Harnessing Chemical Space Neural Networks to Systematically Annotate GPCR ligands

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

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

Abstract

Drug-target interaction databases comprise millions of manually curated data points, yet there are missed opportunities for repurposing established interaction networks to infer novel drug-target interactions by interpolating on chemical neighborhoods. To address this gap, we collect drug-target interactions on 128 unique GPCRs across 187K molecules and establish an all-vs-all chemical space network, which can be efficiently calculated and parameterised as a graph with 32.4 billion potential edges. Beyond testing state-of-the-art machine learning approaches, we develop a chemical space neural network (CSNN) to infer drug activity classes with up to 98% accuracy, by leveraging the graph of chemical neighbourhoods. We combine this virtual library screen with a fast and cheap experimental platform to validate our predictions and to discover 14 novel drug-GPCR interactions. Altogether, our platform integrates virtual library screening and experimental validation to facilitate fast and efficient coverage of missing drug-target interactions.

1 Introduction

At the intersection of machine learning, public drug-target databases, and human G protein-coupled receptors (hGPCRs), lies an opportunity for drug-repurposing and understanding adverse drug effects. hGPCRs are a highly promiscuous group of drug targets, and a single drug may affect more than 30 different targets [1–3]. 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]. Naturally, this is largely due to the significant economic and technical challenges involved at this scale. For example the drug database, ChEMBL, is growing at 780K activities per year (Supplementary Figure S1), with 1,071,688 of 20,334,684 activities reported on hGPCRs [7]. The data, however, is exceptionally sparse with only 1.5% of all possible drug-target combinations being documented. In this vast uncharted space there exist critical undocumented adverse drug effects as well as underexplored opportunities for drug repurposing [8, 9].

Drug databases form the basis of large datasets with relational knowledge between chemical compounds and hGPCRs, as well as accelerated virtual drug screens using machine learning (ML) and deep learning (DL) algorithms. These have proven effective in gigascale chemical spaces at the heart of modern drug discovery [10, 11]. Building upon drug-target databases, ML and DL have been extensively applied to (i) predict bioactivity from a compound representation, (ii) predict interaction of a given compound and target pair, and (iii) predict new targets given a compound representation (See Overview in Table 1). This is a first principles approach, as the prediction itself requires no additional data for inference, but only uses the representations of target and compound to infer bioactivity. Statistically, the task may be defined as the probability of activity y_i given $P(y_i|c_i)$ and $P(y_i|c_i, t_j)$ for activity prediction and drug-target interaction (DTI) prediction, respectively (c_i: compound i, t_j: target j, and y_i: activity of compound i). This, however, introduces an unnecessary disadvantage to current ML/DL algorithms, which consequently lack an important predictive parameter, because the chemical neighbourhood can serve to infer drug targets.

Chemical space networks (CSNs) and similar methods have been developed to rationally engineer drugs by considering the chemical neighbourhood and building structure–activity relationship (SAR) [12]. However, these methods are difficult to scale due to the quadratic dependence on number of compounds and arbitrary design choices in their construction [13, 14]. Chemical similarity thresholds are essential in limiting the connectivity of CSNs to enable interpretation. [13]. These network-based methods have found their way to established chemical bioinformatics companies (e.g. Chemaxon with Neo4j[1]) and have been key tools for SAR campaigns [12]. Preliminary results on successfully utilising CSNs for lead-compound identification were recently demonstrated [15]. Nevertheless, interpolation remains challenging, as the chemical space is insufficiently annotated and sparse.

Here, we investigate the task of predicting activity y_i given a compound c_i, its target t_j, and its chemical neighbourhood $\mathcal{N}_i$. Unlike traditional ML/DL methods, we use public data during inference to inform the prediction of activity, thus tackling the interpolation problem in the sparsely annotated chemical space for drug-repurposing and adverse drug effect predictions. We term our method chemical space neural network (CSNN) and validated it on an experimentally curated test-set of 539 compounds across 7 hGPCRs on a high-throughput [16] and low-cost yeast platform, which aligned with up to 98% accuracy to documented DTIs in ChEMBL and the IUPHAR/BPS guide to pharmacology. We show that our pipeline can rapidly predict and experimentally validate new DTIs in a way that is compatible with home-computers and standard laboratory equipment.

[1] `https://chemaxon.com/blog/presentation/neo4j_presentation`

2 Results

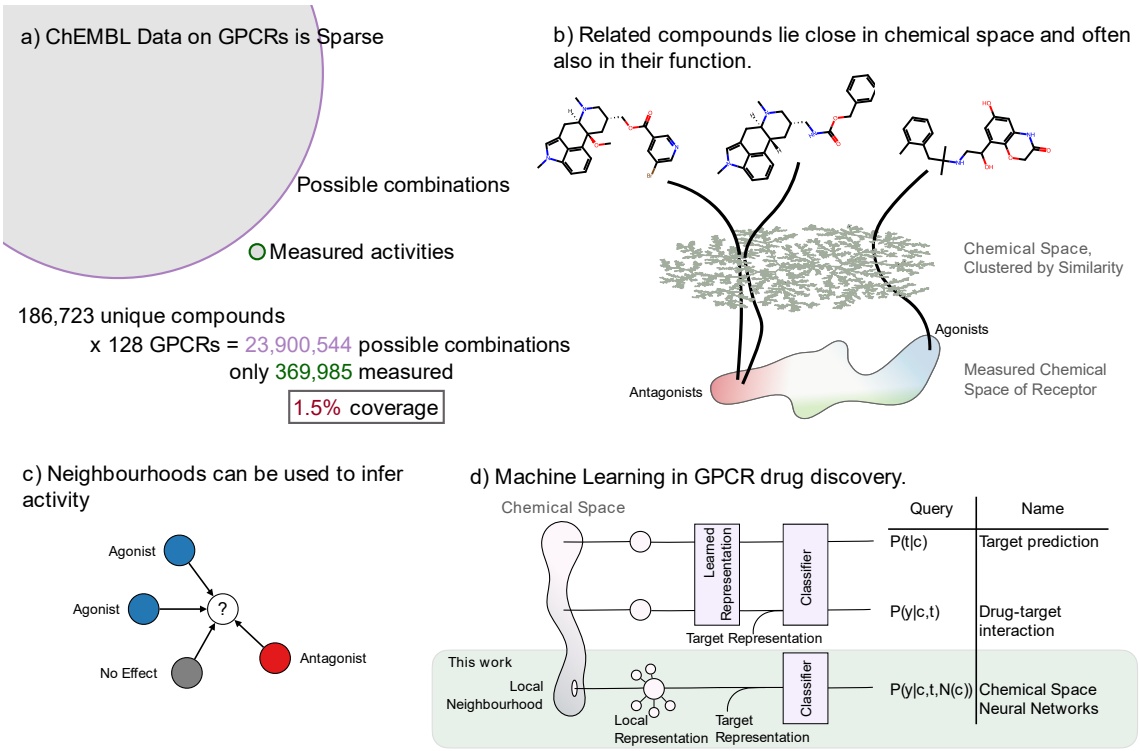

Figure 1: **hGPCR Data Overview in Chemical Neighbourhoods. (a)** A high-quality annotated data set was procured from ChEMBL and IUPHAR/BPS Guide to Pharmacology, comprising 128 hGPCRs and 186,723 unique compounds. 23 million possible DTIs exist for an all-vs-all interaction network, yet public databases cover merely 1.5% of possible interactions. **(b)** Chemical space representing the space of all possible compounds. Molecular structure similarity frequently correlated with function among compounds. **(c)** Homogeneity in the chemical space implied that predictions of bioactivity could be based on neighbouring compounds. **(d)** Prior work on machine learning in drug discovery focused primarily on compound-to-activity predictions. Our work utilised the graph neighbourhood to form a local representation and to inform predictions. We term this method chemical space neural network (CSNN).

2.1 Data curation

As a basis for our work, we compiled a dataset containing all bioactivity data on hGPCRs available in ChEMBL[17] and the IUPHAR/BPS guide to pharmacology[18]. In total, the dataset comprises 186,723 unique molecules represented as SMILES, labeled with known bioactivity classes against the 128 respective hGPCRs (Supplementary File S1).

