SSnet - Secondary Structure based End-to-End Learning model for Protein-Ligand Interaction Prediction

Niraj Verma,[†,§] Xingming Qu,[‡,§] Francesco Trozzi,[†,§] Mohamed Elsaied,[¶,§] Yunwen Tao,[†,§] Eric C. Larson,[‡,§] and Elfi Kraka[*,†,§]

†*Department of Chemistry, Southern Methodist University, Dallas TX USA*

‡*Department of Computer Science, Southern Methodist University, Dallas TX USA*

¶*Department of Engineering Management and Information System, Southern Methodist University, Dallas TX USA*

§*Southern Methodist University*

E-mail: ekraka@gmail.com

Abstract

Computational prediction of bioactivity has become a critical aspect of modern drug discovery as it mitigates the cost, time, and resources required to find and screen new compounds. Deep Neural Networks (DNN) have recently shown excellent performance in modeling Protein-Ligand Interaction (PLI). However, DNNs are only effective when physically sound descriptions of ligands and proteins are fed into the neural network for processing. Furthermore, previous research has not incorporated the secondary structure of proteins in a meaningful manner. In this work, we developed a DNN framework that utilizes the secondary structure information of proteins which is extracted as the curvature and torsion of the protein backbone to predict PLI. We demonstrate how our

model outperforms previous machine and non-machine learning models on four major datasets: humans, C.*elegans*, DUD-E, and BindingDB. Visualization of the intermediate layers of our model shows a potential latent space for proteins which extracts important information about the bioactivity. We further investigate the inner workings of our model by visualizing the most important aspects in a protein that the model finds influential. We observed that our model learns important information about possible locations where a ligand would bind including binding sites, allosteric sites and cryptic sites, regardless of the conformation used. This work opens the door to the exploration of secondary structure based deep learning in general, which is not just confined to protein-ligand interactions.

Introduction

The interaction between proteins and small molecules is a complex mechanism which biological systems adopt to regulate many fundamental operations. These interactions are governed by a multitude of factors including hydrogen bonding[1,2], π-interactions[3], hydrophobicity[4] etc. The experimental validation of Protein-Ligand Interaction (PLI) is the state-of-the-art method, however, it is time-consuming and expensive. Computational methods can significantly boost time and save resources, however, due to the complex nature of PLI its prediction is a challenging computational enterprise. However, reliable PLI predictions could significantly reduce the discovery time for new treatments, eliminate toxic drug candidates and efficiently guide medicinal chemistry efforts[5]. It is therefore imperative to develop computational methods to predict PLI despite the complexity of the problem.

Traditional PLI relies on high-throughput screening which is an experimental technique with high cost and low efficiency[6]. Virtual screening (VS) accelerates the PLI process while greatly reducing time and resources. Broadly, VS can be divided in two major categories: Ligand Based Virtual Screening (LBVS) and Structure Based Virtual Screening (SBVS)[7]. LBVS applies known sets of ligands to a target of interest and therefore, its capability

to find novel chemotypes is limited. SBVS uses the 3D structure of a given target and therefore is a better choice for the discovery of novel active compounds[8]. However, SBVS has a somewhat poor performance, sometimes not being able to distinguish active from inactive compounds[9]. Over the last few decades, many classical techniques such as force field, empirical and knowledge based[5] PLI predictions have been applied, however, often showing low performance and in some cases even discrepancies when compared with experimental bioactivities[10].

Machine learning (ML) and Deep learning (DL) approaches have recently received attention in this field. Various reviews summarize the application of ML/DL in drug design and discovery[11–15]. Machine learning based PLI prediction has been developed from a chemogenomics perspective[16] that considers interactions in a unified framework from chemical space and genomic space. Jacob and Vert[17] used tensor-product based features and applied Support Vector Machines (SVM) to predict PLI. Yamanishi et al.[18] minimized Euclidean distances over common features derived by mapping ligands and proteins. Wallach et al.[19] used a 3D grid for proteins along with 3D convolutional networks. Tsubaki et al.[20] used a combination of convolutional network for proteins and graph network for ligands. Li et al.[21] used Bayesian additive regression trees to predict PLI. Lee et al.[22] applied deep learning with convolution on protein sequences.

Most of the protein structure based models (ML/DL) for PLI predictions achieve low accuracy as i) high resolution protein-ligand pair for training is mostly absent, ii) the 3D grid for the target is large and sparse matrix, which hinder ML/DL models to learn and predict PLI. In this work, we propose a secondary structure based representation for proteins with a 1D vector contrary to 3D[19] or 2D[23] based representations. The 1D representation is based on the curvature and torsion of protein backbone and therefore does not necessarily require high resolution geometry information. Mathematically, curvature and torsion are sufficient to reasonably represent the 3D structure of a protein[24] and therefore contains compact information about it's function.

Tremendous work has been carried out to represent the ligands. Gómez-Bombarelli et al.[25] created a model to generate Continuous Latent Space (CLP) from sparse Simplified Molecular-Input Line-Entry System (SMILES) strings (i.e., a string representation of a molecule) based on a variational autoencoder similar to word embedding[26]. Scarselli et al.[27] proposed a Graph Neural Network (GNN) to describe molecules. Rogers and Hahn[28] developed Extended-Connectivity Fingerprints (ECF), which includes the presence of substructures (and therefore also includes stereochemical information) to represent molecules. Riniker and Landrum[29] proposed a fingerprint based on substructure and their similarity (Avalon).

End-to-end learning, a powerful ML/DL technique to exploit drug discovery and development, has gained interest in recent years[30]. The end-to-end learning technique involves i) embedding inputs to lower dimensions, ii) formulating various neural networks depending on the data available, and iii) using backpropagation over the whole architecture to minimize loss and update weights. In this work, we utilize such an architecture where the proteins and the ligands are transformed into lower dimensions with the help of convolutional and fully connected dense networks. We coin this neural network based end-to-end learning model SSnet. A general overview of the SSnet model is shown as Figure 2 of the supporting information.

The goal of this work is to predict binary PLI (1 is active, 0 is inactive). Any other analysis carried out is to show the robustness and versatility of the model. This limits the scope of this research to perform indepth analysis on various demeanor that can be carried out utilizing secondary structure information such as locating binding site, prediction of selectivity among ligands, pose prediction etc.

We analyzed the SSnet model by utilizing the Grad-CAM method[31]. Grad-CAM is a useful technique to visualize heatmaps of the activation from networks that maximally excite the input feature. In other words, it shows the important data points in the input feature that are responsible for the prediction. We show that our model learns the important

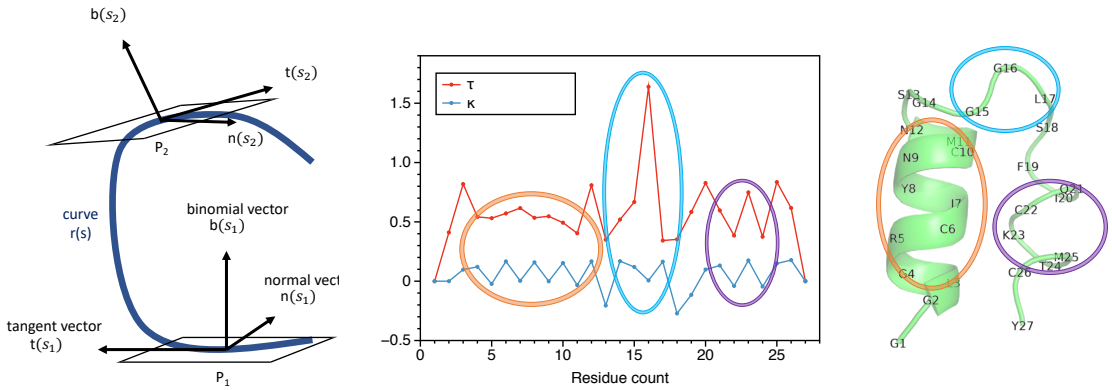
residues in the protein which the machine finds most important for PLI prediction. With the heatmap information we tend to answer why curvature and torsion is an important tool for representation of proteins for ML/DL models. This analysis also proves that SSnet is not learning irrelevant information and therefore can generalize for unknown proteins and ligands.

Often in virtual screening, only few conformations for an unknown protein are available that requires immediate attention (eg. in epidemics such as SARS-CoV-2[32], Marburg Virus[33] etc.). We tested several conformations of a protein and observed that SSnet can predict similar results from apo protein (protein without ligand) or from a different conformation of the protein (protein with a different ligand). We further observed that SSnet can also predict PLI for cryptic proteins (proteins that undergo a significant conformational change upon ligand binding which results in the opening of cryptic binding sites). SSnet thus shows a versatile nature in predicting PLI and therefore is not confined to require a unique conformation of protein-ligand complex.

In the following we first demonstrate how the secondary structure of proteins can be used in ML/DL. Then we discuss the representation of ligands following the introduction of SSnet model, possible evaluation criteria, its merits and demerits. Then we explain the datasets used in this work and analyze the performance and outcomes of SSnet in the Results and Discussion section. We then unbox the SSnet model by visualizing heatmaps of the proteins. Finally, we summarize and conclude our work and provide a future perspective.

Representation of Proteins

Protein structures exhibit a large conformational variety which has traditionally been modeled through complex combinations of primary, secondary, tertiary, and quaternary levels. Many automated and manual sorting databases like SCOP[34], CATH[35], DALI[36], and programs like DSSP[37] STRIDE[38], DEFINE[39], and KAKSI[40] etc. have provided protein classifications based on the secondary structure. However, these classifications are often conflict-



(a) (b)

Figure 1: a) The tangent vector **t**, normal vector **n** and the binormal vector **b** of a Frenet frame at points P_1 and P_2 respectively for a curve $\mathbf{r}(s)$. b) Representation of protein backbone in terms of scalar curvature κ and torsion τ respectively. The ideal helix, turn and non-ideal helix is shown in orange, cyan and magenta respectively. The curvature and torsion pattern captures the secondary structure of the protein.

ing[41]. A more promising approach to determine protein fold based on secondary structure has been introduced by Ranganathan et al.[24] and Guo and Cremer[42]; based on Frenet-Serret frame and coordinates with the aim to classify secondary structure elements in proteins they coined the Automated Protein Structure Analysis (APSA). The original idea behind APSA is based on Unified Reaction Valley Approach (URVA) developed by Cremer and Kraka[43] where a reaction path is characterized by arc length, curvature, and torsion. Inspired by the representation of features in APSA, we tend to answer what, if, and how can we utilize the secondary structure in ML/DL approaches for Protein Ligand Interaction (PLI) prediction.

