availability

We make the data, scripts, and trained models available on the GitHub repository (https://github.com/BHSAI/neuroGPCRs).

Declarations

Ethics approval and consent to participate

The modeling results presented herein were independently assessed for reproducibility.

Competing interests

The authors declare no competing interests.

Author details

[1]Department of Defense Biotechnology High Performance Computing Software Applications Institute, Defense Health Agency Research & Development, Medical Research and Development Command, Fort Detrick, MD 21702-5012, USA. [2]The Henry M. Jackson Foundation for the Advancement of Military Medicine, Inc., Bethesda, MD, USA.

Received: 17 February 2026 Accepted: 4 July 2026
Published online: 15 July 2026

References

1. Waring MJ, Arrowsmith J, Leach AR, Leeson PD, Mandrell S, Owen RM et al (2015) An analysis of the attrition of drug candidates from four major pharmaceutical companies. Nat Rev Drug Discov 14(7):475–486
2. Hay M, Thomas DW, Craighead JL, Economides C, Rosenthal J (2014) Clinical development success rates for investigational drugs. Nat Biotechnol 32(1):40–51
3. Cook D, Brown D, Alexander R, March R, Morgan P, Satterthwaite G et al (2014) Lessons learned from the fate of AstraZeneca's drug pipeline: a five-dimensional framework. Nat Rev Drug Discov 13(6):419–431
4. Basketter D, Clewell H, Kimber I, Rossi A, Blaauboer B, Burrier R et al (2012) A roadmap for the development of alternative (non-animal) methods for systemic toxicity testing. Altex 29(1):3–91
5. Schmeisser S, Miccoli A, von Bergen M, Berggren E, Braeuning A, Busch W et al (2023) New approach methodologies in human regulatory toxicology - not if, but how and when! Environ Int 178:108082
6. Avila AM, Bebenek I, Bonzo JA, Bourcier T, Davis Bruno KL, Carlson DB et al (2020) An FDA/CDER perspective on nonclinical testing strategies: classical toxicology approaches and new approach methodologies (NAMs). Regul Toxicol Pharmacol 114:104662
7. Shukla SJ, Huang R, Austin CP, Xia M (2010) The future of toxicity testing: a focus on *in vitro* methods using a quantitative high-throughput screening platform. Drug Discov Today 15(23–24):997–1007
8. Hauser AS, Attwood MM, Rask-Andersen M, Schioth HB, Gloriam DE (2017) Trends in GPCR drug discovery: new agents, targets and indications. Nat Rev Drug Discov 16(12):829–842
9. Roth B, Sheffler D, Kroeze W (2004) Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia. Nat Rev Drug Discov 3(4):353–359
10. Bender A, Scheiber J, Glick M, Davies JW, Azzaoui K, Hamon J et al (2007) Analysis of pharmacology data and the prediction of adverse drug reactions and off-target effects from chemical structure. ChemMedChem 2(6):861–873
11. Wiholm B, Mortimer O, Boethius G, Häggström JE (1984) Tardive dyskinesia associated with metoclopramide. Br Med J (Clin Res Ed) 288(6416):545–547
12. Rao A, Camilleri M (2010) Review article: metoclopramide and tardive dyskinesia. Aliment Pharmacol Ther 31(1):11–19

