Data-Driven AI Frameworks toward Computational Drug Discovery

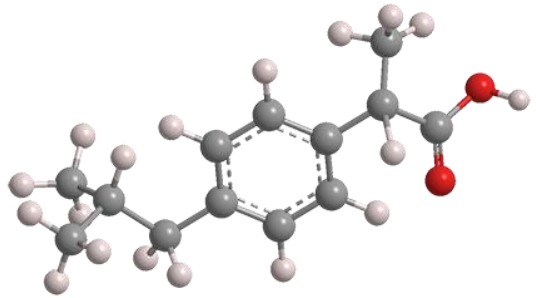
Md Hossain Shuvo, Ph.D.
Assistant Professor
Department of Computer Science
Prairie View A&M University

Oct 8, 2025

Introduction

What is drug?

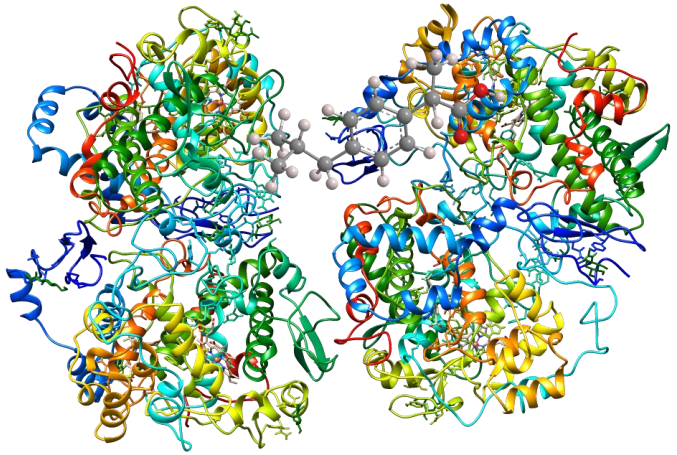
A drug is a chemical substance that alters the normal operations of an organism's body



Ibuprofen: C13H18O2

How does a drug work?

A drug works by binding to a specific biological target to block its activity

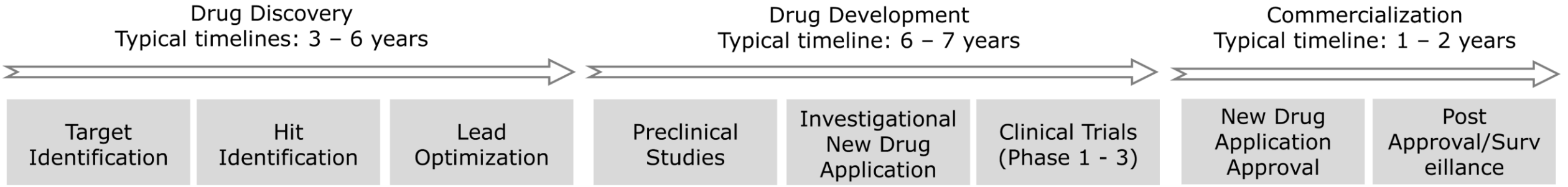


CYCLOOXYGENASE-2
PDB ID: 1CX2

Computational Drug discovery

Computational drug discovery leverages AI and advanced computing power to identify, design, and optimize potential drug candidates before laboratory testing

Drug discovery and development

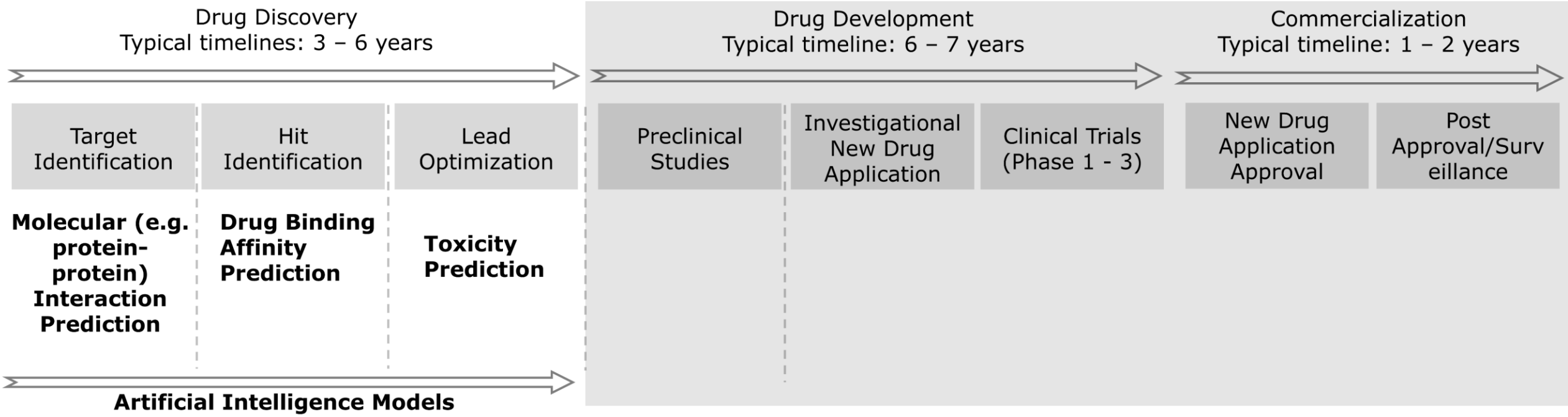


- Drug development
  - A lengthy and costly process
  - Takes ~15 years to bring a new drug to market
  - Development costs can exceed $1–2.6B per drug
  - Failure rates as high as 90%

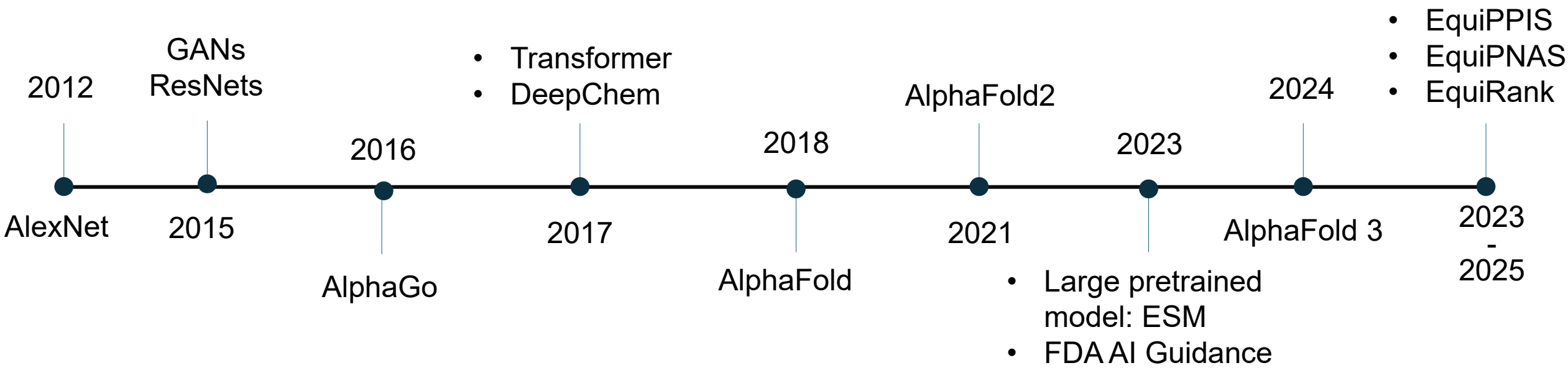
- Eroom's Law
  - Moore's Law states that computing power grows exponentially.
  - Eroom's Law states that drug development costs double roughly every nine years.
  - Reverse Moore's Law highlights the increasing inefficiency of traditional drug development.