In this approach a protein is represented by the α carbons (CA atoms) of the backbone as it defines a unique and recognizable structure, especially for protein categorization[44]. In fact, a significant amount of information about the protein is embedded in the secondary structure elements such as helices, β sheets, hairpins, coils, and turns etc. Therefore, it is imperative to incorporate the secondary structure of the protein when representing its features for machine learning approaches. Otherwise, the algorithms will be blind to interactions dependent on the secondary structure.

The secondary structure information can be retrieved by a smooth curve generated by a cubic spline fit of the CA atoms. Figure 1a shows the arc length s, scalar curvature κ and scalar torsion τ which define the 3D curve $\mathbf{r}(s)$. The scalar curvature κ is expressed as a function of arc length s

$$\kappa(s) = |\mathbf{r}''(s)| \tag{1}$$

and the scalar torsion

$$\tau(s) = \frac{\langle \mathbf{r}'(s), \mathbf{r}''(s), \mathbf{r}'''(s) \rangle}{|\mathbf{r}''(s))|^2} \tag{2}$$

where $|\cdot|$ is the norm and $\langle \cdot \rangle$ is the vector triple product. A protein can then be represented by considering the curvature and torsion at the anchor points (locations of CA) forming a 1D vector with twice the the number of amino acids.

Figure 1b shows the decomposition of a protein found in Conus villepinii (PDB ID - 6EFE) into scalar curvature κ and torsion τ respectively. The residues 5 through 10 show a near ideal α helix type secondary structure, which is represented as an oscillation of τ with smooth κ. Similarly, the turn (residues 15 to 17) and a non-ideal α helix (residues 20 to 25) are captured in the decomposition plot via unique patterns. Because the curvature and torsion information of the secondary structure of proteins are encoded as patterns, machine learning techniques may be powerful tools to predict PLI through efficiently learned representations of these patterns. More specifically, we hypothesized that, using convolution, varying sized filters may be excellent pattern matching methods for discerning structure from these decomposition plots. More analysis on protein representation is provided in the section SSnet.

Representation of Ligands

A molecule can be represented by the SMILES string, which represents its various bonds and orientations. However, the SMILES string is sparse and does not necessarily provide information about the ligand structure in an efficient way. Therefore, SMILES strings are

difficult for machine learning algorithms to effectively learn from. A number of alternative representations for ligands have been proposed that model varying aspects of the ligand in a more machine readable format. The hope has been that machine learning algorithms can more effectively use these representations for prediction. Since ligand representation is an ongoing research topic, we consider four different methods: CLP[25], GNN[27], Avalon[29], and ECF[28]. CLP was generated by the code provided by Gómez-Bombarelli et al.[25]; Avalon and ECF were generated from RDKit[45]; and GNN was implemented as proposed by Tsubaki et al.[20].

SSnet model

Figure 2 shows the SSnet model developed in this work. We provide here a general overview of the network and more details about its specific design operation is given in the later part of this section. As denoted in the left upper branch of Figure 2 after conversion into the Frenet-Serret frame and the calculation of curvature κ and torsion τ, κ and τ data (i.e decomposition data) is fed into the neural network. We denote this input as a 2D matrix (1D vector with curvature and torsion reshaped to contain curvature in one row and torsion in the other), $\mathbf{X}^{(0)}$, where each column represents a unique residue and the rows corresponding the curvature and torsion. The first layer is a branch convolution with varying window sizes. That is, each branch is a convolution with a filter of differing length. We perform this operation so that patterns of varying lengths in the decomposition plot can be recognized by the neural network. Each branch is then fed to more convolutions of same window size. This allows the network to recognize more intricate patterns in $\mathbf{X}^{(0)}$ that might be more difficult to recognize with a single convolution. The output of these convolutional branches are concatenated, pooled over the length of the sequence, and fed to a fully connected dense layer. The rightmost upper branch of Figure 2 shows a ligand vector which is generated and fed to a fully connected dense layer. The output of this layer is typically referred to as an embedding. Intuitively, this embedding is a reduced dimensionality representation of the

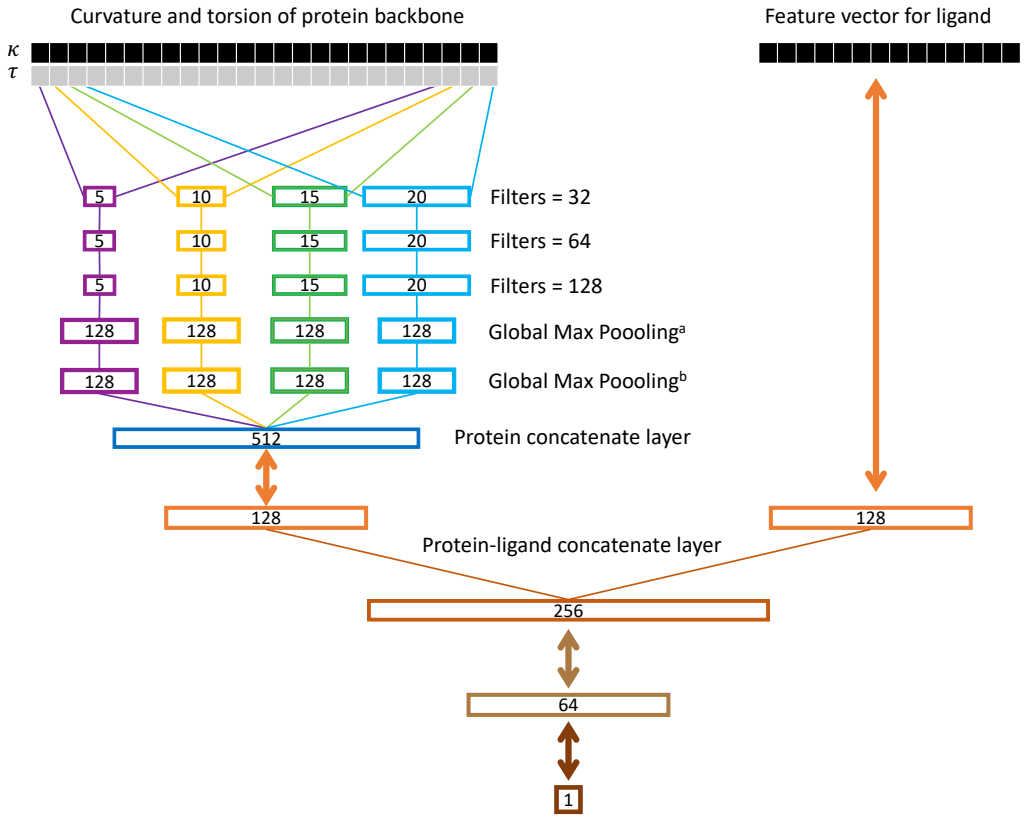


Figure 2: SSnet model. The curvature and torsion pattern of a protein backbone is fed through multiple convolution networks with varying window sizes as branch convolution. Each branch further goes through more convolution with same window size (red, orange, green and light blue boxes). A global max pooling layer is implemented to get the protein vector. The ligand vector is directly fed to the network. Each double array line implies a fully connected dense layer. The number inside a box represents the dimension of the corresponding vector. In the case of GNN, the ligand vector is replaced by a graph neural network as implemented by Tsubaki et al.[20]

protein and ligand. The outputs of the protein embedding and the ligand embedding are then concatenated and fed to further dense layers to predict the PLI.

The convolutional network in this research uses filter functions over the protein vector $\mathbf{X}^{(0)}$. To define the convolution operation more intuitively, we define a reshaping operation as follows:

$$\mathbf{c}_i^{(0)} = \text{flat}\left(\mathbf{X}^{(0)}_{\text{row}=i:i+K,\forall\text{col}}\right)$$

where the flattening operation reshapes the row of $\mathbf{X}^{(0)}$ from indices i to $i+K$ to be a column vector $\mathbf{c}_i^{(0)}$. This process is also referred to as vectorization. The size of the filter will then be of length K. We define the convolution operation as:

$$\mathbf{X}^{(1)}_{\text{row}=i,\forall\text{col}} = f(\mathbf{W}^{(0)}_{\text{conv}}\mathbf{c}_i^{(0)} + \mathbf{b}^{(0)}_{\text{conv}}) \tag{3}$$

where f is a function known as the rectified linear unit (ReLU), $\mathbf{W}^{(0)}_{\text{conv}}$ is the weight matrix and $\mathbf{b}^{(0)}_{\text{conv}}$ is the bias vector. This operation fills in the columns of the output of the convolution, $\mathbf{X}^{(1)}_{\text{row}=i,\forall\text{col}}$ (also called the activation or feature map). Each row of $\mathbf{W}^{(0)}_{\text{conv}}$ is considered as a different filter and each row of $\mathbf{X}^{(1)}$ is the convolutional output of each of these filters. These convolutions can be repeated such that the nth activation is computed as:

$$\begin{aligned}\mathbf{c}_i^{(n)} =& \text{flat}\left(\mathbf{X}^{(n)}_{\text{row}=i:i+K,\forall\text{col}}\right)\\ \mathbf{X}^{(n)}_{\text{row}=i,\forall\text{col}} =& f(\mathbf{W}^{(n-1)}_{\text{conv}}\mathbf{c}_i^{(n-1)} + \mathbf{b}^{(n-1)}_{\text{conv}})\end{aligned} \tag{4}$$

