Labels as a Feature: Network Homophily for Systematically Discovering human GPCR Drug-Target Interactions.

Frederik G. Hansson[1†] **Niklas Gesmar Madsen**[1†] **Lea G. Hansen**[2] **Tadas Jakočiūnas**[1]

Bettina Lengger[1] **Jay D. Keasling**[3 4 5 6] **Michael K. Jensen**[1 2] **Carlos G. Acevedo-Rocha**[1*]

Emil D. Jensen[1*]

[1] The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark

[2] Biomia ApS

[3] Joint BioEnergy Institute, Emeryville, CA, USA

[4] Biological Systems and Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

[5] Department of Chemical and Biomolecular Engineering, Department of Bioengineering, University of California, Berkeley, CA, USA

[6] Center for Synthetic Biochemistry, Institute for Synthetic Biology, Shenzhen Institutes of Advanced Technologies, Shenzhen, China

[*] Corresponding authors;
emdaje@biosustain.dtu.dk
cargac@biosustain.dtu.dk

[†] These authors contributed equally to the study

Abstract

Machine learning (ML) has revolutionized drug discovery by enabling the exploration of vast, uncharted chemical spaces essential for discovering novel patentable drugs. Despite the critical role of human G protein-coupled receptors (hGPCRs) in FDA-approved drugs, exhaustive in-distribution drug-target interaction (DTI) testing across all pairs of hGPCRs and known drugs is rare due to significant economic and technical challenges. This often leaves off-target effects unexplored, which poses a considerable risk to drug safety. In contrast to the traditional focus on out-of-distribution (OOD) exploration (drug discovery), we introduce a neighborhood-to-prediction model termed Chemical Space Neural Networks (CSNN) that leverages network homophily and training-free graph neural networks (GNNs) with Labels as Features (LaF). We show that CSNN's ability to make accurate predictions strongly correlates with network homophily. Thus, LaFs strongly increase a ML model's capacity to enhance in-distribution prediction accuracy, which we show by integrating labeled data during inference. We validate these advancements in a high-throughput yeast biosensing system (3773 DTIs, 539 compounds, 7 hGPCRs) to discover novel DTIs for FDA-approved drugs and to expand the general understanding of how to build reliable predictors to guide experimental verification.

1 Introduction

Machine learning (ML) methods have significantly advanced high-throughput drug discovery pipelines by enabling virtual compound screening for large chemical spaces, the vastness of which are inaccessible to fully explore experimentally. Furthermore, public drug-target interaction (DTI) databases are growing at an exponential rate, hence enabling the further enhancement of ML methods at predicting compound bioactivity (*1*, *2*). Human G protein-coupled receptors (hGPCRs) are one of the most important classes of drug targets and include over 286 non-olfactory receptors (*3*), which are targeted by approximately 35% of all FDA-approved pharmaceuticals (*4*).

Despite the critical role of hGPCRs in drug discovery and their documented promiscuity (*3*, *5*), it remains uncommon to exhaustively test all hGPCRs against all drugs in a library (*6*). Illustratively, there are a lot of empty *in-distribution* labels missing (Figure 1a). For instance, for 186 thousand unique hGPCR-targeting compounds across 128 hGPCRs, only 1.5 % of all possible activities have been documented in publicly sourced data. Thus, 98.5 % of DTIs remain in the dark. Naturally, this is largely due to the significant economic and technical challenges involved at this scale. Nevertheless, most approved hGPCR-targeting drugs are accompanied by famously extensive medicine package leaflets documenting the incidence rate of side-effects, which are likely due to off-target drug-interaction(s) (*7*, *8*).

Solving this problem by exhaustively documenting off-target effects is thus the opposite of drug discovery pipelines: These slowly move toward undocumented regions of chemical space (out-of-distribution (OOD) and patentability). However, in this case, robust and reliable *in-distribution* prediction is required to narrow down the DTIs selected for experimental verification.

A critical inductive bias in many ML models is the assumption of smoothness or locality, where there is a presumed continuity or proximity between the input space and the output space. This implies proximity between similar compounds in chemical space. In chemoinformatics, this is known as the chemical homophily principle (*9*) and is more generally known as network homophily. Intuitively, "similarity breeds connection" and is found in a range of social, economic behavior as well as network based databases (*10*). By analogy, similar molecules have similar bioactivity and, if the network is homophilous, prediction is straight-forward when a chemical neighbourhood is known (Figure 1b).

Yet, ML for drug discovery has primarily focused on compound-to-prediction architectures rather than considering the chemical data available in the chemical neighbourhood during inference (neighbourhood-to-prediction) (See Table S10 for a comparison). If indeed DTI data is homophilous for hGPCRs, then it would be beneficial to include data during inference to increase reliability and robustness of predictions.

An emerging paradigm known as training free graph neural networks (TFGNNs) (*11*), show how labels as a features (LaF) is an admissible operation in transductive node classification tasks: labels of neighbouring nodes are used to update the learned representations. The paradigm proves how this approach increases the expressive power over Graph Neural Networks (GNNs). Empirically, the results show that even without training, prediction accuracy is incredibly high on common benchmark GNN datasets. Here, we demonstrate similar training free predictions that are on-par or outperform conventional trained compound-to-prediction architectures. Instead of framing the task as transductive node classification, we re-frame it as a graph classification task with neighbourhood labels as edge features (Figure 1d) in a directed (transductive) graph, which integrates available data during inference. We demonstrate strong network homophily in hGPCR bioactivity data and show a strong correlation between network homophily and machine learned prediction accuracy. By exploiting the graph homophily of chemical space networks (CSN) we develop Chemical Space Neural Networks (CSNN) and combine the method with a high-throughput yeast biosensing system

for experimentally validating novel-DTIs. Finally, we correlate predictions with experimental results and discover 14 novel DTIs potentially linked with off-target effects.

2 Results

Building the architecture of Chemical Space Neural Networks.

As a basis for our work, we compiled a dataset containing cleaned bioactivity data on hGPCRs available in ChEMBL(*12*) and the IUPHAR/BPS guide to pharmacology(*13*). In total, the dataset comprises 186,723 unique molecules represented as SMILES and labeled with known bioactivity classes against the 128 respective hGPCRs (Supplementary File S1).

Figure 1a shows that ChEMBL has >369,000 annotated activities across 128 hGPCRs. This is, however, a vanishing fraction of all possible interactions (1.5 % coverage), leaving knowledge of the majority (98.5 %) of DTIs in the dark. ChEMBL is growing at a rate of 780 K activities per year (Supplementary Figure S17), which illustrates the need for constructing reliable predictive algorithms to fill in the missing interactions.

Given the homophily principle assumption, compounds which are close in chemical space often have similar bioactivity class on a given hGPCR, it is thus conceivable that a chemical neighbourhood is strongly indicative of a compound's bioactivity class. This is illustrated in Figure 1b, d, where we extended upon works by Yu-Chen Lo et. al (*14*) and Zengrui Wu et. al (*6*, *15*) to develop Chemical Space Neural Networks (CSNN). The key advantage is illustrated in Figure S1: Commonly, methods using ML for hGPCR drug-discovery are compound-to-prediction architectures. This takes a compound from the chemical space and directly infers its target through some learned or fixed representation (e.g., chemical fingerprints) with ML models. CSNN, developed in this work, takes a local chemical neighbourhood by querying a database during inference, learning a local representation using LaF, and then using the constructed neighbourhood graph for predicting DTIs (regression and classification).

Figure 1c illustrates the inference process of CSNN, which operates akin to the transductive node classification task. First, a query compound i is represented by a bit-vector, then a one-vs-all database search (by Tanimoto similarty) returns chemically similar neighbours (related compounds). From a pre-trained directed message-passing neural network (D-MPNN, ChemProp) a dense representation is obtained for the query and neighbouring compounds and edge features are compiled to a neighbourhood graph $\mathcal{G}_i$ as illustrated in Figure 1d. Here, neighbouring compound labels are used as edge features. Then a trained MPNN takes the graph $\mathcal{G}_i$ mapping it to it's y-value (classification/regression): $f_\theta : \mathcal{G}_i \rightarrow \mathbf{y}_i$. Effectively, the method is composing MPNN layers at the level of compounds (MPNN on atoms) and again at the neighbourhood level (MPNN between DTIs) as shown in Figure 1e. To overcome the current limitations of CSNs, we develop an extremely fast way of constructing the chemical neighbours by reducing the computation of the 186 K x 186 K adjacency matrix (CSN) to 20 min (described in methods and additional results in S8). Inference for a one-vs-all (1-vs-186 K) lookup takes 0.187 $\pm$ 0.002 seconds.

Chemical Neighbourhoods enable training free and confident predictions of hGPCR bioactivity classes.

With an established all-vs-all CSN (Fig. 2a) from which one can derive important neighbourhood statistics, we can from the visualised example presented in Figure 2b see that the the query molecule has primarily 'Agonists' in its neighbourhood, which is in fact the true label of the compound for the hGPCRs shown. More generally, across 201,120 unique DTIs and across all 128 hGPCRs and 6 bioactivity classes, the neighbourhood achieved an F1 score of up to 93 % and shows remarkable performance across bioactivity classes (See Figure 2c and S3). Here we simply calculate the frequency of each label in the neighbourhood of a query and take the most frequent label as the prediction (Argmax operation), then compare with the true

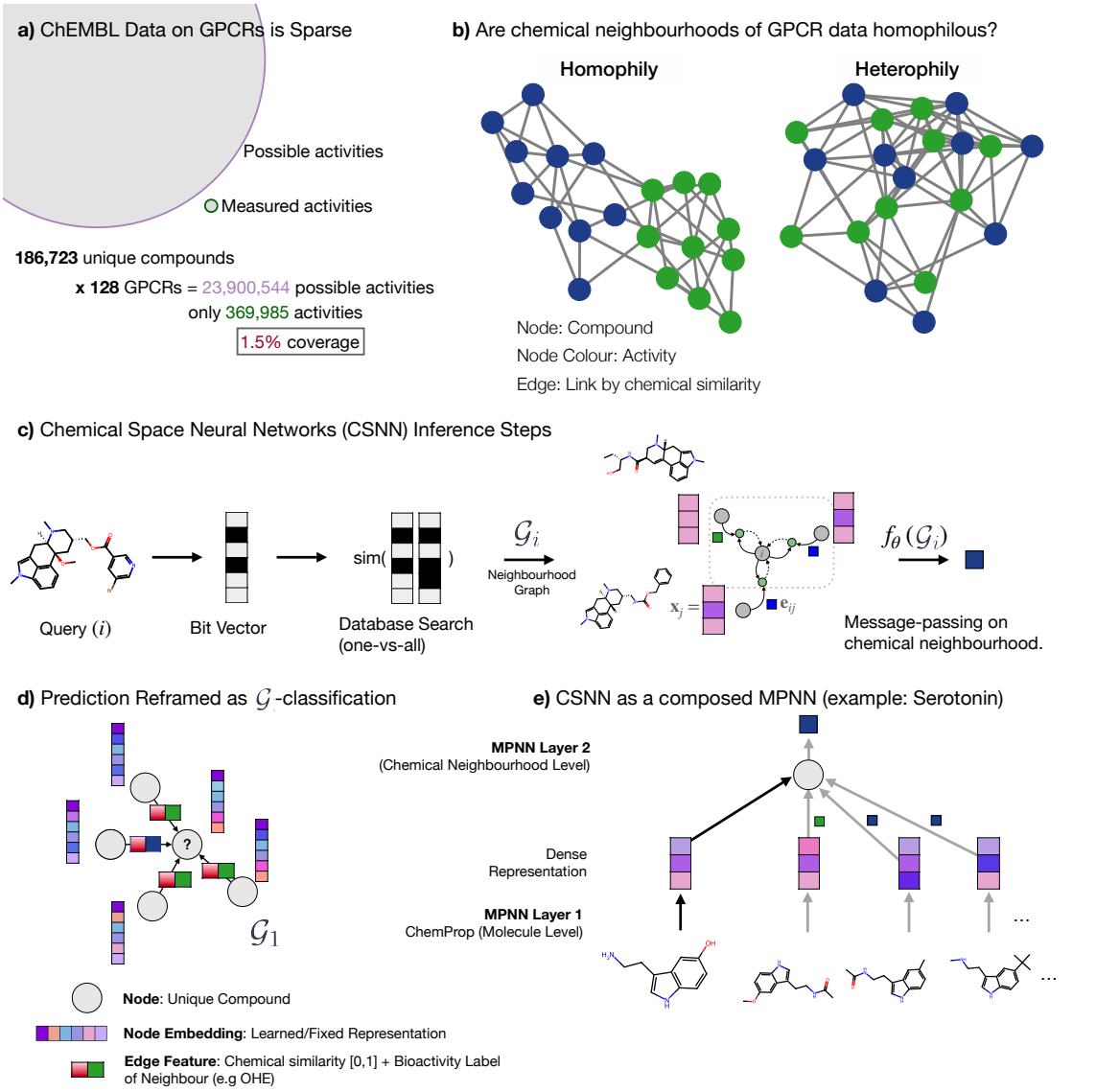


Figure 1: **Introducing Network Homophily and Transductive Node Classification**. **(a)** Collected data on bioactivity classes for 186 K unique hGPCR-targeting compounds across 128 hGPCRs only has 369 K of 23.9 M possible activities. This indicates the sparse annotation in public databases. **(b)** A visual example to introduce network homophily: "similarity breeds connection" (*10*). **(c)** Inference for CSNN to illustrate how labels as a feature are used. First the query compound is encoded by a bit-vector, then a one-vs-all database search returns compounds, which are chemically similar and labels on those compounds. A neighbourhood graph $\mathcal{G}_i$ is constructed that is fed to the ML network (f_θ). **(d)** Instead of transductive node classification, we simplify the task to transductive graph classification by introducing a one-hop directed graph (incident on the query) with neighbourhood LaFs as edge attributes. **(e)** The CSNN framework can be viewed a as composed message-passing neural network (MPNN): First at the molecule level (MPNN on atoms) and then again at the neighbourhood level including LaFs.

label on a given hGPCR. Thus the neighbourhood labels becomes a feature for inference. This demonstrates the strong homophily of the hGPCR-DTI space and shows a training free GNN approach with LaFs. To enable ML on these neighbourhoods, we structure the data into the standardised open-source PyTorch Geometric framework (*16*) (Figure 2d).

