DeorphaNN: Virtual screening of GPCR peptide agonists using AlphaFold-predicted active state complexes and deep learning embeddings

Larissa Ferguson[1,*], Sébastien Ouellet[2], Elke Vandewyer[3], Christopher Wang[2,4], Zaw Wunna[1,5], Tony K.Y. Lim[6], William R. Schafer[1,3,§], Isabel Beets[3,§]

[1] Neurobiology Division, MRC Laboratory of Molecular Biology, Cambridge, United Kingdom

[2] Independent Researcher, Ottawa, Ontario, Canada

[3] Department of Biology, KU Leuven, Leuven, Belgium

[4] Current affiliation: University College London, United Kingdom

[5] Current affiliation: University of Leeds, United Kingdom

[6] Department of Pharmacology, University of Cambridge, Cambridge, United Kingdom.

[§] Senior authors (equal contribution)

* Corresponding Author

ORCID: LF (0000-0003-4274-8634), SO (0000-0002-3098-4907), TKYL (0000-0003-1843-0060), WRS (0000-0002-6676-8034), IB (0000-0002-0968-4360)

CLASSIFICATIONS

Biological Sciences—Biophysics and Computational Biology

KEYWORDS

G protein-coupled receptors (GPCRs), AlphaFold, deorphanization, peptide agonists, neuropeptides, peptide hormone, protein representations, active state structures, protein embeddings, structural bioinformatics

ABSTRACT

G protein-coupled receptors (GPCRs) are important cell surface receptors involved in numerous physiological processes. Although peptides are the cognate ligands for many of these receptors, identifying endogenous peptide agonists for GPCRs remains a significant challenge. Deep learning-based protein structure prediction algorithms, such as AlphaFold (AF) have utility in non-structural tasks including protein-protein interaction prediction, suggesting they may be useful for predicting GPCR-peptide agonist interactions. Leveraging a dataset of experimentally validated agonist and non-agonist GPCR-peptide interactions from *Caenorhabditis elegans*, we show that AF-Multimer confidence metrics enable partial discrimination between GPCR-agonist and non-agonist complexes. To better reflect agonist-bound conformations, AF-Multistate templates are used to produce active-state GPCR-peptide complexes, improving discriminatory power. Embeddings from the final hidden layer of AF-Multimer's neural network, which capture structural and interaction patterns, were used to train random forest classifiers to assess whether AF-Multimer protein representations can distinguish agonist from non-agonist complexes. Feature performance analysis reveals that AF-Multimer's pair representations outperform single representations, with distinct subregions of the pair representation providing complementary predictive signals. Building on these findings, we developed DeorphaNN—a graph neural network that integrates active-state GPCR-peptide structural predictions, interatomic interactions, and pair representations to predict agonist identity. DeorphaNN's predictive utility generalizes to datasets outside of *C. elegans*, including annelids and humans, and experimental validation of predicted agonists for two orphan GPCRs uncovers their cognate agonists. Our approach offers a resource to accelerate GPCR deorphanization through the *in silico* identification of receptor-agonist candidates for AI-guided experimental validation.

SIGNIFICANCE STATEMENT

This study addresses the challenge of identifying peptide agonists for G protein-coupled receptors (GPCRs), which play crucial roles in numerous physiological processes. We have developed a machine learning model that combines predicted active state complexes, interatomic interactions, and deep learning protein representations obtained from AlphaFold-Multimer to identify potential peptide agonists for GPCRs. We then validated this approach by identifying and experimentally confirming novel ligands for two GPCRs. This approach can accelerate the identification of promising peptide ligands for GPCRs, offering a valuable tool to guide GPCR deorphanization efforts and improve our understanding of critical signalling pathways involved in health and disease.

INTRODUCTION

G protein-coupled receptors (GPCRs) are a large and diverse family of membrane proteins critical for cellular signalling. Defined by their characteristic seven-transmembrane domain structure, GPCRs couple extracellular ligand binding to the regulation of intracellular signal transducers, mainly G proteins and arrestins[1]. A key class of GPCR agonists includes small peptides, such as neuropeptides and peptide hormones, which regulate physiological functions ranging from homeostasis and metabolism[2] to complex behaviours like sleep, mechanosensation, and learning[3–6]. However, identifying peptide-GPCR interactions

remains a challenge, and many GPCRs across species are still “orphan” receptors without known endogenous agonists. Phylogenetic analyses have proven to be effective in predicting ligands of peptide GPCRs that are conserved in widely divergent phyla[7–11], but these approaches are less effective for GPCRs with no clear orthologues. Large-scale experimental screens that systematically evaluate extensive libraries of putative bioactive peptides derived through peptidomic analysis[12] or computational tools[13–15] have advanced peptide agonist discovery and GPCR deorphanization[9,11,15,16]. However, systematic screening of putative peptide-GPCR interactions can be resource-intensive, because animal genomes commonly encode dozens to over 100 peptide GPCRs and hundreds of bioactive peptides[9,15–17].

Computational prediction methods for identifying potential GPCR agonists can address these limitations by prioritizing high-probability candidates for experimental validation. Indeed, given the therapeutic importance of GPCRs, numerous machine learning tools have been developed to predict small-molecule GPCR ligands. For example, AiGPro predicts agonist and antagonist activities for human GPCRs using a multi-task framework that combines protein representations (from multiple sequence alignments) and ligand representations (from SMILES strings)[18]. DeepGPCR employs a graph convolutional network to classify GPCR interactions with small molecule ligands[19], while another approach uses a multitask model to predict half-maximal effective concentration (EC_{50}) values for chemical-GPCR pairs[20]. Additionally, structure-based docking tools that simulate molecular interactions using physics-based models are also effective for identifying small molecule ligands[21,22]. Despite advances in machine learning applications for GPCRs and small-molecule docking, computational tools for *in silico* peptide agonist screening are still emerging, in part because the complexity of peptides—larger, more flexible ligands than small molecules—poses a challenge[23]. Docking methods are able to predict peptide-binding poses to GPCRs[24], however they struggle to identify peptide agonists for GPCRs[25], underscoring the need for next-generation computational tools specifically designed to address the unique structural and mechanistic complexities of GPCR-peptide interactions.

Recent advances in protein structure prediction tools, including AlphaFold2 (AF2), are rapidly reshaping structure-based discovery of protein complexes and interactions. The AF2 algorithm[26] predicts protein structure by implicitly learning biophysical relationships between residues through a combination of coevolutionary data and deep learning[27]. These biophysical properties, encoded as statistical patterns in AF2’s protein representations, enable applications beyond structural modelling, such as ligand-binding site prediction[28], protein function prediction[29], and recombinant protein expression[30]. The development of AF-Multimer[31], an extension of AF2, enables accurate prediction of multimeric complexes, including peptide-protein interactions[32,33]. In addition, system-wide experimental screens are expanding our knowledge of GPCR-peptide interactions in diverse animal species, providing high-quality datasets for training machine learning models. In *Caenorhabditis elegans*, we recently constructed a comprehensive resource of over 400 cognate GPCR-peptide agonist pairs and numerous non-agonist pairs by screening > 55 000 GPCR-peptide interactions[9]. Together, these advances—high-quality GPCR-peptide interaction data, generalizable AF2-derived protein representations, and AF-Multimer’s capacity to perform peptide-protein docking—provide a foundation for developing computational methods to predict peptide agonists for GPCRs.

In this study, we harness AF-Multimer-derived structural predictions and hidden layer protein representations to address the challenge of classifying agonist and non-agonist GPCR-peptide interactions. Leveraging a system-wide, experimentally validated *C. elegans* dataset of GPCR-peptide agonist and non-agonist pairs[9], we first evaluate AF-Multimer confidence metrics as indicators of agonist activity, revealing their modest discriminative capability. This capability is enhanced when active-state GPCR-peptide complexes—modelled using templates derived from AF-Multistate[34]—are utilized, aligning receptor confirmations with functional binding sites. Next, we explore AF-Multimer's hidden layer protein representations, revealing that pair representations outperform single representations in distinguishing agonists. We find that subregions within the pair representations contribute distinct, complementary predictive signals, emphasizing the value of multi-faceted feature integration.

Drawing on these insights, we developed DeorphaNN, a graph neural network (GNN) that integrates AF-Multimer-derived structural predictions, interatomic interactions, and protein representations to classify agonist interactions. We demonstrate that this pipeline generalizes to non-*C. elegans* systems, including *Platynereis* and human datasets. Experimental validation of top-ranked peptide agonists—prioritized by DeorphaNN—identifies peptide ligands of two orphan *C. elegans* GPCRs. These results highlight DeorphaNN's ability to accelerate GPCR deorphanization and enhance conventional screening through machine learning guidance.

RESULTS

Dataset preprocessing

We leveraged a previously published dataset that employed a high-throughput CHO $G\alpha_{16}$-mediated calcium signalling assay to screen peptide agonist activity across 161 putative peptide-activated GPCRs from *C. elegans*[9]. We only included GPCRs that showed concentration-dependent activation by at least one peptide, as a lack of activity could have been due to factors such as non-functional expression, misfolding, or failure to couple with $G\alpha_{16}$, rather than the absence of a true agonist. Due to the constraints of AF-Multimer in regard to posttranslational modifications, modified residues were removed from peptide sequences. Following additional filtering—such as removal of short peptides (< 3 residues) and redundant GPCR and peptide sequences (see Methods)—our final dataset contained 65 GPCRs from 55 genes and 339 unique peptides derived from 93 genes (**Figure 1A**). Combining each GPCR with each peptide ("GPCR-peptide complex") yielded 20 035 combinations, all of which were experimentally confirmed to be either agonist or non-agonist matches[9]. The median number of agonists per GPCR in the dataset is 3 and ranges from 1 to 61 (**Figure 1B**). 38 of the 56 GPCRs (68%) with more than one agonist are responsive to peptides derived from a single gene (**Supplemental Figure 1A**). Peptides range in size from 3 residues to 37 residues, with a median length of 11 residues (**Supplemental Figure 1B**). The names and primary amino acid sequences of all GPCRs and peptides used in this study are provided in **Supplemental Data 1**.

AF-Multimer positions peptides in GPCR binding pocket with minimal discrimination

We modelled each GPCR-peptide complex using AF-Multimer to evaluate whether the outputs could distinguish agonists from non-agonists. As experimentally determined structures show that peptide agonists typically interact with the helical cavity ("binding pocket") and extracellular loops of GPCRs[35], we hypothesized that AF-Multimer would

predominantly model agonist peptides within the GPCR binding pocket, while non-agonist peptides might be modelled in less physiologically relevant locations (**Figure 1C**). To standardize our analysis across different receptors, we defined a putative GPCR binding pocket as residues within ±5 positions from the DeepTMHMM[36]-predicted extracellular-membrane boundary. We then computed the minimum peptide-to-pocket residue distance (using a 12.5 Å cutoff) to classify peptides as being "inside" or "outside" the GPCR binding pocket. Our analysis revealed that while AF-Multimer does indeed routinely position agonist peptides in the GPCR binding pocket, it frequently does the same for non-agonists. On average, < 5% of peptides per GPCR were predicted to be further than 12.5 Å from the nearest binding pocket residue, though this varied by GPCR (**Figure 1D**). Across the dataset, only one agonist peptide was modelled outside of the binding pocket of its respective GPCR, while approximately 3.4% of non-agonist peptides were modelled outside the binding pocket (**Figure 1E**). Due to the prevalence of peptides positioned in the binding pocket irrespective of ligand class, this feature alone does not provide a reliable basis for distinguishing between agonist and non-agonist GPCR-peptide complexes.

AF-Multimer confidence metrics partially discriminate peptide agonists from non-agonists

We next evaluated the ability of AF-Multimer confidence metrics (interface pTM [ipTM], predicted local distance difference test [pLDDT], and predicted aligned error [PAE]) to discriminate peptide agonists from non-agonists. The ipTM is a metric that assesses the accuracy of the interface between proteins in the predicted multimeric complex and is given as a single number per multimer, ranging from 0 (low accuracy) to 1 (high accuracy)[31]. The pLDDT is a measure of confidence output as a per-residue score ranging from 0 (low confidence) to 100 (high confidence). Since the transmembrane regions of GPCRs are typically predicted with high confidence by AF2, we focused on confidence metrics of the peptide. pLDDT scores were averaged across peptide residues to yield a single aggregated "peptide pLDDT" score for each GPCR-peptide multimer. Lastly, the PAE quantifies the expected positional error of each residue in the multimer in a pairwise alignment with every other residue, output as a 2D array with values ranging from 0 Å (low positional error) to a cap at 31.75 Å (high positional error). Per-residue PAE scores for the peptide aligned on the GPCR binding pocket residues were averaged to yield a single "pocket PAE" score per GPCR-peptide multimer. For each GPCR-peptide complex, we selected the model with the highest ipTM, highest peptide pLDDT, or lowest pocket PAE from the five models output by AF-Multimer, depending on the metric being analysed.

