# Prediction of GPCR activity using Machine Learning

Prakarsh Yadav,[†,§] Parisa Mollaei,[†,§] Zhonglin Cao,[†,‖] Yuyang Wang,[†,‖] and Amir Barati Farimani[*,†,‡,¶]

†*Department of Mechanical Engineering, Carnegie Mellon University, USA*

‡*Department of Biomedical Engineering, Carnegie Mellon University, USA*

¶*Machine Learning Department, Carnegie Mellon University, USA*

§*Joint First Authorship*

‖*Equal Contribution*

E-mail: barati@cmu.edu

**Abstract**

GPCRs are the target for one-third of the FDA-approved drugs, however; the development of new drug molecules targeting GPCRs is limited by the lack of mechanistic understanding of the GPCR structure-activity-function relationship. To modulate the GPCR activity with highly specific drugs and minimal side-effects, it is necessary to quantitatively describe the important structural features in the GPCR and correlate them to the activation state of GPCR. In this study, we developed 3 ML approaches to predict the conformation state of GPCR proteins. Additionally, we predict the activity level of GPCRs based on their structure. We leverage the unique advantages of each of the 3 ML approaches, interpretability of XGBoost, minimal feature engineering for 3D convolutional neural network, and graph representation of protein structure for graph neural network. By using these ML approaches, we are able to predict the GPCRs

activation state with high accuracy (91%-95%) and also predict the activation state of GPCRs with low error (MAE of 7.15-10.58). Furthermore, the interpretation of the ML approaches allows us to determine the importance of each of the features in distinguishing between the GPCRs conformations.

# Introduction

G-Protein Coupled Receptors (GPCRs) are important proteins for signaling networks, such as the Ras signaling pathway and neurotransmission.[1–5] Through these networks the GPCR proteins regulate the physiological processes and pathogenesis of diseases associated with such signaling networks.[6–8] The GPCR proteins have been a subject of significant academic attention, which is aimed at elucidating the mechanism of activation for novel GPCRs by structural biology and biochemical methods.[9] The ubiquitous presence of GPCRs in signaling networks makes them an important target of drug molecule-based therapeutics.[10] Approximately, 34% of all the drugs approved by the FDA target GPCRs with the objective of either activating (agonist) or deactivating (antagonist) the receptor.[11] With the developments in computational biology, biotechnology and pharmacology, there is an increased emphasis on efforts to regulate GPCRs activity via allosteric sites.[12,13] To accomplish this task and design therapeutics specific to the conformation of GPCRs, it is important to have a quantitative description of the protein conformations. Such quantitative description will allow for the improvements in rational drug design and accelerate the development of highly specific therapeutics with minimal side effects.[14]

The GPCR proteins are clustered together as a protein superfamily, which is further divided into subclasses based on the protein sequence similarity. The common feature in all[15] GPCRs is the presence of characteristic transmembrane helices, which create the ligand binding site in the extracellular domain and the binding site for G-protein or arrestin in the intracellular domain.[16] The binding of the ligand to the receptor causes conformational changes in the ligand binding site. These conformational changes are communicated across

the protein towards the intracellular region and results in large translations and bending of the transmembrane helices.[17] The comprehensive understanding of the amino acid level conformational changes, which cause transmembrane helix movement, is critical for the design of effective therapeutics.[18] The GPCR protein structures can be represented as quantitative models of their features, such as the dihedral angles and contact distances. These quantitative models of GPCRs are expected to be high-dimensional and difficult to interpret without high-throughput computational methods.[19,20] This leads to an interesting problem, can the features extracted from a GPCR structure model be used to predict its conformation?

Machine Learning (ML) can help answer the question of predicting GPCR state by learning the important features which distinguishes between the different conformations of the GPCR proteins. The developments in ML, especially Deep Learning (DL) methods, have increased their applicability to solve niche biological problems. For example, AlphaFold[21] and RoseTTAFold[22] have leveraged DL approaches to predict protein structure from sequence input. AlphaFold has also made available the predicted structure model of more than 350,000 protein sequences. Other works have used Convolutional Neural Network (CNN) for the prediction of protein-protein interactions,[23] protein-ligand binding,[24,25] protein folding,[26] protein phosphorylation site,[27] and protein structure classification.[28] However, CNN is designed for structured data like images, while the features which describe the GPCR conformation are unstructured.

To this end, Graph Neural Networks (GNNs)[29–32] are introduced for modeling the nodes and their relationships (edges) within the unstructured data. Modern GNNs learn the representations of graphs via aggregated message passing between the nodes.[33] Recently, GNNs have been leveraged in multiple domains concerning proteins, including protein-compound interaction,[34–36] protein folding prediction,[21,22] and function estimation.[37–39] Also, the application of ML has been further extended to discovery of GPCR agonist[40] and GPCR bioactive ligands.[41] We build upon the motivation of previous works utilizing ML for biological problems and create ML models which can predict the state of a given GPCRs protein structure

and also describe the extent of its activation. To accomplish this goal we require a large amount of training data comprising the GPCRs proteins in different conformations. We used the structure models for more than 500 GPCRs proteins determined by experimental methods, which includes the active and inactive conformations of the GPCRs, in addition to some intermediate conformations which may represent the transition between the two key states. We use the structure information for refined models of GPCRs from the GPCRdb server to develop quantitative models for the different conformations of the GPCR proteins, by predicting their state (active, inactive, or intermediate).[42] In this work, we present 3 approaches for this task, biophysics-aware feature engineering followed by shallow ML methods, 3D Convolutional Neural Networks (CNNs) with voxelization, and graph representation of protein followed by Graph Neural Networks (GNNs). The shallow ML method takes the biophysics-aware features as input, while the DL methods, including CNN and GNN, automatically extract features from the GPCR protein structures, thus requiring little or no domain knowledge.

We interpret and rank the importance of all the engineered features. The biophysics-aware feature engineering allowed us to discover and incorporate the important residue interactions and contact distances with the ML models. The ranking of engineered features also enabled us to identify the residue positions which are important for elucidating the GPCR conformation. The accuracy and robustness of the three approaches are benchmarked against each other in the tasks of predicting the activation state (classification) and the percentage of activation of the conformation (regression).

# Methods

## Machine Learning Models and GPCR structure dataset

The GPCRdb contains information about experimental data, phylogenetic diagrams, structures, and analysis tools for GPCRs.[42] This database provides insights into the molecular

mechanisms of GPCR activation, signal transduction, protein binding, and allosteric modulation.[42] From the GPCRs listed on the GPCRdb, we collect the protein data bank (PDB) structures of the biological complexes containing the G Protein Coupled Receptor.[43] Our dataset contains 555 PDB structures coming from 105 unique receptors' types included in GPCRdb (See supporting information for the structures and their properties). Each PDB structure of a GPCR is then converted to 3 different representations, including manually selected features, voxelization representation, and graph representation. These representations are used as input for different ML models (i.e. engineered features for XGBoost,[44] voxel representation for 3D convolutional neural network and graph representation for the graph neural network) to predict the activation state or the activity level (percentage activation) (Figure 1).

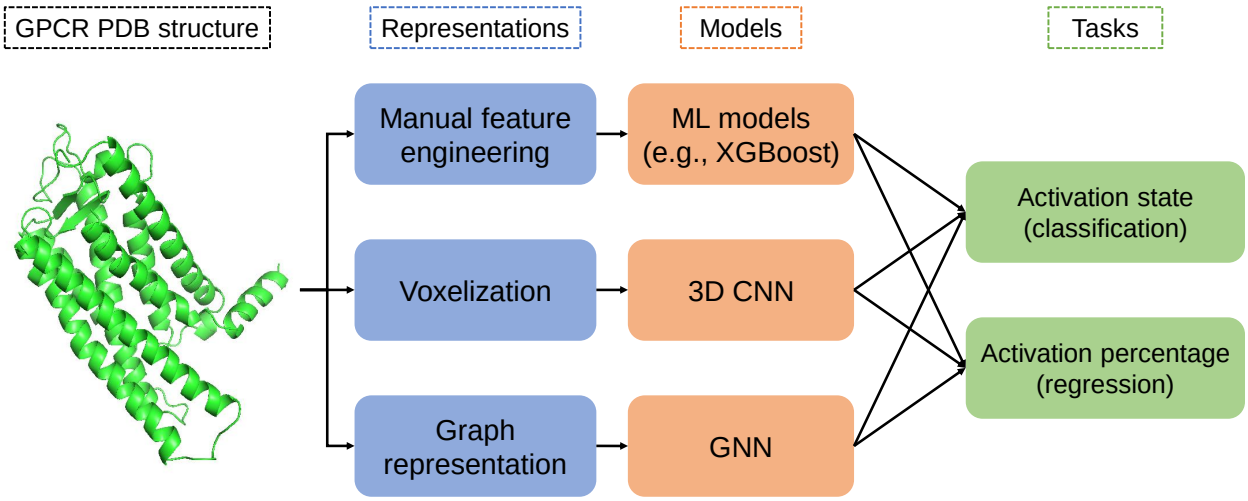


Figure 1: The three frameworks that we developed to predict the GPCR activation state. The PDB structure of each GPCR is converted into a representation (i.e. manually selected features, voxelized space, or graph representation) before being used as input to each machine learning model. The model will then output either the state or activation percentage of the GPCR.