Figure 1a shows that ChEMBL has 369K annotated activities across 128 hGPCRs. This is, however, a vanishing small 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 780K activities per year (Supplementary Figure S1), which illustrates the need for constructing reliable predictive algorithms to fill in the missing interactions.

Given that compounds which are proximal in the chemical space often have similar bioactivity class on a given hGPCR (Figure 1b), it is conceivable that a chemical neighbourhood is strongly indicative of a compound's bioactivity class. This is illustrated in Figure 1c, where we extended upon works by Yu-Chen Lo et. al [19] and Zengrui Wu et. al [6, 8] to develop Chemical Space Neural Networks (CSNN). The key advantage is illustrated in Figure 1d: Commonly, methods using ML for hGPCR drug-discovery, takes a compound from the chemical space and directly infers its target through some learned or fixed representation (e.g., chemical fingerprints). Predicted DTIs infer bioactivity (y) given the compound representation and

a target representation (P(y|c,t)). CSNN, developed in this work, takes a local neighbourhood, learns a representation, and then uses it for predicting interactions.

2.2 Chemical Space Network Construction in 20 Minutes

To realise CSNN, we first constructed CSNs from an adjacency matrix (Figure 2), which encodes the graph structure connecting similar dots (Figure 2). We tackled three major challenges in large CSN utilisation:

(i) Accelerated CSN construction By using available tools to parallelize the calculation of Jaccard/Tanimoto similarity [20] we reduced the calculation of an all-vs-all CSN for 187K compounds to 20 minutes on a home-computer (32.4 billion edges possible on a MacBook Pro 32 GB RAM, M1 Max @ 2.0 GHz x 10 cores). This makes the method accessible and can be done for any compound library. For moderately sized compound libraries (10K), a CSN was established in 0.952 ± 0.002 seconds. In head-to-head comparison, published methods reported a 1 hour computation for 80K compounds [21], our methods completed the same in 3.7 minutes, which was 16-fold faster than state-of-the art. The resulting computation was a sparse adjacency matrix which describes the connectivity of the graph (See Figure 2a) and forms the basis for querying and visualisation of data with CSNN.

(ii) Open-access CSN tools to Query Available Data We enabled the broader use of CSNs by developing a standalone user-friendly script, akin to [21], for visualising and utilising chemical neighbourhoods to understand the rapidly expanding datasets (Supplementary Figure S1), which are being procured. Our script queries the CSN and automates visual summaries of the neighbourhoods (Figure 2b). The visual neighbourhoods give an overview of available data and help infer possible DTIs of a given compound. The script contains a module for maximum common substructures, thereby additionally providing an overview of the derivatizations performed on a given compound and its neighbourhood (Figure 2c). These tools Querying of the hGPCR dataset and visual reporting was achieved in < 6 seconds, and an output .csv was produced with associated DTIs, ChEMBL ID, and compound information as well as the figures illustrated in Figure 2.

(iii) The Chemical Neighbourhood is Predictive of DTIs Given the CSN, queries about whether the neighbourhood was predictive of DTIs could be answered. This was a multi-class classification task, with 6 bioactivity classes (agonist, antagonist, partial agonist, inverse agonist, allosteric binding or no effect). Initially, we simply summed the frequency of each class in the neighbourhood, irrespective of chemical similarity, and took the maximum argument (argmax(.)). For example, if there are 6 neighbouring compounds which are 'Agonist' and two 'No Effect', then argmax(.) would predict the compound to be an 'Agonist' on a given hGPCR. We compared the predicted labels with that of true data on a given compound and determined the accuracy, precision, recall, and F1 score. For a total of 201,120 unique comparisons across all 128 hGPCRs and 6 bioactivity classes, the neighbourhood achieved a F1 score of 89.7 % (See Figure 2d and S1). This illustrated the inherent graph homophily in the chemical space, meaning that similar and thus connected compounds had similar bioactivity classes on a given hGPCR. The scores in Figure 2d further show that accuracy was reduced for bioactivity classes, which occurred less frequently in the database ('Inverse Agonist', 'Allosteric binding' and 'Partial Agonist'). Despite its remarkable simplicity, argmax(.) set a challenging benchmark for ML in terms of predictive power, while offering complete traceability to the underlying DTIs used for the prediction.

2.3 Developing Chemical Space Neural Networks

Given the strong results of the "white-box" argmax(.) method in predicting the bioactivity class of a given compound, it was necessary to understand if it could compete with state-of-the-art "black-box" ML. To do the benchmark, we structured data into the open-source PyTorch Geometric framework [22].

-

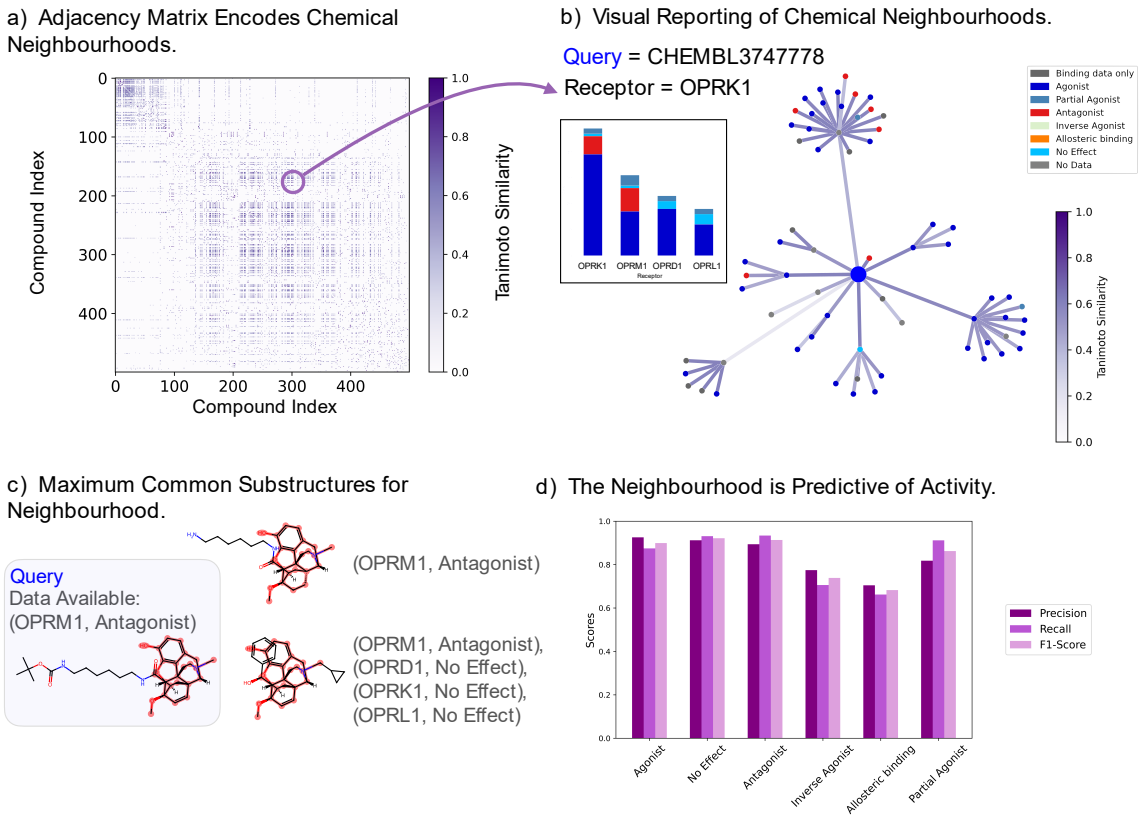

Figure 2: **Rapid Calculation and Visualization of Chemical Neighbourhoods**. **(a)** The adjacency matrix encoded the neighbourhood graph and could be calculated in 20 min for 187K compounds. Searching the chemical space network was rapid (1-hop neighbourhood: 0.187 ± 0.002 seconds) the displayed Adjacency matrix is a subset, displaying pairwise similarity across 500 randomly selected compounds. **(b)** Visual reporting tools enabled quick identification of close compounds for a given query and corresponding data, only edges that expand radially out from the query are shown. **(c)** Maximum common substructures indicates potential activities of query. **(d)** The neighbourhood of a query was found to be predictive of it's activity illustrated by the precision, recall, and F1 scores for the 6 bioactivity classes across 201,120 DTIs (See metrics in S1). This demonstrated that the chemical neighbourhood was homophilous.

