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 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 network for further processing. Furthermore, previous research has not incorporated the secondary structure of the protein in a meaningful manner. In this work, we utilize secondary structure information of the protein which is extracted as the curvature and torsion of the backbone of protein to predict PLI. We demonstrate how our model outperforms

previous machine and non-machine learning models on three major datasets: humans, C.*elegans*, and DUD-E. Visualization of the intermediate layers of our model shows a potential latent space for proteins which extracts important information about the activity of the protein. We further investigate the inner workings of our model by visualizing heatmaps through Grad-CAM. This analysis is adapted to visualize the most important aspects of the protein that the algorithm has learned. We observed that the important residues highlighted by Grad-CAM are the ones responsible for non-covalent interactions with a ligand and is not just confined to the binding site as it also includes allosteric sites and other locations where a ligand interacts. Our new model opens the door in exploration of DNN based on the secondary structure which is not just confined to protein ligand interactions.

Introduction

The interaction of a protein with small molecules is a complex mechanism controlling many fundamental operations in a biological system.[1–12] These interactions are governed by a multitude of factors[10–12] including hydrogen bonding,[1–5] π-interactions,[6–8] hydrophobicity[9] 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. Therefore, it is imperative to develop computational methods to predict PLI. Reliable PLI predictions could significantly reduce the discovery time for new treatments, eliminate toxic drug candidates and efficiently guide medicinal chemistry efforts.[13]

Traditional PLI relies on high-throughput screening which is an experimental technique with high cost and low efficiency.[14] 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).[15]

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.[16] However, SBVS has a somewhat poor performance, sometimes not being able to distinguish active from non-active compounds.[17] Over the last few decades, many classical techniques such as force field, empirical and knowledge based[18] PLI predictions have been applied, however, often showing low performance and in some cases even discrepancies when compared with experimental bioactivities.[19]

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.[20–35] Machine learning based PLI prediction has been developed from a chemogenomics perspective[36] that considers interactions in a unified framework from chemical space and genomic space. Jacob et al.[37] used tensor-product based features and applied Support Vector Machines (SVM) to predict PLI. Yamanishi et al.[38] minimized Euclidean distances over common features derived by mapping ligands and proteins. Izhar et al.[39] used a 3D grid for proteins along with 3D convolutional networks. Masashi et al.[40] used a combination of convolutional network for proteins and graph network for ligands. Li et al.[41] used Bayesian additive regression trees to predict PLI. Ingoo et. al[42] 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 a big and sparse matrix, which hinder ML/DL models to learn and predict PLI. In this work, for the first time we propose a secondary structure based model for proteins with a 1D vector contrary to 3D[39] based representations. The 1D representation is based on the curvature and torsion of protein backbone. Mathematically, curvature and torsion are sufficient to reasonably represent the 3D structure of a protein.[43]

Tremendous work has been carried out to represent the ligands. Rafel at el.[44] 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.[45–47] Esben et al.[48] proposed an improved version of the latent space autoencoder for de novo molecule generation. Scarselli et al.[49] proposed a Graph Neural Network (GNN) to describe molecules. David et al.[50] developed Extended-Connectivity Fingerprints (ECF), which includes the presence of substructures (and therefore also includes stereochemical information) to represent molecules. Sereina et al.[51] 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.[52] 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 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 1 of the supporting information.

We analyzed the SSnet model by utilizing the Grad-CAM method.[53] 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 maximally interact with the ligand.

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,[54] CATH,[55] DALI,[56] and programs like DSSP,[57] STRIDE,[58] DEFINE,[59] and KAKSI[60] have provided protein classifications based on the secondary structure. However, these classifications are often conflicting.[61] A more promising approach has been introduced by Ranganathan et al.[43] based on Frenet-Serret frame and coordinates to classify secondary structure elements in proteins.

In this approach a protein is represented by the α carbons of the backbone (CA atoms) because the backbone defines a unique and recognizable structure, especially for protein categorization.[62] Even so, a significant amount of information about the protein is embedded in the secondary structure elements such as helices, β sheets, hairpins, coils, and turns. Therefore it is imperative to incorporate the secondary structure of the protein when representing its features for machine learning algorithms. 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 1 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.