We in our SSnet model use four different branches with filter sizes of $\kappa = 5, 10, 15$ and 30. The final convolutional activations for layer N can be referred to as $\mathbf{X}_\kappa^{(N)}$ where κ denotes the branch. The activation $\mathbf{X}_\kappa^{(N)}$ is often referred to as the latent space because it denotes the latent features of the input sequence. The number of columns in $\mathbf{X}_\kappa^{(N)}$ is dependent upon the size of the input sequence. To collapse this unknown size matrix into a fixed size vector, we apply a maximum operation along the rows of $\mathbf{X}_\kappa^{(N)}$. This is typically referred to as a Global Max Pooling layer in neural networks and is repeated R times for each row in $\mathbf{X}_\kappa^{(N)}$:

$$\mathbf{d}_\kappa = \begin{bmatrix} \max\left(\mathbf{X}^{(N)}_{\kappa,\text{row}=1,\forall\text{col}}\right) \\ \max\left(\mathbf{X}^{(N)}_{\kappa,\text{row}=2,\forall\text{col}}\right) \\ \dots \\ \max\left(\mathbf{X}^{(N)}_{\kappa,\text{row}=R,\forall\text{col}}\right) \end{bmatrix} \tag{5}$$

where $\mathbf{d}_\kappa$ is a length R column vector regardless of the number of columns in the latent space $\mathbf{X}^{(N)}_\kappa$. This maximum operation, while important, has the effect of eliminating much of the information in the latent space. To better understand the latent space, we can further process $\mathbf{X}^{(N)}_\kappa$ to understand how samples are distributed. For example, a simple operation would be to define another column vector $\mathbf{v}$ that denotes the total variation in each row of the latent space:

$$\mathbf{v}_\kappa = \begin{bmatrix} \max\left(\mathbf{X}^{(N)}_{\kappa,\text{row}=1,\forall\text{col}}\right) - \min\left(\mathbf{X}^{(N)}_{\kappa,\text{row}=1,\forall\text{col}}\right) \\ \max\left(\mathbf{X}^{(N)}_{\kappa,\text{row}=2,\forall\text{col}}\right) - \min\left(\mathbf{X}^{(N)}_{\kappa,\text{row}=2,\forall\text{col}}\right) \\ \dots \\ \max\left(\mathbf{X}^{(N)}_{\kappa,\text{row}=R,\forall\text{col}}\right) - \min\left(\mathbf{X}^{(N)}_{\kappa,\text{row}=R,\forall\text{col}}\right) \end{bmatrix} \tag{6}$$

The concatenation of vectors $\mathbf{d}$ and $\mathbf{v}$ help elucidate how the samples are distributed in the latent space. As such, we can use these concatenated vectors as inputs to a fully connected dense layer which can learn to interpret the latent space. This output is referred to as the embedding of the protein, $\mathbf{y}_{\text{prot}}$, and is computed as

$$\mathbf{y}_{\text{prot}} = f(\mathbf{W}_{\text{prot}} \cdot [\mathbf{d}_5^T, \mathbf{v}_5^T, \mathbf{d}_{10}^T, \mathbf{v}_{10}^T, \mathbf{d}_{15}^T, \mathbf{v}_{15}^T, \mathbf{d}_{20}^T, \mathbf{v}_{20}^T]^T + \mathbf{b}_{\text{prot}}) \tag{7}$$

where $\mathbf{W}_{\text{prot}}$ is the learned weight matrix and $\mathbf{b}_{\text{prot}}$ is the bias vector of a fully connected network.

The method described above is similar to a technique recently used in speech verification systems where the window sizes need to be dynamic because the length of audio snippet is unknown. In speech systems, the latent space is collapsed via mean and standard de-

viation operations and the embeddings provided for these operations are typically referred to as D-Vectors[46] or X-Vectors[47]. In proteins we have a similar problem as the length of the decomposition sequence to consider the active site(s) of protein is dynamic and are of unknown sizes. By including the window sizes of 5, 10, 15 and 20 (number of residues to consider at a time), we ensure that the network is able to extract different sized patterns from backbones of varying length.

After embedding the protein and the ligand, we concatenate the vectors together and feed them into the final neural network branch, resulting in a prediction of binding, $\hat{y}$, which is expected to be closer to "0" for ligands and proteins that do not bind and closer to "1" for proteins and ligands that do bind. This final branch consists of two layers:

$$\hat{y} = \sigma\left(\mathbf{W}_2 \cdot f\left(\mathbf{W}_1 \cdot [\mathbf{y}^T_{\text{prot}}, \mathbf{y}^T_{\text{ligand}},]^T + \mathbf{b}_1\right) + \mathbf{b}_2\right) \tag{8}$$

where σ refers to a sigmoid function that maps the output to $[0, 1]$. If we denote the ground truth binding as y, which is either 0 or 1, and denote all the parameters inside the network as $\mathbf{W}$ then the loss function for the SSnet model can be defined as binary cross entropy, which is computed as:

$$l(\mathbf{W}) = -\frac{1}{M}\sum_{i}^{M}\left[y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i)\right] \tag{9}$$

where M is the number of samples in the dataset. By optimizing this loss function the neural network can learn to extract meaningful features from the protein and ligand features that relate to binding. At first all weights are initialized randomly and we use back propagation to update the parameters and minimize loss. All operations defined are differentiable, including the collapse of the latent space with Global Max Pooling such that errors in the loss function can back propagate through the network to update all parameters, including the convolutional operations.

The hyperparameters optimized for the model and speed of execution are provided in

section 1 of the supporting information.

Grad-CAM method for heatmap generation

A neural network generally exhibits a large number of weights to be optimized so that complex information can be learned, however, some of this information could be irrelevant to a prediction task. For example, consider the task of identifying if a certain image contains a horse or not. If all horse images also contain a date information on the image and images without horse does not contains date information, the machine will quickly learn to detect the date rather than the goal object (a horse in this case). Therefore, it is essential to verify what a neural network considers "influential" for classification after training. Selvaraju et al.[31] proposed a Gradient-weighted Class Activation (Grad-CAM) based method to generate a heatmap which shows important points in the feature data, based on a particular class of prediction. That is, this method uses activations inside the neural network to understand what portions of an image are most influential for a given classification. In the context of protein structures, this methods can help to elucidate which portions of the decomposition plot are most important for a given classification. These influential patterns in the decomposition plot can then be mapped to specific sub-structures in the protein.

Grad-CAM is computed by taking the gradient weight α_k for all channels in a convolutional layer as

$$\alpha_k = \frac{1}{Z} \sum_i -\frac{\delta \hat{y}}{\delta \mathbf{X}^{(N)}_{\text{row}=k,\text{col}=i}} \tag{10}$$

where k is the row in the final convolutional layer, Z is a normalization term, $\mathbf{X}^{(N)}$ is the activation of the final convolutional layer, and $\hat{y}$ is the final layer output. The heatmap $\mathbf{S}$ is then computed by the weighted sum of final layer activations:

$$\mathbf{S}_i = \frac{1}{\mathbf{S}_{max}} \sum_k \alpha_k \mathbf{X}^{(N)}_{\text{row}=k,\text{col}=i} \tag{11}$$

This heatmap **S** specifies the important portions in the input sequence that are most responsible for a particular class activation. For each convolutional branch, we can apply this procedure to understand which portions of the input decomposition sequence are contributing the most, according to each filter size $K = 5, 10, 15, 20$. In this way, we can then map the most influential portions onto locations on the backbone of the protein. To the best of our knowledge, this procedure has never been applied to protein (or ligand) structures because Grad-CAM has been rarely applied outside of image processing.

Evaluation criteria

The evaluation criteria for PLI in general is presented by the area under the receiver operating characteristics (AUC)[48]. The receiver operating characteristic curve is the plot of true positive rate vs false positive rate and the area under this curve is AUC. Thus AUC greater than 0.5 suggests that the model performs better than chance. However, AUC faces the early recognition problem (high positive rate for highest ranked ligands which are assayed first) and therefore, may incorrectly judge a model. Enrichment Factor[49,50] (EF) and Boltzman-Enhanced Discrimination of Receiver Operating Characteristic (BEDROC) considers early recognition problem. EF is defined as

$$EF_{X\%} = \frac{Compounds_{selected}/N_{X\%}}{Compounds_{total}/N_{total}} \tag{12}$$

where $N_{X\%}$ is the number of ligands in the top $X\%$ of the ranked ligands. EF thus considers an estimate on a random distribution for how many more actives can be found withing the early recognition threshold. BEDROC is interpreted as the probability that an active is ranked before a ligand taken from random probability distribution in an ordered list. An exponent factor α determines the shape of the distribution. In this work, we chose $\alpha = 20$ following the bench-marking of fingerprints for ligand-based virtual screening[29].

Datasets

The dataset used for the evaluation of PLI prediction models is of critical importance. The dataset should have credible positive and negative samples (i.e., protein-ligand pairs that interact and protein-ligand pairs that do not interact, respectively). However, most of the datasets applied currently for PLI prediction use randomly generated negative sample[51,52] which creates noise in the data as there might be false negatives (i.e. a sample that is categorized as negative by the model but is positive).

The highly credible negative samples datasets human and C.*elegans* were created by Liu et al.[53] by using the concept that a protein dissimilar to a known target will likely not be targeted by a ligand of the latter and vice-versa. The positive samples were created by DrugBank 4.1[54] and Matador[55]. The human dataset contains 1052 and 852 unique compounds and proteins, respectively, for a total of 3369 positive interactions, while the C.*elegans* dataset contains 1434 and 2504 unique compounds and proteins, respectively, for a total of 4000 positive interactions. For comparison with other state-of-the-art models for PLI predictions, we considered the experimental setting suggested by Tabei and Yamanishi[56], where the number of positive to negative samples were 1:1, 1:3 and 1:5 respectively. The negative samples were extracted from human and C.*elegans* as the top candidates based on their scores obtained by Liu et al.[53]. A five fold cross validation was done for evaluation.

