Empowering Heterogeneous Networks for Drug-Target Affinity Prediction

Selen Parlar
January 27, 2022

Outline

- Drug Design
- Background
- WideDeepDTA
- Results
- Conclusion

Drug Design

- Discovers candidate chemicals
- Evaluates against various proteins

→ Costly and time-consuming process

- Rational design with computers
- Selecting and improving ligands from a molecule library

→ Huge search space

~ 100M chemicals

~ 190M proteins

Drugs & Proteins

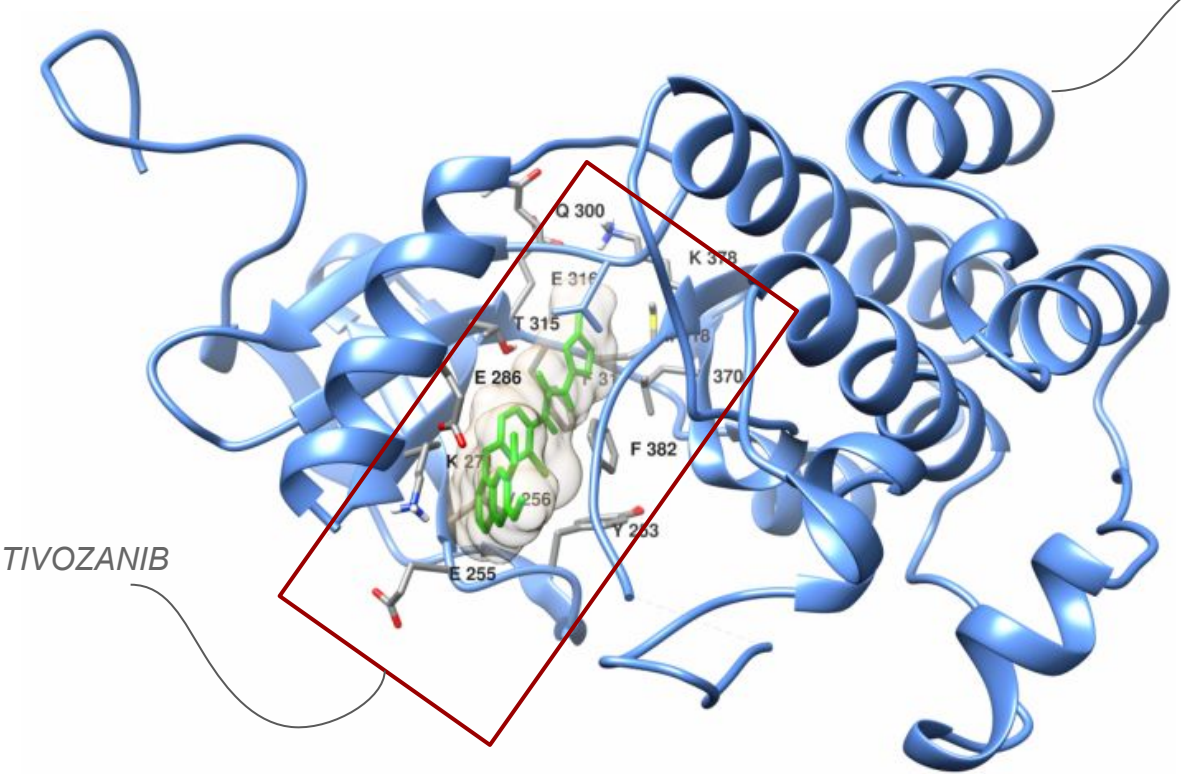
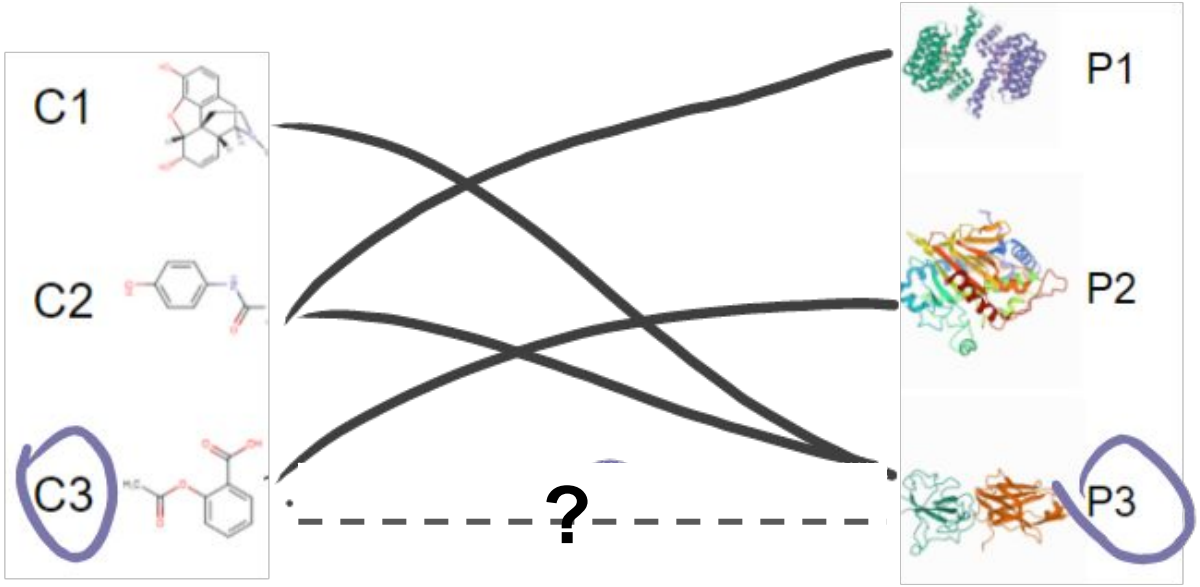


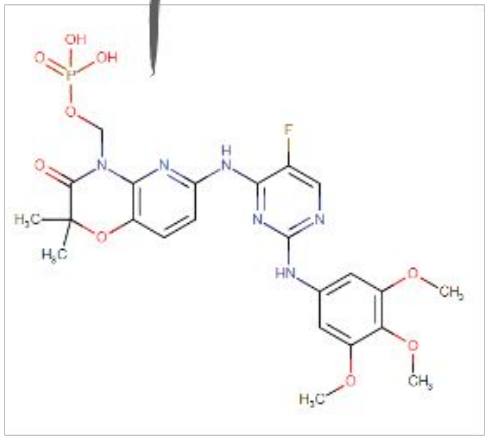
Image source: Machine learning guides drug-target mapping efforts

Searching High-Affinity Pairs

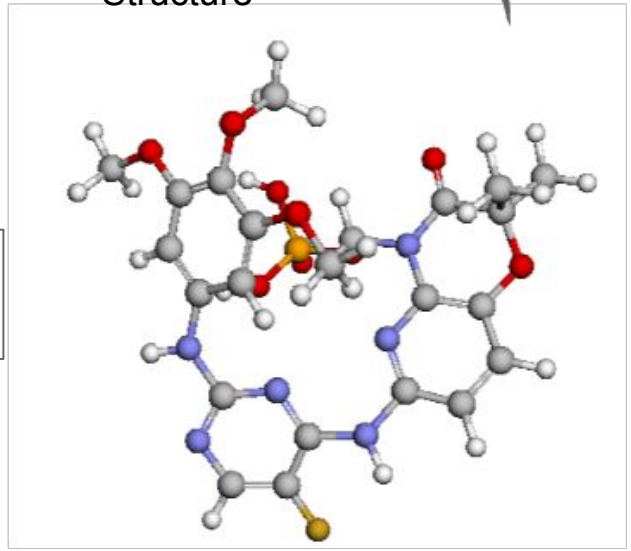


Data

2D: Molecular Structure



3D: Molecular Structure



Fostamatinib
$C_{23}H_{26}FN_6O_9P$

COc1cc(Nc2ncc(F)c(Nc3ccc4c(n3)N(COP(=O)(O)O)C(=O)C(C)(C)O4)n2)cc(OC)c1OC

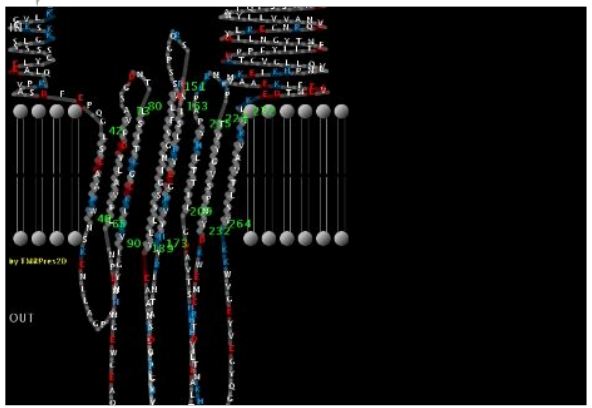
1D:SMILES

1D:InChI

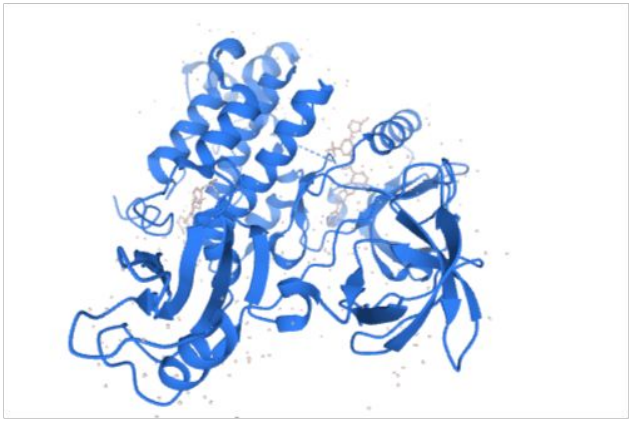
1S/C23H26FN6O9P/c1-23(2)21(31)30(11-38-40(32,33)34)20-14(39-23)6-7-17(28-20)27-19-13(24)10-25-22(29-19)26-12-8-15(35-3)18(37-5)16(9-12)36-4/h6-10H,11H2,1-5H3,(H2,32,33,34)(H2,25,26,27,28,29)

