**Source of datasets:**

Davis dataset: Comprehensive analysis of kinase inhibitor selectivity. (doi:10.1038/nbt.1990)

KIBA dataset: Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. (doi: 10.1021/ci400709d)

**Summary of datasets:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Proteins | Compounds | Interactions |
| Davis | 442 | 68 | 30056 |
| KIBA | 229 | 2111 | 118254 |

**Document description:**

├── data

│   ├── davis

│   │   ├── Davis\_binding\_affinity.txt

│   │   ├── Davis\_compound\_sim.txt

│   │   ├── Davis\_compound\_smiles.txt

│   │   ├── Davis\_proteins\_seq.txt

│   │   └── Davis\_proteins\_sim.txt

│   └── kiba

│   ├── kiba\_binding\_affinity.txt

│   ├── kiba\_compound\_sim.txt

│   ├── kiba\_compound\_smiles.txt

│   ├── kiba\_proteins\_seq.txt

│   └── kiba\_proteins\_sim.txt

|  |  |  |  |
| --- | --- | --- | --- |
| File name | Description | Data format |  |
| binding\_affinity | Compounds-Proteins binding affinity | 2D matrix | Row: compound, same order with file compound\_smiles  Column: proteins, same order with file proteins\_seq |
| compound\_sim | The similarity between compounds, calculated by KronRLS method. | 2D matrix | Row: compound, same order with file compound\_smiles  Column: same with row |
| compound\_smiles | The SMILES (Simplified Molecular Input Line Entry System) representation of compounds | json |  |
| proteins\_seq | Proteins sequence | json |  |
| proteins\_sim | The similarity between proteins, calculate by Smith–Waterman (S–W) similarity | 2D matrix | Row: proteins, same order with file proteins\_seq  Column: same with row |

**Attention:**

1. Two datasets use different method to calculate binding affinity, so the scale of the binding affinity varies between two datasets.
2. The “nan” value indicates that there is no experimental value for the compound-protein pair.