**HyperAttentionDTI: improving drug–protein interaction prediction by sequence-based deep learning with attention mechanism**

**(Supplementary Data)**

Qichang Zhao1, Haochen Zhao1, Kai Zheng1, and Jianxin Wang1\*

1Hunan Provincial Key Lab on Bioinformatics, School of Computer Science and Engineering, Central South University, Changsha,410083, China.

\* Contact: jxwang@mail.csu.edu.cn

1. The details of all baselines

We compare our model with seven state-of-the-art deep learning baselines. Table S1 lists the feature extractors of these baselines.

Table S1. The details of all baselines.

|  |  |  |  |
| --- | --- | --- | --- |
| Baselines | Year | Drug extractor | Protein extractor |
| TransformerCPI | 2020 | GNN | CNN, Transformer |
| MolTrans | 2021 | Transformer | Transformer |
| DeepConV-DTI | 2019 | ECFP, DNN | CNN |
| GNN-CPI | 2019 | GNN | CNN |
| GNN-PT | 2020 | GNN | CNN, Transformer |
| DeepEmbedding-DTI | 2021 | GNN | LSTM, Transformer |
| GraphDTA | 2020 | GNN | CNN |

1. Performance evaluation on KIBA and Davis datasets under de novo setting

We compare our model with the baselines under these *de nove* settings on the Davis and KIBA datasets. The results on KIBA dataset under the settings E2, E3, and E4 are in Table S2, S3, and S4, respectively. And the results on Davis dataset under the settings E2, E3, and E4 are in Table S5, S6, and S7, respectively.

Table S2: Comparison results on the KIBA dataset under the setting E2.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Methods | Accuracy(std) | Precision(std) | Recall(std) | AUC(std) | AUPR(std) |
| TransformerCPI | 0.830(0.035) | 0.562(0.032) | 0.498(0.034) | 0.799(0.033) | 0.555(0.032) |
| MolTrans | 0.834(0.021) | 0.618(0.023) | 0.484(0.025) | 0.781(0.022) | 0.525(0.024) |
| DeepConV-DTI | 0.833(0.022) | 0.620(0.023) | 0.471(0.026) | 0.788(0.023) | 0.547(0.025) |
| GNN-CPI | 0.801(0.029) | 0.511(0.098) | 0.472(0.095) | 0.770(0.023) | 0.503(0.044) |
| GNN-PT | 0.804(0.006) | 0.507(0.054) | 0.395(0.046) | 0.732(0.025) | 0.474(0.040) |
| DeepEmbedding-DTI | **0.846**(0.009) | **0.664**(0.062) | 0.408(0.064) | 0.809(0.018) | 0.586(0.037) |
| GraphDTA | 0.836(0.001) | 0.659(0.091) | 0.279(0.055) | 0.754(0.037) | 0.495(0.045) |
| HyperAttentionDTI | 0.806(0.011) | 0.494(0.031) | **0.698**(0.027) | **0.839**(0.025) | **0.640**(0.038) |

Table S3: Comparison results on the KIBA dataset under the setting E3.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Methods | Accuracy(std) | Precision(std) | Recall(std) | AUC(std) | AUPR(std) |
| TransformerCPI | 0.833(0.032) | 0.567(0.037) | 0.538(0.036) | 0.822(0.035) | 0.581(0.035) |
| MolTrans | 0.847(0.021) | 0.626(0.027) | 0.584(0.035) | 0.817(0.024) | 0.573(0.023) |
| DeepConV-DTI | 0.837(0.022) | 0.623(0.023) | 0.599(0.031) | 0.812 (0.022) | 0565(0.024) |
| GNN-CPI | 0.820(0.010) | 0.534(0.028) | 0.448(0.036) | 0.792(0.035) | 0.512(0.052) |
| GNN-PT | 0.828(0.015) | 0.528(0.045) | 0.455(0.032) | 0.814(0.017) | 0.522(0.022) |
| DeepEmbedding-DTI | 0.808(0.021) | 0.508(0.059) | 0.461(0.064) | 0.792(0.013) | 0.507(0.032) |
| GraphDTA | **0.849**(0.001) | **0.639**(0.048) | 0.492(0.006) | 0.838(0.015) | 0.619(0.031) |
| HyperAttentionDTI | 0.808(0.010) | 0.497(0.027) | **0.701**(0.025) | **0.854**(0.012) | **0.644**(0.028) |