H., & Alam, T. (2023)

Computational approaches can accelerate the process

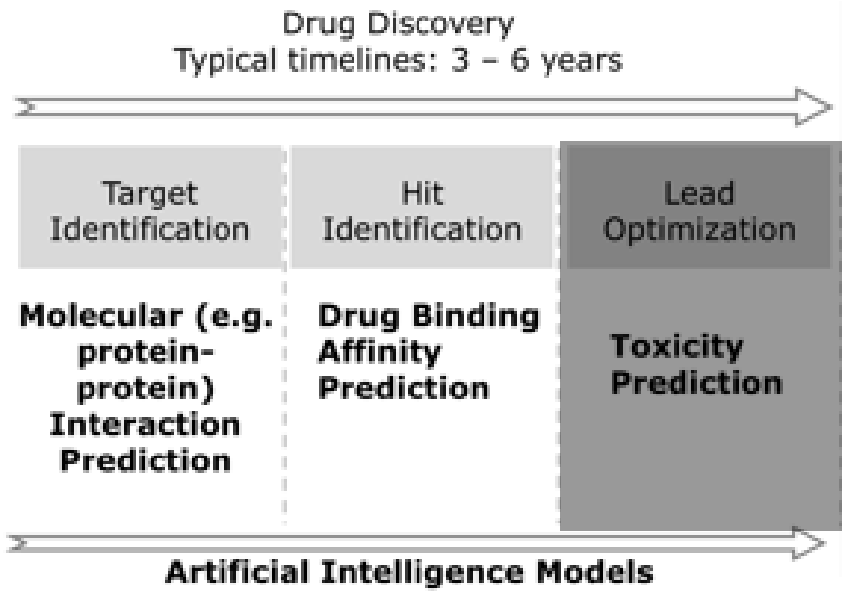


The rise of AI and deep learning in drug discovery

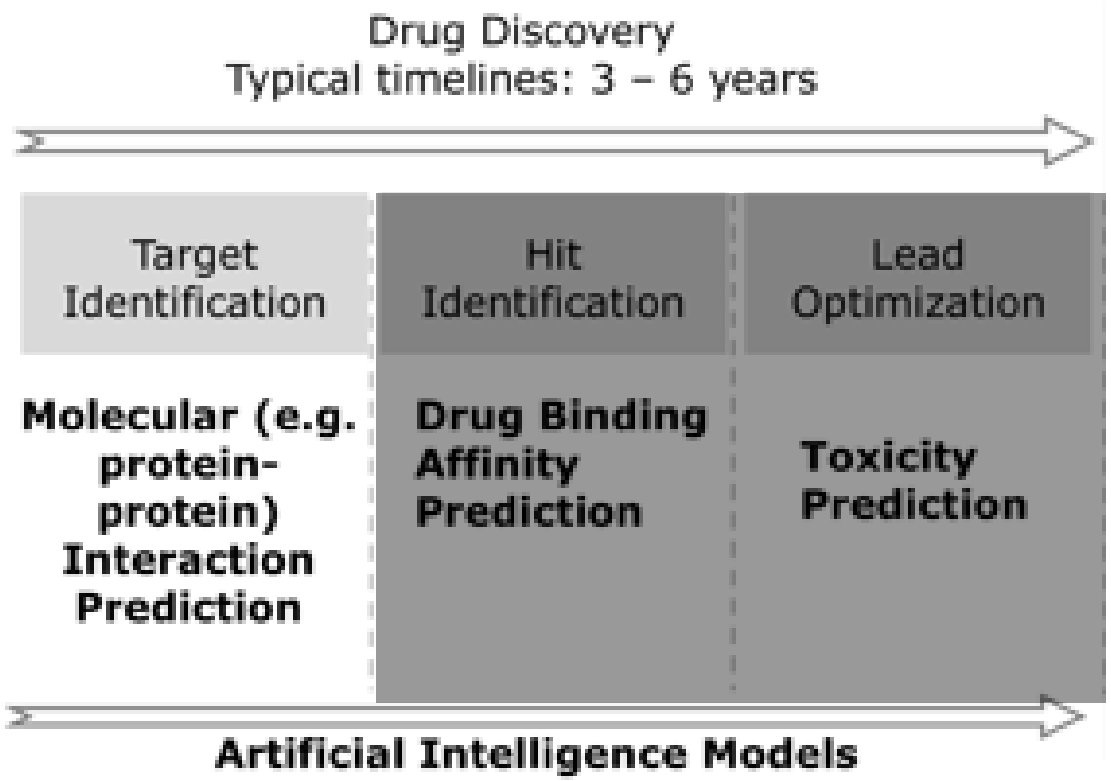


In this presentation…

- Drug Discovery
  - Target identification
    - Methods for protein-protein/nucleic acids interaction prediction
    - **Our methods:** EquiPPIS, EquiPNAS, EquiRank
  - Hit identification
    - Methods for drug-target interactions/affinity prediction
    - Representative models: **GraphDTA**, **iNGNN-DTI**,