## Feature Extraction and Shallow Machine Learning

The PDB structures for GPCRs often contain co-crystallized lipid molecules, non-GPCR proteins, and drug molecules. Therefore, the PDB structures are preprocessed before being used to generate the input for the ML methods. As a data preprocessing step, we extract only the structure of the transmembrane (TM) region of the receptors and removed all other segments of the proteins from the complex. The activation state labels for training the ML model are obtained from GPCRdb. The labels for the classification task correspond to 3 states of the receptors: active state, inactive state, and intermediate state. The intermediate state corresponds to the structure transitioning from between the active and inactive conformation. The labels for the regression task show the activation level for each receptor, which were also obtained from the GPCRdb. The truncated structures are then aligned with respect to each other to identify the position of amino acids in GPCRs with different peptide sequences. The aligned structures are used for feature extraction to describe the receptors' state. The input features for the shallow ML models comprise of the contact distances and distribution of reciprocal interatomic distances (DRID) of the identified amino acid pairs and the TM helix pairs. To assess the performance of shallow ML models, we train and evaluate different ML models, including Random Forest, Support Vector Machine, XG-Boost, and Logistic Regression. 5-fold cross validation is performed for these models at both the classification task and the regression task.

To manually engineer the protein features for shallow ML prediction, we select 420 residue pairs by randomly picking 20 residue pairs on each of the 21 possible TMs pairs. The average closest heavy-atom distance of the residue pairs is then calculated from all of the 555 protein structures in the dataset. It is observed that for a given contact distance pair, a larger difference between the average values of active and inactive conformation ($\Delta M$, measured as a distance between peaks of histograms in active (green plot) and inactive (red plot) states, Figure 6c-f) correlates with higher prediction accuracy (Figure 6a). To achieve high classification and regression prediction accuracy, we rank the 420 random residue pairs based

on the prediction accuracy by using each individual feature. We then choose the top 105 features (5 features per TM pair) to train our model (Table S3). Furthermore, in the list of 105 pairs, we observe 21 residue pairs containing at least one residue belongs to the polar network of GPCR activation (Table S2). For these features we extract the $C-\alpha$ contact distance and the Distribution of Inverse Reciprocal Distances (DRID). The concatenated vector of contact distances and DRID is the input vector for the ML models. XGBoost model and Random Forest classifiers are implemented to predict the states of the GPCRs and corresponding regressors are implemented to predict the activation level. XGBoost is a gradient boosting framework which utilizes second order derivatives and Random Forest is an ensemble method which uses decision trees and takes the mode of the trees as the output.[44,45] We implement these algorithms from the Scikit-learn python library.[46] Random Forest classifier is implemented with 25 decision trees in the ensemble, bootstrapped initialization of the trees and Gini impurity for calculation of information gain. The XGBoost classifier is implemented with the gbtree boosting method, learning rate of 0.3, max tree depth of 6, and uniform sampling of the training instances.

## Voxelization and 3D Convolutional Neural Network

Voxelization is a technique for mapping the continuous 3D space to a discrete 3D mesh grid using unit cubic cells (voxels).[47] It has been used to generate geometrical representation of 3D objects for applications such as computer-aided design model classification[48] and 3D vision.[49,50] Recently, voxelization has been introduced to the domain of molecular/atomic property predictions. The voxelized 3D space can preserve 3D atomic structural information while making ideal input to CNN.[23] CNN has achieved outstanding performance in applications such as image classification.[51–54] The convolutional layers in the CNN function as automatic feature extractors that detect important features from the input without human supervision. Combining the voxelization and 3D CNN creates a framework to learn the representation via filters in an efficient way. CNN can automatically extract relevant

features from the input so that feature engineering is no more needed. The combination of voxelization and CNN achieved high accuracy in tasks such as predicting bioactivity of small molecules,[55] interatomic force and potential prediction,[56,57] and prediction of binding affinity of protein-ligand complexes.[58] The combination of 3-dimensional voxelization and CNN can be another approach to featurize the GPCRs.

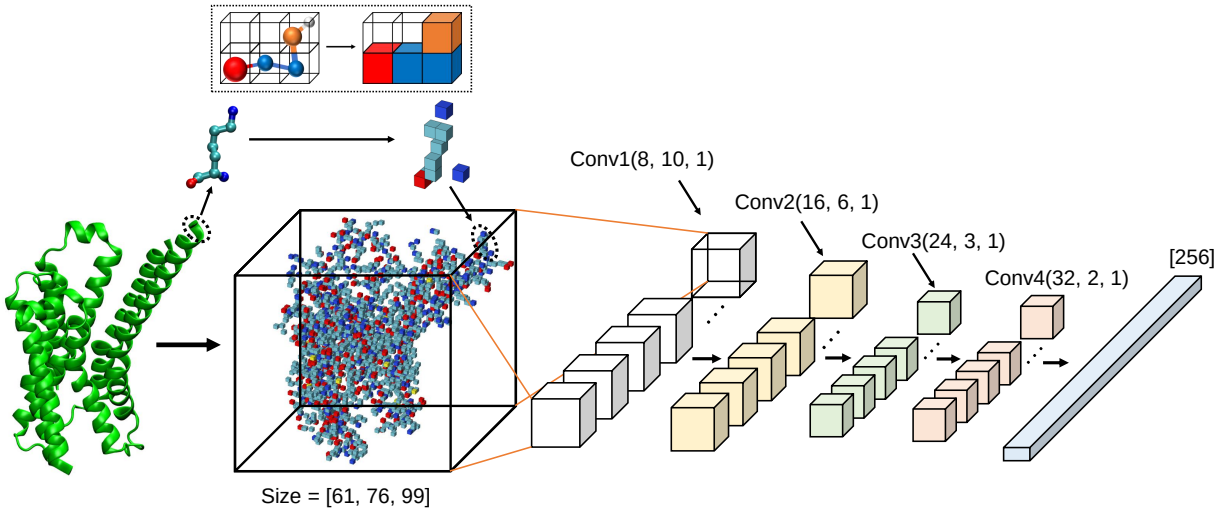


Figure 2: Voxelization and 3D convolutional neural network for GPCRs. At the top of the figure is a cartoon showing the voxelization process of atomic structures. The left part of the figure shows the voxelized 3D structure of a GPCR. Cyan, blue, yellow, and red voxels represent carbon, nitrogen, sulfur, and oxygen atoms, respectively. Voxels containing hydrogen atoms are not shown to avoid an over-crowded view. A zoom-in view of the voxelization of one residue is shown. A 4-layer 3D CNN extracts a feature vector from the voxelized GPCR structure for activity prediction.

The voxelization of the GPCRs is shown by the cartoon at the top of Figure. 2. The 3D space including and around an atomic structure is represented using a grid of voxels, called the voxelized space. To enforce the consistency of the size of input to the CNN, We take the maximum value of the length, width, and height of all the GPCRs in the dataset and construct a voxelized space that can fit all GPCRs with the size of 61Å $\times$ 76Å $\times$ 99Å in the $x$, $y$ and $z$ dimension. In this work, each voxel is a cube with the side length of 1Å, thus there are 458964 voxels in total. All voxels have an initial value of 0. If the cartesian coordinates of an atom are within a voxel, the voxel is then given a value of the atomic

number of the atom in it (e.g. 6 for carbon and 8 for oxygen atom). In some rare cases when two atoms can appear in the same voxel, for example, a hydrogen atom is diagonally opposite to another atom in one voxel, the voxel is given the value with the greater atomic number of the two atoms. Since most overlapping happens between a non-hydrogen atom and a hydrogen atom, which has less effect on the GPCR property,[23] taking the greater atomic number can minimize the information loss. 3D CNN differs from 2D CNN from the dimension of the convolution filters. For this specific work, we choose to use 3D CNN instead of 2D CNN because the former can better extract 3D structural features from the GPCRs data.[23,25,26] The voxelized GPCR structures are then fed into a 4-layer 3D CNN to extract features. There are 8, 16, 24, and 32 filters, and the kernels are cubes with sizes of 10, 6, 3, and 2 for each layer of the CNN, respectively. (see Figure 2 The stride for the convolution calculation in all layers is 1. 3D batch normalization and average pooling are applied after each convolutional layer.[59] The feature vector output from the CNN has a size of 256. A 3-layer fully-connect neural network, which has 400, 256, and 64 neurons in each layer, is used to predict the activity of GPCR using the feature vector. ReLU activation[60] is used in both the CNN and the fully-connected neural network.