For comparison with ML/DL and non-machine learning models we considered the Database of Useful Decoys : Enhanced (DUD-E) dataset[57]. The dataset contains 102 unique proteins and 22,886 positive interactions with an average of 224 compounds per target. For each active compounds, 50 decoys with similar 1D physico-chemical properties were considered to remove bias which have dissimilar 2D topology and therefore are likely non-binders. We randomly divided 102 targets to 72 targets for training and 30 targets for testing. The dataset was then balanced by considering equal number of negative samples (selected randomly) to positive samples for each target, which led to the final evaluation dataset with 22,886 positive and 22,886 negative interactions.

The decoys generated computationally faces the problem of false negatives and therefore an experimental validation of SSnet is mandatory. We considered the BindingDB dataset[58] which is a public, web-accessible database of measured experimental binding affinities and contains around 1.3 million data records. We created a database by considering the following properties for each data entry

1. The target has PDB ID cross-referenced as 3D structure. The first annotated structure is taken as reference PDB file.

2. The ligand has SMILES representation in the entry.

3. Record has IC50 value (a measure of strength of binding) and is either less than 10 nM (positive) or greater than 10 μM (negative).

The dataset contains 1887 unique proteins, and 249,854 (105690 active and 144,164 inactive) unique PLIs. The dataset was divided similarly to DUD-E dataset: 100 proteins for testing and 1787 proteins for training the model.

Results and discussion

The model proposed in the present study, SSnet, takes as inputs the PDB format file (to extract curvature κ and torsion τ) for targets and SMILES strings for ligands. The SMILES string is further converted into molecular descriptors utilizing the methods ECF[28], Avalon[29], GNN[27], and CLP[25] respectively. DNNs with convolutional neural networks (CNN) have a large number of weights to optimize and therefore require a large number of data instances to be able to learn. However, in the human and C.*elegans* dataset the number of instances are insufficient, causing SSnet to overfit (shown as Figure 3 in the supporting information). To overcome this problem, we ignored the convolution layer and directly fed the proteins' curvature and torsion to the fully connected dense layer (similarly to ligands processing) for human and C.*elegans* dataset.

Table 1: Model comparison on the human dataset for various molecular descriptors

Ligand descriptors	AUC
GNN	0.974
ECF	0.982
CLP	0.966
Avalon	0.984

Table 1 shows the performance of the SSnet model in the human dataset (balanced dataset with 1:1 ratio of positive to negative samples) when different ligand descriptors were employed. GNN is built up based on convolution neural networks which require ample amount of data to make sense of the spatial information provided to the model. The model might also suffer if essential information such as functional groups are deeply embedded in the input data and are not directly accessible to the network. This might be one of the reasons for a lower performance of GNN in terms of AUC when compared to ECF and Avalon. CLP gives a descent AUC score of 0.966, however, it is the lowest performing model. CLP is based on autoencoder which is trained to take an input SMILES string, converts it to a lower dimension, and reproduces the SMILES string back. In this way CLP is able to generate a lower dimensional vector for a given SMILES string. However, relevant information required for the prediction of PLI might have lost from the ligands which explains its lower performance. ECF and Avalon have almost similar AUC score as they both directly provide the information of the atoms and functional groups a ligand contains. For further comparison on human and C.*elegans* datasets we considered Avalon as ligand descriptor.

Table 2 shows the comparison of various traditional machine learning models on the human and C.*elegans* datasets. We adapted the same experimental setting as Liu et al.[53] (all models were performed on the same experimental setting) for comparison which was obtained by Tsubaki et al.[20]. The SSnet model outperforms other models in both balanced (1:1) and unbalanced (1:3 and 1:5) datasets. This suggests that the SSnet model is robust

and is able to learn useful information about the protein and ligand pairs. If the model failed to learn robust features, it is likely that the classifier would simply predict the majority class, having a correspondingly low AUC score. We do not observe this. We further compared our

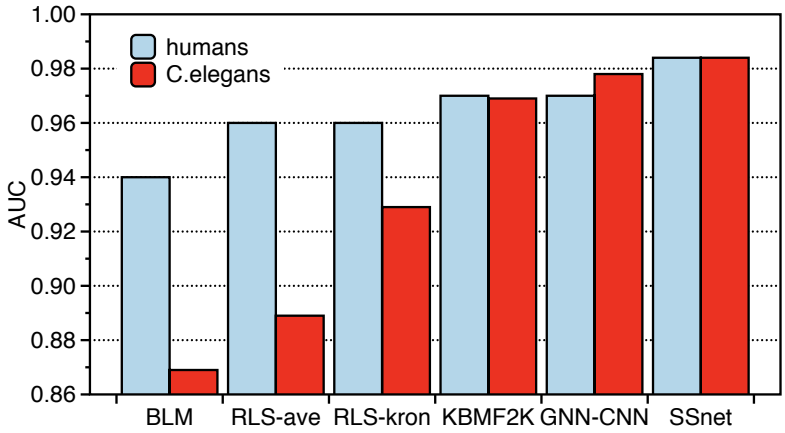


Figure 3: Model comparison on various PLI specific methods with AUC.

model with PLI specific methods : BLM[59], RLS-avg and RLS-Kron[60], KBMF2K-classifier, KBMF2K-regression[61] and GNN-CNN[20] as shown in Figure 3 (performed on the same experimental setting as Liu et al.[53]). It is important to note that BLM, RLS-avg, RLS-Kron, KBMF2K-classifier, and KBMF2K-regression are modeled on properties such as chemical structure similarity matrix, protein sequence similarity matrix and PLI matrix. Despite such preorganized inputs, SSnet was able to outperform them in terms of AUC. The GNN-CNN uses a protein sequence and a SMILES string for ligand based end-to-end learning model. A graph neural network (GNN) is employed over ligands which obtains low-dimensional real-valued vector representations from molecular graphs, and convolutional neural network over sequences of protein. The applicability range of GNN-CNN is superior as it requires only sequence information for a protein compared to SSnet which requires 3D information. However, when compared to accuracy in predicting actives, the superior description of proteins with simpler and direct representation of ligands such as Avalon or ECF, SSnet model was able to outperform the GNN-CNN model. The input description of GNN-CNN model being sequence, limits it's capability to extract crucial information that might be embedded in the

secondary structure.

Table 2: Data comparison (AUC) on balanced and unbalnced datasets

Dataset	k-NN	RF	L2	SVM	GNN-CNN	SSnet
humans (1:1)	0.860	0.940	0.911	0.910	0.970	**0.984**
humans (1:3)	0.904	0.954	0.920	0.942	0.950	**0.978**
humans (1:5)	0.913	0.967	0.920	0.951	0.970	**0.976**
C.*elegans* (1:1)	0.858	0.902	0.892	0.894	0.978	**0.984**
C.*elegans* (1:3)	0.892	0.926	0.896	0.901	0.971	**0.983**
C.*elegans* (1:5)	0.897	0.928	0.906	0.907	0.971	**0.983**

Note: k-nearest neighbour (k-NN), random forest (RF), L2-logistic (L2) and SVM results were obtained by Liu et al.[53].

In addition, we evaluated our model with the DUD-E dataset as shown in Figure 4. The best performing molecular descriptor ECF was employed for ligands and branch convolution neural network was employed for proteins. The performance of other molecular descriptors are shown as Table 1 in the supporting information. The power of convolution neural networks is now visible as the dataset contains enough instances of PLI. We compared our model with Autodock Vina[62] and Smina[63] as non-machine learning methods and Atomnet[19], 3D-CNN[64] and GNN-CNN[20] as machine learning models. Autodock Vina is an open-source program which is based on the classical force fields to predict docking in PLI. Smina is an advanced version of Autodock Vina that adds more control of scoring function and minimization. Atomnet is a machine learning model that considers vectorized versions of 1 Å 3D grids as input vectors for a protein-ligand complex (Note: Atomnet requires 3D information of protein-ligand complexes). A DNN framework is built based on 3D convolutional layers to predict binary PLI. Similar to GNN-CNN, 3D-CNN also takes fixed size 3D grid (24 Å) from the centre of the binding site (requires protein-ligand complex) as input which is converted to density distribution around the centre of each atom. The density distribution is composed of piecewise combination of a Gaussian and a quadratic function. These

information is then fed to convolutional neural networks to predict PLI. All the mentioned 3D models are based on all atoms in the protein ligand complex. Although a satisfactory information is provided to the model, the large number of input features create noise which makes binary prediction of PLI a challenge. With limited amount of information, SSnet was able to outperform all these models in terms of AUC with an AUC score of 0.984. These results suggest that the curvature and torsion information accumulates compact information for PLI prediction tasks. The learning curve of loss over epochs is shown as Figure 8a of the supporting information.

As described in the evaluation criteria section, accuracy and AUC are not the best metric for PLI prediction evaluation. The Receiver operating characteristics (ROC) curve on the DUD-E dataset for each single protein in the test set (30 proteins) is shown as Figure 4 in the supporting information.

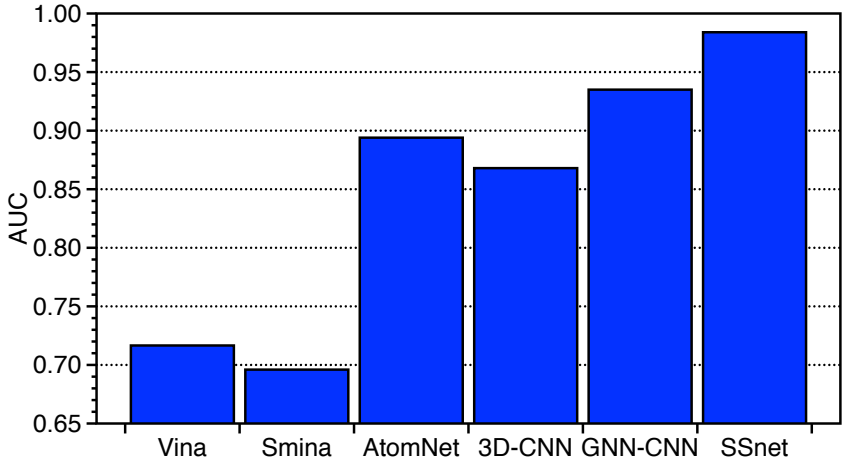


Figure 4: Model comparison of various non-machine and machine learning methods. The AUC score for the methods mentioned are derived from the literature[19,20,64].