Figure 2 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

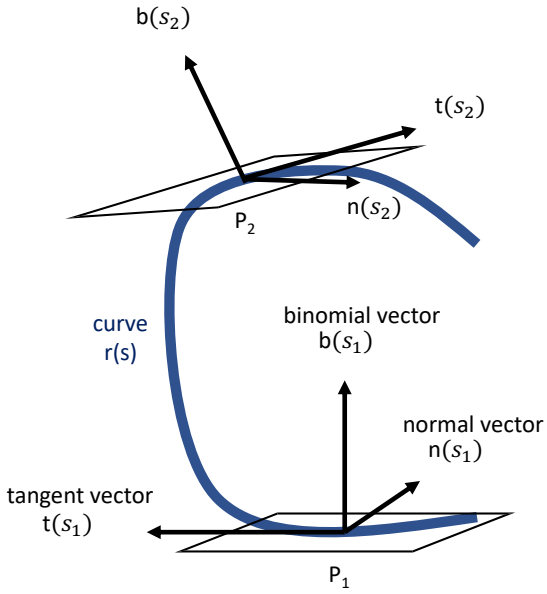


Figure 1: The tangent vector $\mathbf{t}$, normal vector $\mathbf{n}$ and the binormal vector $\mathbf{b}$ of a Frenet frame at points P_1 and P_2 respectively for a curve $\mathbf{r}(s)$

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.

Representation of Ligands

A molecule can be represented by the Simplified Molecular-Input Line-Entry System (SMILES) in the form of strings, which represent 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

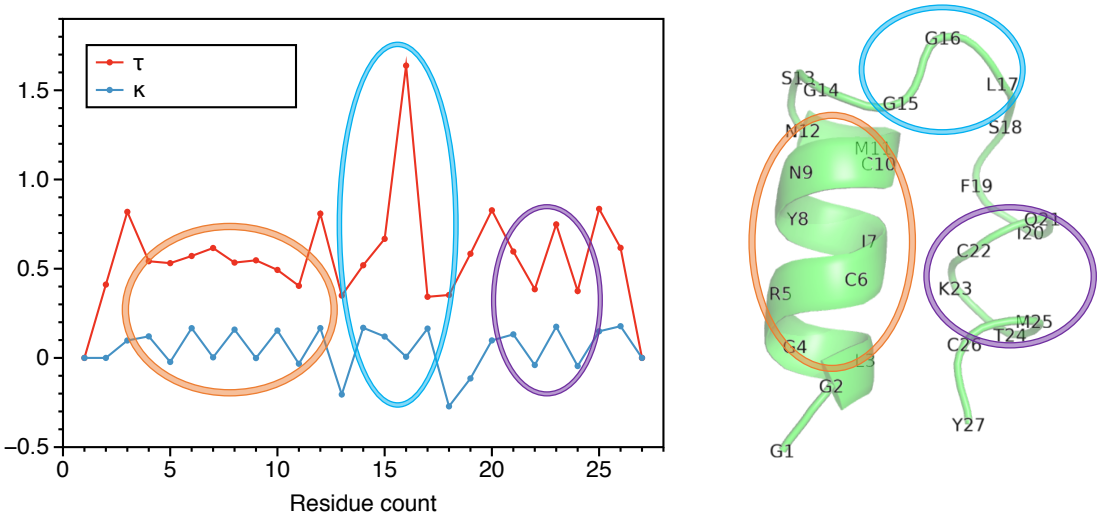


Figure 2: Representation of protein backbone in terms of scalar curvature κ and torsion τ respectively. The ideal helix, turn and non-ideal helix is shown in green, cyan and magenta respectively. The curvature and torsion pattern captures the secondary structure of the protein.

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,[44] GNN,[49] Avalon,[51] and ECF.[50] CLP was generated by the code provided by Rafel at el.;[44] Avalon and ECF were generated from RDKit;[63] and GNN was implemented as proposed by Masashi et al.[40]

SSnet model

Figure 3 shows the end-to-end learning SSnet model developed in this work. First a general overview of the network is given followed by more details about its specific design operation in the remainder of this section. As denoted in the left upper branch of Figure 3 after