CNN performance scale very well with the size of training sets. In addition, CNN is translation-rotation invariant, making data augmentation a viable solution and technique for enhancing their performance. We perform data augmentation[53,61,62] on the voxelized GPCR structures. Each voxelized GPCR structure is augmented by flipping along either $x$, $y$, or $z$ axis, and the augmented samples are given the same label (Figure S2). The size of the dataset is increased by 3 times through this method. During the 5-fold cross validation, original GPCR structures and their augmented samples in the training set are used for training the model, while only the original GPCR structures in the test set are used for testing.

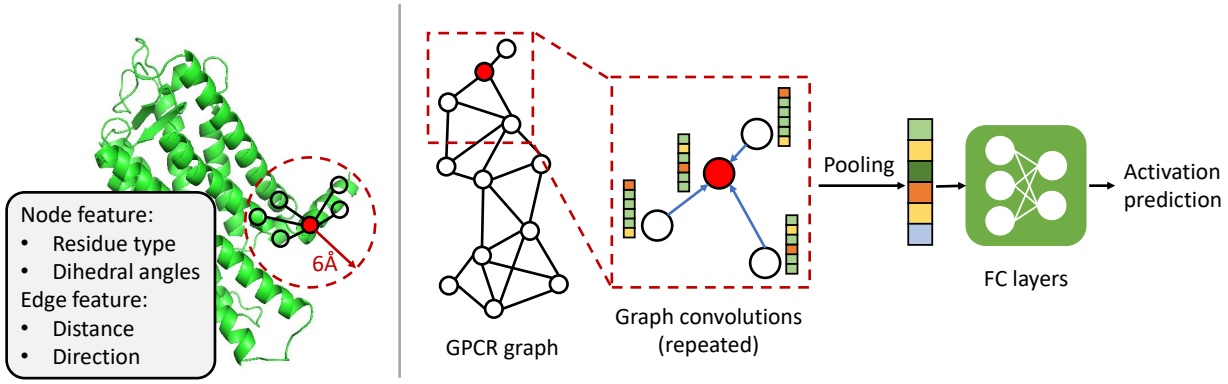


Figure 3: Overview of the Graph Neural Network (GNN) framework for GPCR activation prediction. Left: Each node in the GPCR graph represents a residue and edges are created to connect residue pairs within 6Å. Nodes are described by the amino acid type and the dihedral angles, while edges are defined by the distance and direction between residues. Right: The GNN model takes in the GPCR graph and sequentially runs through the graph convolutional layers, a pooling layer, and fully-connect layers to predict the activation.

## Graph Neural Network

Previous ML and 3D CNN models require either manually engineered features or voxelization to process proteins into Euclidean data, which can lose important information. Graph Neural Networks (GNNs),[29,31] on the other hand, are designed to learn representation directly from unstructured graphs, such as chemical compounds.[21,24,32] GNN is built upon the graphical data which consists of different numbers of nodes and edges in between. Such a strategy provides a more flexible way to model the protein structure and helps keep structural information. In our case, a GPCR is considered as a graph $G$ , where $V$ denotes all the residues (nodes) within a GPCR and $E$ denotes connections (edges) between residues.[34,63,64] We further define $X_v$ as the node attribute for $v \in V$ and $e_{uv}$ as the edge attribute for $(u, v) \in E$. In the GPCR graph as shown in Figure 3, each node represents a residue and each edge represents the distance between neighboring residues. We define the distance between residues as the distance of alpha carbons ($C_\alpha$), which captures the backbone structure of GPCRs. Neighboring residues within 6Å are connected by edges, which nicely covers the adjacent residues and meanwhile excludes remote residues to increase computational efficiency.

For each node $v$, $X_v$ is defined as the 20-dimensional one-hot encoding of all the 20 amino acid type, $A_v$, together with the dihedral angles $\phi_v$ and $\psi_v$, namely $X_v = [A_v, \phi_v, \psi_v]$. Each edge attribute $e_{uv}$ is defined as $e_{uv} = [d_{uv}, x_{uv}, y_{uv}, z_{uv}]$, where $d_{uv}$ is the distance and $[x_{uv}, y_{uv}, z_{uv}]$ is the normalized directional vector between node $u$ and $v$. In the experiments, combinations of different node and edge features are investigated. Namely, we compare the test accuracy of GNN models with different features included to explore which input features contribute more to the prediction of GPCR activation. For example, the GNN model is trained with only node amino acid type as the node feature in comparison with the model with both amino acid type and dihedral angles included.

The Graph Neural Network (GNN) takes in the node and edge attributes and update the node representations iteratively through aggregation and combination operations.[29,65–67] Let $h_v^{(k)}$ denote the representation of node $v$ at the $k$-layer in the GNN and $h_v^{(0)}$ is initialized as $h_v^{(0)} = X'_v = l_v(X_v)$, where $l_v$ is a linear projection which maps $X_v$ to the embedding dimension. Similarly, edge attributes are mapped to the same dimension through a linear projection $l_e$, where $e'_{uv} = l_e(e_{uv})$. We build our GNN following the Graph Isomorphism Network (GIN)[68] with edge attributes included[69] as this has been demonstrated a powerful GNN model in various applications.[33,70] The update rule of node representations in each graph convolution layer is defined as Equation 1:

$$h_v^{(k)} = f^{(k)}\left(h_v^{k-1} + \sum_{u \in \mathcal{N}(v)} \sigma(h_u^{(k-1)} + e'_{uv})\right), \tag{1}$$

where $\mathcal{N}(u)$ denotes all the nodes directly connected to $u$ through edges, $\sigma(\cdot)$ is the activation function, and $f^{(k)}(\cdot)$ is the non-linear update function. In our case, ReLU[60] is developed as activation function and $f^{(k)}(\cdot)$ is modeled by fully-connected layers. After $K$ layers of graph convolutions, we obtain the updated node representations $h_v^{(K)}$ for $v \in V$. An average

pooling is implemented to extract the graph representation $h_G$ as given in Equation 2

$$h_G = \frac{1}{|V|} \sum_{v \in V} h_v^{(K)}. \tag{2}$$

Another prediction head is developed using fully connected layers which takes $h_G$ and outputs the prediction of the activation, which can be either the classification of whether the GPCR graph is activated or the regression of the activation percentage.

In our GNN model, we develop 4 graph convolutional layers with the dimension of node representation 128 and a batch normalization layer[59] is added after each graph convolutional layer. The prediction head contains 2 hidden layers with dimensions 64 and 32, respectively. The model is trained for 100 epochs with batch size 32. Adam is leveraged to update the weights with an initial learning rate of 0.001 and weight decay $10^{-5}$. The learning rate decays to 0.0001 after training for the first 50 epochs.[71] Similarly to training the shallow machine learning and 3D CNN models, we run 5-fold validation and within each fold the whole dataset is randomly split to training/validation set by the ratio of 4:1.

# Results and discussion

## Experimental Results

Table 1 lists the GPCR state classification results of different machine learning methods on the 5-fold validation. We implement accuracy and F1 score as two metrics to evaluate the classification performance of each model. The accuracy measures the ratio of correctly predicted instances. The F1 score is defined as the harmonic mean of precision and recall of the predictions, where precision is the fraction of true positives and all positive predictions and recall is the fraction of true positive and the total positive samples. Comparing to accuracy, F1 score provides a better metric when data is imbalanced, i.e., number of data varies in each category. In the shallow ML tests two protein featurization schemes were

used; the distance between any two non hydrogen atoms in 420 random pairs of residues and selected 105 pairs providing the most ML accuracies and those residues which are involved in the polar network of proteins. XGBoost model with 105-pairs surpasses the other models on both prediction accuracy and F1 score, indicating that the selected features capture the important residue networks which reflect the GPCR activation state. Comparing to XGBoost, GNN also achieves comparable performance while requiring much less feature engineering compared to the XGBoost and learning to extract important features related to the activity of GPCRs. The regression results for the percentage of GPCR activation, are also shown in Table 2 with both rooted mean square error (RMSE) and mean absolute error (MAE) reported. The results show that 3D CNN fails to compete with other methods.

