# Supplementary Information

## A Dual-Branch Graph Neural Network Architecture for Drug-Target Binding Affinity Prediction

March 5, 2026

# SI 1 Materials and methods

## S1.1 Data pre-processing

The DAVIS and KIBA datasets were processed through a standardized pipeline designed for drug-target interaction prediction tasks. For the DAVIS dataset, which contains kinase inhibition profiles, the processing involved downloading a compressed archive from the following source [1], extracting affinity matrices stored in text format, and parsing JSON files containing SMILES strings for drugs and amino acid sequences for targets. The dataset was transformed from matrix form into paired drug-target interactions, with each entry consisting of a SMILES string, protein sequence, and corresponding binding affinity value measured in Kd (nM). By default, affinity values were converted to logarithmic pKd units (-log10(Kd/1e9)) to facilitate regression task.

Similarly, the KIBA dataset, which provides integrated binding scores combining multiple affinity types, underwent comparable processing with download from a GitHub repository [2], extraction of tab-separated affinity matrices, and parsing of accompanying JSON files for drug and target identifiers. The KIBA scores, being pre-integrated values, were used directly without logarithmic transformation for regression tasks. Both datasets maintained complete pairing where every drug was matched with every target, resulting in comprehensive interaction matrices that were then flattened into individual drug-target pair instances for machine learning input.

For both datasets, the train-test partitioning followed a standard approach employing a 70:10:20 split ratio, allocating 70% of drug-target pairs for training, 10% for validation during model development, and 20% for final testing. This partitioning utilized random splitting by default, ensuring statistical representativeness through seed-controlled randomization. Moreover, the reproducibility of these splits was ensured through deterministic random seed, allowing exact replication of data partitions across experimental runs.

[1] *http://staff.cs.utu.fi/ aatapa/data/DrugTarget/DAVIS.zip*

[2] *https://github.com/futianfan/DeepPurpose_Data/blob/main/KIBA.zip*

## S1.2 Dataset characteristics

In this section we have given dataset characteristics about Davis dataset.