Data

2D: Protein Structure



3D: Protein Structure

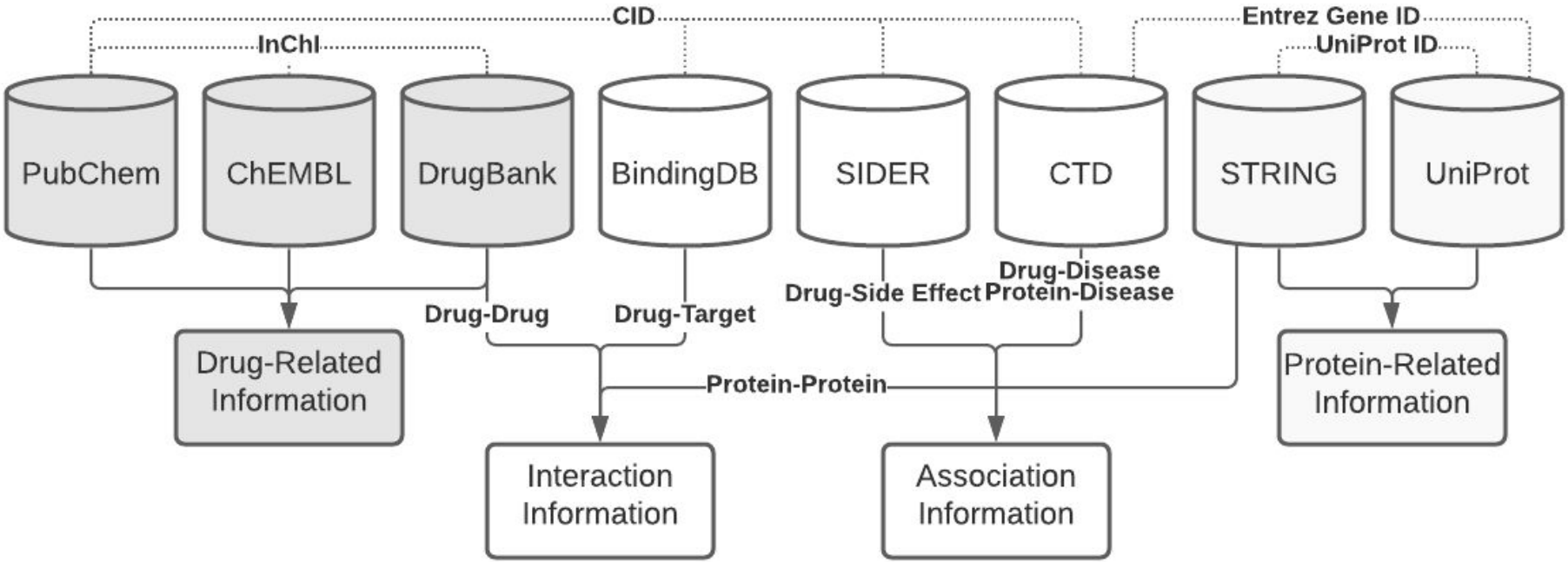


Tyrosine-protein kinase ABL1

1D:Amino Acid Sequence

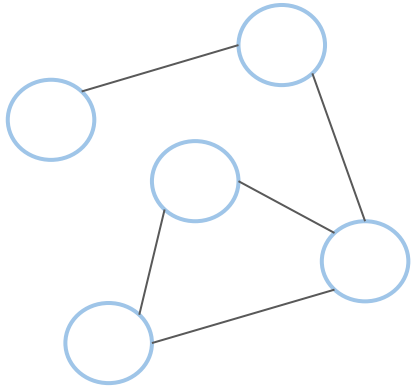
MLEICLKLVG CKSKKGLSSS SSCYLEEALQ RPVASDFEPQ GLSEAARWNS KENLLAGPSE NDPNLFVALY DFVASGDNTL SITKGEKLRV LGYNHNGEWC EAQTKNGQGW VPSNYITPVN SLEKHSWYHG PVSRNAAEYL LSSGINGSFL VRESESSPGQ RSISLRYEGR VYHYRINTAS DGKLYVSSES RFNTLAELVH HHSTVADGLI TTLHYPAPKR NKPTVYGVSP NYDKWEMERT DITMKHKLGG GQYGEVYEGV WKKYSLTVAV KTLKEDTMEV EEFLKEAAVM KEIKHPNLVQ LLGVCTREPP FYIITEFMTY GNLLDYLREC NRQEVNAVVL LYMATQISSA MEYLEKKNFI HRDLAARNCL VGENHLVKVA DFGLSRLMTG DTYTAHAGAK FPIKWTAPES LAYNKFSIKS DVWAFGVLLW EIATYGMSPY PGIDLSQVYE LLEKDYRMER PEGCPEKVYE LMRACWQWNP SDRPSFAEIH QAFETMFQES SISDEVEKEL GKQGVRGAVS TLLQAPELPT KTRTSRRAAE HRDTTDVPEM PHSKGQGESD PLDHEPAVSP LLPRKERGPP EGGLNEDERL LPKDKKTNLF SALIKKKKKT APTPPKRSSS FREMDGQPER RGAGEEEGRD ISNGALAFTP LDTADPAKSP KPSNGAGVPN GALRESGGSG FRSPHLWKKS STLTSSRLAT GEEEGGGSSS KRFLRSCSAS CVPHGAKDTE WRSVTLPRDL QSTGRQFDSS TFGGHKSEKP ALPRKRAGEN RSDQVTRGTV TPPPRLVKKN EEAADEVFKD IMESSPGSSP PNLTPKPLRR QVTVAPASGL PHKEEAGKGS ALGTPAAAEP VTPTSKAGSG APGGTSKGPA EESRVRRHKH SSESPGRDKG KLSRLKPAPP PPPAASAGKA GGKPSQSPSQ EAAGEAVLGA KTKATSLVDA VNSDAAKPSQ PGEGLKKPVL PATPKPQSAK PSGTPISPAP VPSTLPSASS ALAGDQPSST AFIPLISTRV SLRKTRQPPE RIASGAITKG VVLDSTEALC LAISRNSEQM ASHSAVLEAG KNLYTFCVSY VDSIQQMRNK FAFREAINKL ENNLRELQIC PATAGSGPAA TQDFSKLLSS VKEISDIVQR

Additional Data

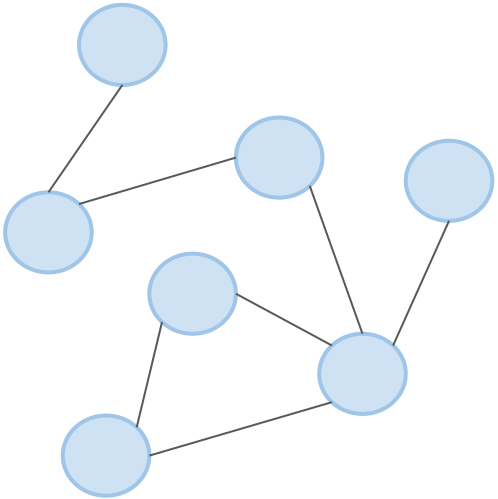


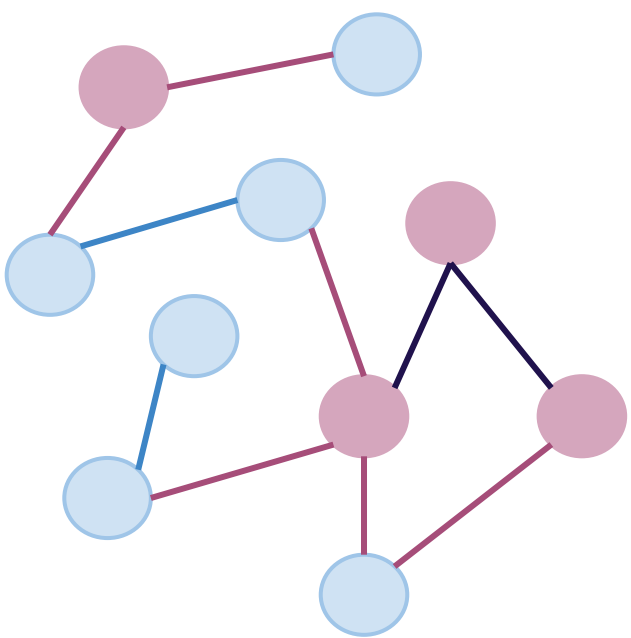
Graphs

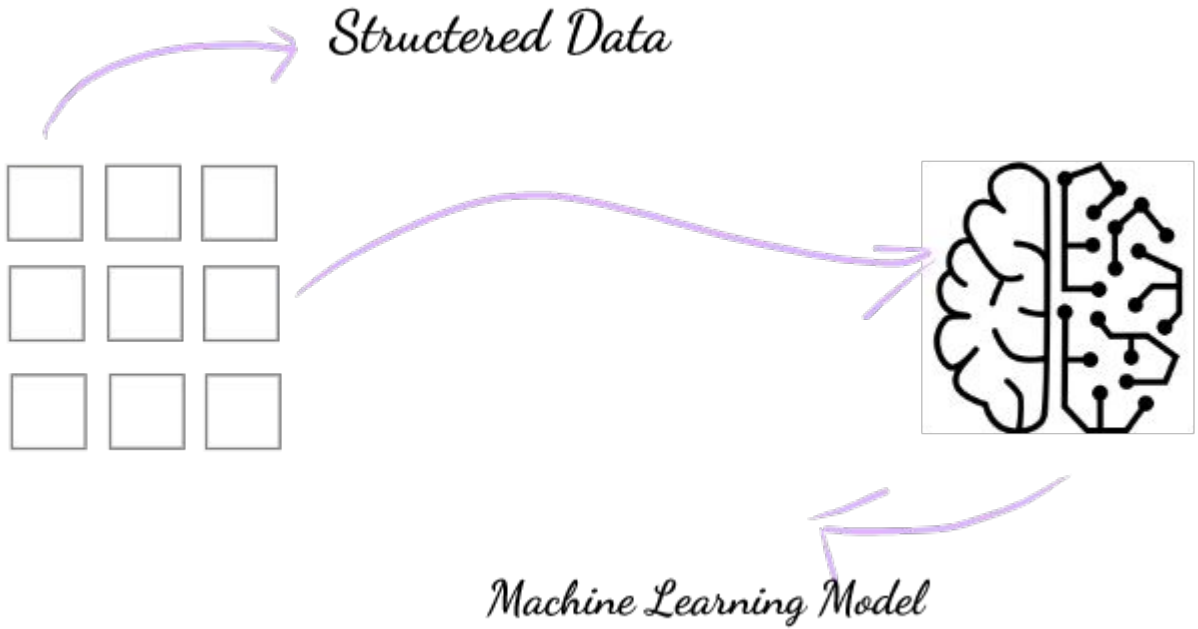
- A graph is a non-linear data structure consisting of nodes and edges.
- G = (V, E, T) comprising:
  - A set of nodes V,
  - A set of edges E, from $u \in V$ to $v \in V$, with relation type τ as $(u,\tau,v) \in E$.