One possible explanation for the lower performance of 3D CNN is that the voxelization of amino acids may lead to loss of important information about the type of residues and its stereo-chemical properties. Additionally, the structures for GPCRs from X-Ray diffraction and Cryo-EM can feature missing side chain information for certain residues and in certain cases has no information for the hydrogen atoms. Since the graph featurization method does not require atomic positions of all atoms, it is agnostic to the lack of such information. The graphical featurization of proteins captures the global structural features of GPCRs, such as the TM3-TM6 distance, in addition to the individual amino acid features. Therefore, even in the cases of missing residues from X-Ray structures, this method is able to accurately classify the GPCR structure as active, inactive and intermediate. Additionally, in the case of GPCRs, the global conformation changes are encoded into the relative distance between residue pairs. The absence of side-chain atom information will not eliminate the positional information of the residues. Since, in the manual featurization scheme we have calculated the closest heavy atom distances, the relative distance between residue pairs is still encoded into the XGBoost model input. This ensures high robustness against missing residues in the model input, as even if the positional information of a few atoms is missing, the remainder of the residue is sufficient for capturing the global features of GPCR conformation. In GNN,

we conduct mean pooling which averages over all the updated node features to extract the feature for the whole GPCR structure. Thus, even some residues may be missing in the data. This will not affect the implementation of GNN models. Finally, even the 3D CNN methodology, which explicitly voxelizes the atomic positions of all residues, can capture the global features very well and generates accurate predictions for the GPCRs. To verify the impact of hydrogens on prediction accuracy of 3D CNN, we created an augmented dataset of GPCR structures with all hydrogens removed. We then evaluated the pre-trained model performance specifically for the structures for both instances (GPCRs with and without hydrogens). The results from Table. S5 and Table. S6, demonstrate that the performance of 3D CNN slightly drops for both classification and regression tasks with the absence of atomic information of hydrogen atoms, indicating that the presence of hydrogen is beneficial to 3D CNN in predicting activation of the GPCRs. To further boost the performance of 3D CNN more training data can be incorporated. It is noteworthy that the data augmentation improves the classification accuracy of 3D CNN from 0.8829 to 0.9117, and the regression MAE from 0.131 to 0.1058, respectively. Therefore we can expect a better performance of 3D CNN from a larger training dataset.

To better understand the impact of mutations on GPCR conformations and sensitivity of the presented methods to such mutations, we analyzed cases of reported mutations in the GPCR and their corresponding effect on protein conformation. For instance, in the case of Neurotensin 1 Receptor (NTSR1), the experimental method of directed evolution has been used to obtain the NTSR1 protein in the inactive conformation (PDB: 3ZEV). Directed evolution caused the protein to acquire 11 point mutations, (namely A86L, H103D, H105Y, A161V, R167L, R213L, V234L, I253A, H305R, F358V, and S362A).[72,73] These mutations were attributed to higher stability of the inactive conformation of NTSR1 and they have a cumulative effect on the inactivation of NTSR1. Similarly, in the case of Glucagon like Peptide 1 Receptor (GLP-1R), the active (PDB: 5VAI) and the inactive (PDB: 6LN2) structures have a sequence identity of 90.48%,[74,75] indicating that during the experimental processes

of determining these structures, mutations were accumulated. However, the featurization schemes proposed in this work can explicitly encode such mutation information in the inputs to the GNN, 3D CNN and XGBoost models. The featurizing schemes are able to incorporate atomic level information and the spatial orientation of amino acids, therefore, the feature space is influenced by the mutations occurring in the structures. Voxelization can encode spatial organization of atoms and atoms types, where even a single atom change is encoded into the model input. In the case of GNN, each node denotes the residue. Therefore, the mutation information is naturally encoded in the GNN models, as mutated GPCR possess different residues at certain positions. Finally, the inter-residue distance based featurization for XGBoost encodes the spatial atomic organization by considering the closest heavy atom distance between the amino acid pairs. These featurization schemes allow the encoding of point mutations into the model inputs and then predict the GPCR conformation, and percentage of activation with high accuracy.

Table 1: Performance of different models for GPCR state classification. We report the mean and standard deviation of the accuracy and F1 score in 5-fold validation.

| Model | Accuracy | F1 score |
|---|---|---|
| XGBoost for 105-pairs | 0.9586 (0.0044) | 0.9571 (0.0033) |
| XGBoost for 21-pairs | 0.9369 (0.0098) | 0.9339 (0.0102) |
| 3D CNN | 0.9117 (0.0438) | 0.7708 (0.1411) |
| 3D CNN w/o data augmentation | 0.8829 (0.0446) | 0.7604 (0.1634) |
| GNN | 0.9585 (0.0176) | 0.9386 (0.0296) |

Table 2: Performance of different models for regression to degree of GPCR activation. We report the mean and standard deviation of the RMSE and MAE in 5-fold validation.

| Model | RMSE | MAE |
|---|---|---|
| XGBoost for 105-pairs | 0.1291 (0.0701) | 0.0715 (0.0074) |
| XGBoost for 21-pairs | 0.1605 (0.0682) | 0.0969 (0.0077) |
| 3D CNN | 0.1420 (0.0394) | 0.1058 (0.0289) |
| 3D CNN w/o data augmentation | 0.1750 (0.0238) | 0.1310 (0.0231) |
| GNN | 0.1449 (0.0048) | 0.0897 (0.0041) |

## Misclassified GPCRs and tSNE visualization

To further compare the machine learning models, we investigate the distribution of misclassified GPCRs by class and the confusion matrix for each model. Figure 4a shows that GPCRs of intermediate state are difficult for the 3D CNN model to classify. 3D CNN misclassified 33, 6, and 10 GPCRs in intermediate, inactive, and active states, respectively, indicating that intermediate state GPCRs are very hard for 3D CNN to distinguish. Similar to 3D CNN, XGBoost model also misclassified more GPCRs of intermediate state than of the other two states. Compared with 3D CNN and XGBoost model, GNN demonstrates the same level of accuracy for GPCRs of all activation states. The confusion matrix of each model is normalized over the ground truth condition (every row). Although 3D CNN model achieves 98% and 95.5% accuracy in classifying inactive and active GPCRs (Figure 4d), the model misclassifies 75% of the intermediate GPCRs to the inactive state. The low accuracy for 3D CNN in classifying the intermediate state GPCRs not only corresponds to its lower F1 score (Table 1) compared with the other 2 models but also hinders it to achieve higher accuracy as the GNN. Moreover, the XGBoost and GNN model tend to misclassify inactive of active GPCRs to the opposite activation state instead of the intermediate state. For example, XGBoost model misclassifies none of the inactive or active GPCRs as an intermediate state. On the other hand, the 3D CNN tends to misclassify the inactive or active GPCRs to the intermediate state instead of the opposite activate state. A reason for the difference of models in the misclassification is that the 3D CNN makes predictions based on the extracted structural features of the GPCRs, and the structural features can be very different between the inactive and active GPCRs.

The t-SNE[76] visualization of the features learnt by 3D CNN and GNN can help us rationalize the superior performance of the GNN method compared to 3D CNN. t-SNE is a dimensionality reduction technique that preserves the local structure of the data points (i.e. data points are similar to each other if they are clustered). It has been vastly adopted in the field of bioinformatics.[77–79] A latent feature vector for each GPCR is extracted during