An overview of the data is given in Figure 3a and b, which shows that each data-point is a graph $\mathcal{G}$ with node features $\mathbf{x}$, edges $\mathcal{E}$, edge features $\mathbf{e}$, hGPCR embedding $\mathbf{p}$, and the activity label $\mathbf{y}$. Node features (compound representations) were obtained from a pre-trained message-passing neural network (MPNN) from [23], allowing for message-passing on the level of chemical neighbours rather than compounds. Edges restricted the information flow on the graph and were constructed such that it was a directed graph with edges incident on the node to be predicted. Critically, we reformulated CSNNs as a graph classification task (Figure 3a), where the edge features represented the activity label of neighbouring molecules as well as their chemical similarity. Akin to drug-target prediction, we used a protein language model embedding to represent the target (specifically ESM-2 [24]). The activity label was a one-hot encoding of the 6 aforementioned classes, thus a natural loss-function was categorical cross-entropy.

A CSNN prediction was thus $f_\theta : \mathcal{G} \rightarrow \mathbb{R}^6$ (alternatively formulated as $P(\mathbf{y}|\mathcal{G})$), where the graph $\mathcal{G}$ encoded the relevant information of the chemical neighbourhood and hGPCR target. In extended form, a CSNN predicted the bioactivity of compound $\mathbf{y_i}$ given it's representation $\mathbf{x}_i$, the target representation $\mathbf{p}$, and information on the chemical neighbourhood $\mathcal{N}_i$. Training, testing, and validation splits were done according to the Murcko Scaffold using established methods [25], thus preventing data leakage and testing generalisation beyond the training set (See Methods). Additionally, chemical neighbourhoods for subsequent experimental validation were completely withheld during training to test generalisation performance.

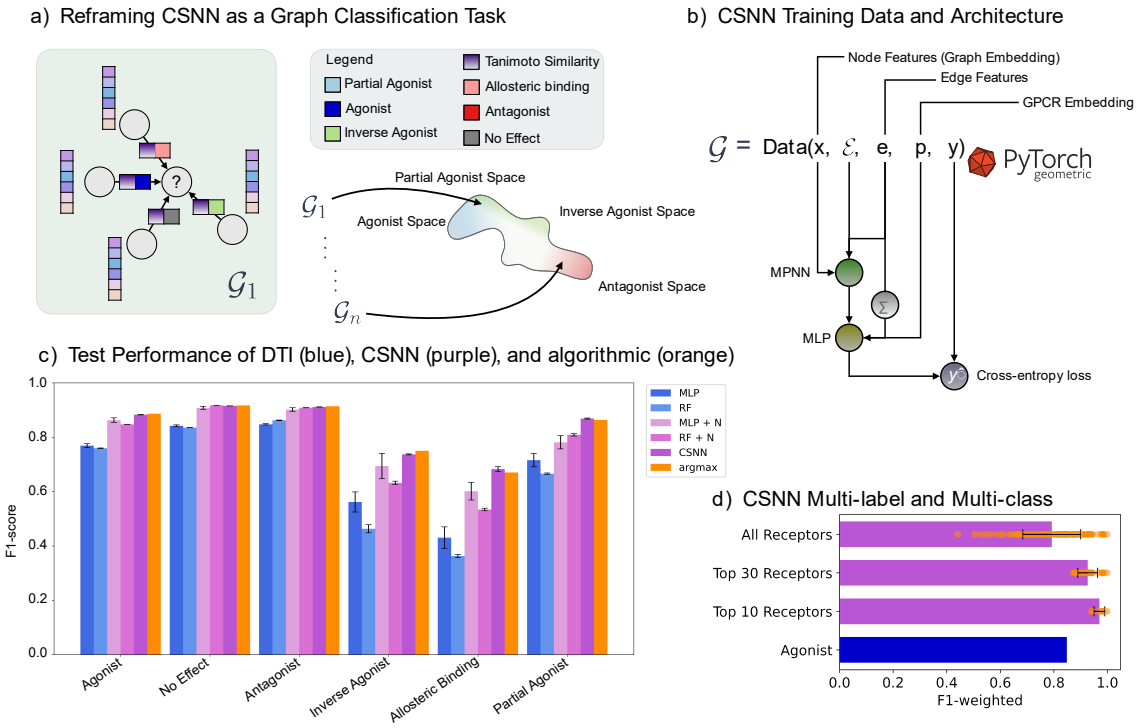

Figure 3: **Chemical Space Neural Networks Generalise Beyond the Training Set (a)** A CSNN task could be re-framed as a graph classification task as $f_\theta : \mathcal{G} \rightarrow \mathbb{R}^6$, where the graph encoded the chemical neighbourhood, compound representation, and bioactivity classes for neighbours as edge features. **(b)** The data was packaged into the accessible PyTorch format, enabling future improvements by the community. **(c)** Test performance of methods that considered only compound and target (DTI, blue), that considered the chemical neighbourhoods (CSNN), and the argmax(.) baseline (orange). The F1 scores indicate that to generalise to unseen scaffolds, utilising available data was highly advantageous as they out-perform DTI methods significantly. **(d)** The CSNN could be formulated as a multi-label and multi-class prediction $f : \mathcal{G} \rightarrow \mathbb{R}^{128x6}$ predicting the class-wise probabilities for all 128 hGPCRs at once, which utilised all available data in the chemical neighbourhood. The top-10 scoring hGPCRs had a class-weighted F1 score of 0.970 ± 0.019.

Supplementing the algorithmic argmax(.) baseline, which utilised the chemical neighbourhood, we also used a drug-target interaction (DTI) method, which uses only the compound representation and the target representation to make predictions.

Figure 3c summarises the F1-scores for each bioactivity class across all hGPCRs for the DTI (blue), CSNN (purple), and algorithmic argmax(.) (orange) methods. Due to the Murcko scaffold splitting, the test molecules evaluated generalisation to unseen molecules, and it was clear that DTI methods lagged behind CSNN and the algorithmic methods, which are aware of chemical neighbourhoods. Drug classes with small statistical support (e.g allosteric binding, and inverse agonist) were difficult to predict, reflected by lower mean F1 scores. The message-passing CSNN outperformed standard DTI methods by $\approx 10\%$, while including neighbourhoods without message-passing was 4% less accurate.

Interestingly, the CSNN performance was on par with a simple algorithmic argmax(.) over the chemical neighbourhood. This questions the advantage of using ML/DL methods to impute missing data if neighbouring compounds with data exist. The CSNN with message-passing achieved a F1-score of 0.8829 ± 0.0009 and 0.9152 ± 0.0002 on the agonist and no effect classes, respectively (See Supplementary Table S2 for complete class summaries), but performed less effectively on classes with lower statistical representation.

Ultimately, to leverage data from all hGPCRs in the chemical neighbourhood, a CSNN was formulated, which took a graph $\mathcal{G}$ as input and predicted bioactivity as $\mathbf{y} \in \mathbb{R}^{128\times6}$, which represented class-wise probabilities for all hGPCRs at once. The input graph contained all available data on all 128 hGPCRs

as edge features, thus allowing for mixing between data channels. This was intuitively advantageous, as related hGPCRs showed similar signalling responses to the same molecule, thus enhancing the predictive performance by integrating all available data. Results of this method are reported in Figure 3d, which shows that for the top-10 scoring hGPCRs, classification was near perfect (class weighted F1-score of 0.970 $\pm$ 0.019). Across all hGPCRs, the classification F1-score was 0.79 $\pm$ 0.11, whereas if only agonist class assignment is considered, the F1-score was 0.848.