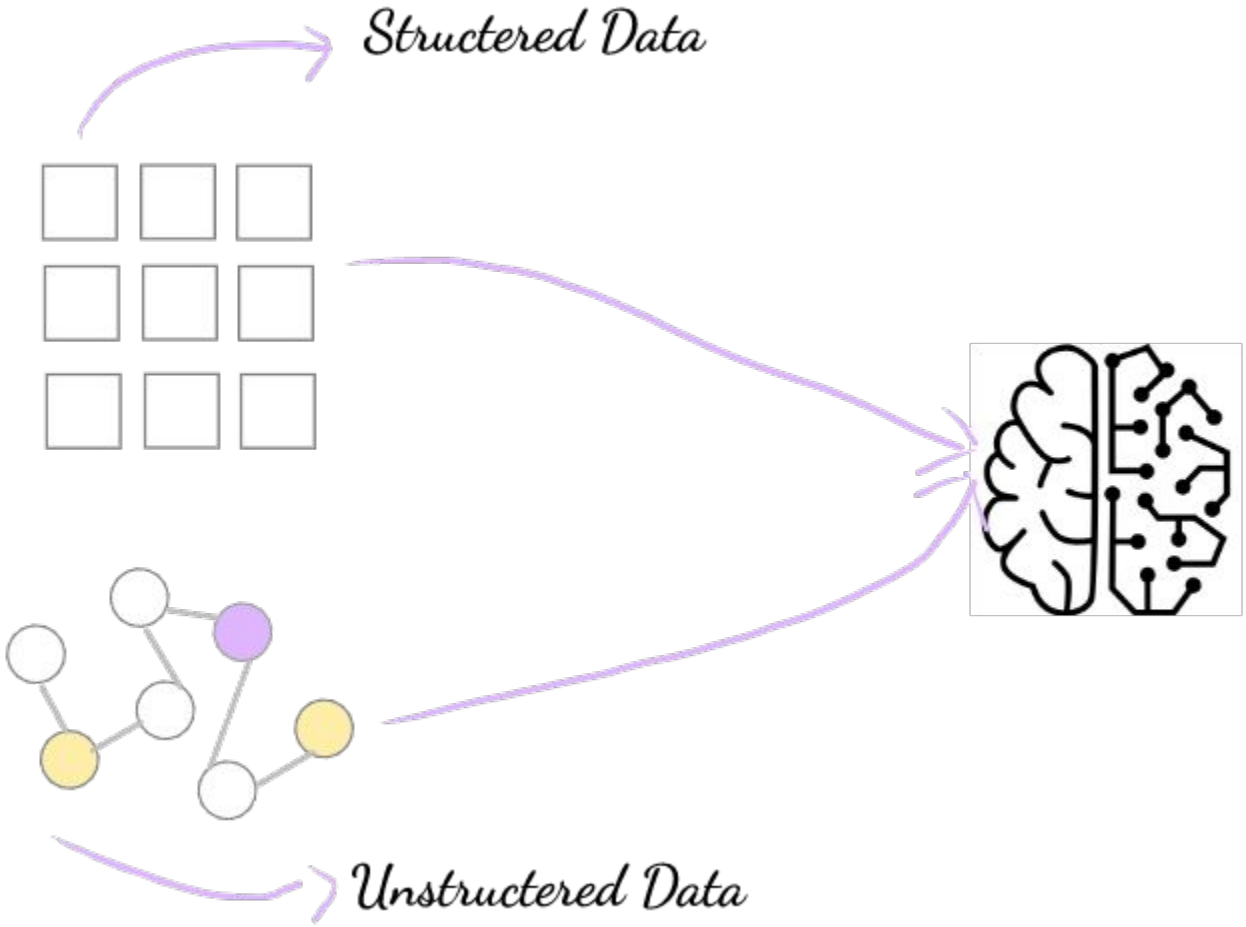


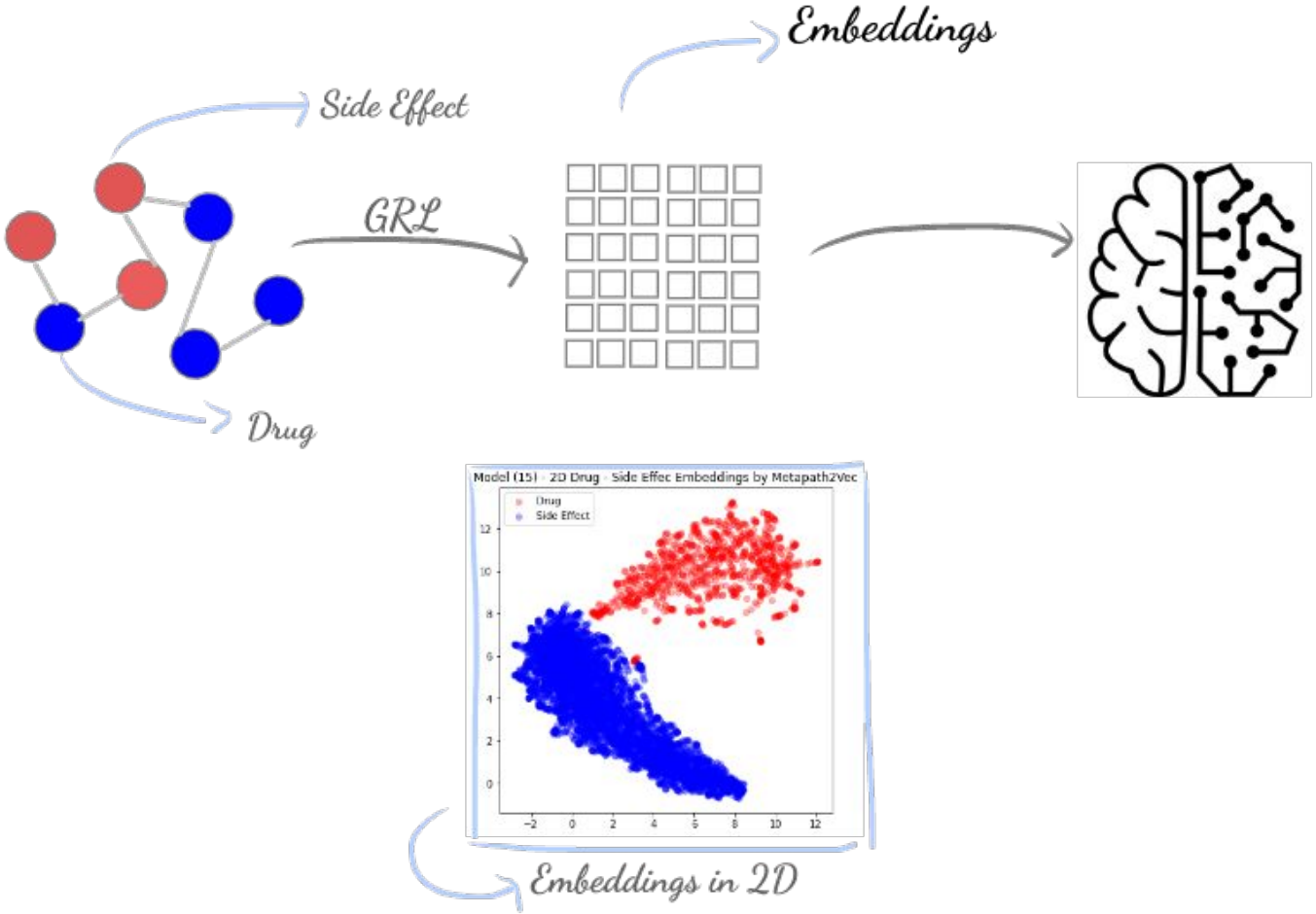
Homogeneous & Heterogeneous Graphs





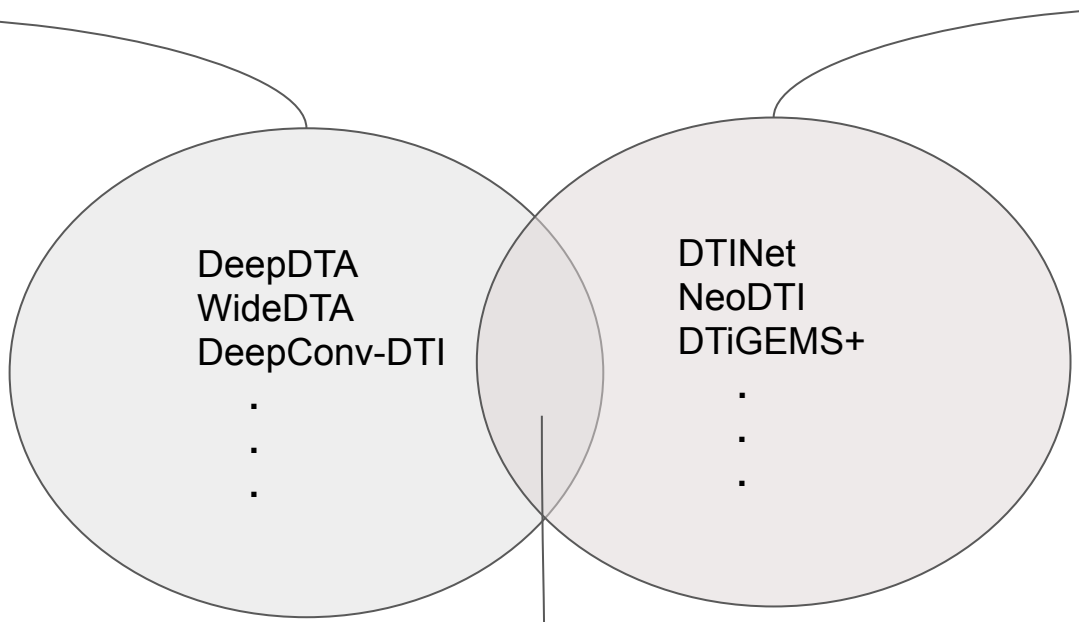

Structered Data
Machine Learning Model


Structered Data
Unstructered Data


Embeddings
Side Effect
GRL
Drug
Model (15) - 2D Drug - Side Effec Embeddings by Metapath2Vec
Drug
Side Effect
Embeddings in 2D

Employs text-based representation.

Employs heterogeneous representation.



WideDeepDTA

Öztürk, H., Özgür, A., & Ozkirimli, E. (2018). DeepDTA: deep drug–target binding affinity prediction. *Bioinformatics*, *34*(17), i821-i829.

Öztürk, H., Ozkirimli, E., & Özgür, A. (2019). WideDTA: prediction of drug-target binding affinity. *arXiv preprint arXiv:1902.04166*.

Lee, I., Keum, J., & Nam, H. (2019). DeepConv-DTI: Prediction of drug-target interactio… learning with convolution on protein sequences. *PLoS computational biology*, *15*(6), e1…

Wan, F., Hong, L., Xiao, A., Jiang, T., & Zeng, J. (2019). NeoDTI: neural integration of neighbor information from a heterogeneous network for discovering new drug–target interactions. *Bioinformatics*, *35*(1), 104-111.

