nature machine intelligence

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s42256-020-0152-y

In the format provided by the authors and unedited.

Predicting drug–protein interaction using quasi-visual question answering system

In the format provided by the authors and unedited

Supplementary Materials

Dataset details

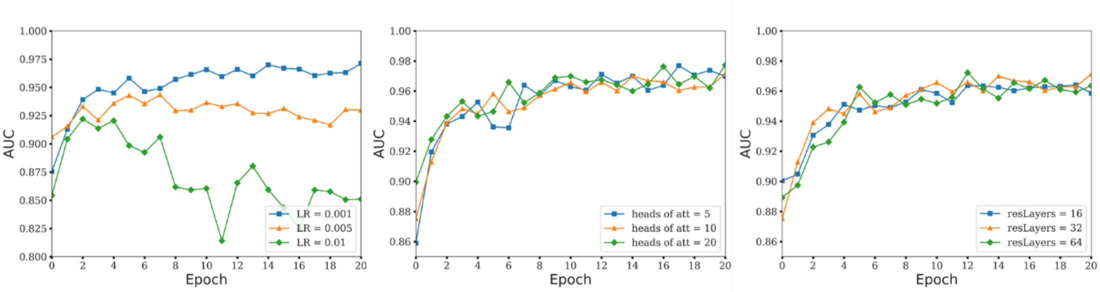
Considering the limited memory of the used GPU (GTX1080Ti 12GB), we chose a maximum length of 1000 characters (amino acids) for protein sequences. The maximum lengths cover 100%, 91%, and 93% of the original DUD-E, BindingDB, and Human datasets, respectively. The origin datasets could be easily downloaded from the literatures. We retrieved the structure data from the Protein Data Bank (PDB). For proteins that don't have crystal structure, we used blast to retrieve homologous proteins with the highest sequence identity from PDB. The proteins with sequence identity less than 40% were excluded.

Neural network training and performance details

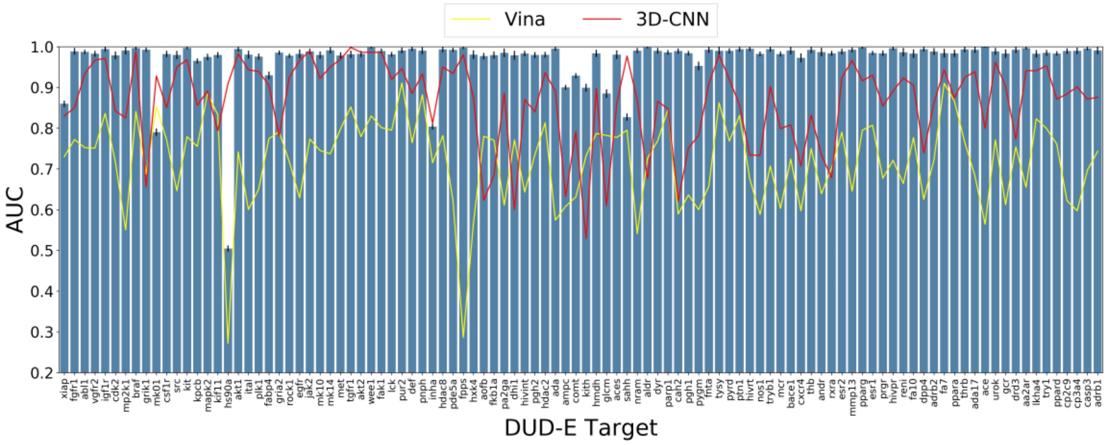
We searched the hyperparameter on a space defined in Supplementary Table 1. The set of the best hyperparameters evaluated on the Human validation set is highlighted in bold. Learning curves with various hyperparameters on the Human validation set were shown in Supplementary Figure 1. We note that the rest of the hyperparameters on other two datasets were not tuned, as we found the model performance was not sensitive to reasonable settings except the learning rate.

Supplementary Table 1. Hyperparameters space, parameters for the best model in bold.