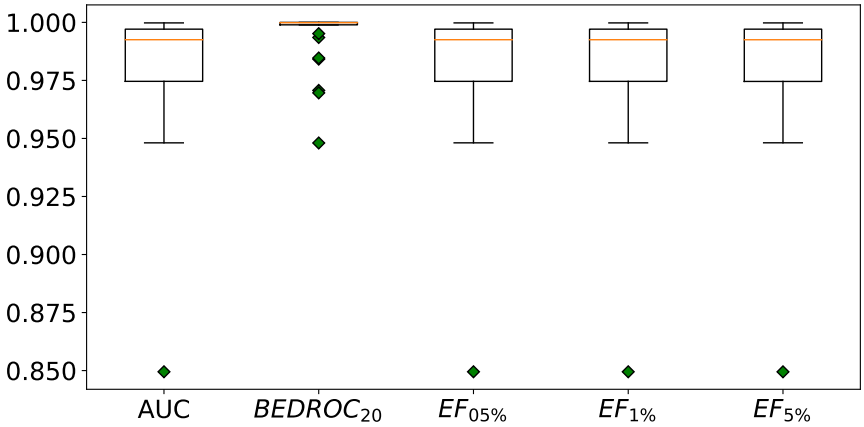


Figure 5: Box plot for the evaluation of predictions for each protein (30) in the test set of DUD-E dataset.

Figure 5 shows the box plot for various metrics applied to evaluate the performance of the SSnet model. The mean AUC, $BEDROC_{20}$, $EF_{0.5}$, $EF_{1.0}$, and $EF_{5.0}$ were 0.982, 0.995, 0.982, and 0.982 respectively. We obtained similar conclusions as with AUC using $EF_{0.5}$, $EF_{1.0}$, and $EF_{5.0}$ respectively. Similarly, $BEDROC_{20}$ is better with a higher mean, confirming that the SSnet model displays superior performance regardless of evaluation criteria.

For validation, using an experimental dataset taken from BindingDB, we benchmarked GNN-CNN, which has outperformed not only partial information (sequence etc.) based model but 3D descriptor based models as well, to predict PLI. We retrained/tested GNN-CNN with the same train/test dataset used for SSnet. The hyperparameters for GNN-CNN are provided in section 2 of the supporting information.

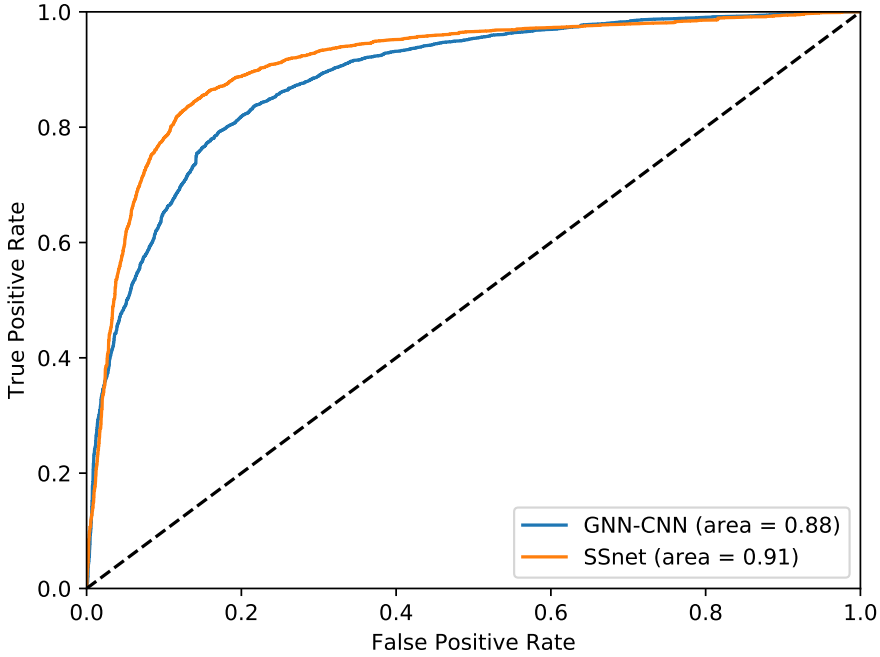


Figure 6: ROC plot for the prediction of PLIs in the test set of BindingDB.

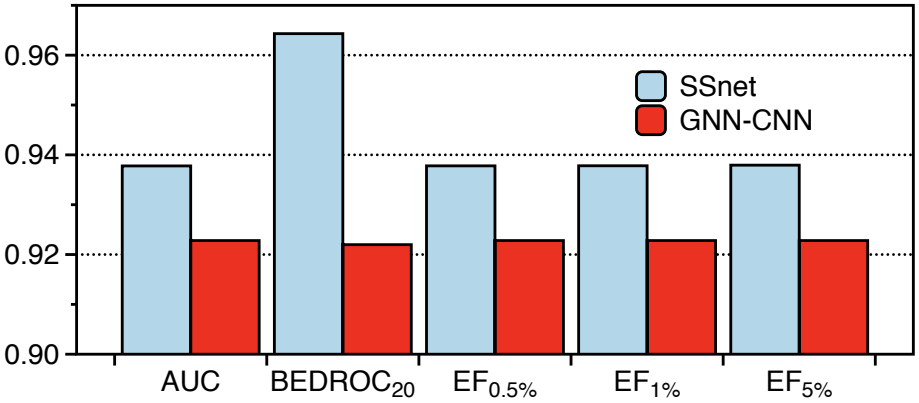


Figure 7: Average AUC, BEDROC and EF on the test set of BindingDB dataset.

Figure 6 shows the comparison of ROC curves for prediction of all PLIs in the test set of BindingDB dataset. SSnet outperforms GNN-CNN in terms of AUC. Figure 7 shows the average AUC, BEDROC and EF on the test set of BindingDB. Similar with AUC, SSnet outperforms GNN-CNN in all metric criteria. The box plot for SSnet and GNN-CNN is shown as Figure 9 and 10 respectively in the supporting information. The learning curve of

loss over epochs is shown as Figure 8b of the supporting information.

Latent space for proteins

In order to start the unraveling of the inner workings of our model we analysed the outputs from the final layers in the global max pooling layer (GMP) (Protein concatenate layer in Figure 2) using the t-distributed Stochastic Neighbor Embedding (t-SNE). The t-SNE Embedding (t-SNE) of GMP is shown as Figure 7 in the supporting information. t-SNE is a method to embed high-dimensional points in low dimensions by retaining similarities between points. In this way, similar data points forms a cluster and are distinguishable with other data points. We tested the proteins in the test set of DUD-E dataset (# of unique proteins = 30) and consider all their interaction with the ligands. The results of t-SNE clearly distinguishes all the proteins (# of clusters = 30). It is important to note that our model had no prior information about the proteins as they were not in the training set. The fact that t-SNE clearly distinguishes all the protein suggests that the information gathered by the convolution layers are not general (such as α helix or β sheet type patterns). More specifically, our model gathers necessary information required for PLI prediction. Based on these results we conclude that our model is able to create a latent space which encodes important information about the bioactivity of the protein. Since the model was trained to predict activity of a protein based on several ligands, such latent space will encode important information about its binding site (including allosteric site) and therefore can be a potential scoring function to compare proteins based on their activity. With the development of latent space we opened a door for future exploration in this direction.

Visualization of heatmap using Grad-CAM

ML/DL are black box models and therefore it is imperative to investigate what features the model learned. In most of the previous studies[20], a neural attention layer is added to the network to understand the important points in the feature space that get higher