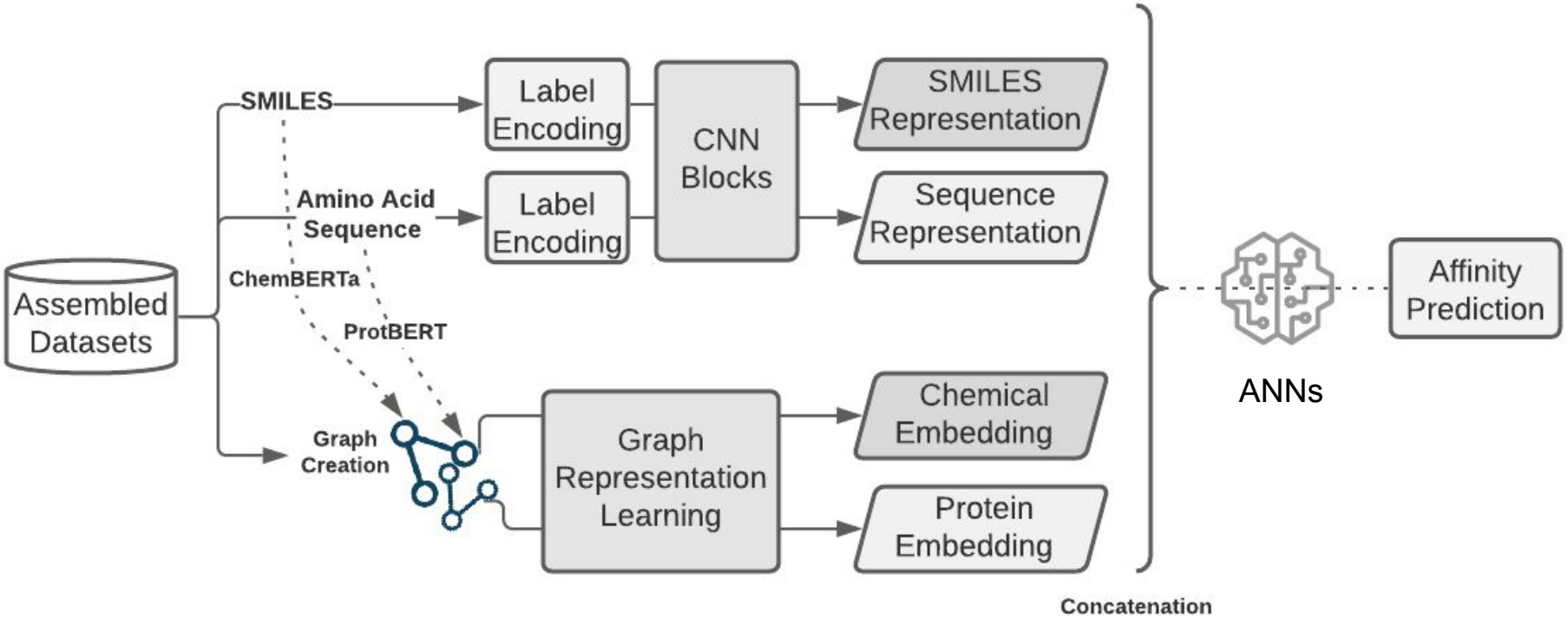
Luo, Y., Zhao, X., Zhou, J., Yang, J., Zhang, Y., Kuang, W., ... & Zeng, J. (2017). A network …tion approach for drug-target interaction prediction and computational drug repositioning …eterogeneous information. *Nature communications*, *8*(1), 1-13

… M. A., Olayan, R. S., Ashoor, H., Albaradei, S., Bajic, V. B., Gao, X., ... & Essack, M. (2020). DTiGEMS+: drug–target interaction prediction using graph embedding, graph mining, and similarity-based techniques. *Journal of Cheminformatics*, *12*(1), 1-17..

WideDeepDTA

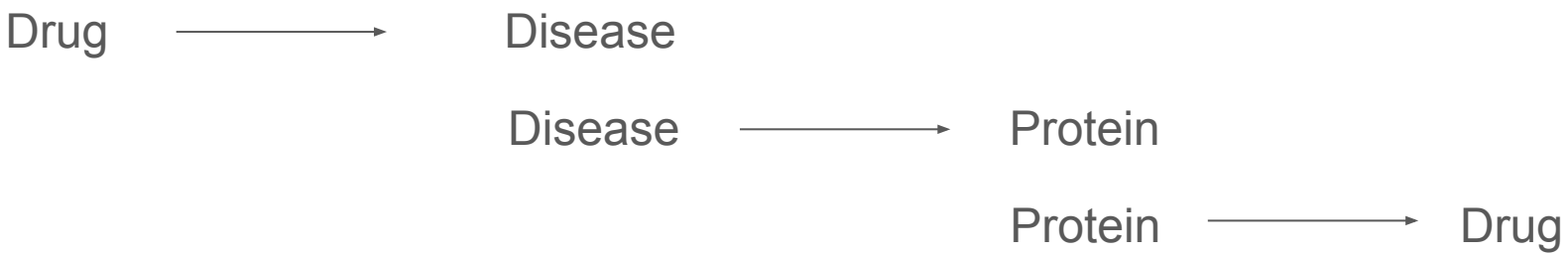
The first drug-target affinity prediction framework empowers heterogeneous networks with text-based representations.

WideDeepDTA



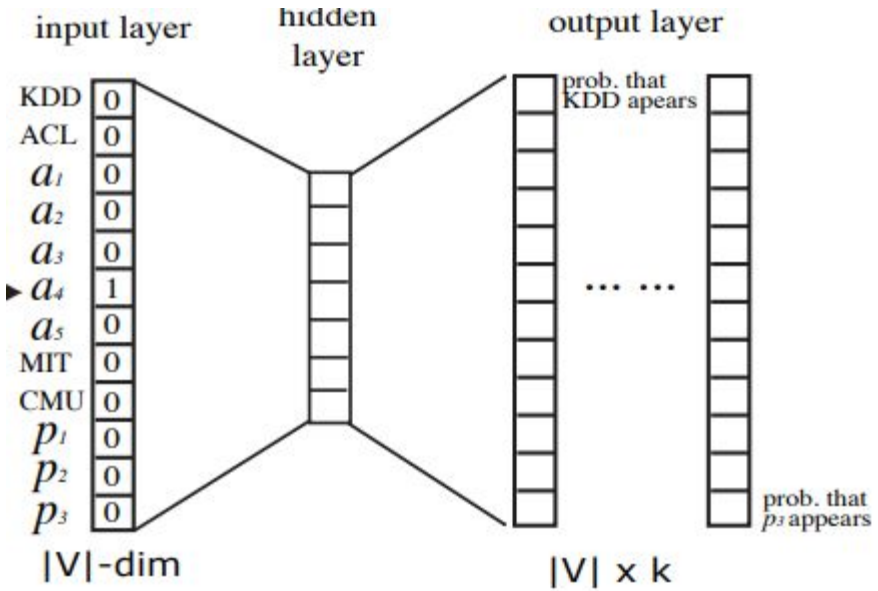
Metapath2Vec

- A **meta-path** in a heterogeneous graphs, is a path following a specific meta path scheme P.



Metapath2Vec

- Given an heterogeneous network G = (V, E, T) and a meta-path scheme P, the Metapath2Vec model calculates the transition probability. And finds the next node, given the current node, the type of the node and the scheme P..



Dong, Y., Chawla, N. V., & Swami, A. (2017, August). metapath2vec: Scalable representation learning for heterogeneous networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 135-144).

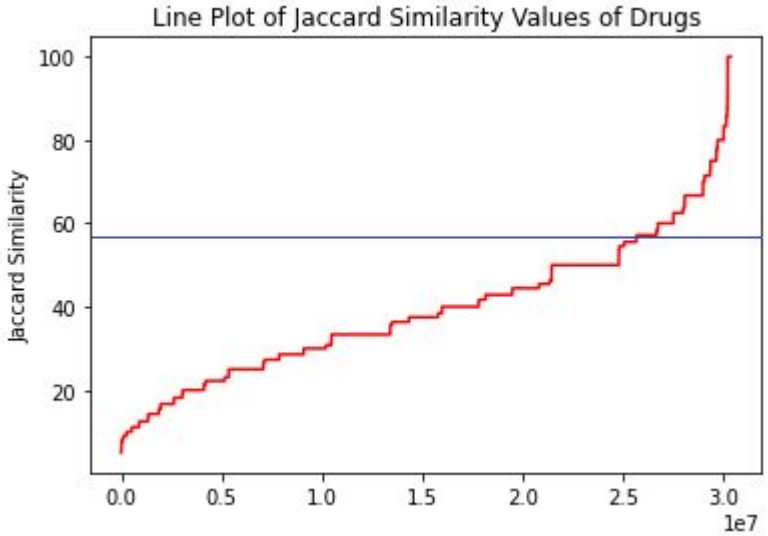
Graph Creation

1. Chemical Related Information:
    a. SMILES strings
    b. Drug-Side Effect Association
    c. Drug-Disease Association
    d. Text-Based Similarity
    e. ChemBERTa Embeddings
    f. BioBERT Embeddings for diseases and side effects

Graph Creation

1. Chemical Related Information:
    a. SMILES strings
    b. Drug-Side Effect Association
    c. Drug-Disease Association
    d. Text-Based Similarity
    e. ChemBERTa Embeddings
    f. BioBERT Embeddings for diseases and side effects

$$Jaccard(A, B) = \frac{|A \cap B|}{|A \cup B|} \times 100$$



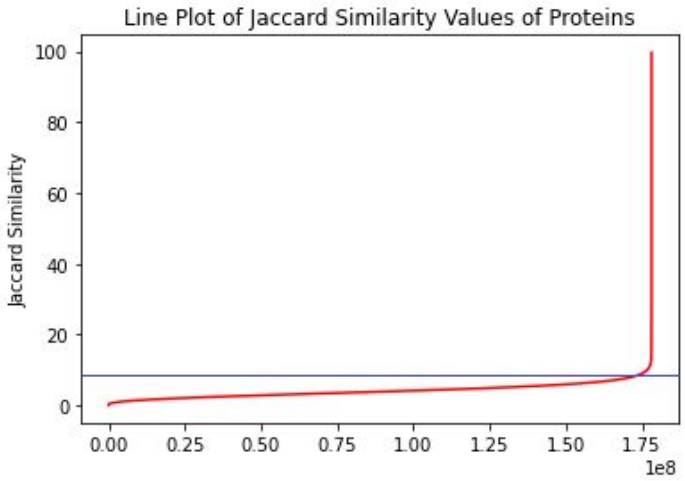
Graph Creation

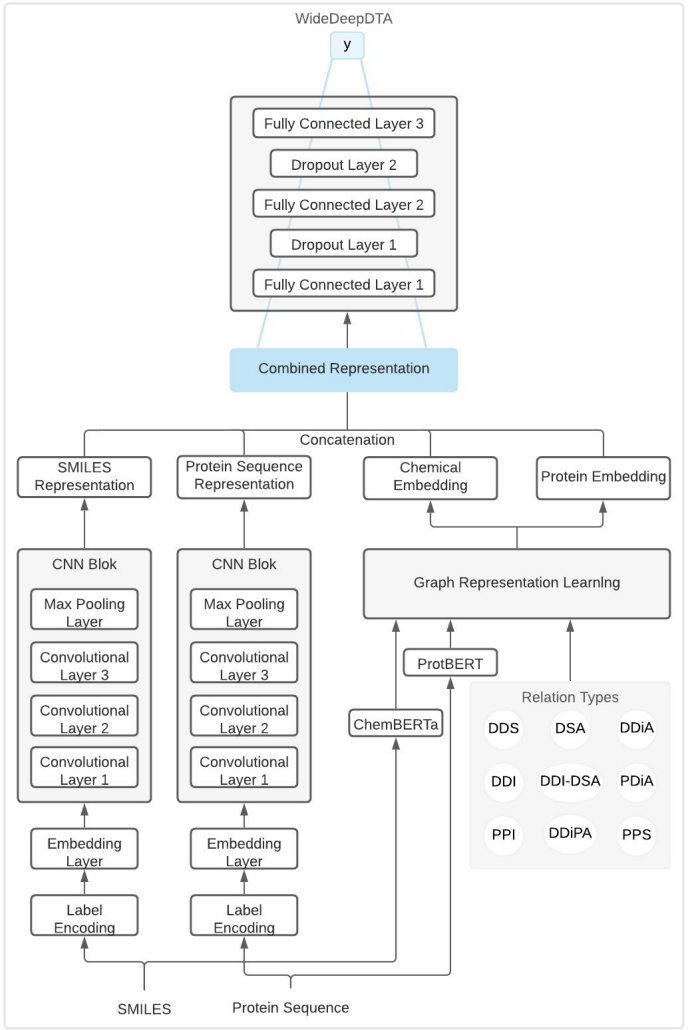
2. Protein Related Information:
    a. Amino acid sequences
    b. Protein-Disease Association
    c. Text-Based Similarity
    d. ProtBERT Embeddings
    e. BioBERT Embeddings for diseases