Target Identification

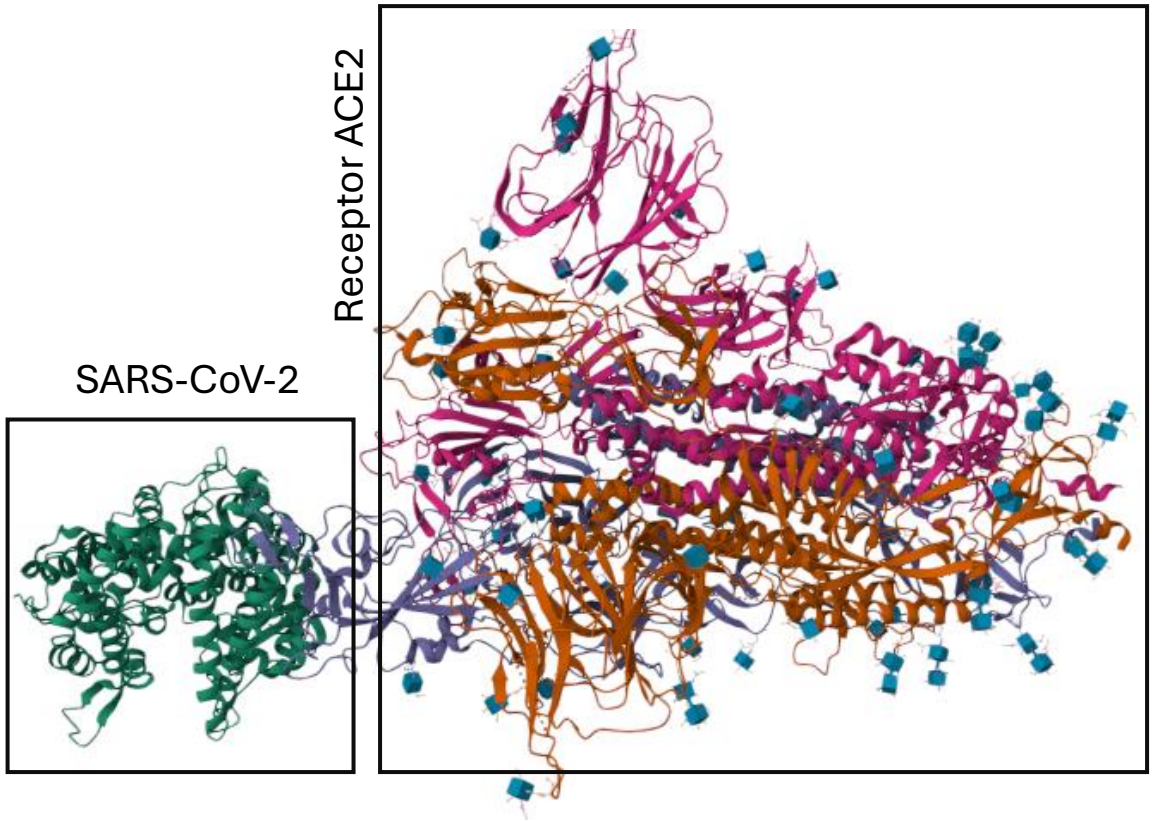


Target identification

- Target identification is the process of finding biological targets, such as proteins, nucleic acids, or interaction interfaces, that play a key role in a disease and could be modulated for therapeutic purposes
- In vivo/experimental methods (low throughput)
  - Yeast two-hybrid
  - CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats)
- In silico / computational methods
  - Deep learning methods that can predict a potential targets:
    - Protein–protein interactions (our method: EquiPPIS)
    - Protein–nucleic acid interactions (our method: EquiPNAS)
    - Ranking scores or probabilities for candidate interactions (Our method: EquiRank)

1. Protein-Protein interaction prediction

- Protein–Protein Interaction (PPI) prediction identifies how proteins bind and interact
- Discover and map critical protein–protein interactions that drive
- Blocking the interaction can prevent infection and halt disease progression.



1. EquiPPIS: Protein-Protein interaction site prediction

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

E(3) equivariant graph neural networks for robust and accurate protein-protein interaction site prediction

Rahmatullah Roche, Bernard Moussad, Md Hossain Shuvo, Debswapna Bhattacharya

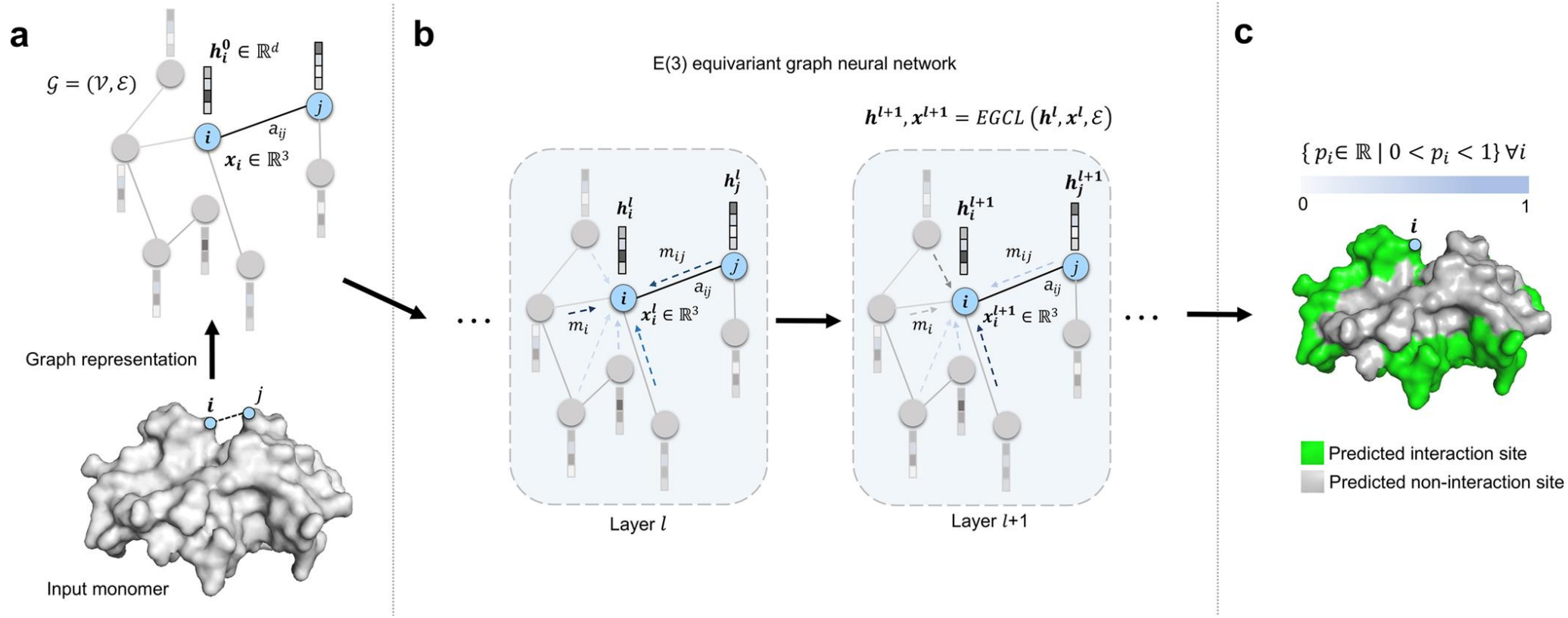
Version 2 Published: August 31, 2023 • https://doi.org/10.1371/journal.pcbi.1011435

EquiPPIS is freely accessible at https://github.com/Bhattacharya-Lab/EquiPPIS

Roche et al., 2023

Drug discovery task
Target identification

1. EquiPPIS: Method architecture



- Trained on 335 redundancy removed proteins collected from three public PPI datasets (Dset_186, Dset_72, Dset_164)
- **Notable features:** Evolutionary, Language model (ESM2), and Orientation and Geometry

Drug discovery task
Target identification

1. EquiPPIS: Results

Tested on 60 non-redundant proteins from Dset_186, Dset_72, and Dset_164

Method	Accuracy	Precision	Recall	F1	MCC	ROC-AUC	PR-AUC
PSIVER	0.561	0.188	0.534	0.278	0.074	0.573	0.190
ProNA2020	0.738	0.275	0.402	0.326	0.176	N/A	N/A
SCRIBER	0.667	0.253	0.568	0.350	0.193	0.665	0.278
DLPred	0.682	0.264	0.565	0.360	0.208	0.677	0.294
DELPHI	0.697	0.276	0.568	0.372	0.225	0.699	0.319
DeepPPISP	0.657	0.243	0.539	0.335	0.167	0.653	0.276
SPPIDER	0.752	0.331	0.557	0.415	0.285	0.755	0.373
MaSIF-site	0.780	0.370	0.561	0.446	0.326	0.775	0.439
GraphPPIS	0.776	0.368	0.584	0.451	0.333	0.786	0.429
EquiPPIS	**0.787**	**0.389**	**0.615**	**0.477**	**0.366**	**0.805**	**0.467**

Note: Except EquiPPS, results for the other methods are obtained directly from the published work of GraphPPIS; values in bold represent the best performance.