Key Parameters	Possible values
Residual blocks of CNN	16, **32**, 64
Hidden units of BiLSTM	**64**, 128, 256
Hidden units of attention	50, **100**, 150
Heads of attention	5, 10, **20**
Learning rate	**0.001**, 0.003, 0.01
Output dropout	**0.2**, 0.5, 0.8



Supplementary Figure 1. Learning curves with various hyperparameters on the Human validation set. In all learning curves, unless otherwise noted, we use the following hyperparameters: hidden units of BiLSTM = 64, output dropout = 0.2, hidden units of attention = 100, L2 regularization 0.001.

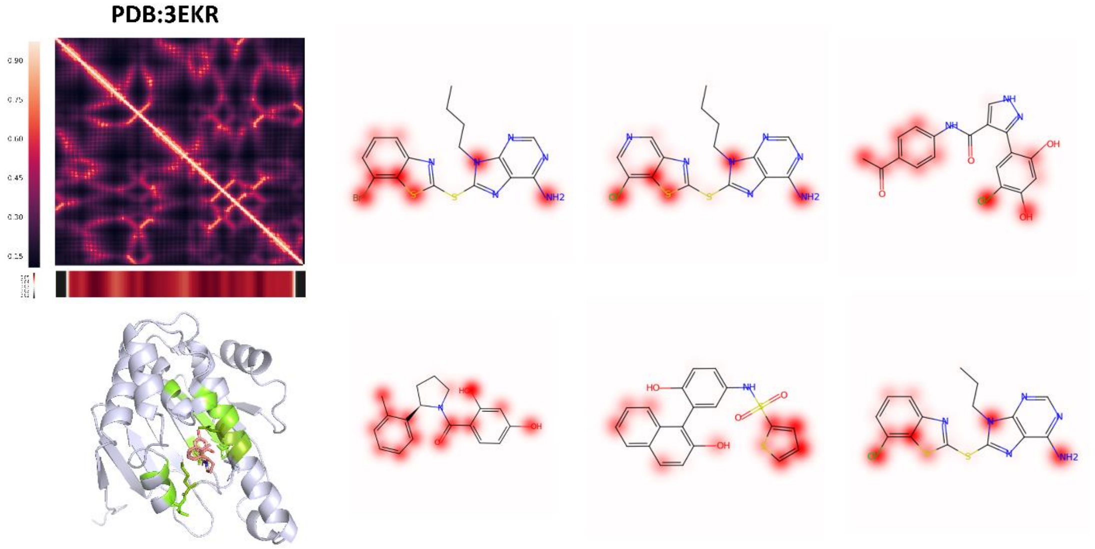


Supplementary Figure 2. Cross-validation performance of DrugVQA model on the DUD-E benchmark compared to the Vina scoring function and 3D-CNN Model.