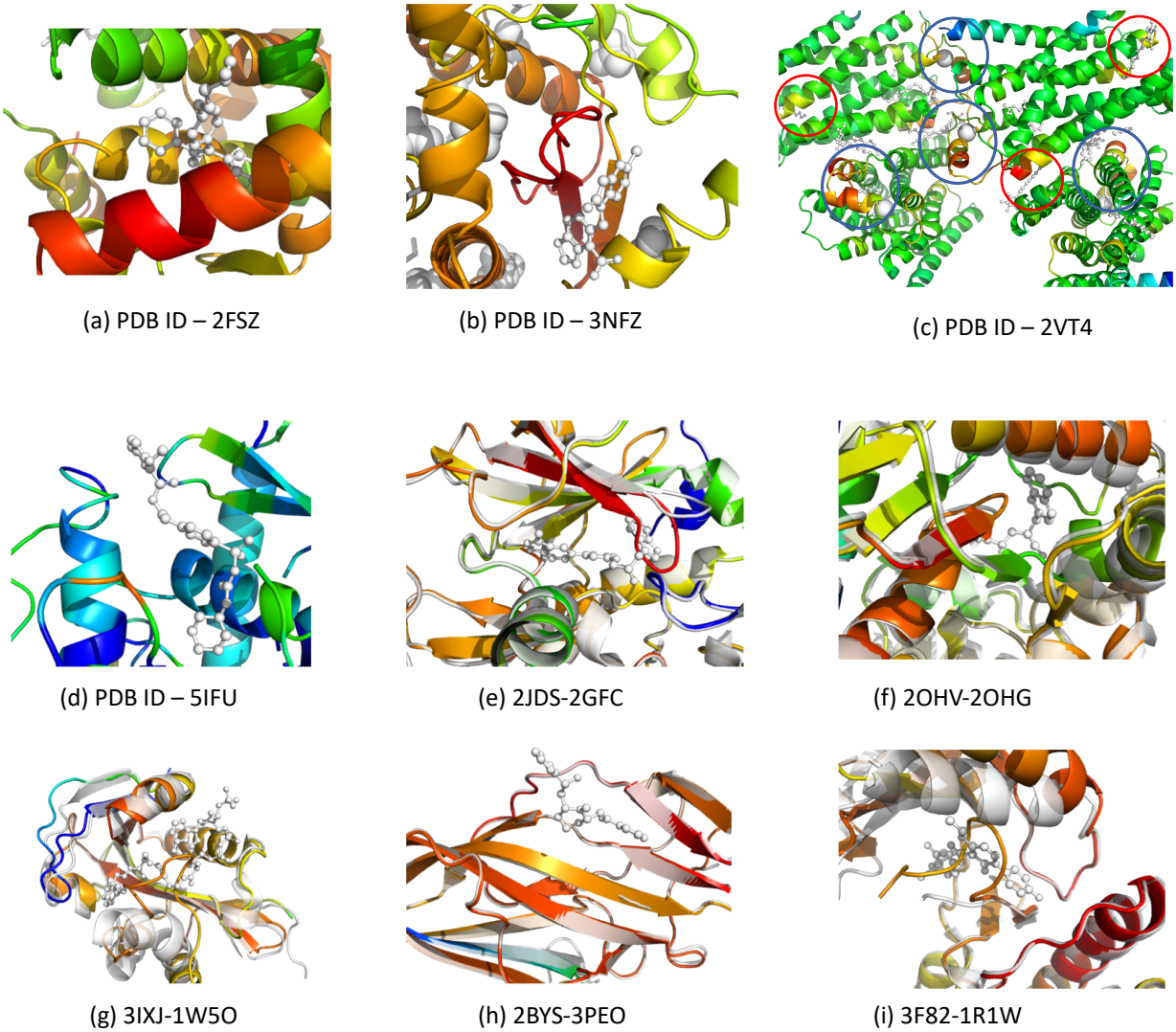


Figure 8: Grad-CAM visualization of the heatmap for nine different proteins. The heatmap is a rainbow mapping with violet as the lowest and red as the highest value. The ligand and other small molecules are shown in white. a) and b) shows that the heatmap generates high value for residues near the binding site. c) shows a protein with four active sites and the model detects them all. The model also detects places where small molecules interact apart from the active site. d) shows a protein's allosteric site which inhibits toxicity from drugs binding at the active site. e), f) and g) shows cryptic sites with the protein-ligand complex shown in white. The heatmap is shown only for the apo protein which is aligned with the protein-ligand complex. Similarly, h) and i) are also cryptic sites, however, in the allosteric location

attention relative to the others. However, the information that is learned from a neural attention layer could be misinterpreted since it adds an additional layer, increasing the complexity of the network. To tackle this problem we opted for Grad-CAM, due to its capability to provide a better insight for which points in the feature are important to predict a particular class based on convolution outputs. A convolution layer is used as it contains most of the spatial information which is lost in a fully connected dense layer. Figure 8a and 8b show the heatmaps generated by Grad-CAM on a homo sapiens protein classified as hormone/growth factor[65] (PDB ID - 2FSZ, taken from test set of DUD-E dataset) with the ligand 4-hydroxytamoxifen (HT) and on a mus musculus organism classified as hydrolase[66] (PDB ID - 3NFZ, not in the training set of DUD-E dataset) with an acetylated substrates as the ligand. Grad-CAM clearly highlights the important residues for ligand recognition. In both cases the ligand forms several non-covalent interactions such as hydrogen bonding, van der Walls interaction etc. with the side chains of the highlighted part in the protein. On a closer look, we observed that the SSnet model in general highlights a weighted probability density of the binding sites present in a protein (heatmaps generated via Grad-CAM for 10 different proteins is shown in section 3 of the supporting information). We further evaluated the protein found in meleagris gallopavo classified as receptor[67] (PDB ID - 2VT4, taken from test set of DUD-E dataset) as shown in Figure 8c. The protein contains four active sites which are comprised of 15 amino acids side chains. The loops define the entrance channel for PLI and are stabilized by sodium ions. SSnet model is able to capture all the 4 sites shown as high values in the heatmap and encircled in blue color. Apart from the active sites, the SSnet model also captures sites which are not necessarily the protein known active sites (encircled in red color). The identification of all possible sites is a useful information for drug discovery as complementary binding sites might regulate various protein functions. Another example can be found in a Plasmodium falciparum protein shown in Figure 8d (PDB ID - 5IFU, not in the training set of DUD-E dataset) in complex with glyburide. It is important to note that glyburide does not bind to the known primary binding site of the

protein but in a distinct secondary binding site. The interaction between glyburide and this secondary binding site causes the inhibition of the protein, making this ligand an allosteric inhibitor. It has been shown that the presence of glyburide overcomes the toxicity related to drugs binding at the active site of this protein[68], which is one of the characteristics of an allosteric inhibitor. The fact that SSnet model is able to highlight the portion of the protein where glyburide could bind, which is not the known orthosteric binding site but an allosteric one, represents one of the various applications that can be carried out with this model. These results also suggest that the SSnet model is learning the relevant information and generalizing the data learned to predict PLI for both unknown proteins and ligands.

Applicability of SSnet

In this work, a ML/DL method that is excellent tools in recognizing patterns has been developed. As we mentioned in section 2, curvature and torsion form various patterns that a machine finds easier to recognize and includes intricate details such as turns, β sheets, α helixes and many more. These patterns are a consequence of the various interactions within a protein (such as side chain interactions). However, the involvement of a ligand perturbs the secondary structure and might cause a slight difference from the original unbound protein structure. To check if SSnet can discover drugs based on an unbound protein structure or a slightly different conformation of the protein, we focused on the following:

- Can SSnet predict the same results using an unbound protein structure or a different conformation of the same protein?
- Can SSnet detect cryptic sites based on unbound protein structures?

Table 3 shows the results of prediction when different proteins conformations are utilized for 10 randomly selected targets in the test set of DUD-E dataset. The first and second columns denote the PDB ID for a protein ligand complex (PLC) in the DUD-E dataset and PDB ID of a different conformation (DC) of the same protein, respectively. The first five

Table 3: Comparison on performance of SSnet on different conformations of a protein

PLC[a]	DC[b]	State[c]	RMSD[d]	Actives		Inactives		Total PLIs
				PLC	DC	PLC	DC	
1B9V	1B9S	Bound	0.267	98	98	98	98	196
1C8K	1A8I	Bound	0.279	76	76	78	78	154
1MV9	1MVC	Bound	0.175	131	131	131	131	262
1Q4X	2J4A	Bound	0.463	103	103	103	103	206
1QW6	1QWC	Bound	0.237	98	98	102	102	200
1BCD	2FNM	Unbound	0.270	493	493	491	491	984
1E66	1EEA	Unbound	0.343	448	448	453	453	901
1H00	4EK3	Unbound	0.178	477	477	471	471	948
1J4H	5HT1	Unbound	0.666	111	111	111	111	222
1KVO	1MF4	Unbound	0.565	99	99	99	99	198

[a]Protein-ligand complex from test set of DUD-E dataset
[b]Different conformation of PLC
[c]Bound refers to DC with a different ligand and unbound refers to DC with no ligand
[d]Root-mean-sqared distance between PLC and DC
*All PLC and DC rows refer to 4 letter PDB ID.

rows have DC with the same protein in PLC bound with a different ligand, and the remaining are apo proteins of the PLCs. The presence of a ligand changes the secondary structure of the protein and therefore we observe a range of root-mean-squared-distance (RMSD) from 0.175 to 0.666 between a PLC-DC pair. The prediction results for each PLC-DC pair in predicting actives and inactives are exactly the same. These results suggest that SSnet is able to predict similar results for a given protein regardless its specific conformation.

Some proteins have binding sites that are not easily detectable as a ligand could induce a significant conformation change. The cryptic proteins are one such example where a cryptic site is defined as a binding site that is present in a protein-ligand complex but not in the apo protein (unbound protein).[69] The change in conformation upon ligand binding is a dynamical phenomena and has been well-known in the literature.[70] Figure 8(e-i) shows cryptic sites for 5 different proteins taken from CryptoSite set.[71] Figure 8e shows the bound (protein with heatmap) and unbound (grey) proteins (cAMP-dependent protein kinase known as Protein Kinase A) respectively. SSnet was utilized to predict the heatmap of unbound protein. The unbound structure with PDB ID 2GFC has the activation loop that protrudes into the active site, closing the binding pocket. The bound structure with PDB- ID 2JDS resides in partially hydrophobic outer region of the active site. SSnet predicts from the unbound structure that the ligand will bind strongly to this protein and highlights location closer to the actual binding site. Retrieving such information is of critical importance as these sites are practically impossible to detect using classical virtual screening methods. Another example is provided by the protein Glutamate Racemase (GluR) shown in Figure 8f with PDB IDs 2OHV and 2OHG as bound and unbound pair, respectively. Similarly, Figure 8g shows the bound-unbound proteins for protease beta-secretase 1 (BACE-1) with PDB IDs 3IXJ-1W50, respectively, where the unbound structure has a low score of binding as the loop comprising residues 71-74 is away from the active site. In both cases SSnet was able to detect that a ligand would be active when tested on unbound protein despite significant differences in the protein structure. SSnet provided a high probability score of more than 0.98 in both scenarios

and highlighted high importance regions near the binding site. Figure 8(h-i) shows cryptic sites for bound-unbound pairs with PDB IDs 2BYS-3PEO (Nicotinic acetylcholine receptor) and 3F82-1R1W (Tyrosine kinase domain of C-MET) respectively. These are the allosteric sites. The allosteric communication is dynamic in nature and the allosteric modulators bind at flexible regions in the protein even without pre-formed binding sites. Allosteric sites are important for drug discovery[72,73] and are a prime motivation when analysing cryptic sites. SSnet being able to detect such sites based on apo protein suggests the versatile nature of the model and shines light on the potential future of fast drug screening models.

