Explainable Deep Relational Networks for Predicting Compound-Protein Affinities and Contacts

Mostafa Karimi,[†,‡,§] Di Wu,[†,§] Zhangyang Wang,[¶] and Yang Shen[∗,†,‡]

†Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843, United States

‡TEES-AgriLife Center for Bioinformatics and Genomic Systems Engineering, Texas A&M University, College Station, TX 77840, United States

¶Department of Computer Science and Engineering, Texas A&M University, College Station, TX 77843, United States

§Co-first authors

E-mail: yshen@tamu.edu

Abstract

Predicting compound-protein affinity is critical for accelerating drug discovery. Recent progress made by machine learning focuses on accuracy but leaves much to be desired for interpretability. Through molecular contacts underlying affinities, our large-scale interpretability assessment finds commonly-used attention mechanisms inadequate. We thus formulate a hierarchical multi-objective learning problem whose predicted contacts form the basis for predicted affinities. We further design a physics-inspired deep relational network, DeepRelations, with intrinsically explainable architecture. Specifically, various atomic-level contacts or "relations" lead to molecular-level

1

affinity prediction. And the embedded attentions are regularized with predicted structural contexts and supervised with partially available training contacts. DeepRelations shows superior interpretability to the state-of-the-art: without compromising affinity prediction, it boosts the AUPRC of contact prediction 9.5, 16.9, 19.3 and 5.7-fold for the test, compound-unique, protein-unique, and both-unique sets, respectively. Our study represents the first dedicated model development and systematic model assessment for interpretable machine learning of compound-protein affinity.

Introduction

Current drug-target interactions are predominantly represented by the interactions between small-molecule compounds as drugs and proteins as targets.[1] The enormous chemical space to screen compounds is estimated to contain 10^{60} drug-like compounds.[2] And these compounds act in biological systems of millions or more protein species or "proteoforms" (considering genetic mutations, alternative splicing, and post-translation modifications of proteins).[3,4] Facing such a combinatorial explosion of compound-protein pairs, drug discovery calls for efficient characterization of compound efficacy and toxicity, and computational prediction of compound-protein interactions (CPI) addresses the need.

Recently computational CPI prediction has made major progress beyond predicting whether compounds and proteins interact. Indeed, thanks to increasingly abundant molecular data and advanced deep-learning techniques, compound-protein affinity prediction is reaching unprecedented accuracy, with inputs of compound-protein structures,[5–7] compound identities (such as SMILES and graphs) and protein structures (see a relevant problem of binding classification[8,9]), or even just compound-protein identities.[10–12] As previously summarized,[10] structure-based affinity-prediction methods are limited in applicability due to the often-unavailable structures of compound-protein pairs or even proteins alone, whereas structure-free methods, being broadly applicable, could be limited in interpretability.

Interpretability remains a major gap between the capability of current compound-affinity

predictors and the demand of rational drug discovery. The central question about interpretability is whether and how methods (including machine learning models) could explain *why* they make certain predictions (affinity level for any compound-protein pair in our context). This important topic is rarely addressed for affinity prediction. DeepAffinity[10] has embedded joint attentions over compound-protein component pairs and uses such joint attentions to assess origins of affinities (binding sites) or specificities. Additionally, attention mechanisms have been used for predictions of CPI,[13] chemical stability[14] and protein secondary structures.[15] Assessment of interpretability for all these studies was either lacking or limited to few case studies. We note a recent work proposing *post-hoc* attribution-based test to determine whether a model learns binding mechanisms.[16]

We raise reasonable concerns on how much attention mechanisms can reproduce natural contacts in compound-protein interactions. Attention mechanisms were originally developed to boost the performance of seq2seq models for neural machine translations.[17] And they have gained popularity for interpreting deep learning models in visual question answering,[18] natural language processing[19] and healthcare.[20] However, they were also found to work differently from human attentions in visual question answering.[21]

Representing the first effort dedicated to interpretability of compound-protein affinity predictors, our study is focused on how to define, assess, and enhance such interpretability as follows.

How to define interpretability for affinity prediction. Interpretable machine learning is increasingly becoming a necessity[22] for fields beyond drug discovery. Unlike interpretability in a generic case,[22] what interpretability actually means and how it should be evaluated is much less ambiguous for compound-protein affinity prediction. So that explanations conform with scientific knowledge, human understanding, and drug-discovery needs, we define interpretability of affinity prediction as the ability to explain predicted affinity through underlying atomic interactions (or contacts). Specifically, atomic contacts of various types are known to constitute the physical basis of intermolecular interactions,[23] modeled in force fields

3

to estimate interaction energies,[24] needed to explain mechanisms of actions for drugs,[25,26] and relied upon to guide structure-activity research in drug discovery.[27,28] We emphasize that simultaneous prediction of affinity and contacts does not necessarily make the affinity predictors intrinsically interpretable unless predicted contacts form the basis for predicted affinities.

How to assess interpretability for affinity prediction. Once interpretability of affinity predictors is defined through atomic contacts, it can be readily assessed against ground truth known in compound-protein structures, which overcomes the barrier for interpretable machine learning without ground truth.[29] In our study, we have curated a dataset of compound-protein pairs, all of which are labeled with K_d values and some of which with contact details; and we have split them into training, test, compound-unique, protein-unique, and both-unique (or double-unique) sets. We measure the accuracy of contact prediction over various sets using area under the precision-recall curve (AUPRC) which is suitable for binary classification (contacts/non-contacts) with imbalanced classes (far less contacts than non-contacts). We have performed large-scale assessments of attention mechanisms in various molecular data representations (protein amino-acid sequences and structure-property annotated sequences[10] as well as compound SMILES and graphs) and corresponding neural network architectures (convolutional and recurrent neural networks [CNN and RNN] as well as graph convolutional and isomorphism networks [GCN and GIN]). And we have found that current attention mechanisms inadequate for interpretable affinity prediction, as their AUPRCs were merely 50% more than chance.

How to enhance interpretability for affinity prediction. We have made three main contributions to enhance interpretability.

The first contribution, found to be the most impactful, is to design intrinsic explainability into the architecture of a deep "relational" network. Inspired by physics, we explicitly model and learn various types of atomic interactions (or "relations") through deep neural networks and embed attentions at the levels of residue-atom pairs and relation types. This was

4

motivated by relational neural networks first introduced to learn to reason in computer vision[30,31] and subsequent interaction networks to learn the relations and interactions of complex objects and their dynamics.[32,33] Moreover, we combine such deep relational modules in hierarchy to progressively focus attention from putative protein surfaces, binding-site k-mers and residues, to putative residue-atom binding pairs.

The second contribution is to incorporate physical constraints into data representations, model architectures, and model training. (1) To respect the sequence nature of protein inputs and to overcome the computational bottlenecks of RNNs, inspired by protein folding principles, we represent protein sequences as hierarchical k-mers and model them with hierarchical attention networks (HANs). (2) To respect the structural contexts of proteins, we predict from protein sequences solvent exposure over residues and contact maps over residue pairs; and we introduce novel structure-aware regularizations for structured sparsity of model attentions.

The third contribution is to supervise attentions with partially available contact data and train models accordingly. For interpretable and accurate affinity prediction, we have formulated a hierarchical multi-objective optimization problem where contact predictions form the basis for affinity prediction. We utilize contact data available to a minority (around 7.5%) of training compound-protein pairs and design hierarchical training strategies accordingly.

The rest of the paper is organized as follows. The aforementioned contributions in defining, measuring, and enhancing intepretable affinity prediction will be detailed in Methods. In Results, compared to state-of-the-art first, the resulting framework of DeepRelations is found to drastically boost interpretability robustly over default test, protein-unique, compound-unique, and double-unique sets, without sacrificing accuracy. Ablation studies then reveal the most contributing methodological contribution — the intrinsically explainable model architecture of our deep "relational" networks. Case studies provide further insights into the pattern of interpreted contacts.

Methods

Toward genome-wide prediction of compound-protein interactions (CPI), we assume that proteins are only available in 1D amino-acid sequences, whereas compounds are available in 1D SMILES or 2D chemical graphs. We start the section with the curation of a dataset of compound-protein pairs with known pK_d values, a subset of which is of known intermolecular contacts. We will introduce the state-of-the-art and our newly-adopted neural networks to predict from such molecular data. These neural networks will be first adopted in our previous framework of DeepAffinity[10] (supervised learning with joint attention) so that the interpretability of attention mechanisms can be systematically assessed in CPI prediction. We will then describe our physics-inspired, intrinsically explainable architecture of deep relational networks where aforementioned neural networks are used as basis models. With carefully designed regularization terms, we will explain multi-stage deep relational networks that increasingly focus attention on putative binding-site k-mers, binding-site residues, and residue-atom interactions, for the prediction and interpretation of compound-protein affinity. We will also explain how the resulting model can be trained using compound-protein pairs with affinity values but not necessarily with atomic interaction details.