We evaluated the predictive capability of each metric using mean average precision (mAP). For each GPCR, average precision (AP) measures how agonists rank relative to non-agonists within the receptor's dataset. mAP is then calculated by averaging these AP scores across all receptors, ensuring that GPCRs with few agonists contribute equally to the metric. This approach avoids biases inherent in methods that assign disproportionately higher weight to receptors with many agonists and reduces sensitivity to class imbalance inherent in the dataset. Furthermore, by prioritizing agonists at the top of the ranked peptide list, AP better reflects practical screening scenarios. Analysis of the AP for AF-Multimer confidence metrics (peptide pLDDT, ipTM, and pocket PAE) revealed their ability to distinguish agonists from non-agonists (**Figure 1F, Supplemental Figure 1C-D**). Peptide pLDDT significantly outperformed ipTM and pocket PAE, establishing it as the most reliable indicator of GPCR-peptide agonist interactions among the tested metrics.

Active state GPCR-peptide complexes enhance the ability of peptide pLDDT to discriminate agonists

AF2 is inherently biased towards modelling GPCRs in an inactive conformation[34], likely due to having been exposed to predominantly inactive-state GPCR structures during training. In contrast, agonist binding stabilizes the active conformation of GPCRs, promoting the initiation of intracellular signalling pathways important for signal transduction[37]. Therefore, we hypothesized that biasing the GPCR in each GPCR-peptide complex toward an active conformation would preferentially improve the confidence of AF-Multimer predictions for agonists.

To model active-state GPCR-peptide complexes, we generated GPCR-specific templates using AF-Multistate[34] and refined them through targeted adjustments (**Figure 2A**). First, we trimmed extracellular regions from the template, as these domains mediate ligand binding. This allows AF-Multimer to model peptide interactions without interference from pre-defined extracellular structures. Second, we preserved the transmembrane helices, as shifts to an active-state conformation are driven by displacement of the transmembrane helices[38–40]. Third, we removed low-confidence residues in the intracellular domain (involved in G-protein interactions), preventing inaccuracies in the template from distorting AF-Multimer's predictions of active state GPCR-peptide complexes.

Examples of how active-state predicted GPCR structures impact complex predictions is shown in **Figures 2B–2C**. **Figure 2B** shows an example of a GPCR-peptide agonist complex modelled with and without an active state template, while **Figure 2C** shows the same peptide modelled with a GPCR for which it is not an agonist. Overall, using active state templates improved agonist rank for a subset of low-ranked GPCR-agonist complexes, significantly improving mAP compared to the same complexes modelled without templates (**Figure 2D**). Additionally, AF-Multimer positioned an increased proportion of non-agonist peptides outside of the binding pocket of the GPCR when using active state templates (**Figure 2E-F**). These findings—improved mAP and enhanced exclusion of non-agonist peptides from the binding pocket—suggest that active-state GPCR templates improve discrimination between agonist and non-agonist complexes. Therefore, we used active-state templates to generate GPCR-peptide complexes for all subsequent analyses.

Pairwise protein embeddings outperform single embeddings in discriminating GPCR-peptide agonists

In the process of predicting the 3D structure of proteins, AF2 generates "single" (residue-level) and "pair" (pairwise residue-level) representations that are transformed and updated in the EvoFormer block of the algorithm[26]. Each single representation is a matrix in $\mathbb{R}^{\{n \times 256\}}$ where n is the number of residues in the GPCR-peptide complex; similarly, each pairwise representation is a tensor in $\mathbb{R}^{\{n \times n \times 128\}}$ (**Figure 3A**).

We applied local pooling to residue subregions in each representation to identify features useful for classifying GPCR-peptide agonists. For single representations, we averaged single-residue embeddings across the entire GPCR, the GPCR binding pocket residues, or the peptide residues, producing three different 256-dimensional feature vectors. For pair representations, we averaged pairwise residue embeddings involving GPCR residues, GPCR binding pocket residues, peptide residues, or the intersection of GPCR binding pocket and peptide residues, yielding four distinct 128-dimensional feature vectors.

To assess feature importance across subregions of the single and pair representations, we trained random forest classifiers using stratified group 10-fold cross-validation. This ensures that GPCRs from the same gene gamily are not split across training/validation folds, preventing overfitting to specific families. We found that features derived from the peptide subregions (in both single and pair representations) outperform those from GPCR subregions as measured by receiver operating characteristic (ROC) (**Figure 3B**) and precision-recall (PR) (**Figure 3C**) area under the curve (AUC). While pair representations averaged over GPCR residues or binding pockets significantly outperformed single representations from the same regions in ROC-AUC and PR-AUC, peptide-averaged pair and single representations showed no significant difference in these metrics. However, our dataset exhibits significant variation in agonist distribution: some promiscuous GPCRs have many agonists, while most have few (**Figure 1B**). Metrics like ROC-AUC and PR-AUC weight each GPCR-agonist complex equally, which disproportionately prioritizes promiscuous GPCRs in evaluation. To address this, we used AP, which is computed per GPCR family and weights each family equally regardless of agonist count. This ensures that families with few agonists are not overshadowed by those with many agonists. Our results showed that peptide-averaged pair representations achieve higher mAP scores than single representations (**Figure 3D**). This indicates that pair representations capture more discriminative information for distinguishing peptide agonists across diverse GPCRs than single representations.

Distinct subregions of pair representations capture complementary information relevant to successful classification of GPCR-peptide agonists

Our analysis of different subregions in single and pair representations revealed that pair representations provide valuable information for predicting GPCR-peptide agonists. To further investigate, we compared three key subregions of the pair representations: (1) the interaction subregion (pairwise residue between the GPCR binding pocket and the peptide), (2) the peptide subregion (peptide residues only), and (3) the binding pocket subregion (GPCR binding pocket residues only).

We found that models trained on the different pair representation subregions achieved similar overall mAP scores (**Figure 3E**). However, performance varied significantly across GPCR families depending on the region used for training (**Figure 3F**). For example, the GPCR, SEB-3, showed highest AP when trained on features from the interaction subregion or peptide subregion, but lower precision with models trained on the binding pocket subregion. Conversely, the GPCR, FRPR-15, was best predicted by models trained on features derived from the binding pocket subregion and poorly predicted by models trained on peptide or interaction subregion features. This variability suggests that features from different regions are complementary, with no single subregion universally optimal for all GPCRs.

GNN integration of multimodal GPCR-peptide features for agonist classification

Given the complementary nature of features from different subregions, we generated graph representations of individual GPCR-peptide complexes to train DeorphaNN—a GNN designed to classify agonist interactions by integrating AF-Multimer-derived structural predictions, interatomic interactions, and pairwise protein embeddings (**Figure 4A**). This approach leverages the ability of GNNs to model complex relational data (e.g., residue interactions), enabling agonist classification across diverse GPCR structures by capturing both local and global structural patterns critical for distinguishing agonists and non-agonists.

The graph representations consists of two distinct sets of nodes: GPCR residue nodes and peptide residue nodes. To define the edges between these residue nodes, we use Arpeggio[41], a computational tool designed to identify biophysically meaningful interatomic interactions in protein structures by analysing spatial and physicochemical relationships (e.g., hydrogen bonds, ionic interactions, cation-π interactions, etc.) (**Figure 4B**). Only GPCR residues directly connected to peptide residues via Arpeggio-identified interactions (or within one degree of such interactions) are selected as nodes, ensuring the graphs focus on biophysically relevant regions of each GPCR. In contrast, all peptide residues are included as nodes, as the smaller size of peptides enables their inclusion without excessive computational burden. Additionally, intramolecular connectivity is preserved by connecting nodes with edges based on their primary sequences.

GPCR nodes are loaded with embeddings derived by residue-specific pooling of the pair representation's GPCR subregion, while peptide nodes are loaded with embeddings generated similarly from the peptide subregion (**Figure 4C**). Specifically, for each residue i, we average the i-th row and the i-th column across all features of the respective subregion matrix, yielding a 128-dimensional embedding that captures residue-level structural and interaction information from the pair representation. Edges between GPCR and peptide residues are loaded with pairwise interaction embeddings derived from the interaction subregion of the pair representation. This design ensures that node embeddings encode residue-specific information, incorporating both local and global context from the pair representation, while interaction edges encode intermolecular residue-residue interaction context.

Impact of graph structure and edge embeddings on GNN performance

To evaluate the impact of graph structure and edge features on performance, we conducted an ablation study (**Figure 4D**). We first replaced Arpeggio-defined interaction edges with distance-based edges (residues ≤6 Å apart) and observed a significant reduction in mAP (evaluated via shuffle-groups-out cross-validation with 11 splits). Similarly, removing edge features from Arpeggio-defined edges also decreased performance (**Figure 4D, Supplemental Figure 2A)**. Interestingly, when both changes were combined—using distance-based edges and omitting edge features—the model outperformed either single ablation alone (i.e., distance edges with features or Arpeggio edges without features), although this combined ablation still achieved significantly lower mAP than the complete graph structure (Arpeggio edges with edge features).

These results suggest that Arpeggio-defined edges and edge features play distinct yet complementary roles. Distance-based edges introduce non-functional contacts that add noise when paired with edge features, likely because spatial proximity lacks the physicochemical specificity captured by Arpeggio. Conversely, Arpeggio edges, while more biologically relevant, are sparser and limit information propagation; edge features compensate for this sparsity by encoding interaction context from the pair representation, enabling the GNN to learn meaningful patterns even with fewer edges. When both distance edges and edge features are omitted, the model benefits from reduced noise but lacks interaction-specific guidance, accounting its intermediate performance. Overall, the complete graph structure—using Arpeggio-defined edges and interaction embeddings—outperformed all ablated versions, demonstrating the synergistic value of biophysically meaningful edges and interaction-aware edge features for agonist classification. This framework enables

DeorphaNN to learn complex relational patterns that differentiate agonists from non-agonists (**Figure 4E**). Additionally, DeorphaNN identifies agonists missed by pLDDT (**Figure 4F**).

Validation on phylogenetically diverse data and comparison to state-of-the-art

To assess whether DeorphaNN captures conserved mechanisms of GPCR-peptide agonism rather than species-specific patterns, we evaluated its performance on a dataset from the bilaterian marine annelid, *Platynereis dumerilii*[11]. This dataset contains 17 GPCRs and 124 peptides that were tested in a similar experimental pipeline as the *C. elegans* training dataset (high-throughput CHO $G\alpha_{16}$-mediated calcium signalling assay). We utilized DeorphaNN to rank *Platynereis* peptides for each GPCR based on predicted agonist potential. The mAP across all *Platynereis* GPCRs was significantly greater than the baseline expected with random ranking (**Figure 5A**), demonstrating that DeorphaNN generalizes to non-*C. elegans* datasets.

To assess the model's generalization to human GPCR-peptide interactions, we utilized the subset of human GPCRs from a literature-curated dataset of experimentally confirmed agonist interactions across species[42]. While this dataset includes human GPCR-agonist pairs, it lacks non-agonist interactions. To address this gap, we generated synthetic non-agonist GPCR-peptide pairs using a dissimilarity-based approach inspired by principles from prior work[42]. Specifically, we leveraged ESM-2[43] language model embeddings to quantify peptide dissimilarity. By calculating the Euclidean distance between each peptide's ESM-2 representation and that of the cognate agonists, we selected peptides with a distance exceeding an empirical threshold—chosen to ensure at least 3 non-agonists per GPCR.

To validate the biological relevance of these synthetic pairs, we trained a model using the DeorphaNN framework on the human dataset (augmented with synthetic non-agonist interactions) and tested its performance on the *C. elegans* dataset, which contained no synthetic data (**Figure 5B**). The human-trained GNN achieved a mAP that was significantly higher than random baseline. This cross-species generalization confirms that the human dataset—which combines confirmed agonist pairs with ESM-2-guided synthetic non-agonists—reliably captures real-world interaction patterns and supports the augmented dataset's utility for benchmarking.

We then benchmarked DeorphaNN (*C. elegans*-trained) on the augmented human dataset, achieving a mAP of 0.72—substantially higher than the random baseline of 0.30 (**Figure 5C**). This result further underscores DeorphaNN's ability to generalize across species. For comparison, we evaluated SpatialPPIv2, a state-of-the-art model designed for protein-protein interaction prediction that leverages large language models and graph attention networks to capture structural patterns[44]. On the same augmented human dataset, SpatialPPIv2 achieved a mAP of 0.33, not significantly different from random ranking. DeorphaNN significantly outperformed SpatialPPIv2, demonstrating greater accuracy in predicting GPCR-peptide agonist interactions. This difference highlights that our specialized GNN—which integrates species-specific structural predictions and interaction data—captures biologically relevant features specific to GPCR-peptide agonism more effectively than generalized protein-protein interaction models.