Conclusion

In this work we have approached the prediction of protein-ligand interaction based on the secondary structure of proteins and the molecular description of ligands. The protein's secondary structure is acknowledged as the curvature and torsion of the backbone of protein. The curvature and torsion are comprised of 1D data and therefore has compact information that the machine finds easier to learn. In comparison of 3D or 2D feature representation for proteins, the information provided to the machine is sparse and therefore these models have poor performance. The curvature and torsion have unique patterns which are detected by convolution network added in the model. We showed that the machine learns important points in proteins, with the help of a convolution network, where it should look for predicting a protein-ligand interaction. The molecular descriptors were accessed based on several previous studies (graph neural network, variational auto encoder, morgan/circular fingerprints and Avalon fingerprints). Inspired by the t-SNE results for the last layer in protein embedding we propose a possible latent space for proteins that encodes important information about the protein bioactivity and further exploration would result in a metric to compare proteins based on their bioactivity. Our SSnet model outperforms previous models and it predicts protein-ligand interactions in the human, C.*elegans*,DUD-E, and BindingDB

dataset. Our model also shows a strong potential in detecting active sites of proteins even for proteins with multiple binding sites. It also finds all possible locations including allosteric sites and cryptic sites where a ligand might interact, which is an important information for chemists to regulate various properties of proteins such as mitigating toxicity.

It is important to note that SSnet is an ML/DL based method and therefore, is much faster than traditional methods such as Vina[62] or Smina[63]. The SSnet model utilizes secondary structure information of the protein and, since it only processes CA atoms, it does not necessarily require high resolution structural information. The preliminary analysis done on bound-unbound proteins show that SSnet can predict similar results even with different conformations of the protein including cryptic sites. SSnet requires a single conformation (apo protein or protein-ligand complex) to predict whether a ligand is active or not even if the protein-ligand complex has different secondary structure. However, to derive general rules, a more rigorous testing would be required. Another important issue that needs to be addressed is to distinguish between better and lesser binders. Our study suggests that end-to-end learning models based on the secondary structure of proteins have great potential in bioinformatics which is not just confined to protein ligand prediction and can be extended to various biological studies such as protein-protein interaction, protein-DNA interaction, protein-RNA interactions etc. We leave these explorations of the SSnet model for future work.

Acknowledgement

The authors thank SMU for generous supercomputer resources. The authors would also like to thank **Saeedi Mohammadi** for formulating the idea of SSnet and **Brian Zoltowski** for giving helpful tips.

Funding

This work was financially supported by National Science Foundation Grants CHE 1464906.

References

(1) Chen, D.; Oezguen, N.; Urvil, P.; Ferguson, C.; Dann, S. M.; Savidge, T. C. Regulation of Protein-Ligand Binding Affinity by Hydrogen Bond Pairing. *Science Advances* **2016**, *2*, e1501240.

(2) Itoh, Y.; Nakashima, Y.; Tsukamoto, S.; Kurohara, T.; Suzuki, M.; Sakae, Y.; Oda, M.; Okamoto, Y.; Suzuki, T. N+-C-H$\cdots$O Hydrogen Bonds in Protein-ligand Complexes. *Scientific Reports* **2019**, *9*, DOI: 10.1038/s41598-018-36987-9.

(3) Kumar, K.; Woo, S. M.; Siu, T.; Cortopassi, W. A.; Duarte, F.; Paton, R. S. Cation$-\pi$ Interactions in Protein–ligand Binding: Theory and Data-mining reveal Different Roles for Lysine and Arginine. *Chem. Sci.* **2018**, *9*, 2655–2665.

(4) Patil, R.; Das, S.; Stanley, A.; Yadav, L.; Sudhakar, A.; Varma, A. K. Optimized Hydrophobic Interactions and Hydrogen Bonding at the Target-Ligand Interface Leads the Pathways of Drug-Designing. *PLoS ONE* **2010**, *5*, e12029.

(5) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and Scoring in Virtual Screening for Drug Discovery: Methods and Applications. *Nat. Rev. Drug Discovery* **2004**, *3*, 935–949.

(6) Hughes, J.; Rees, S.; Kalindjian, S.; Philpott, K. Principles of Early Drug Discovery. *Br. J. Pharmacol.* **2011**, *162*, 1239–1249.

(7) Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E. W. Computational Methods in Drug Discovery. *Pharmacol. Rev.* **2013**, *66*, 334–395.

(8) Lyne, P. D. Structure-based Virtual Screening: An Overview. *Drug Discovery Today* **2002**, *7*, 1047–1055.

(9) Chen, Y.-C. Beware of Docking! *Trends Pharmacol. Sci.* **2015**, *36*, 78–95.

(10) Pagadala, N. S.; Syed, K.; Tuszynski, J. Software for Molecular Docking: A Review. *Biophys. Rev.* **2017**, *9*, 91–102.

(11) Rifaioglu, A. S.; Atas, H.; Martin, M. J.; Cetin-Atalay, R.; Atalay, V.; Doğan, T. Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases. *Brief. Bioinform* **2018**, 1–36.

(12) Dana, D.; Gadhiya, S.; Surin, L. S.; Li, D.; Naaz, F.; Ali, Q.; Paka, L.; Yamin, M.; Narayan, M.; Goldberg, I.; Narayan, P. Deep Learning in Drug Discovery and Medicine; Scratching the Surface. *Molecules* **2018**, *23*, 2384.

(13) Shen, C.; Ding, J.; Wang, Z.; Cao, D.; Ding, X.; Hou, T. From Machine Learning to Deep Learning: Advances in Scoring Functions for Protein–ligand Docking. *WIREs Comput Mol Sci* **2019**, e1429.

(14) Hong, Y.; Hou, B.; Jiang, H.; Zhang, J. Machine Learning and Artificial Neural Network Accelerated Computational Discoveries in Materials Science. *Wiley Interdiscip. Rev. Comput. Mol. Sci* **2019**, e1450.

(15) Kulik, H. J. Making Machine Learning a Useful Tool in the Accelerated Discovery of Transition Metal Complexes. *Wiley Interdiscip. Rev. Comput. Mol. Sci* **2019**, e1439.

(16) Bredel, M.; Jacoby, E. Chemogenomics: An Emerging Strategy for Rapid Target and Drug Discovery. *Nat. Rev. Genet* **2004**, *5*, 262–275.

(17) Jacob, L.; Vert, J.-P. Protein-ligand Interaction Prediction: An Improved Chemogenomics Approach. *Bioinformatics* **2008**, *24*, 2149–2156.

(18) Yamanishi, Y.; Araki, M.; Gutteridge, A.; Honda, W.; Kanehisa, M. Prediction of Drug-target Interaction Networks from the Integration of Chemical and Genomic Spaces. *Bioinformatics* **2008**, *24*, i232–i240.

(19) Wallach, I.; Dzamba, M.; Heifets, A. AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-based Drug Discovery. 2015; arXiv:1510.02855.

(20) Tsubaki, M.; Tomii, K.; Sese, J. Compound–protein Interaction Prediction with End-to-end Learning of Neural Networks for Graphs and Sequences. *Bioinformatics* **2018**, *35*, 309–318.

(21) Li, L.; Koh, C. C.; Reker, D.; Brown, J. B.; Wang, H.; Lee, N. K.; haw Liow, H.; Dai, H.; Fan, H.-M.; Chen, L.; Wei, D.-Q. Predicting Protein-Ligand Interactions based on Bow-Pharmacological Space and Bayesian Additive Regression Trees. *Sci. Rep.* **2019**, *9*, 7703.

(22) Lee, I.; Keum, J.; Nam, H. DeepConv-DTI: Prediction of Drug-target Interactions via Deep Learning with Convolution on Protein Sequences. *PLoS Comput. Biol* **2019**, *15*, e1007129.

(23) Zheng, S.; Li, Y.; Chen, S.; Xu, J.; Yang, Y. Predicting Drug Protein Interaction using Quasi-Visual Question Answering System. **2019**, DOI: 10.1101/588178.

(24) Ranganathan, S.; Izotov, D.; Kraka, E.; Cremer, D. Description and Recognition of Regular and Distorted Secondary Structures in Proteins using the Automated Protein Structure Analysis Method. *Proteins* **2009**, *76*, 418–438.

(25) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4*, 268–276.

(26) Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. 2013; arXiv:1301.3781.

(27) Scarselli, F.; Gori, M.; Tsoi, A. C.; Hagenbuchner, M.; Monfardini, G. The Graph Neural Network Model. *IEEE Trans. Neural Netw.* **2009**, *20*, 61–80.

(28) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model* **2010**, *50*, 742–754.

(29) Riniker, S.; Landrum, G. A. Open-source Platform to Benchmark Fingerprints for Ligand-based Virtual Screening. *J. Cheminf.* **2013**, *5*, 26.

(30) Ekins, S.; Puhl, A. C.; Zorn, K. M.; Lane, T. R.; Russo, D. P.; Klein, J. J.; Hickey, A. J.; Clark, A. M. Exploiting Machine Learning for End-to-end Drug Discovery and Development. *Nat. Mater* **2019**, *18*, 435–441.

(31) Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. 2016; arXiv:1610.02391.

(32) Zou, L. et al. SARS-CoV-2 Viral Load in Upper Respiratory Specimens of Infected Patients. *New Eng J Med.* **2020**, *382*, 1177–1179.

(33) Amman, B. R. et al. Seasonal Pulses of Marburg Virus Circulation in Juvenile Rousettus aegyptiacus Bats Coincide with Periods of Increased Risk of Human Infection. *PLoS Pathogens* **2012**, *8*, e1002877.

(34) Murzin, A. SCOP: A Structural Classification of Proteins Database for the Investigation of Sequences and Structures. *J. Mol. Biol.* **1995**, *247*, 536–540.