In Figure 2e, f and g we illustrate the performance metrics of ML methods that operate on chemical neighbourhoods contrasted with normal compound-to-prediction methods (without a neighbourhood of LaFs). In Figure 2e we compare common models (Multilayer Perceptron: MLP, and Random Forest: RF) without the neighbourhood using only the compound representation and with the neighbourhood (+ N) frequency vector (LaF concatenated). Just this LaF inclusion significantly improves performance. Interestingly, the developed CSNN method (GNN on chemical neighbourhood) is only on par with the Argmax prediction. We refer to this method hence-forth as NN_θ^6 as it produces logits for the six bioactivity classes.

However, if one only includes highly-confident predictions (logits are strongly peaked, high class probability), the performance outperforms Argmax classification metrics (Fig. 2f,g). This falls in line with the intuition provided in (*11*), that LaFs improve the expressive power of GNNs. At a high class probability (> 0.8) the classification is almost without error on the test set (see Fig. S2a for complete class summaries and Fig. S2b,c for ROC-AUC curves). The advantage of neural network over Argmax predictions is that the former are probabilistic and that the model confidence directly correlates with the mean ROC-AUC (Fig. S5). For further and more complete investigations of the NN_θ^6 model, see supplementary text and Figures S2 - S5.

To generalise the transductive node classification method to include all data on all hGPCRs in the neighbourhood and reduce the number of forward passes to predict labels for all compound-hGPCR pairs, we develop a NN_θ^{128} CSNN model, which in a single forward pass produces bioactivity class logits for all 128 hGPCRs at once. The CSNN takes a graph $\mathcal{G}$ as input and predicts bioactivity as $\mathbf{y} \in \mathbb{R}^{1\times128\times6}$, which represents class-wise logits. The input graph contains all available data on all 128 hGPCRs as edge features (LaF), thus allowing for mixing between data channels. This was intuitively advantageous, as related hGPCRs show similar signalling responses to the same molecule, thus enhancing the predictive performance by integrating all available data in the sparsely label dataset. Results of this method are reported in Figure 2h, which shows that for the top-10 scoring hGPCRs, classification was near perfect (class weighted F1-score of 0.978 ± 0.012). Across all hGPCRs, the classification weighted F1-score was 0.83 ± 0.10, whereas if only the agonist class assignment is considered, the F1-score was 0.877. Further quality metrics and full class summaries are provided in Figure S6a,b,c, where we demonstrate a strong correlation between testing F1-score and training support for each bioactivity class and ROC-AUC > 0.95 for all classes, thus outperforming the NN_θ^6 model.

Benchmarking Neighbourhood Methods on Regression Tasks and newly generated datasets

As LaF showed strong promise in predicting class labels, we perform a benchmark study on a recently published method and dataset (pdCSM (*17*)) in the more challenging regression setting for binding affinity predictions. We compute a CSN on all compounds and allow training nodes to refer to each other, while restricting testing compounds only to refer to the training data during inference (admissible transductive framework). Figure 3a, shows a schematic training example with three nodes (compounds) and their true $-\log_{10}(K_i)$ values. Taking a mean over the neighbourhood (6.20) is an acceptable prediction of the true regression label (6.14). Figure 3b shows that the mean prediction $< Y_{neigh} >$ is accurate across all hGPCRs in the dataset (MSE = 0.685, MAE = 0.624, Pearson = 0.757). Regression plots split by hGPCR are found in Figure S9 and as a function of aggregation technique (mean, max, min, weighted) in Figure S10.

To benchmark our method, we reconstruct molecule representations used for pdCSM with available data from (*17*) (see dataset summary in Table S9, in which training/test data points 47013/7114 were used based

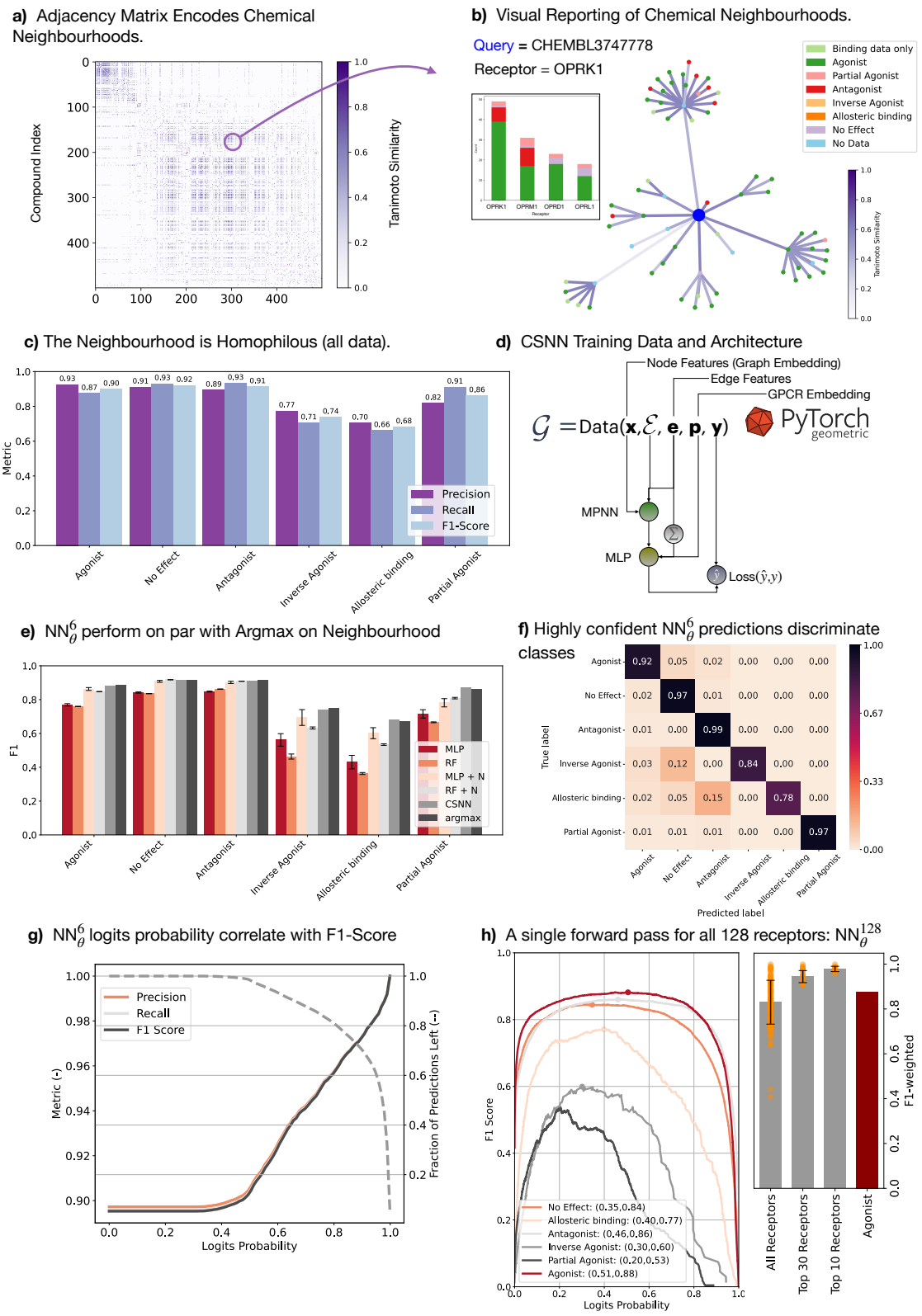


Figure 2: **Benchmarking Bioactivity Label Prediction with LaFs**. **(a)** The adjacency matrix (186 K $\times$ 186 K possible connections) parameterised the all-vs-all chemical space neighbourhood. **(b)** Querying the CSN can return data on related compounds already illustrating graph homophily (agonist are over represented and the true label for OPRK1 is agonist). **(c)** Training free prediction metrics using the most frequent LaF in the neighbourhood, demonstrates the strong network homophily. **(d)** Structuring the dataset into an accessible format and CSNN architecture. **(e)** Prediction metrics on test set for ML methods: without LaFs (MLP, RF), with LaFs (MLP + N, CSNN), and a full MPNN on chemical neighbourhoods (CSNN) compared to the training free prediction (Argmax). **(f)** When the class label is confident probability >0.8, the CSNN method (referred to as NN$^6_\theta$) produces high-quality class discrimination. **(g)** As NN$^6_\theta$ class label is filtered by logits probability, the performance metrics tend toward perfect predictions. **(h)** NN$^1_\theta$28 CSNN model (one forward pass for class labels across all 128 hGPCRs using LaFs) shows strong performance metrics for most hGPCRs.

