Rapid and accurate deorphanization of ligand-receptor pairs using AlphaFold

Niels Banhos Danneskiold-Samsøe[1,2*], Deniz Kavi[1], Kevin M. Jude[3], Silas Boye Nissen[1], Lianna W. Wat[1,4], Laetitia Coassolo[1,4], Meng Zhao[1,4], Galia Asae Santana-Oikawa[1], Beatrice Blythe Broido[1], K. Christopher Garcia[3], Katrin J. Svensson[1,4,5*]

[1]Department of Pathology, Stanford University School of Medicine, Stanford, CA, USA.

[2]Department of Biology, University of Copenhagen, Denmark

[3]Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, CA, USA

[4]Stanford Diabetes Research Center, Stanford University School of Medicine, Stanford, CA, USA

[5]Stanford Cardiovascular Institute, Stanford University School of Medicine, CA, USA.

*Co-corresponding authors: nbds@stanford.edu and katrinjs@stanford.edu

Abstract

Secreted proteins are extracellular ligands that play key roles in paracrine and endocrine signaling, classically by binding cell surface receptors. Experimental assays to identify new extracellular ligand-receptor interactions are challenging, which has hampered the rate of novel ligand discovery. Here, using AlphaFold-multimer, we developed and applied an approach for extracellular ligand-binding prediction to a structural library of 1,108 single-pass transmembrane receptors. We demonstrate high discriminatory power and a success rate of close to 90 % for known ligand-receptor pairs where no *a priori* structural information is required. Importantly, the prediction was performed on *de novo* ligand-receptor pairs not used for AlphaFold training and validated against experimental structures. These results demonstrate proof-of-concept of a rapid and accurate computational resource to predict high-confidence cell-surface receptors for a diverse set of ligands by structural binding prediction, with potentially wide applicability for the understanding of cell-cell communication.

Many secreted proteins, polypeptides, and peptides constitute signaling molecules that control intercellular communication by binding and activating membrane receptors[1,2]. Upon receptor binding, these molecules directly coordinate short or long-distance signaling responses and biological functions such as cell growth, survival, and metabolism[1,3,4]. The human secretome contains at least 2,000 secreted proteins, not counting posttranslationally processed fragments and peptides[5]. The vast majority of these ligands have no assigned function or cognate receptor. Single-pass transmembrane receptors, also known as bitopic proteins, represent more than 1300 proteins in humans[6] and include receptor tyrosine kinases (RTKs), cytokine receptors, enzymes, and extracellular matrix proteins[4,7–9]. Surprisingly, while single-pass transmembrane receptors constitute up to 50 % of all transmembrane proteins[10], most ligands for these receptors remain unknown. Deorphanization of protein and peptide ligands and their functional receptors can open up entirely new fields in biology and offer new therapeutic avenues[11].

Performing experimental screens to identify ligand-receptor pairs is challenging for several reasons. Mapping interactions at the cell surface is inherently more difficult than identifying intracellular interactions. This is because extracellular ligand-receptor interactions often have low affinity and fast dissociation rates, making high-throughput screening methods such as affinity purification challenging[12,13]. Similarly, binding screens using an individual ligand applied to a receptor in solution are time-consuming, not applicable for all receptor types, and may lack the cellular environment necessary for posttranslational modifications or co-receptor binding[13,14]. Lastly, cell-based CRISPR screens have recently been utilized for the deorphanization of ligands, but are limited by the ability to gain sufficient receptor expression and the lack of expression of essential coreceptors[13,15].

With the revolutionizing ability to predict protein 3D structures from their amino acid sequences, AlphaFold has become an omnipresent tool in the field of structural biology[16,17]. As the 3D structure of a protein is closely related to its function and interactions with other

molecules, AlphaFold has been tremendously useful in predicting intracellular protein-protein interactions, heterodimeric protein complex[18–20], as well as extracellular interactions using structure and topology prediction[21,22]. However, a reliable method to assign the binding of secreted ligands to single-pass transmembrane receptors has not previously been developed.

Here, we demonstrate that protein ligands for single-pass transmembrane receptors can be predicted using AlphaFold. We describe the computational and structural requirements for the prediction screen, performance and success rate, and provide proof-of-principle evidence of identification of high-confidence binders. This work is likely to be relevant to a wide variety of fields and provide a useful resource for future investigations.

Results

Construction of a structural library of 1,108 single-pass transmembrane receptors

To test the ability of AlphaFold to predict cell surface receptors for secreted proteins, we first established a library of single-pass transmembrane proteins using sequences obtained from UniProt (**Fig. 1a**). Single-pass transmembrane receptors span the membrane once and are classified into types I, II, II, or IV, depending on their transmembrane topology (**Fig. 1b**)[6,23]. To limit computation time, we excluded receptors with duplicated gene names, entries without a gene name, entries without an annotated extracellular domain, and receptors with an extracellular domain > 3,000 amino acids. Since AlphaFold was trained on sequences longer than 15 amino acids, we also excluded entries with an extracellular domain < 16 amino acids. This resulted in a library of 1,108 receptors. The majority of entries in the library constitute type 1 (86.4 %) single-pass transmembrane receptors with type II, III, and IV at progressively decreasing fractions of 12.2, 1.4, and 0.1 %, respectively (**Fig. 1c**). To assess the composition of the library we mapped the phylogeny as annotated in the membraneome database[24]. The largest group of proteins in the library are defined as receptors followed by structural/adhesion

proteins and receptor ligands/regulators at 45%, 24%, and 12% (**Fig. 1d and Extended Data Table S1**). We also investigated the expression of the receptors and their top GO terms across the tissues annotated in the Human Protein Atlas[25], finding that 48.5 % of the receptors in the library were tissue-enhanced, that 25 % had low-tissue specificity, and 0.9 % were not detected (**Extended Data Fig. 1a**). The number of receptors expressed, at any level, was high and constant across tissues (**Fig. 1e**, ($p < 0.001$) and cell types (**Extended Data Fig. 1c**, ($p < 0.001$), demonstrating broad applicability of the library. The library is enriched in tissues known to respond to many secreted cues, including the spleen, lymph nodes, intestines, the liver, the kidney and adipose tissue. GO analysis showed that cytokine receptor activity and immune receptor activity were among the top ten enriched terms (**Extended Data Fig. 1d**). To gauge the applicability of the screen, we investigated ligand gene lengths for previously annotated ligand-receptor complexes stratified by receptor type[21,26]. The median ligand gene length was 284 (quartiles: 189-416) amino acids for single-pass receptor ligands compared to 103 (quartiles: 77-152) amino acids for ligands that bind multi-pass receptors ($p = 10^{-14}$) (**Fig. 1f**). These data demonstrate that the receptor type may to some degree be inferred by ligand size.