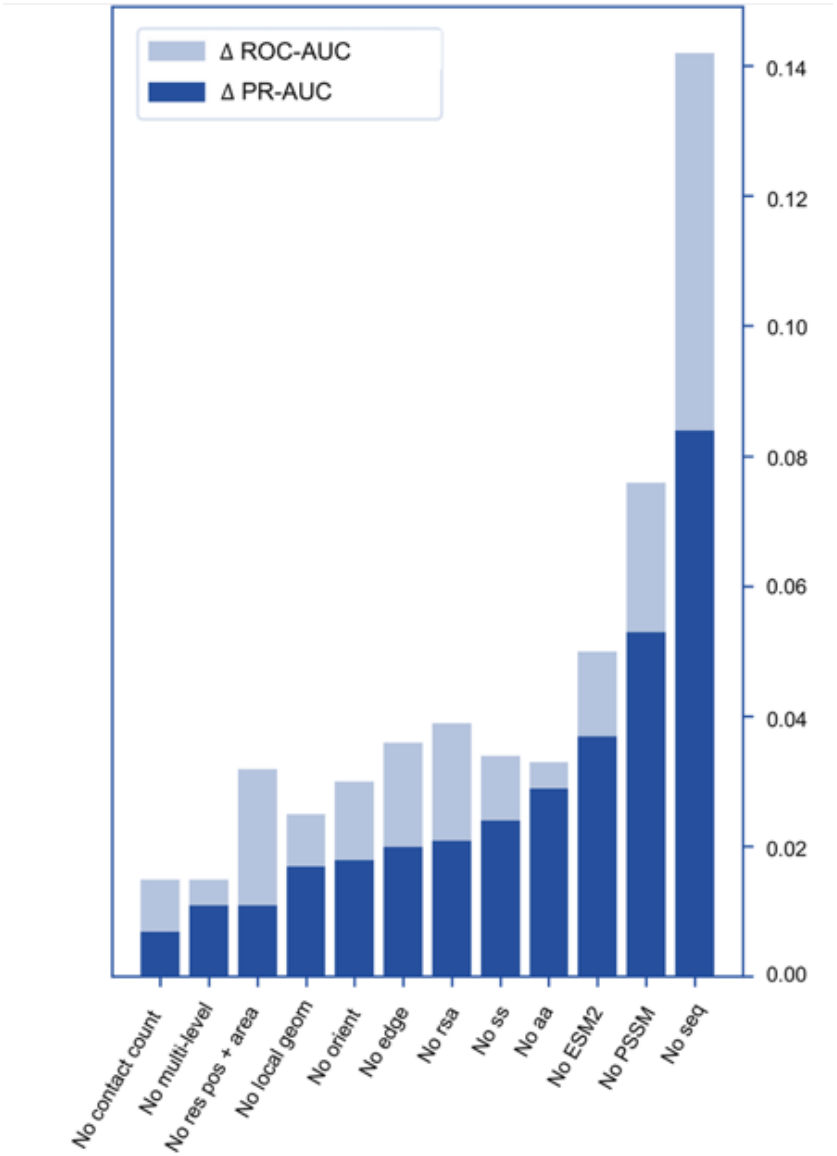
https://doi.org/10.1371/journal.pcbi.1011435.t001

EquiPPIS outperforms state-of-the-art approaches

Drug discovery task
Target identification

1. EquiPPIS: Model interpretation

- All features in EquiPPIS contribute to its accuracy
- Sequence-based features like PSSM and ESM2 are more important than simple residue types, and removing them significantly reduces performance.
- Structural and Physicochemical features have complementary contribution to the improved performance