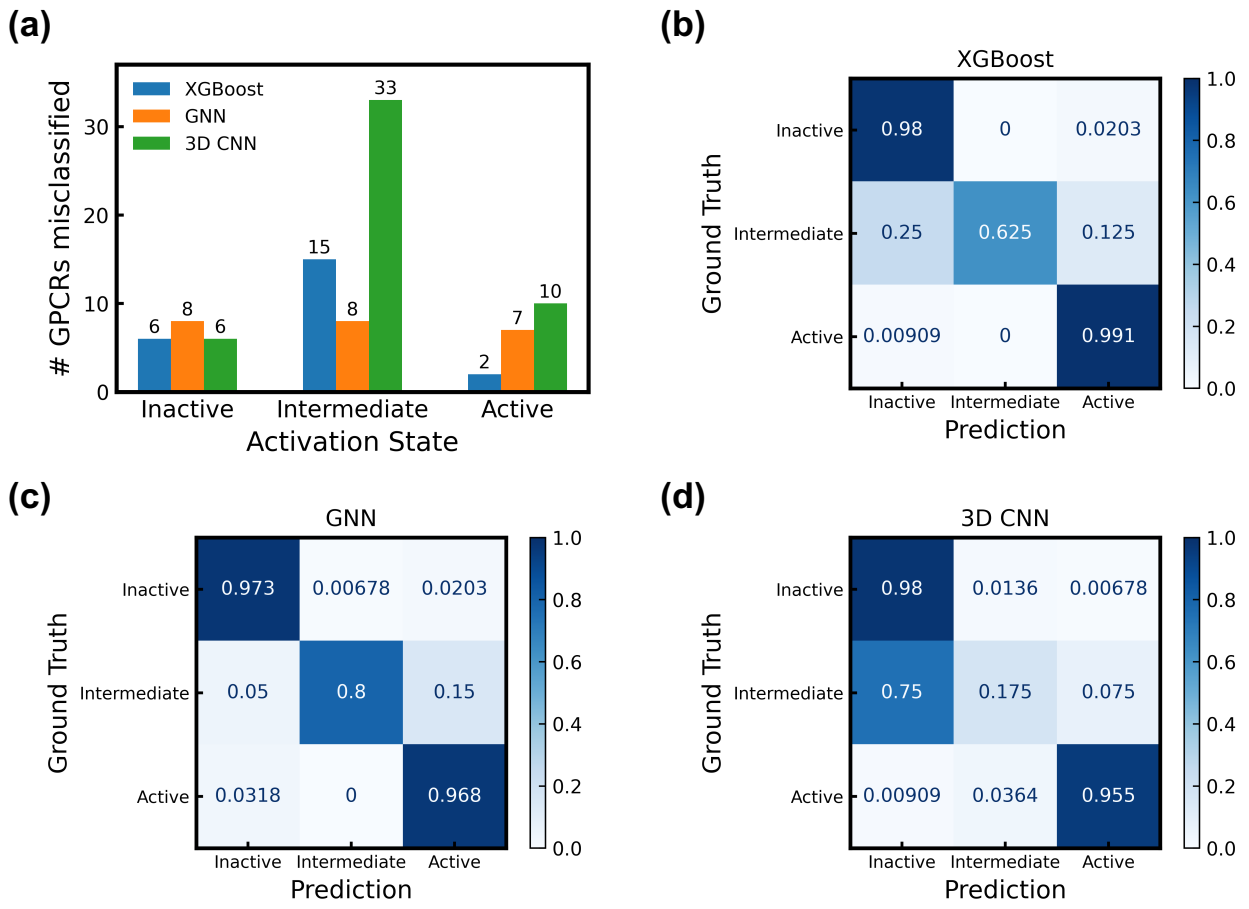


Figure 4: **(a)** Number of misclassified GPCRs in each activation state for the three models. **(b)-(d)** Normalized confusion matrix of each model.

the training of GNN and 3D CNN models. The feature maps learnt by GNN and 3D CNN can be passed through a series of Fully Connected layers of neural networks (Multi Layer Perceptron, MLP). A latent feature vector is the output of the last layer of the FC layers. The output of the last layer is a dense feature representation of the inputs as it represents all the information learnt by the preceding layers of the network. This latent feature vector can be used as an input for visualization in 2D space by using t-SNE.

Here, t-SNE is used to map the latent feature vector of all GPCRs into a 2D space (Figure 5). In this 2D space, the t-SNE embedding 1 is the mapping of the latent feature vector onto the first dimension and t-SNE embedding 2 is the mapping of the latent feature vector onto the second dimension. Although the active and inactive GPCRs are well-separated in

the 2D visualization of the latent feature learned by 3D CNN (Figure 5a), the boundary of the intermediate state GPCRs cluster is not obvious. This corresponds to the high misclassification rate of 3D CNN on intermediate state GPCRs because the feature extracted by 3D CNN does not distinguish the intermediate state from the others. On the other hand, in the case of GNN latent feature visualization, almost all GPCRs are clustered with other members of the same activation class. The comparison between the t-SNE visualizations of 3D CNN and GNN extracted features shows that GNN can maximize the difference between GPCRs of different activate states, which results in the higher accuracy of GNN in the classification task. Moreover, the t-SNE visualization can help to find mislabeled GPCRs. For example, the GLP1R_5NX2 GPCRs (bottom right corner of both Figure 5a and 5b) is labeled as an intermediate state but clustered with active state GPCRs using both GNN and 3D CNN learnt features. The classification for intermediate class is also particularly challenging due to the lack of consensus in the biophysics literature for defining intermediate state of GPCRs.[80,81] The resultant ambiguity in the intermediate label creates confusing classification task for both the 3D CNN and GNN models.

## Feature Engineering and XGBoost

We investigate the prediction accuracy using those features involved in polar network in GPCR activation (Table S1).[82] Such features include the hydrogen bonds stabilizing both the active and inactive states of opioid receptors which have to be rearranged to achieve the active conformation. Most of the residues engaged in the polar network are conserved, suggesting that they may have similar functions in GPCR activation. We observe that the list of top 105-pairs includes 21 residue pairs where at least one residue belongs to the polar network in GPCRs (Table S2). By using these 21-pairs of contact distances as input, we were able to predict the GPCR state with 92.25% accuracy for the Random Forest model and 93.69% accuracy for the XGBoost model. The comparison of prediction accuracies through Random Forest and XGBoost model reveals that the difference between their prediction

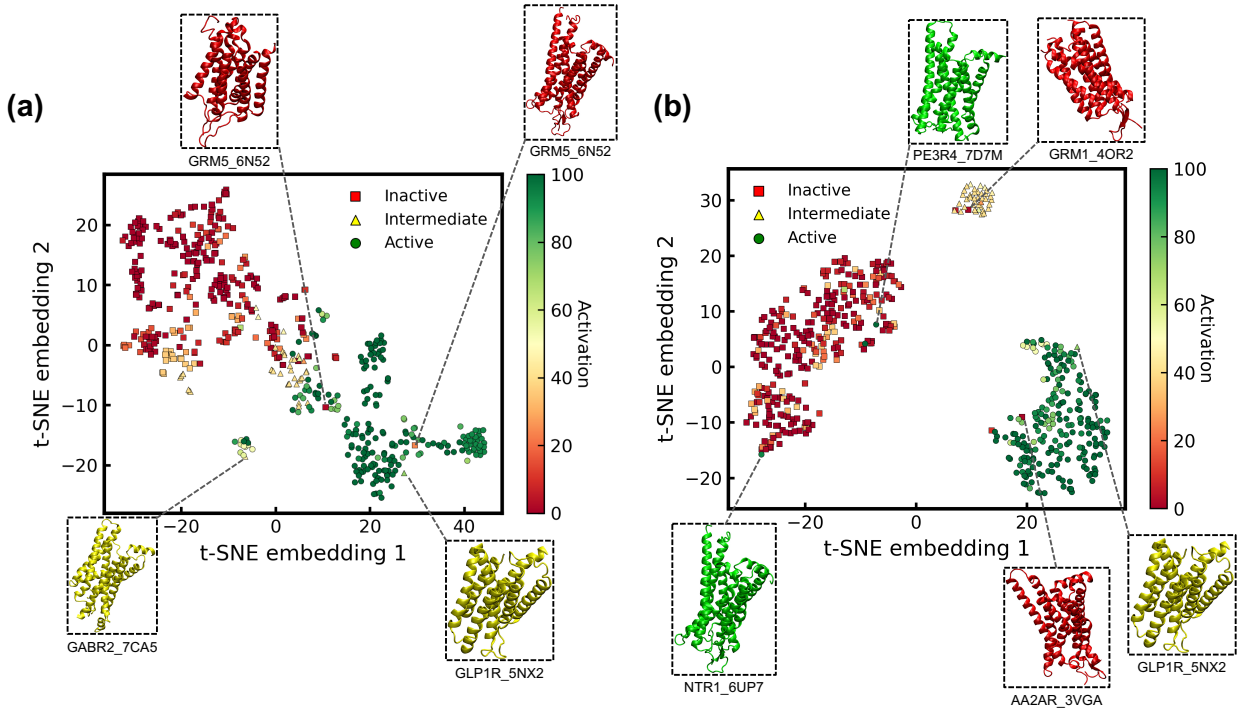


Figure 5: t-SNE dimension-reduced visualization of GPCRs using features learned by **(a)** 3D CNN and **(b)** GNN. Each GPCR is colored based on its activate level. Rectangle, triangle, and circle scatter points represent inactive, intermediate and active GPCRs, respectively. Snapshots of some GPCRs that are clustered to other activation states are selected to be shown.

accuracies for different datasets is no more than 1.44 percent (Figure 6b).