2.4 Pairing Virtual Library Screening with Experimental Validation

With the argmax(.) and CSNN methods, we were able to expand the total coverage of labeled hGPCR-compound combinations from 1.5% to 5.6% and 100%, respectively. Naturally, the CSNN method contained a prediction confidence, thus implying that confident predictions do not ensure total coverage. The robust argmax(.) prediction still left 94.4% of the combinations, which can not currently be predicted using white-box methods, as the data is too sparse to impute the missing data. Nevertheless, argmax(.) easily increased data coverage 4-fold from 370K to 1.3 million unique drug-hGPCR interactions. Therefore we investigated both bioinformatic and experimental ways of expanding the coverage further.

Bioinformatically, we observed that clustering of hGPCRs was conserved between MSA-based phylogeny and clustering based on Cramer's V statistic on the shared DTIs between hGPCRs (Supplementary Figures S3, S2). This information indicated that some gaps might be filled by taking into account the target's phylogeny in addition to the protein language models (PLMs) target representation. In fact, the clustering of low-dimensional projections from PLMs implicitly captured the hGPCR class and endogenous ligand class (See Supplementary Figure S2). However, as this approach would only marginally increase the available data for CSNN, we decided to focus more on expanding the dataset with experimentally validated DTIs.

By introducing experimental validations, it is intuitive that the performance of CSNNs will increase, while the argmax(.) method will become applicable to a larger area of the chemical space. We therefore sought to pair our argmax(.) and CSNN methods with an experimental hGPCR assay, to both validate the models on an external dataset, and to provide a complete platform for high-throughput coverage of the chemical space.

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[16, 26] (See Table S3 and supplementary file S2). We therefore constructed a small set of yeast platform strains for 7 hGPCRs previously reported to signal in yeast [26–31], 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) [32] yet with NanoLUC [33] as a reporter instead of GFP for increased assay sensitivity [16]. We confirmed their performances by measuring dose-response curves (DRC) with agonists on each hGPCR to further evaluate the expected quality of gathered data using the yeast platform (See Supplementary Figures S4, S5 and table S5). The EC50 values observed for CHRM3 and ADRA2B were comparable to the most sensitive screens reported (Supplementary table S5[17, 34]). The dynamic ranges were in 6 out of 7 cases higher in yeast than observed for mammalian cell assays (Supplementary Figure S4, and table S5).

We then to assembled a chemically diverse library of 539 unique compounds, 390 sourced from the prestwick library and filtered based on any reported cell surface interactions, and 149 sourced from Chemfaces natural product library, picked from a substructure search for an indole ring. We screened the library against all 7 hGPCRs and by normalising the data with receptor-wise Z-score statistics, we were able to set a unified threshold of $|Z|>3$ (p-value < 0.003) for hit-determination, and the data could then be compared to prior art registered in our annotated dataset (Figure 4a and b). The distribution of hits across chemical space are available in Supplementary Figure S6. Across all tested hGPCRs, we observed 92% correspondence between prior art, consisting of 474 individual DTIs, and the yeast platform (Figure 4b)

.

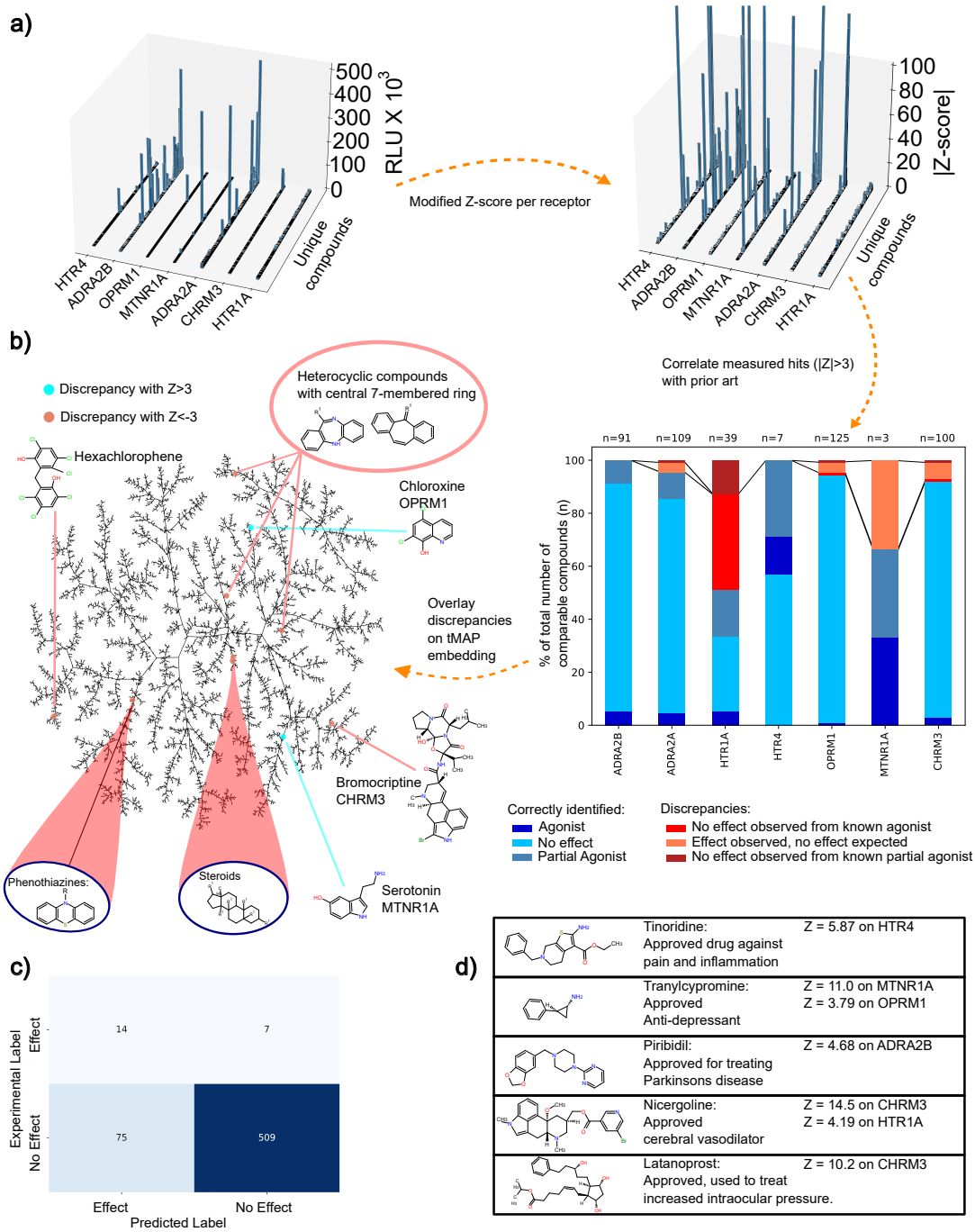

Figure 4: **The Yeast Platform Validates Novel DTIs (a)** Data processing from raw RLU values to modified Z-scores allowed for a unified threshold (|Z|>3) of activity for all 7 hGPCRs. **(b)** Evaluation of yeast's ability to correctly identify previously described hits for each hGPCR. The plot represents percentages of molecules in each group compared to the total number of matches (n) with compounds previously tested on each hGPCR. In cases where an effect was observed, despite literature suggesting that no activity should be expected (salmon), the chemicals in question were plotted as overlaid in a Tmap projection of our complete dataset. **(c)** Confusion matrix over CSNN predictions vs. experimental values, filtered by DTIs for which the chemical neighbourhood was not empty, and for DTIs where the CSNN class prediction was confident (>0.6). **(d)** Highlighted novel hits discovered with no close neighbouring compounds tested on the respective hGPCRs

381 of the 474 comparable DTIs are labelled with "no effect", and if only partial agonism and agonism is investigated, the correspondance drops to 68%. However, the majority of all discrepancies are caused by a single hGPCR, as HTR1A performed exceptionally poorly, missing out on 19 out of 28 expected agonists and partial agonists, likely due to the reduced sensitivity and low dynamic range observed in the DRC (Figure 4b

and Supplementary Figure S4). When excluding HTR1A, the total correspondance rises to 89% for agonists and partial agonists, and to 95% including no effect DTIs. These results demonstrate the conservation of hGPCR signalling in yeast across a variety of DTIs.