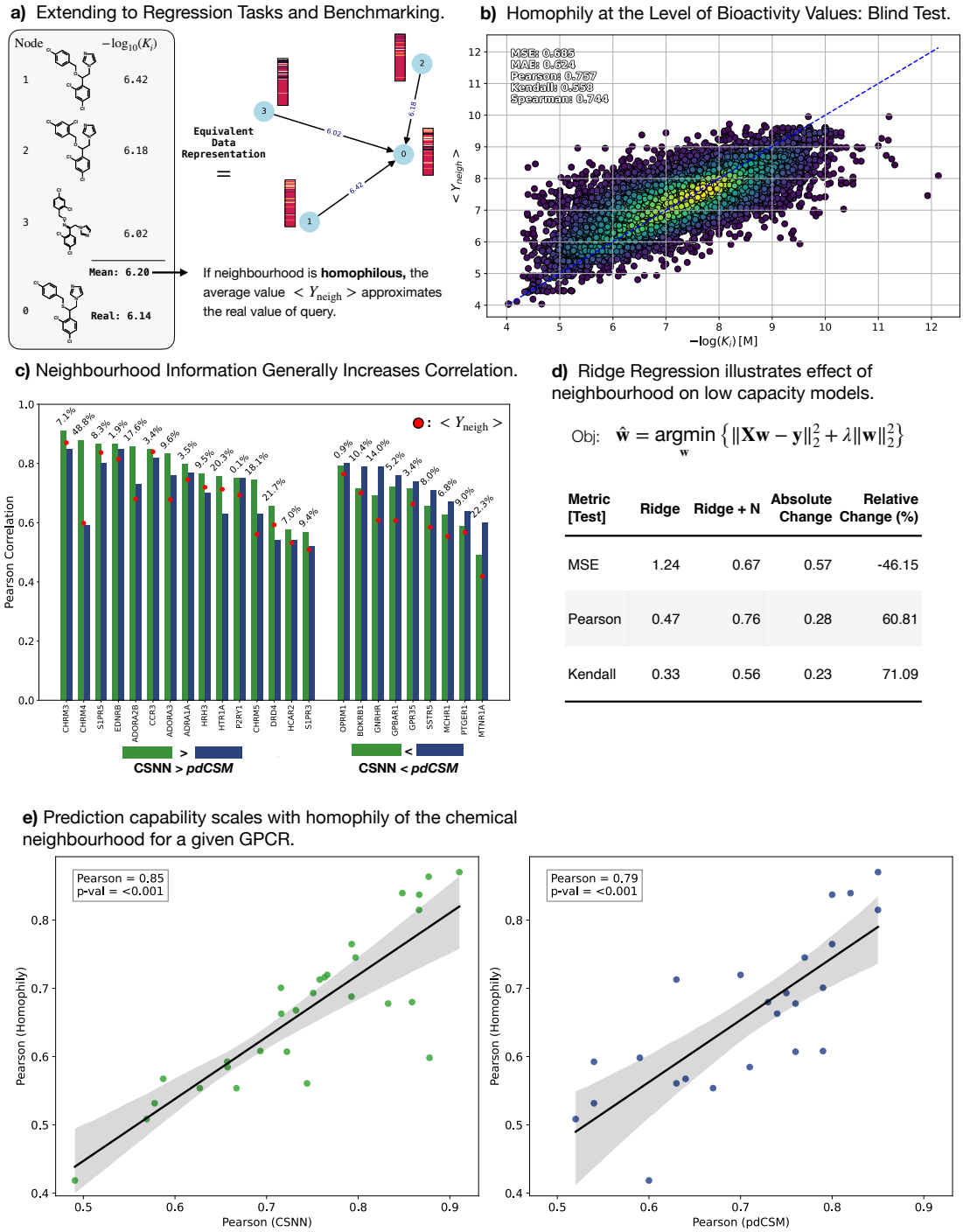
on the published data). We chose pdCSM due to the relative dataset size and published source data. We instantiate a CSNN using a random forest (RF) model as done in the published benchmark method (pdCSM), but include the neighbourhood mean value prediction, which allows a comparison between methods without and with neighbourhood information (LaF). Unlike the pdCSM method, which has no available training or inference code, we do not perform feature selection as their code was not available to replicate the method. Nevertheless, CSNN outperforms pdCSM on 15/24 targets (Fig. 3c, and Figs. S13, S14), and requires only an increase in feature dimension by one (the mean-value of the neighbourhood). Furthermore, CSNN have a guard-rail against OOD predictions, as without a neighbourhood no predictions are made. This is not the case for the benchmark method, which will predict K_i values for *any* SMILES compound. Interestingly, the maximum Pearson correlation achieved on a given hGPCR is highly correlated with the mean-value prediction derived from the homophily assumption (Fig. 3c, e), implying that the more homophilous the neighbourhood (mean-value is a good approximation) is, the better the ML model performs. Accordingly, the pdCSM method only outperforms a training-free prediction ($< Y_{neigh} >$) on 17/24 hGPCR targets, while CSNN outperforms on all (24/24) hGPCRs. Overall, these results confirm the smoothness inductive bias from the introduction between the input and output space ($f_\theta : \mathbb{R}^n \rightarrow \mathbb{R}$), allowing one to predict the utility of a ML-model depending on the homophily of the dataset.

The effect of having LaFs can be observed in a low capacity model, like Ridge regression as judged by the change in MSE, Pearson correlation and Kendall correlation without and with (+N) the neighbourhood information using the same fixed representation (Fig. 3d). The MSE is reduced by 46 % and the Pearson correlation coefficient increases by 60 %. Again, LaF has a strong effect on reducing the complexity of the learning task, as postulated in (*11*). The results are less striking but still apparent for RF models with or without LaF, it drives the MSE error down by 7.89 % and increases Pearson correlation by 2 % (Fig. S8) for a fixed model size.

Argmax and CSNN methods are able to expand the total coverage of labeled hGPCR-compound DTIs from 1.5 % to 5.6 % and 100 %, respectively. Naturally, the CSNN method contains a prediction confidence, thus implying that confident predictions do not ensure total coverage. The robust Argmax prediction still left 94.4 % of the DTIs unknown, which can not currently be predicted in a training free manner, as the data is too sparse to impute the missing data. Nevertheless, Argmax easily increased data coverage 4-fold from 370 K to 1.3 million unique drug-hGPCR interactions.

Given that LaFs is a strong, admissible, and a powerful inductive bias of network homophily, which is inherent in the DTI-hGPCR space, we sought to experimentally develop a high-throughput platform to compliment model predictions and to validate both the class probability and regression model.

Heterologous expression of hGPCRs in yeast offers advantage over traditional hGPCR screening methods because it is both cost-efficient and allow for accelerated screening of chemical libraries (*18*, *19*) (See Supplementary Notes, Table S5 and supplementary file S2). We therefore constructed a small set of yeast platform strains for 7 hGPCRs previously reported to signal in yeast (*18*, *20–24*), which were chosen based on their immediate potentials as drug targets (supplementary table S4). We based our design on the system presented by Shaw et al. (2019) (*25*) yet with NanoLUC (*26*) as a reporter of relative luminescence units (RLU) instead of GFP for increased assay sensitivity (*19*) (See Figure 4a for schematic overview). We confirmed their performances by measuring dose-response curves (DRC) with agonists on each hGPCR (See Figure 4b) to further evaluate the expected quality of gathered data using the yeast platform (Supplementary Figures S20, S21 and table S6). The EC50 values observed for CHRM3 and ADRA2B were comparable to the most sensitive screens reported (Supplementary table S6(*12*, *27*)). The dynamic ranges were in 5 out of 7 cases higher in yeast than observed for mammalian cell assays (Supplementary Figure S20, and table S6).



Metric [Test]	Ridge	Ridge + N	Absolute Change	Relative Change (%)
MSE	1.24	0.67	0.57	-46.15
Pearson	0.47	0.76	0.28	60.81
Kendall	0.33	0.56	0.23	71.09

Figure 3: **Benchmarking LaFs in the Regression Setting**. **(a)** An illustrated training example: node 0 is the query, nodes 1-3 are neighbouring compounds and their bioactivity label (LaF). The mean over the neighbourhood is 6.20, which is close to the true label of 6.14 (units: $-\log_{10}(K_i)$). **(b)** Training free mean value ($< Y_{neigh} >$) predictions on the test set using LaFs across all hGPCRs in the pdCSM dataset shows strong performance metrics. **(c)** Comparing the published pdCSM method (RF on RDKit compound representations) with LaFs (CSNN) on the same compound representations. In most cases LaFs improve the Pearson correlation coefficient. The $< Y_{neigh} >$ prediction closely track the top performing method. **(d)** Illustrating the effect of a low-capacity Ridge regression model without (column: Ridge) and with LaFs (column: Ridge + N). The input dimension differs only by one column (the mean value over neighbourhood). **(e)** Interestingly, the test metrics on a given hGPCR correlates strongly with the Pearson correlation under the homphily assumption (mean value is a good prediction). The LaF methods outperform none LaF methods (compound-to-prediction architectures).

We then assembled a chemically diverse library of 539 unique compounds, with 386 sourced from the Prestwick library and filtered based on any reported cell surface interactions, and 153 sourced from Chemfaces natural product library based on a substructure search for an indole ring. We screened the library against all 7 hGPCRs, and a comparison to prior art revealed that the reduced dynamic range and sensitivity of HTR1A in yeast made this part of the dataset unreliable. Furthermore, 36 compounds were found to dampen luminescence independently from hGPCR expression (Supplementary Figure S22-S23), rendering some negative hits (Z-score<-3) meaningless. This left a slightly reduced (3,018 remaining out of 3,773 measured DTIs), but high-quality, dataset with a 98.7% overall correspondence to prior art, i.e., from mammalian reporter cell lines (Figure 4c and d, supplementary Figure).

Correlating Predictions with Experimental Values

Figure 4e and f relate the predicted labels with the experimental labels, which shows that CSNN are a viable pre-screening method to reduce the experimental load by removing non-interacting DTIs with a high specificity. In Figure 4e argmax predictions have a specificity of 0.992 for the 'No Effect' class, showing a high discriminative ability. With the $\mathrm{NN}^1_\theta 28$ label prediction model, a higher number of DTI's can be predicted due to inference considering all data in the chemical neighbourhood. Still here we found a high specificity (0.979) for the 'No Effect' class. Remarkably, however, other labels which should show a high correspondence with experimental labels ('Partial Agonist' and 'Agonist') have a low specificity and precision. Many predicted agonists have non-significant hGPCR activation. Labels such as 'Antagonist' and 'Inverse Agonist' are not expected to show any correspondence, due to the nature of the assay. Nevertheless, the methods developed can reliably be used to sort out true negative DTIs, which reduces the experimental workload to validate adverse or otherwise off-target effects significantly.

Finally, in Figure 4f, the experimental and predicted data is used to retrospectively assess the performance as a pre-screening methodology for DTI predictions. A experimentally significant 'Hit' is defined as a Z-score > 3, with the predicted label being assigned as a hit if (1) the label is not 'No Effect' and (2) the predicted K_i < 100 nM (using the regression model). The resulting confusion matrix shows that the methods are accurate for filtering out true negative DTI pairs (specificity = 0.948), but cannot reliably discriminate true positive hits (precision = 0.182). When combining the class label and regression model predictions, one can significantly enhance the DTI hit-rate (Fig. 4f), from 3.3 % to 18 %, thus reducing the experimental load to validate DTIs. This demonstrates CSNN as a useful, pre-screening methodology which can assist, but not replace, validation experimentally.

Discovering Novel DTIs

In addition to DTI predictions made by CSNN, the underlying CSN can be used to probe the degree of novelty of measured DTIs with significant RLU, as all public database knowledge is query-able and accessible in a single CSN. Using this, we filter the experimental DTIs for significant hits (|Z-score| > 3) and without any neighbourhood data on the given hGPCR. We identified 14 novel hits without prior art on the hGPCR in the chemical neighbourhood. Six of these hits are highlighted in Figure 4g, and each of these represents new information on approved drugs. Tinoridines' association with HTR4 adds indications for treating gastrointestinal disorders (*28*), whereas Nicergoline interaction with CHRM3 indicates potential uses for treating Alzheimer's disease and schizophrenia (*29*) and at the same time adds a potential mechanism of action to Nicergoline's primary use, as CHRM3 is directly tied to blood-pressure control (*30*). Similarly, the mechanism of action of Latanoprost for mitigating increased intraoccular pressure may be partially explained by the observed interaction with CHRM3. Some of these interactions also correspond well with known side-effects of the drugs, such as disturbances in the circadian rhythm associated with overdosing of Tranylcypromide, which is likely caused by the discovered interaction with MTNR1A(*31*)