Discovery of novel peptide agonists for orphan GPCRs identified by virtual screening

Many putative peptide-activated GPCRs in *C. elegans* remain orphan despite having been tested in prior large-scale screening experiments. One possibility for lack of agonist activity is that their cognate endogenous agonists were simply absent from the screened peptide

386 libraries. Peptidomics and comparative genomics studies have since expanded the number
387 of peptides known in *C. elegans* beyond those included in the training dataset[45–47], providing
388 an opportunity to reassess *C. elegans* GPCRs with a broader set of candidate ligands.

389 To test whether DeorphaNN can predict novel *C. elegans* GPCR-peptide agonist complexes
390 that were not included in the training dataset, we first applied the model to newly
391 deorphanized receptors, NPR-34 and SEB-2. Neither GPCR was found to have agonist-
392 induced activity with any of the peptides in the training dataset, but subsequent experiments
393 demonstrated that both receptors respond to peptides recently identified through
394 phylogenetic approaches[9]. As shown in **Table 1**, the confirmed agonists for these GPCRs
395 are ranked first out of 364 candidates (339 peptides from the training dataset plus 25
396 recently identified peptides [**Supplemental Data 1**]) by DeorphaNN's predictions and were
397 among the only peptides exceeding the agonist threshold (prediction > 0) for those GPCRs.

398 We then applied DeorphaNN to two orphan *C. elegans* GPCRs with no known ligands,
399 H23L24.4 and NPR-33. The highest-ranked candidate agonists for these receptors were
400 experimentally tested using a CHO $G\alpha_{16}$-mediated calcium signalling assay. We observed
401 robust activation of H23L24.4 by NLP-69-1 (EC_{50} value of 231.7 nM), a peptide that is
402 ranked 5th out of 364 peptides by DeorphaNN (**Table 1, Figure 6A, Supplemental Figure**
403 **2C-D**). Similarly, NPR-33 was potently activated by NLP-70-2 (EC_{50} value of 76.3 nM), the
404 highest-ranked peptide in our predictions (**Table 1, Figure 6B, Supplemental Figure 2C**).
405 Given that GPCRs are often activated by multiple peptides within the same gene family
406 (**Supplemental Figure 1A**), we also tested NLP-70-1 (ranked 156/364 by DeorphaNN,
407 **Table 1**), another putative peptide encoded by the *nlp-70* gene. Although we found
408 activation of the receptor, the EC_{50} value (1.16 μM; **Figure 6C, Supplemental Figure 2C**) is
409 significantly higher than that of the NLP-70-2 peptide, suggesting that DeorphaNN may be
410 intrinsically tuned to prioritize high-potency agonist interactions.

411 **DISCUSSION**

412 In this study, we present DeorphaNN—a GNN-based pipeline for predicting GPCR-peptide
413 agonist interactions *in silico*—offering a scalable strategy for GPCR deorphanization.
414 Previous studies have shown that pLDDT from AlphaFold predictions correlate with
415 interaction affinity[31,32,48], aligning with our findings that pLDDT has partial ability to
416 discriminate agonists from non-agonists. However, affinity alone is insufficient to determine
417 agonist function[49,50]. While affinity is a prerequisite for agonism, it fails to capture additional
418 mechanisms like intrinsic efficacy[51]—the propensity of a ligand to induce conformational
419 changes in the GPCR and activation of downstream G protein signalling.

420 Agonist binding to GPCRs induces conformational changes that drive conserved
421 transmembrane (TM) rearrangements, a hallmark of receptor activation[38,52,53]. A key feature
422 of this process is the outward movement of TM6[54], facilitated by kinking at conserved motifs
423 and disruption of ionic interactions between TM3 and TM6/7[55–57]. These rearrangements
424 reconfigure the surface of the orthosteric binding pocket[58,59], enhancing affinity for
425 agonists[60]. To capitalize on this mechanism, DeorphaNN incorporates active-state structural
426 templates generated using AF-Multistate[34]. This enables the modelling of GPCR-peptide
427 complexes in their active conformations, leveraging AF-Multimer's ability to prioritize
428 peptides that both bind and stabilize the active-state binding pocket.

429 DeorphaNN builds on AF-Multimer by analyzing patterns in the predicted active-state
430 complex to predict whether the bound peptide can activate the receptor. Agonists activate

GPCRs through interactions with key “toggle switch” residues in the GPCR binding pocket[61,62]. Triggering these toggle switches is thought to require two distinct types of intermolecular interactions: “anchor” interactions that provide stability for ligand binding, and “driver” interactions that directly engage with toggle switches to induce conformational changes[63–65]. To enable DeorphaNN to learn patterns corresponding to these interactions, we engineered DeorphaNN to utilize Arpeggio-defined edges, which capture interatomic interactions[41], and AF pairwise embeddings, which encode implicit biophysical relationships between residues[27]. By integrating these features, DeorphaNN learns to distinguish agonists from non-agonists, likely through patterns that correspond to the anchor and driver interactions essential for receptor activation and intrinsic efficacy. In contrast, SpatialPPIv2—a state-of-the-art model for predicting general protein-protein interactions[44]—lacks this context, explaining its poor performance in differentiating agonist from non-agonist interactions (**Figure 5C-D**).

Despite being trained on a relatively small dataset of 457 GPCR-agonist pairs from *C. elegans*, DeorphaNN’s predictions generalize across species. This success reflects the effective application of inductive transfer learning principles, which have revolutionized fields like natural language processing and computer vision by enabling models pretrained on large datasets to be fine-tuned for specialized tasks with limited data[66,67]. In our work, DeorphaNN leverages the AF2 algorithm, which was pretrained on approximately 170 000 protein structures from diverse species[26], and adapts it for the specialized task of GPCR-peptide agonism prediction. The model’s ability to predict agonist interactions in phylogenetically distinct species such as *Platynereis dumerilii* and humans suggests that the implicit biophysical principles learned by AF2 during pretraining were successfully transferred to this specialized task.

Compared to existing methods, DeorphaNN represents a significant advancement in GPCR-peptide agonism prediction. To our knowledge, the only other model that predicts peptide agonists for GPCRs utilizes a support vector machine algorithm trained on peptide descriptors (PD-incorporated SVM), representing sequence patterns of one to five residues that are then one-hot encoded into a numeric array for machine learning[42]. Unlike a purely sequence-based approach, DeorphaNN incorporates structural predictions from AF-Multimer, deep learning-derived protein representations, and intermolecular interaction data. These features enable DeorphaNN to learn relationships from three-dimensional structures and biophysical interaction patterns that are not captured by the primary sequence, providing a more comprehensive and biologically relevant approach to GPCR-peptide agonism prediction across diverse receptors.

Despite its strengths, several factors may limit DeorphaNN’s broader applicability. First, because the dataset used for training was derived from a single experimental screen, it may have missed interactions dependent on specific downstream pathways—such as those involving distinct G protein coupling—potentially misclassifying those complexes as non-agonists. Second, our model’s performance depends on AF2’s ability to accurately predict GPCR-peptide complexes—therefore, inaccurate structural predictions may fail to capture the true nature of GPCR-peptide relationships, impacting DeorphaNN’s effectiveness. Indeed, AF2-predicted GPCR structures demonstrate deviations in the shapes of ligand-binding pockets and extracellular domains when compared to experimentally determined structures[68,69]. Additionally, AF2 cannot account for posttranslational modifications of GPCRs and peptides. For GPCRs, extracellular posttranslational modifications such as tyrosine

477 sulfation, glycosylation, and proteolytic cleavage may alter ligand binding[70]; for peptides,
478 posttranslational modifications such as C-terminal amidation or pyroglutamation are often
479 important for agonist activity[49]. Although direct interactions between GPCRs and the C-
480 terminal amidated group of peptide ligands have not yet been identified[49], amidation has
481 been shown to influence biological activity[71], and the omission of such posttranslational
482 modifications in AF2 predictions may result in lost interaction information. However, with the
483 recent open-source release of AlphaFold3[72], these limitations may be addressed in future
484 work.

485 In this study, we successfully applied DeorphaNN to identify agonists for two previously
486 orphan GPCRs in *C. elegans*. Our results demonstrate that DeorphaNN's predictive
487 capabilities extend beyond this model organism, establishing it as a valuable tool for GPCR
488 deorphanization across diverse species. Furthermore, we anticipate that DeorphaNN may
489 prove instrumental in identifying cognate GPCRs for orphan peptides. In addition to use
490 cases that facilitate the discovery of novel signalling pathways and potential therapeutic
491 targets, DeorphaNN may also be used to accelerate drug discovery efforts. Recent
492 advances in machine learning-based synthetic peptide design have produced tools for
493 generating synthetic ligands for GPCRs[73,74]. DeorphaNN complements these approaches by
494 prioritizing promising candidate agonists for experimental validation, creating a synergistic
495 workflow that accelerates the discovery of GPCR peptide agonists and reduces the need for
496 exhaustive screening.

497 This work highlights the potential of AF-Multimer-derived structural predictions and hidden
498 layer protein representations for identifying peptide agonists of GPCRs. By integrating
499 active-state biasing and pairwise embeddings in a GNN framework, we present a novel deep
500 learning approach to GPCR deorphanization that offers a powerful complement to traditional
501 screening methods. Our model's ability to generalize across species suggests that the
502 predicted structures and protein embeddings capture fundamental principles of GPCR-
503 peptide interactions, positioning this approach as a valuable tool for GPCR deorphanization,
504 the characterization of novel cell signalling pathways, and the identification of novel targets
505 for drug discovery efforts.

506

507 **MATERIALS AND METHODS**

508 **Datasets**

509 The *C. elegans* training dataset was derived from a previously published system-wide
510 screen using CHO $G\alpha_{16}$-mediated calcium signalling assays to assess endogenous peptide
511 agonist activity across *C. elegans* GPCRs[9]. To accommodate AF-Multimer's limitations
512 regarding posttranslational modifications, we preprocessed peptide sequences by removing
513 C-terminal glycine residues for amidated peptides and N-terminal glutamine residues for
514 pyroglutamated peptides. After removing duplicate sequences and peptides shorter than 3
515 residues, 339 unique peptides remained. To minimize the possibility that non-agonist
516 classifications were not artefacts of experimental limitations, we excluded GPCRs lacking
517 concentration-dependent agonist responses in the screening data. Additionally, we excluded
518 DMSR-8-1, which has the same agonist interaction profile as the other DMSR-8 isoform and
519 only differs in sequence by an addition of four residues at the N-terminus. This filtering
520 process resulted in a final training dataset comprising 65 unique GPCRs, each with validated
521 agonist interactions demonstrating concentration-dependent responses. The full list of

522 GPCRs and peptides included in the *C. elegans* training dataset is provided in
523 **Supplemental Data 1**.

524 For the *Platynereis dumerilii* dataset used for cross-species validation, we processed all
525 peptides with modified residues as above and omitted peptides with D-amino acids, resulting
526 in 122 peptides. We omitted GPCRs that had no confirmed agonists among the included
527 peptides, resulting in 18 receptors and 23 agonist interactions.

528 For the human dataset, we utilized a previously published list of GPCR-peptide agonist
529 interactions compiled across diverse species, including vertebrates and inverterbates[42]. We
530 limited our analysis to agonist interactions involving human GPCRs, while using peptides
531 from all species (excluding those longer than 50 residues) to augment the dataset with
532 synthetic non-agonist interactions. Specifically, we generated ESM2-150M[43] embeddings for
533 each peptide and identified, for each GPCR, peptides with Euclidean distances > 3.9 from
534 known agonists, ensuring at least 3 putative non-agonists per GPCR. This process yielded
535 an augmented dataset containing 345 agonist and 3370 non-agonist GPCR-peptide pairs
536 across 82 human GPCRs.

537 In all datasets, we omitted GPCRs that were not predicted by DeepTMHMM to have the
538 characteristic GPCR structure (seven transmembrane domains, with an extracellular N-
539 terminal and an intracellular C-terminal). The full list of GPCRs and peptides included in the
540 validation datasets (*Platynereis* and human) are provided in **Supplemental Data 2**.

541 **AlphaFold2**

542 Primary amino acid sequences for each GPCR-peptide complex were input together into AF-
543 Multimer[26,31] using local ColabFold[75] on a high-performance computing cluster. All modelling
544 settings were left at default, resulting in the generation of 5 predicted structures,
545 corresponding confidence metrics, and hidden layer protein representations (single and
546 pair).

547 For confidence metric analyses, we utilized the model with the highest ipTM, highest peptide
548 pLDDT, or lowest pocket PAE, depending on the metric being analysed. To determine
549 whether the peptide was modelled within the binding pocket, we carried out analysis on the
550 predicted structure with the highest peptide pLDDT, measuring the minimum distance of
551 peptide to binding pocket residues.

552 For all protein representation analyses, we averaged representations across all five AF-
553 Multimer models to prevent learning artefacts from individual model variations.
554 Representations averaged across all models outperformed those derived solely from the top-
555 ranked model (determined by AF-Multimer confidence or by peptide pLDDT, **Supplemental**
556 **Figure 2B**), validating our approach.