The largest group of discrepancies were entries in ChEMBL and IUPHAR/BPS guide to pharmacology being classified as "no effect", which our screen identified as hits (Figure 4b). Hexachlorophene and chloroxine (known antifungals), also generally dampened the signal to a similar degree across all hGPCRs (Except OPRM1 for chloroxine). Finally, bromocriptine displayed a strongly dampening effect on the CHRM3 hGPCR only. By re-screening these compounds on the parental yeast strain ScFH237, we discovered that halogenated phenothiazines, low hydrophilic-lipophilic balance steroids, and all compounds carrying a 7-membered ring, dampened the luminescence signal independently from hGPCR signalling (Supplementary Figure S7) and were consequently excluded from our downstream analyses. We also observed cases where hGPCR activation was measured despite prior art suggesting otherwise. For example, chloroxine activated OPRM1 (Z=8.8), and serotonin activated MTNR1A (Z=5.0). Both have previously been classified as having "no effect" on the respective hGPCRs (CHEMBL1201862 [35], and CHEMBL1130532 [36], respectively).

In summary, our findings revealed 92% correspondence between prior research in mammalian cells and the yeast platform, underscoring the viability of yeast as a cost-effective substitute for human cell lines (Supplementary Table S3). HTR1A accounted for most discrepancies due to poor signalling in yeast, while other variations stemmed from compounds that reduced the reporter signal irrespective of hGPCRs, indicating yeast-specific effects. By excluding hGPCR-independent signal-dampening compounds and HTR1A, we achieve a 98% correspondence between prior studies and the yeast platform.

2.5 Chemical Space Networks Based Predictions Validated Experimentally

With 6 out of 7 tested hGPCRs determined to be reliable in yeast, and problematic compound groups excluded, we proceeded to correlate bioactivity class predictions on the chemical library against the observed hits. We initially looked at the fold-enrichment of Z-scores for predictions of Effect/No-Effect. For example, for compounds assayed against HTR4, the mean Z-score for no effect predictions were -0.46 $\pm$ 0.25 whereas for effect predictions it was 6.37 $\pm$ 2.40, indicating a fold-enrichment of 13.79. Thus, pre-screening a library virtually to filter for effect predictions with CSNNs increased the hit-rate in an hGPCR-dependent manner. We then optimised the class probability threshold to obtain a maximum classification AUROC. We computed a confusion matrix, which showed an overall classification weighted F1-score of 90.2%, but was heavily dominated by correct 'No Effect' predictions (AUROC = 0.769). The false-positive rate was 12.4%, and the recall was 66.6%, indicating the difficulty of generalising to unseen data even with CSNNs.

As with other virtual drug screening methods CSNN has the potential to significantly reduce the experimental library required, thereby reducing the number of experiments required for full coverage. For example, if one took only confident 'Effect' predictions, the library is reduced to 398 compound-target interactions, reduced from the 1607 DTIs with chemical neighbourhood data. This is a 4-fold reduction of experimental testing.

Additionally, the CSN allows for rapid identification of novel experimental hits (i.e., no prior data available on a given hGPCR in the chemical neighbourhood of the hit). We identified 14 novel hits with | Z-score | > 3 and no prior data on the hGPCR in the chemical neighbourhood (Figure 4c and d). Seven of these hits are highlighted on Figure 4d, and each of these represents new information on approved drugs. Tinoridines' association with HTR4 adds indications for treating gastrointestinal disorders [37], whereas Nicergoline interaction with CHRM3 indicates potential uses for treating Alzheimer's disease and schizophrenia [38] 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 [39]. 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[40] (See Supplementary Table S4).

3 Discussion

We have developed chemical space neural networks, CSNNs, which consistently fills out the gaps among the 23 million possible compound-hGPCR interactions across ChEMBL and IUPHAR/BPS guide to pharmacology. We show that DTI methods relying on ML are less robust and generalise poorly as compared to CSNNs, and then train a CSNN to predict bioactivity classes for all 128 hGPCRs in one forward-pass, allowing for mixing between data channels. We demonstrate that a white-box algorithm, argmax(.), performs on par with CSNNs, thus illustrating the marginal gains of black-box algorithms in this application area. We robustly increase data coverage of the chemical space surrounding all 128 hGPCRs from 1.5% to 5.6% with a purely white-box algorithmic method (i.e., argmax(.)), and offer tentative predictions on all possible DTIs within this framework using CSNN. Extending on the predictions experimentally, our work has established yeast as a reliable alternative to the current market options for hGPCR reporter assays by recapitulating 98% of known agonists, partial agonists and no effect data for ADRA2A, ADRA2B, HTR4, MTNR1A, OPRM1, and CHRM3, thus showing high conservation between the yeast platform and mammalian cell line reporters. Finally, we identify hGPCR-independent reporter-dampening effects that explain observed mismatches, which originate from steroid class compounds, halogentated phenothiazines and 7-membered ring-containing heterocyclic compounds. These results highlight the persistent need for including proper controls in HTS campaigns.

CSNs face significant criticism, as subtle molecular modifications can change the toxicological and pharmacological behaviours of compounds [13]. However, our results indicate that reliable conclusions can be drawn from chemical neighbourhoods to correctly assign a bioactivity class, as we demonstrate an AUROC of 0.769 for CSNN predictions for the experimental data set tested on the yeast platform. Additionally, our visualisation tools for combining maximum common substructures and CSN neighborhood visualisations makes it possible to quickly summarise the results of past and future derivatization attemps.

Using CSNNs as a virtual pre-screening method prior to *in vitro* experimental validation, the required number of experimental testing could be reduced by up to an order of magnitude. A recent publication by Davide Boldini et al. [41] demonstrated how changing the molecular fingerprint-encoding can lead to very different similarity scores. This could potentially be used to improve upon the methods presented here, by iterating over several encoding methods to ascertain which predictions remain significant across all encodings. Nevertheless, we identified 14 completely novel compound-hGPCR hits, with no known prior art in the chemical neighbourhoods available, and then tentatively linked these DTIs to biological functions, underpinning the importance of partnering *in silico* methods with experimental testing. We open-source all data to encourage future advancements, including 3,773 additional experimental measurements. Taken together, these findings demonstrate that partnering network-based predictions *in silico* with a robust experimental validation platform facilitates both efficient and cost-effective searches into the dark areas of the chemical space.

Some of the observed discrepancies between prior art and the yeast platform can be explained by the comparatively high dynamic ranges on the yeast platform, which may have allowed it to pick up on DTIs that other screens might have missed. Chloroxine is known to bind OPRK1 [42], so based on phylogeny, it is not surprising that it also binds to OPRM1 as shown in our screen, despite the conflicting information from prior art. Activation of MTNR1A by serotonin has already been shown in other publications relying on yeast platforms [26, 43], however, this information is not deposited on either ChEMBL or the IUPHAR/BPS

guide to pharmacology, which is why an older, and presumably less sensitive assay, is the only reference on this combination in the dataset[36]. This demonstrates how the high dynamic ranges on the yeast platform (Supplementary figure S4) can help catch DTIs, which other screening campaigns have missed.

We found that hGPCRs, which are phylogenetically related, also share similar chemical spaces targeting them. This has important implications toward algorithmic white-box methods that can transfer data between hGPCRs to further expand the coverage of the chemical space in the near future. We envision that (i) neural networks that operate on the graph structure of chemical neighbourhoods rather than the graph structure of compounds will enhance interpolations, thus reliably identifying molecules that may be implicated in adverse drug effects, 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 investigations. Point (i) is underway with established algorithmic[8, 44] and neural[45] network-based methods, yet is limited to paired compounds. Here we take a significant step toward CSNN. What remains to be done are end-to-end differentiable and explainable [46] CSNNs, rather than using precomputed compound representations. Point (ii) is also well-underway, with over 50 hGPCRs already reported to signal in yeast [47]. CSNNs should additionally benefit biotechnology by rapidly identifying a suitable hGPCR for a given ligand to accelerate the sustainable biosynthesis of complex natural products and derivatives by utilising state-of-the-art microbial cell factories [43, 48].