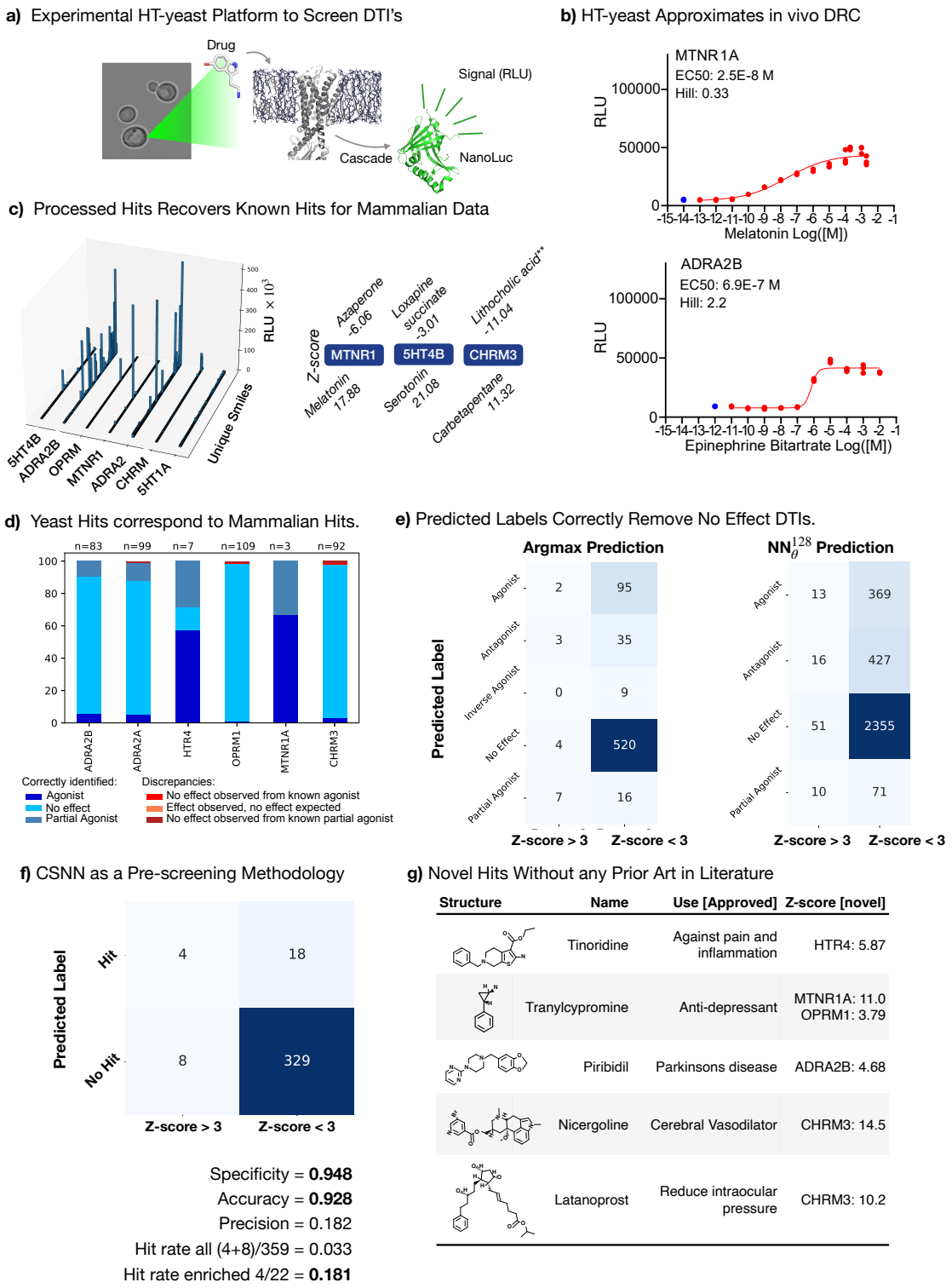


Figure 4: **Experimental Validation using a Developed HT-yeast Biosensing Platform**. **(a)** Overview of experimental HT-yeast platform based on designs by Shaw et al. 2019 (*25*). The drug may induce the hGPCR and the signal is measured by a reporter in relative luminescence units (RLU). **b** Dose-response curves (DRCs) reveal large differences in dynamic range and sensitivity (See all hGPCR DRCs in supplementary Figure S20). **(c)** Overview over all-vs-all 3773 DTIs measured (539 compounds and 7 hGPCRs, left) with top hits and their associated Z-score (right), which recovers known agonists. **(d)** Using the CSN as a knowledge graph, we compare the signal response in HT-yeast of DTI with that found in mammalian systems and find a high-correspondance. **(e)** Heatmap of predicted label and the experimental Z-score for DTIs. Argmax predictions and neural network predictions can correctly filter out no effect DTIs. Naturally, the assay cannot capture some of the predicted labels in the Z-score and thus does not correlate with the Z-score bin. **(f)** CSNN can effectively be used to enrich the positive hits (Z-score > 3), with high specificity (0.948) and an enriched hit rate of 18 %. A predicted hit was defined the CSNN prediction for K_i < 100 nM and the label is not 'No Effect'. The precision is markedly low (0.182), showing difficulty in positive hits. **(g)** Novel hits without prior art in literature. We use the CSN as a knowledge graph, sort for significant hits (|Z-score| > 3), without any database knowledge in the chemical neighbourhood. We further validate by a manual literature search. See additionally Figure S27 and Table S4.

(See Supplementary Table S4). Thus, we further demonstrated the complimentary between a robust ML method and a high-throughput yeast biosensing system to discover and validate novel DTIs.

3 Discussion

hGPCR drug-activity databases are growing at astonishing rates, yet most compound-hGPCR interactions are absent. This tension remains unresolved by ML, which is routinely used to predict activities of new compounds, but rarely addresses how to impute missing data robustly in these databases. We have developed neighbourhood-to-prediction architectures which use labels as a feature and consistently filled out the gaps among the 23 million possible compound-hGPCR interactions. We further increased coverage from 1.5 % (370 K) to 5.6 % (1.3 million) of all possible DTIs in the database, robustly using a transparent training free prediction.

ML methods, akin to database sizes, are also growing at astonishing rates, yet often with little-to-none experimental validation as to the utility of the model. Furthermore, few guard-rails are imposed in terms of OOD detection, and public web-servers will produce predictions for any SMILES compound, regardless of the distance from the training set. We document this effect here and develop CSNN to prevent such OOD predictions, in which "No neighbourhood" implies no prediction. OOD predictions are problematic for experimental scientists which readily run into null-results. We demonstrate the applicability of neighbourhood-to-prediction methods to reduce OOD predictions, increasing the reliability and robustness of ML methods for drug-discovery. LaFs are strong inductive biases in homophilous networks and enable training free predictions which are robust, accurate, and interpretable. Finally, we complemented the *in silico* DTI discovery pipeline with a synergising HT-yeast platform to uncover novel DTIs. In particular, we utilise how our method is able to predict both mode of action and binding affinity to show a Pearson correlation of 0.75 between measured DTIs in yeast and the binding affinity of predicted agonists. Furthermore, by using the chemical neighbourhoods as a knowledge graph, we validated 14 novel compound-hGPCR interactions not described in literature or public databases, tentatively linking some with adverse effects. We imaging this utility could be used to pinpoint underrepresented chemical groups for future HTS campaigns on hGPCRs, which is relevant particularly when searching for novel patentable drug candidates, and at the same time validates off-target DTIs.

Chemical space networks face significant criticism, as subtle molecular modifications can change the toxicological and pharmacological behaviours of compounds (*32*). However, our results indicate that reliable conclusions can be drawn from chemical neighbourhoods (in a training free manner) to correctly assign a bioactivity class and regression labels. This demonstrates the inherent network homophily in the hGPCR-compound space and offers efficient ways to increase coverage on the otherwise sparsely annotated drug-target interactions.

The results also link with a more fundamental discussion on whether graph neural networks require network homophily (*33*, *34*). The recent LaF framework demonstrates that advanced GNNs are on par in performance with training free GNNs, but this paradigm only works on homophilous graphs (*11*). As we show in a benchmark with pdCSM, architectures without LaFs are outperformed by a training free prediction with LaFs on 15 out of 24 targets, and that this scales with network homophily. Thus, we demonstrate the utility of CSNN in using this advantage. Work has begun characterising "good" and "bad" network heterophily and building GNNs which are more robust to these properties (*33*, *34*). This is likely an interesting future line of work in the DTI setting for reliable database driven inference to uncover unknown DTIs.

The results presented here clearly demonstrate network homophily, while general discussion about whether this is inherent in chemical space is ongoing (*9*, *35*). Publicly available data does not sample chemical space independent and identically distributed (i.i.d), and deposited data is often a result of significant derivatization efforts to increase ligand efficiency, reduce toxicological properties, and increase availability pharmacokinetically. Thus, homophily is 'over-sampled' as often derivitisation efforts maintain the bioactivity class

(e.g., Agonism), while improving other properties of the compound. This sampling bias then overestimates the true network homophily on which ML models are trained, which then retains the smoothness or locality inductive bias between the input and output spaces. This extends into train/test data splitting, leading to significant system leakage and an overestimation of model performance (*36*). CSNN models outperform non-neighbourhood methods, but imply stricter requirements for reducing system leakage and thus evaluating models fairly. In the paradigm of LaFs, it will be increasingly critical to fairly evaluate and develop robust ML systems for DTI discovery in combination with experimental validation assays to assess true performance.

We anticipate that (i) neural networks that operate on the graph structure of chemical neighbourhoods rather than the graph structure of compounds using LaFs will enhance robust identification of in-distribution DTIs and limit OOD outliers, and (ii) that an increasing number of hGPCRs will become available to yeast and expand on the robust and cost-effective yeast platform to accommodate future DTI discovery. In fact, LaFs are not entirely new to biology: Alphafold2 has a 'template' enabled mode (*37*), which can be viewed as a label of a neighbour. The label provides an approximate answer, which may be integrated to form a refined answer. By analogy, LaFs for DTI predictions provide an approximate answer for the bioactivity class and value of a DTI, which may be refined by a ML model. Point (i) is underway with established algorithmic (*15*, *38*) and neural (*39*) network-based methods, yet remains limited to paired compounds in the latter case. Here, we take a significant step toward using LaFs on an arbitrarily large number of neighbours by using a one-vs-all database search during inference by CSNN. What remains to be done are end-to-end differentiable and explainable (*40*) neighbourhood-to-predictions. Point (ii) is also well-underway, with over 70 hGPCRs already reported to signal in yeast (*41*). Neighbourhood-to-predictions architectures should additionally benefit biotechnology by rapidly identifying a suitable hGPCR for a given ligand to accelerate the development of sustainable biosynthesis of complex natural products and derivatives by utilising state-of-the-art microbial cell factories (*42*, *43*).

Acknowledgements

This project has received funding from the Novo Nordisk Foundation, grant number NNF20CC0035580, and European Union Horizon 2020 research and innovation program (grant number 814645 (MIAMi) to MKJ). We would like to thank William Shaw for providing access to his yeast platform strains, which we used to build our hGPCR signalling yeast strains. This work was further supported by Danish Data Science Academy (to NGM), which is funded by the Novo Nordisk Foundation (NNF21SA0069429) and VILLUM FONDEN (40516).

Funding

This work was supported by

The Novo Nordisk Foundation Center for Biosustainability (FGH, NGM, LGH, BL, TJ, CGAR, MKJ, EDJ)

The Danish Data Science Academy (NGM)

European Union Horizon 2020 research and innovation program (MKJ)

Author contributions

Conceptualization: FGH, NGM, MKJ, CGAR, EDJ

Methodology: FGH, NGM, BL

Investigation: FGH, NGM

Visualization: FGH, NGM

Chemical Library assembly: LGH, TJ

Funding Acquisition: MKJ, JDK, CGAR, EDJ

Supervision: MKJ, CGAR, EDJ

Writing - original draft: FGH, NGM, EDJ

Writing - review and editing: FGH, NGM, MKJ, CGAR, EDJ

Competing interests:

J.D.K., L.G.H., and M.K.J. are inventors on pending patent applications (patent applicant: Technical University of Denmark; application number: PCT/EP2023/063481). L.G.H., J.D.K. and M.K.J. have financial interests in Biomia. J.D.K. also has financial interests in Amyris, Lygos, Demetrix, Napigen, Apertor Pharmaceuticals, Maple Bio, Ansa Biotechnologies, Berkeley Yeast and Zero Acre Farms. All other authors have no competing interests.

Material and Methods

Computational methods

Graph Construction of Chemical Space Networks

The chemical space network can be formalised as a graph: $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with vertices (nodes) $\mathcal{V}$ and edges $\mathcal{E}$, which may be parameterized by the adjacency matrix $\mathbf{A} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$. Each unique compound corresponds with a unique node in $\mathcal{V}$. To obtain this adjacency matrix, a similarity metric is required between any two compounds ($sim(d_i, d_j)$, where $sim : \mathcal{V} \times \mathcal{V} \rightarrow \mathbb{R} \in [0, 1]$) associated with a real-valued vector ($\mathbf{d}_i, \mathbf{d}_j \in \mathbb{R}^p$). A natural definition is the Tanimoto similarity, which may be written in vector notation as:

$$T(d_i, d_j) = \frac{\langle \mathbf{d}_i, \mathbf{d_j} \rangle}{||\mathbf{d_i}||^2 + ||\mathbf{d_j}||^2 - \langle \mathbf{d}_i, \mathbf{d_j} \rangle} \tag{1}$$

where $\langle \mathbf{d}_i, \mathbf{d_j} \rangle$ is the inner product, and $||\mathbf{d_i}||^2$ is the Euclidean norm squared. A näive implementation would calculate all pairwise Tanimoto similarity for the set of objects in $D = \{d_1, d_2, ..., d_n\}$ (scaling as $\mathcal{O}(n^2)$). A graph could then be constructed by considering neighbours for which $sim(d_i, d_j) \geq \epsilon$. This is known as the *all-pairs similarity search* (APSS) problem, which Anastasiu et al. (*44*) have optimised by incorporating similarity bounds, thus significantly reducing the number of floating point operations. The approach, naturally, further exploits the commutative property (symmetry) of the Tanimoto similarity metric ($sim(d_i, d_j) = sim(d_j, d_i)$). Given a minimum threshold similarity (ϵ), one can exclude objects whose norm are either too small or too large to be neighbours. We found that an all-vs-all pairs similarity search was necessary to accurately construct chemical neighbourhoods as low-dimensional embedding distances did not necessarily represent chemical similarity (Figure S26).