Table S4: Comparison results on the KIBA dataset under the setting E4.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Methods | Accuracy(std) | Precision(std) | Recall(std) | AUC(std) | AUPR(std) |
| TransformerCPI | 0.795(0.043) | 0.448(0.043) | 0.367(0.046) | 0.717(0.036) | 0.416(0.042) |
| MolTrans | 0.807(0.031) | 0.476(0.034) | 0.445(0.033) | 0.726 (0.035) | 0.428(0.032) |
| DeepConV-DTI | 0.806(0.032) | 0.486(0.038) | 0.436(0.037) | 0.732 (0.037) | 0.438(0.033) |
| GNN-CPI | 0.732(0.042) | 0.359(0.032) | 0.446(0.350) | 0.686(0.028) | 0.357(0.053) |
| GNN-PT | 0.755(0.025) | 0.322(0.055) | 0.352(0.068) | 0.651(0.025) | 0.312(0.036) |
| DeepEmbedding-DTI | 0.780(0.031) | 0.419(0.049) | 0.329(0.093) | 0.706(0.021) | 0.379(0.021) |
| GraphDTA | **0.821**(0.005) | **0.533**(0.077) | 0.181(0.039) | 0.678(0.027) | 0.374(0.046) |
| HyperAttentionDTI | 0.765(0.014) | 0.393(0.016) | **0.526**(0.056) | **0.737**(0.019) | **0.448**(0.021) |

Table S5: Comparison results on the Davis dataset under the setting E2.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Methods | Accuracy(std) | Precision(std) | Recall(std) | AUC(std) | AUPR(std) |
| TransformerCPI | **0.807**(0.037) | 0.671(0.032) | 0.628(0.041) | 0.842(0.035) | 0.702(0.042) |
| MolTrans | 0.774(0.031) | 0.633(0.031) | 0.498(0.034) | 0.791(0.032) | 0.621(0.036) |
| DeepConV-DTI | 0.800(0.028) | 0.656(0.034) | 0.608(0.034) | 0.845(0.032) | 0.692(0.040) |
| GNN-CPI | 0.785(0.011) | 0.675(0.043) | 0.471(0.036) | 0.812(0.030) | 0.647(0.032) |
| GNN-PT | 0.771(0.016) | 0.607(0.068) | 0.479(0.057) | 0.787(0.026) | 0.601(0.072) |
| DeepEmbedding-DTI | 0.796(0.020) | 0.670(0.038) | 0.567(0.044) | 0.830(0.009) | 0.691(0.018) |
| GraphDTA | 0.789(0.002) | **0.712**(0.055) | 0.402(0.083) | 0.802(0.027) | 0.641(0.064) |
| HyperAttentionDTI | 0.804(0.012) | 0.629(0.021) | **0.752**(0.026) | **0.875**(0.021) | **0.754**(0.022) |

Table S6: Comparison results on the Davis dataset under the setting E3.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Methods | Accuracy(std) | Precision(std) | Recall(std) | AUC(std) | AUPR(std) |
| TransformerCPI | 0.697(0.035) | 0.471(0.039) | 0.483(0.038) | 0.685(0.041) | 0.479(0.037) |
| MolTrans | 0.711(0.034) | 0.485(0.036) | 0.349(0.033) | 0.661(0.031) | 0.452(0.036) |
| DeepConV-DTI | 0.701(0.034) | 0.479(0.031) | 0.442(0.039) | 0.673(0.032) | **0.455**(0.038) |
| GNN-CPI | 0.634(0.036) | 0.379(0.037) | 0.400(0.038) | 0.592(0.035) | 0.406(0.033) |
| GNN-PT | 0.734(0.033) | 0.543(0.171) | 0.385(0.117) | 0.705(0.091) | 0.522(0.153) |
| DeepEmbedding-DTI | **0.727**(0.043) | **0.555**(0.148) | 0.387(0.101) | 0.705(0.031) | 0.517(0.087) |
| GraphDTA | 0.695(0.002) | 0.492(0.133) | 0.262(0.109) | 0.690(0.035) | 0.451(0.086) |
| HyperAttentionDTI | 0.681(0.038) | 0.455 (0.036) | **0.508**(0.035) | **0.689**(0.029) | **0.491**(0.031) |