Curation of a CPI Relational Benchmark Set

We have previously curated affinity-labeled compound-protein pairs[10] based on BindingDB.[34] The compound data were in the format of canonical SMILES as provided in PubChem[35] and the protein data are in the format of FASTA sequences. In this study, we used those data with amino-acid sequences no more than 300 from the pK_d-labeled set,[10] which corresponds to 1,926 compound-protein pairs. We also converted SMILES to graphs with RDKit.[36]

The pK_d-labeled data only shows the affinity strength between proteins and compounds, but it lacks the details on where and how the pairs interact. We have thus curated a subset of the pK_d-labeled data with atomic-level intermolecular contacts (or "relations") derived from

compound-protein co-crystal structures in PDB,[37] as ground truth for the interpretablity of affinity prediction. Specifically, we cross-referenced aforementioned compound-protein pairs in PDBsum[38] and used its LigPlot service to collect high-resolution atomic contacts or relations. These relations are given in the form of contact types (hydrogen bond or hydrophobic contact), atomic pairs, and atomic distances.

The resulting dataset of 1,926 pK_d-labeled compound-protein pairs (including 144 pairs with atomic-contact data) corresponds to 137 proteins and 1,376 compounds. We randomly split them into four folds where fold 1 do not overlap with fold 2 in compounds, do not do so with fold 3 in proteins, and do not do so with fold 4 in either compounds or proteins. Folds 2, 3, and 4 are referred to as compound-unique, protein-unique, and double unique sets for generalization tests; and they contain 201(19), 191(14), and 192(10) compound pairs (including those with contact details in the parentheses). Fold 1 was randomly split into training (70%) and test (30%) sets where 10% of the training set was set aside as the validation or development set. The training (including validation) and test sets contain 974(74) and 368(27) compound-protein pairs (with contact details). The split of the whole dataset is illustrated in Figure 1 below.

Although monomer structures of proteins are often unavailable, their structural features can be predicted from protein sequences alone with reasonable accuracy. We have predicted the secondary structure and solvent accessibility of each residue using the latest SCRATCH[39,40] and contact maps for residue pairs using RaptorX-contact.[41] These data provide additional structural information to regularize our machine learning models. If protein structures are available, actual rather than predicted such data can be used instead.

Data Representation and Corresponding Basis Neural Networks

Baseline: CNN and RNN for 1D protein and compound sequences.

When molecular data are given in 1D sequences, these inputs are often processed by convolutional neural networks (CNN)[11,42] and by recurrent neural networks (RNN) that are more

7

Figure 1: The complete data set consists of training, test, compound-unique, protein-unique, and double unique sets with compound-protein counts provided (including those with contact details in parentheses).

suitable for sequence data with long-term interactions.[10]

Challenges remain in RNN for compound strings or protein sequences. For compounds in SMILES strings, the descriptive power of such strings can be limited. In this study, we overcome the challenge by representing compounds in chemical formulae (2D graphs) and using two types of graph neural networks (GNN). For proteins in amino-acid sequences, the often-large lengths demand deep RNNs that are hard to be trained effectively (gradient vanishing or exploding and non-parallel training).[43] We previously overcame the second challenge by predicting structure properties from amino-acid sequences and representing proteins as a much shorter structure property sequences where each 4-letter tuple corresponds to a secondary structure.[10] This treatment however limits the resolution of interpretability to be at the level of protein secondary structures (multiple neighboring residues) rather than individual residues. In this study, we overcome the second challenge while achieving residue-level interpretability by using biologically-motivated hierarchical RNN (HRNN).

8

Proposed: GCN and GIN for 2D compound graphs.

Compared to 1D SMILES strings, chemical formulae (2D graphs) of compounds have more descriptive power and are increasingly used as inputs to predictive models.[10,12–14,44] In this study, compounds are represented as 2D graphs in which vertices are atoms and edges are covalent bonds between atoms. Suppose that n is the maximum number of atoms in our compound set (compounds with smaller number of atoms are padded to reach size n). Let's consider a graph $G = (\mathcal{V}, \mathcal{X}, \mathcal{E}, \mathcal{A})$, where $\mathcal{V} = \{v_j\}_{j=1}^n$ is the set of n vertices (each with d_g features), $\mathcal{X} \in R^{n \times d_g}$ that of vertex features, $\mathcal{E}$ that of edges, and $\mathcal{A} \in \{0,1\}^{n \times n}$ is unweighted symmetric adjacency matrix. Let $\hat{\mathcal{A}} = \mathcal{A} + \mathcal{I}$ and $\hat{\mathcal{D}}$ be the degree matrix (the diagonals of $\hat{\mathcal{A}}$).

We used Graph Convolutional Network (GCN)[45] and Graph Isomorphism Network (GIN)[46] which are the state of art for graph embedding and inference. GCN consists of multiple layers and at layer l the model can be written as:

$$\mathcal{H}^{(l)} = \text{ReLU}(\hat{\mathcal{D}}^{-\frac{1}{2}} \hat{\mathcal{A}} \hat{\mathcal{D}}^{-\frac{1}{2}} \mathcal{H}^{(l-1)} \Theta^{(l)}), \tag{1}$$

where $\mathcal{H}^{(l)} \in R^{n \times d_g^{(l)}}$ is the output, $\Theta^{(l)} \in R^{d_g^{(l-1)} \times d_g^{(l)}}$ the trainable parameters, and $d_g^{(l)}$ the number of features, all at layer l. Initial conditions (when $l = 0$) are $\mathcal{H}^{(0)} = \mathcal{X}$ and $d_g^{(0)} = d_g$.

GIN is the most powerful graph neural network in theory: its discriminative or representational power is equal to that of the Weisfeiler-Lehman graph isomorphism test.[47] Similar to GCN, GIN consists of multiple layers and at layer l the model can be written as a multi-layer perceptron (MLP):

$$\mathcal{H}^{(l)} = \text{MLP}^{(l)}(\bar{\mathcal{A}}^{(l)} \mathcal{H}^{(l-1)}), \tag{2}$$

where $\bar{\mathcal{A}}^{(l)} = \mathcal{A} + \epsilon^{(l)} \mathcal{I}$, $\epsilon^{(l)}$ can be either a trainable parameter or a fixed hyper-parameter. Each GIN layer has several nonlinear layers compared to GCN layer with just a ReLU per layer, which might improve predictions but suffer in interpretability.

The final representation for a compound is $\mathcal{Y} = \mathcal{H}^{(L)}$ if GCN or GIN has L layers. In

this study, vertex features are as in,[14] with few additional features detailed later for physics-inspired relational modules.

Proposed: HRNN for 1D protein sequences.

We aim to keep the use of RNN that respects the sequence nature of protein data and mitigate the difficulty of training RNN for long sequences. To that end, inspired by the hierarchy of protein structures, we model protein sequences using hierarchical attention networks (HANs). Specifically, during protein folding, sequence segments may fold separately into secondary structures and the secondary structures can then collectively pack into a tertiary structure needed for protein functions. We exploit such hierarchical nature by representing a protein sequence of length easily in thousands as tens or hundreds of k-mers (consecutive sequence segments) of length k ($k = 15$ in this study). Accordingly we process the hierarchical data with hierarchical attention networks (HANs)[48] which have been proposed for natural language processing. We also refer to it as hierarchical RNN (HRNN).

Given a protein sequence $\mathbf{x}$ with maximum length m (shorter sequences are padded to reach length m) partitioned into T groups of k-mers, we use two types of RNNs (specifically, LSTMs here) in hierarchy for modeling within and across k-mers. We first use an embedding layer to represent the ith residue in t^{th} k-mer as a vector $\mathbf{x}_{it}$. And we use a shared LSTM for all k-mers for the latent representation of the residue: $\mathbf{h}_{it} = \text{LSTM}(\mathbf{x}_{it})$ ($t = 1, \ldots, T$). We then summarize each k-mer as $\mathbf{k}_t$ with an intra-k-mer attention mechanism:

$$
\begin{aligned}
u_{it} &= \mathbf{v_1}\tanh(\Theta_1 \mathbf{h}_{it} + \mathbf{b_1}) \ \ \forall\, i, t \\
u'_{it} &= \frac{\exp(u_{it})}{\sum_{i'} \exp(u_{i't})} \ \ \forall\, i, t \\
\mathbf{k}_t &= \sum_i u'_{it} \mathbf{h}_{it} \ \ \forall\, t
\end{aligned}
\tag{3}
$$

Then we use another LSTM for for $\mathbf{k}_t$ and reach $\mathbf{h}_t = \text{LSTM}(\mathbf{k_t})$ ($t = 1, \ldots, T$).

The final representation for a protein sequence is the collection of $\mathbf{h}_t$.

10

Joint attention over protein-compound atomic pairs for interpretability.