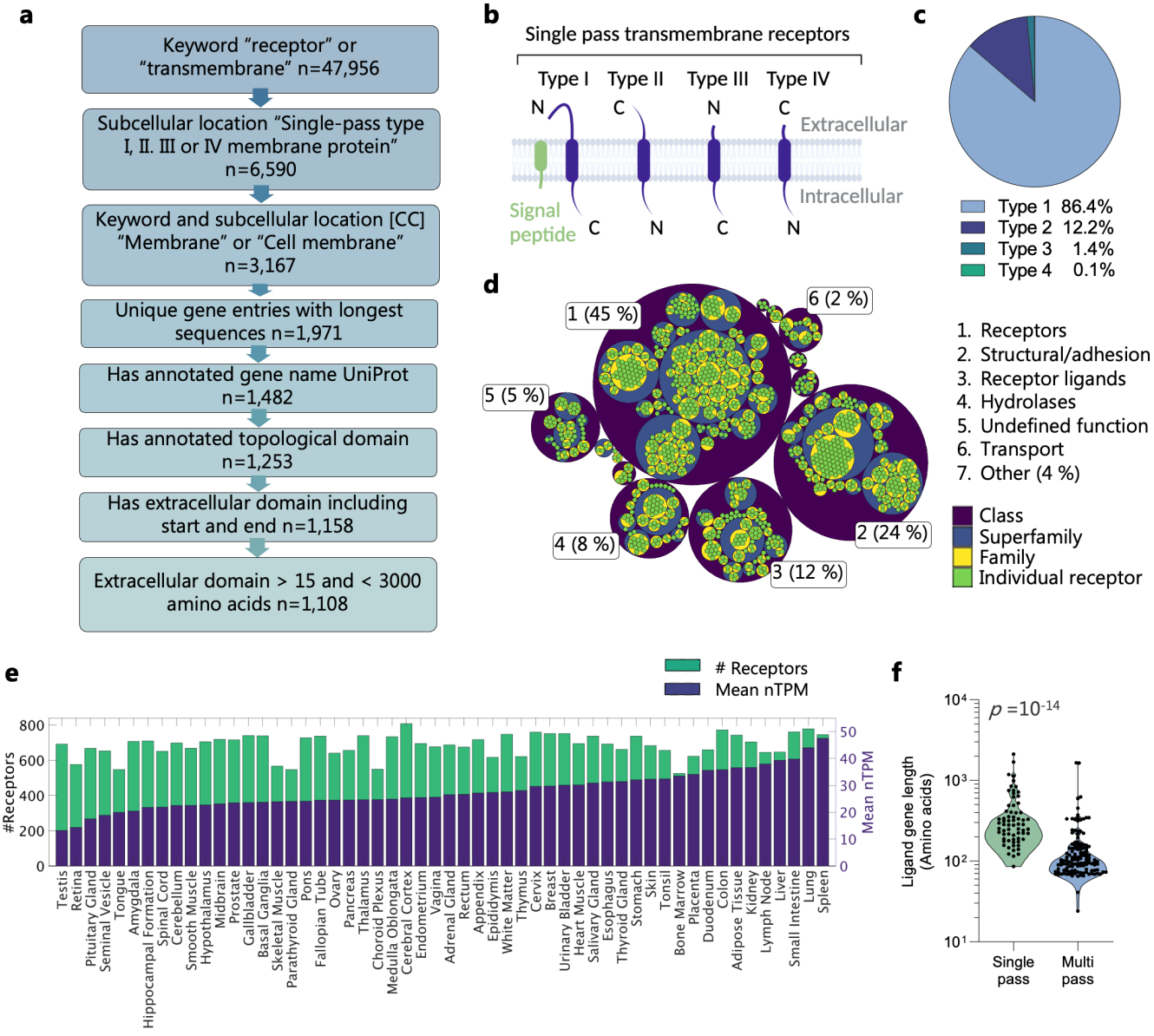


Figure 1. Construction and properties of a structural library of 1,108 single-pass transmembrane receptors. **a**) Schematic of the receptor library construction. 1: Extract human entries with the keyword either "receptor" or "transmembrane" n=47,956. 2: Retain entries with subcellular location [CC] either "Single-pass type I, II, III or IV membrane protein" n=6,590. 3: Retain entries with keyword and subcellular location [CC] either "Membrane" or "Cell membrane" n=3,167. 4: Exclude duplicated gene names, for each gene retaining entries with longest sequences n=1,971. 5: Remove entries without an annotated gene name according to UniProt n=1,482. 6: Retain entries with an annotated topological domain n=1,253. 7: Remove entries without an extracellular domain including start and end n=1,158. 8: Exclude receptors with an extracellular domain shorter than 16 amino acids and longer than 3,000 amino acids, n=1,108. **b**) Schematic diagram of single-pass transmembrane receptor classified by type. **c**) Pie diagram of receptor type distribution in the library. **d**) Phylogeny distribution of receptor library as defined by membraneome.org **e**) Receptor expression (mean nTPM) relative to the number of receptors (# Receptors) across tissues. **f**) Canonical protein sequence length for ligands that bind either multi-pass or single-pass receptors expressed as amino acids (KS test p =10^{-14}), n=173 multi-pass, n=64 single-pass. Significant differences in ligand length for known ligand-receptor pairs were calculated using the Kolmogorov–Smirnov test.

Ligand-receptor binding prediction accuracy is dependent on the sequence input

AlphaFold demonstrates unprecedented prediction of protein-protein interactions, but single-pass transmembrane receptors may produce spurious results with intertwined transmembrane, intracellular, or extracellular domains which might interfere with ligand binding prediction[24]. To establish the parameters for the highest binding prediction strength, we hypothesized that removing the intracellular and transmembrane parts of the receptor would improve the prediction of ligand binding. To predict structures, we used Alphafold2 (AF2) using precomputed (multiple sequence alignments) MSAs. Importantly, to avoid any learning-based bias by AF2, we selected ligand-receptor pairs for genes where crystal structures from the Protein Data Bank (PDB) had not been released at the point of AF2 training. Prediction of ligand-receptor binding associations was performed using either the full-length receptor consisting of the extracellular domain (ECD), the transmembrane domain (TMD), the intracellular domain (ICD), or the ECD alone. For the ligand input, we used either the full-length ligand (secreted protein with or without the pro-region) without the signal peptide, or the processed ligand cleaved from a precursor protein. For qualitative assessment of the ligand-receptor binding prediction, we used the interface template modeling (ipTM) score for modeling protein complexes where a value closer to 1 reflects a likely protein complex with a high probability of correct interface modeling, while values lower than 0.2 indicate two randomly chosen proteins[27,28]. Importantly, the ipTM score is not influenced by the size of the protein. To test the effect of the ligand input sequence, we compounded four test ligands that all had annotated chains according to UniProt, and where the ligand-receptor structure was missing at AF2 training (**Extended Data Table S2**). The test set included the following ligand-receptor pairs: bone morphogenic protein 10 (BMP10) with its receptors bone morphogenetic protein receptor type-1A, B (BMPR1A and B) and activin A receptor like type 1 (ACVRL1), the ligand anti-Mullerian hormone (AMH) and its receptor anti-Mullerian hormone type-2 receptor (AMHR2) [29,30], the receptor tyrosine kinase ligand ALK and LTK ligand 1

(ALKAL1) and its receptors ALK and LTK, and the secreted CD160 antigen (CD160), a cell surface ligand for herpes virus entry mediator (TNFRSF14/HVEM). As expected, the ipTM value reflected the ability of AF2 to detect similar binding residues to the matching crystal structures[29,31] as depicted for BMP10-ACVRL1 **(Fig. 2a-e)** and AMH-AMHR2 **(Extended Data Fig. 2a-e)**. In cases where predicted binding residues were erroneous (**Fig. 2b and 2e**), we expectedly observed lower ipTM values (**Fig. 2f) and Extended Data Fig. 2e**). The highest prediction strength was observed when predicting the full or secreted ligand in combination with only the ECD of the receptor, which led to average ipTM values above 0.7 for AMH-AMHR2, ALKAL1-LTK, and CD160-TNFRSF14 (**Fig. 2f**). For BMP10, the ipTM was over 0.7 for the binding to its receptor bone morphogenetic protein receptor type-1B (BMPR1B), but 0.6 when binding to bone morphogenetic protein receptor type-1A (BMPR1A) using the full ligand and only the ECD (**Fig. 2f**). Similarly, the ipTM value dropped to ~0.6 for the ALKAL1-ALK complex when using the full ligand. Predicting the binding using both the ECD and ICD domains of the receptor with the full ligand consistently led to lower prediction strength as demonstrated by a median ipTM of ~0.3 for AMH-AMHR2, ~0.6 for ALKAL1-LTK, ~0.2 for BMP10-BMPR1A, and ~0.3 for BMP10-BMPR1B (**Fig. 2f**). In contrast, for the BMP10-ACVRL1 complex, the median ipTM value was higher (~0.6) using the full ligand compared with the secreted ligand (~0.2). In conclusion, predicting the ligand-receptor structure using either the secreted ligand or full ligand in combination with the ECD of the receptor, led to excellent binding prediction, while including the ICD worsened prediction strength.