For a similarity threshold of $\epsilon > 0.4$ we compute an all-vs-all similarity metric using Tanimoto similarity. Values less than ≤ 0.4 are set to zero to keep a sparse adjacency matrix (otherwise a computer runs into memory limitations 186K $\times$ 186K possible edges). We take inspiration from (*45*, *46*). Compounds are encoded with a 4096 bit Morgan fingerprint in RDKit (Chirality=True, radius=3) and concatenated with the 166 bit MACCS key substructure fingerprint. The bit strings undergo lossless compression by packing each successive 8 bits into a single byte while padding the end to ensure it is a multiple of eight, thus reducing the dimensionality of the input vector to $\mathbf{d_i} \in \mathbb{Z}^{533}$. As bytes may have shared bits, the Tanimoto similarity of the packed bits is always T(A',B') $\leq$ T(A,B), where A' and B' are packed bits and A and B are unpacked bits. The similarity metric is computed in batches using SimSIMD[1]. The obtained adjacency matrix is upper triangular and is made symmetric by $A_{sym} = A + A^\top$. This adjacency matrix has 187,254 unique nodes corresponding to all unique compounds in the acquired ChEMBL data with the addition of the 550 compounds for the experimental library. To query related compounds the n-hop neighbourhood is used $\mathcal{N}_j(n) = \{i \mid d(j, i) \leq n, \forall i \in \mathcal{V}\}$ where $d(j, i)$ is the graph distance between vertex i and j. Practically, we restrict lookup to the 40 nearest neighbours for the training/testing/validation data due to compute speed in constructing neighbourhoods and to a 2-hop neighbourhood for visualisation tasks.

Data partitioning for Training, Testing and Validation

For all 550 compounds in the screened chemical library, we identify all molecules within a 2-hop neighbourhood and remove any compounds with over a 0.65 Tanimoto similarity to prevent data leakage for the experimental validation. Then, remaining compounds are assigned a Murcko scaffold index, and the data is split into a 80/10/10 train/test/val split using code adapted from `https://github.com/chainer/chainer-chemistry/blob/master/chainer_`

[1] `https://github.com/ashvardanian/SimSIMD`

`chemistry/dataset/splitters/scaffold_splitter.py`. This ensures that scaffolds in the test dataset are not present during training.

For each compound in each split we explode the pandas array such that each unique compound-receptor interaction(DTI) is a separate row. This facilitates a comparison with conventional drug-target interaction methods. We then precompute a compound representation $\mathbf{x}_i$ based on a pre-trained message-passing neural network from (*47*). This numerically represents the chemical graph (MPNN at atom level with ChemProp) allowing for message-passing on the level of chemical neighbours rather than compounds. We choose learned representations as opposed to chemical fingerprints as they consistently outperform them (*40*, *48*). In the case of the benchmark in 2, we re-implement the fixed RDKit molecule representations (without feature selection) using the Supporting information from the pdCSM paper. This allows for a direct comparison, removing the effect of the compound representation. Due to lacking training and inference code, however, we could not entirely replicate their workflow. Correspondence with authors was unsuccessful (no response).

Likewise, we pre-compute target representations $\mathbf{p}_i$ using the ESM-2 protein language model and the canonical UniProt sequence (esm.pretrained.esm2_t6_8M_UR50D()) (*49*).

For each unique compound-receptor data-point (i) we prepare a graph $\mathcal{G}_i$ and label $\mathbf{y}_i$ using the accessible graph data framework by PyTorch (*16*). The graph is structured as $\mathcal{G} = \mathrm{Data}(\mathbf{x}, \mathcal{E}, \mathbf{e}, \mathbf{p}, \mathbf{y})$, where $\mathbf{x}$ are the compound representations for the chemical neighbourhood and compound i, $\mathcal{E}$ are the unweighted graph edges, $\mathbf{e}$ are the edge features, $\mathbf{p}$ is the target representation, and $\mathbf{y}$ is the label. The compound representations are $\mathbf{x} \in \mathbb{R}^{|\mathcal{V}| \times 300}$. The edges in $\mathcal{E}$ are always directed and incident on the node to be classified, as illustrated in Figure 1d. The edge features are $\mathbf{e} \in \mathbb{R}^{|\mathcal{E}| \times 7}$ which includes the one-hot encoding of the neighbouring compound bioactivity (LaF in effect) as well as the Tanimoto similarity defined in the adjacency matrix. This allows for the node-classification task to be rephrased as a graph classification task. This also enables the Label as a Feature framework delineated in Section S3. As PyTorch requires $\mathbf{p} \in \mathbb{R}^{|\mathcal{V}| \times d}$ we broadcast the target representation, which is only of dimension 1 x 320 (ESM-2, (*49*)). $\mathbf{y}$ is a one-hot encoding of the bioactivity class such that $\mathbf{y} \in 0, 1^6$. We provide the torch files dataset_train.pt, dataset_test.pt, dataset_val.pt to enable subsequent use and bench-marking of future methods (*50*). An i.i.d sampled version is also provided, the results are comparable for both splitting methods.

LaFs: Training-free argmax(.) predictions

To demonstrate the efficacy of training-free predictions using LaFs we compute the most likely label algorithmically. We sum over the all $\mathbf{e}$ for neighbouring nodes excluding the column, which represents the Tanimoto similarity, then divide by the number of edges to obtain a normalised output (frequency). The output for a given graph is a vector $\mathbf{e}_{tot} \in \mathbb{R}^6$ which the argmax($\mathbf{e}_{tot}$) may be taken and compared with the ground truth label argmax($\mathbf{y}$) to determine the performance metrics. Intuitively, if the chemical neighbourhood has a majority of agonists, the compound in question is likely also an agonist. If network homophily is the case, then training-free prediction using LaFs will be a good approximation of the true label.

In Figure S3, the metrics are computed across all graphs in test, train, and validation to obtain the predicted and true labels for each graph. As the output is deterministic, no uncertainty is reported. Then common performance metrics are calculated (See later section).

LaF: RF + N

Drug-target interaction (DTI) commonly has compound-to-prediction architectures, where a compound is represented with a numerical vector and passed through a ML architecture. In the case of simply ML models (RF or MLPs) we extend this to include LaF, by simply concatenating the compound representation,

receptor representation, and the aforementioned frequency vector of the neighbourhood (RF + N) or simply the receptor and compound (RF) to evaluate the effect of +N (LaF).

The input vector is of dimension 300 + 320 (pretrained compound representation and ESM-2 target representation) and the output is the one-hot encoded bioactivity class label $\mathbf{y} \in 0, 1^6$. We use scikit-learn (*51*) and the ensemble random forest (RF) classifier as well as a simple multi-layer perceptron (MLP) as the baseline models. A hyper-parameter grid search is carried out with 2-fold cross validation:

$$\begin{aligned}
\text{param_grid_RF} = \{&\text{'max_depth'} : [10, 20, 30, 40, 50], \\
&\text{'min_samples_leaf'} : [1, 2, 4, 8], \\
&\text{'n_estimators'} : [100, 500, 1000], \} \\
\text{param_grid_MLP} = \{&\text{'hidden_layer_sizes'} : [(1024,), (512, 64), (256, 32), (256, 32, 8)], \\
&\text{'activation'} : ['relu'], \\
&\text{'alpha'} : [1e-3, 1e-6]\}
\end{aligned}$$

and the best parameters are chosen. Fixed parameters for the RF are class_weight='balanced'. For the MLP max_iter=2000, early_stopping=True, validation_fraction=0.1, n_iter_no_change=200.

An additional neighbourhood based RF + N and MLP + N is trained using the same hyper-parameters where the input vector is additionally concatenated with the normalised $\mathbf{e}_{tot}$ presented previously. This adheres to the permutation invariance property desired and is thus a valid although crude construction.

Five independent training's are carried out to calculate the mean and standard deviation on the performance metrics, thus assessing the stability of training.

LaF: increasing the expressive power of GNNs with CSNN

The chemical space neural network (CSNN) is designed as illustrated in Figure 2d. We instantiate CSNN as a weaker form of message-passing (*52*), namely a graph convolutional neural network (GCN) with edge attributes [2]. In the first block, a local representation is computed by aggregating information from neighbouring nodes and LaF using edge attributes to form $\mathbf{r}_i$ from the local latent representations $\mathbf{x}'_i$:

$$\mathbf{r}_i = h_\Theta \left((1+\epsilon) \cdot \mathbf{x}_i + \sum_{j \in \mathcal{N}(i)} \mathrm{ReLU}(\mathbf{x}_j + \mathbf{e}_{j,i}) \right) \tag{2}$$

As the graph is directed upon the node with a missing label-to-be predicted (with only a 1-hop neighbourhood), no graph pooling is needed and a single GCN layer suffices: as the output dimension is already one-dimensional and represents a pooled representation.

We use an output dimension of 256 for the local representation using a hidden size of 256 and apply a ReLU() non-linearity. Then the local representation is concatenated with the target representation and the normalised $\mathbf{e}_{tot}$, which acts as skip connection, to form $\mathbf{r}'_i$. Finally, $\mathbf{r}'_i$ is transformed by a MLP (hidden size 512, 2 hidden layers, output dimension: 6) as:

$$\hat{\mathbf{z}}_i = \sigma\left(\mathbf{W}\mathbf{r}'_i + \mathbf{b}\right), \quad \hat{\mathbf{z}}_i \in \mathbb{R}^6 \tag{3}$$

[2] https://pytorch-geometric.readthedocs.io/en/latest/generated/torch_geometric.nn.conv.GINEConv.html#torch_geometric.nn.conv.GINEConv

to form the predictions from which the logits are used to compute the cross-entropy loss. We compute a batch-wise loss and train with a batch size of 16 for 100 epochs with the Adam optimiser (*53*), a learning rate of 0.00001 and weight decay of 0.0001:

$$\mathcal{L}_{\text{batch}} = -\frac{1}{N_{batch}} \sum_{i=1}^{N_{batch}} \sum_{c=1}^{C} y_{i,c} \log \left(\frac{e^{z_{i,c}}}{\sum_{j=1}^{C} e^{z_{i,j}}} \right) \tag{4}$$

where C is the number of classes and N_{batch} is the number of batches of batch size 16.

LaF: Multi-label and Multi-class

The previous method predicts only class-wise probabilities for a given drug-target interaction (receptor encoded by ESM-2 embedding), specified by the compound representation, target representation, and it's chemical neighbourhood. To predict class-wise probabilities for all 128 hGPCR receptors and all six classes (multi-label & multi-class), we adapt the previous method by including all available data in the edge features (128 + 1) dimension with the class index at each position and 0 if no data is present. This is expanded to a one-dimensional vector of (128 x 7 = 896) to allow for a one-hot encoding of neighbouring data for the six-classes and the no data index. We use the same skip connection as previously but a output size of the GCN as 64. The concatenated $\mathbf{r}'_i$ is of dimension 960 before transformation with a MLP. The final output is a one-dimensional vector of dimension 896 (128 x 7), which may be reshaped to a tensor of dimension 1 x 128 x 7 with a softmax applied along the third axis to obtain receptor-wise class probabilities. The true output is of dimension 1 x 128 and represents the class (in integers) for all 128 receptors. As no compounds have data for all receptors, the positions which are zero (the no data index) are masked when computing the loss on the output logits. We train with the Adam optimiser, a learning rate of 0.00005, and a weight decay of 0.0001. While inference produces class-wise predictions for all receptors, the performance metrics naturally only consider those with data available by using a mask.

3.0.1 LaF: Regression Benchmark

We source the pdCSM training and test set from `https://biosig.lab.uq.edu.au/pdcsm_gpcr/` and obtain the performance metrics on the blind test set from the supplementary information in (*17*). This was done because training and inference code was not available, neither provided upon reasonable request. Thus, performance in the comparison has pdCSM models with feature selection vs. our equivalent models (RF + N) on the reproduced compound representations (using their SI).

To uphold the transductive node classification restriction, we allowed training data to dynamically refer to each other during inference and training: labels of neighbouring compounds within the training set are used as features. For the test set, the neighbourhood graph is constructed using only training data compounds and labels as features.

Of key importance, the pdCSM method trains a single model for each receptor. Instead, we concatenate the ESM-2 representation and train a single model on all training data across all receptors. We found that pooling the data in this fashion lead to increased performance (a single model see's more data).

In Figure 3e the Pearson correlation coefficient is given per receptor (one point = one receptor) under the homophily assumption. This uses the mean-value of the labels in the neighbourhood as the prediction. Then the correlation coefficient is calculated per receptor between the predicted and true values. As illustrated, the models performance scales linearly with the degree of network homophily (for which the mean value is a good approximation). This is a highly receptor dependent effect: CHRM3 has a $\rho = 0.87$ while MTNR1A is more heterophilic $\rho = 0.42$. The receptor-wise breakdown is given in the table below:

Performance Metrics

For the compound-to-prediction methods (RF and MLP), the corresponding neighbourhood-to-prediction (RF + N, MLP + N,), and CSNN using LaFs, the performance metrics are calculated for each bioactivity class i as:

$$\text{Precision}_i = \frac{TP_i}{TP_i + FP_i} \tag{5}$$

where TP_i is the number of true positives and FP_i is the number of false positives. The recall is calculated as

$$\text{Recall}_i = \frac{TP_i}{TP_i + FN_i} \tag{6}$$

where FN_i is the number of false negatives. The F1 score is calculated as:

$$\text{F1}_i = 2 \times \frac{\text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \tag{7}$$

To summarise across all bioactivity classes, a weighted F1 score is computed:

$$\text{F1-weighted} = \sum_{i=1}^{N} w_i \times \text{F1}_i \tag{8}$$

where w_i is the proportion of true instances for class i out of all instances, and N is the number of classes. For regression tasks, the Pearson correlation, Kendall correlation, and Spearman correlation coefficient, mean average error (MAE), and mean square-root error (MSE) are calculated with sklearn.