Once the representation of protein sequences ($\mathbf{h}_t$ where $t = 1, \ldots, T$ is the index of protein k-mer) and that of compound sequences or graphs ($\mathbf{y}_j$ where $j = 1, \ldots, n$ is the index of compound atom) are defined, they are processed with a joint k-mer–atom attention mechanism to interpret any downstream prediction:

$$
\begin{aligned}
N_{tj} &= \tanh(\mathbf{h}_t \Theta_2 \mathbf{y}_j) \ \ \forall \, t, j \\
\mathcal{W}'_{tj} &= \frac{\exp(\mathbf{N}_{tj})}{\sum_{t',j'} \exp(\mathbf{N}_{t'j'})} \ \ \forall \, t, j
\end{aligned}
\tag{4}
$$

With $\mathcal{W}'_{tj}$, the joint attention between the tth k-mer and the jth atom, we can combine it with the intra-k-mer attention over each residue i in the tth k-mer and reach $\mathcal{W}_{ij}$, the joint attention between the ith protein residue and the jth compound atom:

$$
\mathcal{W}_{ij} = u'_{it} \mathcal{W}'_{tj} \ \ \forall \, i, j
\tag{5}
$$

This joint attention mechanism is an extension of our previous work[10] where a protein sequence was represented as a single, "flat" RNN rather than multiple, hierarchical RNNs.

DeepRelations

Overall architecture.

We have developed an end-to-end "by-design" interpretable architecture named DeepRelations for joint prediction and interpretation of compound-protein affinity. The overall architecture is shown in Figure 2.

There are three relational modules (Rel-CPI) corresponding to three stages. Their attentions are trained to progressively focus on putative binding k-mers, residues, and pairs; and earlier-stage attentions guide those in the next stage through regularization. In each Rel-CPI module, there are six types of atomic "relations" or interactions (including elec-
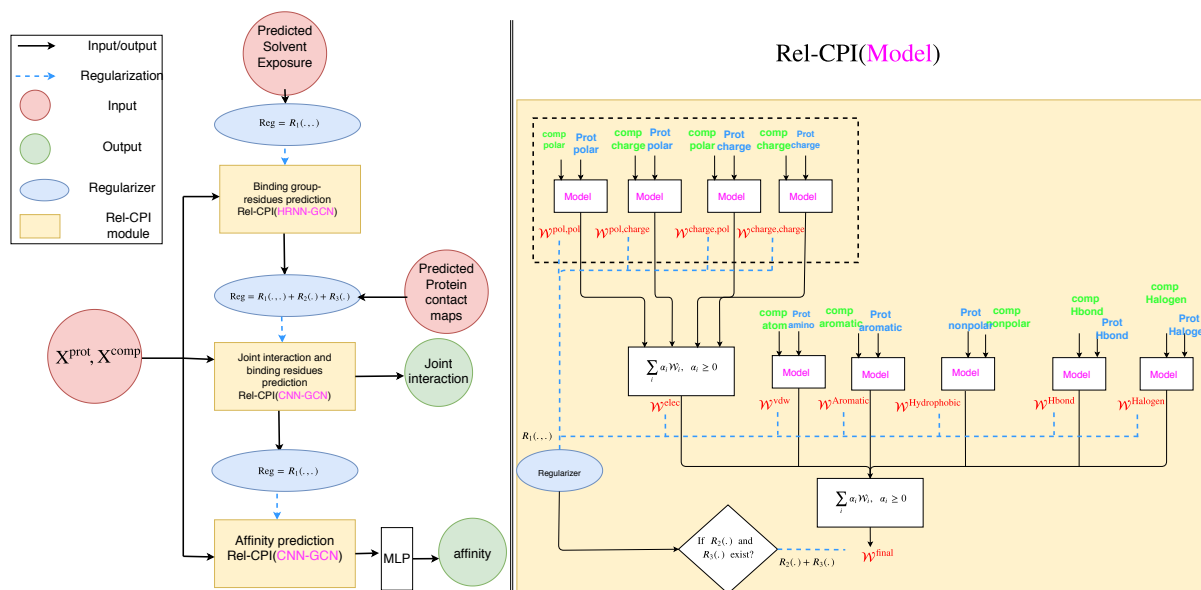
11

Figure 2: Schematic illustration of DeepRelations, an intrinsically explainable neural network architecture for predicting compound-protein interactions.

trostaics as the non-negative linear combination of four sub-types). And each (sub)type of relation is modelled by aforementioned neural network pairs with joint attentions. For instance, the first Rel-CPI uses HRNN-GCN (HRNN for protein sequences and GCN for compound graphs) and the next two use CNN-GCN (dilated causal CNN for proteins and GCN for compounds). And the non-negative linear combination of six individual relations' attention matrices $\mathcal{W}_i$ (where i corresponds to the six relation types) gives the overall joint attention matrix $\mathcal{W}_{\text{final}}$ (or $\mathcal{W}$ in short) in each module.

Physics-inspired relational modules

The relational modules are inspired by physics. Specifically, atomic "relations" or interactions constitute the physical bases and explanations of compound-protein interaction affinities and are often explicitly modelled in force fields. We have considered the following six types relations with attentions paid on and additional input data defined for.

- *Electrostatic interactions*: A non-negative linear combination of four subtypes of compound-protein interactions through attentions: 1) charge-charge, 2) charge-dipole, 3) dipole-

12

charge, and 4) dipole-dipole interactions. The input feature for the charge of a protein residue or a compound atom is the CHARMM27 parameter and the atomic formal charge, respectively. That for the dipole of a protein residue or a compound atom is the residue being polar/nonpolar or the Gasteiger atomic partial charge.

- *Hydrogen bond*: Non-covalent interaction between an electronegative atom as a hydrogen "acceptor" and a hydrogen atom that is covalently bonded to an electronegative atom called a hydrogen "donor". Therefore, if a protein residue or compound atom could provide a hydrogen acceptor/donor, its hydrogen-bond feature is -1/+1; otherwise the feature value is 0. A protein residue is allowed to be both hydrogen-bond donor and acceptor.

- *Halogen bond*: A halogen bond is very similar to hydrogen bond except that a halogen (rather than hydrogen) atom (often found in drug compounds) is involved in such interactions. If a protein residue or a compound atom has/is a halogen atom such as iodine, bromine, chlorine and fluorine, its halogen-bond feature is assigned +4, +3, +2 and +1, respectively, for decreasing halogen-bonding strength. If it can be a halogen acceptor, the feature is -1. If it can be neither, the feature is simply set at 0.

- *Hydrophobic interactions*: The interactions between hydrophobic protein residues and compound atoms contribute significantly to the binding energy between them. If a protein residue is hydrophobic, it is represented as 1 and otherwise as 0. Moreover, the non-polar atoms are represented as 1 and the polar one as -1.

- *Aromatic interactions*: Aromatic rings in tryptophans, phenylalanines, and tyrosines participate in "stacking" interactions with aromatic moieties of a compound (π-π stacking). Therefore, if a protein residue has an aromatic ring, its aromatic feature is set at 1 and otherwise at 0. Similarly, if a compound atom is part of a ring, the feature is set at 1 and otherwise at 0.

13

- *VdW interactions*: Van der Waals are weaker interactions compared to others. But the large amount of these interactions contribute significantly to the overall binding energy between a protein and a compound. We consider the amino-acid type and the atom element as their features and use an embedding layer to derive their continuous representations.

For each (sub)type of atomic relations, corresponding protein and compound features are fed into basis neural network models such as HRNN for protein sequences and GNN for compound graphs. All features are made available to baseline methods (DeepAffinity+ variants) as well for fair comparison.

Physical constraints as regularization.

The joint attention matrices $\mathcal{W}$ in each Rel-CPI module, for individual relations or overall, are regularized with the following two types of physical constraints.

Focusing regularization   In the first regularization, a constraint input is given as a matrix $\mathcal{T} \in [0, 1]^{m \times n}$ to penalize the attention matrices $\mathcal{W}_i$ for all the 10 (sub)types of relations if they focus on the undesired regions of proteins. In addition, an L1 sparsity regularization is on the attention matrices $\mathcal{W}_i$ for all relations to promote interpretability as a small portion of protein residues interact with compounds. Therefore, this "focusing" penalty can be formalized as:

$$R_1(\mathcal{W}) = \lambda_{\text{relation}} \sum_{i=1}^{10} ||(\mathbf{1} - \mathcal{T}) \odot \mathcal{W}_i||_2 + \lambda_{\text{L1}} \sum_{i=1}^{10} ||\mathcal{W}_i||_1, \qquad (6)$$

where the $\mathcal{T}$ term, a parameter, can be considered as soft thresholding and its penalty only incurs when $\mathcal{T}_{ij} = 0$.

The first regularization is used for all three Rel-CPI modules or stages with increasingly focusing $\mathcal{T}$. Let $\mathcal{T}^{[k]}$ be the constraint matrix and $\mathcal{W}^{[k]}$ the learned attention matrix for a

given relation in the kth stage. In the first stage, $\mathcal{T}_{ij}^{[1]}$ is one only for any residue i predicted to be solvent-exposed in order to focus on surfaces. In the second stage, $\mathcal{T}_{ij}^{[2]} = \max_{j'} \mathcal{W}_{ij'}^{[1]}$ to focus on putative binding residues hierarchically learned for k-mers and residues in stage 1. In the last stage, $\mathcal{T}_{ij}^{[3]} = \mathcal{W}_{ij}^{[2]}$ focuses on putative contacts between protein residues and compound atoms. The focusing regularization is enforced on attentions for every relation (sub)type in the current implementation and can be done only for given (sub)types in future.