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.

Funding

This work was supported by

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

European Union Horizon 2020 research and innovation program (MKJ)

4 Methods

Compiling bioactivity data

Bioactivity data from ChEMBL on all human GPCRs with more than 500 unique compound associations were downloaded and saved to csv files, totalling 128 receptors. All entries where no SMILES annotations were provided were removed. Salts were then removed from all smiles, and using a custom sorting algorithm, we classified all entries into one of the following bioactivity classes; agonist, antagonist, partial agonist, inverse agonist, binding data only, allosteric binding or no effect. This was done based on the information provided in the columns of the ChEMBL data structure, in particular, the columns; "Assay Description", "Comments", "ACTION TYPE", "Properties", "Standard relation", "Standard value", "Standard units", and "Standard type" were taken into account.

Then the IUPHAR/BPS database on the same 128 hGPCRs was pulled using their API: "https://www.guidetopharmacology.org/services/", and sorted into the previously listed bioactivity classes based on the provided functional data. The two datasets were combined and entries that failed to be categorised were removed from the dataset. Duplicate measurements for receptor-ligand combinations were removed, keeping the highest level of information in the following order; "Inverse Agonist", "Partial Agonist", "Antagonist", "Agonist", "Allosteric binding", "Binding", "No effect". The python scripts used to perform above data cleanup and the resulting dataset will be made publicly available upon publication in a peer-reviewed journal, the dataset is available upon request as Supplementary File S3.

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. [49] 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 S10).

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). We take inspiration from [12, 21]. 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[2]. 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_chemistry/dataset/splitters/scaffold_splitter.py`. This ensures that scaffolds in the test dataset are not present during training. The splits will be made available upon publication in a peer-reviewed journal.

For each compound in each split we explode the pandas array such that each unique compound-receptor interaction 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 [23]. This numerically represents the chemical graph 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 [46, 50]. Likewise, we precompute target representations $\mathbf{p}_i$ using the ESM-2 protein language model and the canonical UniProt sequence (esm.pretrained.esm2_t6_8M_UR50D()) [24].

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 [22]. 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 3a. The edge features are $\mathbf{e} \in \mathbb{R}^{|\mathcal{E}| \times 7}$ which includes the one-hot encoding of the neighbouring compound bioactivity (6 classes) 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. As PyTorch requires $\mathbf{p} \in \mathbb{R}^{|\mathcal{V}| \times d}$ we broadcast the target representation, which is only of dimension 1x320 (ESM-2). $\mathbf{y}$ is a one-hot encoding of the bioactivity class such that $\mathbf{y} \in \mathbb{Z}^6$. We provide the torch files dataset_train.pt, dataset_test.pt, dataset_val.pt to enable subsequent use and bench-marking of future methods. As well as i.i.d sampled versions which are used as in subsequent analysis. I.i.d sampling and retraining was carried out before predicting activities of the yeast compound library. We still withhold the chemical neighbourhoods of the library, by removing any compounds in the one-hop neighbourhood.

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

Predictions: argmax(.)

To determine 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. 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.

In Figure 2 the metrics are computed across all graphs in test, train, and validation to obtains 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).

In Figure 3, only the performance metrics for the test set is calculated allowing for comparison with the methods presented.

Predictions: Drug-target interaction

For the drug-target interaction (DTI) mode, we simply extract the node representation for the graph and the target representation so to create a concatenated form $\mathbf{x}_i \parallel \mathbf{p}_i$ which is used as a one dimensional input vector to standard machine learning tools. 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 $\mathbf{y} \in \mathbb{Z}^6$. 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.

Predictions: CSNN

The chemical space neural network (CSNN) is designed as illustrated in Figure 3b. We instantiate CSNN as a weaker form of message-passing [52], namely a graph attention network (GAT) [53]. In the first block, a local representation is computed by aggregating information from neighbouring nodes with attention to form the updated $\mathbf{r}_i$ from the local latent representations $\mathbf{x}'_i$:

$$\mathbf{x}'_i = \alpha_{i,i}\mathbf{\Theta}_s\mathbf{x}_i + \sum_{j\in\mathcal{N}(i)} \alpha_{i,j}\mathbf{\Theta}_t\mathbf{x}_j \tag{2}$$

with the attention coefficient $\alpha_{i,j}$ that is calculated utilising the edge-features $\mathbf{e}_{i,j}$ as:

$$\alpha_{i,j} = \frac{\exp(\text{LeakyReLU}(\mathbf{a}_s^\top\mathbf{\Theta}_s\mathbf{x}_i + \mathbf{a}_t^\top\mathbf{\Theta}_t\mathbf{x}_j + \mathbf{a}_e^\top\mathbf{\Theta}_e\mathbf{e}_{i,j}))}{\sum_{k\in\mathcal{N}(i)\cup\{i\}}\exp(\text{LeakyReLU}(\mathbf{a}_s^\top\mathbf{\Theta}_s\mathbf{x}_i + a_t^\top\mathbf{\Theta}_t\mathbf{x}_k + \mathbf{a}_e^\top\mathbf{\Theta}_e\mathbf{e}_{i,k}))} \tag{3}$$

A graph mean pool for $\mathcal{G}_i$ is computed as the average:

$$\mathbf{r}_i = \frac{1}{N_i}\sum_{n=1}^{N_i}\mathbf{x}'_i, \quad \mathbf{r}_i \tag{4}$$

We use an output dimension of 64 for the local representation using two attention heads and only a single GAT layer as each node is only one edge from one another.

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 (256, 6 hidden layers) as

$$\hat{\mathbf{z}}_i = \sigma\left(\mathbf{W}\mathbf{r}_i + \mathbf{b}\right), \quad \hat{\mathbf{z}}_i \in \mathbb{R}^6 \tag{5}$$

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 [54], a learning rate of 0.00001 and weight decay of 0.0001:

$$\mathcal{L}_{\text{batch}} = -\frac{1}{N_{batch}}\sum_{i=1}^{N_{batch}}\sum_{c=1}^{C} y_{i,c}\log\left(\frac{e^{z_{i,c}}}{\sum_{j=1}^{C} e^{z_{i,j}}}\right) \tag{6}$$

Predictions: CSNN Multi-label and Multi-class

The previous method predicts only class-wise probabilities for a given drug-target interaction, 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 (128x7 = 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. The concatenated $\mathbf{r}'_i$ is of dimension 1024 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 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.

Performance Metrics

For the DTI methods (RF and MLP), the DTI with neighourhoods (RF + N, MLP + N,), and CSNN the performance metrics are calculated for each bioactivity class i as

$$\text{Precision}_i = \frac{TP_i}{TP_i + FP_i} \tag{7}$$

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

$$\text{Recall}_i = \frac{TP_i}{TP_i + FN_i} \tag{8}$$

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

$$\text{F1}_i = 2 \times \frac{\text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \tag{9}$$

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

$$\text{F1-weighted} = \sum_{i=1}^{N} w_i \times \text{F1}_i \tag{10}$$

where w_i is the proportion of true instances for class i out of all instances, and N is the number of classes.

Input	Task	Explanation	Notation	Type	Refs
c_i	y_i	Does the compound have GPCR activity?	$P(y_i\|c_i)$	Regression, Classification	[6]
c_i, t_j	y_i	Given a compound and it's target representation, is the compound active?	$P(y_i\|c_i, t_j)$	Regression, Classification	
c_i, c_j	y_i, y_j	Given two compounds, how active are they?	$P(y_i, y_j\|c_i, c_j)$	Regression, Classification	[45]
c_i	t_j	Given a compound, predict its target.	$P(t_i\|c_i)$	Classification	[8]
c_i	a_j	Given a compound, predict adverse drug events.	$P(a_i\|c_i)$	Regression, Classification	[9]
Inference from Neighbourhood					
$c_i, t_j, N_i(y_j)$	y_i	Given a compound, it's neighbourhood activity on one target j, and a target representation t_j, is it active?	$P(y_i\|c_i, N_i(y_j), t_j)$	Classification	(ours)
$c_i, N_i(y_n), t_n$	$y_{i,n}$	Given a compound, it's neighbourhood activity on all n targets, and target representations, on which target is c_i active?	$P(y_{n,i}\|c_i, N_i(y_n), t_j)$	Classification	(ours)

Table 1: Overview of current efforts on drug activity inference methods

Chemical library

390 compounds were sourced from the Prestwick library, selected by reported receptor associations and chemical diversity, stock concentrations at 10mM. 160 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.