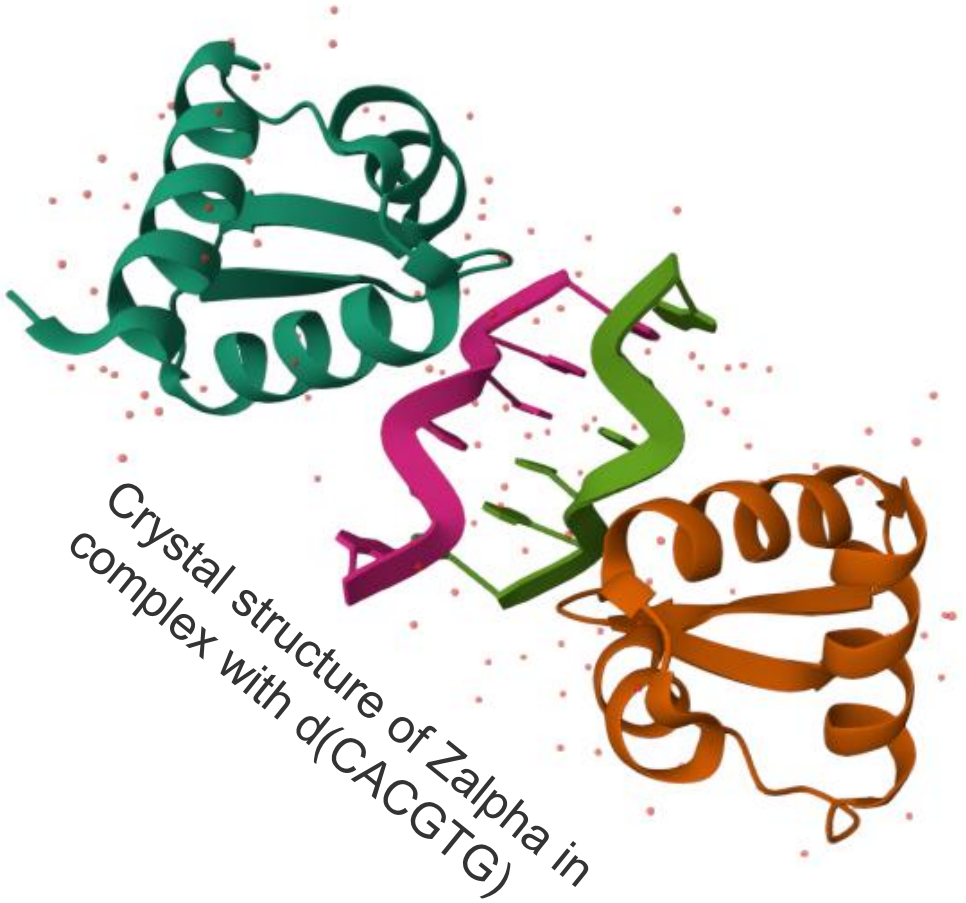


PSSM: Position Specific Sequence Matrix
ESM: Evolutionary Scale Modeling

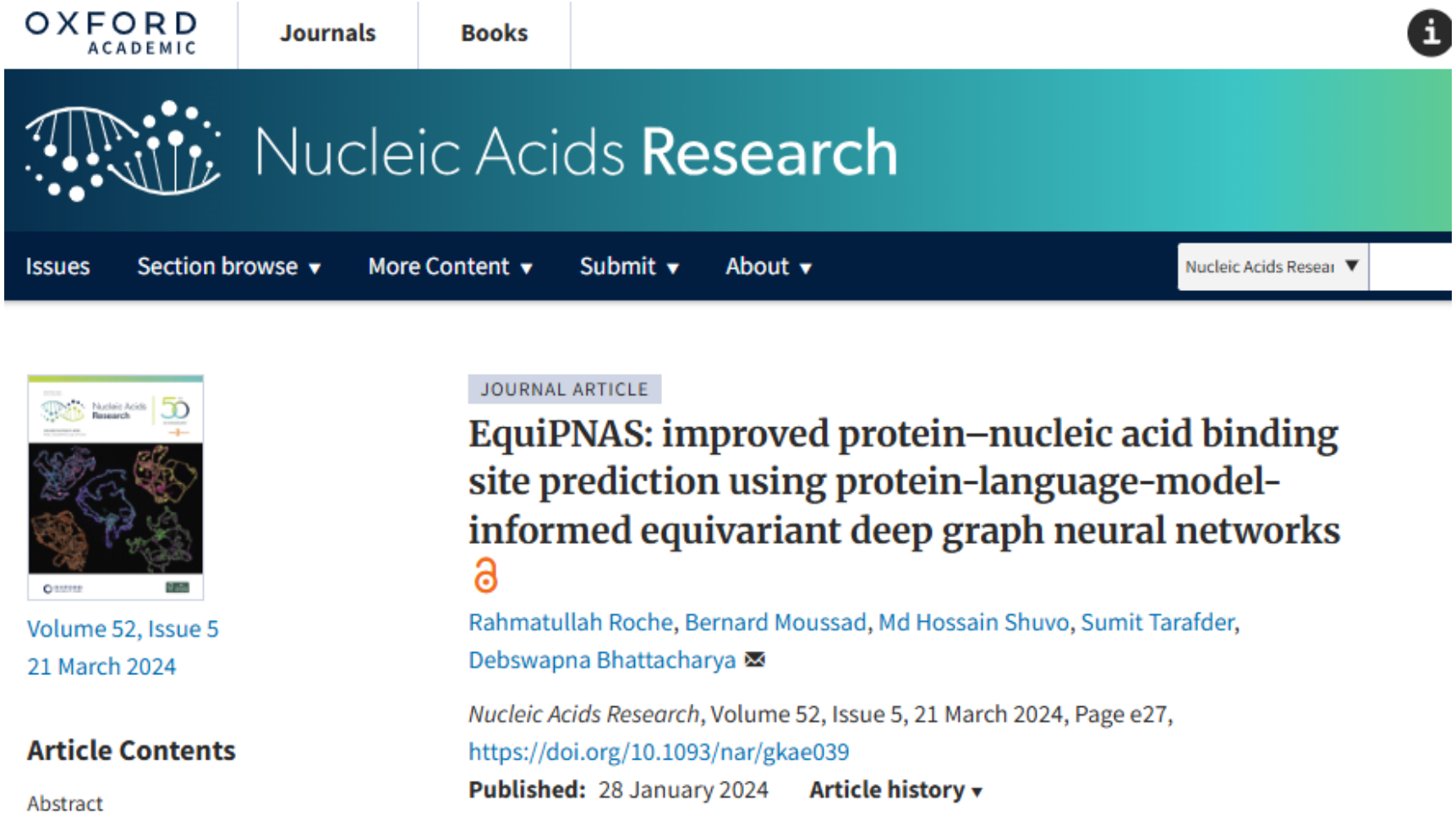
Drug discovery task
Target identification

2. Protein-Nucleic Acids interaction prediction

- Protein–Nucleic acid interaction prediction shows exactly where proteins bind to DNA or RNA
- helps identify important interactions that control gene activity or allow viruses to replicate
- Stopping these interactions can block disease



2. Protein-Nucleic Acids interaction side prediction

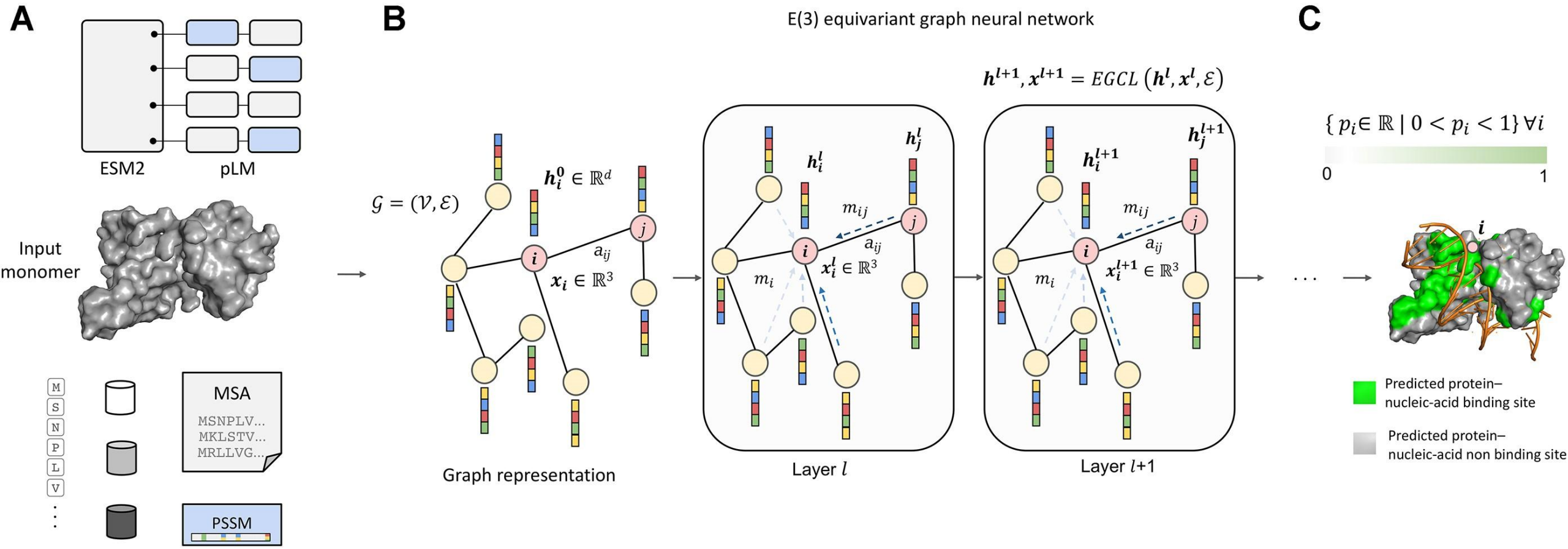


EquiPNAS is freely available at https://github.com/Bhattacharya-Lab/EquiPNAS

Roche et al., 2023

Drug discovery task
Target identification

2. EquiPNAS: Method architecture



- Trained on:
  - Protein–DNA with 573 unique protein chains from the BioLiP database
  - Protein–RNA with 495 unique protein chains from the BioLiP database

Drug discovery task
Target identification

2. EquiPNAS: Results

Tested on protein–DNA using the Test_129 and Test_181 datasets, and protein–RNA using the Test_117 dataset