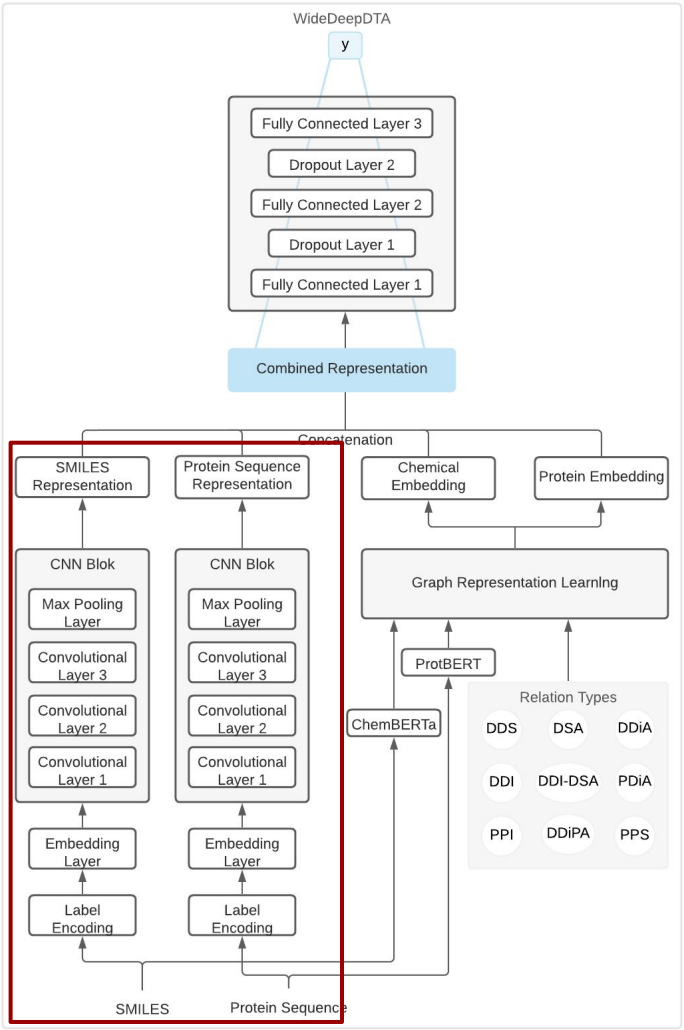
Graph Creation

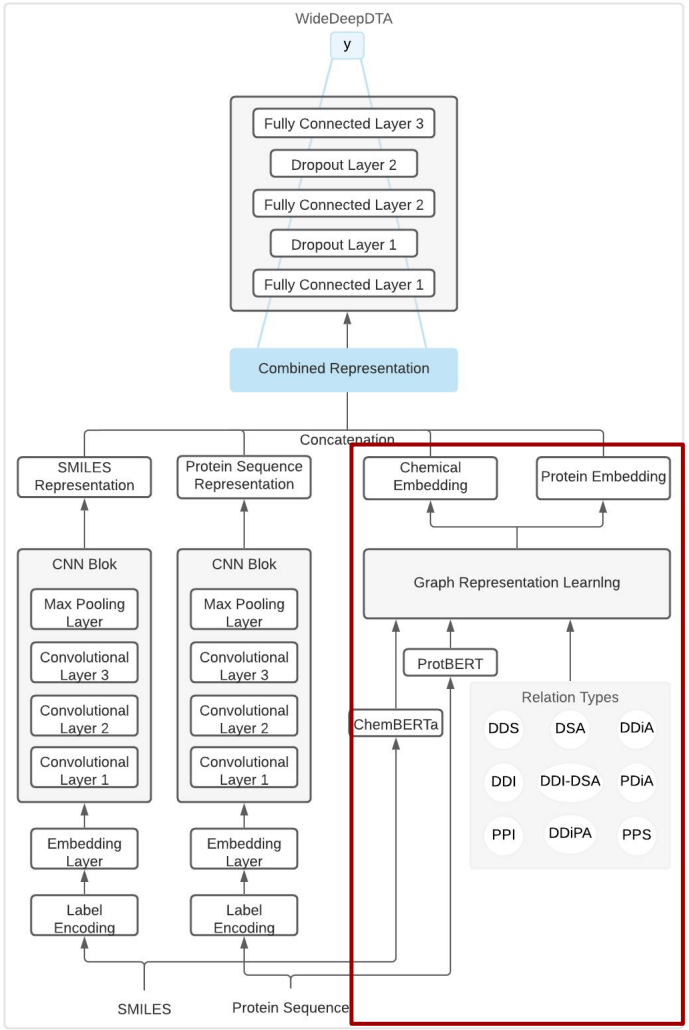
2. Protein Related Information:
    a. Amino acid sequences
    b. Protein-Disease Association
    c. Text-Based Similarity
    d. ProtBERT Embeddings
    e. BioBERT Embeddings for diseases

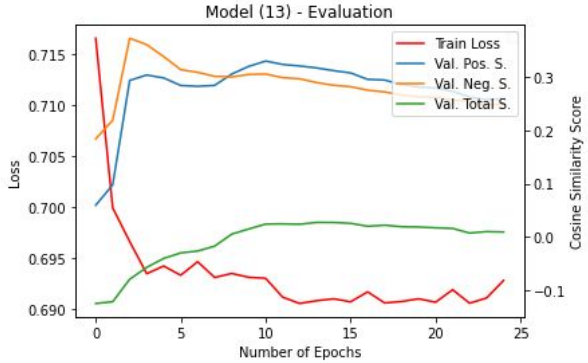
$$Jaccard(A, B) = \frac{|A \cap B|}{|A \cup B|} \times 100$$

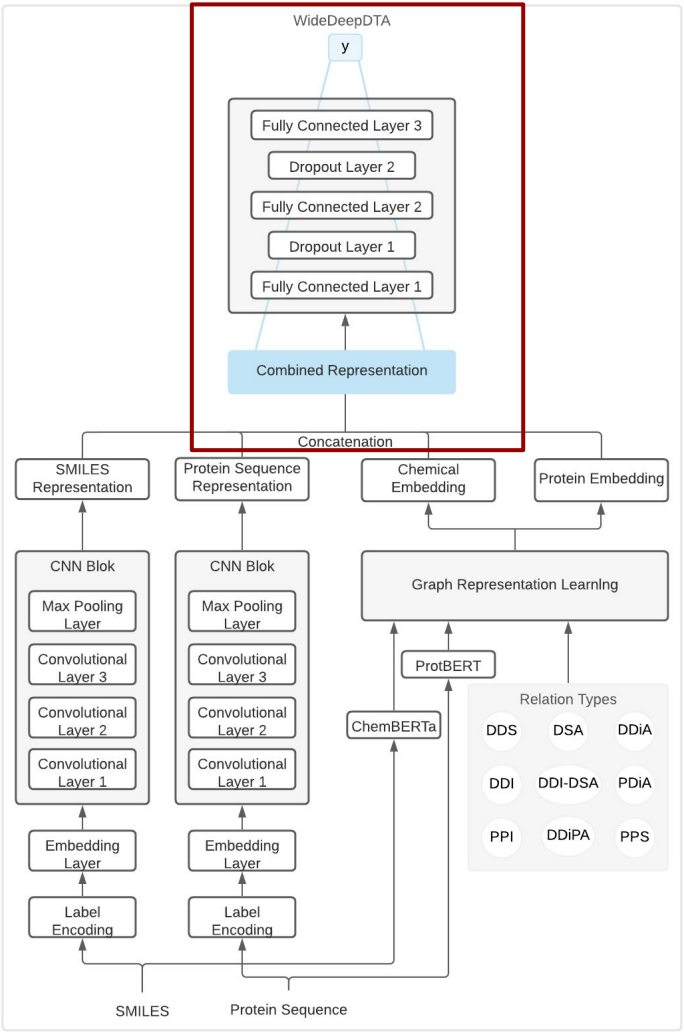



WideDeepDTA
y
Fully Connected Layer 3
Dropout Layer 2
Fully Connected Layer 2
Dropout Layer 1
Fully Connected Layer 1
Combined Representation
Concatenation
SMILES Representation
Protein Sequence Representation
Chemical Embedding
Protein Embedding
CNN Blok
Max Pooling Layer
Convolutional Layer 3
Convolutional Layer 2
Convolutional Layer 1
CNN Blok
Max Pooling Layer
Convolutional Layer 3
Convolutional Layer 2
Convolutional Layer 1
Graph Representation Learnlng
ProtBERT
ChemBERTa
Relation Types
DDS
DSA
DDiA
DDI
DDI-DSA
PDiA
PPI
DDiPA
PPS
Embedding Layer
Label Encoding
Embedding Layer
Label Encoding
SMILES
Protein Sequence


WideDeepDTA
y
Fully Connected Layer 3
Dropout Layer 2
Fully Connected Layer 2
Dropout Layer 1
Fully Connected Layer 1
Combined Representation
Concatenation
SMILES Representation
Protein Sequence Representation
Chemical Embedding
Protein Embedding
CNN Blok
Max Pooling Layer
Convolutional Layer 3
Convolutional Layer 2
Convolutional Layer 1
CNN Blok
Max Pooling Layer
Convolutional Layer 3
Convolutional Layer 2
Convolutional Layer 1
Graph Representation Learnlng
ProtBERT
ChemBERTa
Relation Types
DDS
DSA
DDiA
DDI
DDI-DSA
PDiA
PPI
DDiPA
PPS
Embedding Layer
Embedding Layer
Label Encoding
Label Encoding
SMILES
Protein Sequence


WideDeepDTA
y
Fully Connected Layer 3
Dropout Layer 2
Fully Connected Layer 2
Dropout Layer 1
Fully Connected Layer 1
Combined Representation
Concatenation
SMILES Representation
Protein Sequence Representation
Chemical Embedding
Protein Embedding
CNN Blok
Max Pooling Layer
Convolutional Layer 3
Convolutional Layer 2
Convolutional Layer 1
CNN Blok
Max Pooling Layer
Convolutional Layer 3
Convolutional Layer 2
Convolutional Layer 1
Graph Representation Learnlng
ProtBERT
ChemBERTa
Relation Types
DDS
DSA
DDiA
DDI
DDI-DSA
PDiA
PPI
DDiPA
PPS
Embedding Layer
Embedding Layer
Label Encoding
Label Encoding
SMILES
Protein Sequence


Model (13) - Evaluation
Train Loss
Val. Pos. S.
Val. Neg. S.
Val. Total S.
Loss
Cosine Similarity Score
Number of Epochs
0.715
0.710
0.705
0.700
0.695
0.690
0.3
0.2
0.1
0.0
−0.1
0
5
10
15
20
25