557 To bias AF-Multimer predictions towards GPCR active-state conformations, we generated
558 active state GPCR templates using AF2-Multistate[34]. This approach leverages state-
559 annotated GPCR databases to guide predictions with a restricted MSA protocol. The top
560 ranked active state structures were processed by trimming intracellular residues with pLDDT
561 values below 70 and all extracellular residues, based on DeepTMHMM[36] annotations. These
562 processed structures served as GPCR-specific templates in subsequent AF-Multimer
563 predictions for each GPCR-peptide pair, generating state-biased predictions. MSA reduction

was not necessary as the templates were derived from the same sequence as the target GPCRs[76], and retaining the full MSA preserves coevolutionary signals that are vital for driving accurate inter-protein contact predictions[77]. The final Amber relaxation step was omitted in the AF-Multimer pipeline and performed *post hoc* prior to Arpeggio processing.

Mean Average Precision (mAP)

Ranking performance was quantified according to recommendation system metrics[78], using average precision (AP) computed across all peptides (*N*) per GPCR or GPCR family. For each receptor, AP aggregates precision at every rank containing an agonist:

$$AP = \frac{\sum_{k}^{N} \left(\frac{\text{agonists in top } k}{k} \right) \times \ rel@k}{total\ agonists}$$

where $rel@k$ equals 1 if the k-th ranked peptide is relevant (i.e. an agonist) and 0 otherwise. Mean average precision (mAP) averages AP equally across all GPCRs.

To establish a random baseline, we randomized peptide rankings 100 times per GPCR or GPCR family and computed the AP. The average of the 100 APs was then used as the random baseline AP.

Supervised learning algorithms

We implemented random forest classifiers using scikit-learn[79] with default hyperparameters, leveraging their ability to perform implicit feature selection and assess feature importance. To maintain consistency across all cross-validation splits, we used identical random seed and data partitions. For Figures 3B and 3C, we employed stratified group 10-fold cross-validation. Grouping by GPCR was performed to minimize leakage of receptor-specific information between folds. Isoforms from the same GPCR gene were treated as a single unit, ensuring they appear together (55 unique GPCR genes), and class distribution was balanced by stratification. For Figures 3D and 3E, we employed leave-one-group-out cross-validation.

Arpeggio

The top-ranked GPCR-peptide structure according to peptide pLDDT was examined. Peptides located more than 12.5 Å from the binding pocket were excluded from further analysis. Structures were relaxed by Amber[80] and analysed with Arpeggio[41] to identify interatomic interactions between the GPCR and peptide. Specifically, we identified hydrophobic interactions, polar bonds, weak polar bonds, hydrogen bonds, weak hydrogen bonds, ionic bonds, aromatic ring interactions, van der Waals' forces, and amide group interactions.

Graph construction

Graphs were constructed from GPCR-peptide interaction data obtained from AlphaFold2 structural predictions and pair representations. Each graph represents a GPCR-peptide complex, with nodes corresponding to individual residues. All peptide residues were included as nodes, while GPCR residues were included if they directly interacted with peptide residues or were within one degree of such interactions. Nodes were connected by two

types of edges: intramolecular edges connecting nodes according to primary sequence, and intermolecular edges connecting GPCR and peptide nodes according to Arpeggio-identified interactions. For Figure 4D, spatial proximity-based edges (< 6 Å) were also utilized.

Node features were derived using pair representations averaged over 5 models for each residue position. First, we processed the tensor $\boldsymbol{P} \in \mathbb{R}^{n \times n \times 128}$—where n is the total residue count in the GPCR-peptide complex—to extract three sub-tensors:

$$\boldsymbol{P}^{\mathrm{GPCR}} \in \mathbb{R}^{n_{GPCR} \times n_{GPCR} \times 128}$$

$$\boldsymbol{P}^{\mathrm{pep}} \in \mathbb{R}^{n_{pep} \times n_{pep} \times 128}$$

$$\boldsymbol{P}^{\mathrm{int}} = \frac{1}{2}\left(\left(\boldsymbol{P}^{\mathrm{GPCR,pep}} \in \mathbb{R}^{n_{GPCR} \times n_{pep} \times 128}\right) + \left(\boldsymbol{P}^{\mathrm{pep,\,GPCR}} \in \mathbb{R}^{n_{pep} \times n_{GPCR} \times 128}\right)^{T}\right)$$

where n_{GPCR} and n_{pep} are residues of the GPCR and peptide, respectively.

We derived per-residue embeddings for GPCR and peptide nodes by applying cross-axis pooling independently to sub-tensors, $\boldsymbol{P}^{\mathrm{GPCR}}$ and $\boldsymbol{P}^{\mathrm{pep}}$. For each residue i in the sub-tensor $X \in \{\mathrm{GPCR}, \mathrm{pep}\}$, we computed its embedding by averaging within each of the 128 features across the i-th row and the i-th column while correcting for (i, i) duplication:

$$z_i^X = \frac{\left(\sum_{k=1}^{n_X} P_{k,i,:}^X + \sum_{k=1}^{n_X} P_{i,k,:}^X\right) - P_{i,i,:}^X}{2n_X - 1}$$

Intermolecular edge features were derived from the interaction region of the pair representations ($\boldsymbol{P}^{\mathrm{int}}$), with each intermolecular edge loaded with a 128-dimensional feature vector representing the interaction between the residue pairs connected by that edge. Intramolecular edges were assigned a uniform feature vector, computed as the mean-pooled representation of all intermolecular edge features across the graph.

Model architecture and training

The GNN architecture is based on the Graph Attention Network (GAT) framework, specifically utilizing the GATv2Conv[81] layer implementation from PyTorch Geometric. The network begins with a Batch Normalization layer to standardize input features, ensuring stable training dynamics. This is followed by a GATv2Conv layer with 10 attention heads, a dropout rate of 0.5, and 64 hidden channels. The attention mechanism allows the model to dynamically weigh the importance of neighbouring nodes differently when aggregating information, enabling it to focus on relevant structural and interaction patterns within the graph.

After the convolutional layer, a rectified linear unit (ReLU)[82] activation function introduces non-linearity to the model. For graph-level representation learning, a global mean pooling operation aggregates node embeddings across the entire graph, producing a fixed-size vector representation regardless of graph size. This pooled representation is then passed through a final linear layer with a dropout rate of 0.5 for regularization before being fed into a softmax function for classification. The model is trained using the AdamW optimizer[83] with a learning rate of 0.0005 and employs CrossEntropyLoss[84] with label smoothing to mitigate overfitting and improve generalization. During training, graphs are batched and processed using PyTorch Geometric's DataLoader, with edge features incorporated where applicable.

For each validation fold, the training set was subsampled to maintain a balance between positive and negative examples. Specifically, for each positive example (agonist GPCR-peptide complex), four negative examples (non-agonist GPCR-peptide complex) were randomly selected, creating a balanced training set with a 1:4 ratio of positive to negative samples. The model selection was based on validation performance using group shuffle-split cross-validation, with early stopping to prevent overfitting. In this approach, the dataset was divided into groups based on GPCR families, and multiple random splits were generated where each split maintained a specified proportion of groups in the training and validation sets. This method ensures that the model is evaluated on previously unseen GPCR families while maintaining biological relevance in the validation process. The final model for each fold was chosen based on the highest validation average precision score achieved during training. Training was performed for 30 epochs, with the model evaluated periodically on the validation set to monitor convergence and generalization performance.

The final predictions reported were generated using an ensemble of 10 GNNs trained via bootstrap aggregating, with each model trained on a distinct 90% bootstrap sample of the training data. The ensemble outputs were aggregated to produce the final prediction scores.

Aequorin-based GPCR activation assay

An aequorin-based calcium mobilization assay was used to assess GPCR activation following peptide application, as previously described[9]. In brief, reverse pharmacology screening was performed using CHO-K1 cells expressing mitochondrial-targeted apo-aequorin and promiscuous human $G\alpha_{16}$. CHO cells were transfected with pcDNA3.1 plasmids containing the GPCR of interest and harvested two days after transfection. BSA served as a negative control for peptide-evoked calcium responses; ATP activates an endogenous receptor in CHO cells and was used as a positive control. Maximum calcium responses for normalization were obtained by lysing cells at the end of the assay.

Statistical analysis

Statistical analysis was performed in GraphPad Prism version 10.2.2 for macOS (GraphPad Software, San Diego, CA). All statistical tests were two-tailed.

RESOURCE AVAILABILITY

Lead Contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Larissa Ferguson (LariFerg@gmail.com).

Materials Availability

This study did not generate new unique reagents.

Data and Code Availability

All datasets utilized in this study are available as Supplemental Materials and in a Hugging Face repository at https://huggingface.co/datasets/lariferg/DeorphaNN. The DeorphaNN code is available at https://github.com/Zebreu/DeorphaNN.

ACKNOWLEDGEMENTS

We thank Jake Grimmett, Toby Darling, and Ivan Clayson of LMB Scientific Computing. Additional thanks to members of the Schafer lab for helpful discussions. Z.W. acknowledges support from the Max Perutz Fund as part of the GenerationResearch project. W.R.S. acknowledges support from the Medical Research Council (MRC) core grant MC-A023-5PB91 and the Research Foundation – Flanders (FWO) grant G050825N. I.B. acknowledges support from the KU Leuven Research Council grant C16/19/003, the Research Foundation – Flanders (FWO) grant G036524N, and the Baillet Latour Fund.

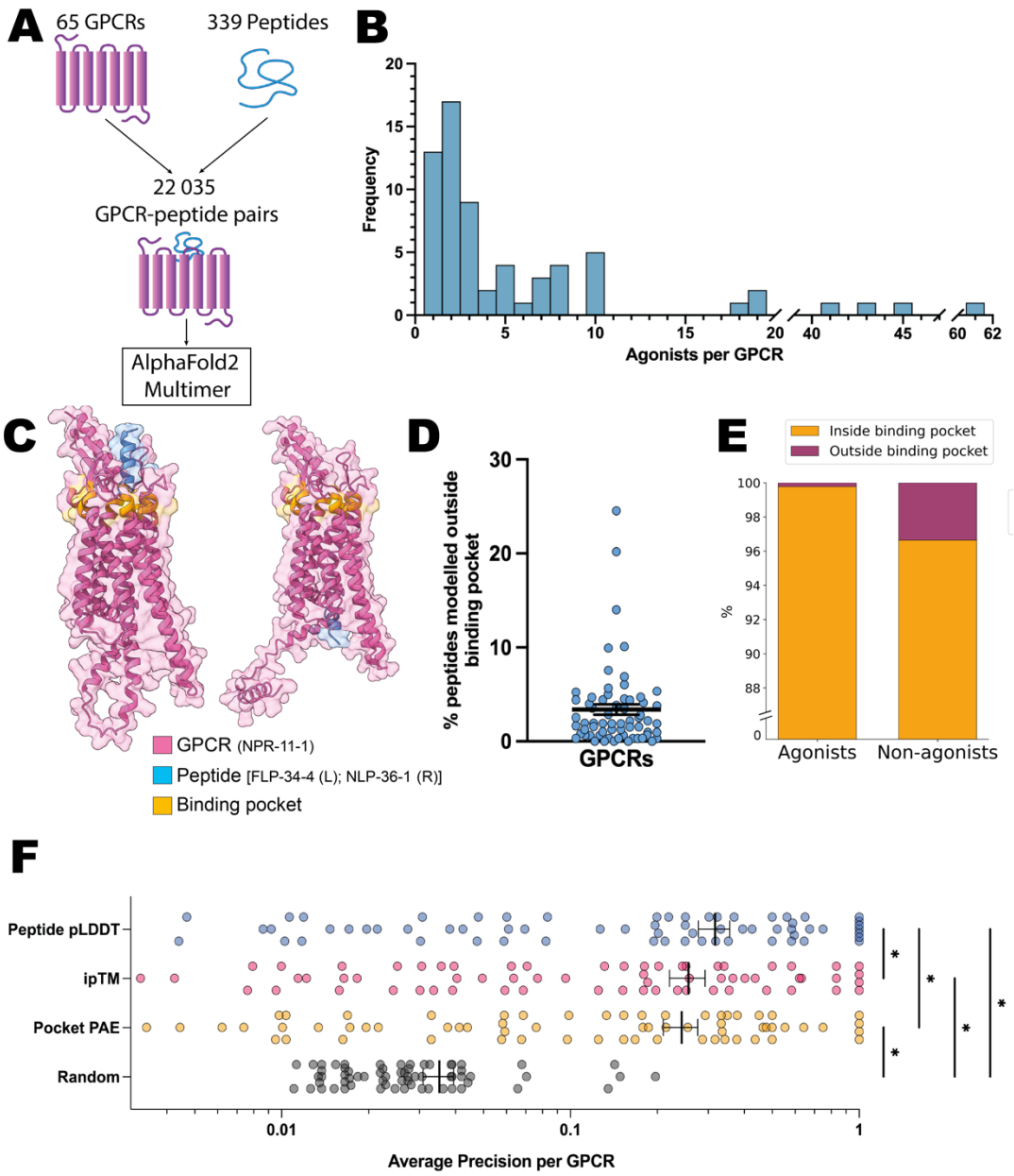
AUTHOR CONTRIBUTIONS

L.F., S.O., T.K.L., W.R.S., and I.B. conceptualized the work; L.F., S.O., and T.K.L developed methodology; L.F., S.O., and T.K.L. performed formal analysis; L.F. and T.K.L. contributed to visualization; S.O., L.F., and C.W. developed software; L.F., S.O., E.V., C.W., Z.W. conducted investigation and experiments; L.F. wrote the original manuscript draft; T.K.L., I.B., S.O., W.R.S., and C.W. reviewed and edited the manuscript; W.R.S., I.B., and L.F. supervised the project. W.R.S. and I.B. provided funding and infrastructure.