Structure-aware sparsity regularization over protein contact maps We further develop a structure aware sparsity constraints based on known or RaptorX-predicted contact maps of the unbound protein. As sequentially distant residues might be close in 3D and form binding sites for compounds, we define overlapping groups of residues where each group consists of a residue and its spatially close neighboring residues. Just in the second stage, we introduce Group Lasso for spatial groups and the Fused Sparse Group Lasso (FSGL) for sequential groups on the overall, joint attention matrix $\mathcal{W}$:

$$R_2(\mathcal{W}) = \lambda_{\text{group}}||\mathcal{W}||_{\text{group}} + \lambda_{\text{fused}}||\mathcal{W}||_{\text{fused}} + \lambda_{\text{L1-overall}}||\mathcal{W}||_1. \tag{7}$$

The group Lasso penalty will encourage a structured group-level sparsity so that few clusters of spatially close residues share similar attentions within individual clusters. The fused sparsity will encourage local smoothness of the attention matrix so that sequentially close residues share similar attentions with compound atoms. The L1 term maintains the sparsity of the overall attention matrix $\mathcal{W}$, since the L1 sparsity of attention matrices $\mathcal{W}_i$ for individual relations do not guarantee that their linear combination remains sparse.

Supervised attention.

It has been shown in visual question answering that attention mechanisms in deep learning can differ from human attentions.[21] As will be revealed in our results, they do not necessarily focus on actual atomic interactions (relations) in compound-protein interactions either.

We have thus curated a relational subset of our compound-protein pairs with affinities, for which known ground-truth atomic contacts or relations are available. We summarize actual contacts of a pair in a matrix $\overline{\mathcal{W}}^{\text{true}}$ of length $\overline{m} \times \overline{n}$ ($\overline{m}$ and $\overline{n}$ are actual numbers of protein residues and compound atoms, respectively, for a given pair), which is a binary pairwise interaction matrix normalized by the total number of nonzero entries. We have accordingly introduced a third regularization term to supervise $\overline{\mathcal{W}}$, the non-padded submatrix of attention matrix $\mathcal{W}$, in the second stage:

$$R_3(\mathcal{W}) = \lambda_{\text{bind}} \ c \ ||\overline{\mathcal{W}} - \overline{\mathcal{W}}^{\text{true}}||_F^2, \tag{8}$$

where c is a normalization constant across batches. Suppose that, in any batch, a given pair's actual interaction matrix is of size $m_1 \times n_1$ and the smallest such size across all batches is $m_{\min} \times n_{\min}$. Then $c = \frac{m_{\min} \times n_{\min}}{m_1 \times n_1}$ for this pair.

Training strategy for hierarchical multi-objectives

Accuracy and interpretability are the two objectives we pursue at the same time. In our case, the two objectives are hierarchical: compound-protein affinity originates from atomic-level interactions (or "relations") and better interpretation in the latter potentially contributes to better prediction of the former.

Challenges remain in solving the hierarchical multi-objective optimization problem. First, optimizing for both objectives simultaneously (for instance, through weighted sum of them) does not respect that the two objectives do no perfectly align with each other and are of different sensitivities to model parameters. Second, ground-truth data for interpretability of affinity prediction, i.e., compound-protein contacts, is rare. In fact, merely 7.5% of our compound-protein pairs labeled with K_d affinities are with contact data.

To overcome the aforementioned challenges, we consider the problem as multi-label machine learning facing missing labels. And we design hierarchical training strategies to solve

16

the corresponding hierarchical multi-objective optimization problem. The whole DeepRelations model, including the three Rel-CPI modules, are trained end-to-end.[49] In the first stage, we "pre-trained" DeepRelations to minimize mean squared error (MSE) of pK_d regression alone, with physical constraints turned on; in other words, attentions were regularized (through $R_1(\cdot)$ and $R_2(\cdot)$) but not supervised in this stage. We tuned combinations of all hyperparameters except λ_{bind} in the discrete set of $\{10^{-5}, 10^{-4}, \ldots, 10^{-1}\}$, with 400 epochs at the learning rate of 0.001. Over the validation set, we recorded the lowest RMSE for affinity prediction and chose the hyperparameter combination with the highest AUPRC for contact prediction subjective to that the corresponding affinity RMSE (root mean square error) does not deteriorate from the lowest by more than 10%.

In the second stage, with the optimal values of all hyperparameters but λ_{bind} fixed, we loaded the corresponding optimized model in the first stage and "fine-tuned" the model to minimize MSE additionally regularized by supervised attentions (through $R_1(\cdot)$, $R_2(\cdot)$, and $R_3(\cdot)$). As only 7.5% training examples are with known contacts, we used the their average and ignored the other examples for $R_3(\cdot)$ in each batch. We used a slower learning rate (0.0001) and less training epochs (200) in the fine-tuning stage; and we tuned λ_{bind} in the set of $\{10^{-1}, 10^{-4}, \ldots, 10^{-5}\}$ following the same strategy as in the pre-training stage.

In the end, we chose $\lambda_{\text{relation}} = 10^{-3}$, $\lambda_{\text{L1}} = 10^{-5}$, $\lambda_{\text{group}} = 10^{-2}$, $\lambda_{\text{fused}} = 10^{-3}$, $\lambda_{\text{L1-overall}} = 10^{-5}$ and $\lambda_{\text{bind}} = 10^1$ for DeepRelations.

We did similarly for hyper-parameter tuning while constraining (and supervising) attentions to make DeepAffinity+ variants. For HRNN-GCN_cstr (modeling protein sequences with HRNN and compound graph with GCN, regularized by physical constraints in $R_2(\cdot)$), we chose $\lambda_{\text{group}} = 10^{-5}$, $\lambda_{\text{fused}} = 10^{-4}$, and $\lambda_{\text{L1-overall}} = 10^{-5}$; and for its supervised version HRNN-GCN_cstr_sup, the additional $\lambda_{\text{bind}} = 10^4$. For HRNN-GIN_cstr (modeling protein sequences with HRNN and compound graph with GIN, regularized by physical constraints in $R_2(\cdot)$), we chose $\lambda_{\text{group}} = 10^{-3}$, $\lambda_{\text{fused}} = 10^{-4}$, and $\lambda_{\text{L1-overall}} = 10^{-4}$; and for its supervised version HRNN-GIN_cstr_sup, the additional $\lambda_{\text{bind}} = 10^4$. $R_1(\cdot)$ was for attentions on indi-

vidual relations in DeepRelations and not applicable for DeepAffinity+ variants, although a surface-focusing regularization on overall attentions could be introduced.

Results

Attentions alone are inadequate for interpreting compound-protein affinity prediction

Our first task is to systematically assess the adequacy of attention mechanisms for interpreting model-predicted compound-protein affinities. To that end, we adopt various data representations and corresponding state-of-the-art neural network architectures in our framework of DeepAffinity. To model proteins, we have adopted RNN using protein SPS[10] as input data as well as CNN and newly developed HRNN using protein amino-acid sequences. To model compounds, we have adopted RNN using SMILES as input data as well as GCN and GIN using compound graphs with node features and edge adjacency.[14] In the end, we have tested six DeepAffinity variants for protein-compound pairs, including RNN-RNN, RNN-GCN, CNN-GCN, HRNN-RNN, HRNN-GCN, and HRNN-GIN. The first two (RNN-RNN and RNN-GCN), where protein SPS sequences are modeled by RNN and compound SMILES or graphs are modeled by RNN or GCN, are essentially our previous models[10] except that no unsupervised pretraining is used in this study. Whereas these two models' attentions on proteins are at the secondary structure levels (thus not assessed for interpretability here), the rest have joint attentions at the level of pairs of protein residues and compound atoms.