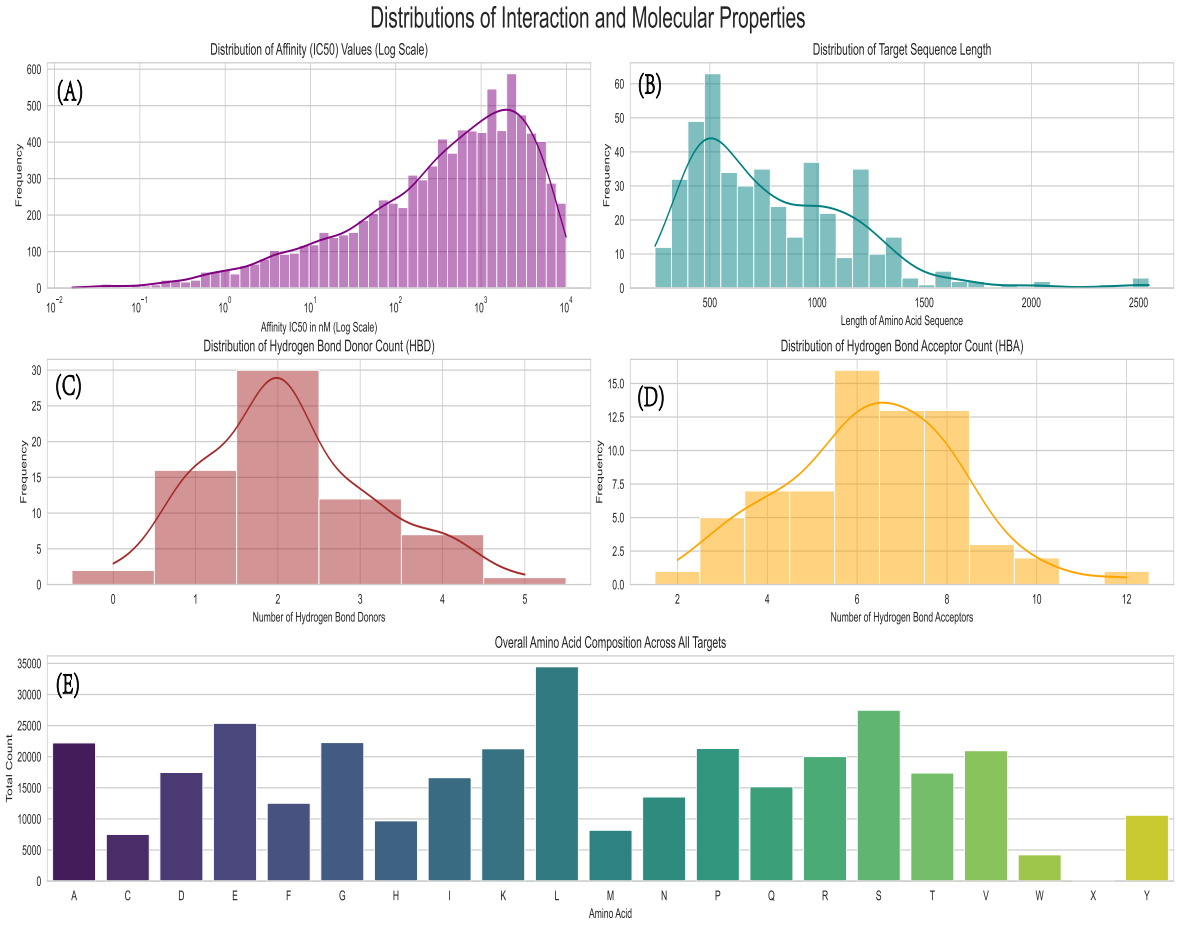


Figure S1: Comprehensive characterization of drug-target interaction properties and molecular descriptors. Subfigure (A) Distribution of binding affinity values (pIC50) showing the potency profile of drug-target pairs, with inactive compounds (IC50 $\geq$ 10,000 nM) highlighted. (B) Molecular weight distribution of small molecule compounds, indicating the size diversity of the chemical library. (C) Target protein sequence length distribution, reflecting the variability in protein domain architectures. (D) Hydrogen bond acceptor (HBA) count distribution across compounds, illustrating hydrogen bonding capacity for target recognition. (E) Amino acid composition analysis of target proteins, revealing residue preferences in the drug-target interactome.

Molecular descriptors are quantitative representations of molecular properties that play crucial roles in determining the behavior of drug molecules and their interactions with biological targets. Here we present some characteristics of our selected dataset.

## Distribution of Affinity (IC50)

Fig.S1 (A) shows the histogram of affinity score IC50 at log scale. It tells us the concentration of a drug required to inhibit the activity of a specific biological target by 50%. A lower IC50 value signifies a more potent compound because a smaller amount of the drug is needed to achieve the desired inhibitory effect.

## Distribution of Target Sequence Length

Fig.S1 (B) shows the distribution of the lengths of the target protein sequences, measured by the number of amino acid residues. **Relevance to Drug Discovery:** While not a direct measure of druggability, sequence length provides context about the targets.

- **Target Diversity:** A wide distribution of lengths suggests that the targets are diverse, ranging from smaller, single-domain proteins to larger, multi-domain complexes.
- **Complexity:** Larger proteins often have more complex structures and regulation mechanisms, which can present both challenges and opportunities for designing selective inhibitors.
- **Computational Modeling:** The size of a protein is a practical consideration for computational methods like molecular docking and molecular dynamics simulations, as larger systems require significantly more computational resources.

## Hydrogen Bond Donor Count (HBD)

$$\text{HBD} = \sum_{i=1}^{n} \delta(\text{atom}_i \in \{\text{N-H}, \text{O-H}\}) \tag{S1}$$

Hydrogen Bond Donor (HBD) count represents the number of atoms in a molecule that can donate hydrogen bonds. These are typically atoms with polar hydrogen atoms attached to electronegative atoms such as nitrogen (N-H) or oxygen (O-H). The Fig.S1 (C) shows the HBD count histogram.

**Relevance to Drug Discovery:** Hydrogen bonds are key non-covalent interactions that contribute to the binding affinity and specificity of a drug to its target protein.