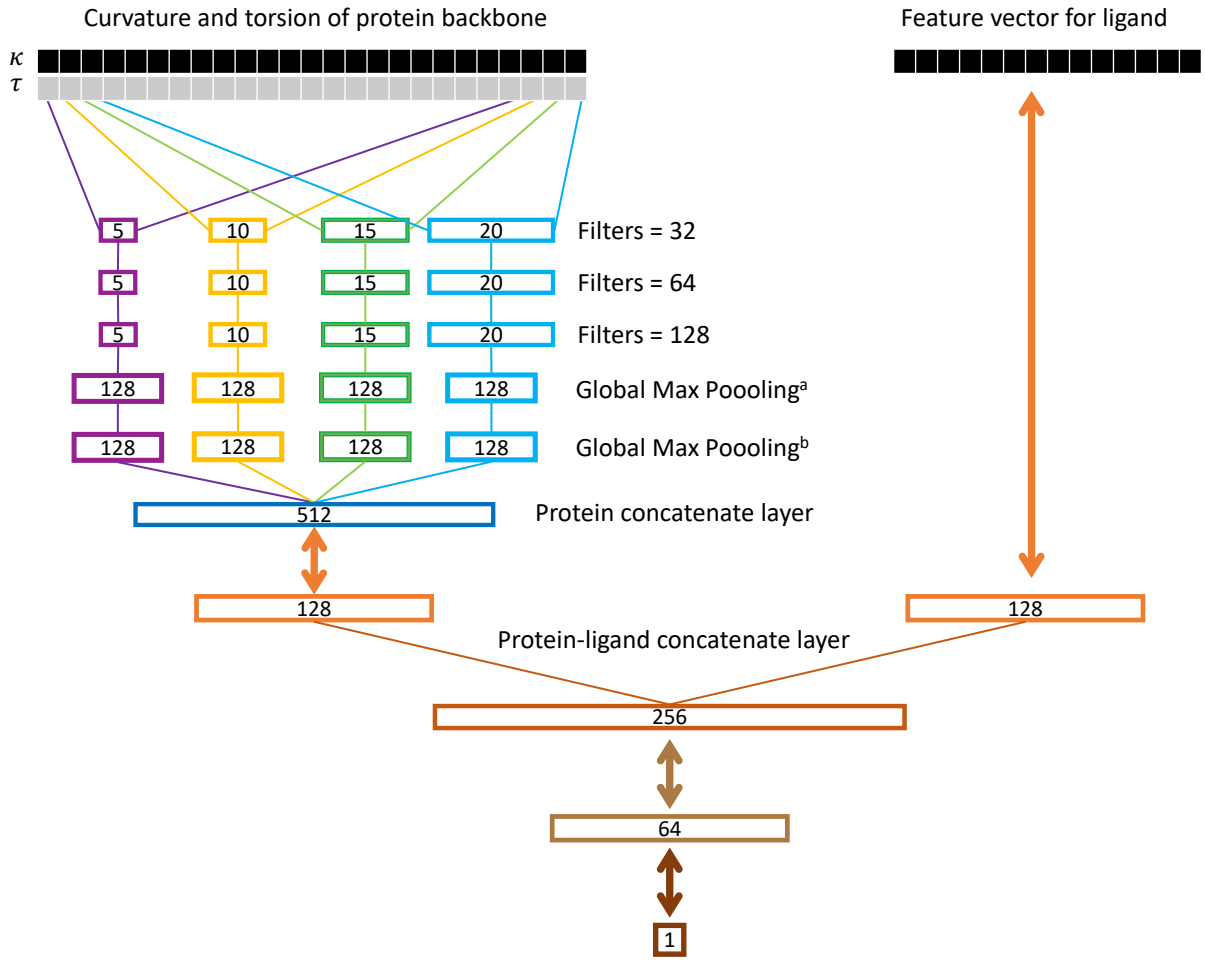


Figure 3: SSnet model. The curvature and torsion pattern of a protein backbone is fed through multiple convolution networks with varying window sizes as branch convolution. Each branch further goes through more convolution with same window size (red, orange, green and light blue boxes). A global max 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 Masashi et al.[40]

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, $\mathbf{X}^{(0)}$, 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 3 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 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 n^{th} 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 $\mathbf{X}^{(N)}_{\kappa}$ where κ denotes the branch. The activation $\mathbf{X}^{(N)}_{\kappa}$ is often referred to as the latent space because it denotes the latent features of the input sequence. The number of columns in $\mathbf{X}^{(N)}_{\kappa}$ 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 $\mathbf{X}^{(N)}_{\kappa}$. This is typically referred to as a Global Max Pooling layer in neural networks and is repeated R times for each row in $\mathbf{X}^{(N)}_{\kappa}$:

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

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:

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

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 that 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.[64–66] In speech systems, the latent space is collapsed via mean and standard deviation operations and the embeddings provided for these operations are typically referred to as D-Vectors[64] or X-Vectors.[65,66] 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:

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

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:

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

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.

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 imperative to verify what a neural network considers "influential" for classification after training. Ramprasaath et al.[53] 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 $\mathbf{S}$ 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).[67] 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[68,69] (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 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.[70]

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[71,72] 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.[73] by using the concept that proteins dissimilar to known target are much likely to be targeted by a compound and vice-versa. The positive samples were created by DrugBank 4.1[74] and Matador.[75] The human dataset contains 1052 and 852 unique compounds and proteins respectively for 3369 positive interactions and the C.*elegans* dataset contains 1434 and 2504 unique compounds and proteins respectively for 4000 positive interactions. For comparison with other state-of-the-art models for PLI predictions, we considered the experimental setting suggested by Tabei et al.,[76] 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.[73] 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.[77] 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. However, the decoys can have false negatives and therefore we performed Autodock Vina[78] on all the decoys and considered 80 % (11,28,971) of the top scoring decoys. 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.