The accuracy of affinity prediction, measured by RMSE (root mean squared error) in pK_d, is summarized for the DeepAffinity variants in the top panel of Figure 3. Overall, all variants have shown pK_d error between 1.1 and 1.3, a level competitively comparable even to the state-of-the-art affinity predictor using compound-protein co-crystal structures.[7] These models have robust accuracy profiles across the default, compound-unique, protein-unique, and double-unique test sets, suggesting their generalizability beyond training compounds or

18

proteins. Modeling compound SMILES with RNN seems to have slightly worse performance compared to modeling compound graphs with GCN or GIN, although less features are used for SMILES strings compared to node features for compound graphs.
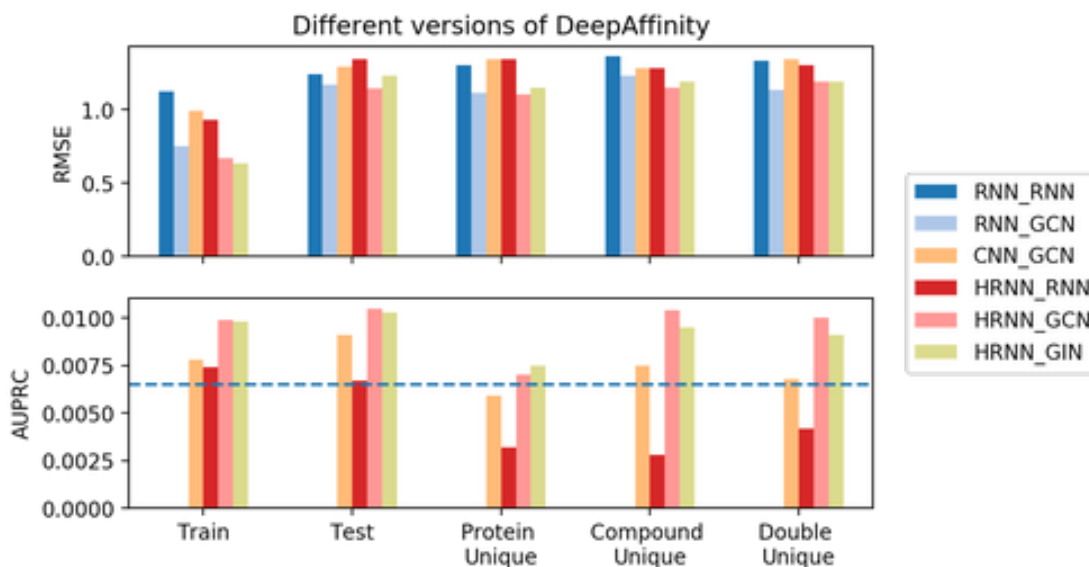

Figure 3: Comparing accuracy and interpretability among various versions of DeepAffinity with (unsupervised) joint attention mechanisms. Separated by underscores in legends are neural network models for proteins and compounds respectively.

The interpretability of affinity prediction is assessed against ground truth of contacts, as in the bottom panel of Figure 3. Specifically, we use joint attention scores to classify all possible residue-atom pairs into contacts or non-contacts. As contacts only represent a tiny portion (0.0061 ± 0.0023 in our dataset) of all possible pairs, we use the area under the precision-recall curve (AUPRC), instead of the area under the receiver operating characteristic curve (AUROC), to assess such binary classification. Here AUPRC is averaged over all pairs involved in the corresponding set. Interestingly, compared to chance (AUPRC=0.0061), modeling protein amino-acid sequences through CNN or modeling compound SMILES through RNN had comparable or even worse contact prediction (or interpretability here). Modeling protein amino-acid sequences through hierarchical RNN and compound graphs as GNN (GCN or GIN here) would improve the AUPRC by around 50% for all test sets except the protein-unique one. However, even with 50% relative improvement,

the absolute accuracy level of contact prediction, or the absolute level of model interpretability, remains low (AUPRC around 0.01). Moreover, unlike the case of affinity accuracy, the interpretability results have shown some sensitivity to training data, especially when the test proteins are not contained in the training set.

From the results above, we conclude that attention mechanisms alone are inadequate for the interpretability of compound-protein affinity predictors, regardless of the choice of commonly used, generic neural network architectures.

Regularizing attentions with physical constraints modestly improves interpretability.

Our next task is to enhance the interpretability of compound-protein affinity prediction beyond the level achieved by attention mechanisms alone. The first idea is to incorporate domain-specific physical constraints into model training. The rationale is that, by bringing in the (predicted) structural contexts of proteins and protein-compound interactions, attentions can be guided in their sparsity patterns accordingly for better interpretability.

We start with the two best-performing DeepAffinity variants so far (HRNN-GCN and HRNN-GIN) where protein amino-acid sequences are modeled by hierarchical RNN and compound graphs by various GNNs (including GCN and GIN). And we introduce structure-aware sparsity regularization $R_2(\cdot)$ to the two models to make "DeepAffinity+" variants. The resulting HRNN-GCN_cstr and HRNN-GIN_cstr models with physical constraints are assessed in Figure 4. Compared to the the non-regularized counterparts in Figure 3, both models achieved similar accuracy levels across various test sets for affinity prediction, but their interpretability improved. Specifically, HRNN-GCN after constraints, compared to that before constraints, had AUPRC improvement of 5.7%, 2.9%, 19.2%, and 20.0% for default test, protein-unique, compound-unique, and double-unique sets, respectively. However, the interpretability improvements from physical constraints were modest especially when the absolute level of AUPRC remained around 0.01. These results suggest that incorporating

20

physical constraints to structurally regularize the sparsity of attentions is useful for improving interpretability but may not be enough.
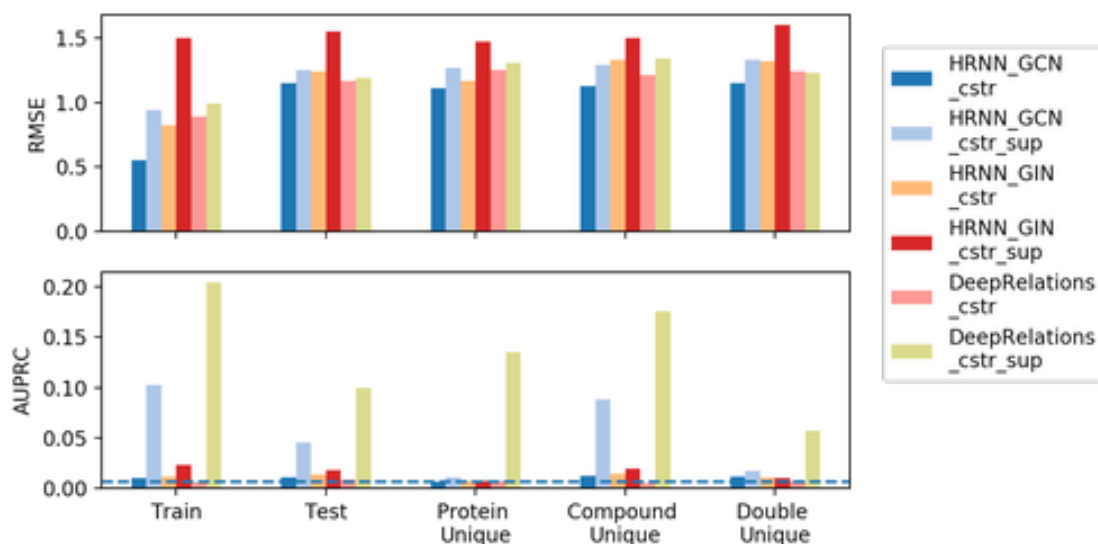

Figure 4: Comparing accuracy and interpretability among various versions of DeepAffinity+ (DeepAffinity with regularized and supervised attentions) and DeepRelations. "cstr" in legends indicates physical constraints imposed on attentions through regularization term $R_2(\cdot)$, whereas "sup" indicates supervised attentions through regularization term $R_3(\cdot)$.

Supervising attentions significantly improves interpretability.

As regularizing attentions with physical constraints was not enough to enhance interpretability, our next idea is to additionally supervise attentions with ground-truth contact data available to some but not all training examples. Again we introduce "DeepAffinity+" models starting with HRNN-GCN and HRNN-GIN, by both regularizing and supervising attentions (using $R_2(\cdot)$ and $R_3(\cdot)$).

The performances of resulting HRNN-GCN_cstr_sup and HRNN-GIN_cstr_sup models are shown in Figure 4. Importantly, HRNN-GCN_cstr_sup (light blue) significantly improves interpretability of affinity prediction without the sacrifice of accuracy. The average AUPRC improved to 0.0455, 0.0106, 0.0883, and 0.0175 for the default test, protein-unique, compound-unique, and double-unique test sets, representing a relative improvement of 309%

21

(645%), 92% (73%), 600% (1347%), and 46% (186%), respectively, compared to the constrained counterparts (chance). Interestingly, supervising attentions in HRNN-GIN did not see as significant improvement in interpretability.

Building explainability into DeepRelations architecture further drastically improves interpretability.

Toward better interpretability, besides regularizing and supervising attentions, we have further developed an explainable, deep relational neural network named DeepRelations. Here atomic "relations" constituting physical bases and explanations of compound-protein affinities are explicitly modeled in the architecture with multi-stage gradual "zoom-in" to focus attentions. In other words, the model architecture itself is intrinsically explainable by design.

The superior performances of the resulting DeepRelations (with both regularized and supervised attentions) are shown in Figure 4 (yellow-green "DeepRelations_cstr_sup"). With equally competitive accuracy in affinity prediction as all previous models, DeepRelations achieved drastic improvements in interpretability. Strikingly, the average AUPRC further improved to 0.0996, 0.1350, 0.1754, and 0.0571 for the default test, protein-unique, compound-unique, and double-unique test sets, representing a relative improvement of 121% (1532%), 1173% (2113%), 98% (2775%), and 226% (836%), respectively, compared to the previous best DeepAffinity+ variant (chance).