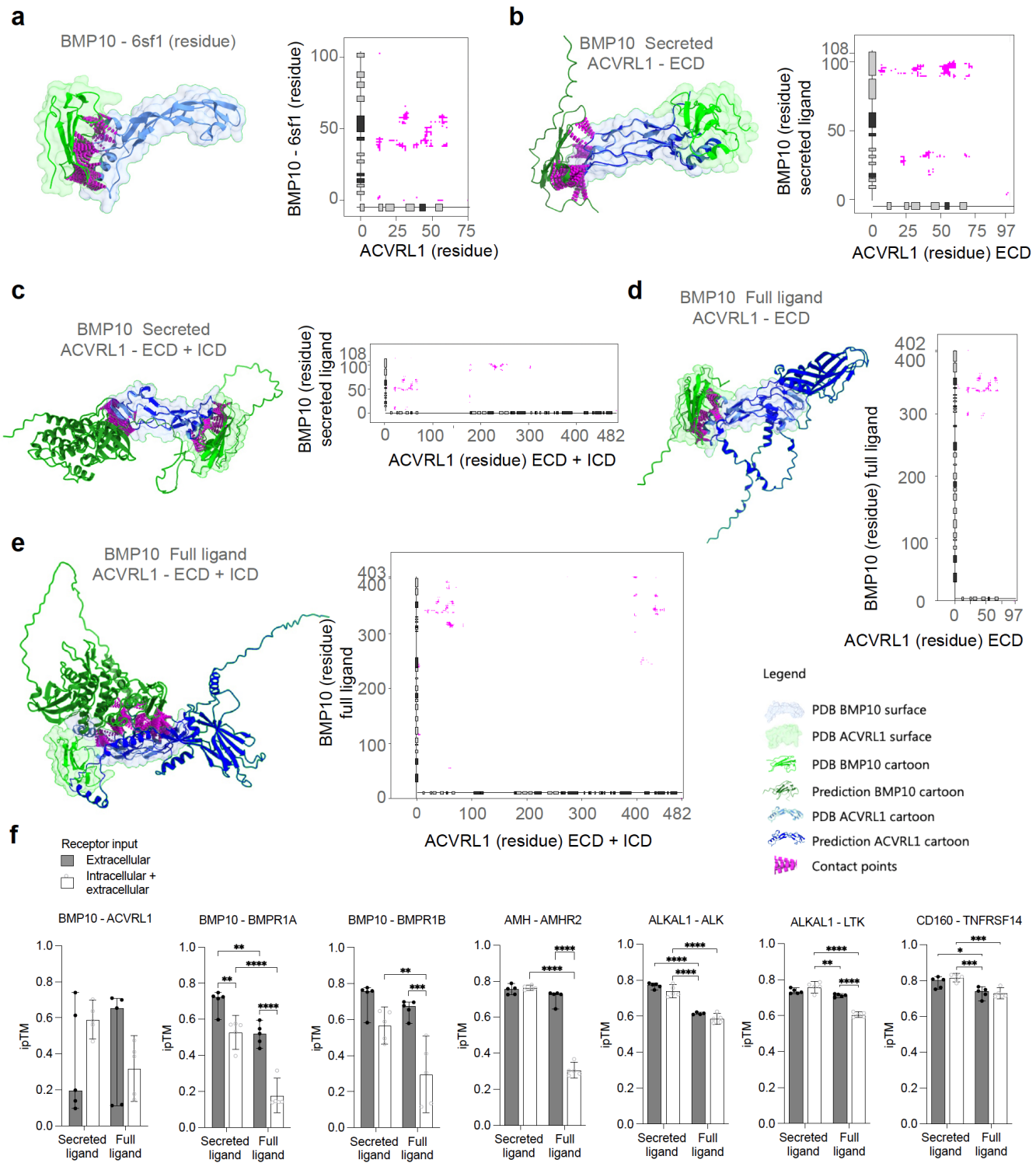


Figure 2. Ligand-receptor binding prediction accuracy is dependent on the sequence input. a) pdb structure (PDB id: 6sf1) and contact map of the BMP10-ACVRL1 complex complementary to representative predictions in b-e. **b-e**) Structural binding prediction and corresponding contact maps (Distances below <8 Å were considered contacts) of ligand-receptor pairs comparing full or truncated chains of ligand and receptor. Annotation for a-e: light green cartoon and surface: PDB database receptor, dark green cartoon: AF2 predicted receptor, light blue cartoon and surface: PDB database ligand, dark blue cartoon: AF2 predicted ligand, magenta: contact points. **f**) ipTM of the ligand-receptor pairs BMP10-ACVRL1, BMP10-BMPR1A, BMP10-BMPR1B, AMH-AMHR2, ALKAL1-ALK, ALKAL1-LTK, and CD160-TNFRSF14 predicted by AF2 using either of the following annotated regions from UniProt: secreted ligand/chain, full ligand (pro-region and secreted ligand/chain without signal-peptide),

extracellular (extracellular without signal peptide), intracellular + extracellular (full canonical sequence without signal peptide). The predictions are the median± 95% CI of five independent predictions for each ligand-receptor pair (n=5). Two-way ANOVA followed by Turkey's test was used for multiple comparisons of differences in ipTM values between different input conditions for ligand-receptor pairs in GraphPad Prism version 9.5.0, *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001.

Validation of accurate prediction of single-pass transmembrane receptors for known ligands

Having established the parameters for binding prediction strength, we next tested if AF2 could predict the correct receptors for known ligands using the single-pass transmembrane library. Known receptor-ligand pairs where structures for both proteins were absent at the time of AF2 training and where the ligand was annotated as secreted by UniProt were included. To increase the number of test cases we also included a manually curated list of ligand-receptor pairs where the receptor, but not the ligand, had a structure available at the AF2 training cutoff[26]. Searching the PDB for these conditions resulted in eight ligand-receptor pairs (**Extended Data Table S2**). Next, we predicted protein-protein interactions between one ligand against all the receptors in the library. For many orphan ligands, information on pro-peptide presence and location is unknown. Because of this, and since our binding analyzes showed that the use of the full ligand and the ECD of the receptor generally resulted in ipTM values >0.6, we proceeded to use this combination for the ranking prediction (**Fig. 3a)**. To rank predictions we used the ipTM value of five predictions from AF2 and penalized receptors with high variation as previously described[22]. Remarkably, with one exception, the correct receptors for the eight test cases consistently ranked among the top three receptors with ipTM values between 0.6-0.8 for all ligand-receptor pairs. We accurately predicted the type II receptor AMHR for the ligand AMH as the top predicted receptor (**Fig. 3b)**. For the ligand BMP10, its receptor ACVRL1 is ranked number two while BMPR1B and BMPR1A ranked first and sixths, respectively (**Fig. 3c)**. Furthermore, ALKAL1, known to bind the tyrosine kinase receptor ALK[32], was predicted as number one in the screen, while the other known receptor, LTK, ranked third (**Fig 3d**). Importantly, other RTKs displayed

low ipTMs. Given that monomeric ALKAL1 is known to lead to homodimerization upon ALK binding[33], these results suggest that binding prediction is independent of conformation changes. Furthermore, the prediction correctly identified the cytokine receptors IL17RA and IL17RB for interleukin-25 (IL25)[34](**Fig 3e**), CD160 for TNFRSF14/CD270[35] (**Fig. 3f**), the secreted metalloproteinase fetuin-b (FETUB) for meprin A (MEP1A)[36] (**Fig. 3g),** IL27 for IL27RA (**Fig 3h**). For one ligand, neural EGFL like 2 (NELL2), AF2 did not predict any binding (ipTM ~ 0.11) to the proposed receptor ROBO3[37] (**Fig. 3i and Extended Data Fig. 3a**). True positive and negative ligand-receptor interactions were distinctly classified by the ipTM score (**Fig 3j**), yielding a ROC curve with an excellent area under the curve (AUC) value of 0.947 (**Fig 3k**). Given that the basis for the ipTM values is built on the structural binding prediction of the ligand-receptor pairs, the predicted ligand-receptor interactions by AF2 demonstrated excellent overlap with the known PDB structures for IL27-IL27RA, ALKAL1-LTK, IL25-IL17RA/B, CD160-TNFRSF14, and FETUB-MEP1A complexes (**Fig. 3l-p).** The ranking of receptors was not significantly different when using the average ipTM, median ipTM, penalized ipTM, or pDockQ[22,38], demonstrating robust prediction with low variability (**Extended Data Fig. 3b**). In conclusion, we show that we can rapidly and reliably predict the receptors for a broad range of ligands with a success rate of 87.5 % for the eight ligands tested.

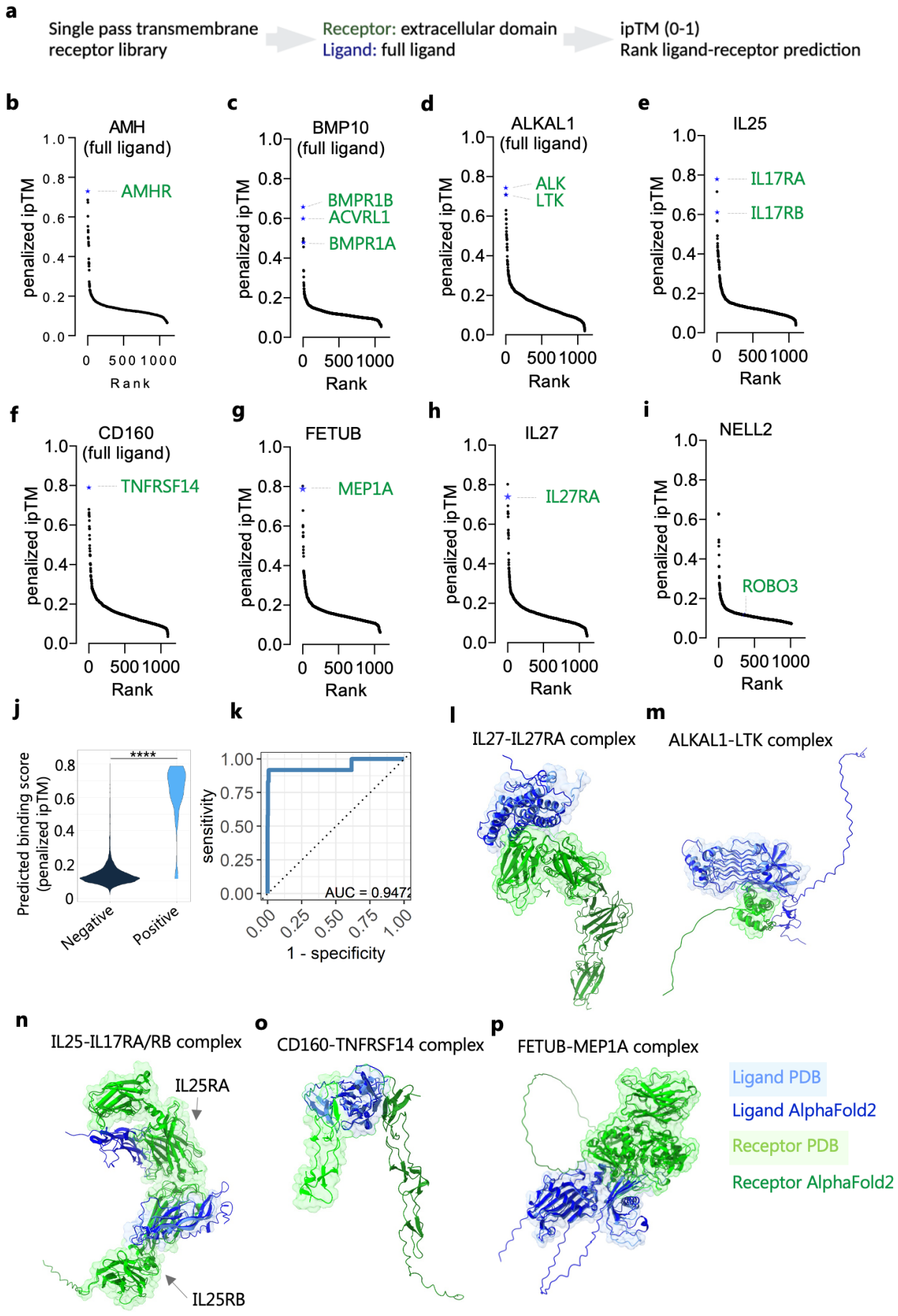


Figure 3. Accurate prediction of single-pass receptors for known secreted ligands. **a)** Approach and metric for scoring of binding of ligands against the single-pass transmembrane receptor library. **b-i)** Binding prediction of anti-Mullerian hormone (AMH) (**b**), bone morphogenic protein 10 (BMP10) (**c**), ALK and LTK ligand 1 (ALKAL1) (**d**), interleukin-25 (IL25) (**e**), CD160 (**f**), Fetuin-B (FETUB) (**g**), IL27 (**h**), and neural EGFL like 2 (NELL2) (**i**) to the receptor library. Values are expressed as ranked penalized ipTM. The predictions are the median minus median absolute deviation of five independent predictions for each ligand-receptor pair. **j**) Performance of ipTM in predicting true positives. IpTM values include all prediction across the test set. Positives were defined as all validated ligand-receptor pairs in b-i, **k**) Performance of ipTM for sensitivity and specificity. ROC curve including area under the curve (AUC) = 0.9742. **l-p**) Representative structural binding prediction of ligand-receptor pairs comparing the PDB structures with AF2 (light blue: ligand in PDB, light green: receptor in PDB, dark blue: predicted ligand, dark green: predicted receptor) for IL27-IL27Ra (**l**), ALKAL1-LTK (**m**), IL25-IL17RA/B (**n**), CD160-TNFRSF14 (**o**), and FETUB-MEP1A (**p**). PDB structures used: IL17-IL27Ra (7u7n), ALKAL1-LTK (7nx0), IL25-IL17RA/B (7uwj), CD160-TNFRSF14 (7msg), FETUB-MEP1A (7uai). Wilcoxon signed-rank test was used to compare differences in ipTM between non-binders (negative) and validated binders (positive) in R version 4.2.1. *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001.