Results and discussion

Our model takes the curvature κ and torsion τ for proteins and SMILES strings for ligands as input. The SMILES string is further converted into molecular descriptors utilizing the methods ECF,[50] Avalon,[51] GNN[49] and CLP[44] respectively. DNNs with convolutional neural networks (CNN) have a large number of weights to optimize and therefore requires a large number of data instances to learn. However, in the human and C.*elegans* datset the number of instances are very few and therefore the proposed model (SSnet) overfits (shown as Figure 2 in the supporting information). To overcome this problem we ignored the convolution layer and directly fed the curvature and torsion of proteins to fully connected dense layer (similar to ligands) for human and C.*elegans* dataset.

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

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

employed. GNN is built up based on convolution neural networks which requires ample amount of data to make sense of the spatial information provided to the model. 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, convert it to lower dimensions, and reproduce 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.[73] (all models were performed on the same experimental setting) for comparison which was obtained from Masashi et al.[40] The SSnet model outperforms all 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.

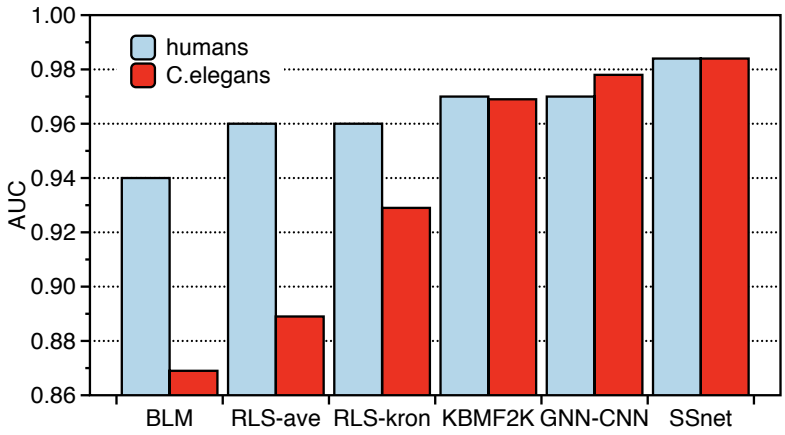
We further compared our model with PLI specific methods : BLM,[79] RLS-avg and RLS-Kron,[80] KBMF2K-classifier, KBMF2K-regression[81] and GNN-CNN[40] as shown in Figure 4a (performed on the same experimental setting as Liu et al.[73]). It is important to note that BLM, RLS-avg, RLS-Kron, KBMF2K-classifier, 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 superior description of proteins in SSnet model was also able to outperform the state-of-the-art model GNN-CNN.

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

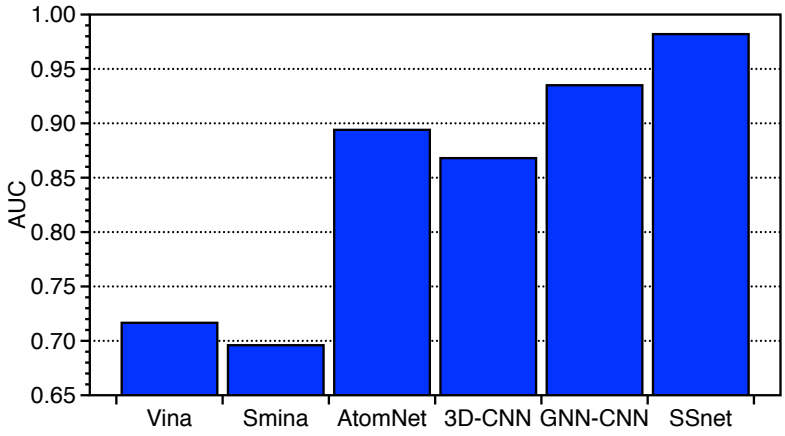
Dataset	k-NN	RF	L2	SVM	GNN-CNN	SSnet
humans (1:1)	0.860	0.940	0.911	0.910	0.970	**0.984**
humans (1:3)	0.904	0.954	0.920	0.942	0.950	**0.978**
humans (1:5)	0.913	0.967	0.920	0.951	0.970	**0.976**
C.*elegans* (1:1)	0.858	0.902	0.892	0.894	0.978	**0.984**
C.*elegans* (1:3)	0.892	0.926	0.896	0.901	0.971	**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.[73]

In addition we evaluated our model with the DUD-E dataset as shown in Figure 4b. The best performing molecular descriptor GNN 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 are now visible as the dataset contains enough instances of PLI. We compared our model with Autodock Vina[78] and Smina[83] as non-machine learning methods and Atomnet,[39] 3D-CNN[82] and GNN-CNN[40] 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 mini-



(a) human and C.*elegans* datasets



(b) DUD-E dataset

Figure 4: Model comparison on various methods to predict PLI in terms of AUC for a) human and C.*elegans* and b) DUD-E datasets. The AUC score for the methods mentioned are derived from the literature.[39,40,82]

mization. Atomnet and 3D-CNN are DNN based models which considers 3D information of proteins and ligands. Our model outperforms all these models including the state-of-the-art model GNN-CNN in terms of AUC.