We further assessed contact prediction (or interpretability) of DeepAffinity+ variants and DeepRelations using the precision, sensitivity, and odds ratio (or enrichment factor) of their top K predictions (where K ranged from 5 to 50). Figure 5 shows that DeepRelations drastically outperforms other methods in all assessment measures considered. The precision and sensitivity levels may not appear impressive, largely due to the very strict definition of "true contacts" in our study, as will be revealed in a case study. Note that all atomic-level contact predictions were made with the inputs of protein sequences and compound graphs alone.
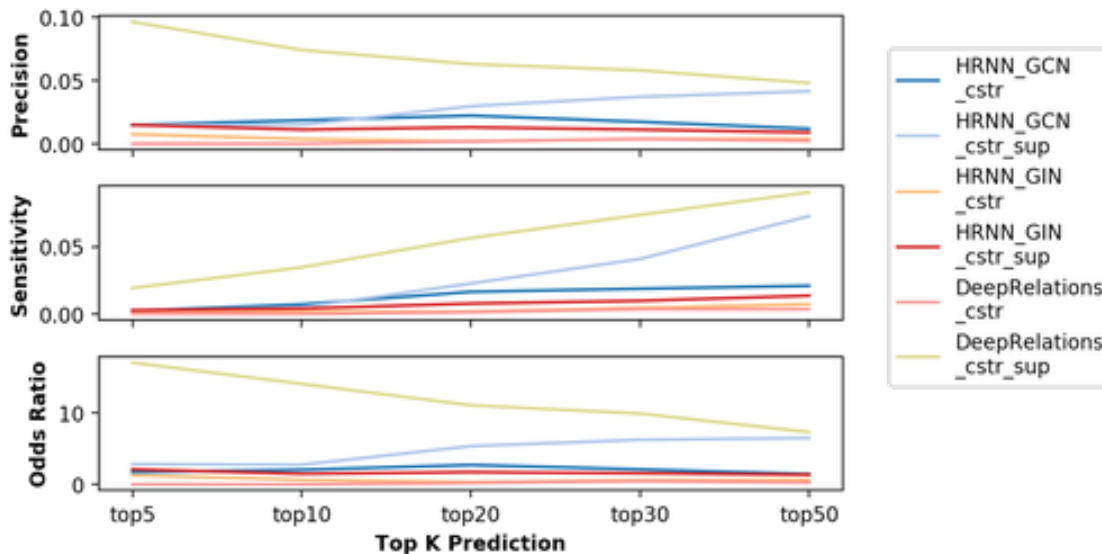
Figure 5: Comparing precision, sensitivity, and odds ratio (enrichment) of affinity-interpreting contacts predicted by various versions of DeepAffinity+ and DeepRelations.

For fair comparison, all DeepAffinity+ variants and DeepRelations were using the same set of features. A negative control experiment in the subsequent ablation study further validated this. Therefore, the architecture of DeepRelations, being intrinsically explainable, is the major contributor to its superior interpretability. From the machine learning perspective, DeepAffinity+ variants have various molecular features lumped into general-purpose neural networks, which makes it very hard to learn governing physics laws from the molecular affinity data. Instead, DeepRelations directly builds the physics laws into its model architecture and carefully structure various features into corresponding atomic relations and eventually the overall binding affinity.

Ablation study for DeepRelations

To disentangle various components of DeepRelations and understand their relative contributions to DeepRelations' superior interpretability, we removed components from DeepRelations and made "DeepRelations-" variants. Besides regularized and supervised attentions, we believe that the main contributions in the architecture itself are (1) the multi-stage "zoom-

in" mechanisms that progressively focus attentions from surface, binding k-mers, binding residues to binding residue-atom pairs; and (2) the explicit modeling of atomic relations that can explain the structure feature-affinity mappings consistently with physics principles. We thus made three DeepRelations- variants: DeepRelations without multi-stage focusing, without explicit atomic relations, or without both.

We compared the three intermediate "DeepRelations-" versions with the best DeepAffinity+ (regularized and supervised HRNN-GCN) and DeepRelations in Figure 6. Consistent with our conjecture, we found that, the explicit modeling of atomic relations was the main reason for DeepRelations' superior interpretability, as the removal of this component alone reduced the average AUPRC down to a similar level of the best DeepAffinity+ (except for the protein-unique case). Removing both components essentially reproduced the best Deep-Affinity+ (again, except that it still outperforms the latter in the protein-unique case), which served well as a "negative control" case here.
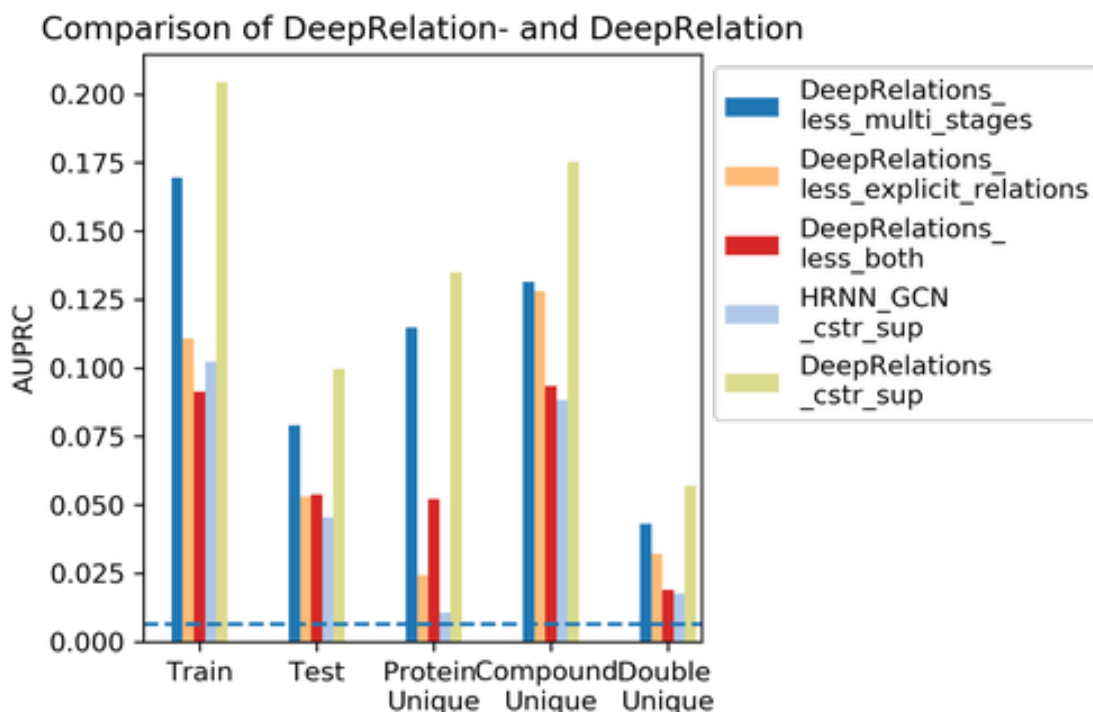

Figure 6: Comparing interpretability between DeepRelations and DeepRelations- (DeepRelations without multi-stage focusing, explicitly-modeled relations, or both).

Case Study

Now that we have established how drastically DeepRelations improves the interpretability of compound-protein affinity prediction and explained why it achieves so by design, we went on to examine the pattern in which DeepRelations contact prediction outperforms the best DeepAffinity variant HRNN-GCN (and leaves room for further improvement). We thus randomly chose a compound-protein test pair with known contacts for case study: carbonic anhydrase II inhibitor with its compound AL1 (PDB ID:1BNN).

As shown in Figure 7, DeepRelations (middle) not only made more correct contact predictions than HRNN-GCN (left) but also showed much improved contact *pattern*. In particular, HRNN-GCN could focus attention on residue-atom pairs that are actually as far as above 20Å away, and the attended residues could be dispersed at two sides of a protein. In contrast, DeepRelations predictions were correctly focused in the binding site of the protein and many of its "incorrect" predictions may correspond to residue-atom pairs within 10Å or less, which could be partially attributed to the physical constraints introduced as regularization.

To further examine the possible benefit of explicitly modeled atomic relations, we examined the overall attention matrix and found that the most contributions originate from electrostatic relations. We therefore examined the top-10 predicted electrostatic contacts according to the electrostatic attention matrix alone and found four true electrostatic interactions associated with the same protein residue (Hisidine 94).