Receptor distances and clustering

ESM-2 distance metrics were obtained by calculating the cosine similarity between all-pairs. Obtaining a distance metric from a shared chemical space between any two receptors was calculated as follows: For a given receptor pair i, j the dataframe was reduced to only consider compounds for which data was available for both receptors. Then we calculate a corrected Cramer's V statistic, which measures correlation between two variables which are categorical. If less than 5 co-occurring or no co-occurring compounds existed, the distance was set to its maximum (1). Highly correlated hGPCR receptors were assigned a value of (1 - Cramer's V), thus a distance of zero for completely correlated receptors. The phylogenetic trees in Supplementary Figure S19 were generated from distances calculated between all 128 selected receptors based on; 1. ESM-LLM distance metrics (*49*), 2. from the chemical space they are able/unable to sense, and 3. from a multiple sequence alignment using muscle 5.1(*54*). Three distinct trees were then generated using the Bio.Phylo package(*55*), and visualised using ETE V3(*56*).

Z-score Analysis

For each molecule-receptor pair a single-point relative luminescence unit (RLU) was measured as described above, then $\log_{10}$ transformed, subtracted by the mean (μ) RLU for the given receptor, and then divided by the standard deviation σ for the given receptor according to:

$$\text{Z} - \text{score} = \frac{x_i - \mu}{\sigma} \tag{9}$$

where x_i is the $\log_{10}$ raw RLU signal. $\log_{10}$ transformation prior to Z-score analysis was necessary to attain a Gaussian distribution, which is amenable to Z-score analysis.

To rescue outliers a modified Z-score was used, as defined by:

$$\mathrm{Z-score}_{mod} = 0.6745 \cdot \frac{x_i - \mathrm{m}}{\mathrm{MAD}}, \quad MAD = \mathrm{med}(|x_i - \mathrm{m}|), \quad \mathrm{m} = \mathrm{med}(X) \tag{10}$$

where x_i is the log-transformed RLU signal, MAD is the median absolute deviation, and m is the median value for a given receptor. The effect of the modified Z-score is qualitatively shown in Supplementary Figure S24, which illustrates a broadening of the kernel density estimate fit with respect to the canonical Z-score. The figure further shows the cut-off for the hit/no-hit threshold which is set at $\mid \mathrm{Z-score} \mid > 3$.

Experimental methods

Chemical library

386 compounds were sourced from the Prestwick library, selected by reported receptor associations and chemical diversity, stock concentrations at 10mM. 153 compounds were sourced from Chemfaces chemical library, selected by a substructure search for the indole ring moiety, stock concentrations acquired at 1mM. All chemicals (Supplementary File S1) had a reported purity of 95% or higher

Gene Synthesis

All biosynthetic genes used in the current study are listed in Supplementary Table S7. Genes were synthesized by IDT (San Diego, USA) and if stated, optimized for expression in S. cerevisiae.

Cloning and yeast transformation

All plasmids were constructed by USER cloning (*57*) and propagated in E. coli DH5α competent cells. Genes for integrative Easy Clone - marker free vectors were prepared according to Jessop-Fabre et al. (*58*) and the gRNA expression cassettes as previously described (*59*). All plasmids were verified by Sanger sequencing prior to yeast transformation. The integrative plasmids were linearized by NotI enzyme (New England Biolabs, MA, USA). The yeast transformations were done using standard Li-acetate methods according to Gietz and Schiestl (*60*). The integration of heterologous genes was verified as described in CasEMBLR (*58*, *59*), The strains generated in this study are listed in supplementary Table S8.

Yeast strains and cultivation

Yeast strains ScS1-7 expresses NanoLuc upon stimulation of their respective human G-protein coupled receptors. The parental strains ScFH236 and ScFH237 were generated by changing the sfGFP reporter to Nano-Luciferase (NanoLuc) in the platform strains yWS2266 and yWS2267 published by W. Shaw et al. (*25*). This was achieved by transformation of a transient Cas9 plasmid (pCfB5270), transient sfGFP specific gRNA plasmid (pFH69), and a repair template with homology to the upstream promoter and downstream part of sfGFP, containing the NanoLuc coding sequence and the tCYC1 terminator, which was linearised from a plasmid vector (pHT9). ScS1,2,3,5, 6, and 7 were generated by integrating 5HTR1A, 5HTR4, ADRA2A, OPRM1, CHRM3, or MTNR1A, respectively into ScFH237, ScS4 was generated by integrating ADRA2B into ScFH236. This design combines advantages from two of the most well known yeast GPCR biosensors, described in Kapolka et al. and Miettinen et al. (*18*, *19*). To generate biomass for the assays, yeast was inoculated from cryostock to 5 mL SC-trp and incubated ON at 30 °C 250 rpm. The next day, 20 hours before the assay, 5ml of saturated culture was added to 45ml SC-trp in a 500ml shakeflask and incubated overnight 30C 200RPM. Next 25 mL culture was centrifuged at 5000 x g for 5 min in 50ml falcon tubes,

and the supernatant was discarded. The pellet was then resuspended in pH buffered SC-trp at pH 7.2 (*61*). OD600 was measured and adjusted to 5 before continuing with the procedures described in "Bioactivity assays"

Chemical preparations

For the DRCs shown in Figure 4d and supplemantary Figure S20, DAMGO-Enkephalin acetate salt, Serotonin HCL, Acetylcholine chloride, Epinephrine bi-tartrate and Melatonin were acquired from Sigma-Aldrich and 10X stocks of each were prepared in sterile mq water containing 10% DMSO, in the concentrations listed in supplementary file S3, including a no-ligand control of only 10%DMSO in sterile Milli-Q water. The ChemFaces and Prestwick chemical libraries were acquired pre-solubilised in DMSO at 1mM and 10mM respectively. 10X stocks were prepared in 96 well plates (Greiner 655160) by diluting a fraction of the supplied stock to 100μM in sterile mq water, with a final DMSO concentration of 10%. One positive control and three negative control wells were reserved on all plates containing the chemical libraries.

Bioactivity assays

For the chemical library screen, all 10X stocks aswell as a no-ligand control were diluted 1:10 in 96 well plates (Greiner 655160) in cell cultures prepared according to "yeast strains and cultivation", resulting in 1 well pr. receptor / compound combination with 100μL of sensing cells and 10μM of compounds. A gap of 4 minutes between loading of each plate was maintained to allow sequential measurements in a SynergyMX microtiter plate reader (BioTek). For the DRCs shown in Figure 4, 3 replicates of each concentration of ligands with relevant receptors were prepared aswell as a no-receptor strain control for each ligand. For supplementary Figure S20, a 1% v/v final concentration of DMSO control was run in triplicate along with triplicates of each of the selected compounds from our chemical library at 10μM concentrations (1% v/v DMSO). All plates were covered with a breathe-easy breathable membrane and incubated for 4 h at 30 °C 300 rpm. Twenty minutes before incubation was completed, a lysis mix of 1.33% v/v Furimazine stock (NanoLuc substrate) from Promega in CelLyticTM Y cell lysis reagent (Sigma-Aldrich) was mixed and 12 μL was distributed to wells of white small volume 96-well plates (Greiner 675083). Following incubation, the plates were vortexed for 5 sec. before the membrane was removed and 4 μL of each well containing sensing cells was transferred to the plate containing the lysis mix and the plate was incubated at RT for 18 minutes and then placed in a SynergyMX microtiter plate reader (BioTek). The luminescence was measured at 3 time points for the chemical library screen: 18.5, 20, and 21.5 min. after mixing, and 2 time points for the DRCs: 20 and 21 min. after mixing. The settings were filter-luminescence, with gain 150 and 0.5 sec. integration pr. well, the average of each of those time points is reported as single replicates pr. well.

Dose response curve calculation

All DRCs shown in Figure 4d and supplementary Figure S20 were generated in Graphpad Prism 9.5.0. The models chosen for HTR4, ADRA2A, ADRA2B and MTNR1A, were all standard 4-parameter non-linear regressions. The models chosen for HTR1A and CHRM3 were a bell shaped DRC, and a bi-phasic DRC respectively. Since the DRC of OPRM1 did not plateau at the higher concentrations, a 5-parameter non-linear regression was applied for the best fit for visualising the trend. All additional information related to the fitted curves is available in supplementary file S3.

References

1. D. B. Catacutan, J. Alexander, A. Arnold, J. M. Stokes, Machine learning in preclinical drug discovery. en, *Nat Chem Biol* **20**, Publisher: Nature Publishing Group, 960–973, ISSN: 1552-4469, (2024; https://www.nature.com/articles/s41589-024-01679-1) (Aug. 2024).
2. J. Vamathevan, D. Clark, P. Czodrowski, I. Dunham, E. Ferran, G. Lee, B. Li, A. Madabhushi, P. Shah, M. Spitzer, S. Zhao, Applications of machine learning in drug discovery and development. en, *Nat Rev Drug Discov* **18**, Publisher: Nature Publishing Group, 463–477, ISSN: 1474-1784, (2024; https://www.nature.com/articles/s41573-019-0024-5) (June 2019).
3. K. Alhosaini, A. Azhar, A. Alonazi, F. Al-Zoghaibi, GPCRs: The most promiscuous druggable receptor of the mankind. *Saudi Pharmaceutical Journal* **29**, 539–551 (2021).
4. K. Sriram, P. A. Insel, G Protein-Coupled Receptors as Targets for Approved Drugs: How Many Targets and How Many Drugs? *Mol. Pharmacol.* **93**, 251–258 (2018).
5. P. A. Insel, C.-M. Tang, I. Hahntow, M. C. Michel, Impact of GPCRs in clinical medicine: monogenic diseases, genetic variants and drug targets. *Biochim. Biophys. Acta* **1768**, 994–1005 (2007).
6. Z. Wu, W. Li, G. Liu, Y. Tang, Network-Based Methods for Prediction of Drug-Target Interactions. *Front. Pharmacol.* **9**, 1134 (2018).
7. I. M. Jamu, H. Okamoto, Recent advances in understanding adverse effects associated with drugs targeting the serotonin receptor, 5-HT GPCR. eng, *Front Glob Womens Health* **3**, 1012463, ISSN: 2673-5059 (2022).
8. D. Yang, Q. Zhou, V. Labroska, S. Qin, S. Darbalaei, Y. Wu, E. Yuliantie, L. Xie, H. Tao, J. Cheng, Q. Liu, S. Zhao, W. Shui, Y. Jiang, M.-W. Wang, G protein-coupled receptors: structure- and function-based drug discovery. *Sig Transduct Target Ther* **6**, 1–27 (2021).
9. M. Vogt, D. Stumpfe, G. M. Maggiora, J. Bajorath, Lessons learned from the design of chemical space networks and opportunities for new applications. en, *J Comput Aided Mol Des* **30**, 191–208, ISSN: 1573-4951, (2024; https://doi.org/10.1007/s10822-016-9906-3) (Mar. 2016).
10. M. McPherson, L. Smith-Lovin, J. M. Cook, Birds of a Feather: Homophily in Social Networks. en, *Annual Review of Sociology* **27**, Publisher: Annual Reviews, 415–444, ISSN: 0360-0572, 1545-2115, (2024; https://www.annualreviews.org/content/journals/10.1146/annurev.soc.27.1.415) (Aug. 2001).
11. R. Sato, *Training-free Graph Neural Networks and the Power of Labels as Features*, Aug. 2024, (2024; http://arxiv.org/abs/2404.19288).
12. D. Mendez, A. Gaulton, A. P. Bento, J. Chambers, M. De Veij, E. Félix, M. P. Magariños, J. F. Mosquera, P. Mutowo, M. Nowotka, M. Gordillo-Marañón, F. Hunter, L. Junco, G. Mugumbate, M. Rodriguez-Lopez, F. Atkinson, N. Bosc, C. J. Radoux, A. Segura-Cabrera, A. Hersey, A. R. Leach, ChEMBL: towards direct deposition of bioassay data. *Nucleic acids research* **47**, D930–D940, ISSN: 0305-1048 (2019).
13. S. D. Harding, J. F. Armstrong, E. Faccenda, C. Southan, S. P. H. Alexander, A. P. Davenport, M. Spedding, J. A. Davies, The IUPHAR/BPS Guide to PHARMACOLOGY in 2024. *Nucleic Acids Research* **52**, D1438–D1449 (2023).
14. Y.-C. Lo, S. Senese, C.-M. Li, Q. Hu, Y. Huang, R. Damoiseaux, J. Z. Torres, Large-scale chemical similarity networks for target profiling of compounds identified in cell-based chemical screens. *PLoS Comput. Biol.* **11**, e1004153 (2015).
15. Z. Wu, H. Ma, Z. Liu, L. Zheng, Z. Yu, S. Cao, W. Fang, L. Wu, W. Li, G. Liu, J. Huang, Y. Tang, wSDTNBI: a novel network-based inference method for virtual screening. *Chem. Sci.* **13**, 1060–1079 (2022).
16. M. Fey, J. E. Lenssen, *Fast Graph Representation Learning with PyTorch Geometric*, 2019.
17. J. P. L. Velloso, D. B. Ascher, D. E. V. Pires, pdCSM-GPCR: predicting potent GPCR ligands with graph-based signatures. *Bioinformatics Advances* **1**, ISSN: 2635-0041, (2024; https://doi.org/10.1093/bioadv/vbab031) (Jan. 2021).
18. N. J. Kapolka, G. J. Taghon, J. B. Rowe, W. M. Morgan, J. F. Enten, N. A. Lambert, D. G. Isom, DCyFIR: a high-throughput CRISPR platform for multiplexed G protein-coupled receptor profiling and ligand discovery. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 13117–13126 (2020).
19. K. Miettinen, N. Leelahakorn, A. Almeida, Y. Zhao, L. R. Hansen, I. E. Nikolajsen, J. B. Andersen, M. Givskov, D. Staerk, S. Bak, S. C. Kampranis, A GPCR-based yeast biosensor for biomedical, biotechnological, and point-of-use cannabinoid determination. *Nat. Commun.* **13**, 3664 (2022).
20. I. Erlenbach, E. Kostenis, C. Schmidt, F. F. Hamdan, M. H. Pausch, J. Wess, Functional expression of M(1), M(3) and M(5) muscarinic acetylcholine receptors in yeast. *J. Neurochem.* **77**, 1327–1337 (2001).
21. Brown, A.J., Dyos, S.L., Whiteway, M.S., White, J.H.M., Watson, M.-A.E.A., Marzioch, M., Clare, J.J., Cousens, D.J., Paddon, C., Plumpton, C., Romanos, M.A. and Dowell, S.J., Functional coupling of mammalian receptors to the yeast mating pathway using novel yeast/mammalian G protein a-subunit chimeras. *Yeast* **16**, 11–22 (2000).