We explore the normalized histogram of the closest distance between two non-hydrogen atoms of the residue pairs over 555 proteins and the prediction accuracy out of XGBoost model, with corresponding PyMOL[83] representation of the residue pair on NTS1 protein (Figure 6c-f). Here $\Delta M$ is measured as the distance between peaks of histograms in active (green plot) and inactive (red plot) states. 3.40-7.49 residue pair achieve 72% prediction accuracy where 7.49 residue belongs to the polar network of GPCR activation. On NTS1 protein, $N^{7.49}$ rotates 46.6° and translates 1.6Å. On the other side, $A^{3.40}$ rotates 6.6° and translates 1.5Å (Figure 6c). The pair of 6.38-7.46 in which 7.46 residue engaged to the polar network obtained 69% prediction accuracy. The residue pair rearrangement on NTS1 protein is associated with a 3.6Å translation and 69.4° rotation of $S^{7.46}$ and 4.1Å translation and 32.5° rotation of $R^{6.38}$ (Figure 6d). 6.31-3.32 residue pair, where 3.32 residue belongs

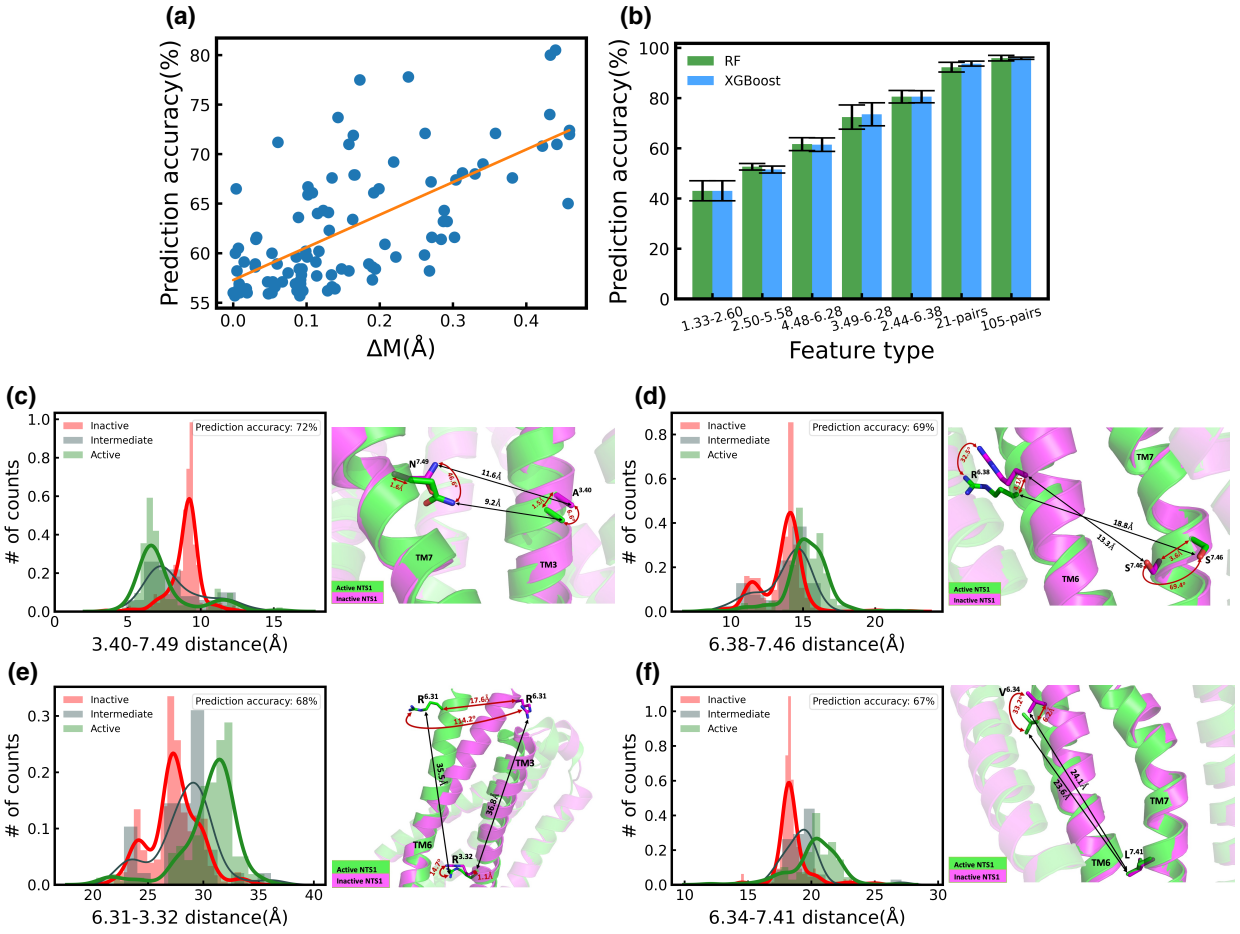


Figure 6: **(a)** Correlation between prediction accuracy and difference of the mean of the closest heavy distances in active and inactive states ($\Delta M$). **(b)** Comparison between prediction accuracy of Ransom Forest and XGBoost methods. The 105-pairs are 105 residue pairs with top accuracies over 420 random residue pairs. The 21-pairs is 21 residue pairs out of 105-pairs with at least one residue belongs to the polar network in GPCR activation. Each data point is taken over 555 proteins. **(c)**,**(d)**,**(e)**,**(f)** Normalized histogram of closest distance of two non-hydrogen atoms of a residue pair over 555 proteins in active, intermediate, and inactive states with corresponding representation of conformational changes of the residue pairs on NTS1 protein. The Ballesteros-Weinstein numbers are used to label the amino acids.

to the polar network provides 68% prediction accuracy. $R^{6.31}$ translates 17.6Å and rotates 114.2° while $R^{3.32}$ has a 14.7° rotation and 1.1Å translation on NTS1 protein (Figure 6e). Residue pair of 6.34-7.41 provides 67% prediction accuracy where 7.41 residue is involved in the polar network. On NTS1 protein, $L^{7.41}$ has no significant conformational changes while $V^{6.34}$ translates and rotates for 6.2Å and 33.2° respectively (Figure 6f).

Comparing the accuracy of XGBoost model for 21-pairs and 105-pairs reveals that higher dimension features improve the prediction accuracy from 93.69% to 95.86% and the regression MAE from 0.0969 to 0.0715 (Table 1, 2). The manually engineered features improve the performance of XGBoost model, making it comparable to the GNN model at the classification task (95.86% vs 95.85%) and better than the GNN model at the regression task (0.0715 vs 0.0897, Table 1, 2). In addition to using ML methods, correlating important residue positions with the GPCR structure can provide insights into the GPCR conformational landscape. By comparing the ranked list of features from our feature engineering method with the important residue interactions in literature, we conclude that the feature engineering method is able to identify the important descriptors of the GPCR states and conformations. The top ranking features which best distinguish between the active and inactive conformation of GPCRs, can be the potential targets of therapeutic molecules to regulate the GPCR structure-activity-function relationship.

# Conclusion

In this work, we have developed 3 ML-based approaches to predict the discrete state and activation level of all the GPCR structures. To learn the structure activity relations in GPCRs, we developed 3D CNN, GNN, and XGBoost models. The 3D CNN approach requires minimal feature engineering and can extract important features from the voxelized representation of the protein structure. However, the 3D CNN approach also has the lowest accuracy of the 3 methods that we have developed (91%). The GNN approach, in comparison to 3D CNN, offers improvement in the prediction accuracy for both the classification (95.85%) and regression task (MAE 0.0897). The GNN approach incorporates the notion of feature engineering by generating a graph representation of the protein structure by encoding the residue type, dihedral angles. In the third approach, we have designed and engineered biophysics-aware features and rank these features. The top features were used to train the XGBoost model

for the classification and regression tasks. The biophysics-aware features perform almost similar to the GNN model for the classification task (95.86%) and outperforms the GNN method for the regression task (7.15%). We then interpret the important features learnt by our models. We use the feature importance to identify the residue pairs which can be used to distinguish between the active, inactive, and intermediate conformation of GPCRs. Finally, we propose a list of residue pairs that can be used to develop a quantitative description of the GPCR states. This work and its conclusions can be extended and applied to understand the transition between the active and inactive conformation of a GPCR protein and design therapeutics to identify the important reside pairs for the transition.

## Acknowledgement

The authors gratefully acknowledge the use of the supercomputing resource Arjuna provided by the Pittsburgh Supercomputing Center (PSC). This work is supported by Center for Machine Learning in Health (CMLH) at Carnegie Mellon University and start-up fund from Mechanical Engineering Department at CMU.

## References

(1) Liebmann, C. Regulation of MAP kinase activity by peptide receptor signalling pathway: paradigms of multiplicity. *Cellular signalling* **2001**, *13*, 777–785.

(2) van Blesen, T.; Hawes, B. E.; Luttrell, D. K.; Krueger, K. M.; Touhara, K.; Porfflri, E.; Sakaue, M.; Luttrell, L. M.; Lefkowitz, R. J. Receptor-tyrosine-kinase-and G$\beta\gamma$-mediated MAP kinase activation by a common signalling pathway. *Nature* **1995**, *376*, 781–784.