- **Binding Affinity:** HBD groups form specific hydrogen bonds with complementary acceptor sites on target proteins, significantly enhancing binding affinity
- **Specificity:** Hydrogen bonding contributes to molecular recognition and specificity by creating directional interactions
- **Lipinski's Rule of Five:** HBD count $\leq 5$ is one of the criteria for good oral bioavailability
- **Solubility:** Higher HBD count generally increases aqueous solubility.

## Hydrogen Bond Acceptor Count (HBA)

$$\text{HBA} = \sum_{i=1}^{n} \delta(\text{atom}_i \in \{\text{N}, \text{O}, \text{F}\} \text{ with lone pairs}) \tag{S2}$$

Hydrogen Bond Acceptor (HBA) count represents the number of atoms that can accept hydrogen bonds. These are typically electronegative atoms with lone electron pairs, such as oxygen, nitrogen, or fluorine. Fig.S1 (D) shows histogram distribution of our selected dataset.

**Relevance to Drug Discovery:**

- **Complementary Interactions:** HBA sites interact with hydrogen bond donors on target proteins, creating stable complexes
- **Binding Energy:** Each hydrogen bond contributes approximately 1-5 kcal/mol to binding energy
- **Molecular Recognition:** Spatial arrangement of HBA groups determines complementarity with target binding sites
- **Polar Surface Area:** HBA count correlates with polar surface area, affecting membrane permeability.

## Partition Coefficient (LogP)

$$\text{LogP} = \log\left(\frac{[\text{drug}]_{\text{octanol}}}{[\text{drug}]_{\text{water}}}\right) \tag{S3}$$

LogP is the logarithm of the partition coefficient between n-octanol and water, measuring the molecule's hydrophobicity/hydrophilicity balance. Fig.S2 (C) gives the distribution of LogP of our dataset.

**Relevance to Drug Discovery:**

- **Membrane Permeability:** Optimal LogP (typically 1-5) facilitates passive diffusion through lipid membranes
- **Binding Pocket Compatibility:** Determines compatibility with hydrophobic or hydrophilic binding pockets
- **Solubility:** Low LogP indicates high water solubility; high LogP indicates lipophilicity
- **Pharmacokinetics:** Affects absorption, distribution, metabolism, and excretion (ADME) properties
- **Toxicity:** Extremely high LogP may lead to non-specific binding and toxicity

