Machine Learning Methods to Predict Olfactory Receptor-Ligand Binding without Docking

Sophia Leeman, Molly Borowiak, Jake Spruance, Ben Peng, Teodora Petkova, Austin Nicolas, Justice Lu, Hiroaki Matsunami, Fred Dietrich

Abstract

G-protein coupled receptors are the largest family of **drug targets**[12,] with **olfactory receptors** (ORs) being the largest class of these receptors[20]. However, little is known how odors bind to ORs. We developed 3 novel **machine learning** architectures to predict and gain insights about OR-odor binding, informed by the **largest existing experimental dataset** of deorphanized receptors and new capabilities to model protein structure with **AlphaFold**. Our models overcome the limitations of existing models by considering protein and ligand sequence and structure without requiring the structure of the ligand and protein bound together.

Objectives

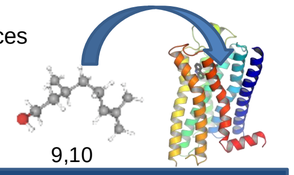
Predict OR-odor pairs

- Take advantage of refined structural data available from AlphaFold to supplement feature matrices

Develop alternative prediction models to offer a spectrum of viewpoints toward binding

- Compare model accuracy and interpretability

Gain insights about features important for binding



9,10

SL SRF: String-based Random Forest MB

Data Preparation

Features are **kmer** (k-length substring) frequencies of the input strings

- Ligands are represented as **SMILES** strings[7,14,17,18,19]
- Amino acids are categorized by **side-chain properties**[3]
- 3Di sequences are generated from AlphaFold structures via the novel structure comparison software, **FoldSeek,** and represent **tertiary interactions**[16]

Focus on features near the **binding pocket** of the protein

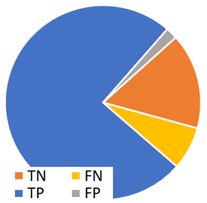
Model Architecture

Has a flexible **filtering method** that only admits kmers with a high likelihood of distinguishing between classified pairs[3]

- Kmer filtering + removing duplicate frequencies significantly reduces data but improves efficiency
- 273 protein/ligand interactions informed the model

Training algorithm is a **Random Forest** with balanced class weights

Results



Distribution of model predictions

Metric	Score
Balanced Accuracy	0.903
Binary Cross Entropy	0.338
Area under ROC Curve	0.991

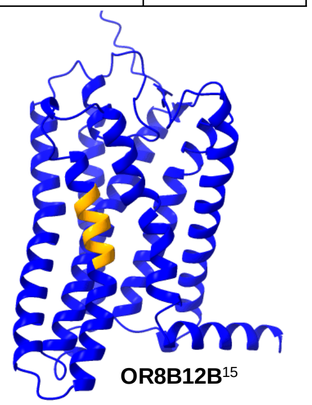
Interpretability

Model identified the highlighted region in **TM6** as the most important to binding

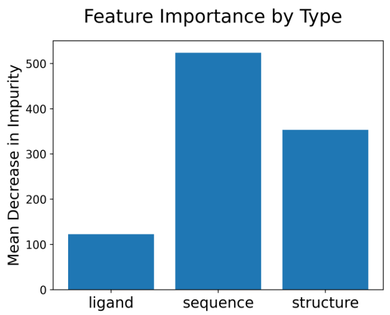
- TM6 is the region in GPCRs with the most **variation**[4]

This important region also includes the conserved motif, **FYG**.

- This motif is known to be associated with the **toggle switch,** which is important in protein activation[2]



OR8B12B[15]



Protein sequence was determined most important to the machine decision

The addition of structure **increases balanced accuracy** by 0.06 compared with sequence alone

JS CNN: Convolutional Neural Network BP

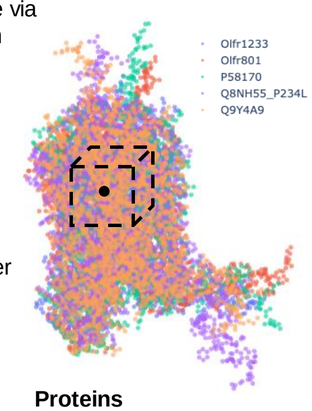
Data Preparation

Ligand and protein structures were represented in 3D space via "voxelization."[13] Weights were assigned to different points in space based on protein/ligand qualities:

- Aromaticity
- Hydrophobicity/hydrophilicity
- Hydrogen bond donor/acceptor status

Proteins were aligned by minimizing the root mean squared distance between each receptor and a template olfactory receptor, OR5D18

All olfactory receptors were centered on the geometric center of OR5D18's binding pocket