DECLARATION OF INTERESTS

The authors declare no competing interests.

699 **FIGURES**



700

701 **Figure 1: Agonist GPCR-peptide complexes demonstrate higher AF-Multimer**
702 **confidence metrics than non-agonist complexes**

703 (**A**) 65 GPCRs and 339 endogenous peptides (22 035 GPCR-peptide pairs) from *C. elegans* were
704 modelled using AF2-Multimer. (**B**) Frequency distribution of the number of agonists per GPCR in the
705 dataset. Most GPCRs have few agonists (median = 3), though a small number of promiscuous
706 GPCRs account for the majority of agonist interactions. (**C**) Example AF-Multimer predicted structures
707 of NPR-11-1 in complex with an agonist (FLP-34-4, left) and a non-agonist (NLP-36-1, right),
708 illustrating two distinct binding configurations. The GPCR is shown in pink, binding pocket residues
709 (defined as ±5 residues from the extracellular-membrane boundary) in yellow, and peptides in blue.
710 (**D**) The percentage of peptides modelled outside the binding pocket for each GPCR ranges from 0 –
711 24.5%, with a mean of 3.4% (±0.6% SEM).(**E**) 99.8% of agonists are modelled within the GPCR

binding pocket, compared to 96.6% of non-agonists. (**F**) Peptide pLDDT significantly outperforms ipTM, pocket PAE, and random ranking in distinguishing agonists from non-agonists (Friedman test with FDR correction using the Benjamini, Krieger, and Yekutieli method, $q < 0.05$). Mean average precision (mAP) values: peptide pLDDT (0.32), ipTM (0.26), pocket PAE (0.24), random ranking (0.035). Error bars indicate mean ± SEM.

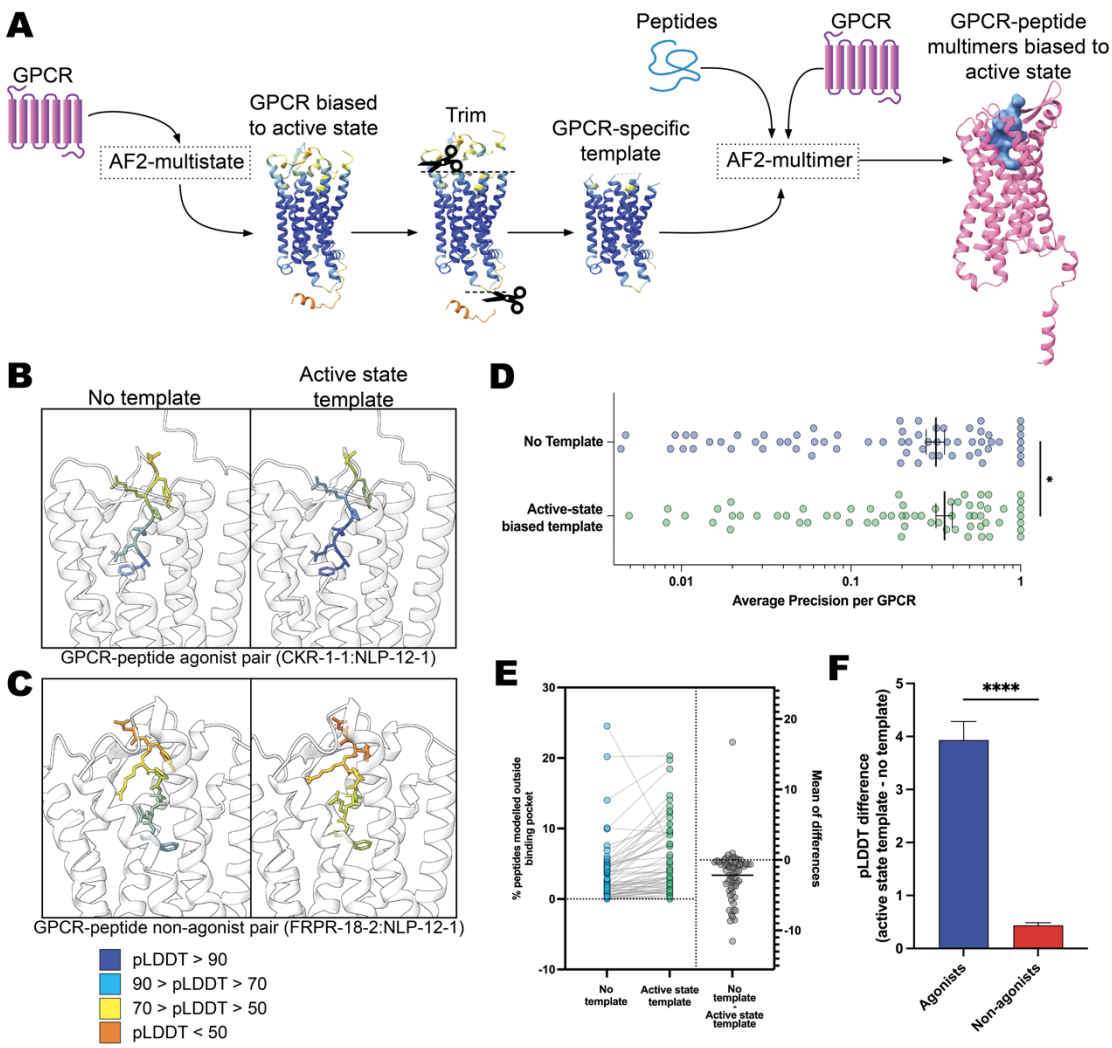
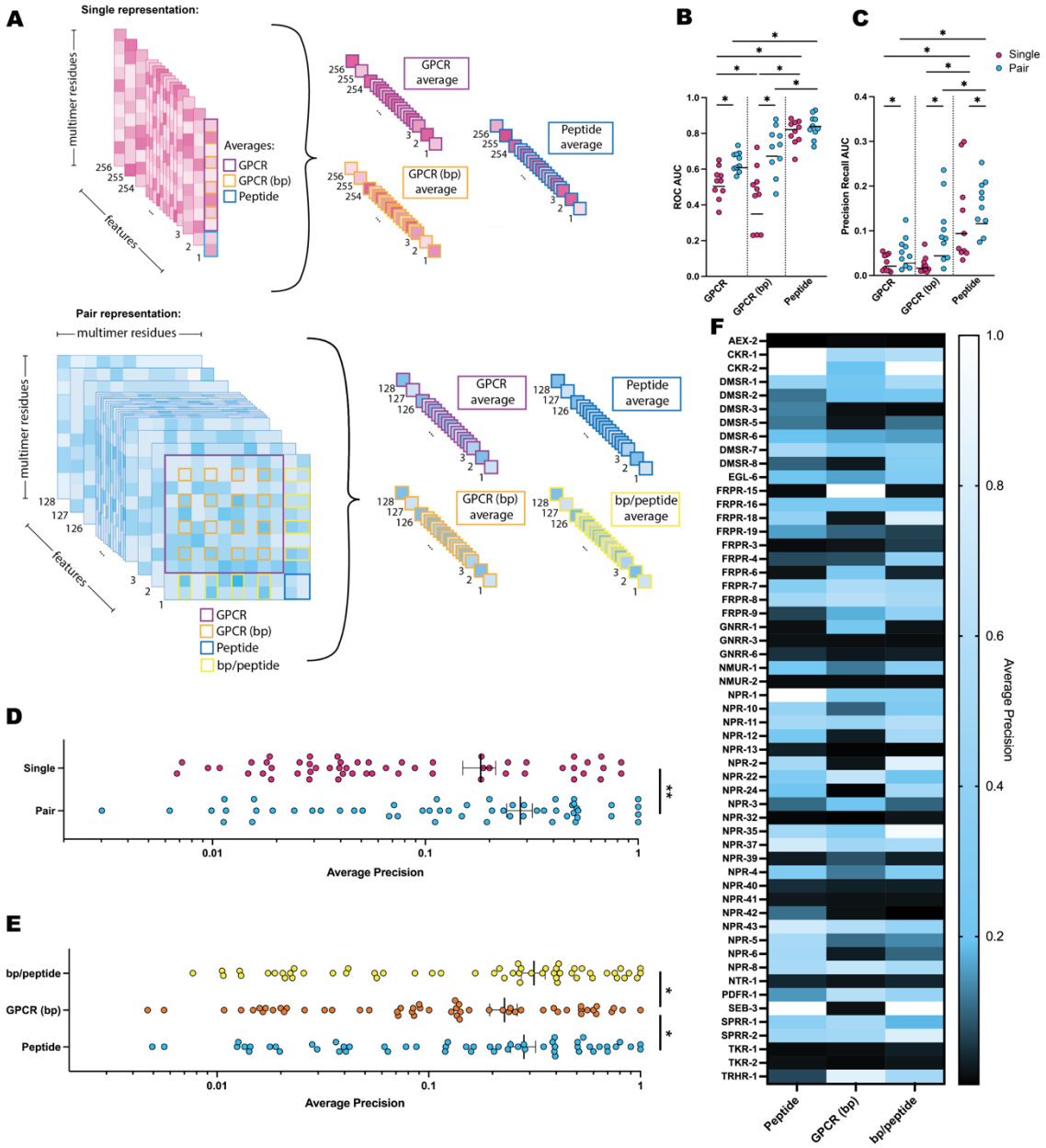


Figure 2: Active state templates enhance AF-Multimer's discrimination of agonist and non-agonist GPCR-peptide interactions

(**A**) Predicted active-state structures were generated for each GPCR using AF-Multistate. After trimming extracellular regions and low confidence (pLDDT < 70) intracellular regions, these templates were then used in AF-Multimer to model GPCR-peptide complexes. (**B**) Representative agonist complex: GPCR CKR-1-1 is modelled with agonist peptide NLP-12-1 without (left) and with (right) an active state template. The use of an active-state template improved peptide pLDDT. (**C**) Representative non-agonist complex: GPCR FRPR-18-2 is modelled with non-agonist NLP-12-1 (the same peptide shown in B) with (left) and without (right) an active-state template. The use of an active-state template does not alter peptide pLDDT. (**D**) Average precision (AP) of agonist peptides ranked by peptide pLDDT significantly improved when using active state templates (mAP: 0.36 vs. 0.32 without templates; Wilcoxon signed-rank test, $p < 0.05$). Error bars indicate mean ± SEM. (**E**) Active state templates increased the percentage of peptides modelled outside the GPCR binding pocket (mean: 5.5% ±0.6% SEM vs. 3.4% ±0.6% without templates; paired t-test, $p < 0.0001$). (**F**) The improvement in peptide pLDDT with active state templates was significantly greater for agonists compared to non-agonists (Welch's t-test, $p < 0.0001$). Error bars represent mean ± SEM.