The genes

All biosynthetic genes used in the current study are listed in Supplementary Table S6. 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 [55] and propagated in E. coli DH5α competent cells. Genes for integrative Easy Clone - marker free vectors were prepared according to Jessop-Fabre et al. [56] and the gRNA expression cassettes as previously described [57]. 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 [58]. The integration of heterologous genes was verified as described in CasEMBLR [56, 57]. The strains generated in this study are listed in supplementary Table S7

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.[32]. 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 hGPCR biosensors, described in Kapolka et al. and Miettinen et al.[26][16]. 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[59]. OD_{600} was measured and adjusted to 5 before continuing with the procedures described in "Bioactivity assays"

Chemical preparations

For the DRCs shown in Figure 4, 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 S4, 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 S7, 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 CelLytic™ 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.

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 S3 were generated from distances calculated between all 128 selected receptors based on; 1. ESM-LLM distance metrics [24], 2. from the chemical space they are able/unable to sense, and 3. from a multiple sequence alignment using muscle 5.1[60]. Three distinct trees were then generated using the Bio.Phylo package[61], and visualised using ETE V3[62].

Dose response curve calculation

All DRCs shown in Figure 4A 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[63]. The models chosen for HTR1A and CHRM3 were a bell shaped DRC[64], and a bi-phasic DRC[65] 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[66]. All additional information related to the fitted curves is available in supplementary file S4.

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:

$$\mathrm{Z-score} = \frac{x_i - \mu}{\sigma} \tag{11}$$

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:

$$\mathrm{Z-score}_{mod} = 0.6745 \cdot \frac{x_i - \mathrm{m}}{\mathrm{MAD}}, \quad MAD = \mathrm{med}(|x_i - \mathrm{m}|), \quad \mathrm{m} = \mathrm{med}(X) \tag{12}$$

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 S8, 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 $\mid \mathrm{Z-score} \mid > 3$.

References

[1] D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, J. Woolsey, *Nucleic Acids Res.* **2006**, *34*, D668–72.

[2] Clozapine DB00363, `https://go.drugbank.com/drugs/DB00363`, Accessed: 2024-1-23.

[3] K. Alhosaini, A. Azhar, A. Alonazi, F. Al-Zoghaibi, *Saudi Pharmaceutical Journal* **2021**, *29*, 539–551.

[4] K. Sriram, P. A. Insel, *Mol. Pharmacol.* **2018**, *93*, 251–258.

[5] P. A. Insel, C.-M. Tang, I. Hahntow, M. C. Michel, *Biochim. Biophys. Acta* **2007**, *1768*, 994–1005.

[6] Z. Wu, W. Li, G. Liu, Y. Tang, *Front. Pharmacol.* **2018**, *9*, 1134.

[7] 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, *Nucleic Acids Research* **2024**, *52*, D1180–D1192.

[8] Z. Wu, H. Ma, Z. Liu, L. Zheng, Z. Yu, S. Cao, W. Fang, L. Wu, W. Li, G. Liu, J. Huang, Y. Tang, *Chem. Sci.* **2022**, *13*, 1060–1079.

[9] Z. Yu, Z. Wu, W. Li, G. Liu, Y. Tang, *Brief. Bioinform.* **2022**, *23*.

[10] A. V. Sadybekov, V. Katritch, *Nature* **2023**, *616*, Number: 7958 Publisher: Nature Publishing Group, 673–685.

[11] G. Opassi, A. Gesù, A. Massarotti, *Drug Discovery Today* **2018**, *23*, 565–574.

[12] H. Matter, C. Buning, D. D. Stefanescu, S. Ruf, G. Hessler, *J. Chem. Inf. Model.* **2020**, *60*, Publisher: American Chemical Society, 6120–6134.

[13] R. Kunimoto, J. Bajorath, *ACS Omega* **2018**, *3*, Publisher: American Chemical Society, 3768–3777.

[14] N. Amoroso, *Scientific Reports* **2023**.

[15] J. Yi, S. Lee, S. Lim, C. Cho, Y. Piao, M. Yeo, D. Kim, S. Kim, S. Lee, *Computational and Structural Biotechnology Journal* **2023**, *21*, 4187–4195.

[16] 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, *Nat. Commun.* **2022**, *13*, 3664.

[17] 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, *Nucleic acids research* **2019**, *47*, D930–D940.

[18] S. D. Harding, J. F. Armstrong, E. Faccenda, C. Southan, S. P. H. Alexander, A. P. Davenport, M. Spedding, J. A. Davies, *Nucleic Acids Research* **2023**, *52*, D1438–D1449.

[19] Y.-C. Lo, S. Senese, C.-M. Li, Q. Hu, Y. Huang, R. Damoiseaux, J. Z. Torres, *PLoS Comput. Biol.* **2015**, *11*, e1004153.

[20] A. Vardanian, SimSIMD, https://github.com/ashvardanian/SimSIMD, **2023**.

[21] V. F. Scalfani, V. D. Patel, A. M. Fernandez, *Journal of Cheminformatics* **2022**, *14*, 87.

[22] M. Fey, J. E. Lenssen, Fast Graph Representation Learning with PyTorch Geometric, **2019**.

[23] A. Dalkıran, A. Atakan, A. S. Rifaioğlu, M. J. Martin, R. Ç. Atalay, A. C. Acar, T. Doğan, V. Atalay, *Bioinformatics* **2023**, *39*, i103–i110.

[24] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, et al., *bioRxiv* **2022**.

[25] E. Heid, K. P. Greenman, Y. Chung, S.-C. Li, D. E. Graff, F. H. Vermeire, H. Wu, W. H. Green, C. J. McGill, *J. Chem. Inf. Model.* **2024**, *64*, Publisher: American Chemical Society, 9–17.

[26] N. J. Kapolka, G. J. Taghon, J. B. Rowe, W. M. Morgan, J. F. Enten, N. A. Lambert, D. G. Isom, *Proc. Natl. Acad. Sci. U. S. A.* **2020**, *117*, 13117–13126.

[27] I. Erlenbach, E. Kostenis, C. Schmidt, F. F. Hamdan, M. H. Pausch, J. Wess, *J. Neurochem.* **2001**, *77*, 1327–1337.

[28] 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., *Yeast* **2000**, *16*, 11–22.

[29] A. M. Ehrenworth, T. Claiborne, P. Peralta-Yahya, *Biochemistry* **2017**, *56*, 5471–5475.

[30] G. Gaibelet, E. Meilhoc, J. Riond, I. Saves, T. Exner, L. Liaubet, B. Nürnberg, J.-M. Masson, L. J. Emorine, *European Journal of Biochemistry* **1999**, *261*, 517–523.

[31] T. Kokkola, M. A. Watson, J. White, S. Dowell, S. M. Foord, J. T. Laitinen, *Biochem. Biophys. Res. Commun.* **1998**, *249*, 531–536.

[32] W. M. Shaw, H. Yamauchi, J. Mead, G.-O. F. Gowers, D. J. Bell, D. Öling, N. Larsson, M. Wigglesworth, G. Ladds, T. Ellis, *Cell* **2019**, *177*, 782–796.e27.

[33] C. G. England, E. B. Ehlerding, W. Cai, *Bioconjugate Chemistry* **2016**, *27*, PMID: 27045664, 1175–1187.

[34] 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, *J. Med. Chem.* **2017**, *60*, 9239–9250.

[35] 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, *Nucleic Acids Res.* **2023**, *51*, D1373–D1380.

[36] P.-K. Li, G.-H. Chu, M. L. Gillen, T. Parekh, P. A. Witt-Enderby, *Bioorg. Med. Chem. Lett.* **1997**, *7*, 2409–2414.