Proteins superimposed over the geometric center of OR5D18's binding pocket

Model Architecture

> 10000 known protein/ligand interactions were determined from literature

Dual-input convolutional neural network followed by 3 fully connected layers

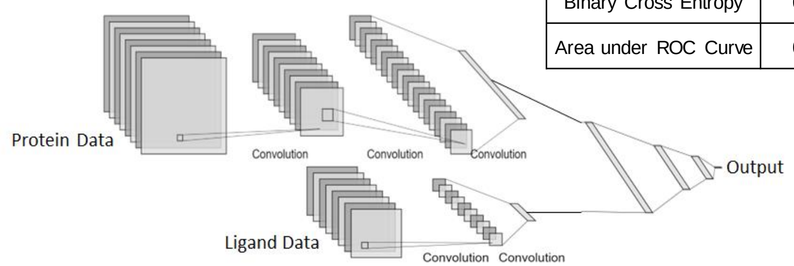
Inputs:

- Protein: 32 Å³ cube around each protein as aligned to OR5D18's binding pocket
- Ligand: 15 Å³ cube about each ligand's geometric center

Binary output indicating whether binding is expected

Metric	Score
Balanced Accuracy	0.824
Binary Cross Entropy	0.195
Area under ROC Curve	0.931

Results



TP GNN: Graph Neural Network AN

Data Preparation

The protein and ligand are represented by a graph embedded in 2D space with an adjacency and an atom feature matrix

Atom features are extracted from Alphafold prediction files using OpenBabel

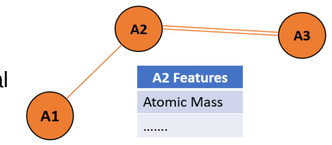
Zoom in on central 3,000 atoms of the protein

Model Architecture

4 Input Graph Convolutional Neural Network with both **regression** and **classification** outputs

Data from the 4 matrices is processed in 4 tensors using 4 different models which are concatenated and processed to get an output.

Grid Search method used to optimize hyperparameters for the regression model.



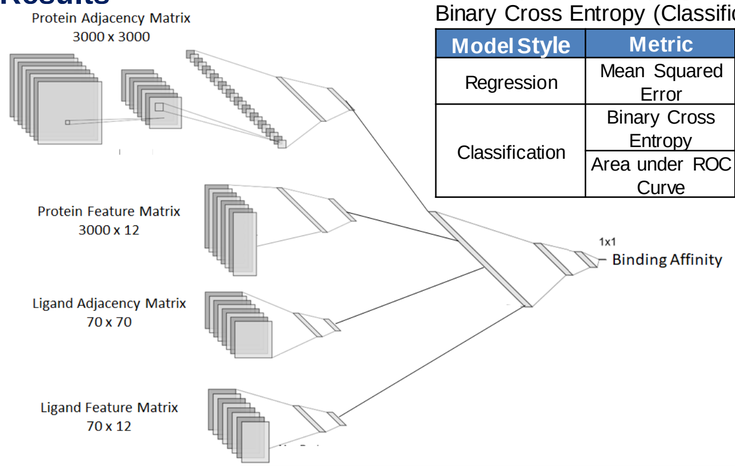
A1 Features
Atomic Mass
.....

A2 Features
Atomic Mass
.....

A3 Features
Atomic Mass
Atom Type
Exact Mass
Formal Charge
Hybridization
Isotope
Number of non-hydrogens attached
Partial Charge
Spin Multiplicity
Confidence Score

Adjacency Matrix			
	A1	A2	A3
A1	0	1	0
A2	1	0	2
A3	0	2	0

Results



Loss Functions:

Mean Squared Error (Regression)

Binary Cross Entropy (Classification)

Model Style	Metric	Score
Regression	Mean Squared Error	1.597
Classification	Binary Cross Entropy	0.33
	Area under ROC Curve	0.64

Conclusions

Each of the developed models comes with its own **advantages**

- SRF is highly accurate, interpretable, and efficient
- CNN utilizes the largest dataset and achieves the lowest binary cross entropy
- GNN provides both regression and classification output

Protein **sequence** made a higher contribution to impurity decrease and model prediction score[3]

- This could be a consequence of inaccurate structure prediction by AlphaFold
- The chemical interactions between a protein and ligand may be more important to binding

Utilizing information from **both sequence and structure** improves prediction score[3]

The neighboring residues to the **FYG region in TM6** were found to be a promising area for future research

Acknowledgements/Reference

Thank you to Brandon Fain for mentorship on this project and to the Duke Department of Computer Science for financial and computing resources.

References