We extended the analysis of the patterns of predicted contacts over all test cases. Considering that the true contacts are defined rather strictly, we assess distance-distributions of residue-atom pairs predicted by HRNN-GCN, HRNN-GCN with regularized attention, HRNN-GCN with regularized and supervised attention, and DeepRelations (also with regularized and supervised attention). As seen in Figure 8, DeepRelations outperforms competitors in all distance ranges over all test sets (except the 4Å~10Å range for the seemingly most challenging protein-unique case). Impressively, among top-50 contacts predicted by DeepRelations, around 33%, 39%, 27%, and 33% were actually within 10Åin the default
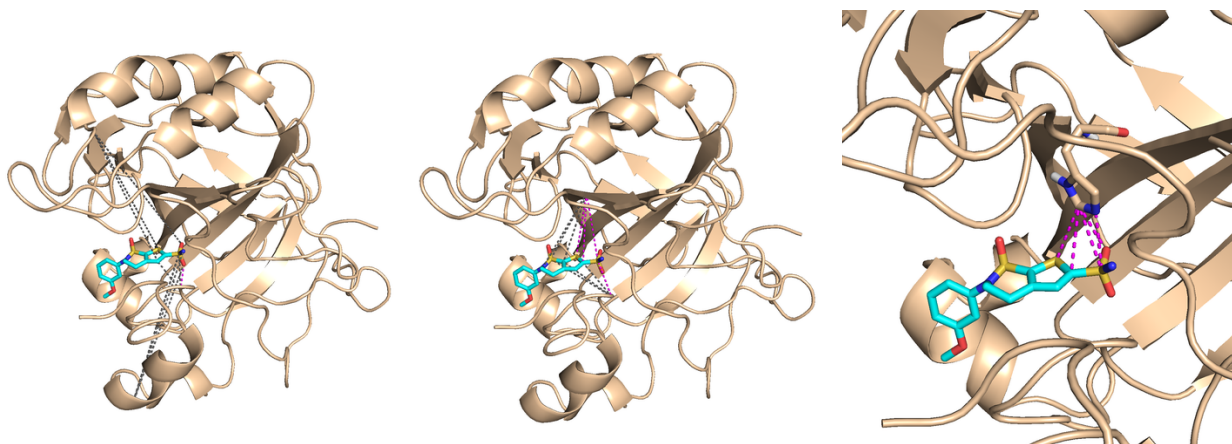
Figure 7: Structural visualization of top-10 intermolecular contacts predicted by HRNN-GCN (left) and DeepRelations (middle and right) for a test case. The protein (carbonic anhydrase II inhibitor) is shown in wheat cartoons and compound (AL1) in cyan sticks. Dashed lines represent the top-10 predicted relations (interactions between protein residues and compound atoms) in either case. Their magenta and gray colors indicate correct and incorrect relations/contacts, respectively, according to ground truth strictly defined as in PDBsum. In the right panel, we examined the dominant electrostaic attention matrix alone and found four true positives, all formed with Histidine 94 (wheat sticks), in the top-10 predicted electrostatic contacts.

test, compound-unique, protein-unique, and double-unique test sets, respectively.

Conclusions

Toward accurate and interpretable machine learning of compound-protein affinity, we have curated an affinity-labeled dataset with partially annotated contact details, assessed the adequacy of current attention-based deep learning models for both accuracy and interpretability, and developed novel machine-learning models and training strategies to drastically enhance interpretability without sacrificing accuracy. This is the first study with dedicated model development and systematic model assessment for interpretability in affinity prediction.

Our study has found that commonly-used attention mechanisms alone, although better than chance in most cases, are not satisfying in interpretability: the most attended contacts in affinity prediction do not reveal true contacts underlying affinities at a useful level. We have tackled the challenge with three innovative, methodological advances. First, we in-
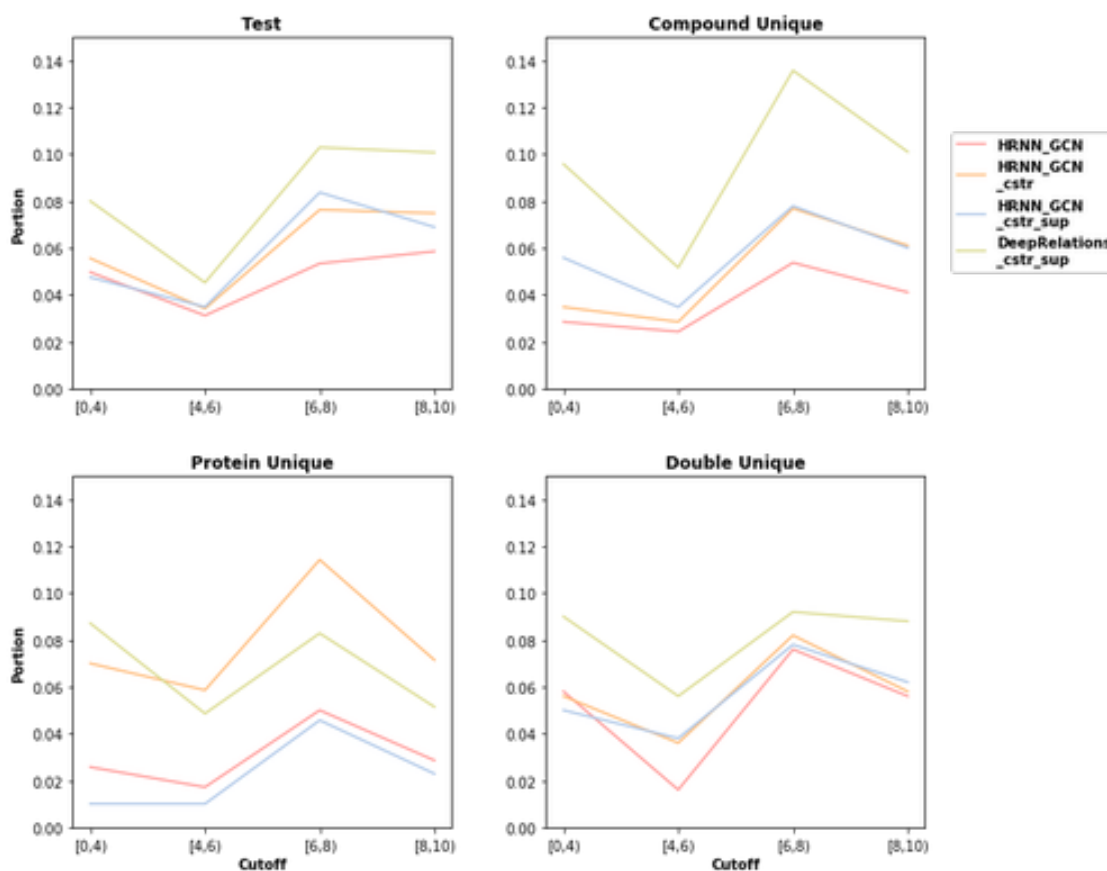
Figure 8: Distributions of top-50 contacts, predicted by DeepAffinity, DeepAffinity+, and DeepRelations, in various distance ranges (unit: Å).

troduce domain-specific physical constraints to regularize attentions (or guide their sparsity patterns), in which structural contexts such as sequence-predicted protein surfaces and protein contact maps are utilized. Second, we exploit partially available ground-truth contacts to supervise attentions. Lastly, we build intrinsically explainable model architecture where various atomic relations, reflecting physics laws, are explicitly modeled and aggregated for affinity prediction. Joint attentions are embedded over residue-atom pairs for individual and overall relations. And a multi-stage hierarchy, trained end-to-end, progressively focuses attentions on protein surfaces, binding k-mers and residues, and residue-atom contact pairs.

Empirical results demonstrate the superiority of DeepRelations in interpretability without sacrificing accuracy. Compared to the best DeepAffinity variant with joint attention (HRNN-GCN), the AUPRC for contact prediction was boosted to 9.48, 16.86, 19.28, and

27

5.71-fold for the default test, compound-unique, protein-unique, and double-unique cases. Importantly, the interpretability of DeepRelations proves robust and generalizable, as the margins of improvement were even higher when compounds or/and proteins are not present in the training set. Ablation studies demonstrate that the explainable relational network architecture was the major contributor to such performances. Case studies suggest that DeepRelations predict not only more correct but also more well-patterned contacts. And many "incorrect" predictions due to the strict definition of contacts were within reasonable ranges — in fact, around one third of the top-50 predicted contacts correspond to residue-atom pairs within 10Å.

An additional benefit of DeepRelations is its broad applicability toward the vast chemical and proteomic spaces. It does not rely on 3D structures of compound-protein complexes or even protein monomers when such structures are often unavailable. The only inputs needed are protein sequences and compound graphs. Meanwhile, it adopts the latest technology to predict structural contexts for protein sequences (such as surfaces, secondary structures, and residue-contact maps) and incorporatea such structural contexts into affinity and contact predictions. When structure data are available, DeepRelations can readily integrate such data by using actual rather than predicted structural contexts.

Our study demonstrates that, it is much more effective to directly build explainability into machine learning model architectures (as DeepRelations models underlying atomic relations explicitly) than to infer explainability from general-purpose architectures (as Deep-Relations variants learn attentions from data alone). In other words, designing intrinsically interpretable machine learning models, although more difficult, can be much more desired than pursuing interpretability in a *post hoc* manner.

Acknowledgement

28

Part of the computing time was provided by the Texas A&M High Performance Research Computing.