Reverse screen for receptors against a ligand library

Since all correct ligand-receptor pairs had ipTM values above 0.47 and, due to thinning out in hits above this value, we hypothesized that the screen using AF2 could also be done in reverse to identify ligands for specific receptors. We constructed a ligand library and predicted ligands for receptors that were not in the PDB database at the cut-off date for AlphaFold training. The ligand library was generated by including entries for genes annotated in UniProt predicted to be secreted, and have a sequence length between 15-2000 amino acids, excluding gene names containing IGH, IGKC, IGKV, IGLC, or IGLV. The ligand library comprises 1,862 unique entries (**Fig. 4a-b**). This prediction accurately identified AMH as the ligand for AMHR as the second-ranked hit (**Fig. 4c**), IL27 as the first ligand for IL27RA (**Fig. 4d**), ALKAL1 and ALKAL2 as the top two ligands for LTK (**Fig. 4e**), and FETUB as the top ligand for MEP1A (**Fig. 4f**). In conclusion, we show that we can also reliably predict the ligands for receptors with a success rate of 100 % for the four ligands tested.

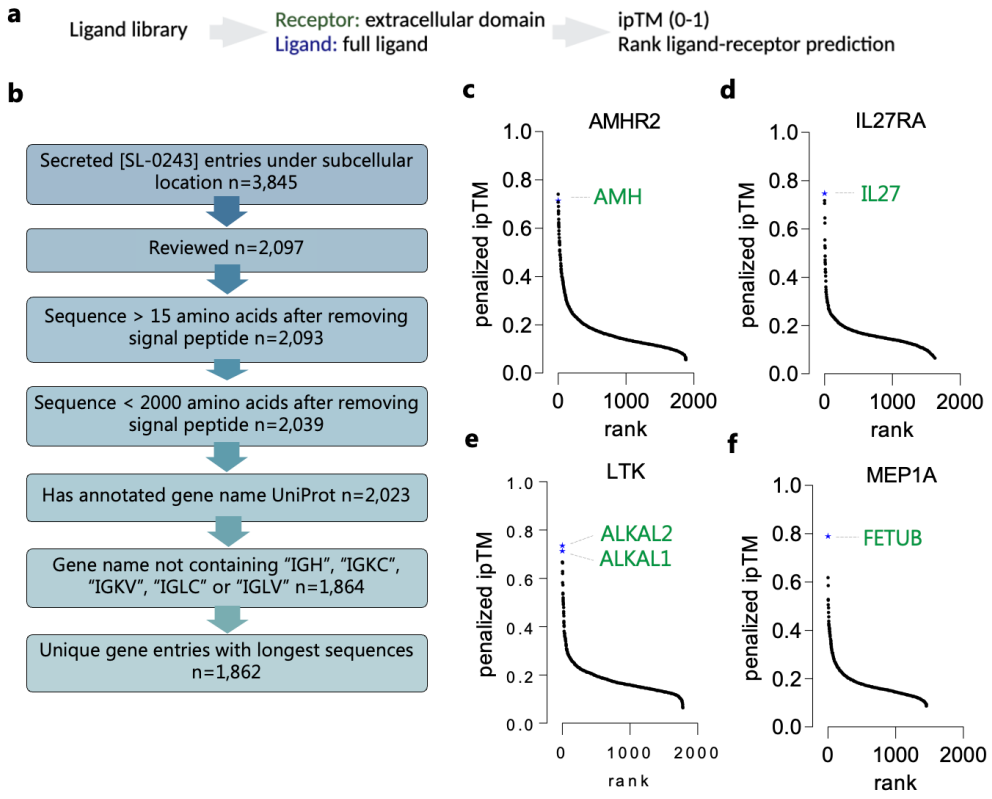


Figure 4. AF2 can be used to screen receptors against a ligand library. a) Approach and metric for scoring of binding of single-pass transmembrane receptors against a ligand library, **b)** Schematic of ligand library construction. 1: Extract human entries annotated as secreted [SL-0243] under subcellular location n=3,845. 2: Retain reviewed entries n=2,097. 3: Exclude secreted peptides/proteins with an extracellular domain shorter than 16 amino acids n=2,093 and 4: longer than 2,000 amino acids n=2,039. 5: Keep proteins with annotated gene names n=2,023. 6: Remove immunoglobulins with gene names either including "IGH, "IGKC", "IGKV", "IGLC" or "IGLV" n=1,864. 6: exclude duplicated gene names retaining entry with the longest sequence, n=1,862. Binding prediction of **c)** AMHR2, **d)** IL27RA, **e)** LTK and **f)** MEP1A to the ligand library. Values are expressed as ranked penalized ipTM. The predictions are the median minus median absolute deviation of five independent predictions for each ligand-receptor.

Prediction of single-pass transmembrane receptors for orphan secreted ligands

To test if we could predict receptors for orphan secreted ligands using the single-pass transmembrane library, we tested the screen using 15 ligands based on a previously curated library of potentially high-value orphan secreted proteins[15] (**Extended Data Table S3**). Due to limited graphics processing unit (GPU) resources, we limited our test to binding partners that could be predicted using central processing units (CPUs). We qualitatively scored binding predictions with a penalized ipTM value > 0.6 taking into consideration tissue expression, known activities, and known or predicted structure (**Fig. 5a-o and Extended Data Table S3**). We

predict that the ligand IL-40 (C17orf99), likely binds Interleukin-13 receptor subunit alpha-1 (IL13RA1) (penalized ipTM=0.65), the prolactin receptor (PRLR) (penalized ipTM=0.68), and CMRF35-like molecule 1 (CD300LF) and) (**Fig. 5a**). We predict the connective tissue mitoattractant CCN family member 2 (CCN2) to bind Killer cell lectin-like receptor subfamily B member 1 (KLRB1) (penalized ipTM=0.63) (**Fig. 5b**). For the cerebral dopamine neurotrophic factor (CDNF), we find that it likely binds the activin receptor type-2B (ACVR2B) (penalized ipTM=0.74) and activin receptor type-2A (ACVR2A) (penalized ipTM=0.63) (**Fig. 5c**). Moreover, the extracellular matrix protein 2 (ECM2) likely binds to Myelin-oligodendrocyte glycoprotein (MOG) (penalized ipTM=0.69) and Disintegrin and metalloproteinase domain-containing protein 23 (ADAM23) (penalized ipTM=0.68) (**Fig. 5d**), while lymphocyte antigen 6H (LY6H) is predicted to bind to Semaphorin-4B (SEMA4B), B-cell antigen receptor complex-associated protein alpha chain (CD79A), Leucine-rich repeat-containing protein 15 (LRRC15) and HLA class II histocompatibility antigen, DR beta 3 chain (HLA-DRB3) (penalized ipTM=0.71-0.60) (**Fig. 5e**). We predict that the secreted ligand Meteorin (METRN) binds to neurogenic locus notch homolog protein 2 (NOTCH2), lysosome-associated membrane glycoprotein 5 (LAMP5), and neurogenic locus notch homolog protein 1 (NOTCH1)(penalized ipTM=0.64-0.61) (**Fig. 5f**). For midkine (MK), we find five potential binding partners including Protocadherin alpha-C1 (PCDC1), Paired immunoglobulin-like type 2 receptor beta (PILRB), Immunoglobulin superfamily member 6 (IGSF6), Killer cell lectin-like receptor subfamily G member 1 (KLRG1) and Basal cell adhesion molecule (BCAM)(penalized ipTM=0.69-0.61) (**Fig. 5g**). For leucine-rich glioma-inactivated protein 1 (LGI1), we identified an already experimentally validated receptor in the disintegrin and metalloproteinase domain-containing protein 22 (ADAM22) (penalized ipTM=0.72) where the crystal structure was released after the AF2 cut-off date[39], supporting our prediction analysis. We further find likely that ADAM11 and ADAM23 are also likely receptors for LGI1 (**Fig. 5h**).