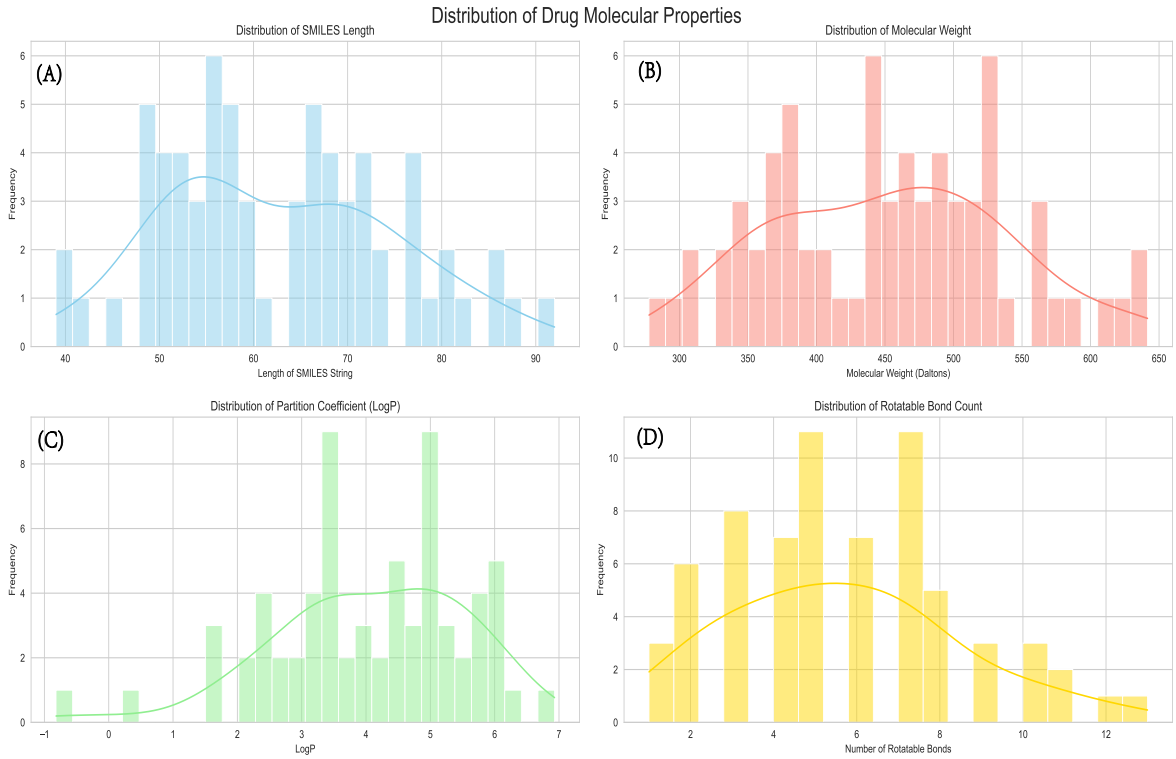


Figure S2: This figure shows drug molecular properties. Subfigure (A) shows SMILE string length distribution. (B) A histogram showing the frequency of compounds across different molecular weights. (C) A histogram of the calculated LogP values for the compound library. (D) A histogram of number of rotable bonds.

## Rotatable Bond Count

$$\text{Rotatable Bonds} = \sum \delta(\text{bond is non-ring, non-terminal, non-amide, non-ester}) \quad \text{(S4)}$$

Rotatable bond count represents the number of bonds that allow free rotation, typically single bonds not in rings and not adjacent to terminal atoms. Fig.S2 (D) shows the histogram of different rotatable bonds.

**Relevance to Drug Discovery:**

- **Conformational Flexibility:** Determines the molecule's ability to adopt different conformations.
- **Binding Entropy:** Fewer rotatable bonds reduce entropy loss upon binding, enhancing affinity.
- **Target Adaptation:** Flexible molecules can adapt to different binding site geometries.
- **Oral Bioavailability:** Rotatable bond count $\leq 10$ is preferred for good oral bioavailability.
- **Synthetic Accessibility:** Fewer rotatable bonds often correlate with easier synthesis.

### Optimal Ranges for Drug-Like Properties

Table S1 summarizes the optimal ranges of molecular and interaction characteristics used in drug discovery.

Table S1: Optimal ranges for molecular descriptors in drug discovery

| Descriptor | Optimal Range | Role in DTI | Impact |
|---|---|---|---|
| HBD Count | 0-5 | Hydrogen bonding | Binding affinity, specificity |
| HBA Count | 2-10 | Hydrogen bonding | Solubility, recognition |
| LogP | 1-5 | Hydrophobicity | Permeability, binding |
| Rotatable Bonds | $\leq 10$ | Flexibility | Bioavailability, entropy |

### Amino Acid Composition

Fig.S1 (E) illustrates the overall frequency of each of the amino acids across all target protein sequences in the dataset combined.

**Relevance to Drug Discovery:** The amino acid composition influences the overall properties of the target proteins and their binding sites.

- **Binding Site Character:** A high prevalence of hydrophobic residues (like Leucine, Isoleucine, Valine) might suggest that lipophilic drugs would be more effective. Conversely, a high number of charged or polar residues (like Aspartate, Glutamate, Lysine) would favor drugs capable of forming salt bridges or hydrogen bonds.

- **Target Class Identification:** Certain protein classes have characteristic amino acid compositions. For example, kinases often have conserved residues in their ATP-binding pocket. Analyzing this composition can help classify the targets.

- **Druggability Assessment:** Proteins rich in "hot spot" residues (like Tyrosine, Tryptophan, Arginine), which contribute significantly to binding energy, are often considered more "druggable."