Table S7: Comparison results on the Davis dataset under the setting E4.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Methods | Accuracy(std) | Precision(std) | Recall(std) | AUC(std) | AUPR(std) |
| TransformerCPI | 0.672(0.047) | 0.414(0.043) | 0.368(0.041) | 0.639(0.047) | 0.412(0.046) |
| MolTrans | 0.648(0.042) | 0.420(0.045) | 0.365(0.045) | 0.642(0.047) | 0.419(0.044) |
| DeepConV-DTI | 0.640(0.040) | 0.405(0.049) | 0.359(0.045) | 0.640(0.045) | 0.420(0.048) |
| GNN-CPI | 0.624(0.045) | 0.341(0.040) | 0.319(0.047) | 0.548(0.043) | 0.378(0.049) |
| GNN-PT | 0.656(0.051) | 0.391(0.167) | 0.284(0.092) | 0.612(0.073) | 0.408(0.074) |
| DeepEmbedding-DTI | **0.716**(0.044) | **0.539**(0.157) | 0.273(0.118) | 0.642(0.049) | 0.426(0.082) |
| GraphDTA | 0.675(0.004) | 0.301(0.174) | 0.157(0.110) | 0.605(0.056) | 0.360(0.084) |
| HyperAttentionDTI | 0.674(0.048) | 0.455(0.040) | **0.383**(0.041) | **0.648**(0.047) | **0.437**(0.044) |

1. Case studies on GABAR proteins

To evaluate the reliability of our model, we perform a case study using actual FDA-approved drugs targeting specific proteins, the human Gamma-aminobutyric acid receptors (GABARs). The number of drugs that have interaction with 16 GABAR proteins and the number of drugs divided in the training and test sets are described in Table S8. Furthermore, the top 10 candidate results of 16 GABAR proteins are listed in Table S9.

Table S8. Summary of the drug numbers in the case study.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Name | UniProt ID | DrugBank2020 | DrugBank2021 | Train | Test | Top10 |
| GABAR alpha-1 | P14867 | 84 | 99 | 84 | 15 | 4 |
| GABAR alpha-2 | P47869 | 76 | 92 | 76 | 16 | 5 |
| GABAR alpha-3 | P34903 | 77 | 93 | 77 | 16 | 6 |
| GABAR alpha-4 | P48169 | 66 | 84 | 66 | 21 | 5 |
| GABAR alpha-5 | P31644 | 73 | 90 | 73 | 17 | 5 |
| GABAR alpha-6 | Q16445 | 66 | 84 | 66 | 21 | 6 |
| GABAR beta-1 | P18505 | 56 | 76 | 56 | 20 | 4 |
| GABAR beta-2 | P47870 | 55 | 76 | 55 | 21 | 5 |
| GABAR beta-3 | P28472 | 58 | 78 | 58 | 20 | 5 |
| GABAR delta | O14764 | 54 | 74 | 54 | 20 | 6 |
| GABAR epsilon | P78334 | 54 | 74 | 54 | 20 | 4 |
| GABAR gamma-1 | Q8N1C3 | 58 | 77 | 58 | 19 | 5 |
| GABAR gamma-2 | P18507 | 63 | 78 | 63 | 17 | 5 |
| GABAR gamma-3 | Q99928 | 57 | 77 | 57 | 20 | 3 |
| GABAR pi | O00591 | 53 | 74 | 53 | 21 | 4 |
| GABAR theta | Q9UN88 | 49 | 73 | 49 | 24 | 6 |