13. Kay GG, Ebinger U (2008) Preserving cognitive function for patients with overactive bladder: evidence for a differential effect with darifenacin. Int J Clin Pract 62(11):1792–1800
14. Coupland CAC, Hill T, Dening T, Morriss R, Moore M, Hippisley-Cox J (2019) Anticholinergic drug exposure and the risk of dementia: a nested case-control study. JAMA Intern Med 179(8):1084–1093
15. Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin II, Cronin M et al (2014) QSAR modeling: where have you been? Where are you going to? J Med Chem 57(12):4977–5010
16. Tropsha A (2010) Best practices for QSAR model development, validation, and exploitation. Mol Inform 29(6–7):476–488
17. Overington JP, Al-Lazikani B, Hopkins AL (2006) How many drug targets are there? Nat Rev Drug Discov 5(12):993–996
18. Yamanishi Y, Araki M, Gutteridge A, Honda W, Kanehisa M (2008) Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. Bioinformatics 24(13):i232–i240
19. Cortés-Ciriano I, Ain QU, Subramanian V, Lenselink EB, Méndez-Lucio O, Ijzerman AP et al (2015) Polypharmacology modelling using proteochemometrics (PCM): recent methodological developments, applications to target families, and future prospects. MedChemComm 6(1):24–50
20. Pahikkala T, Airola A, Pietila S, Shakyawar S, Szwajda A, Tang J et al (2015) Toward more realistic drug-target interaction predictions. Brief Bioinform 16(2):325–337
21. Ozturk H, Ozgur A, Ozkirimli E (2018) DeepDTA: deep drug-target binding affinity prediction. Bioinformatics 34(17):i821–i829
22. Nguyen T, Le H, Quinn TP, Nguyen T, Le TD, Venkatesh S (2021) GraphDTA: predicting drug-target binding affinity with graph neural networks. Bioinformatics 37(8):1140–1147
23. Huang K, Xiao C, Glass LM, Sun J (2021) MolTrans: molecular interaction transformer for drug-target interaction prediction. Bioinformatics 37(6):830–836
24. Chen L, Tan X, Wang D, Zhong F, Liu X, Yang T et al (2020) TransformerCPI: improving compound-protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. Bioinformatics 36(16):4406–4414
25. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O et al (2021) Highly accurate protein structure prediction with AlphaFold. Nature 596(7873):583–589
26. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W et al (2023) Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 379(6637):1123–1130
27. Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L et al (2022) ProtTrans: toward understanding the language of life through self-supervised learning. IEEE Trans Pattern Anal Mach Intell 44(10):7112–7127
28. Ross J, Belgodere B, Chenthamarakshan V, Padhi I, Mroueh Y, Das P (2022) Large-scale chemical language representations capture molecular structure and properties. Nat Mach Intell 4(12):1256–1264
29. Chithrananda S, Grand G, Ramsundar B (2020) ChemBERTa: large-scale self-supervised pretraining for molecular property prediction. arXiv preprint arXiv:2010.09885
30. Wu H, Liu J, Jiang T, Zou Q, Qi S, Cui Z et al (2024) AttentionMGT-DTA: a multi-modal drug-target affinity prediction using graph transformer and attention mechanism. Neural Netw 169:623–636
31. Zhang S, Huo D, Horne RI, Qi Y, Pujalte Ojeda S, Yan A et al (2025) Sequence-based virtual screening using transformers. Nat Commun 16(1):6925
32. Shi MH, Zhang SW, Zhang QQ, Han Y, Zhang S (2025) PLAGCA: predicting protein-ligand binding affinity with the graph cross-attention mechanism. J Biomed Inform 165:104816
33. Lu Z, Song G, Zhu H, Lei C, Sun X, Wang K et al (2025) DTIAM: a unified framework for predicting drug-target interactions, binding affinities and drug mechanisms. Nat Commun 16(1):2548
34. Wang G, Zhang H, Shao M, Feng Y, Cao C, Hu X et al (2024) DeepTGIN: a novel hybrid multimodal approach using transformers and graph isomorphism networks for protein-ligand binding affinity prediction. J Cheminform 16(1):147
35. Zhang L, Yao T, Luo J, Yi H, Han X, Pan W et al (2024) ChemNTP: advanced prediction of neurotoxicity targets for environmental chemicals using a Siamese neural network. Environ Sci Technol 58(51):22646–22656
36. Rathore AS, Jain S, Choudhury S, Raghava GPS (2025) A large language model for predicting neurotoxic peptides and neurotoxins. Protein Sci 34(8):e70200
37. Pang X, He X, Yang Y, Wang L, Sun Y, Cao H et al (2025) NeuTox 2.0: a hybrid deep learning architecture for screening potential neurotoxicity of chemicals based on multimodal feature fusion. Environ Int 195:109244
38. Ahmed M, Hasani HJ, Kalyaanamoorthy S, Barakat K (2021) GPCR_LigandClassify.py; a rigorous machine learning classifier for GPCR targeting compounds. Sci Rep 11(1):9510
39. Deflorian F, Perez-Benito L, Lenselink EB, Congreve M, van Vlijmen HWT, Mason JS et al (2020) Accurate prediction of GPCR ligand binding affinity with free energy perturbation. J Chem Inf Model 60(11):5563–5579
40. Remington JM, McKay KT, Beckage NB, Ferrell JB, Schneebeli ST, Li J (2023) GPCRLigNet: rapid screening for GPCR active ligands using machine learning. J Comput Aided Mol Des 37(3):147–156
41. Burggraaff L, van Veen A, Lam CC, van Vlijmen HWT, IJerman AP, van Westen GJP (2020) Annotation of allosteric compounds to enhance bioactivity modeling for class A GPCRs. J Chem Inf Model 60(10):4664–4672
42. Yamane H, Ishida T (2023) Helix encoder: a compound-protein interaction prediction model specifically designed for class A GPCRs. Front Bioinform 3:1193025
43. Brahma R, Moon S, Shin JM, Cho KH (2025) AiGPro: a multi-tasks model for profiling of GPCRs for agonist and antagonist. J Cheminform 17(1):12
44. Gorantla R, Kubincova A, Weisse AY, Mey A (2024) From proteins to ligands: decoding deep learning methods for binding affinity prediction. J Chem Inf Model 64(7):2496–2507
45. Mack CM, Tsui-Bowen A, Smith AR, Jensen KF, Kodavanti PRS, Moser VC (2024) Identification of neural-relevant toxcast high-throughput assay intended gene targets: applicability to neurotoxicity and neurotoxicant putative molecular initiating events. Neurotoxicology 103:256–265
46. Viviani B, Midali M, Serafini MM, Mangas I (2025) Implementing AOPs use for pesticide neurotoxicity assessment. Curr Opin Toxicol 41:100505
47. Bento AP, Hersey A, Felix E, Landrum G, Gaulton A, Atkinson F et al (2020) An open source chemical structure curation pipeline using RDKit. J Cheminform 12(1):51
48. Oprea TI, Bologa CG, Brunak S, Campbell A, Gan GN, Gaulton A et al (2018) Unexplored therapeutic opportunities in the human genome. Nat Rev Drug Discov 17(5):317–332
49. Butina D (1999) Unsupervised data base clustering based on daylight's fingerprint and tanimoto similarity: a fast and automated way to cluster small and large data sets. J Chem Inf Comput Sci 39(4):747–750
50. Devlin J, Chang M-W, Lee K, Toutanova K (2019) Bert: pre-training of deep bidirectional transformers for language understanding. In: North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 1: 4171–4186
51. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. J Chem Inf Model 50(5):742–754
52. Singh R, Sledzieski S, Bryson B, Cowen L, Berger B (2023) Contrastive learning in protein language space predicts interactions between drugs and protein targets. Proc Natl Acad Sci U S A 120(24):e2220778120
53. Nga HC, Pham P, Hy TS (2025) LANTERN: leveraging large language models and transformers for enhanced molecular interactions. bioRxiv:2025.02.10.637522
54. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al (2017) Attention is all you need. In: Advances in Neural Information Processing Systems, 30
55. Kurata H, Tsukiyama S (2022) ICAN: interpretable cross-attention network for identifying drug and target protein interactions. PLoS ONE 17(10):e0276609
56. Chen T, Guestrin C (2016) XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, p 785–794
57. Akiba T, Sano S, Yanase T, Ohta, T, Koyama M (2019) Optuna: a next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, p 2623–2631
58. Bergstra J, Bardenet R, Bengio Y, Kégl B (2011) Algorithms for hyper-parameter optimization. In: Advances in Neural Information Processing Systems, 24