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 3 in the supporting information. Figure 5 shows various metrics applied to evaluate the performance of the SSnet model. The mean AUC, BEDROC_{20}, $\text{EF}_{0.5}$, $\text{EF}_{1.0}$, and $\text{EF}_{5.0}$ were 0.986, 0.995, 0.986, and 0.986 respectively. We obtained similar conclusions as with AUC using $\text{EF}_{0.5}$, $\text{EF}_{1.0}$, and $\text{EF}_{5.0}$ respectively. Similarly, BEDROC_{20} is better with a higher mean, confirming that the SSnet model performs superior regardless of evaluation criteria.

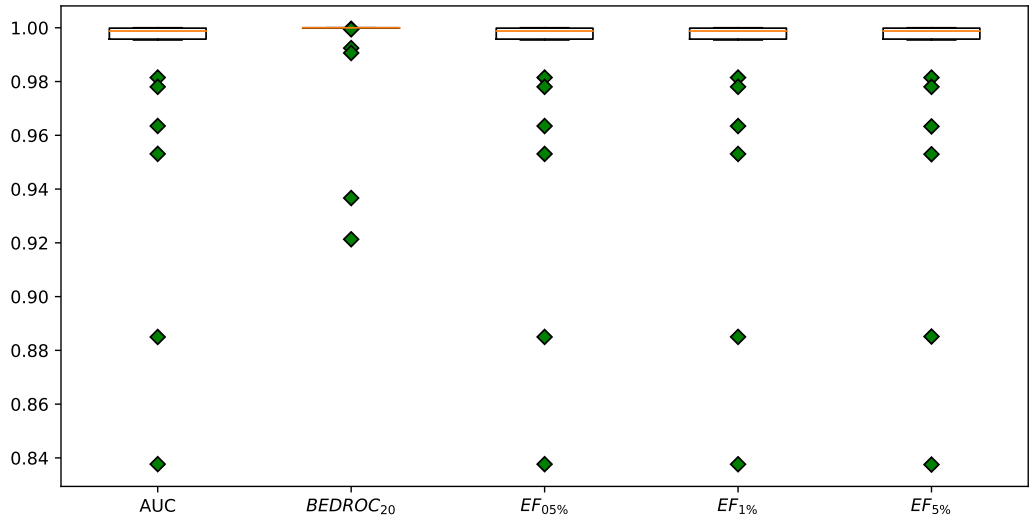


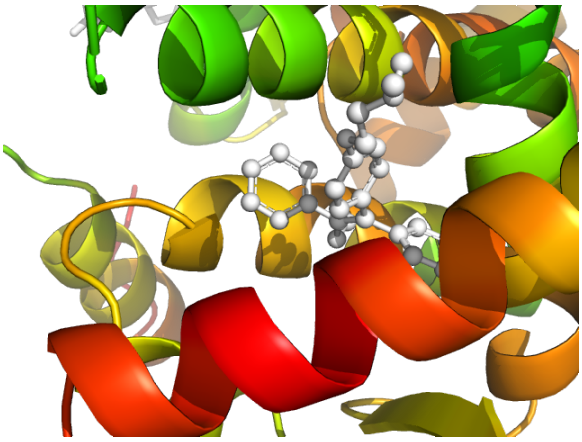
Figure 5: Box plot for the evaluation of predictions for each protein (30) in the test set of DUD-E dataset.

Latent space for proteins

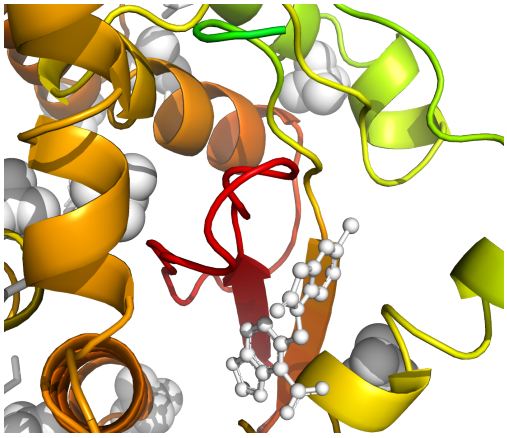
In order to further test the intrinsic mechanism of our model we analysed the outputs from the final layers in the global max pooling layer (GMP) (Protein concatenate layer in Figure

3). The t-distributed Stochastic Neighbor Embedding (t-SNE) of GMP is shown as Figure 4 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 is 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 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 activity of the protein. Since the model was trained to predict activity of the protein based on several ligands, such latent space will encode important information about the binding site (including allosteric site) of the protein and therefore can be a potential scoring function to compare proteins based on its activity. With the development of latent space we opened a door for future exploration in this direction.