Table S9: The top 10 candidate results of 16 GABAR proteins.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| UniProt ID | DrugBank ID | Evidence | UniProt ID | DrugBank ID | Evidence |
| P14867 | DB01215 | - | P47869 | DB01558 | - |
|  | DB01511 | DrugBank |  | DB01511 | DrugBank |
|  | DB13837 | DrugBank |  | DB01559 | - |
|  | DB01559 | - |  | DB01587 | DrugBank |
|  | DB00384 | - |  | DB01589 | - |
|  | DB01553 | DrugBank |  | DB09166 | - |
|  | DB07712 | - |  | DB13837 | DrugBank |
|  | DB13335 | DrugBank |  | DB13872 | - |
|  | DB00169 | - |  | DB15489 | - |
|  | DB06875 | - |  | DB01553 | DrugBank |
| P34903 | DB13872 | DrugBank | P48169 | DB00897 | - |
|  | DB00546 | DrugBank |  | DB01559 | DrugBank |
|  | DB01559 | DrugBank |  | DB01587 | DrugBank |
|  | DB01110 | - |  | DB13345 | - |
|  | DB00683 | DrugBank |  | DB00546 | DrugBank |
|  | DB01189 | DrugBank |  | DB00934 | - |
|  | DB00246 | - |  | DB09118 | - |
|  | DB00237 | DrugBank |  | DB01511 | DrugBank |
|  | DB00774 | - |  | DB00321 | - |
|  | DB02709 | - |  | DB14719 | DrugBank |
| P31644 | DB01110 | - | Q16445 | DB00546 | DrugBank |
|  | DB01511 | DrugBank |  | DB00690 | - |
|  | DB01587 | DrugBank |  | DB01587 | DrugBank |
|  | DB01589 | - |  | DB09205 | - |
|  | DB09166 | DrugBank |  | DB01489 | DrugBank |
|  | DB13837 | DrugBank |  | DB02709 | - |
|  | DB09237 | - |  | DB01594 | DrugBank |
|  | DB01553 | DrugBank |  | DB06217 | - |
|  | DB13953 | - |  | DB13335 | DrugBank |
|  | DB00381 | - |  | DB13837 | DrugBank |
| P18505 | DB00683 | DrugBank | P47870 | DB08691 | - |
|  | DB01489 | DrugBank |  | DB01587 | DrugBank |
|  | DB13346 | - |  | DB01594 | - |
|  | DB04444 | - |  | DB01489 | DrugBank |
|  | DB05710 | - |  | DB00508 | - |
|  | DB13335 | DrugBank |  | DB04633 | - |
|  | DB00909 | - |  | DB13335 | DrugBank |
|  | DB06333 | - |  | DB00875 | - |
|  | DB07819 | - |  | DB00842 | - |
|  | DB13837 | DrugBank |  | DB01511 | DrugBank |
| P28472 | DB01589 | - | O14764 | DB00683 | DrugBank |
|  | DB01587 | DrugBank |  | DB00690 | - |
|  | DB04534 | - |  | DB00801 | - |
|  | DB01489 | DrugBank |  | DB01489 | DrugBank |
|  | DB06616 | - |  | DB00829 | - |
|  | DB00308 | - |  | DB13335 | DrugBank |
|  | DB00683 | DrugBank |  | DB13837 | DrugBank |
|  | DB01511 | DrugBank |  | DB01594 | - |
|  | DB13335 | DrugBank |  | DB01511 | DrugBank |
|  | DB03332 | - |  | DB01545 | DrugBank |
| P78334 | DB00683 | DrugBank | Q8N1C3 | DB00683 | DrugBank |
|  | DB00690 | - |  | DB00690 | - |
|  | DB00801 | - |  | DB00897 | - |
|  | DB00829 | - |  | DB01489 | DrugBank |
|  | DB00842 | - |  | DB01511 | DrugBank |
|  | DB00897 | - |  | DB01587 | - |
|  | DB01489 | DrugBank |  | DB14719 | DrugBank |
|  | DB01511 | DrugBank |  | DB09088 | - |
|  | DB01544 | DrugBank |  | DB01544 | DrugBank |
|  | DB01215 | - |  | DB06152 | - |
| P18507 | DB00721 |  | Q99928 | DB15489 | - |
|  | DB01110 |  |  | DB09166 | DrugBank |
|  | DB01189 | DrugBank |  | DB13837 | DrugBank |
|  | DB01559 | DrugBank |  | DB09304 | - |
|  | DB11311 |  |  | DB11273 | - |
|  | DB13872 | DrugBank |  | DB13345 | - |
|  | DB02709 |  |  | DB13872 | DrugBank |
|  | DB00228 | DrugBank |  | DB09236 | - |
|  | DB01594 | DrugBank |  | DB08814 | - |
|  | DB00578 |  |  | DB04552 | - |
| O00591 | DB07819 | - | Q9UN88 | DB00774 | - |
|  | DB01587 | DrugBank |  | DB01559 | DrugBank |
|  | DB00909 | - |  | DB01189 | DrugBank |
|  | DB13335 | DrugBank |  | DB02462 | - |
|  | DB00201 | - |  | DB01594 | DrugBank |
|  | DB13837 | DrugBank |  | DB00618 | - |
|  | DB00651 | - |  | DB13872 | DrugBank |
|  | DB00711 | - |  | DB02340 | - |
|  | DB02709 | - |  | DB09166 | DrugBank |
|  | DB00683 | DrugBank |  | DB13335 | DrugBank |

1. T-distributed stochastic neighbor embedding results of GABAR proteins

To explore the structural similarity between the top-ranked candidates and the approved drugs, we visualized the ECFP of compounds. We conducted t-distributed stochastic neighbor embedding (t-SNE) for dimension reduction and visualized the approved drugs, the top-ranked candidates, and compounds with low interaction probabilities.



Figure S1: t-SNE results with compound features. Each color corresponds to a compound class. The red dots represent the approved drugs, the green dots represent the top-ranked candidates, and the blue dots represent compounds with low interaction probabilities.