[37] E. J. Jeong, S. Y. Chung, H. N. Hong, S.-W. Oh, J. Y. Sim, *Br. J. Pharmacol.* **2018**, *175*, 485–500.

[38] S. P. Moran, J. Maksymetz, P. J. Conn, *Trends Pharmacol. Sci.* **2019**, *40*, 1006–1020.

[39] A. W. Cowley, Jr, *Hypertension* **2018**, *72*, 588–591.

[40] S. Rivara, M. Mor, A. Bedini, G. Spadoni, G. Tarzia, *Curr. Top. Med. Chem.* **2008**, *8*, 954–968.

[41] D. Boldini, D. Ballabio, V. Consonni, R. Todeschini, F. Grisoni, S. A. Sieber, *J. Cheminform.* **2024**, *16*, 35.

[42] J. Saxena, F. Nicolini, M. Lockley, *Cancer Res.* **2019**, *79*, 3801–3801.

[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, *Nat. Commun.* **2022**, *13*, 6201.

[44] F. Cheng, C. Liu, J. Jiang, W. Lu, W. Li, G. Liu, W. Zhou, J. Huang, Y. Tang, *PLoS Comput. Biol.* **2012**, *8*, e1002503.

[45] K. Amara, R. Rodríguez-Pérez, J. Jiménez-Luna, *Journal of Cheminformatics* **2023**, *15*, 67.

[46] J. Jiménez-Luna, F. Grisoni, G. Schneider, *Nat Mach Intell* **2020**, *2*, Number: 10 Publisher: Nature Publishing Group, 573–584.

[47] B. Lengger, M. K. Jensen, *FEMS Yeast Res.* **2020**, *20*.

[48] 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, *Nat Chem Biol* **2023**, *19*, Number: 12 Publisher: Nature Publishing Group, 1551–1560.

[49] D. C. Anastasiu, G. Karypis, *Int J Data Sci Anal* **2017**, *4*, 153–172.

[50] 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, *J. Chem. Inf. Model.* **2019**, *59*, Publisher: American Chemical Society, 3370–3388.

[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, *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

[52] M. M. Bronstein, J. Bruna, T. Cohen, P. Velickovic, *CoRR* **2021**, *abs/2104.13478*.

[53] P. P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, Y. Bengio in International Conference on Learning Representations, **2018**.

[54] D. Kingma, J. Ba in International Conference on Learning Representations (ICLR), San Diega, CA, USA, **2015**.

[55] H. H. Nour-Eldin, B. G. Hansen, M. H. H. Nørholm, J. K. Jensen, B. A. Halkier, *Nucleic Acids Res.* **2006**, *34*, e122.

[56] M. M. Jessop-Fabre, T. Jakočiūnas, V. Stovicek, Z. Dai, M. K. Jensen, J. D. Keasling, I. Borodina, *Biotechnol. J.* **2016**, *11*, 1110–1117.

[57] T. Jakočiūnas, I. Bonde, M. Herrgård, S. J. Harrison, M. Kristensen, L. E. Pedersen, M. K. Jensen, J. D. Keasling, *Metab. Eng.* **2015**, *28*, 213–222.

[58] R. D. Gietz, R. H. Schiestl, *Nat. Protoc.* **2007**, *2*, 38–41.

[59] R. C. Prins, S. Billerbeck, *BMC Microbiol.* **2021**, *21*, 127.

[60] R. C. Edgar, *bioRxiv* **2021**.

[61] E. Talevich, B. M. Invergo, P. J. A. Cock, B. A. Chapman, *BMC Bioinformatics* **2012**, *13*, 209.

[62] J. Huerta-Cepas, F. Serra, P. Bork, *Mol. Biol. Evol.* **2016**, *33*, 1635–1638.

[63] GraphPad Software, LLC. H. J. Motulsky, Equation: [Agonist] vs. response – Variable slope, GraphPad Curve Fitting Guide. `https://www.graphpad.com/guides/prism/latest/curve-fitting/reg_dr_stim_variable_2.htm`, Accessed: 2023-10-26.

[64] GraphPad Software, LLC. H. J. Motulsky, Equation: [Agonist] vs. response – Bell-shaped dose-response, GraphPad Curve Fitting Guide. `https://www.graphpad.com/guides/prism/latest/curve-fitting/reg_bellshaped_dose_response_2.htm`, Accessed: 2023-10-26.

[65] GraphPad Software, LLC. H. J. Motulsky, Equation: [Agonist] vs. response – Biphasic dose-response, GraphPad Curve Fitting Guide. `https://www.graphpad.com/guides/prism/latest/curve-fitting/reg_biphasic_dose_response_2.htm`, Accessed: 2023-10-26.

[66] GraphPad Software, LLC. H. J. Motulsky, Equation: [Agonist] vs. response –Asymmetrical (five parameter), GraphPad Curve Fitting Guide. `https://www.graphpad.com/guides/prism/latest/curve-fitting/reg_asymmetric_dose_response_ec_2.htm`, Accessed: 2023-10-26.

[67] M. N. Davies, A. Secker, A. A. Freitas, M. Mendao, J. Timmis, D. R. Flower, *Bioinformatics* **2007**, *23*, 3113–3118.

[68] L. van der Maaten, G. Hinton, *Journal of Machine Learning Research* **2008**, *9*, 2579–2605.

[69] D. Probst, J.-L. Reymond, *J. Cheminform.* **2020**, *12*, 12.

[70] 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**.

[71] C. Bergsdorf, C. Kropp-Goerkis, I. Kaehler, L. Ketscher, U. Boemer, K. Parczyk, B. Bader, *Assay Drug Dev. Technol.* **2008**, *6*, 39–53.

[72] Q. Ma, L. Ye, H. Liu, Y. Shi, N. Zhou, *Expert Opin. Drug Discov.* **2017**, *12*, 511–523.

[73] R. H. Oakley, C. C. Hudson, R. D. Cruickshank, D. M. Meyers, R. E. Payne, Jr, S. M. Rhem, C. R. Loomis, *Assay Drug Dev. Technol.* **2002**, *1*, 21–30.

[74] J. Wang, X. Xie, *Acta Pharmacol. Sin.* **2007**, *28*, 1698–1704.

[75] P. G. Strange, *Br. J. Pharmacol.* **2010**, *161*, 1238–1249.

[76] D. B. Bylund, J. D. Deupree, M. L. Toews, *Methods Mol. Biol.* **2004**, *259*, 1–28.

[77] D. B. Bylund, M. L. Toews, *Am. J. Physiol.* **1993**, *265*, L421–9.

[78] H. Chen, C. E. Rosen, J. A. González-Hernández, D. Song, J. Potempa, A. M. Ring, N. W. Palm, *Cell* **2023**, *186*, 3095–3110.e19.

[79] F. Fiorino, B. Severino, E. Magli, A. Ciano, G. Caliendo, V. Santagada, F. Frecentese, E. Perissutti, *J. Med. Chem.* **2014**, *57*, 4407–4426.

[80] 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, *Transl. Psychiatry* **2022**, *12*, 273.

[81] M. Ono, E. Takamura, K. Shinozaki, T. Tsumura, T. Hamano, Y. Yagi, K. Tsubota, *Am. J. Ophthalmol.* **2004**, *138*, 6–17.

[82] A. K. Paul, C. M. Smith, M. Rahmatullah, V. Nissapatorn, P. Wilairatana, M. Spetea, N. Gueven, N. Dietis, *Pharmaceuticals* **2021**, *14*.

[83] Z. P. Khan, C. N. Ferguson, R. M. Jones, *Anaesthesia* **1999**, *54*, 146–165.

[84] 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, *Bioorg. Med. Chem.*, *24*, 3174–3183.

[85] 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, *J. Med. Chem.* **2009**, *52*, 2214–2225.

[86] 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, *J. Med. Chem.* **2006**, *49*, 125–134.

[87] G. Somalo-Barranco, C. Serra, D. Lyons, H. D. Piggins, R. Jockers, A. Llebaria, *J. Med. Chem.* **2022**, *65*, 11229–11240.

[88] 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, *RSC Med Chem* **2020**, *11*, 896–904.