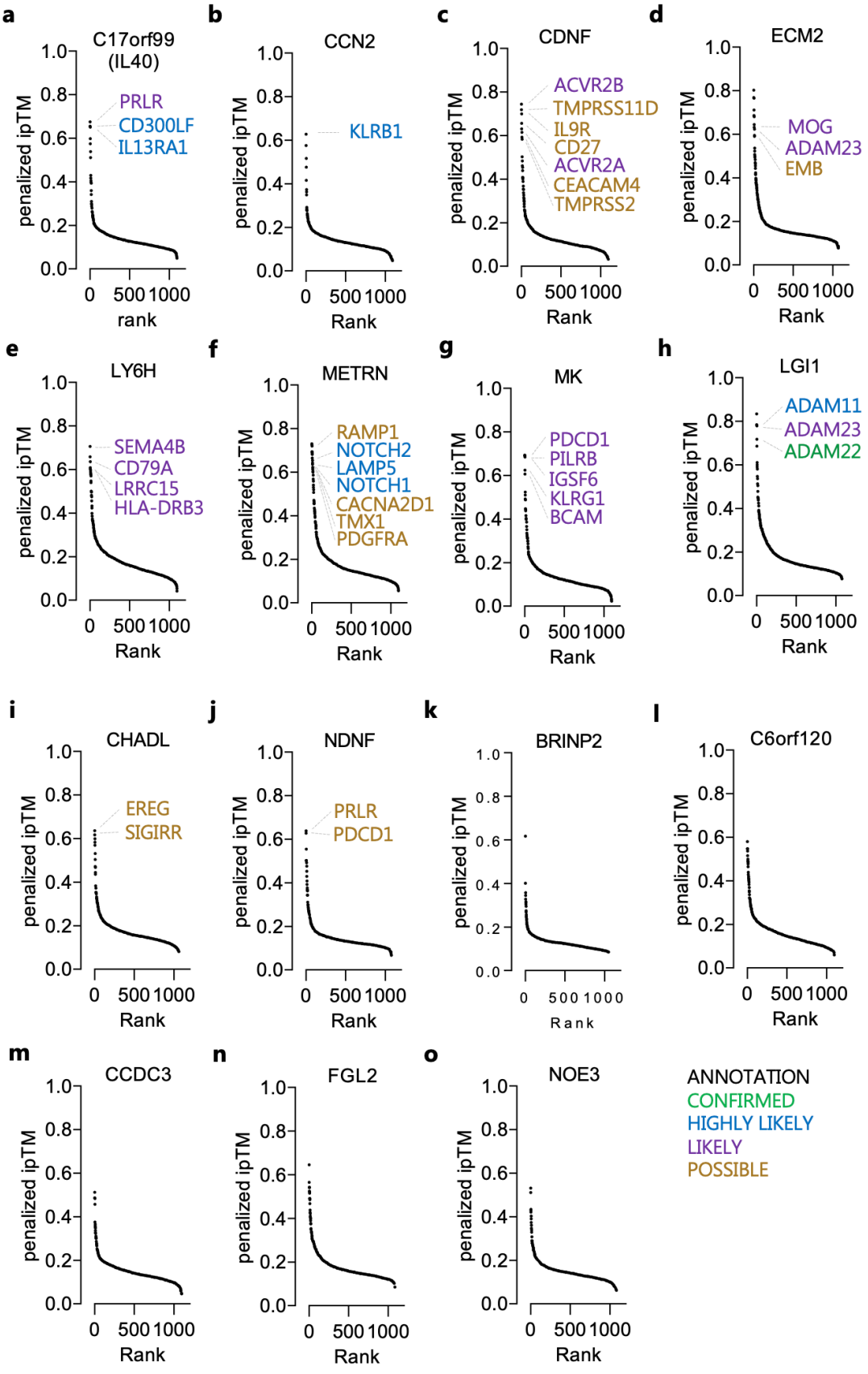


Figure 5. Identification of novel ligand-receptor binding pairs to orphan ligands. a-o) Ranked predicted ligand-receptor binding partners by penalized ipTM value for a) C17orf99, b) CCN2, c) CDNF, d) ECM2, e) LY6H, f) METRN, g) MK, h) LGI1, i) CHADL, j) NDNF, k) BRINP2, l) C6orf20, m) CCDC3, n) FGL2, o) NOE3.

For seven out of fifteen ligands, we find possible binding partners, unlikely receptors, or pairs with a penalized ipTM value < 0.6 (**Fig. 5i-o**). Specifically, for chondroadherin-like protein (CHADL) and protein NDNF (NDNF) we found no likely binding partners (**Fig. 5i-j**). For BMP/retinoic acid-inducible neural-specific protein 2 (BRINP2) and Fibroleukin (FGL2) we find no likely receptors (**Fig. 5k-n**). For UPF0669 protein C6orf120 (C6orf120), Coiled-coil domain-containing protein 3 (CCDC3) and Noelin-3 (NOE3), we find no predicted binding partners with a penalized ipTM > 0.6 (**Fig. 5l-o**). In summary, this method provides a resource for identifying high-confidence orphan ligand-receptor pairs.

Discussion

New therapeutic targets for disease are likely to target receptors or their secreted ligands[40,41]. Yet, for many hundreds of ligands identified through the secreted protein discovery initiative[41] and in the human protein atlas secretome[5], the receptors remain uncharacterized. In this paper, we demonstrate a simple and highly accurate algorithm to predict single-pass receptors for orphan ligands using AlphaFold. We show that the prediction of the receptor can be obtained from full-length canonical ligand sequences without prior knowledge of structural binding. This work presents a major advance for ligand discovery where no à priori knowledge of binding sites is needed and is broadly applicable to a diverse set of secreted ligands including cytokines, hormones, receptor tyrosine kinase ligands, and proteases. The most striking finding was the accuracy in the identification of seven receptor-ligand pairs where the true receptor was identified within the top three receptors in a library with over 1,100 receptors. Based on the data presented in this paper, this resource could potentially be expanded to include other classes of cell-surface receptors, including ion-gated channels, multi-pass receptors, and G-protein coupled receptors (GPCRs)[21,22].

There are some limitations of our method, including that the prediction accuracy is influenced by the quality and completeness of the input data. Approximately 100 single-pass receptors are missing in the library due to a lack of annotated topological domains, as well as proteins without annotated start and end domains in UniProt. These receptors could be included by inferring topological domains using computational prediction[42]. To reduce computation time, we limited the library to single-pass transmembrane receptors. Hence, glycosylphosphatidylinositol (GPI)-anchored proteins are not included because they lack a transmembrane domain. We also selected the canonical isoform of the receptors, which were not always the longest sequences. Given that the full-length ligand performed well, it is possible that the longest splice variant would perform better. Another consideration is that ligands that bind transmembrane proteins can be monomeric, dimeric, or trimeric, or require co-factors for binding[43] which is not accounted for in our binding prediction. While many receptors will have specificity towards one ligand, receptors in the RTK family typically show less specificity, with the ability to bind several ligands with varying affinity. Additionally, this approach may not be applicable to more complex receptors or ligands that interact with multiple receptors, and further studies are needed to determine the generalizability of the approach to different types of ligands and receptors. In one case, we were unable to predict the binding of NELL2 to its receptor ROBO3. The crystal structure for NELL2-ROBO3 includes a truncated part of the ECD of the receptor which might explain the lack of binding prediction for this ligand-receptor pair[37]. Finally, while the approach is effective in predicting ligand-receptor binding, it does not provide information on downstream signaling events, which are also critical for understanding the functional consequences of ligand-receptor interactions.

In conclusion, the potential implications of this research are vast[44], as it has the capacity to serve as a valuable tool for identifying previously unknown ligand-receptor pairs across a diverse range of proteins, thus opening up new possibilities for drug discovery and development.

Acknowledgements

K.J.S. was supported by NIH grants DK125260, the Stanford Diabetes Research Center P30DK116074, the Weintz Family COVID-19 research fund, the Stanford School of Medicine, and the Stanford Cardiovascular Institute (CVI). N.B.D.S. was supported by the Carlsberg Foundation Internationalization Fellowship and the Øllingesøe Foundation. M.Z. was supported by the American Heart Association (AHA) postdoctoral fellowship (905674). L.C. was supported by the Stanford School of Medicine Dean's Postdoctoral Fellowship and the American Heart Association (AHA) postdoctoral fellowship (1011077). L.W.W. was supported by the Stanford School of Medicine Dean's Postdoctoral Fellowship.

Author contributions

Conceptualization: N.B.D.S., K.J.S.; methodology: N.B.D.S., investigation: N.B.D.S., D.K., B.B.B., G.S.O., L.C., M.Z., L.W.W., S.B.N., K.M.J., K.C.G.; supervision and funding acquisition: K.J.S.; Writing – original and revised draft: N.B.D.S., S.B.N., K.J.S.

Competing Interests

The authors do not declare any conflicts of interest.

Data and materials availability

All data generated or analyzed during this study are included in the manuscript, in supporting files and at https://github.com/Svensson-Lab/run-hpc-alphafold. Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contacts, Niels Banhos Danneskiold-Samsøe (nbds@stanford.edu) and Katrin J. Svensson (katrinjs@stanford.edu).