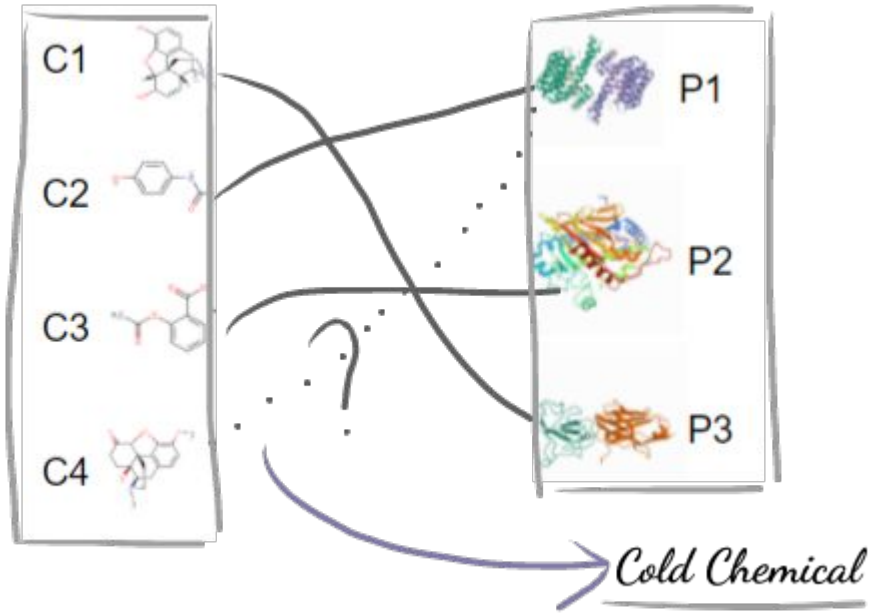

WideDeepDTA
y
Fully Connected Layer 3
Dropout Layer 2
Fully Connected Layer 2
Dropout Layer 1
Fully Connected Layer 1
Combined Representation
Concatenation
SMILES Representation
Protein Sequence Representation
Chemical Embedding
Protein Embedding
CNN Blok
Max Pooling Layer
Convolutional Layer 3
Convolutional Layer 2
Convolutional Layer 1
CNN Blok
Max Pooling Layer
Convolutional Layer 3
Convolutional Layer 2
Convolutional Layer 1
Graph Representation Learnlng
ProtBERT
ChemBERTa
Relation Types
DDS
DSA
DDiA
DDI
DDI-DSA
PDiA
PPI
DDiPA
PPS
Embedding Layer
Label Encoding
Embedding Layer
Label Encoding
SMILES
Protein Sequence

Experiments

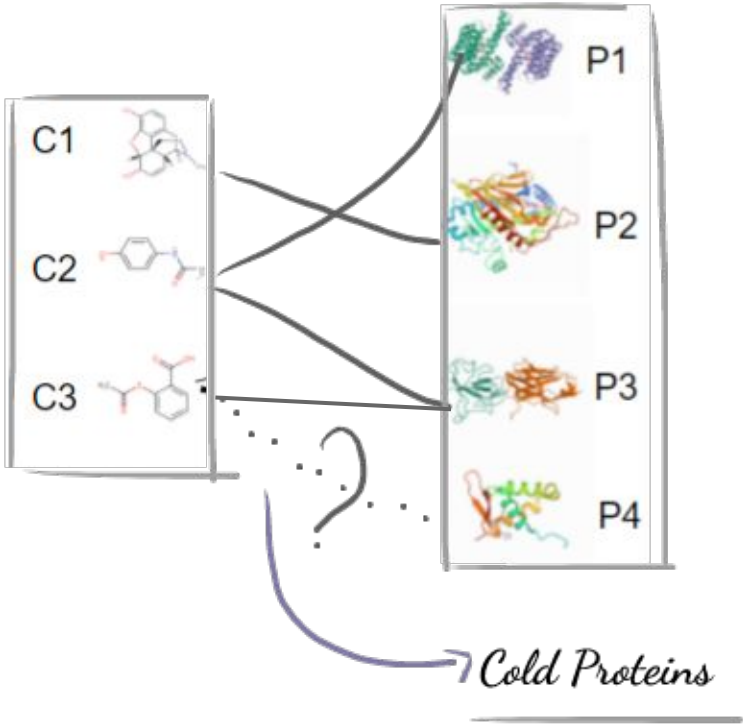
- **BDB:** 490 proteins, 924 ligands, ~30K interactions [1]
- 5-fold cross-validation
- Evaluation: R-squared (R^2) and concordance index (CI)

Özçelik, R., Öztürk, H., Özgür, A., & Ozkirimli, E. (2021). ChemBoost: A Chemical Language Based Approach for Protein–Ligand Binding Affinity Prediction. *Molecular Informatics*, *40*(5), 2000212.

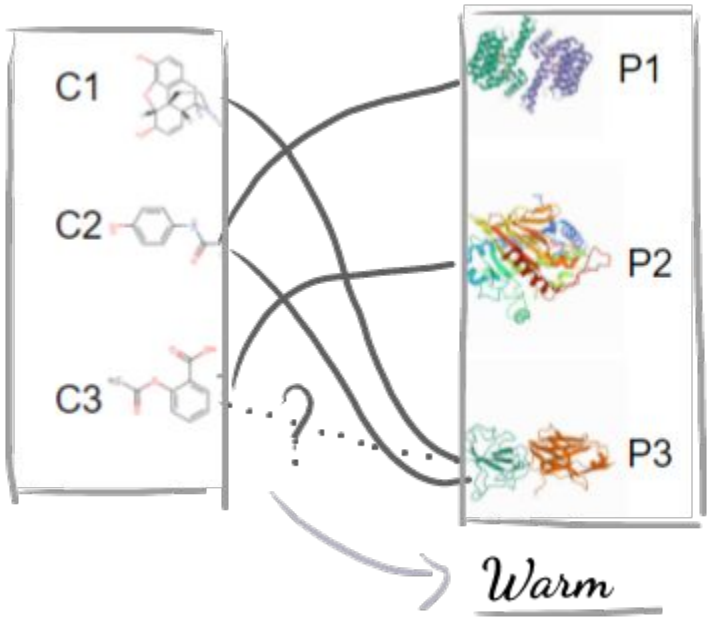
Cold Ligands



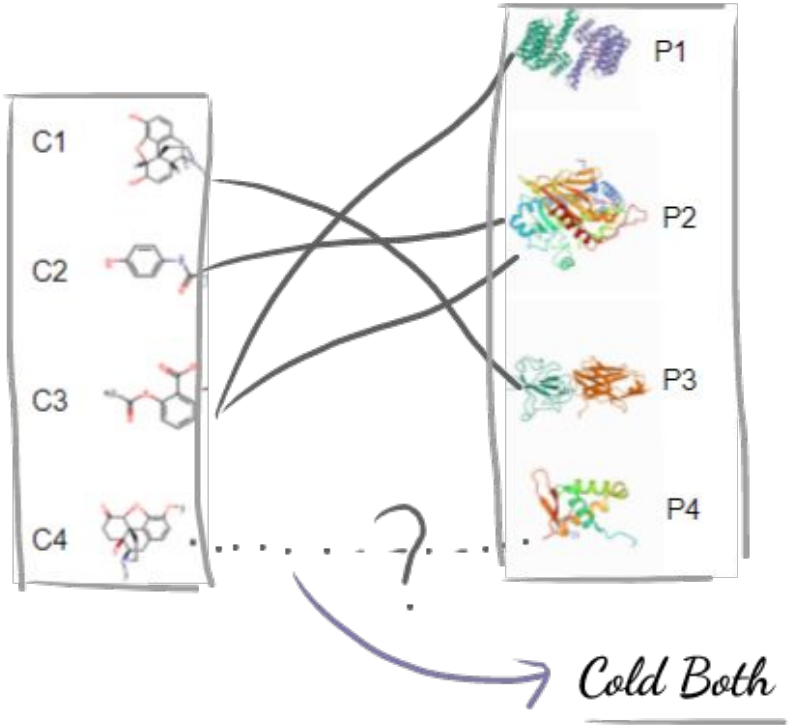
Cold Proteins



Warm Biomolecules

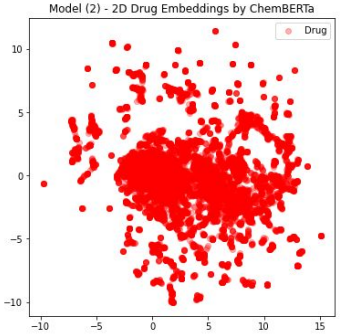


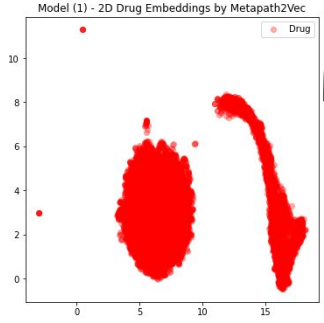
Cold Biomolecules



Results

Biomolecule representation through homogeneous graphs:





Model Name	Number of Nodes	Metapaths	Number of Edges
Model (1), (2)	10935	Drug Interacts with Drug	2196820
Model (3), (4)	6963	Drug Similar to Drug	5848540
Model (5), (6)	12675	Protein Interacts with Protein	124536
Model (7), (8)	465	Protein Similar to Protein	1056

∴ Text-based drug-drug similarity can replace drug-drug Interaction.

		Warm		Cold Ligand		Cold Protein		Cold Both	
	Model	CI	R^2	CI	R^2	CI	R^2	CI	R^2
	DeepDTA	0.888 (0.009)	0.781 (0.028)	**0.687 (0.096)**	**0.039 (0.243)**	0.759 (0.006)	0.315 (0.049)	**0.554 (0.047)**	**-0.154 (0.164)**
DDI	Model1	**0.896 (0.009)**	0.777 (0.026)	0.664 (0.057)	-0.053 (0.210)	**0.779 (0.030)**	**0.375 (0.104)**	0.554 (0.044)	-0.287 (0.184)
DDI	Model2	0.890 (0.014)	0.782 (0.017)	0.640 (0.066)	-0.125 (0.180)	0.768 (0.009)	0.327 (0.072)	0.495 (0.037)	-0.496 (0.211)
DDS	Model3	0.893 (0.005)	0.787 (0.020)	0.651 (0.085)	-0.139 (0.132)	0.775 (0.018)	0.340 (0.082)	0.519 (0.036)	-0.390 (0.235)
DDS	Model4	0.890 (0.006)	**0.789 (0.008)**	0.666 (0.064)	-0.102 (0.330)	0.774 (0.015)	0.327 (0.075)	0.536 (0.068)	-0.274 (0.269)

∴ Text-based drug-drug similarity can replace drug-drug Interaction.

		Warm		Cold Ligand		Cold Protein		Cold Both	
	Model	CI	R^2	CI	R^2	CI	R^2	CI	R^2
	DeepDTA	0.888 (0.009)	0.781 (0.028)	**0.687 (0.096)**	**0.039 (0.243)**	0.759 (0.006)	0.315 (0.049)	**0.554 (0.047)**	**-0.154 (0.164)**
DDI	Model1	**0.896 (0.009)**	0.777 (0.026)	0.664 (0.057)	-0.053 (0.210)	**0.779 (0.030)**	**0.375 (0.104)**	0.554 (0.044)	-0.287 (0.184)
DDI	Model2	0.890 (0.014)	0.782 (0.017)	0.640 (0.066)	-0.125 (0.180)	0.768 (0.009)	0.327 (0.072)	0.495 (0.037)	-0.496 (0.211)
DDS	Model3	0.893 (0.005)	0.787 (0.020)	0.651 (0.085)	-0.139 (0.132)	0.775 (0.018)	0.340 (0.082)	0.519 (0.036)	-0.390 (0.235)
DDS	Model4	0.890 (0.006)	**0.789 (0.008)**	0.666 (0.064)	-0.102 (0.330)	0.774 (0.015)	0.327 (0.075)	0.536 (0.068)	-0.274 (0.269)

∴ Random and PLM initializations are on par with each other.