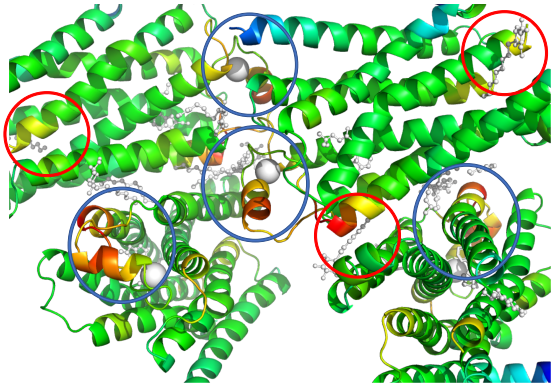
Visualization of heatmap using Grad-CAM



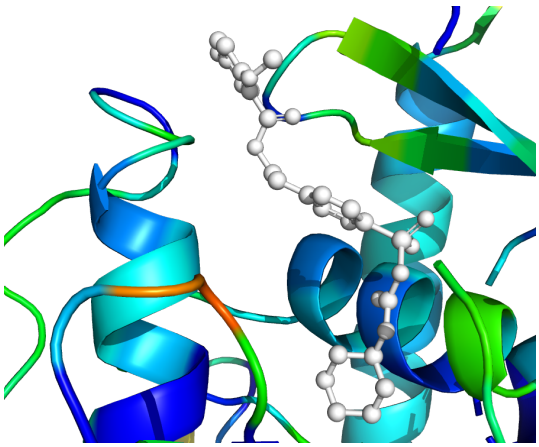
(a) PDB ID - 2FSZ



(b) PDB ID - 3NFZ



(c) PDB ID - 2VT4



(d) PDB ID - 5IFU

Figure 6: Grad-CAM visualization of the heatmap for four different proteins. The heatmap is 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. The model is able to capture all possible places where a ligand could bind including allosteric sites. Note : The model has no prior information of any ligand or small molecules location.

ML/DL are black box models and therefore it is imperative to investigate what features the model learned. In most of the previous studies[40] , a neural attention layer is added to the network to understand the important points in the feature which gets higher attention

relative to others. However, neural attention is a complex network and therefore can result in information that can be misinterpreted (explained in the method section). To tackle this problem we approached towards Grad-CAM which provides a better insight for which points are important in the feature 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 6a and 6b shows the heatmap generated by Grad-CAM on a homo sapiens protein classified as hormone/growth factor[84] (PDB ID - 2FSZ) with the ligand 4-hydroxytamoxifen (HT) and on a mus musculus organism classified as hydrolase[85] (PDB ID - 3NFZ) with the ligand as an acetylated substrates. 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 highlighted part of 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 as Figure 5-15 in the supporting information). We further evaluated the protein found in meleagris gallopavo classified as receptor[86] (PDB ID - 2VT4) as shown in Figure 6c. The protein contains four active sites which is comprised of 15 side chains from amino acid residues. The loops define the entrance channel for PLI and is stabilized by a sodium ion. 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 not necessarily in the active site and is encircled in red color. The identification of all possible sites is a useful information for drug discovery as it regulates various properties of the enzyme. Another such example is of a protein found in Plasmodium falciparum shown in Figure 6d (PDB ID - 5IFU) in complex with glyburide. It is important to note that glyburide is not in the known binding site of the protein and is slightly away from it which acts as 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.[87] The fact that SSnet model highlights the portion of the protein where glyburide could have bound

(not in the active site) represents one of the various applications that can be carried out with our model. These results also suggest that the SSnet model is learning the relevant information from the features for PLI prediction.

Conclusion

In this work we have approached the prediction of protein ligand interaction via end-to-end learning 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 convolution network for where it should look for when 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 activity and further exploration would result in a metric to compare proteins based on their activity. Our SSnet model outperforms previous models which predicts protein ligand interaction in the human, C.*elegans* and the DUD-E 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 where a ligand might interact which is an important information for chemists to regulate various properties of the enzyme 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[78] or Smina.[83] The SSnet model utilizes secondary structure information of the protein and therefore, does not necessarily require high resolution structural information. For validation and reproducibility all codes developed in this work are publicly accessible at CATCO-Github.

Our study suggest that end-to-end learning models based on the secondary structure of proteins has 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.

Associated content

Model overview, training loss on human dataset, results obtained by the SSnet model for all molecular descriptors and statistical results on DUD-E dataset, and Grad-CAM analysis on various proteins.