Online Methods

Construction of a single-pass receptor library

To construct a library of single-pass receptors we searched UniProt human entries for keywords with the terms “receptor” or “transmembrane” available by 11-11-2022 (n=47,956). From this pool, we retained entries stating either “Single-pass type I, II, III or IV membrane protein” under subcellular location [CC], n=6,590. As we restricted our library to secreted proteins, we only retained entries with keyword and subcellular location [CC] either “Membrane” or “Cell membrane”, n=3,167. Next, in the case of duplicated gene names, we prioritized reviewed entries. In case all entries with a duplicated gene name were unreviewed we prioritized the entry with the longest sequence, n=1,971. Shorter sequences were generally truncated versions of the longest FASTA sequence. Due to limited computational resources, we restricted the library to the canonical gene sequence by UniProt avoiding other isoforms. To further limit computational requirements, we removed entries without an annotated gene name, without an annotated topological domain, and without an extracellular domain including start and end according to UniProt retaining 1,158 receptors. Finally, as AlphaFold was trained on sequences longer than 15 amino acids we filtered receptors with extracellular domains equal to or shorter than this. To limit computation we also excluded entries with extracellular domains longer than 3000 amino acids retaining 1,108 receptors in the final library (**Extended Data Table S1)**

Construction of a ligand library

To construct a library of secreted proteins we collected all human entries listed as Secreted [SL-0243] under subcellular location [CC] in UniProt by 01-15-2023, n=3,845. From this pool we kept reviewed entries (n=2,097), and entries longer than 16 amino acids (n=2,093). To limit

computation we also excluded sequences longer than 2,000 amino acids retaining 2,039 entries. We kept only entries with an annotated gene name, n=2,023. We further excluded immunoglobulins by excluding any gene with the name containing either “IGH”, “IGKC”, “IGKV”, “IGLC” or “IGLV” retaining 1,864 entries. In the case of duplicated gene names, we retained only the entry with the longest amino acid sequence retaining 1,862 secreted proteins in the final library.

Predicting structures

We predicted ligand-receptor structures for each ligand against all receptors in the final libraries. Parafold[45] in combination with AlphaFold 2.2.4 without the relaxation step, without template and using the reduced database to generate multiple sequence alignments (MSAs) for both ligands and receptors. To predict structures, we used Alphafold 2.2.4 using precomputed MSAs and the same settings as above predicting five models per ligand-receptor pair. Due to limited GPU access, we first ran predictions using only CPUs restricting it to a maximum of 10.5 hours, 12 CPUs and 64GB of memory on the Danish National Supercomputer Computerome or Sherlock at Stanford University. Results were visualized with ChimeraX version 1.5[46]. A list of PDB IDs for all proteins in the test set and UniProt IDs for all tested ligands is provided in **Extended Data Table S4**. PDB files for predictions with a penalized ipTM value > 0.4 is available at https://github.com/Svensson-Lab/run-hpc-alphafold.

Score prediction

The ipTM scores were extracted from the Alphafold pickle files. Penalized ipTM was calculated by taking the median of available predictions and subtracting the median absolute deviation (MAD) as previously described[22]. The pDockQ score was calculated as previously described[18].

Generation of contact maps

Contact maps were generated using the bio3d package[47] in R version 4.2.1 using either structure from the PDB database or predictions of ligand-receptor pairs as inputs. Distances below 8 Å were considered contacts.

Ligand and receptor characteristics

To determine the amino acid length of ligands that bind to single-pass or multi-pass receptors, we extracted accession numbers for all peptide receptor-ligand pairs in the two databases CellPhoneDB[26] (111 pairs) and GPCRs[21] (86 pairs). We used UniProt[48] to determine whether receptors were single-pass or a multi-pass membrane protein and they were annotated as secreted under subcellular location. We excluded any pairs where both or none are secreted (211 excluded). In addition, we extracted the ligand's gene length, signal peptide length, and chain length. We excluded all receptor-ligand pairs with missing information in UniProt (194 excluded) and receptor-ligand pairs not having exactly one chain each (150 excluded). The final list includes the amino acid lengths of 130 multi-pass membrane proteins and 67 single-pass membrane proteins. The MATLAB script used to obtain and filter data is deposited at https://github.com/Svensson-Lab/danneskiold-samsoe2023.

Expression of receptors across human tissues

To determine the RNA expression of single-pass transmembrane receptors across human tissues, we wrote a script to automatically extract RNA expression of the 1,122 entries in the single-pass receptor library in all 54 tissues and all 79 single cell types found in the Human Protein Atlas (version 22.0)[49]. Here we introduced a lower threshold of 1 normalized transcript expression value (nTPM), defining that any receptor expressed below this threshold is not represented in the cell type/tissue. 26 receptors of the 1,122 entries were not found in the Human

Protein Atlas and were therefore not included for further analysis. This MATLAB script has been deposited at https://github.com/Svensson-Lab/danneskiold-samsoe2023.

Computational cost

Computational requirements were the same as for AF2. To reduce computational costs, we started by calculating MSAs for all receptors using up to 15 hours, 16 CPU cores and 8Gb RAM (**Extended Data Fig. 4a**). Since this step only has to be performed once, the calculation of MSAs significantly reduces computation time. We observed that no ligand-receptor pairs with a penalized ipTM value < 0.1 after the first prediction and < 0.2 for the second prediction obtained a final penalized ipTM > 0.5 (**Extended Data Fig. 4b)**. For the prediction of receptors for orphan ligand we therefore adapted the AF2 to exit after the first predictions in cases where the ipTM value was below these values. As most of the predicted ligand-receptor structures have an ipTM value < 0.2 this also significantly reduces computational cost. All calculations of MSAs and predictions were run using 12 CPU cores and 8Gb RAM. Due to limited GPU access, for the eight test cases in the receptor screen and four test cases in the ligand screen, we first ran 10.5 hours of predictions on CPUs (**Extended Data Fig. 4c)**. For ligand-receptor pairs that did not finish five predictions using CPU, we instead used GPU (**Extended Data Fig. 4d)**.

Code availability

All codes to run the screen can be obtained at https://github.com/Svensson-Lab/run-hpc-alphafold under the Apache License, Version 2.0.

Contact for Reagents and Resource Sharing

Information and requests for resources should be directed to and will be fulfilled by the Lead Contacts, Niels Banhos-Danneskiold-Samsøe (nbds@stanford.edu) and Katrin J. Svensson (katrinjs@stanford.edu).

Statistical analyses

Differences in ligand length for known ligand-receptor pairs were calculated using the Kolmogorov–Smirnov test in MATLAB. We used two-way ANOVA followed by Turkey's test for multiple comparisons of differences in ipTM values between different input conditions for ligand-receptor pairs in GraphPad Prism version 9.5.0, *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001.

References

1. Lefkowitz, R. J. G proteins in medicine. *N. Engl. J. Med.* **332**, 186–187 (1995).
2. McKay, M. M. & Morrison, D. K. Integrating signals from RTKs to ERK/MAPK. *Oncogene* **26**, 3113–3121 (2007).
3. Komolov, K. E. & Benovic, J. L. G protein-coupled receptor kinases: Past, present and future. *Cell. Signal.* **41**, 17–24 (2018).
4. Zhao, M., Jung, Y., Jiang, Z. & Svensson, K. J. Regulation of Energy Metabolism by Receptor Tyrosine Kinase Ligands. *Front. Physiol.* **11**, 354 (2020).
5. Uhlén, M. *et al.* The human secretome. *Sci. Signal.* **12**, (2019).

6. Bugge, K., Lindorff-Larsen, K. & Kragelund, B. B. Understanding single-pass transmembrane receptor signaling from a structural viewpoint-what are we missing? *FEBS J.* **283**, 4424–4451 (2016).
7. Li, E. & Hristova, K. Role of receptor tyrosine kinase transmembrane domains in cell signaling and human pathologies. *Biochemistry* **45**, 6241–6251 (2006).
8. Ramasarma, T. & Joshi, N. V. Transmembrane Domains. in *eLS* (2005). doi:https://doi.org/10.1038/npg.els.0005051.
9. Foxwell, B. M., Barrett, K. & Feldmann, M. Cytokine receptors: structure and signal transduction. *Clin. Exp. Immunol.* **90**, 161–169 (1992).
10. Zviling, M., Kochva, U. & Arkin, I. T. How important are transmembrane helices of bitopic membrane proteins? *Biochim. Biophys. Acta BBA - Biomembr.* **1768**, 387–392 (2007).
11. Ozawa, A., Lindberg, I., Roth, B. & Kroeze, W. K. Deorphanization of novel peptides and their receptors. *AAPS J.* **12**, 378–384 (2010).
12. Honig, B. & Shapiro, L. Adhesion Protein Structure, Molecular Affinities, and Principles of Cell-Cell Recognition. *Cell* **181**, 520–535 (2020).
13. Bushell, K. M., Söllner, C., Schuster-Boeckler, B., Bateman, A. & Wright, G. J. Large-scale screening for novel low-affinity extracellular protein interactions. *Genome Res.* **18**, 622–630 (2008).
14. Taouji, S., Dahan, S., Bossé, R. & Chevet, E. Current Screens Based on the AlphaScreen Technology for Deciphering Cell Signalling Pathways. *Curr. Genomics* **10**, 93–101 (2009).
15. Siepe, D. H. *et al.* Identification of orphan ligand-receptor relationships using a cell-based CRISPRa enrichment screening platform. *eLife* **11**, (2022).
16. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* (2021) doi:10.1038/s41586-021-03819-2.

17. Chang, L. & Perez, A. Ranking Peptide Binders by Affinity with AlphaFold. *Angew. Chem. Int. Ed Engl.* (2022) doi:10.1002/anie.202213362.

18. Bryant, P., Pozzati, G. & Elofsson, A. Improved prediction of protein-protein interactions using AlphaFold2. *Nat. Commun.* **13**, 1265 (2022).

19. Bryant, P. *et al.* Predicting the structure of large protein complexes using AlphaFold and Monte Carlo tree search. *Nat. Commun.* **13**, 6028 (2022).

20. Akdel, M. *et al.* A structural biology community assessment of AlphaFold2 applications. *Nat. Struct. Mol. Biol.* **29**, 1056–1067 (2022).

21. Foster, S. R. *et al.* Discovery of Human Signaling Systems: Pairing Peptides to G Protein-Coupled Receptors. *Cell* **179**, 895-908.e21 (2019).