22. A. M. Ehrenworth, T. Claiborne, P. Peralta-Yahya, Medium-Throughput Screen of Microbially Produced Serotonin via a G-Protein-Coupled Receptor-Based Sensor. *Biochemistry* **56**, 5471–5475 (2017).

23. G. Gaibelet, E. Meilhoc, J. Riond, I. Saves, T. Exner, L. Liaubet, B. Nürnberg, J.-M. Masson, L. J. Emorine, Nonselective coupling of the human μ-opioid receptor to multiple inhibitory G-protein isoforms. *European Journal of Biochemistry* **261**, 517–523 (1999).

24. T. Kokkola, M. A. Watson, J. White, S. Dowell, S. M. Foord, J. T. Laitinen, Mutagenesis of human Mel1a melatonin receptor expressed in yeast reveals domains important for receptor function. *Biochem. Biophys. Res. Commun.* **249**, 531–536 (1998).

25. W. M. Shaw, H. Yamauchi, J. Mead, G.-O. F. Gowers, D. J. Bell, D. Öling, N. Larsson, M. Wigglesworth, G. Ladds, T. Ellis, Engineering a Model Cell for Rational Tuning of GPCR Signaling. *Cell* **177**, 782–796.e27 (2019).

26. C. G. England, E. B. Ehlerding, W. Cai, NanoLuc: A Small Luciferase Is Brightening Up the Field of Bioluminescence. *Bioconjugate Chemistry* **27**, 1175–1187 (2016).

27. I. Fish, A. Stößel, K. Eitel, C. Valant, S. Albold, H. Huebner, D. Möller, M. J. Clark, R. K. Sunahara, A. Christopoulos, B. K. Shoichet, P. Gmeiner, Structure-Based Design and Discovery of New M2 Receptor Agonists. *J. Med. Chem.* **60**, 9239–9250 (2017).

28. E. J. Jeong, S. Y. Chung, H. N. Hong, S.-W. Oh, J. Y. Sim, The novel, potent and highly selective 5-HT4 receptor agonist YH12852 significantly improves both upper and lower gastrointestinal motility. *Br. J. Pharmacol.* **175**, 485–500 (2018).

29. S. P. Moran, J. Maksymetz, P. J. Conn, Targeting Muscarinic Acetylcholine Receptors for the Treatment of Psychiatric and Neurological Disorders. *Trends Pharmacol. Sci.* **40**, 1006–1020 (2019).

30. A. W. Cowley Jr, Chrm3 Gene and M3 Muscarinic Receptors Contribute to Salt-Sensitive Hypertension. *Hypertension* **72**, 588–591 (2018).

31. S. Rivara, M. Mor, A. Bedini, G. Spadoni, G. Tarzia, Melatonin receptor agonists: SAR and applications to the treatment of sleep-wake disorders. *Curr. Top. Med. Chem.* **8**, 954–968 (2008).

32. R. Kunimoto, J. Bajorath, Combining Similarity Searching and Network Analysis for the Identification of Active Compounds. *ACS Omega* **3**, 3768–3777 (2018).

33. J. Zhu, Y. Yan, L. Zhao, M. Heimann, L. Akoglu, D. Koutra, *Beyond Homophily in Graph Neural Networks: Current Limitations and Effective Designs*, Oct. 2020, (2024; `http://arxiv.org/abs/2006.11468`).

34. Y. Ma, X. Liu, N. Shah, J. Tang, *Is Homophily a Necessity for Graph Neural Networks?*, en, July 2023, (2024; `http://arxiv.org/abs/2106.06134`).

35. D. Stumpfe, H. Hu, J. Bajorath, Evolving Concept of Activity Cliffs. *ACS Omega* **4**, Publisher: American Chemical Society, 14360–14368, (2024; `https://doi.org/10.1021/acsomega.9b02221`) (Sept. 2019).

36. J. Durairaj, Y. Adeshina, Z. Cao, X. Zhang, V. Oleinikovas, T. Duignan, Z. McClure, X. Robin, D. Kovtun, E. Rossi, G. Zhou, S. Veccham, C. Isert, Y. Peng, P. Sundareson, M. Akdel, G. Corso, H. Stärk, Z. Carpenter, M. Bronstein, E. Kucukbenli, T. Schwede, L. Naef, *PLINDER: The protein-ligand interactions dataset and evaluation resource*, en, July 2024, (2024; `https://www.biorxiv.org/content/10.1101/2024.07.17.603955v1`).

37. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, Publisher: Nature Publishing Group, 583–589, ISSN: 1476-4687, (2024; `https://www.nature.com/articles/s41586-021-03819-2`) (Aug. 2021).

38. F. Cheng, C. Liu, J. Jiang, W. Lu, W. Li, G. Liu, W. Zhou, J. Huang, Y. Tang, Prediction of drug-target interactions and drug repositioning via network-based inference. *PLoS Comput. Biol.* **8**, e1002503 (2012).

39. K. Amara, R. Rodríguez-Pérez, J. Jiménez-Luna, Explaining compound activity predictions with a substructure-aware loss for graph neural networks. *Journal of Cheminformatics* **15**, 67 (2023).

40. J. Jiménez-Luna, F. Grisoni, G. Schneider, Drug discovery with explainable artificial intelligence. *Nat Mach Intell* **2**, 573–584 (2020).

41. M. Deichmann, F. G. Hansson, E. D. Jensen, Yeast-based screening platforms to understand and improve human health: (Trends in Biotechnology, in press, 2024). en, *Trends Biotechnol.* (May 2024).

42. S. A. Bradley, B. J. Lehka, F. G. Hansson, K. B. Adhikari, D. Rago, P. Rubaszka, A. K. Haidar, L. Chen, L. G. Hansen, O. Gudich, K. Giannakou, B. Lengger, R. T. Gill, Y. Nakamura, T. D. de Bernonville, K. Koudounas, D. Romero-Suarez, L. Ding, Y. Qiao, T. M. Frimurer, A. A. Petersen, S. Besseau, S. Kumar, N. Gautron, C. Melin, J. Marc, R. Jeanneau, S. E. O'Connor, V. Courdavault, J. D. Keasling, J. Zhang, M. K. Jensen, Biosynthesis of natural and halogenated plant monoterpene indole alkaloids in yeast. *Nat Chem Biol* **19**, 1551–1560 (2023).

43. E. D. Jensen, M. Deichmann, X. Ma, R. U. Vilandt, G. Schiesaro, M. B. Rojek, B. Lengger, L. Eliasson, J. M. Vento, D. Durmusoglu, S. P. Hovmand, I. Al'Abri, J. Zhang, N. Crook, M. K. Jensen, Engineered cell differentiation and sexual reproduction in probiotic and mating yeasts. *Nat. Commun.* **13**, 6201 (2022).

44. D. C. Anastasiu, G. Karypis, Efficient identification of Tanimoto nearest neighbors. *Int J Data Sci Anal* **4**, 153–172 (2017).

45. V. F. Scalfani, V. D. Patel, A. M. Fernandez, Visualizing chemical space networks with RDKit and NetworkX. *Journal of Cheminformatics* **14**, 87 (2022).

46. H. Matter, C. Buning, D. D. Stefanescu, S. Ruf, G. Hessler, Using Graph Databases to Investigate Trends in Structure–Activity Relationship Networks. *J. Chem. Inf. Model.* **60**, 6120–6134 (2020).

47. A. Dalkıran, A. Atakan, A. S. Rifaioğlu, M. J. Martin, R. Ç. Atalay, A. C. Acar, T. Doğan, V. Atalay, Transfer learning for drug–target interaction prediction. *Bioinformatics* **39**, i103–i110 (2023).

48. K. Yang, K. Swanson, W. Jin, C. Coley, P. Eiden, H. Gao, A. Guzman-Perez, T. Hopper, B. Kelley, M. Mathea, A. Palmer, V. Settels, T. Jaakkola, K. Jensen, R. Barzilay, Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **59**, 3370–3388 (2019).

49. Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, A. Rives, Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).

50. F. G. Hansson, N. G. Madsen, Harnessing Chemical Space Neural Networks to Systematically Annotate GPCR ligands. (2024).

51. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

52. M. M. Bronstein, J. Bruna, T. Cohen, P. Velickovic, Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges. *CoRR* **abs/2104.13478**, arXiv: `2104.13478`, (`arxiv.org/abs/2104.13478`) (2021).

53. D. Kingma, J. Ba, presented at the International Conference on Learning Representations (ICLR).

54. R. C. Edgar, MUSCLE v5 enables improved estimates of phylogenetic tree confidence by ensemble bootstrapping. *bioRxiv* (2021).

55. E. Talevich, B. M. Invergo, P. J. A. Cock, B. A. Chapman, Bio.Phylo: a unified toolkit for processing, analyzing and visualizing phylogenetic trees in Biopython. *BMC Bioinformatics* **13**, 209 (2012).

56. J. Huerta-Cepas, F. Serra, P. Bork, ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).

57. H. H. Nour-Eldin, B. G. Hansen, M. H. H. Nørholm, J. K. Jensen, B. A. Halkier, Advancing uracil-excision based cloning towards an ideal technique for cloning PCR fragments. en, *Nucleic Acids Res.* **34**, e122 (2006).

58. M. M. Jessop-Fabre, T. Jakočiūnas, V. Stovicek, Z. Dai, M. K. Jensen, J. D. Keasling, I. Borodina, EasyClone-MarkerFree: A vector toolkit for marker-less integration of genes into Saccharomyces cerevisiae via CRISPR-Cas9. en, *Biotechnol. J.* **11**, 1110–1117 (2016).

59. T. Jakočiūnas, I. Bonde, M. Herrgård, S. J. Harrison, M. Kristensen, L. E. Pedersen, M. K. Jensen, J. D. Keasling, Multiplex metabolic pathway engineering using CRISPR/Cas9 in Saccharomyces cerevisiae. en, *Metab. Eng.* **28**, 213–222 (2015).

60. R. D. Gietz, R. H. Schiestl, Large-scale high-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. en, *Nat. Protoc.* **2**, 38–41 (2007).

61. R. C. Prins, S. Billerbeck, A buffered media system for yeast batch culture growth. *BMC Microbiol.* **21**, 127 (2021).