Acknowledgement

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

References

(1) Panigrahi, S. K. Strong and weak hydrogen bonds in protein-ligand complexes of kinases: a comparative study. *Amino Acids* **2008**, *34*, 617–633.

(2) 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.

(3) 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.

(4) Zhou, W.; Yan, H.; Hao, Q. Analysis of surface structures of hydrogen bonding in protein–ligand interactions using the alpha shape model. *Chem. Phys. Lett.* **2012**, *545*, 125–131.

(5) Williams, M. A.; Ladbury, J. E. *Protein-Ligand Interactions*; John Wiley & Sons, Ltd, 2005; Chapter 6, pp 137–161, DOI: 10.1002/3527601813.ch6.

(6) 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.

(7) Brylinski, M. Aromatic interactions at the ligand-protein interface: Implications for the development of docking scoring functions. *Chemical Biology & Drug Design* **2017**, *91*, 380–390.

(8) Gallivan, J. P.; Dougherty, D. A. Cation-π interactions in structural biology. *Proc. Natl. Acad. Sci. U.S.A* **1999**, *96*, 9459–9464.

(9) 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.

(10) Eyers, C. E.; Vonderach, M.; Ferries, S.; Jeacock, K.; Eyers, P. A. Understanding protein–drug interactions using ion mobility–mass spectrometry. *Curr. Opin. Chem. Biol.* **2018**, *42*, 167–176.

(11) Zhou, H.; Sharma, A. Therapeutic protein-drug interactions: plausible mechanisms and assessment strategies. *Expert Opin. Drug Metab. Toxicol* **2016**, *12*, 1323–1331.

(12) West, G. M.; Tucker, C. L.; Xu, T.; Park, S. K.; Han, X.; Yates, J. R.; Fitzgerald, M. C. Quantitative proteomics approach for identifying protein-drug interactions in complex mixtures using protein stability measurements. *Proc. Natl. Acad. Sci. U.S.A* **2010**, *107*, 9078–9082.

(13) 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.

(14) Hughes, J.; Rees, S.; Kalindjian, S.; Philpott, K. Principles of early drug discovery. *Br. J. Pharmacol.* **2011**, *162*, 1239–1249.

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

(16) Lyne, P. D. Structure-based virtual screening: an overview. *Drug Discovery Today* **2002**, *7*, 1047–1055.

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

(18) 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.

(19) Pagadala, N. S.; Syed, K.; Tuszynski, J. Software for molecular docking: a review. *Biophys. Rev.* **2017**, *9*, 91–102.

(20) Lavecchia, A. Machine-learning approaches in drug discovery: methods and applications. *Drug Discov. Today* **2015**, *20*, 318–331.

(21) Lo, Y.-C.; Rensi, S. E.; Torng, W.; Altman, R. B. Machine learning in chemoinformatics and drug discovery. *Drug Discov. Today* **2018**, *23*, 1538–1546.

(22) Lima, A. N.; Philot, E. A.; Trossini, G. H. G.; Scott, L. P. B.; Maltarollo, V. G.; Honorio, K. M. Use of machine learning approaches for novel drug discovery. *Expert Opin. Drug Discovery* **2016**, *11*, 225–239.

(23) Zhang, L.; Tan, J.; Han, D.; Zhu, H. From machine learning to deep learning: progress in machine intelligence for rational drug discovery. *Drug Discov. Today* **2017**, *22*, 1680–1685.

(24) Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The rise of deep learning in drug discovery. *Drug Discov. Today* **2018**, *23*, 1241–1250.

(25) Ghasemi, F.; Mehridehnavi, A.; Pérez-Garrido, A.; Pérez-Sánchez, H. Neural network and deep-learning algorithms used in QSAR studies: merits and drawbacks. *Drug Discov. Today* **2018**, *23*, 1784–1790.

(26) Baskin, I. I.; Winkler, D.; Tetko, I. V. A renaissance of neural networks in drug discovery. *Expert Opin. Drug Discovery* **2016**, *11*, 785–795.

(27) 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.

(28) Panteleev, J.; Gao, H.; Jia, L. Recent applications of machine learning in medicinal chemistry. *Bioorg. Med. Chem. Lett* **2018**, *28*, 2807–2815.

(29) Jing, Y.; Bian, Y.; Hu, Z.; Wang, L.; Xie, X.-Q. S. Deep Learning for Drug Design: an Artificial Intelligence Paradigm for Drug Discovery in the Big Data Era. *The AAPS Journal* **2018**, *20*, 58.

(30) Hessler, G.; Baringhaus, K.-H. Artificial Intelligence in Drug Design. *Molecules* **2018**, *23*, 2520.

(31) 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.

(32) 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.