(3) Zhang, Q.; Liu, B.; Li, Y.; Yin, L.; Younus, M.; Jiang, X.; Lin, Z.; Sun, X.; Huang, R.; Liu, B., et al. Regulating quantal size of neurotransmitter release through a GPCR

voltage sensor. *Proceedings of the National Academy of Sciences* **2020**, *117*, 26985–26995.

(4) Betke, K. M.; Wells, C. A.; Hamm, H. E. GPCR mediated regulation of synaptic transmission. *Progress in neurobiology* **2012**, *96*, 304–321.

(5) Boules, M.; Li, Z.; Smith, K.; Fredrickson, P.; Richelson, E. Diverse roles of neurotensin agonists in the central nervous system. *Frontiers in endocrinology* **2013**, *4*, 36.

(6) Martinez-Fong, D.; Trédaniel, J.; Forgez, P., et al. Neurotensin and its high affinity receptor 1 as a potential pharmacological target in cancer therapy. *Frontiers in endocrinology* **2013**, *3*, 184.

(7) Schimpff, R.; Avard, C.; Fenelon, G.; Lhiaubet, A.; Tenneze, L.; Vidailhet, M.; Rostene, W. Increased plasma neurotensin concentrations in patients with Parkinson's disease. *Journal of Neurology, Neurosurgery & Psychiatry* **2001**, *70*, 784–786.

(8) Hauser, A. S.; Attwood, M. M.; Rask-Andersen, M.; Schiöth, H. B.; Gloriam, D. E. Trends in GPCR drug discovery: new agents, targets and indications. *Nature reviews Drug discovery* **2017**, *16*, 829–842.

(9) Congreve, M.; de Graaf, C.; Swain, N. A.; Tate, C. G. Impact of GPCR structures on drug discovery. *Cell* **2020**, *181*, 81–91.

(10) Hauser, A. S.; Chavali, S.; Masuho, I.; Jahn, L. J.; Martemyanov, K. A.; Gloriam, D. E.; Babu, M. M. Pharmacogenomics of GPCR drug targets. *Cell* **2018**, *172*, 41–54.

(11) Basith, S.; Cui, M.; Macalino, S. J.; Park, J.; Clavio, N. A.; Kang, S.; Choi, S. Exploring G protein-coupled receptors (GPCRs) ligand space via cheminformatics approaches: impact on rational drug design. *Frontiers in pharmacology* **2018**, *9*, 128.

(12) Trzaskowski, B.; Latek, D.; Yuan, S.; Ghoshdastider, U.; Debinski, A.; Filipek, S.

Action of molecular switches in GPCRs-theoretical and experimental studies. *Current medicinal chemistry* **2012**, *19*, 1090–1109.

(13) Deupi, X.; Kobilka, B. K. Energy landscapes as a tool to integrate GPCR structure, dynamics, and function. *Physiology* **2010**, *25*, 293–303.

(14) Latorraca, N. R.; Venkatakrishnan, A.; Dror, R. O. GPCR dynamics: structures in motion. *Chemical reviews* **2017**, *117*, 139–155.

(15) Kooistra, A. J.; Mordalski, S.; Pándy-Szekeres, G.; Esguerra, M.; Mamyrbekov, A.; Munk, C.; Keserű, G. M.; Gloriam, D. E. GPCRdb in 2021: integrating GPCR sequence, structure and function. *Nucleic Acids Research* **2021**, *49*, D335–D343.

(16) Mattedi, G.; Deflorian, F.; Mason, J. S.; de Graaf, C.; Gervasio, F. L. Understanding ligand binding selectivity in a prototypical GPCR family. *Journal of chemical information and modeling* **2019**, *59*, 2830–2836.

(17) Hilger, D.; Masureel, M.; Kobilka, B. K. Structure and dynamics of GPCR signaling complexes. *Nature structural & molecular biology* **2018**, *25*, 4–12.

(18) Kohlhoff, K. J.; Shukla, D.; Lawrenz, M.; Bowman, G. R.; Konerding, D. E.; Belov, D.; Altman, R. B.; Pande, V. S. Cloud-based simulations on Google Exacycle reveal ligand modulation of GPCR activation pathways. *Nature chemistry* **2014**, *6*, 15–21.

(19) Feinberg, E. N.; Farimani, A. B.; Hernandez, C. X.; Pande, V. S. Kinetic Machine Learning Unravels Ligand-Directed Conformational Change of $\mu$ Opioid Receptor. *bioRxiv* **2017**, 170886.

(20) Feinberg, E. N.; Farimani, A. B.; Uprety, R.; Hunkele, A.; Pasternak, G. W.; Majumdar, S.; Pande, V. S. Machine Learning Harnesses Molecular Dynamics to Discover New $\mu$ Opioid Chemotypes. *arXiv preprint arXiv:1803.04479* **2018**,

(21) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A., et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589.

(22) Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Schaeffer, R. D., et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **2021**, *373*, 871–876.

(23) Townshend, R.; Bedi, R.; Suriana, P.; Dror, R. End-to-end learning on 3d protein structure for interface prediction. *Advances in Neural Information Processing Systems* **2019**, *32*, 15642–15651.

(24) Ragoza, M.; Hochuli, J.; Idrobo, E.; Sunseri, J.; Koes, D. R. Protein–ligand scoring with convolutional neural networks. *Journal of chemical information and modeling* **2017**, *57*, 942–957.

(25) Jiménez, J.; Doerr, S.; Martínez-Rosell, G.; Rose, A. S.; De Fabritiis, G. DeepSite: protein-binding site predictor using 3D-convolutional neural networks. *Bioinformatics* **2017**, *33*, 3036–3042.

(26) Derevyanko, G.; Grudinin, S.; Bengio, Y.; Lamoureux, G. Deep convolutional networks for quality assessment of protein folds. *Bioinformatics* **2018**, *34*, 4046–4053.

(27) Luo, F.; Wang, M.; Liu, Y.; Zhao, X.-M.; Li, A. DeepPhos: prediction of protein phosphorylation sites with deep learning. *Bioinformatics* **2019**, *35*, 2766–2773.

(28) de Jesus, D. R.; Cuevas, J.; Rivera, W.; Crivelli, S. Capsule networks for protein structure classification and prediction. *arXiv preprint arXiv:1808.07475* **2018**,

(29) Kipf, T. N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. International Conference on Learning Representations (ICLR). 2017.

(30) Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; Bengio, Y. Graph Attention Networks. International Conference on Learning Representations. 2018.

(31) Xu, K.; Hu, W.; Leskovec, J.; Jegelka, S. How Powerful are Graph Neural Networks? International Conference on Learning Representations. 2019.

(32) Wang, Y.; Wang, J.; Cao, Z.; Farimani, A. B. MolCLR: Molecular contrastive learning of representations via graph neural networks. *arXiv preprint arXiv:2102.10056* **2021**,

(33) Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; Philip, S. Y. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems* **2020**, *32*, 4–24.

(34) Fout, A.; Byrd, J.; Shariat, B.; Ben-Hur, A. Protein Interface Prediction using Graph Convolutional Networks. Advances in Neural Information Processing Systems. 2017.

(35) Stepniewska-Dziubinska, M. M.; Zielenkiewicz, P.; Siedlecki, P. Development and evaluation of a deep learning model for protein–ligand binding affinity prediction. *Bioinformatics* **2018**, *34*, 3666–3674.

(36) Son, J.; Kim, D. Development of a graph convolutional neural network model for efficient prediction of protein-ligand binding affinities. *PloS one* **2021**, *16*, e0249404.

(37) Gainza, P.; Sverrisson, F.; Monti, F.; Rodola, E.; Boscaini, D.; Bronstein, M.; Correia, B. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nature Methods* **2020**, *17*, 184–192.

(38) Sverrisson, F.; Feydy, J.; Correia, B. E.; Bronstein, M. M. Fast end-to-end learning on protein surfaces. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021; pp 15272–15281.

(39) Gligorijević, V.; Renfrew, P. D.; Kosciolek, T.; Leman, J. K.; Berenberg, D.; Vatanen, T.; Chandler, C.; Taylor, B. C.; Fisk, I. M.; Vlamakis, H., et al. Structure-based

protein function prediction using graph convolutional networks. *Nature communications* **2021**, *12*, 1–14.

(40) Tsou, L. K.; Yeh, S.-H.; Ueng, S.-H.; Chang, C.-P.; Song, J.-S.; Wu, M.-H.; Chang, H.-F.; Chen, S.-R.; Shih, C.; Chen, C.-T., et al. Comparative study between deep learning and QSAR classifications for TNBC inhibitors and novel GPCR agonist discovery. *Scientific reports* **2020**, *10*, 1–11.