## Model average running time analysis

In the table we have evaluated the time taken for training on the same datasets and reported in the table S2. We can see MPNN along with Quasi-seq target encoding takes minimum time(10.48 minutes) while GIN_ContextPred along with CNN_RNN target encoding takes maximum time (88.26 minutes). Rest of the models lie within this range.

Table S2: Average Execution Times of Different Models

| **Model(Drug-enc_Target-enc)** | **Time (seconds)** | **Time (minutes)** | **Time (hours)** |
|---|---|---|---|
| MPNN_Quasi-seq | 628.72 | 10.48 | 0.17 |
| MPNN_ESPF | 742.29 | 12.37 | 0.21 |
| MPNN_Conjoint_triad | 745.48 | 12.42 | 0.21 |
| MPNN_AAC | 852.49 | 14.21 | 0.24 |
| MPNN_PseudoAAC | 1549.59 | 25.83 | 0.43 |
| Transformer_Quasi-seq | 1835.40 | 30.59 | 0.51 |
| Transformer_Conjoint_triad | 1861.06 | 31.02 | 0.52 |
| Transformer_ESPF | 1971.01 | 32.85 | 0.55 |
| Transformer_AAC | 2061.97 | 34.37 | 0.57 |
| Transformer_PseudoAAC | 2751.85 | 45.86 | 0.76 |
| GIN_ContextPred_Quasi-seq | 2940.58 | 49.01 | 0.82 |
| GIN_ContextPred_Conjoint_triad | 2971.94 | 49.53 | 0.83 |
| GIN_AttrMasking_Quasi-seq | 3002.24 | 50.04 | 0.83 |
| GIN_AttrMasking_Conjoint_triad | 3167.26 | 52.79 | 0.88 |
| AttentiveFP_Conjoint_triad | 3475.72 | 57.93 | 0.97 |
| GIN_AttrMasking_ESPF | 3498.63 | 58.31 | 0.97 |
| AttentiveFP_Quasi-seq | 3525.28 | 58.75 | 0.98 |
| GCN_Quasi-seq | 3552.36 | 59.21 | 0.99 |
| GIN_ContextPred_ESPF | 3632.94 | 60.55 | 1.01 |
| GCN_Conjoint_triad | 3727.22 | 62.12 | 1.04 |
| GIN_AttrMasking_PseudoAAC | 3741.98 | 62.37 | 1.04 |
| GIN_ContextPred_PseudoAAC | 3761.52 | 62.69 | 1.04 |
| AttentiveFP_ESPF | 3923.95 | 65.40 | 1.09 |
| NeuralFP_Quasi-seq | 4116.39 | 68.61 | 1.14 |
| NeuralFP_Conjoint_triad | 4174.15 | 69.57 | 1.16 |
| AttentiveFP_PseudoAAC | 4197.59 | 69.96 | 1.17 |
| GCN_ESPF | 4362.46 | 72.71 | 1.21 |
| GCN_PseudoAAC | 4436.20 | 73.94 | 1.23 |
| MixHop_ESPF | 4537.40 | 75.62 | 1.26 |
| GIN_AttrMasking_AAC | 4545.43 | 75.76 | 1.26 |
| NeuralFP_AAC | 4566.43 | 76.11 | 1.27 |
| GIN_ContextPred_AAC | 4624.16 | 77.07 | 1.28 |
| NeuralFP_PseudoAAC | 5134.07 | 85.57 | 1.43 |
| Proposed_ESPF | 5504.99 | 91.75 | 1.53 |
| MPNN_CNN | 7205.78 | 120.10 | 2.00 |
| Transformer_CNN | 7369.15 | 122.82 | 2.05 |
| GIN_AttrMasking_CNN | 13253.81 | 220.90 | 3.68 |
| GIN_ContextPred_CNN | 13302.88 | 221.71 | 3.70 |
| NeuralFP_CNN | 13508.47 | 225.14 | 3.75 |
| MPNN_Transformer | 23206.21 | 386.77 | 6.45 |
| Transformer_Transformer | 24611.79 | 410.20 | 6.84 |
| GIN_ContextPred_CNN_RNN | 48015.87 | 800.26 | 13.34 |