62. N. Amoroso, N. Gambacorta, F. Mastrolorito, M. V. Togo, D. Trisciuzzi, A. Monaco, E. Pantaleo, C. D. Altomare, F. Ciriaco, O. Nicolotti, Making sense of chemical space network shows signs of criticality. *Scientific Reports* (2023).

63. J. Yi, S. Lee, S. Lim, C. Cho, Y. Piao, M. Yeo, D. Kim, S. Kim, S. Lee, Exploring chemical space for lead identification by propagating on chemical similarity network. *Computational and Structural Biotechnology Journal* **21**, 4187–4195 (2023).

64. W. Fan, S. Wang, J. Huang, Z. Chen, Y. Song, W. Tang, H. Mao, H. Liu, X. Liu, D. Yin, Q. Li, *Graph Machine Learning in the Era of Large Language Models (LLMs)*, arXiv:2404.14928 [cs], June 2024, (2024; `http://arxiv.org/abs/2404.14928`).

65. P. Chandak, K. Huang, M. Zitnik, Building a knowledge graph to enable precision medicine. en, *Sci Data* **10**, Publisher: Nature Publishing Group, 67, ISSN: 2052-4463, (2024; `https://www.nature.com/articles/s41597-023-01960-3`) (Feb. 2023).

66. A. Santos, A. R. Colaço, A. B. Nielsen, L. Niu, M. Strauss, P. E. Geyer, F. Coscia, N. J. W. Albrechtsen, F. Mundt, L. J. Jensen, M. Mann, A knowledge graph to interpret clinical proteomics data. en, *Nat Biotechnol* **40**, Publisher: Nature Publishing Group, 692–702, ISSN: 1546-1696, (2024; `https://www.nature.com/articles/s41587-021-01145-6`) (May 2022).

67. R. T. Sousa, S. Silva, C. Pesquita, Explaining protein–protein interactions with knowledge graph-based semantic similarity. *Computers in Biology and Medicine* **170**, 108076, ISSN: 0010-4825, (2024; https://www.sciencedirect.com/science/article/pii/S0010482524001604) (Mar. 2024).

68. Z. Lyu, Y. Wu, J. Lai, M. Yang, C. Li, W. Zhou, Knowledge Enhanced Graph Neural Networks for Explainable Recommendation. *IEEE Transactions on Knowledge and Data Engineering* **35**, Conference Name: IEEE Transactions on Knowledge and Data Engineering, 4954–4968, ISSN: 1558-2191, (2024; https://ieeexplore.ieee.org/document/9681226) (May 2023).

69. L. Werner, N. Layaïda, P. Genevès, S. Chlyah, *Knowledge Enhanced Graph Neural Networks for Graph Completion*, arXiv:2303.15487 [cs], Aug. 2023, (2024; http://arxiv.org/abs/2303.15487).

70. T. Beckers, *An Introduction to Gaussian Process Models*, en, arXiv:2102.05497 [cs, eess], Feb. 2021, (2024; http://arxiv.org/abs/2102.05497).

71. A. de Falco, Z. Dezso, F. Ceccarelli, L. Cerulo, A. Ciaramella, M. Ceccarelli, Adaptive one-class Gaussian processes allow accurate prioritization of oncology drug targets. *Bioinformatics* **37**, 1420–1427, ISSN: 1367-4803, (2024; https://doi.org/10.1093/bioinformatics/btaa968) (June 2021).

72. T. van Laarhoven, S. B. Nabuurs, E. Marchiori, Gaussian interaction profile kernels for predicting drug–target interaction. *Bioinformatics* **27**, 3036–3043, ISSN: 1367-4803, (2024; https://doi.org/10.1093/bioinformatics/btr500) (Nov. 2011).

73. M. Fontana, G. Zeni, S. Vantini, Conformal Prediction: a Unified Review of Theory and New Challenges. en, *Bernoulli* **29**, arXiv:2005.07972 [cs, econ, stat], ISSN: 1350-7265, (2024; http://arxiv.org/abs/2005.07972) (Feb. 2023).

74. S. Arvidsson McShane, U. Norinder, J. Alvarsson, E. Ahlberg, L. Carlsson, O. Spjuth, CPSign: conformal prediction for cheminformatics modeling. *Journal of Cheminformatics* **16**, 75, ISSN: 1758-2946, (2024; https://doi.org/10.1186/s13321-024-00870-9) (June 2024).

75. S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang, E. E. Bolton, PubChem 2023 update. *Nucleic Acids Res.* **51**, D1373–D1380 (2023).

76. P.-K. Li, G.-H. Chu, M. L. Gillen, T. Parekh, P. A. Witt-Enderby, The development of a charged melatonin receptor ligand. *Bioorg. Med. Chem. Lett.* **7**, 2409–2414 (1997).

77. J. Saxena, F. Nicolini, M. Lockley, Abstract 3801: Reversing chemoresistance in high grade serous ovarian cancer. en, *Cancer Res.* **79**, 3801–3801 (2019).

78. A. Vardanian, *SimSIMD*, github.com/ashvardanian/SimSIMD, 2023.

79. B. Zdrazil, E. Felix, F. Hunter, E. J. Manners, J. Blackshaw, S. Corbett, M. de Veij, H. Ioannidis, D. M. Lopez, J. F. Mosquera, M. P. Magarinos, N. Bosc, R. Arcila, T. Kizilören, A. Gaulton, A. P. Bento, M. F. Adasme, P. Monecke, G. A. Landrum, A. R. Leach, The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Research* **52**, D1180–D1192 (2024).

80. M. N. Davies, A. Secker, A. A. Freitas, M. Mendao, J. Timmis, D. R. Flower, On the hierarchical classification of G protein-coupled receptors. *Bioinformatics* **23**, 3113–3118 (2007).

81. L. van der Maaten, G. Hinton, Visualizing Data using t-SNE. *Journal of Machine Learning Research* **9**, 2579–2605 (2008).

82. D. Probst, J.-L. Reymond, Visualization of very large high-dimensional data sets as minimum spanning trees. *J. Cheminform.* **12**, 12 (2020).

83. J. M. Remington, K. T. McKay, N. B. Beckage, J. B. Ferrell, S. T. Schneebeli, J. Li, GPCRLigNet: rapid screening for GPCR active ligands using machine learning. *Journal of Computer-Aided Molecular Design* **37**, 147–156, ISSN: 15734951 (3 Mar. 2023).

84. Z. Yu, Z. Wu, W. Li, G. Liu, Y. Tang, ADENet: a novel network-based inference method for prediction of drug adverse events. *Brief. Bioinform.* **23** (2022).

85. F. Fiorino, B. Severino, E. Magli, A. Ciano, G. Caliendo, V. Santagada, F. Frecentese, E. Perissutti, 5-HT(1A) receptor: an old target as a new attractive tool in drug discovery from central nervous system to cancer. *J. Med. Chem.* **57**, 4407–4426 (2014).

86. K. Köhler-Forsberg, B. Ozenne, S. V. Larsen, A. S. Poulsen, E. B. Landman, V. H. Dam, C.-T. Ip, A. Jørgensen, C. Svarer, G. M. Knudsen, V. G. Frokjaer, M. B. Jørgensen, Concurrent anxiety in patients with major depression and cerebral serotonin 4 receptor binding. A NeuroPharm-1 study. *Transl. Psychiatry* **12**, 273 (2022).

87. M. Ono, E. Takamura, K. Shinozaki, T. Tsumura, T. Hamano, Y. Yagi, K. Tsubota, Therapeutic effect of cevimeline on dry eye in patients with Sjögren's syndrome: a randomized, double-blind clinical study. *Am. J. Ophthalmol.* **138**, 6–17 (2004).

88. A. K. Paul, C. M. Smith, M. Rahmatullah, V. Nissapatorn, P. Wilairatana, M. Spetea, N. Gueven, N. Dietis, Opioid Analgesia and Opioid-Induced Adverse Effects: A Review. *Pharmaceuticals* **14** (2021).

89. Z. P. Khan, C. N. Ferguson, R. M. Jones, alpha-2 and imidazoline receptor agonists. Their pharmacology and therapeutic role. *Anaesthesia* **54**, 146–165 (1999).

90. T. Wang, Z. Li, M. E. Cvijic, L. Zhang, C. S. Sum, *Measurement of cAMP for Gαs- and Gαi Protein-Coupled Receptors (GPCRs)* (Eli Lilly & Company and the National Center for Advancing Translational Sciences, 2017).

91. C. Bergsdorf, C. Kropp-Goerkis, I. Kaehler, L. Ketscher, U. Boemer, K. Parczyk, B. Bader, A one-day, dispense-only IP-One HTRF assay for high-throughput screening of Galphaq protein-coupled receptors: towards cells as reagents. *Assay Drug Dev. Technol.* **6**, 39–53 (2008).

92. Q. Ma, L. Ye, H. Liu, Y. Shi, N. Zhou, An overview of Ca2+ mobilization assays in GPCR drug discovery. *Expert Opin. Drug Discov.* **12**, 511–523 (2017).

93. R. H. Oakley, C. C. Hudson, R. D. Cruickshank, D. M. Meyers, R. E. Payne Jr, S. M. Rhem, C. R. Loomis, The cellular distribution of fluorescently labeled arrestins provides a robust, sensitive, and universal assay for screening G protein-coupled receptors. *Assay Drug Dev. Technol.* **1**, 21–30 (2002).

94. J. Wang, X. Xie, Development of a quantitative, cell-based, high-content screening assay for epidermal growth factor receptor modulators. *Acta Pharmacol. Sin.* **28**, 1698–1704 (2007).

95. P. G. Strange, Use of the GTPγS ([35S]GTPγS and Eu-GTPγS) binding assay for analysis of ligand potency and efficacy at G protein-coupled receptors. *Br. J. Pharmacol.* **161**, 1238–1249 (2010).

96. D. B. Bylund, J. D. Deupree, M. L. Toews, Radioligand-binding methods for membrane preparations and intact cells. *Methods Mol. Biol.* **259**, 1–28 (2004).

97. D. B. Bylund, M. L. Toews, Radioligand binding methods: practical guide and tips. *Am. J. Physiol.* **265**, L421–9 (1993).

98. H. Chen, C. E. Rosen, J. A. González-Hernández, D. Song, J. Potempa, A. M. Ring, N. W. Palm, Highly multiplexed bioactivity screening reveals human and microbiota metabolome-GPCRome interactions. *Cell* **186**, 3095–3110 (2023).

99. J. Vucicevic, T. Srdic-Rajic, M. Pieroni, J. M. M. Laurila, V. Perovic, S. Tassini, E. Azzali, G. Costantino, S. Glisic, D. Agbaba, M. Scheinin, K. Nikolic, M. Radi, N. Veljkovic, A combined ligand- and structure-based approach for the identification of rilmenidine-derived compounds which synergize the antitumor effects of doxorubicin. *Bioorg. Med. Chem.* **24**, 3174–3183 (2016).

100. O. Russo, M. Cachard-Chastel, C. Rivière, M. Giner, J.-L. Soulier, M. Berthouze, T. Richard, J.-P. Monti, S. Sicsic, F. Lezoualc'h, I. Berque-Bestel, Design, synthesis, and biological evaluation of new 5-HT4 receptor agonists: application as amyloid cascade modulators and potential therapeutic utility in Alzheimer's disease. *J. Med. Chem.* **52**, 2214–2225 (2009).

101. J. S. D. Kumar, V. J. Majo, S.-C. Hsiung, M. S. Millak, K.-P. Liu, H. Tamir, J. Prabhakaran, N. R. Simpson, R. L. Van Heertum, J. J. Mann, R. V. Parsey, Synthesis and in vivo validation of [O-methyl-11C]2-{4-[4-(7-methoxynaphthalen-1-yl)piperazin- 1-yl]butyl}-4-methyl-2H-[1,2,4]triazine-3,5-dione: a novel 5-HT1A receptor agonist positron emission tomography ligand. *J. Med. Chem.* **49**, 125–134 (2006).

102. G. Somalo-Barranco, C. Serra, D. Lyons, H. D. Piggins, R. Jockers, A. Llebaria, Design and Validation of the First Family of Photo-Activatable Ligands for Melatonin Receptors. *J. Med. Chem.* **65**, 11229–11240 (2022).

103. E. S. Gutman, E. Bow, F. Li, A. Sulima, S. Kaska, R. Crowley, T. E. Prisinzano, Y.-S. Lee, S. A. Hassan, G. H. Imler, J. R. Deschamps, A. E. Jacobson, K. C. Rice, G-Protein biased opioid agonists: 3-hydroxy-N-phenethyl-5-phenylmorphans with three-carbon chain substituents at C9. *RSC Med Chem* **11**, 896–904 (2020).