(35) Orengo, C.; Michie, A.; Jones, S.; Jones, D.; Swindells, M.; Thornton, J. CATH – A Hierarchic Classification of Protein Domain Structures. *Structure* **1997**, *5*, 1093–1109.

(36) Dietmann, S.; Holm, L. Identification of Homology in Protein Structure Classification. *Nat. Struct. Biol.* **2001**, *8*, 953–957.

(37) Kabsch, W.; Sander, C. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-bonded and Geometrical Features. *Biopolymers* **1983**, *22*, 2577–2637.

(38) Frishman, D.; Argos, P. Knowledge-based Protein Secondary Structure Assignment. *Proteins* **1995**, *23*, 566–579.

(39) Richards, F. M.; Kundrot, C. E. Identification of Structural Motifs from Protein Coordinate Data: Secondary Structure and First-level Supersecondary Structure. *Proteins* **1988**, *3*, 71–84.

(40) Martin, J.; Letellier, G.; Marin, A.; Taly, J.-F.; de Brevern, A. G.; Gibrat, J.-F. Protein Secondary Structure Assignment Revisited: A Detailed Analysis of Different Assignment Methods. *BMC Struct. Biol* **2005**, *5*, 17.

(41) Day, R.; Beck, D. A.; Armen, R. S.; Daggett, V. A Consensus View of Fold Space: Combining SCOP, CATH, and the Dali Domain Dictionary. *Protein Sci* **2003**, *12*, 2150–2160.

(42) Guo, Z.; Cremer, D. In *Rev. Comp. Chem.*; Lipkowics, K., Boyd, D., Eds.; John Wiley & Sons, New York, 2016; pp 369–438.

(43) Cremer, D.; Kraka, E. From Molecular Vibrations to Bonding, Chemical Reactions, and Reaction Mechanism. *Curr. Org. Chem* **2010**, *14*, 1524–1560.

(44) Venkatachalam, C. M. Stereochemical Criteria for Polypeptides and Proteins. V. Conformation of a System of three linked Peptide Units. *Biopolymers* **1968**, *6*, 1425–1436.

(45) Landrum, G. RDKit: Open-source Cheminformatics. 2020; http://www.rdkit.org.

(46) Variani, E.; Lei, X.; McDermott, E.; Moreno, I. L.; Gonzalez-Dominguez, J. Deep Neural Networks for Small Footprint Text-dependent Speaker Verification. 2014 IEEE

International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2014; pp 4052–4056.

(47) Snyder, D.; Garcia-Romero, D.; Sell, G.; Povey, D.; Khudanpur, S. X-vectors: Robust dnn embeddings for speaker recognition. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2018; pp 5329–5333.

(48) Triballeau, N.; Acher, F.; Brabet, I.; Pin, J.-P.; Bertrand, H.-O. Virtual Screening Workflow Development Guided by the "Receiver Operating Characteristic" Curve Approach. Application to High-Throughput Docking on Metabotropic Glutamate Receptor Subtype 4. *J. Med. Chem.* **2005**, *48*, 2534–2547.

(49) Empereur-mot, C.; Guillemain, H.; Latouche, A.; Zagury, J.-F.; Viallon, V.; Montes, M. Predictiveness Curves in Virtual Screening. *J. Cheminf.* **2015**, *7*, 52.

(50) Pearlman, D. A.; Charifson, P. S. Improved Scoring of Ligand-Protein Interactions Using OWFEG Free Energy Grids. *J. Med. Chem.* **2001**, *44*, 502–511.

(51) Hamanaka, M.; Taneishi, K.; Iwata, H.; Ye, J.; Pei, J.; Hou, J.; Okuno, Y. CGBVS-DNN: Prediction of Compound-protein Interactions Based on Deep Learning. *Mol. Inf.* **2016**, *36*, 1600045.

(52) Tian, K.; Shao, M.; Wang, Y.; Guan, J.; Zhou, S. Boosting Compound-protein Interaction Prediction by Deep Learning. *Methods* **2016**, *110*, 64–72.

(53) Liu, H.; Sun, J.; Guan, J.; Zheng, J.; Zhou, S. Improving Compound–protein Interaction Prediction by Building up Highly Credible Negative Samples. *Bioinformatics* **2015**, *31*, i221–i229.

(54) Wishart, D. S.; Knox, C.; Guo, A. C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M. DrugBank: A Knowledgebase for Drugs, Drug Actions and Drug Targets. *Nucleic Acids Res.* **2007**, *36*, D901–D906.

(55) Gunther, S. et al. SuperTarget and Matador: Resources for Exploring Drug-target Relationships. *Nucleic Acids Res.* **2007**, *36*, D919–D922.

(56) Tabei, Y.; Yamanishi, Y. Scalable Prediction of Compound-protein Interactions using Minwise Hashing. *BMC Systems Biology* **2013**, *7*, S3.

(57) Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.* **2012**, *55*, 6582–6594.

(58) Gilson, M. K.; Liu, T.; Baitaluk, M.; Nicola, G.; Hwang, L.; Chong, J. BindingDB in 2015: A public Database for Medicinal Chemistry, Computational Chemistry and Systems Pharmacology. *Nucleic Acids Research* **2015**, *44*, D1045–D1053.

(59) Bleakley, K.; Yamanishi, Y. Supervised Prediction of Drug–target Interactions Using Bipartite Local Models. *Bioinformatics* **2009**, *25*, 2397–2403.

(60) van Laarhoven, T.; Nabuurs, S. B.; Marchiori, E. Gaussian Interaction Profile Kernels for Predicting Drug–target Interaction. *Bioinformatics* **2011**, *27*, 3036–3043.

(61) Gonen, M. Predicting Drug-target Interactions from Chemical and Genomic Kernels using Bayesian Matrix Factorization. *Bioinformatics* **2012**, *28*, 2304–2310.

(62) Trott, O.; Olson, A. J. AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J. Comput. Chem.* **2009**, *31*, 455–461.

(63) Koes, D. R.; Baumgartner, M. P.; Camacho, C. J. Lessons Learned in Empirical Scoring with Smina from the CSAR 2011 Benchmarking Exercise. *J. Chem. Inf. Model.* **2013**, *53*, 1893–1904.

(64) Ragoza, M.; Hochuli, J.; Idrobo, E.; Sunseri, J.; Koes, D. R. Protein–Ligand Scoring with Convolutional Neural Networks. *J. Chem. Inf. Model.* **2017**, *57*, 942–957.

(65) Wang, Y.; Chirgadze, N. Y.; Briggs, S. L.; Khan, S.; Jensen, E. V.; Burris, T. P. A Second Binding Site for Hydroxytamoxifen within the Coactivator-binding Groove of Estrogen Receptor Beta. *Proc. Natl. Acad. Sci. U.S.A* **2006**, *103*, 9908–9911.

(66) Hsieh, J. M.; Tsirulnikov, K.; Sawaya, M. R.; Magilnick, N.; Abuladze, N.; Kurtz, I.; Abramson, J.; Pushkin, A. Structures of Aminoacylase 3 in Complex with Acetylated Substrates. *Proc. Natl. Acad. Sci. U.S.A* **2010**, *107*, 17962–17967.

(67) Warne, T.; Serrano-Vega, M. J.; Baker, J. G.; Moukhametzianov, R.; Edwards, P. C.; Henderson, R.; Leslie, A. G. W.; Tate, C. G.; Schertler, G. F. X. Structure of a β1-adrenergic G-protein-coupled Receptor. *Nature* **2008**, *454*, 486–491.

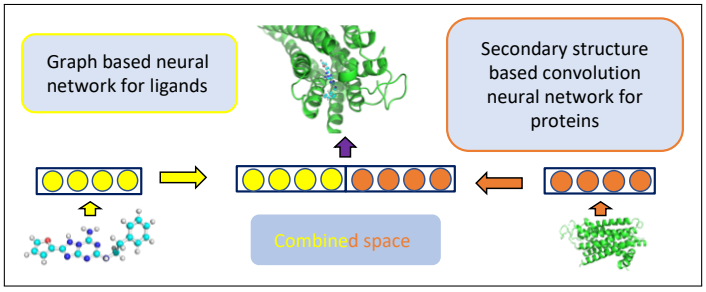
(68) Hewitt, S. N. et al. Biochemical and Structural Characterization of Selective Allosteric Inhibitors of the Plasmodium falciparum Drug Target, Prolyl-tRNA-synthetase. *ACS Infect. Dis.* **2016**, *3*, 34–44.

(69) Cimermancic, P.; Weinkam, P.; Rettenmaier, T. J.; Bichmann, L.; Keedy, D. A.; Woldeyes, R. A.; Schneidman-Duhovny, D.; Demerdash, O. N.; Mitchell, J. C.; Wells, J. A.; Fraser, J. S.; Sali, A. CryptoSite: Expanding the Druggable Proteome by Characterization and Prediction of Cryptic Binding Sites. *J. Mol. Bio.* **2016**, *428*, 709–719.

(70) Laskowski, R. A.; Gerick, F.; Thornton, J. M. The Structural Basis of Allosteric Regulation in Proteins. *FEBS Letters* **2009**, *583*, 1692–1698.

(71) Beglov, D.; Hall, D. R.; Wakefield, A. E.; Luo, L.; Allen, K. N.; Kozakov, D.; Whitty, A.; Vajda, S. Exploring the Structural Origins of Cryptic Sites on Proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2018**, *115*, E3416–E3425.

(72) Bowman, G. R.; Bolin, E. R.; Hart, K. M.; Maguire, B. C.; Marqusee, S. Discovery of Multiple Hidden Allosteric Sites by Combining Markov State Models and Experiments. *Proc. Natl. Acad. Sci. U.S.A* **2015**, *112*, 2734–2739.

(73) Schueler-Furman, O.; Wodak, S. J. Computational Approaches to Investigating Allostery. *Curr. Opin. Chem. Biol.* **2016**, *41*, 159–171.

Graphical TOC Entry