		Warm		Cold Ligand		Cold Protein		Cold Both	
	Model	CI	R^2	CI	R^2	CI	R^2	CI	R^2
	DeepDTA	0.888 (0.009)	0.781 (0.028)	**0.687 (0.096)**	**0.039 (0.243)**	0.759 (0.006)	0.315 (0.049)	**0.554 (0.047)**	**-0.154 (0.164)**
DDI	Model1	**0.896 (0.009)**	0.777 (0.026)	0.664 (0.057)	-0.053 (0.210)	**0.779 (0.030)**	**0.375 (0.104)**	0.554 (0.044)	-0.287 (0.184)
DDI	Model2	0.890 (0.014)	0.782 (0.017)	0.640 (0.066)	-0.125 (0.180)	0.768 (0.009)	0.327 (0.072)	0.495 (0.037)	-0.496 (0.211)
DDS	Model3	0.893 (0.005)	0.787 (0.020)	0.651 (0.085)	-0.139 (0.132)	0.775 (0.018)	0.340 (0.082)	0.519 (0.036)	-0.390 (0.235)
DDS	Model4	0.890 (0.006)	**0.789 (0.008)**	0.666 (0.064)	-0.102 (0.330)	0.774 (0.015)	0.327 (0.075)	0.536 (0.068)	-0.274 (0.269)

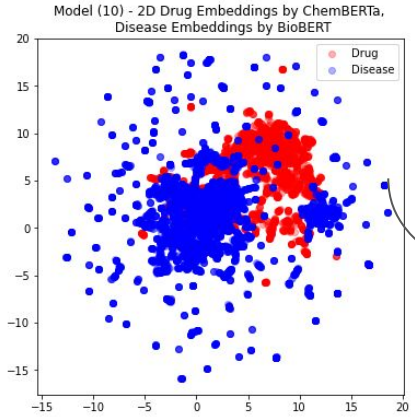
∴ Protein-protein similarity creates the richest representations.

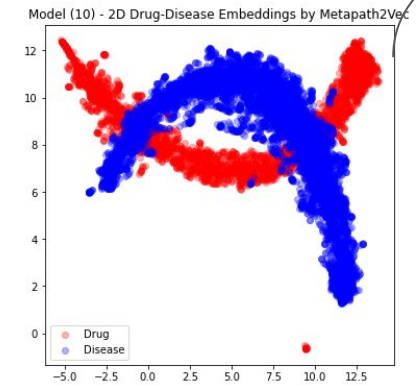
		Warm		Cold Ligand		Cold Protein		Cold Both	
	Model	CI	R^2	CI	R^2	CI	R^2	CI	R^2
	DeepDTA	0.888 (0.009)	0.781 (0.028)	0.687 (0.096)	0.039 (0.243)	0.759 (0.006)	0.315 (0.049)	0.554 (0.047)	-0.154 (0.164)
PPI	Model5	0.888 (0.009)	0.773 (0.025)	0.674 (0.095)	-0.031 (0.290)	0.765 (0.017)	0.323 (0.073)	0.551 (0.049)	-0.269 (0.165)
PPI	Model6	0.888 (0.009)	0.777 (0.012)	0.698 (0.083)	-0.020 (0.347)	0.759 (0.019)	0.299 (0.084)	0.561 (0.031)	-0.363 (0.279)
PPS	Model7	**0.893 (0.006)**	0.775 (0.018)	0.675 (0.083)	-0.067 (0.211)	**0.783 (0.010)**	**0.362 (0.047)**	0.557 (0.036)	-0.260 (0.090)
PPS	Model8	0.892 (0.008)	**0.785 (0.019)**	**0.728 (0.046)**	**0.155 (0.162)**	0.773 (0.022)	0.339 (0.086)	**0.598 (0.061)**	**-0.065 (0.297)**

∴ Random and PLM initializations are on par with each other.

		Warm		Cold Ligand		Cold Protein		Cold Both	
	Model	CI	R^2	CI	R^2	CI	R^2	CI	R^2
	DeepDTA	0.888 (0.009)	0.781 (0.028)	0.687 (0.096)	0.039 (0.243)	0.759 (0.006)	0.315 (0.049)	0.554 (0.047)	-0.154 (0.164)
PPI	Model5	0.888 (0.009)	0.773 (0.025)	0.674 (0.095)	-0.031 (0.290)	0.765 (0.017)	0.323 (0.073)	0.551 (0.049)	-0.269 (0.165)
PPI	Model6	0.888 (0.009)	0.777 (0.012)	0.698 (0.083)	-0.020 (0.347)	0.759 (0.019)	0.299 (0.084)	0.561 (0.031)	-0.363 (0.279)
PPS	Model7	**0.893 (0.006)**	0.775 (0.018)	0.675 (0.083)	-0.067 (0.211)	**0.783 (0.010)**	**0.362 (0.047)**	0.557 (0.036)	-0.260 (0.090)
PPS	Model8	0.892 (0.008)	**0.785 (0.019)**	**0.728 (0.046)**	**0.155 (0.162)**	0.773 (0.022)	0.339 (0.086)	**0.598 (0.061)**	**-0.065 (0.297)**

Heterogeneous representations with disease information





Model Name	Num. of Drug Nodes	Num. of Protein Nodes	Num. of Disease Nodes	Metapaths	Num. of Edges
Model (9) Model (10)	3387	-	7086	Drug Assoc. with Disease Disease Assoc. with Drug	1991308
Model (11) Model (12)	-	32169	126	Protein Assoc.with Disease Disease Assoc.with Protein	64990
Model (13) Model (14)	3387	32169	7087	Drug Assoc. with Disease Disease Assoc.with Protein Protein Assoc.with Disease Disease Assoc.with Drug	2056298

∴ Disease information improves prediction performance over benchmark.

		Warm		Cold Ligand		Cold Protein		Cold Both	
	Model	CI	R^2	CI	R^2	CI	R^2	CI	R^2
	DeepDTA	0.888 (0.009)	0.781 (0.028)	0.687 (0.096)	0.039 (0.243)	0.759 (0.006)	0.315 (0.049)	0.554 (0.047)	-0.154 (0.164)
DDiA	Model9	**0.895 (0.013)**	0.786 (0.023)	0.690 (0.050)	-0.025 (0.202)	0.784 (0.010)	0.349 (0.072)	0.567 (0.047)	-0.187 (0.177)
DDiA	Model10	0.896 (0.006)	0.784 (0.019)	0.695 (0.056)	0.066 (0.122)	0.779 (0.014)	0.333 (0.089)	0.564 (0.031)	-0.201 (0.095)
PDiA	Model11	0.881 (0.007)	0.759 (0.028)	0.695 (0.036)	-0.010 (0.201)	0.762 (0.017)	0.238 (0.233)	0.543 (0.068)	-0.475 (0.506)
PDiA	Model12	**0.895 (0.007)**	**0.794 (0.015)**	0.702 (0.046)	0.030 (0.169)	**0.785 (0.011)**	**0.377 (0.078)**	0.568 (0.027)	**-0.060 (0.146)**
DDiPA	Model13	0.878 (0.009)	0.753 (0.020)	0.664 (0.087)	-0.040 (0.097)	0.749 (0.018)	0.268 (0.061)	0.533 (0.063)	-0.300 (0.161)
DDiPA	Model14	0.882 (0.008)	0.755 (0.022)	**0.703 (0.033)**	**0.030 (0.082)**	0.752 (0.013)	0.282 (0.073)	**0.578 (0.050)**	-0.275 (0.123)

∴ Disease information improves prediction performance over benchmark.

		Warm		Cold Ligand		Cold Protein		Cold Both	
	Model	CI	R^2	CI	R^2	CI	R^2	CI	R^2
	DeepDTA	0.888 (0.009)	0.781 (0.028)	0.687 (0.096)	0.039 (0.243)	0.759 (0.006)	0.315 (0.049)	0.554 (0.047)	-0.154 (0.164)
DDiA	Model9	**0.895 (0.013)**	0.786 (0.023)	0.690 (0.050)	-0.025 (0.202)	0.784 (0.010)	0.349 (0.072)	0.567 (0.047)	-0.187 (0.177)
DDiA	Model10	0.896 (0.006)	0.784 (0.019)	0.695 (0.056)	0.066 (0.122)	0.779 (0.014)	0.333 (0.089)	0.564 (0.031)	-0.201 (0.095)
PDiA	Model11	0.881 (0.007)	0.759 (0.028)	0.695 (0.036)	-0.010 (0.201)	0.762 (0.017)	0.238 (0.233)	0.543 (0.068)	-0.475 (0.506)
PDiA	Model12	**0.895 (0.007)**	**0.794 (0.015)**	0.702 (0.046)	0.030 (0.169)	**0.785 (0.011)**	**0.377 (0.078)**	0.568 (0.027)	**-0.060 (0.146)**
DDiPA	Model13	0.878 (0.009)	0.753 (0.020)	0.664 (0.087)	-0.040 (0.097)	0.749 (0.018)	0.268 (0.061)	0.533 (0.063)	-0.300 (0.161)
DDiPA	Model14	0.882 (0.008)	0.755 (0.022)	**0.703 (0.033)**	**0.030 (0.082)**	0.752 (0.013)	0.282 (0.073)	**0.578 (0.050)**	-0.275 (0.123)

∴ PLM initialization improves performance of heterogeneous graphs.

		Warm		Cold Ligand		Cold Protein		Cold Both	
	Model	CI	R^2	CI	R^2	CI	R^2	CI	R^2
	DeepDTA	0.888 (0.009)	0.781 (0.028)	0.687 (0.096)	0.039 (0.243)	0.759 (0.006)	0.315 (0.049)	0.554 (0.047)	-0.154 (0.164)
DDiA	Model9	**0.895 (0.013)**	0.786 (0.023)	0.690 (0.050)	-0.025 (0.202)	0.784 (0.010)	0.349 (0.072)	0.567 (0.047)	-0.187 (0.177)
DDiA	Model10	0.896 (0.006)	0.784 (0.019)	0.695 (0.056)	0.066 (0.122)	0.779 (0.014)	0.333 (0.089)	0.564 (0.031)	-0.201 (0.095)
PDiA	Model11	0.881 (0.007)	0.759 (0.028)	0.695 (0.036)	-0.010 (0.201)	0.762 (0.017)	0.238 (0.233)	0.543 (0.068)	-0.475 (0.506)
PDiA	Model12	**0.895 (0.007)**	**0.794 (0.015)**	0.702 (0.046)	0.030 (0.169)	**0.785 (0.011)**	**0.377 (0.078)**	0.568 (0.027)	**-0.060 (0.146)**
DDiPA	Model13	0.878 (0.009)	0.753 (0.020)	0.664 (0.087)	-0.040 (0.097)	0.749 (0.018)	0.268 (0.061)	0.533 (0.063)	-0.300 (0.161)
DDiPA	Model14	0.882 (0.008)	0.755 (0.022)	**0.703 (0.033)**	**0.030 (0.082)**	0.752 (0.013)	0.282 (0.073)	**0.578 (0.050)**	-0.275 (0.123)

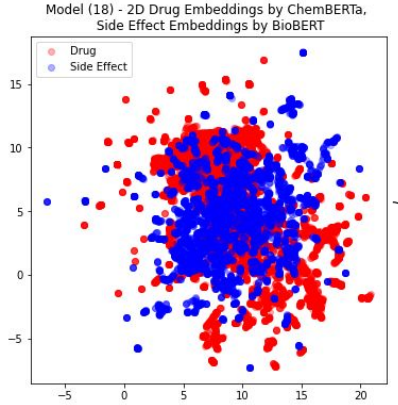
∴ Connecting drugs and proteins through diseases lowers performance.