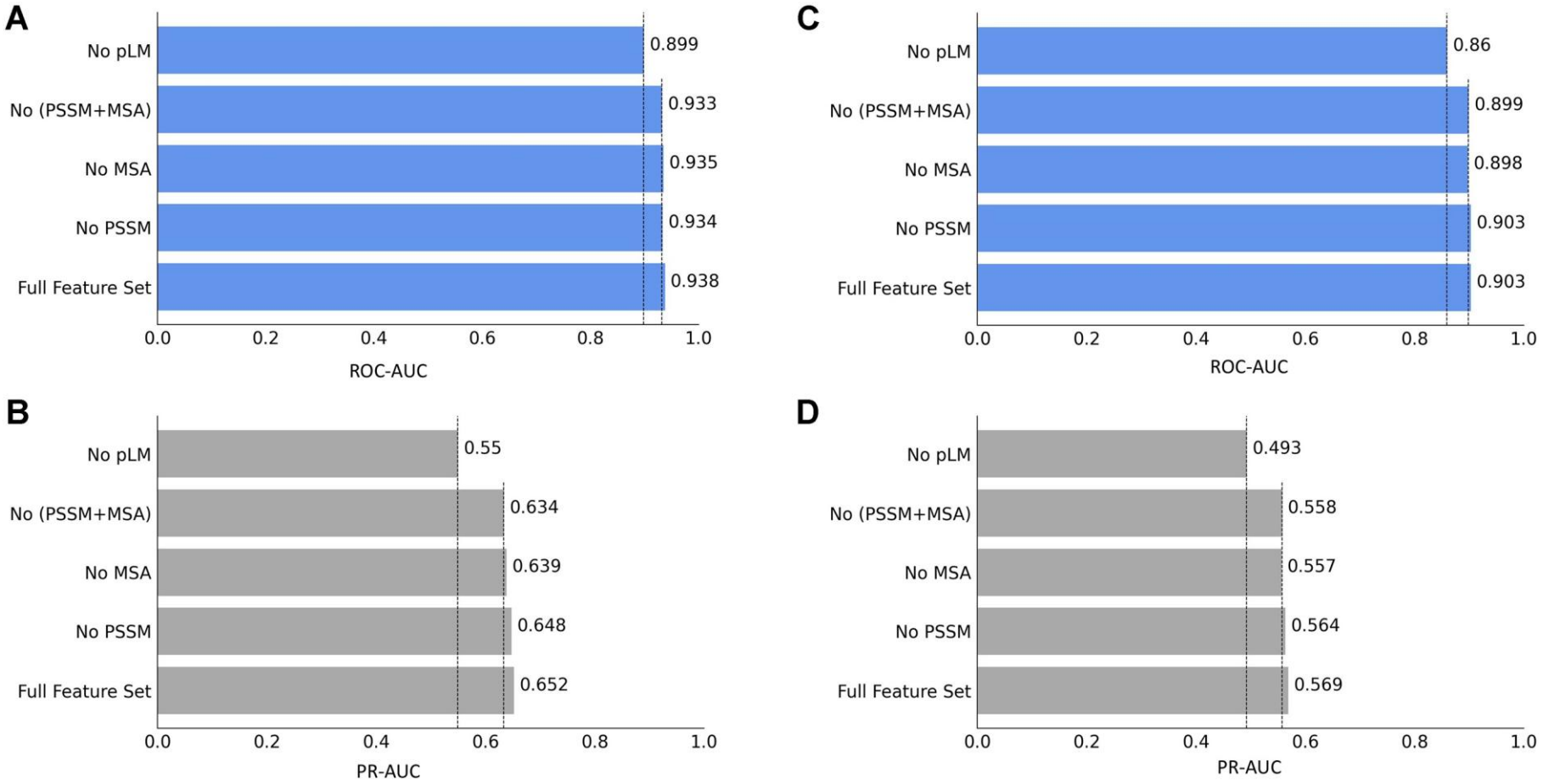
	Datasets	Methods	ROC-AUC	PR-AUC
Protein–DNA	Test_129	GraphBind*	0.916	0.497
		GraphSite*	0.934	0.544
		EquiPNAS	**0.940**	**0.569**
	Test_181	GraphBind*	0.893	0.317
		GraphSite*	0.917	0.369
		EquiPNAS	**0.918**	**0.384**
Protein–RNA	Test_117	GraphBind	0.793	0.204
		EquiPNAS	**0.886**	**0.320**

Note: * Results are obtained directly from the published work of GraphSite.

EquiPNAS consistently outperforms the state-of-the-art methods on both protein–DNA and protein–RNA binding site prediction tasks

Drug discovery task
Target identification

2. EquiPNAS: Model interpretation



For protein–DNA binding site prediction, bar charts representing the performance of the ablated variants in terms of

(A) ROC-AUC and
(B) PR-AUC obtained using 5-fold cross validation are shown.

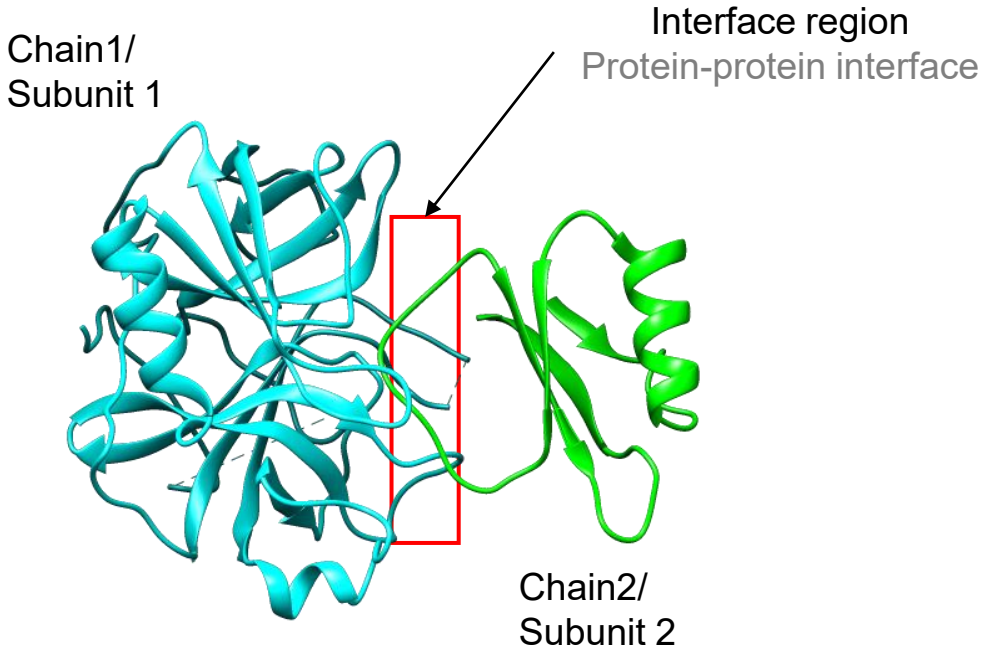
For protein–RNA binding site prediction, bar charts representing the performance of the ablated variants in terms of

(**C**) ROC-AUC and
(**D**) PR-AUC obtained using 5-fold cross validation are shown

Removing protein language model (pLM) features greatly reduces prediction accuracy