Figure 3: Subregions of AF-



Multimer pair representations provide complementary information for classifying agonist and non-agonist peptides

(**A**) Schematic illustration of feature extraction from AF-Multimer representations. Single representations ($\mathbb{R}^{\{n \times 128\}}$) and pair representations ($\mathbb{R}^{\{n \times n \times 128\}}$) were processed by aggregating features via local average pooling over predefined subregions: peptide residues (blue), GPCR residues (purple), GPCR binding pocket (bp) residues (green), and their intersection (yellow; pair representations only). (**B**) Random forest models trained on pair representations significantly outperform single representations. Two-way repeated measures ANOVA shows significant interaction between the representation type ($F(1,56) = 21.72$, $p < 0.0001$) and the subregion used ($F(2,56) = 30.25$, $p < 0.0001$). Multiple comparisons (FDR-corrected by method of Benjamini, Krieger, and Yekutieli) reveal that peptide subregion performance exceeded that of GPCR or GPCR bp subregions, but did not show a significant difference between representation types (* $q < 0.05$). (**C**) Precision-recall (PR) AUC analysis mirrors ROC results. Two-way repeated measures ANOVA shows significant interaction between the representation type ($F(1,56) = 8.602$, $p < 0.01$) and the subregion used ($F(2,56) = 16.88$, $p < 0.0001$). Multiple comparisons (FDR-corrected by method of Benjamini,

752 Krieger, and Yekutieli) reveal that peptide subregion performance exceeded that of GPCR or GPCR
753 bp subregions and showed a significant difference between representation types (* $q < 0.05$). (**D**) A
754 more rigorous evaluation using leave-one-group-out cross-validation (GPCR gene families as groups)
755 further reveals that random forest models trained on peptide pair representations significantly
756 outperform single representations. The average precision (AP) per GPCR family was calculated.
757 Mean AP (mAP) values: single, 0.18; pair, 0.28. Error bars denote mean ± SEM. (**E**) Pair
758 representation subregions contribute distinct predictive signals. Leave-one-group-out analysis
759 revealed that peptide (mAP: 0.28) and peptide-GPCR bp intersection regions (0.31) outperformed
760 GPCR bp alone (0.23) in agonist prediction (*$q<0.01$, Friedman test with FDR correction). (**F**)
761 Heatmap of AP values from (E) illustrates that different GPCR families benefit from information
762 provided by different subregions. Distinct GPCR families rely on different structural features for
763 agonist recognition, with the most informative subregion varying depending on the specific GPCR
764 family.
765

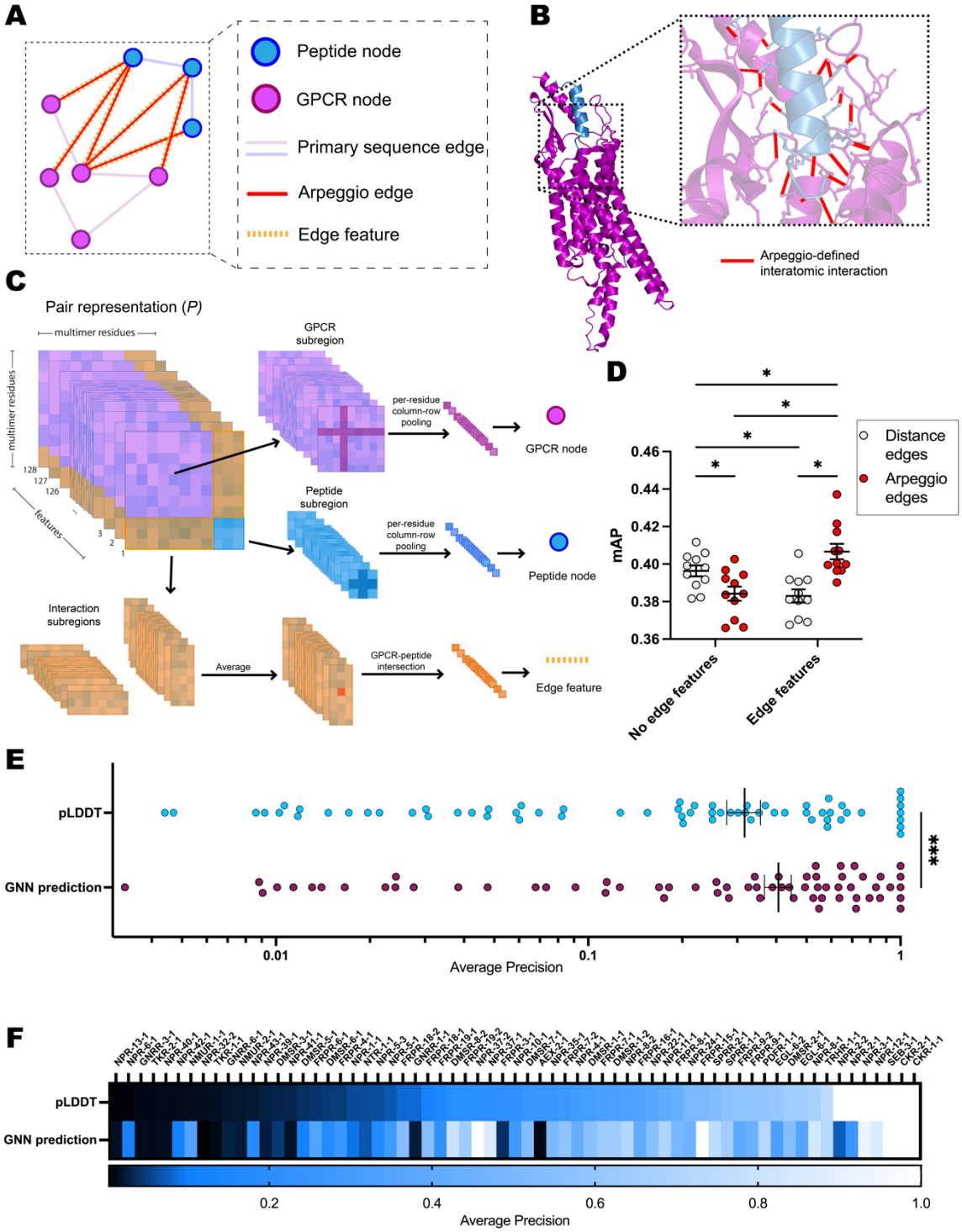


Figure 4: DeorphaNN integrates AF-derived structures and protein representations to predict agonist interactions

(**A**) Graph representations of GPCR-peptide complexes. Intra-protein edges are defined by primary sequence, while inter-protein edges are defined by Arpeggio. Peptide residues (all included) and GPCR residues (those directly connected to peptides via Arpeggio-identified interactions or within one degree of such interactions) serve as nodes. Intermolecular edges connect GPCR and peptide residues, while intramolecular edges preserve primary sequence connectivity. (**B**) Arpeggio identifies biophysically meaningful interatomic interactions (e.g., hydrogen bonds, ionic interactions), which

define intermolecular edges between GPCR and peptide residue nodes. (**C**) AF-Multimer pair representations are split into GPCR, peptide, and interaction subregions. Residue-specific pooling of GPCR and peptide subregions generates 128-dimensional embeddings for respective nodes. Interaction edges are annotated with pairwise interaction embeddings from the interaction subregion. (**D**) Arpeggio-defined edges and use of interaction edge features synergistically improve model performance as determined by shuffle-groups-out cross-validation with 11 splits. Two-way repeated measures ANOVA shows significant interaction between the use of edge features, and method of drawing edges (Distance based cutoff vs. Arpeggio) ($F(1, 10) = 28.75$, $p = 0.0003$. Multiple comparisons (FDR-corrected by method of Benjamini, Krieger, and Yekutieli) reveal that loss of edge features and/or Arpeggio defined edges significantly reduces model performance (* $q < 0.05$). Error bars indicate mean ± SEM. (**E**) DeorphaNN outperforms peptide pLDDT in agonist prediction (mAP: 0.41 vs. 0.32; Wilcoxon signed-rank test, $p < 0.001$). Error bars indicate mean ± SEM. (**F**) Heatmap of AP values demonstrates that DeorphaNN identifies agonists missed by peptide pLDDT across various GPCR families

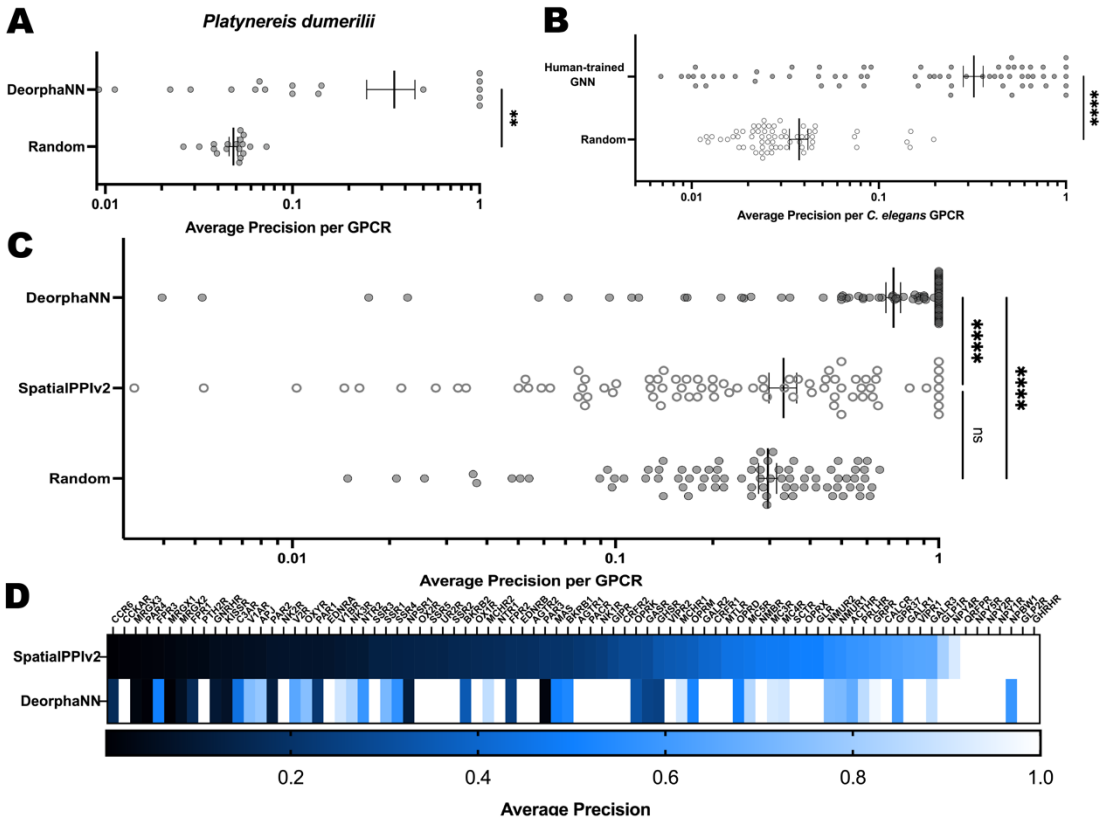


Figure 5: DeorphaNN generalizes to non-*C. elegans* GPCR-peptide agonists

(**A**) DeorphaNN, trained on *C. elegans* data, identifies peptide agonists for GPCRs from *Platynereis dumerilii* with a mAP of 0.35, significantly higher than random ranking. (Wilcoxon matched-pairs signed rank test, ** $p < 0.01$). (**B**) Validation of a human GPCR-peptide agonist dataset, augmented with ESM-2-guided synthetic non-agonists. A model trained on this human dataset achieved a mAP of 0.32 when predicting agonists in the *C. elegans* dataset, which contains no synthetic data. This performance was significantly greater than random ranking (Wilcoxon signed-rank test, **** $p < 0.0001$). (**C**) DeorphaNN significantly outperforms SpatialPPIv2 and random ranking on human GPCR-peptide agonist identification (DeorphaNN mAP = 0.72 vs. SpatialPPIv2 mAP = 0.33 and random mAP = 0.30). Statistical significance was determined using the Friedman test with Dunn's correction (**** $p<0.0001$). Error bars indicate mean ± SEM. (**D**) Heatmap for the per-GPCR AP values from (C), illustrating that DeorphaNN identifies agonists missed by SpatialPPIv2.

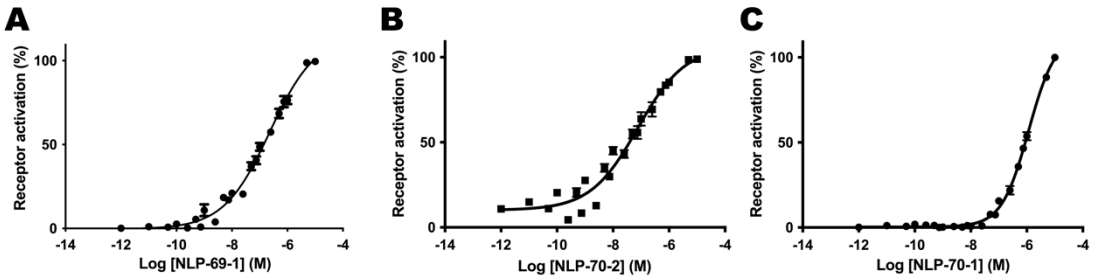


Figure 6: DeorphaNN-guided identification and experimental validation of agonists for two *C. elegans* GPCRs

(**A**) Concentration response curve for CHO cells co-expressing the GPCR H23L24-4-1 with $G\alpha_{16}$ and the calcium indicator aequorin in response to the peptide NLP-69-1. Responses are shown relative to 100% activation after normalization to the total calcium response. EC_{50} = 231.7 nM, n = 6. (**B**) Concentration response curve of the GPCR NPR-33-1 to the peptide, NLP-70-2. EC_{50} = 76.3 nM, n = 6. (**C**) Concentration response curve of the GPCR NPR-33-1 to the peptide NLP-70-1. EC_{50} = 1.16 μM, n = 6.