(41) Jabeen, A.; Ranganathan, S. Applications of machine learning in GPCR bioactive ligand discovery. *Current opinion in structural biology* **2019**, *55*, 66–76.

(42) Vignir Isberg, G. G. M., Christian Munk GPCRdb Documentaion. **2021**,

(43) Rose, P. W.; Prlić, A.; Altunkaya, A.; Bi, C.; Bradley, A. R.; Christie, C. H.; Costanzo, L. D.; Duarte, J. M.; Dutta, S.; Feng, Z., et al. The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic acids research* **2016**, gkw1000.

(44) Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016; pp 785–794.

(45) Breiman, L. Random forests. *Machine learning* **2001**, *45*, 5–32.

(46) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V., et al. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* **2011**, *12*, 2825–2830.

(47) Cohen-Or, D.; Kaufman, A. Fundamentals of surface voxelization. *Graphical models and image processing* **1995**, *57*, 453–461.

(48) Zhang, Z.; Jaiswal, P.; Rai, R. Featurenet: Machining feature recognition based on 3d convolution neural network. *Computer-Aided Design* **2018**, *101*, 12–22.

(49) Su, Z.; Tan, P. S.; Chow, J.; Wu, J.; Cheong, Y.; Wang, Y.-H. DV-ConvNet: Fully Convolutional Deep Learning on Point Clouds with Dynamic Voxelization and 3D Group Convolution. *arXiv preprint arXiv:2009.02918* **2020**,

(50) O'Mahony, N.; Campbell, S.; Krpalkova, L.; Carvalho, A.; Velasco-Hernández, G. A.; Riordan, D.; Walsh, J. Convolutional Neural Networks for 3D Vision System Data: A review. 2018 12th International Conference on Sensing Technology (ICST). 2018; pp 160–165.

(51) He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016; pp 770–778.

(52) Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* **2014**,

(53) Krizhevsky, A.; Sutskever, I.; Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* **2012**, *25*, 1097–1105.

(54) Wang, Y.; Cao, Z.; Farimani, A. B. Efficient water desalination with graphene nanopores obtained using artificial intelligence. *npj 2D Materials and Applications* **2021**, *5*, 1–9.

(55) Wallach, I.; Dzamba, M.; Heifets, A. AtomNet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery. *arXiv preprint arXiv:1510.02855* **2015**,

(56) Barry, M. C.; Wise, K. E.; Kalidindi, S. R.; Kumar, S. Voxelized Atomic Structure Potentials: Predicting Atomic Forces with the Accuracy of Quantum Mechanics Using Convolutional Neural Networks. *The Journal of Physical Chemistry Letters* **2020**, *11*, 9093–9099.

(57) Singh, R.; Sharma, A.; Bingol, O. R.; Balu, A.; Balasubramanian, G.; Johnson, D. D.; Sarkar, S. 3D Deep Learning with voxelized atomic configurations for modeling atomistic potentials in complex solid-solution alloys. *arXiv preprint arXiv:1811.09724* **2018**,

(58) Hassan-Harrirou, H.; Zhang, C.; Lemmin, T. RosENet: improving binding affinity prediction by leveraging molecular mechanics energies with an ensemble of 3D convolutional neural networks. *Journal of chemical information and modeling* **2020**, *60*, 2791–2802.

(59) Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. International conference on machine learning. 2015; pp 448–456.

(60) Maas, A. L.; Hannun, A. Y.; Ng, A. Y., et al. Rectifier nonlinearities improve neural network acoustic models. Proc. icml. 2013; p 3.

(61) Mikołajczyk, A.; Grochowski, M. Data augmentation for improving deep learning in image classification problem. 2018 international interdisciplinary PhD workshop (IIPhDW). 2018; pp 117–122.

(62) Shorten, C.; Khoshgoftaar, T. M. A survey on image data augmentation for deep learning. *Journal of Big Data* **2019**, *6*, 1–48.

(63) Gori, M.; Monfardini, G.; Scarselli, F. A new model for learning in graph domains. Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005. 2005; pp 729–734.

(64) Scarselli, F.; Gori, M.; Tsoi, A. C.; Hagenbuchner, M.; Monfardini, G. The graph neural network model. *IEEE transactions on neural networks* **2008**, *20*, 61–80.

(65) Bruna, J.; Zaremba, W.; Szlam, A.; Lecun, Y. Spectral networks and locally connected

networks on graphs. International Conference on Learning Representations (ICLR2014), CBLS, April 2014. 2014.

(66) Henaff, M.; Bruna, J.; LeCun, Y. Deep convolutional networks on graph-structured data. *arXiv preprint arXiv:1506.05163* **2015**,

(67) Defferrard, M.; Bresson, X.; Vandergheynst, P. Convolutional neural networks on graphs with fast localized spectral filtering. *arXiv preprint arXiv:1606.09375* **2016**,

(68) Xu, K.; Hu, W.; Leskovec, J.; Jegelka, S. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826* **2018**,

(69) Hu, W.; Liu, B.; Gomes, J.; Zitnik, M.; Liang, P.; Pande, V.; Leskovec, J. Strategies for pre-training graph neural networks. *arXiv preprint arXiv:1905.12265* **2019**,

(70) Zhou, J.; Cui, G.; Hu, S.; Zhang, Z.; Yang, C.; Liu, Z.; Wang, L.; Li, C.; Sun, M. Graph neural networks: A review of methods and applications. *AI Open* **2020**, *1*, 57–81.

(71) Bengio, Y. *Neural networks: Tricks of the trade*; Springer, 2012; pp 437–478.

(72) Egloff, P.; Hillenbrand, M.; Klenk, C.; Batyuk, A.; Heine, P.; Balada, S.; Schlinkmann, K. M.; Scott, D. J.; Schütz, M.; Plückthun, A. Structure of signaling-competent neurotensin receptor 1 obtained by directed evolution in Escherichia coli. *Proceedings of the National Academy of Sciences* **2014**, *111*, E655–E662.

(73) Huang, W.; Masureel, M.; Qu, Q.; Janetzko, J.; Inoue, A.; Kato, H. E.; Robertson, M. J.; Nguyen, K. C.; Glenn, J. S.; Skiniotis, G., et al. Structure of the neurotensin receptor 1 in complex with $\beta$-arrestin 1. *Nature* **2020**, *579*, 303–308.

(74) Wu, F.; Yang, L.; Hang, K.; Laursen, M.; Wu, L.; Han, G. W.; Ren, Q.; Roed, N. K.; Lin, G.; Hanson, M. A., et al. Full-length human GLP-1 receptor structure without orthosteric ligands. *Nature communications* **2020**, *11*, 1–10.

(75) Zhang, Y.; Sun, B.; Feng, D.; Hu, H.; Chu, M.; Qu, Q.; Tarrasch, J. T.; Li, S.; Sun Kobilka, T.; Kobilka, B. K., et al. Cryo-EM structure of the activated GLP-1 receptor in complex with a G protein. *Nature* **2017**, *546*, 248–253.

(76) Maaten, L. v. d.; Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research* **2008**, *9*, 2579–2605.

(77) Li, W.; Cerise, J. E.; Yang, Y.; Han, H. Application of t-SNE to human genetic data. *Journal of bioinformatics and computational biology* **2017**, *15*, 1750017.

(78) Kobak, D.; Berens, P. The art of using t-SNE for single-cell transcriptomics. *Nature communications* **2019**, *10*, 1–14.

(79) Linderman, G. C.; Rachh, M.; Hoskins, J. G.; Steinerberger, S.; Kluger, Y. Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. *Nature methods* **2019**, *16*, 243–245.

(80) Swaminath, G.; Xiang, Y.; Lee, T. W.; Steenhuis, J.; Parnot, C.; Kobilka, B. K. Sequential binding of agonists to the $\beta$2 adrenoceptor: kinetic evidence for intermediate conformational states. *Journal of Biological Chemistry* **2004**, *279*, 686–691.

(81) Hauser, A. S.; Kooistra, A. J.; Munk, C.; Heydenreich, F. M.; Veprintsev, D. B.; Bouvier, M.; Babu, M. M.; Gloriam, D. E. GPCR activation mechanisms across classes and macro/microscales. *Nature structural & molecular biology* **2021**, *28*, 879–888.

(82) Huang W, V. A., Manglik A Structural insights into $\mu$-opioid receptor activation. *Nature* **2015**, *524*, 315–321.

(83) Schrödinger, LLC, The PyMOL Molecular Graphics System, Version 1.8. **2015**,