Drug discovery task
Target identification

3. Protein-Protein interaction quality prediction



- # of chains/subunits > 1
- A.K.A. protein complex
- Protein-protein interaction
- Catalyzes biological processes
- Interface quality → multimeric protein quality
- Accurately predicting interface quality helps identify stable protein complexes.
- Predicted interfaces linked to disease can be drug targets, and quality estimation shows prediction reliability.

https://www.rcsb.org/structure/1ACB

Drug discovery task
Target identification

3. Protein-Protein interaction interface prediction

ELSEVIER

Computational and Structural Biotechnology Journal

Volume 27, 2025, Pages 160-170

Research Article

EquiRank: Improved protein-protein interface quality estimation using protein language-model-informed equivariant graph neural networks

Md Hossain Shuvo [a], Debswapna Bhattacharya [b]

Show more

+ Add to Mendeley Share Cite

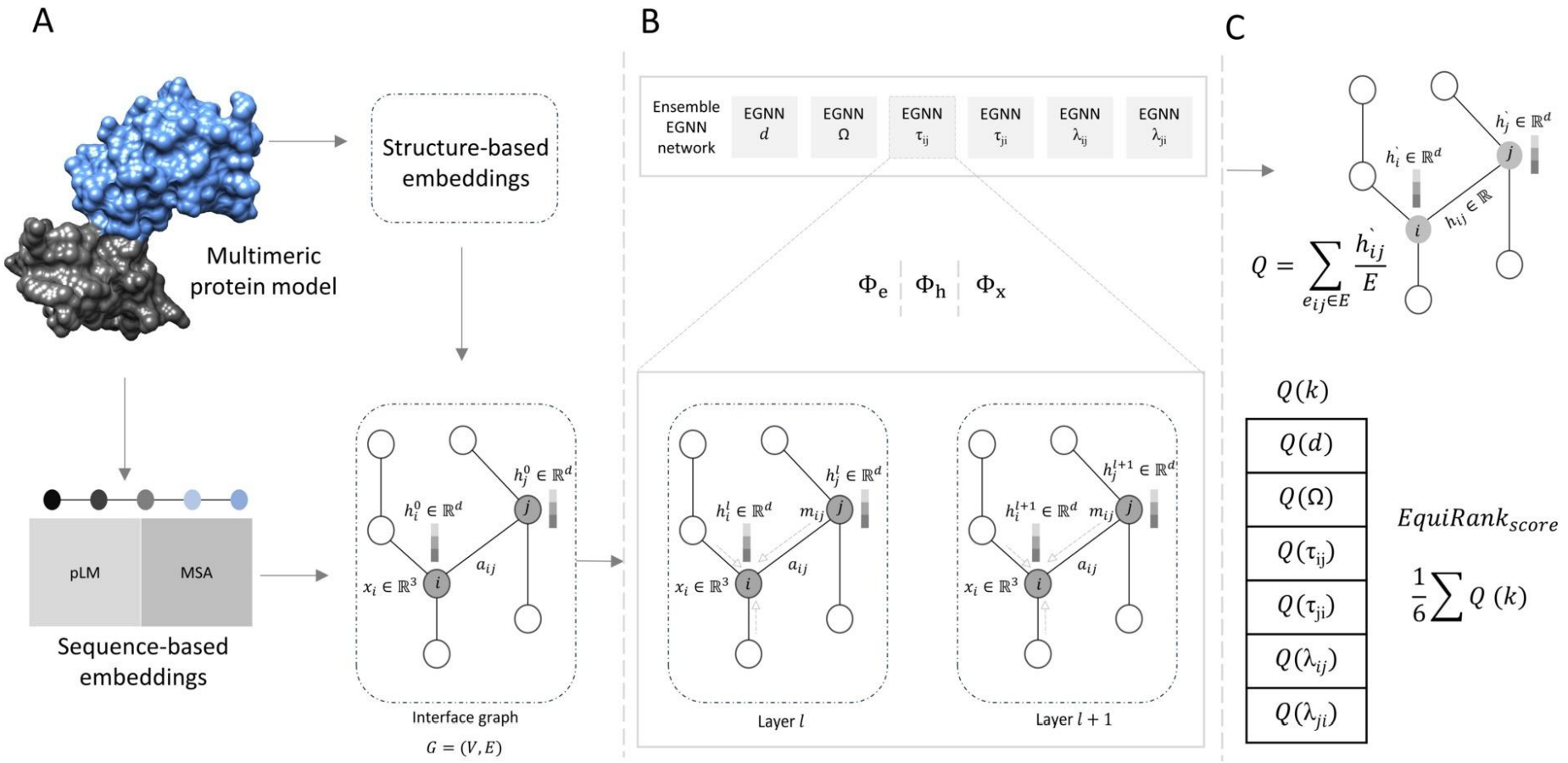
https://doi.org/10.1016/j.csbj.2024.12.015

Get rights and content

EquiRank is freely available at https://github.com/Bhattacharya-Lab/EquiRank

Drug discovery task
Target identification

3. EquiRank: Method architecture



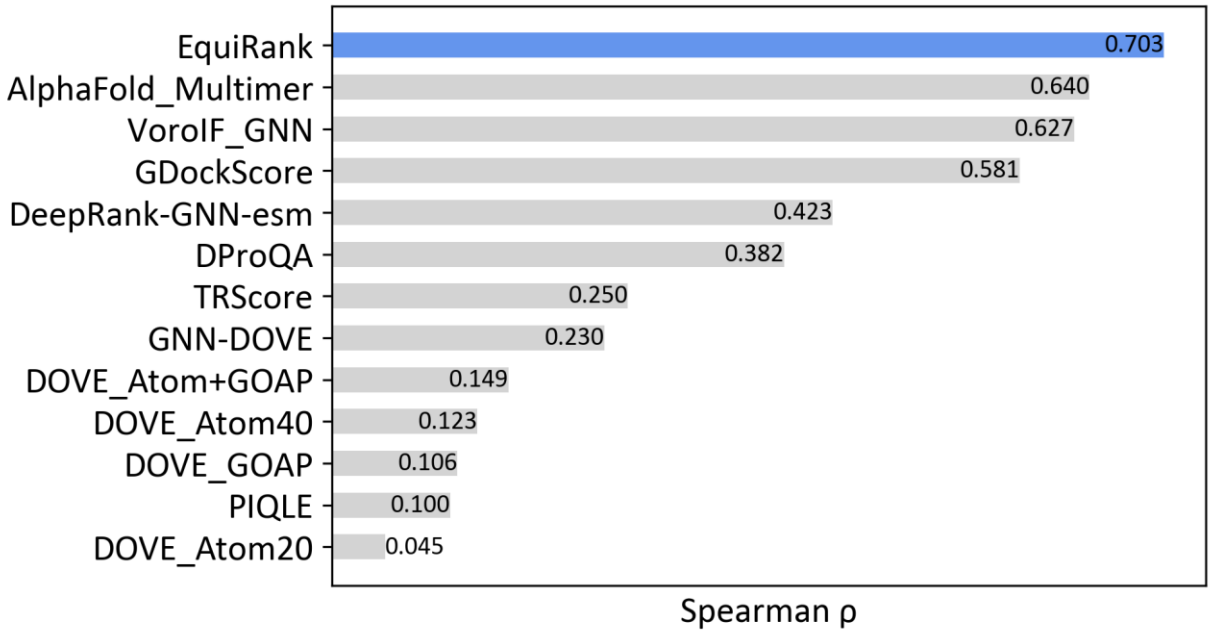
- Trained on 1,127 protein complex targets
- $EquiRank_{score}$ is the probabilistic combination of estimated multimeric distance and orientation error