TABLES

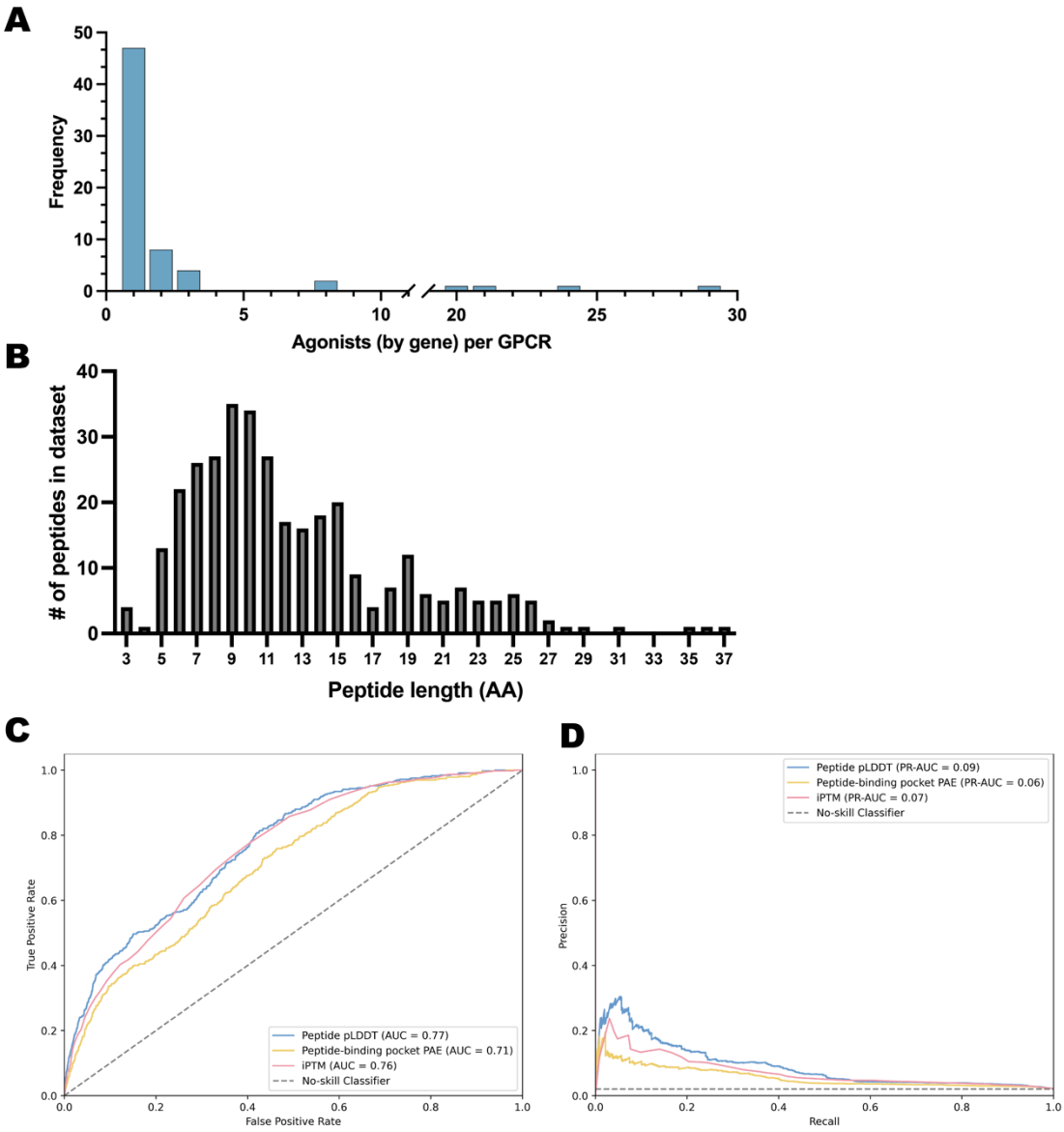
Table 1: GPCR-peptide agonist pairs identified by DeorphaNN

GPCR	GPCR Uniprot ID	Peptide	Peptide sequence	Rank by DeorphaNN	Source
NPR-34	G5EDR6	SNET-1-1	LDCRKFSFAPACRGIML	1/364	Beets et al.[9]
SEB-2	P30650	NLP-73-3	NRQCLLNAGLSQGCDFSDLL HAQTQARKFMSFAGP	1/364	Beets et al.[9]
H23L24.4	Q9N5L5	NLP-69-1	LHRIGGNIVM	5/364	This paper
NPR-33	O62189	NLP-70-2	WYDWQNVPHALQQ	1/364	This paper
		NLP-70-1	WYEWNNDMEIT	156/364	This paper

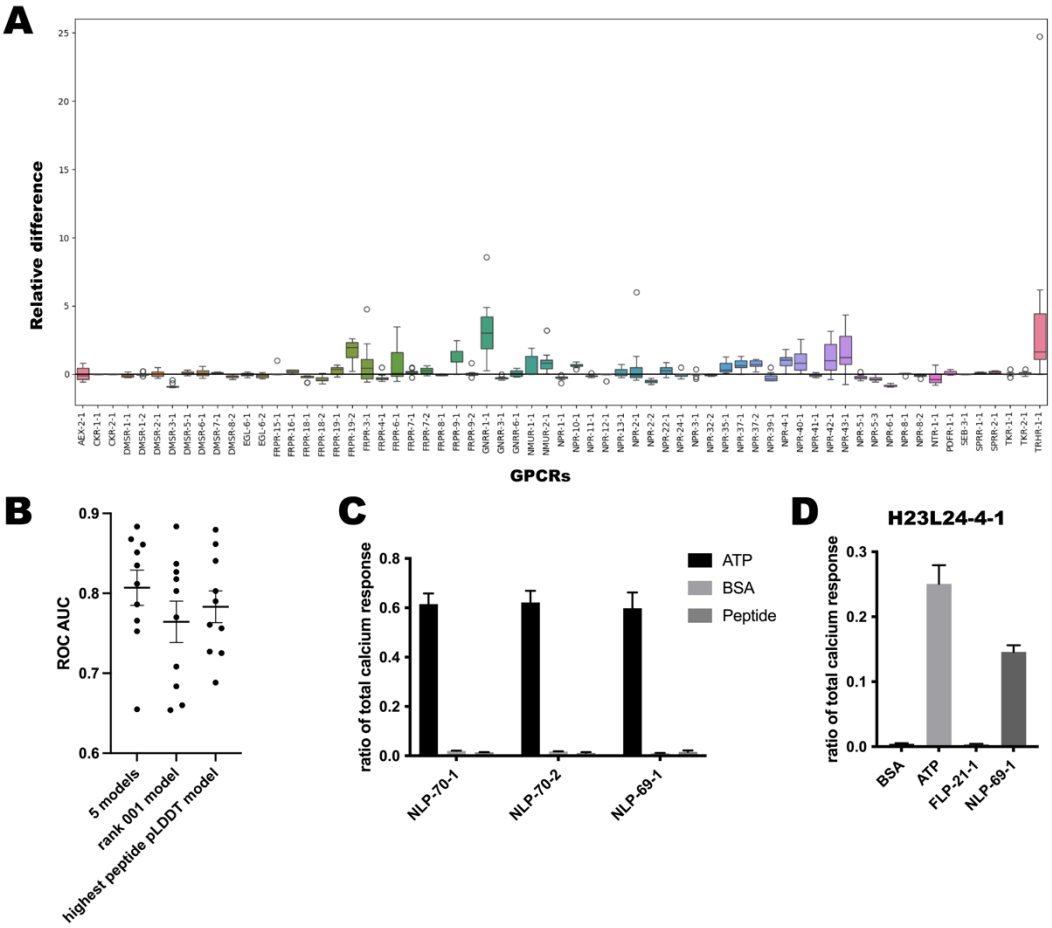
SUPPLEMENTAL INFORMATION

Supplemental Data 1: A list of the names and primary sequences of all GPCRs and peptides in the *C. elegans* training dataset, with a list of the experimentally validated agonist pairs.

Supplemental Data 2: A list of the names and primary sequences of all GPCRs and peptides in the validation datasets (*Platynereis* and human), with agonist pairs labelled.



Supplemental Figure 1
(**A**) Frequency distribution of the number of agonists by gene per GPCR in the *C. elegans* dataset. (**B**) Frequency distribution of the amino acid count per peptide in the *C. elegans* dataset. (**C**) Receiver operating characteristic (ROC) curves and (**D**) Precision-recall (PR) curves for the ipTM, peptide pLDDT, and pocket PAE confidence metrics, illustrating the ability of these confidence metrics to classify agonist and non-agonist GPCR-peptide pairs from the *C. elegans* dataset.



Supplemental Figure 2

(**A**) Relative difference in performance per GPCR between Arpeggio-defined edges with edge features and without edge features, as determined by shuffle-groups-out cross-validation with 11 splits. The improvement in performance varies across GPCRs. (**B**) Random forest classifier trained on the peptide region of the pair representation (as described in Figure 3) performs better when the pair representation is averaged across all 5 models output by AF-Multimer (AUC >0.80) compared to using the "rank 001" model (**C**) No response is seen in cells transfected with empty plasmid (no GPCR) in response to NLP-70-1, NLP-70-2, or NLP-69-1 peptides (10 uM). (**D**) CHO cells expressing the GPCR, H23L24-4-1, have a calcium response to 10 uM of the peptide NLP-69-1, but not FLP-21-1.

REFERENCES

1. Lorente, J. S. *et al.* GPCR drug discovery: new agents, targets and indications. *Nat. Rev. Drug Discov.* (2025) doi:10.1038/s41573-025-01139-y.

2. Atkins, N. *et al.* Functional Peptidomics: Stimulus- and Time-of-Day-Specific Peptide Release in the Mammalian Circadian Clock. *ACS Chem. Neurosci.* **9**, 2001–2008 (2018).

853 3. Bhat, U. S., Shahi, N., Surendran, S. & Babu, K. Neuropeptides and Behaviors: How
854 Small Peptides Regulate Nervous System Function and Behavioral Outputs. *Front. Mol.*
855 *Neurosci.* **14**, 786471 (2021).
856 4. De Fruyt, N., Yu, A. J., Rankin, C. H., Beets, I. & Chew, Y. L. The role of neuropeptides
857 in learning: Insights from C. elegans. *Int. J. Biochem. Cell Biol.* **125**, 105801 (2020).
858 5. Jékely, G. *et al.* The long and the short of it – a perspective on peptidergic regulation of
859 circuits and behaviour. *J. Exp. Biol.* **221**, jeb166710 (2018).
860 6. Chew, Y. L. *et al.* An Afferent Neuropeptide System Transmits Mechanosensory Signals
861 Triggering Sensitization and Arousal in C. elegans. *Neuron* **99**, 1233-1246.e6 (2018).
862 7. Jékely, G. Global view of the evolution and diversity of metazoan neuropeptide signaling.
863 *Proc. Natl. Acad. Sci.* **110**, 8702–8707 (2013).
864 8. Hsueh, A. J. W. & Feng, Y. Discovery of polypeptide ligand-receptor pairs based on their
865 co-evolution. *FASEB J.* **34**, 8824–8832 (2020).
866 9. Beets, I. *et al.* System-wide mapping of peptide-GPCR interactions in C. elegans. *Cell*
867 *Rep.* **42**, 113058 (2023).
868 10. Mirabeau, O. & Joly, J.-S. Molecular evolution of peptidergic signaling systems in
869 bilaterians. *Proc. Natl. Acad. Sci.* **110**, (2013).
870 11. Bauknecht, P. & Jékely, G. Large-Scale Combinatorial Deorphanization of
871 Platynereis Neuropeptide GPCRs. *Cell Rep.* **12**, 684–693 (2015).
872 12. Schulz-Knappe, P., Hans-Dieter, Z., Heine, G., Jurgens, M. & Schrader, M.
873 Peptidomics The Comprehensive Analysis of Peptides in Complex Biological Mixtures.
874 *Comb. Chem. High Throughput Screen.* **4**, 207–217 (2012).
875 13. Ma, Y. *et al.* Identification of antimicrobial peptides from the human gut microbiome
876 using deep learning. *Nat. Biotechnol.* **40**, 921–931 (2022).

877 14. Teufel, F. *et al.* DeepPeptide predicts cleaved peptides in proteins using conditional
878 random fields. *Bioinformatics* **39**, btad616 (2023).
879 15. Foster, S. R. *et al.* Discovery of Human Signaling Systems: Pairing Peptides to G
880 Protein-Coupled Receptors. *Cell* **179**, 895-908.e21 (2019).
881 16. Thiel, D. *et al.* Large-scale deorphanization of Nematostella vectensis neuropeptide G
882 protein-coupled receptors supports the independent expansion of bilaterian and cnidarian
883 peptidergic systems. *eLife* **12**, RP90674 (2024).
884 17. Jékely, G. & Yuste, R. Nonsynaptic encoding of behavior by neuropeptides. *Curr.*
885 *Opin. Behav. Sci.* **60**, 101456 (2024).
886 18. Brahma, R., Moon, S., Shin, J.-M. & Cho, K.-H. AiGPro: a multi-tasks model for
887 profiling of GPCRs for agonist and antagonist. *J. Cheminformatics* **17**, 12 (2025).
888 19. Zhang, H. *et al.* Revolutionizing GPCR–ligand predictions: DeepGPCR with
889 experimental validation for high-precision drug discovery. *Brief. Bioinform.* **25**, bbae281
890 (2024).
891 20. Huang, W.-C., Lin, W.-T., Hung, M.-S., Lee, J.-C. & Tung, C.-W. Decrypting orphan
892 GPCR drug discovery via multitask learning. *J. Cheminformatics* **16**, 10 (2024).
893 21. Ballante, F., Kooistra, A. J., Kampen, S., De Graaf, C. & Carlsson, J. Structure-Based
894 Virtual Screening for Ligands of G Protein–Coupled Receptors: What Can Molecular
895 Docking Do for You? *Pharmacol. Rev.* **73**, 1698–1736 (2021).
896 22. Luttens, A. *et al.* Rapid traversal of vast chemical space using machine learning-
897 guided docking screens. *Nat. Comput. Sci.* (2025) doi:10.1038/s43588-025-00777-x.
898 23. Ciemny, M. *et al.* Protein–peptide docking: opportunities and challenges. *Drug*
899 *Discov. Today* **23**, 1530–1537 (2018).