59. Sheridan RP, Wang WM, Liaw A, Ma J, Gifford EM (2016) Extreme gradient boosting as a method for quantitative structure-activity relationships. J Chem Inf Model 56(12):2353–2360
60. Herrera LPT, Andreassen SN, Caroli J, Rodriguez-Espigares I, Kermani AA, Keseru GM et al (2025) GPCRdb in 2025: adding odorant receptors, data mapper, structure similarity search and models of physiological ligand complexes. Nucleic Acids Res 53(D1):D425–D435
61. Avdiunina P, Jamal S, Gusev F, Isayev O (2025) All that glitters is not gold: importance of rigorous evaluation of proteochemometric models. J Chem Inf Model 65(19):10239–10252
62. Davis MI, Hunt JP, Herrgard S, Ciceri P, Wodicka LM, Pallares G et al (2011) Comprehensive analysis of kinase inhibitor selectivity. Nat Biotechnol 29(11):1046–1051
63. Tang J, Szwajda A, Shakyawar S, Xu T, Hintsanen P, Wennerberg K et al (2014) Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. J Chem Inf Model 54(3):735–743
64. Mosbach M, Andriushchenko M, Klakow D. (2021) On the stability of fine-tuning BERT: misconceptions, explanations, and strong baselines. arXiv preprint arXiv:2006.04884

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.