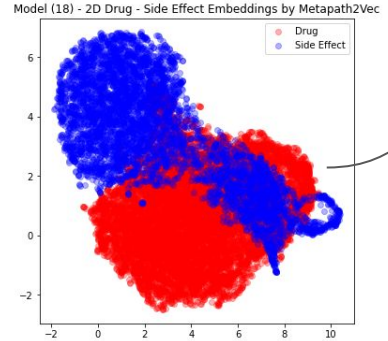
		Warm		Cold Ligand		Cold Protein		Cold Both	
	Model	CI	R^2	CI	R^2	CI	R^2	CI	R^2
	DeepDTA	0.888 (0.009)	0.781 (0.028)	0.687 (0.096)	0.039 (0.243)	0.759 (0.006)	0.315 (0.049)	0.554 (0.047)	-0.154 (0.164)
DDiA	Model9	**0.895 (0.013)**	0.786 (0.023)	0.690 (0.050)	-0.025 (0.202)	0.784 (0.010)	0.349 (0.072)	0.567 (0.047)	-0.187 (0.177)
DDiA	Model10	0.896 (0.006)	0.784 (0.019)	0.695 (0.056)	0.066 (0.122)	0.779 (0.014)	0.333 (0.089)	0.564 (0.031)	-0.201 (0.095)
PDiA	Model11	0.881 (0.007)	0.759 (0.028)	0.695 (0.036)	-0.010 (0.201)	0.762 (0.017)	0.238 (0.233)	0.543 (0.068)	-0.475 (0.506)
PDiA	Model12	**0.895 (0.007)**	**0.794 (0.015)**	0.702 (0.046)	0.030 (0.169)	**0.785 (0.011)**	**0.377 (0.078)**	0.568 (0.027)	**-0.060 (0.146)**
DDiPA	Model13	0.878 (0.009)	0.753 (0.020)	0.664 (0.087)	-0.040 (0.097)	0.749 (0.018)	0.268 (0.061)	0.533 (0.063)	-0.300 (0.161)
DDiPA	Model14	0.882 (0.008)	0.755 (0.022)	**0.703 (0.033)**	**0.030 (0.082)**	0.752 (0.013)	0.282 (0.073)	**0.578 (0.050)**	-0.275 (0.123)

∴ Connecting drugs and proteins through diseases lowers performance.

		Warm		Cold Ligand		Cold Protein		Cold Both	
	Model	CI	R^2	CI	R^2	CI	R^2	CI	R^2
	DeepDTA	0.888 (0.009)	0.781 (0.028)	0.687 (0.096)	0.039 (0.243)	0.759 (0.006)	0.315 (0.049)	0.554 (0.047)	-0.154 (0.164)
DDiA	Model9	**0.895 (0.013)**	0.786 (0.023)	0.690 (0.050)	-0.025 (0.202)	0.784 (0.010)	0.349 (0.072)	0.567 (0.047)	-0.187 (0.177)
DDiA	Model10	0.896 (0.006)	0.784 (0.019)	0.695 (0.056)	0.066 (0.122)	0.779 (0.014)	0.333 (0.089)	0.564 (0.031)	-0.201 (0.095)
PDiA	Model11	0.881 (0.007)	0.759 (0.028)	0.695 (0.036)	-0.010 (0.201)	0.762 (0.017)	0.238 (0.233)	0.543 (0.068)	-0.475 (0.506)
PDiA	Model12	**0.895 (0.007)**	**0.794 (0.015)**	0.702 (0.046)	0.030 (0.169)	**0.785 (0.011)**	**0.377 (0.078)**	0.568 (0.027)	**-0.060 (0.146)**
DDiPA	Model13	0.878 (0.009)	0.753 (0.020)	0.664 (0.087)	-0.040 (0.097)	0.749 (0.018)	0.268 (0.061)	0.533 (0.063)	-0.300 (0.161)
DDiPA	Model14	0.882 (0.008)	0.755 (0.022)	**0.703 (0.033)**	**0.030 (0.082)**	0.752 (0.013)	0.282 (0.073)	**0.578 (0.050)**	-0.275 (0.123)

Heterogeneous representations with side effect information





Model Name	Number of Drug Nodes	Number of Side Effect Nodes	Metapaths	Number of Edges
Model (15), Model (16)	1003	5451	Drug Associates with Side Effect Side Effect Associates with Drug	231742
Model (17) Model (18)	10935	5451	Drug Associates with Side Effect Side Effect Associates with Drug Drug Interacts with Drug	2427902

∴ Side effect information improves prediction performance over benchmark.

		Warm		Cold Ligand		Cold Protein		Cold Both	
	Model	CI	R^2	CI	R^2	CI	R^2	CI	R^2
	DeepDTA	0.888 (0.009)	0.781 (0.028)	0.687 (0.096)	0.039 (0.243)	0.759 (0.006)	0.315 (0.049)	0.554 (0.047)	-0.154 (0.164)
DSA	Model15	0.898 (0.009)	**0.791 (0.016)**	0.683 (0.040)	-0.074 (0.265)	0.774 (0.015)	0.358 (0.059)	0.561 (0.045)	-0.278 (0.293)
DSA	Model16	0.889 (0.011)	0.776 (0.021)	0.664 (0.068)	-0.196 (0.196)	0.766 (0.022)	0.323 (0.097)	0.580 (0.052)	-0.445 (0.284)
DDI-DSA	Model17	0.892 (0.013)	0.781 (0.021)	0.688 (0.080)	-0.034 (0.245)	0.769 (0.019)	0.352 (0.084)	0.569 (0.041)	-0.303 (0.244)
DDI-DSA	Model18	**0.899 (0.008)**	0.785 (0.017)	**0.711 (0.057)**	**0.217 (0.153)**	**0.778 (0.014)**	**0.365 (0.069)**	**0.607 (0.061)**	**-0.018 (0.108)**

∴ Integrating side effect relation and drug-drug interaction creates the richest representations.

		Warm		Cold Ligand		Cold Protein		Cold Both	
	Model	CI	R^2	CI	R^2	CI	R^2	CI	R^2
	DeepDTA	0.888 (0.009)	0.781 (0.028)	0.687 (0.096)	0.039 (0.243)	0.759 (0.006)	0.315 (0.049)	0.554 (0.047)	-0.154 (0.164)
DSA	Model15	0.898 (0.009)	**0.791 (0.016)**	0.683 (0.040)	-0.074 (0.265)	0.774 (0.015)	0.358 (0.059)	0.561 (0.045)	-0.278 (0.293)
DSA	Model16	0.889 (0.011)	0.776 (0.021)	0.664 (0.068)	-0.196 (0.196)	0.766 (0.022)	0.323 (0.097)	0.580 (0.052)	-0.445 (0.284)
DDI-DSA	Model17	0.892 (0.013)	0.781 (0.021)	0.688 (0.080)	-0.034 (0.245)	0.769 (0.019)	0.352 (0.084)	0.569 (0.041)	-0.303 (0.244)
DDI-DSA	Model18	**0.899 (0.008)**	0.785 (0.017)	**0.711 (0.057)**	**0.217 (0.153)**	**0.778 (0.014)**	**0.365 (0.069)**	**0.607 (0.061)**	**-0.018 (0.108)**

DDI	Model1	**0.896 (0.009)**	0.777 (0.026)	0.664 (0.057)	-0.053 (0.210)	**0.779 (0.030)**	**0.375 (0.104)**	0.554 (0.044)	-0.287 (0.184)
DDI	Model2	0.890 (0.014)	0.782 (0.017)	0.640 (0.066)	-0.125 (0.180)	0.768 (0.009)	0.327 (0.072)	0.495 (0.037)	-0.496 (0.211)

∴ PLM initialization improves performance of heterogeneous graphs.

		Warm		Cold Ligand		Cold Protein		Cold Both	
	Model	CI	R^2	CI	R^2	CI	R^2	CI	R^2
	DeepDTA	0.888 (0.009)	0.781 (0.028)	0.687 (0.096)	0.039 (0.243)	0.759 (0.006)	0.315 (0.049)	0.554 (0.047)	-0.154 (0.164)
DSA	Model15	0.898 (0.009)	**0.791 (0.016)**	0.683 (0.040)	-0.074 (0.265)	0.774 (0.015)	0.358 (0.059)	0.561 (0.045)	-0.278 (0.293)
DSA	Model16	0.889 (0.011)	0.776 (0.021)	0.664 (0.068)	-0.196 (0.196)	0.766 (0.022)	0.323 (0.097)	0.580 (0.052)	-0.445 (0.284)
DDI-DSA	Model17	0.892 (0.013)	0.781 (0.021)	0.688 (0.080)	-0.034 (0.245)	0.769 (0.019)	0.352 (0.084)	0.569 (0.041)	-0.303 (0.244)
DDI-DSA	Model18	**0.899 (0.008)**	0.785 (0.017)	**0.711 (0.057)**	**0.217 (0.153)**	**0.778 (0.014)**	**0.365 (0.069)**	**0.607 (0.061)**	**-0.018 (0.108)**

Key Conclusions

- A novel and successful DTA prediction framework is proposed.
- 1D similarity of biomolecules can replace biomolecule interaction information in homogeneous graphs.
- Language model-based vector initialization improves heterogeneous network representations.
- Increasing the heterogeneity of the graph generates richer representations.
- Disease and side effect information yield the largest improvement over the benchmark.

Future Work

- The limited number of protein-protein similarity data limits WideDeepDTA's overall performance.
- Experiments revealed limitations of heterogeneous networks' with pre-trained language models to represent the long protein sequences.
- We encourage further studies to integrate more text-based features into the graphs.

Acknowledgements

Abdullatif Köksal

Asu Büşra Temizer

Cansu Yılmaz

Gökçe Uludoğan

Gönül Aycı

Hakime Öztürk

Hilal Dönmez

Rıza Özçelik

Kutlu Ülgen

Nilgün Kayalı

Nural Özel

Olcay Taner Yıldız

Ömer Benzer

Tuna Tuğcu

Tunga Güngör

Veyis Turgut

TÜBİTAK

TUBITAK ARDEB - 119E133 is gratefully acknowledged

Thank you!