900 24. Badaczewska-Dawid, A. E., Kmiecik, S. & Koliński, M. Docking of peptides to
901 GPCRs using a combination of CABS-dock with FlexPepDock refinement. *Brief.*
902 *Bioinform.* **22**, bbaa109 (2021).
903 25. Ciancetta, A. *et al.* A Multi-Angle Approach to Predict Peptide-GPCR Complexes:
904 The N/OFQ-NOP System as a Successful AlphaFold Application Case Study. *J. Chem.*
905 *Inf. Model.* **64**, 8034–8051 (2024).
906 26. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature*
907 **596**, 583–589 (2021).
908 27. Roney, J. P. & Ovchinnikov, S. State-of-the-Art Estimation of Protein Model
909 Accuracy Using AlphaFold. *Phys. Rev. Lett.* **129**, 238101 (2022).
910 28. Gazizov, A., Lian, A., Goverde, C., Ovchinnikov, S. & Polizzi, N. F. AF2BIND:
911 Predicting ligand-binding sites using the pair representation of AlphaFold2. Preprint at
912 https://doi.org/10.1101/2023.10.15.562410 (2023).
913 29. Hu, M. *et al.* Exploring evolution-aware & -free protein language models as protein
914 function predictors. Preprint at https://doi.org/10.48550/arXiv.2206.06583 (2022).
915 30. Ouellet, S., Ferguson, L., Lau, A. Z. & Lim, T. K. Y. CysPresso: a classification
916 model utilizing deep learning protein representations to predict recombinant expression of
917 cysteine-dense peptides. *BMC Bioinformatics* **24**, 200 (2023).
918 31. Evans, R. *et al.* Protein complex prediction with AlphaFold-Multimer. *bioRxiv*
919 2021.10.04.463034 (2022) doi:10.1101/2021.10.04.463034.
920 32. Tsaban, T. *et al.* Harnessing protein folding neural networks for peptide–protein
921 docking. *Nat. Commun.* **13**, 176 (2022).
922 33. Johansson-Åkhe, I. & Wallner, B. Improving peptide-protein docking with
923 AlphaFold-Multimer using forced sampling. *Front. Bioinforma.* **2**, 959160 (2022).

924 34. Heo, L. & Feig, M. Multi-state modeling of G-protein coupled receptors at
925 experimental accuracy. *Proteins Struct. Funct. Bioinforma.* **90**, 1873–1885 (2022).
926 35. Tikhonova, I. G., Gigoux, V. & Fourmy, D. Understanding Peptide Binding in Class
927 A G Protein-Coupled Receptors. *Mol. Pharmacol.* **96**, 550–561 (2019).
928 36. Jeppe Hallgren, A. K., Konstantinos D. Tsirigos, Mads D. Pedersen, José Juan
929 Almagro Armenteros, Paolo Marcatili, Henrik Nielsen & Winther, O. DeepTMHMM
930 predicts alpha and beta transmembrane proteins using deep neural networks. (2022)
931 doi:https://doi.org/10.1101/2022.04.08.487609.
932 37. Lopez-Balastegui, M. *et al.* Relevance of G protein-coupled receptor (GPCR)
933 dynamics for receptor activation, signalling bias and allosteric modulation. *Br. J.*
934 *Pharmacol.* bph.16495 (2024) doi:10.1111/bph.16495.
935 38. Zhou, Q. *et al.* Common activation mechanism of class A GPCRs. *eLife* **8**, e50279
936 (2019).
937 39. Latorraca, N. R., Venkatakrishnan, A. J. & Dror, R. O. GPCR Dynamics: Structures
938 in Motion. *Chem. Rev.* **117**, 139–155 (2017).
939 40. Weis, W. I. & Kobilka, B. K. The Molecular Basis of G Protein-Coupled Receptor
940 Activation. *Annu. Rev. Biochem.* **87**, 897–919 (2018).
941 41. Jubb, H. C. *et al.* Arpeggio: A Web Server for Calculating and Visualising
942 Interatomic Interactions in Protein Structures. *J. Mol. Biol.* **429**, 365–371 (2017).
943 42. Shiraishi, A. *et al.* Repertoires of G protein-coupled receptors for Ciona-specific
944 neuropeptides. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 7847–7856 (2019).
945 43. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a
946 language model. *Science* **379**, 1123–1130 (2023).

947 44. Hu, W. & Ohue, M. SpatialPPIv2: Enhancing protein–protein interaction prediction
948 through graph neural networks with protein language models. *Comput. Struct. Biotechnol.*
949 *J.* **27**, 508–518 (2025).
950 45. McKay, F. M. *et al.* In silico analyses of neuropeptide-like protein (NLP) profiles in
951 parasitic nematodes. *Int. J. Parasitol.* **52**, 77–85 (2022).
952 46. Van Bael, S. *et al.* A *Caenorhabditis elegans* Mass Spectrometric Resource for
953 Neuropeptidomics. *J. Am. Soc. Mass Spectrom.* **29**, 879–889 (2018).
954 47. Koziol, U. *et al.* De novo discovery of neuropeptides in the genomes of parasitic
955 flatworms using a novel comparative approach. *Int. J. Parasitol.* **46**, 709–721 (2016).
956 48. Chang, L. & Perez, A. Ranking Peptide Binders by Affinity with AlphaFold**.
957 *Angew. Chem. Int. Ed.* **62**, e202213362 (2023).
958 49. Tikhonova, I. G., Gigoux, V. & Fourmy, D. Understanding Peptide Binding in Class
959 A G Protein-Coupled Receptors. *Mol. Pharmacol.* **96**, 550–561 (2019).
960 50. Roche, D., Gil, D. & Giraldo, J. Mechanistic analysis of the function of agonists and
961 allosteric modulators: reconciling two-state and operational models. *Br. J. Pharmacol.*
962 **169**, 1189–1202 (2013).
963 51. Clarke, W. P. & Bond, R. A. The elusive nature of intrinsic efficacy. *Trends*
964 *Pharmacol. Sci.* **19**, 270–276 (1998).
965 52. Venkatakrishnan, A. J. *et al.* Diverse activation pathways in class A GPCRs converge
966 near the G-protein-coupling region. *Nature* **536**, 484–487 (2016).
967 53. Tehan, B. G., Bortolato, A., Blaney, F. E., Weir, M. P. & Mason, J. S. Unifying
968 Family A GPCR Theories of Activation. *Pharmacol. Ther.* **143**, 51–60 (2014).
969 54. Rasmussen, S. G. F. *et al.* Crystal structure of the β2 adrenergic receptor–Gs protein
970 complex. *Nature* **477**, 549–555 (2011).

971 55. Ballesteros, J. A. *et al.* Activation of the β2-Adrenergic Receptor Involves Disruption
972 of an Ionic Lock between the Cytoplasmic Ends of Transmembrane Segments 3 and 6. *J.*
973 *Biol. Chem.* **276**, 29171–29177 (2001).
974 56. Shi, L. *et al.* β2 Adrenergic Receptor Activation. *J. Biol. Chem.* **277**, 40989–40996
975 (2002).
976 57. Venkatakrishnan, A. J. *et al.* Molecular signatures of G-protein-coupled receptors.
977 *Nature* **494**, 185–194 (2013).
978 58. Liu, F. *et al.* GPCR-BSD: a database of binding sites of human G-protein coupled
979 receptors under diverse states. *BMC Bioinformatics* **25**, 343 (2024).
980 59. Liessmann, F., Künze, G. & Meiler, J. Improving the Modeling of Extracellular
981 Ligand Binding Pockets in RosettaGPCR for Conformational Selection. *Int. J. Mol. Sci.*
982 **24**, 7788 (2023).
983 60. Warne, T., Edwards, P. C., Doré, A. S., Leslie, A. G. W. & Tate, C. G. Molecular
984 basis for high-affinity agonist binding in GPCRs. *Science* **364**, 775–778 (2019).
985 61. Filipek, S. Molecular switches in GPCRs. *Curr. Opin. Struct. Biol.* **55**, 114–120
986 (2019).
987 62. Hauser, A. S. *et al.* GPCR activation mechanisms across classes and
988 macro/microscales. *Nat. Struct. Mol. Biol.* **28**, 879–888 (2021).
989 63. Zhang, M. *et al.* G protein-coupled receptors (GPCRs): advances in structures,
990 mechanisms and drug discovery. *Signal Transduct. Target. Ther.* **9**, 88 (2024).
991 64. Nussinov, R. & Tsai, C.-J. Unraveling structural mechanisms of allosteric drug action.
992 *Trends Pharmacol. Sci.* **35**, 256–264 (2014).
993 65. Nussinov, R., Tsai, C.-J. & Liu, J. Principles of Allosteric Interactions in Cell
994 Signaling. *J. Am. Chem. Soc.* **136**, 17692–17701 (2014).

995 66. Howard, J. & Ruder, S. Universal Language Model Fine-tuning for Text
996 Classification. Preprint at https://doi.org/10.48550/arXiv.1801.06146 (2018).
997 67. Panda, A., Panigrahi, D., Mitra, S., Mittal, S. & Rahimi, S. Transfer Learning Applied
998 to Computer Vision Problems: Survey on Current Progress, Limitations, and
999 Opportunities. Preprint at https://doi.org/10.48550/ARXIV.2409.07736 (2024).
1000 68. Karelina, M., Noh, J. J. & Dror, R. O. How accurately can one predict drug binding
1001 modes using AlphaFold models? *eLife* **12**, RP89386 (2023).
1002 69. He, X. *et al.* AlphaFold2 versus experimental structures: evaluation on G protein-
1003 coupled receptors. *Acta Pharmacol. Sin.* **44**, 1–7 (2023).
1004 70. Goth, C. K., Petäjä-Repo, U. E. & Rosenkilde, M. M. G Protein-Coupled Receptors in
1005 the Sweet Spot: Glycosylation and other Post-translational Modifications. *ACS Pharmacol.*
1006 *Transl. Sci.* **3**, 237–245 (2020).
1007 71. Eipper, B. A., Stoffers, D. A. & Mains, R. E. The Biosynthesis of Neuropeptides:
1008 Peptide alpha-Amidation. *Annu. Rev. Neurosci.* **15**, 57–85 (1992).
1009 72. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with
1010 AlphaFold 3. *Nature* **630**, 493–500 (2024).
1011 73. Xue, Y. *et al.* State-specific Peptide Design Targeting G Protein-coupled Receptors.
1012 Preprint at https://doi.org/10.1101/2024.11.27.625792 (2024).
1013 74. Pacesa, M. *et al.* BindCraft: one-shot design of functional protein binders. Preprint at
1014 https://doi.org/10.1101/2024.09.30.615802 (2024).
1015 75. Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nat. Methods*
1016 **19**, 679–682 (2022).
1017 76. Wu, F., Jing, X., Luo, X. & Xu, J. Improving protein structure prediction using
1018 templates and sequence embedding. *Bioinformatics* **39**, btac723 (2023).

77. Ovchinnikov, S., Kamisetty, H. & Baker, D. Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. *eLife* **3**, e02030 (2014).

78. Jadon, A. & Patil, A. A Comprehensive Survey of Evaluation Techniques for Recommendation Systems. Preprint at https://doi.org/10.48550/ARXIV.2312.16015 (2023).

79. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

80. Hornak, V. *et al.* Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins Struct. Funct. Bioinforma.* **65**, 712–725 (2006).

81. Brody, S., Alon, U. & Yahav, E. How Attentive are Graph Attention Networks? Preprint at https://doi.org/10.48550/ARXIV.2105.14491 (2021).

82. Agarap, A. F. Deep Learning using Rectified Linear Units (ReLU). Preprint at https://doi.org/10.48550/ARXIV.1803.08375 (2018).

83. Loshchilov, I. & Hutter, F. Decoupled Weight Decay Regularization. Preprint at https://doi.org/10.48550/ARXIV.1711.05101 (2017).

84. Mao, A., Mohri, M. & Zhong, Y. Cross-Entropy Loss Functions: Theoretical Analysis and Applications. Preprint at https://doi.org/10.48550/ARXIV.2304.07288 (2023).