22. Teufel, F. *et al.* Identifying endogenous peptide receptors by combining structure and transmembrane topology prediction. *bioRxiv* 2022.10.28.514036 (2022) doi:10.1101/2022.10.28.514036.

23. Westerfield, J. M. & Barrera, F. N. Membrane receptor activation mechanisms and transmembrane peptide tools to elucidate them. *J. Biol. Chem.* **295**, 1792–1814 (2020).

24. Lomize, A. L. *et al.* Membranome 3.0: Database of single-pass membrane proteins with AlphaFold models. *Protein Sci. Publ. Protein Soc.* **31**, e4318 (2022).

25. Petryszak, R. *et al.* Expression Atlas update - An integrated database of gene and protein expression in humans, animals and plants. *Nucleic Acids Res.* **44**, D746–D752 (2016).

26. Efremova, M., Vento-Tormo, M., Teichmann, S. A. & Vento-Tormo, R. CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. *Nat. Protoc.* **15**, 1484–1506 (2020).

27. Zhang, Y. & Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins* **57**, 702–710 (2004).

28. Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309 (2005).

29. Hart, K. N. *et al.* Structure of AMH bound to AMHR2 provides insight into a unique signaling pair in the TGF-β family. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).

30. Hinck, A. P., Mueller, T. D. & Springer, T. A. Structural Biology and Evolution of the TGF-β Family. *Cold Spring Harb. Perspect. Biol.* **8**, (2016).

31. Salmon, R. M. *et al.* Molecular basis of ALK1-mediated signalling by BMP9/BMP10 and their prodomain-bound forms. *Nat. Commun.* **11**, 1621 (2020).

32. Li, T. *et al.* Structural basis for ligand reception by anaplastic lymphoma kinase. *Nature* **600**, 148–152 (2021).

33. De Munck, S. *et al.* Structural basis of cytokine-mediated activation of ALK family receptors. *Nature* **600**, 143–147 (2021).

34. Wilson, S. C. *et al.* Organizing structural principles of the IL-17 ligand-receptor axis. *Nature* **609**, 622–629 (2022).

35. Rodriguez-Barbosa, J. I. *et al.* HVEM, a cosignaling molecular switch, and its interactions with BTLA, CD160 and LIGHT. *Cellular & molecular immunology* vol. 16 679–682 Preprint at https://doi.org/10.1038/s41423-019-0241-1 (2019).

36. Hedrich, J. *et al.* Fetuin-A and cystatin C are endogenous inhibitors of human meprin metalloproteases. *Biochemistry* **49**, 8599–8607 (2010).

37. Pak, J. S. *et al.* NELL2-Robo3 complex structure reveals mechanisms of receptor activation for axon guidance. *Nat. Commun.* **11**, 1489 (2020).

38. Bryant, P., Pozzati, G. & Elofsson, A. Author Correction: Improved prediction of protein-protein interactions using AlphaFold2. *Nat. Commun.* **13**, 1694 (2022).

39. Yamagata, A. *et al.* Structural basis of epilepsy-related ligand–receptor complex LGI1–ADAM22. *Nat. Commun.* **9**, 1546 (2018).

40. Stastna, M. & Van Eyk, J. E. Secreted proteins as a fundamental source for biomarker discovery. *Proteomics* **12**, 722–735 (2012).

41. Clark, H. F. *et al.* The secreted protein discovery initiative (SPDI), a large-scale effort to identify novel human secreted and transmembrane proteins: a bioinformatics assessment. *Genome Res.* **13**, 2265–2270 (2003).

42. Möller, S., Croning, M. D. R. & Apweiler, R. Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics* **17**, 646–653 (2001).

43. Hekkelman, M. L., de Vries, I., Joosten, R. P. & Perrakis, A. AlphaFill: enriching AlphaFold models with ligands and cofactors. *Nat. Methods* **20**, 205–213 (2023).

44. Lane, T. J. Protein structure prediction has reached the single-structure frontier. *Nat. Methods* **20**, 170–173 (2023).

45. Zhong, B. *et al.* ParaFold: Paralleling AlphaFold for Large-Scale Predictions. *International Conference on High Performance Computing in Asia-Pacific Region Workshops* 1–9 (2022).

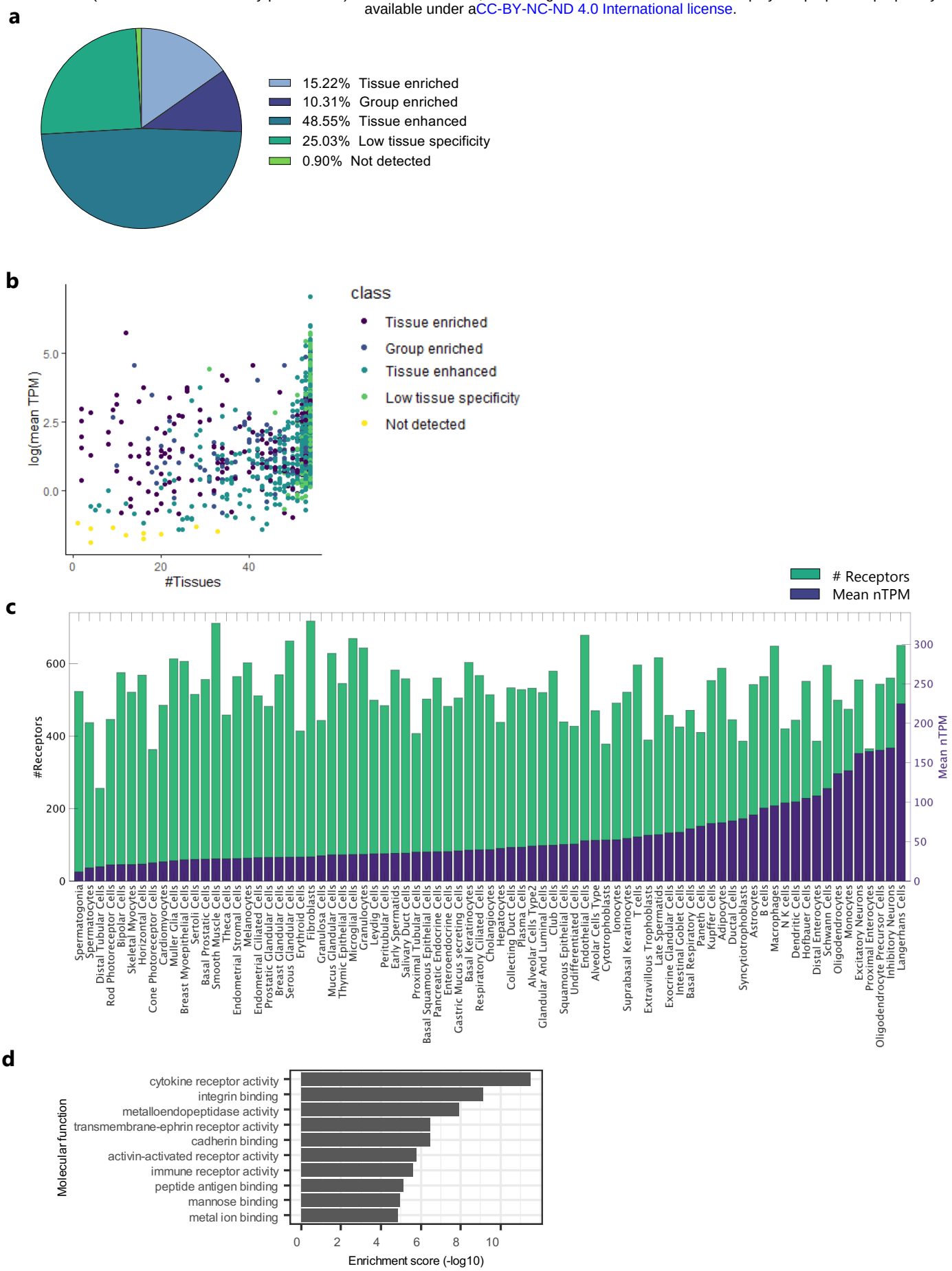
46. Pettersen, E. F. *et al.* UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci. Publ. Protein Soc.* **30**, 70–82 (2021).

47. Grant, B. J., Rodrigues, A. P. C., ElSawy, K. M., McCammon, J. A. & Caves, L. S. D. Bio3d: an R package for the comparative analysis of protein structures. *Bioinforma. Oxf. Engl.* **22**, 2695–2696 (2006).