References

(1) Santos, R.; Ursu, O.; Gaulton, A.; Bento, A. P.; Donadi, R. S.; Bologa, C. G.; Karlsson, A.; Al-Lazikani, B.; Hersey, A.; Oprea, T. I., et al. A comprehensive map of molecular drug targets. *Nature reviews Drug discovery* **2017**, *16*, 19.

(2) Bohacek, R. S.; McMartin, C.; Guida, W. C. The art and practice of structure-based drug design: A molecular modeling perspective. *Medicinal Research Reviews* **1996**, *16*, 3–50.

(3) Ponomarenko, E. A.; Poverennaya, E. V.; Ilgisonis, E. V.; Pyatnitskiy, M. A.; Kopylov, A. T.; Zgoda, V. G.; Lisitsa, A. V.; Archakov, A. I. The Size of the Human Proteome: The Width and Depth. *Int J Anal Chem* **2016**, *2016*, 7436849.

(4) Aebersold, R. et al. How many human proteoforms are there? *Nat. Chem. Biol.* **2018**, *14*, 206–214.

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

(6) Gomes, J.; Ramsundar, B.; Feinberg, E. N.; Pande, V. S. Atomic convolutional networks for predicting protein-ligand binding affinity. *arXiv preprint arXiv:1703.10603* **2017**,

(7) Jiménez, J.; Škalič, M.; Martínez-Rosell, G.; De Fabritiis, G. KDEEP: Protein–Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks. *Journal of Chemical Information and Modeling* **2018**, *58*, 287–296, PMID: 29309725.

(8) Torng, W.; Altman, R. B. Graph Convolutional Neural Networks for Predicting Drug-Target Interactions. *Journal of Chemical Information and Modeling* **2019**, *59*, 4131–4149.

(9) Lim, J.; Ryu, S.; Park, K.; Choe, Y. J.; Ham, J.; Kim, W. Y. Predicting Drug–Target Interaction Using a Novel Graph Neural Network with 3D Structure-Embedded Graph Representation. *Journal of chemical information and modeling* **2019**, *59*, 3981–3988.

(10) Karimi, M.; Wu, D.; Wang, Z.; Shen, Y. DeepAffinity: Interpretable Deep Learning of Compound-Protein Affinity through Unified Recurrent and Convolutional Neural Networks. *arXiv preprint arXiv:1806.07537* **2018**,

(11) Öztürk, H.; Özgür, A.; Ozkirimli, E. DeepDTA: deep drug–target binding affinity prediction. *Bioinformatics* **2018**, *34*, i821–i829.

(12) Feng, Q.; Dueva, E. V.; Cherkasov, A.; Ester, M. PADME: A Deep Learning-based Framework for Drug-Target Interaction Prediction. *CoRR* **2018**, *abs/1807.09741*.

(13) Gao, K. Y.; Fokoue, A.; Luo, H.; Iyengar, A.; Dey, S.; Zhang, P. Interpretable Drug Target Prediction Using Deep Neural Representation. IJCAI. 2018; pp 3371–3377.

(14) Li, X.; Yan, X.; Gu, Q.; Zhou, H.; Wu, D.; Xu, J. DeepChemStable: Chemical Stability Prediction with an Attention-Based Graph Convolution Network. *Journal of chemical information and modeling* **2019**, *59*, 1044–1049.

(15) Uddin, M. R.; Mahbub, S.; Rahman, M. S.; Bayzid, M. S. SAINT: Self-Attention Augmented Inception-Inside-Inception Network Improves Protein Secondary Structure Prediction. *bioRxiv* **2019**, 786921.

(16) McCloskey, K.; Taly, A.; Monti, F.; Brenner, M. P.; Colwell, L. J. Using attribution to decode binding mechanism in neural network models for chemistry. *Proceedings of the National Academy of Sciences* **2019**, *116*, 11624–11629.

(17) Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* **2014**,

(18) Lu, J.; Yang, J.; Batra, D.; Parikh, D. Hierarchical question-image co-attention for visual question answering. Advances In Neural Information Processing Systems. 2016; pp 289–297.

(19) Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. International conference on machine learning. 2015; pp 2048–2057.

(20) Choi, E.; Bahadori, M. T.; Sun, J.; Kulas, J.; Schuetz, A.; Stewart, W. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. Advances in Neural Information Processing Systems. 2016; pp 3504–3512.

(21) Das, A.; Agrawal, H.; Zitnick, L.; Parikh, D.; Batra, D. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding* **2017**, *163*, 90–100.

(22) Doshi-Velez, F.; Kim, B. Towards A Rigorous Science of Interpretable Machine Learning. 2017.

(23) Dill, K.; Bromberg, S. *Molecular driving forces: statistical thermodynamics in biology, chemistry, physics, and nanoscience*; Garland Science, 2012.

(24) Gilson, M. K.; Zhou, H.-X. Calculation of protein-ligand binding affinities. *Annu. Rev. Biophys. Biomol. Struct.* **2007**, *36*, 21–42.

(25) Brzozowski, A. M.; Pike, A. C.; Dauter, Z.; Hubbard, R. E.; Bonn, T.; Engström, O.; Öhman, L.; Greene, G. L.; Gustafsson, J.-Å.; Carlquist, M. Molecular basis of agonism and antagonism in the oestrogen receptor. *Nature* **1997**, *389*, 753.

(26) Congreve, M.; Murray, C. W.; Blundell, T. L. Keynote review: Structural biology and drug discovery. *Drug discovery today* **2005**, *10*, 895–907.

(27) Wlodawer, A.; Vondrasek, J. Inhibitors of HIV-1 protease: a major success of structure-assisted drug design. *Annual review of biophysics and biomolecular structure* **1998**, *27*, 249–284.

(28) Brik, A.; Wong, C.-H. HIV-1 protease: mechanism and drug discovery. *Organic & biomolecular chemistry* **2003**, *1*, 5–14.

(29) Yang, F.; Du, M.; Hu, X. Evaluating explanation without ground truth in interpretable machine learning. *arXiv preprint arXiv:1907.06831* **2019**,

(30) Santoro, A.; Raposo, D.; Barrett, D. G.; Malinowski, M.; Pascanu, R.; Battaglia, P.; Lillicrap, T. A simple neural network module for relational reasoning. Advances in neural information processing systems. 2017; pp 4967–4976.

(31) Lu, C.; Krishna, R.; Bernstein, M.; Fei-Fei, L. Visual relationship detection with language priors. European Conference on Computer Vision. 2016; pp 852–869.

(32) Battaglia, P.; Pascanu, R.; Lai, M.; Rezende, D. J., et al. Interaction networks for learning about objects, relations and physics. Advances in neural information processing systems. 2016; pp 4502–4510.

(33) Hoshen, Y. Vain: Attentional multi-agent predictive modeling. Advances in Neural Information Processing Systems. 2017; pp 2701–2711.

(34) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic acids research* **2006**, *35*, D198–D201.

(35) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.;

Thiessen, P. A.; Yu, B., et al. PubChem 2019 update: improved access to chemical data. *Nucleic acids research* **2018**, *47*, D1102–D1109.

(36) RDKit: Open-source cheminformatics. `http://www.rdkit.org`, [Online; accessed: april 2019].

(37) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The protein data bank. *Nucleic acids research* **2000**, *28*, 235–242.

(38) Laskowski, R. A.; Jabłońska, J.; Pravda, L.; Vařeková, R. S.; Thornton, J. M. PDBsum: Structural summaries of PDB entries. *Protein Science* **2018**, *27*, 129–134.

(39) Cheng, J.; Randall, A. Z.; Sweredoski, M. J.; Baldi, P. SCRATCH: a protein structure and structural feature prediction server. *Nucleic acids research* **2005**, *33*, W72–W76.

(40) Magnan, C. N.; Baldi, P. SSpro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. *Bioinformatics* **2014**, *30*, 2592–2597.

(41) Wang, S.; Sun, S.; Li, Z.; Zhang, R.; Xu, J. Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS computational biology* **2017**, *13*, e1005324.

(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) Trinh, T. H.; Dai, A. M.; Luong, M.-T.; Le, Q. V. Learning longer-term dependencies in rnns with auxiliary losses. *arXiv preprint arXiv:1803.00144* **2018**,

(44) Cortes-Ciriano, I.; Bender, A. KekuleScope: prediction of cancer cell line sensitivity and

compound potency using convolutional neural networks trained on compound images. *Journal of Cheminformatics* **2019**, *11*, 41.

(45) Kipf, T. N.; Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* **2016**,

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

(47) Weisfeiler, B.; Lehman, A. A. A reduction of a graph to a canonical form and an algebra arising during this reduction. *Nauchno-Technicheskaya Informatsia* **1968**, *2*, 12–16.

(48) Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; Hovy, E. Hierarchical attention networks for document classification. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016; pp 1480–1489.

(49) Wang, Z.; Chang, S.; Yang, Y.; Liu, D.; Huang, T. S. Studying very low resolution recognition using deep networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016; pp 4792–4800.