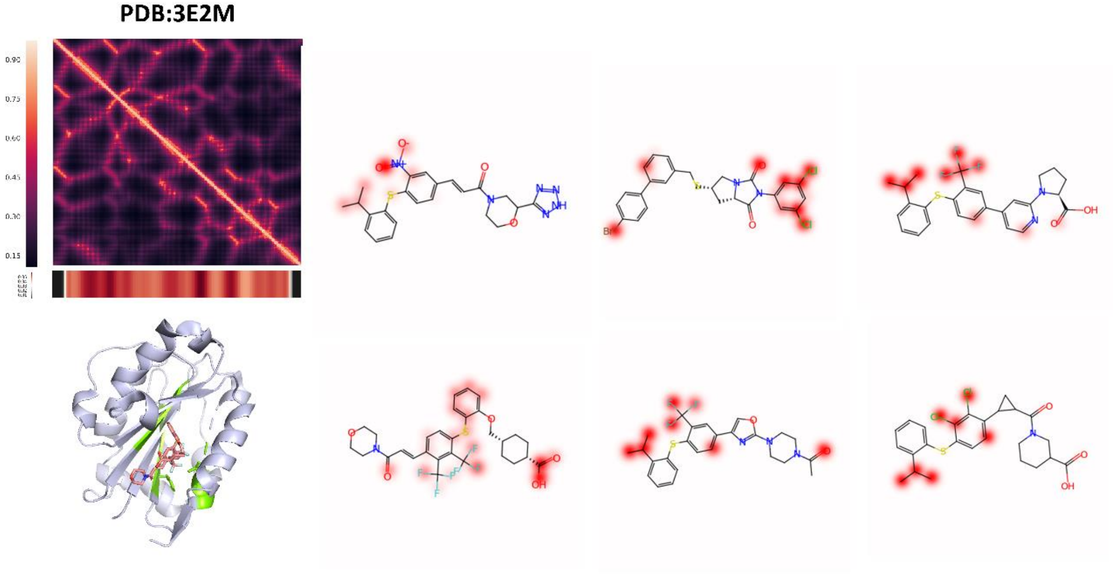
Visualization details

The interpretation of the model is straight forward because of the annotation matrixes A^p and A^m. For each matrix, we summed up over all the annotation vectors, and then normalized the resulting weight vector to a sum of 1. The highest weights correspond to atoms and amino acids in the drug molecules and proteins, which were colored with green and red, respectively.

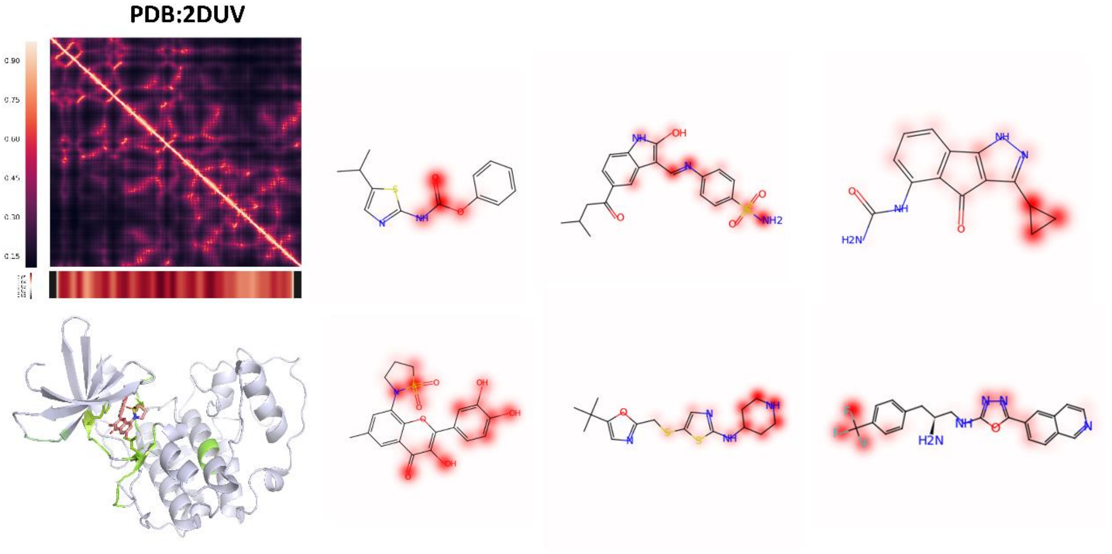
Here we show a few more example visualizations for target CDK2, Hsp90, and LFA-1 with their top predicted actives (Supplementary Figure 3-5). We colored the top fifteen contributing amino acids retrieved from the protein sequential attention map with green, and mapped the molecular attention weight onto atoms of the active compounds with red.



Supplementary Figure 3. Importance visualization of CDK2 (PDB: 3EKR) and its corresponding actives.



Supplementary Figure 4. Importance visualization of LFA-1 (PDB: 3E2M) and its corresponding actives.



Supplementary Figure 5. Importance visualization of Hsp90 (PDB: 2DUV) and its corresponding actives.