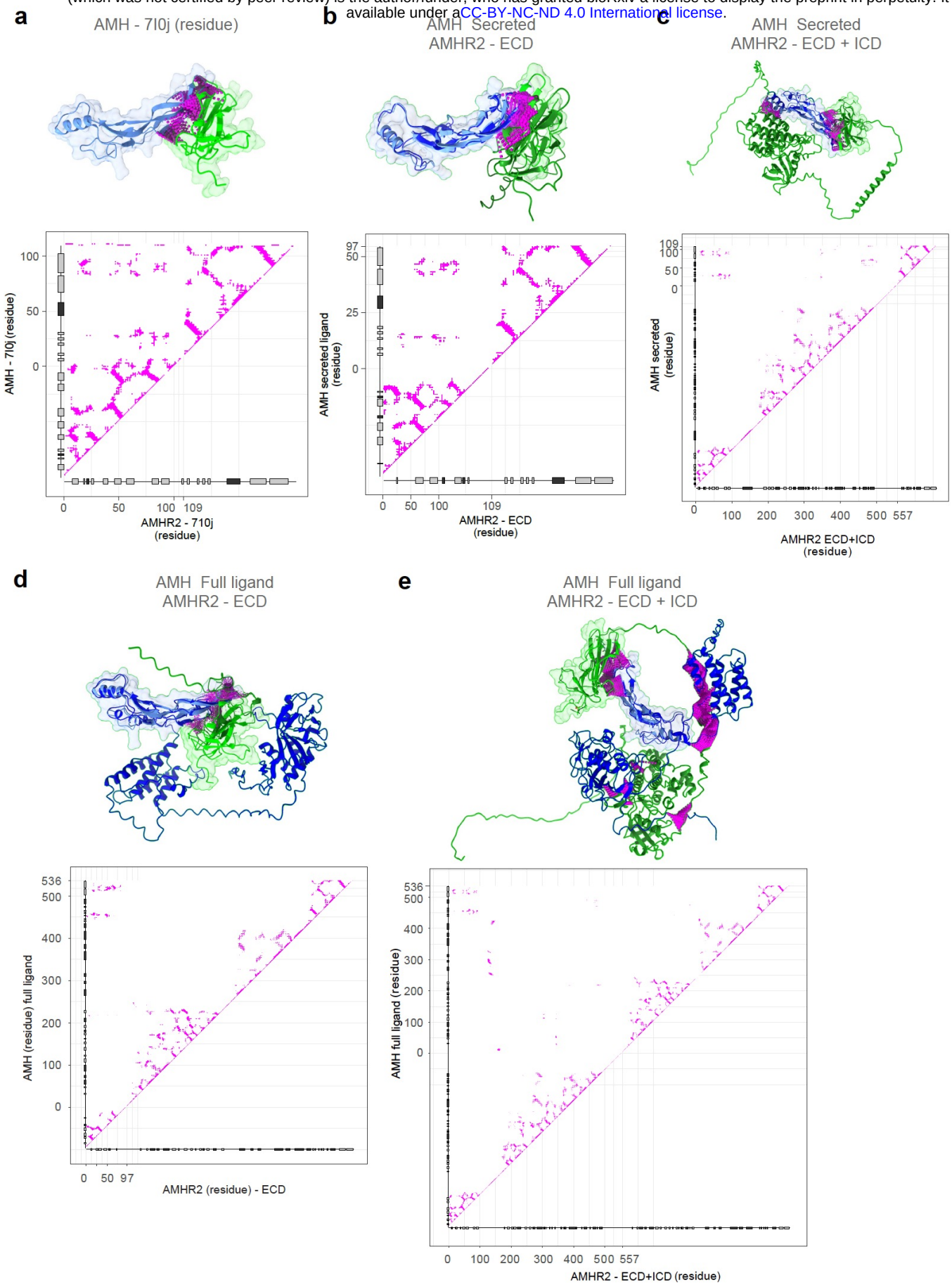
48. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **51**, D523–D531 (2023).

49. Uhlén, M. *et al.* Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).

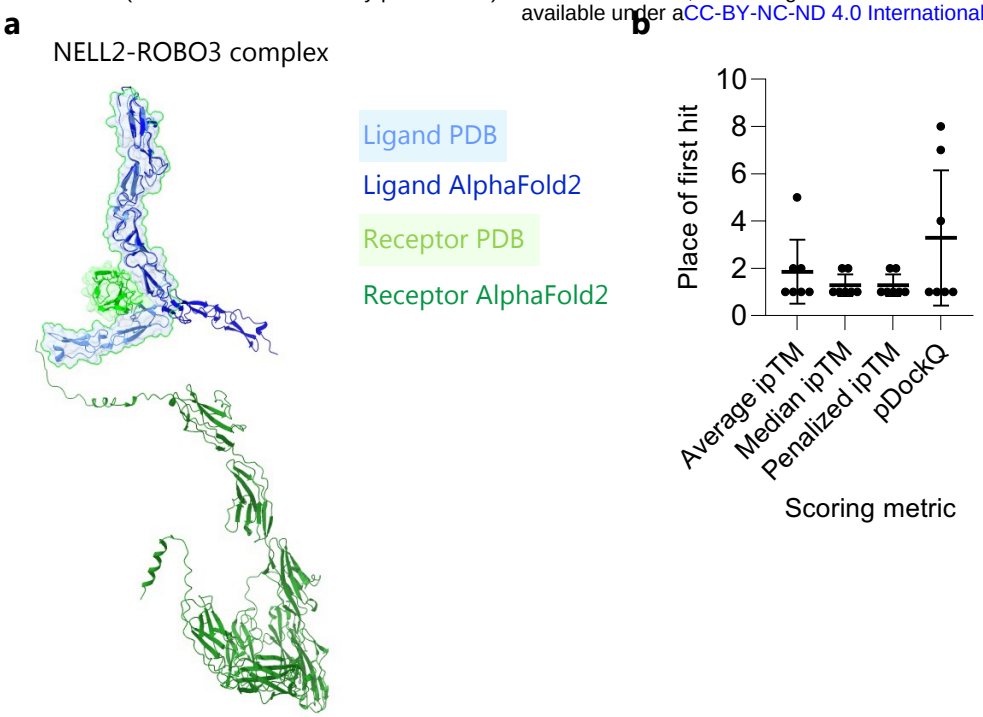
Supplementary information
Danneskiold-Samsøe et al.



Extended Data Fig 1. Composition of the single pass receptor library. a) Pie diagram of tissue distribution of the receptors in the library according to the human protein atlas (HPA). **b**) Receptor expression (mean nTPM) relative to the number of tissues each receptor is expressed in. Each dot represents a receptor. Colors denote tissue distribution according to HPA. **c**) Receptor expression (mean nTPM) relative the number of receptors (# receptors) across cell types. **d**) Classification of molecular functions for receptors in the library.

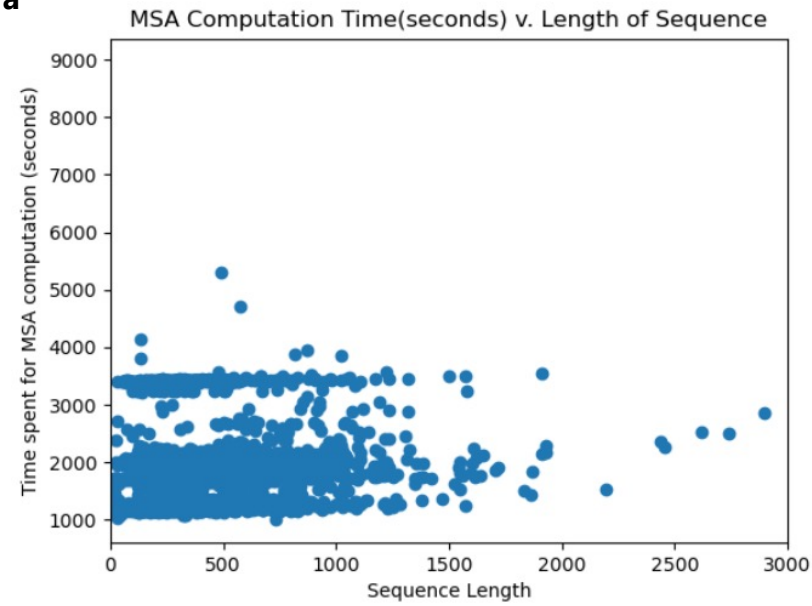


Extended Data Fig 2. Binding prediction of single-pass receptor ligand complexes of AMH-AMHR2 using full or truncated sequences. a-e) Structural binding prediction and corresponding contact maps for, light green: pdb database AMH, dark green: AF2 predicted AMH, light blue: PDB database AMHR2, dark blue: AF2 predicted AMHR2, purple: contacts **a)** PDB entry, **b)** secreted ligand and extracellular domain (ECD) of receptor, **c)** secreted AMH and full receptor including intra (ICD), transmembrane (TCD) and ECD, **d)** full ligand and ECD and **e)** full ligand and full receptor.

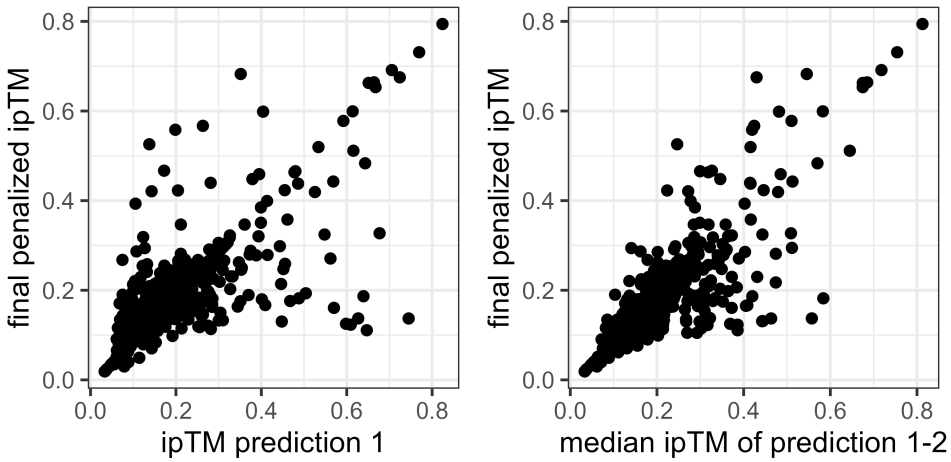


Extended Data Fig 3. AF2 accurately predicts single-pass receptors for secreted ligands. a) predicted structure of the NELL2-ROBO3 complex, **b)** ranking of first hit when using the scoring metrics average ipTM, median ipTM, or penalized ipTM and pDockQ.

a



b



Extended Data Fig 4. Computational cost and mitigation. a) cpu time per receptor in library, **b)** relationship between ipTM value in first predictions and final penalized ipTM value after five predictions for IL27.