(33) Duran-Frigola, M.; Fernández-Torras, A.; Bertoni, M.; Aloy, P. Formatting biological big data for modern machine learning in drug discovery. *Wiley Interdiscip. Rev. Comput. Mol. Sci* **2018**, *9*, e1408.

(34) 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.

(35) 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.

(36) Bredel, M.; Jacoby, E. Chemogenomics: an emerging strategy for rapid target and drug discovery. *Nature Reviews Genetics* **2004**, *5*, 262–275.

(37) Jacob, L.; Vert, J.-P. Protein-ligand interaction prediction: an improved chemogenomics approach. *Bioinformatics* **2008**, *24*, 2149–2156.

(38) 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.

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

(40) 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.

(41) 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.

(42) Lee, I.; Keum, J.; Nam, H. DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences. *PLOS Computational Biology* **2019**, *15*, e1007129.

(43) 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.

(44) 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.

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

(46) Goldberg, Y.; Levy, O. word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. 2014; arXiv:1402.3722.

(47) Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; Dean, J. Distributed Representations of Words and Phrases and their Compositionality. 2013; arXiv:1310.4546.

(48) Bjerrum, E. J.; Sattarov, B. Improving Chemical Autoencoder Latent Space and Molecular De Novo Generation Diversity with Heteroencoders. *Biomolecules* **2018**, *8*, 131.

(49) Scarselli, F.; Gori, M.; Tsoi, A. C.; Hagenbuchner, M.; Monfardini, G. The graph neural network model. *IEEE Trans. Neural Netw. 20*, 61–80.

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

(51) Riniker, S.; Landrum, G. A. Open-source platform to benchmark fingerprints for ligand-based virtual screening. *J. Cheminf.* **2013**, *5*, 26.

(52) 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. *Nature Materials* **2019**, *18*, 435–441.

(53) 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.

(54) Murzin, A. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **1995**, *247*, 536–540.

(55) 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.

(56) Dietmann, S.; Holm, L. Identification of homology in protein structure classification. *Nat. Struct. Biol.* **2001**, *8*, 953–957.

(57) Kabsch, W.; Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22*, 2577–2637.

(58) Frishman, D.; Argos, P. Knowledge-based protein secondary structure assignment. *Proteins* **1995**, *23*, 566–579.

(59) 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.

(60) Martin, J.; Letellier, G.; Marin, A.; Taly, J.-F.; de Brevern, A. G.; Gibrat, J.-F. *BMC Structural Biology* **2005**, *5*, 17.

(61) 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 Science* **2003**, *12*, 2150–2160.

(62) Venkatachalam, C. M. Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units. *Biopolymers* **1968**, *6*, 1425–1436.

(63) Landrum, G. RDKit: Open-source cheminformatics. http://www.rdkit.org.

(64) 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.

(65) 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.

(66) Ramoji, S.; Mohan, A.; Mysore, B.; Bhatia, A.; Singh, P.; Vardhan, H.; Ganapathy, S. The Leap Speaker Recognition System for NIST SRE 2018 Challenge. ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2019; pp 5771–5775.

(67) 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.

(68) Empereur-mot, C.; Guillemain, H.; Latouche, A.; Zagury, J.-F.; Viallon, V.; Montes, M. Predictiveness curves in virtual screening. *J. Cheminf.* **2015**, *7*, 52.

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

(70) Riniker, S.; Landrum, G. A. Open-source platform to benchmark fingerprints for ligand-based virtual screening. *J. Cheminf.* **2013**, *5*, 26.

(71) 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.

(72) Tian, K.; Shao, M.; Wang, Y.; Guan, J.; Zhou, S. Boosting compound-protein interaction prediction by deep learning. *Methods* **2016**, *110*, 64–72.

(73) 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.

(74) 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.

(75) Gunther, S. et al. SuperTarget and Matador: resources for exploring drug-target relationships. *Nucleic Acids Res.* **2007**, *36*, D919–D922.

(76) Tabei, Y.; Yamanishi, Y. Scalable prediction of compound-protein interactions using minwise hashing. *BMC Systems Biology* **2013**, *7*, S3.

(77) 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.

(78) 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.

(79) Bleakley, K.; Yamanishi, Y. Supervised prediction of drug–target interactions using bipartite local models. *Bioinformatics* **2009**, *25*, 2397–2403.

(80) van Laarhoven, T.; Nabuurs, S. B.; Marchiori, E. Gaussian interaction profile kernels for predicting drug–target interaction. *Bioinformatics* **2011**, *27*, 3036–3043.

(81) Gonen, M. Predicting drug-target interactions from chemical and genomic kernels using Bayesian matrix factorization. *Bioinformatics* **2012**, *28*, 2304–2310.

(82) 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.

(83) 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.

(84) 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.

(85) 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.

(86) 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.

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

Graphical TOC Entry