Drug discovery task
Target identification

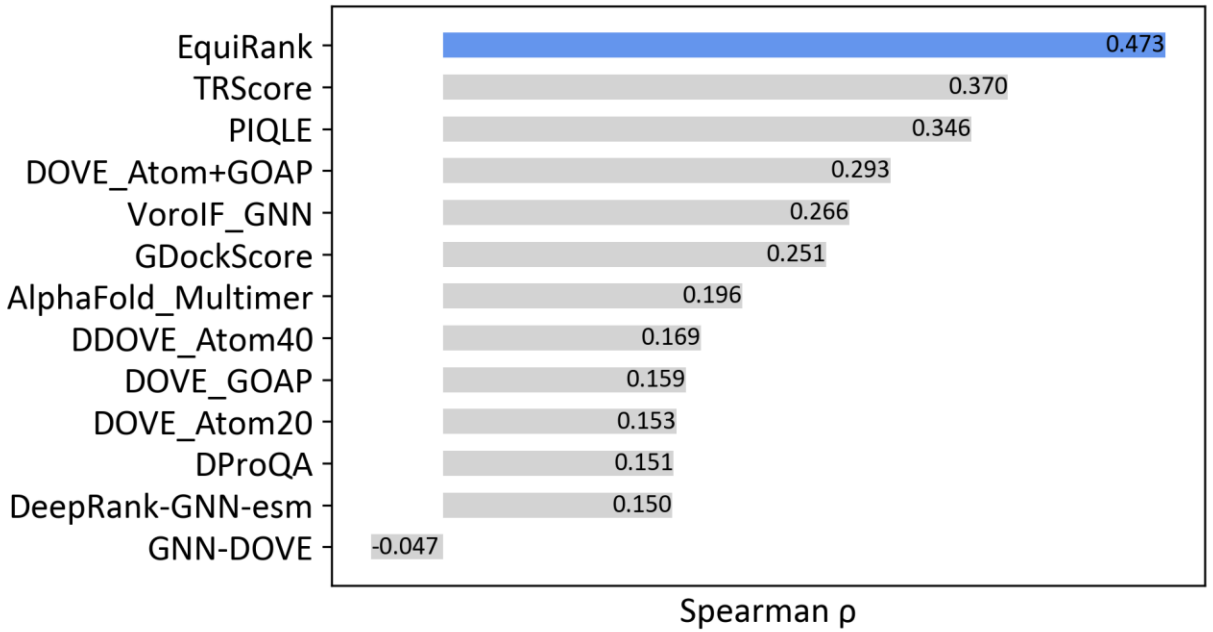
3. EquiRank: Results

Spearman correlation coefficient between predicted and DockQ scores

VoroIF_GNN_test → total decoys: 2,845 Correct: 42% Incorrect: 58%)



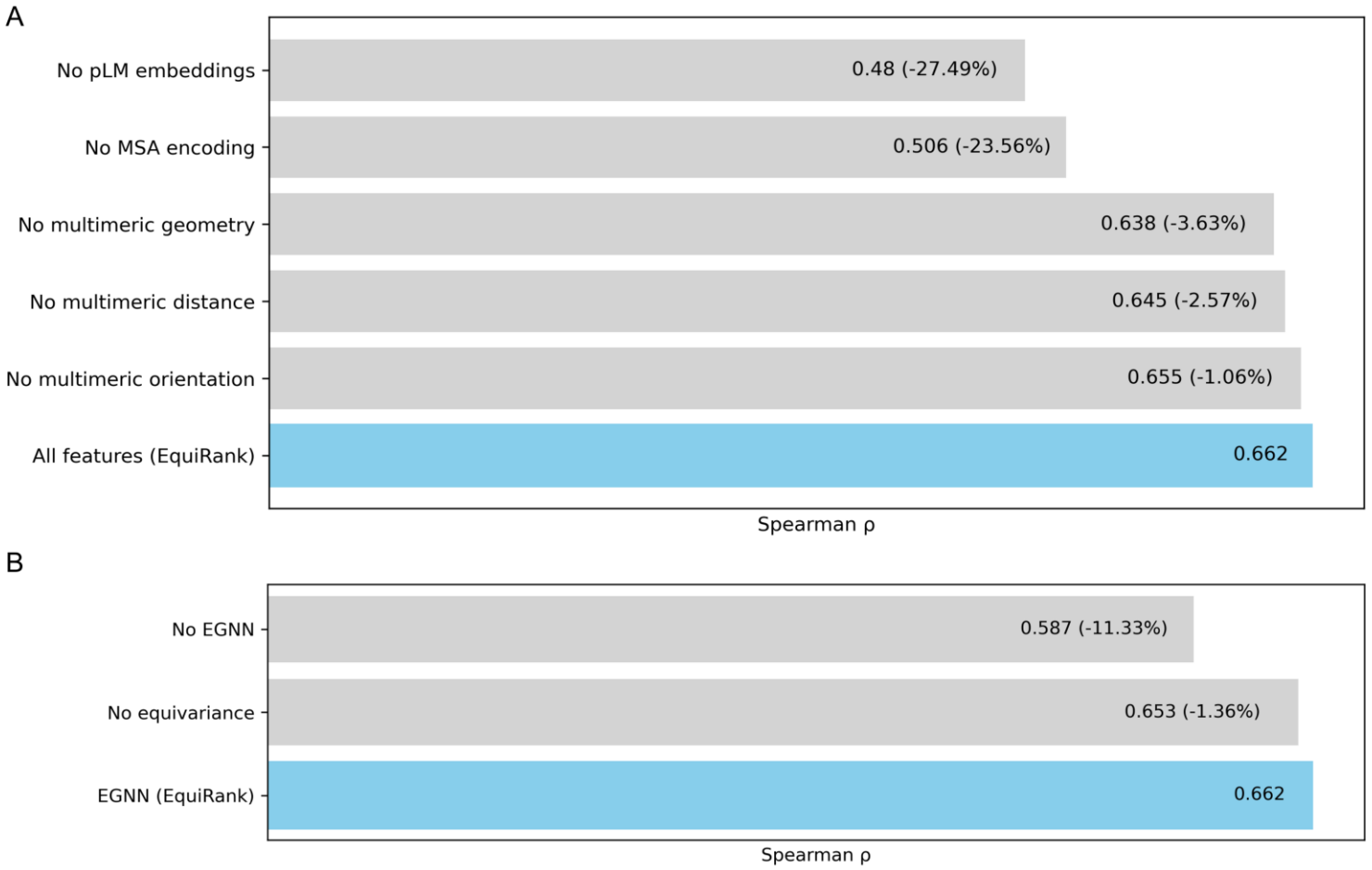
Dockground v1→ total decoys: 2,500 Correct: 10.72% Incorrect: 89.28%



EquiRank is better in reproducing ground truth ranking

Drug discovery task
Target identification

3. EquiRank: Model interpretation



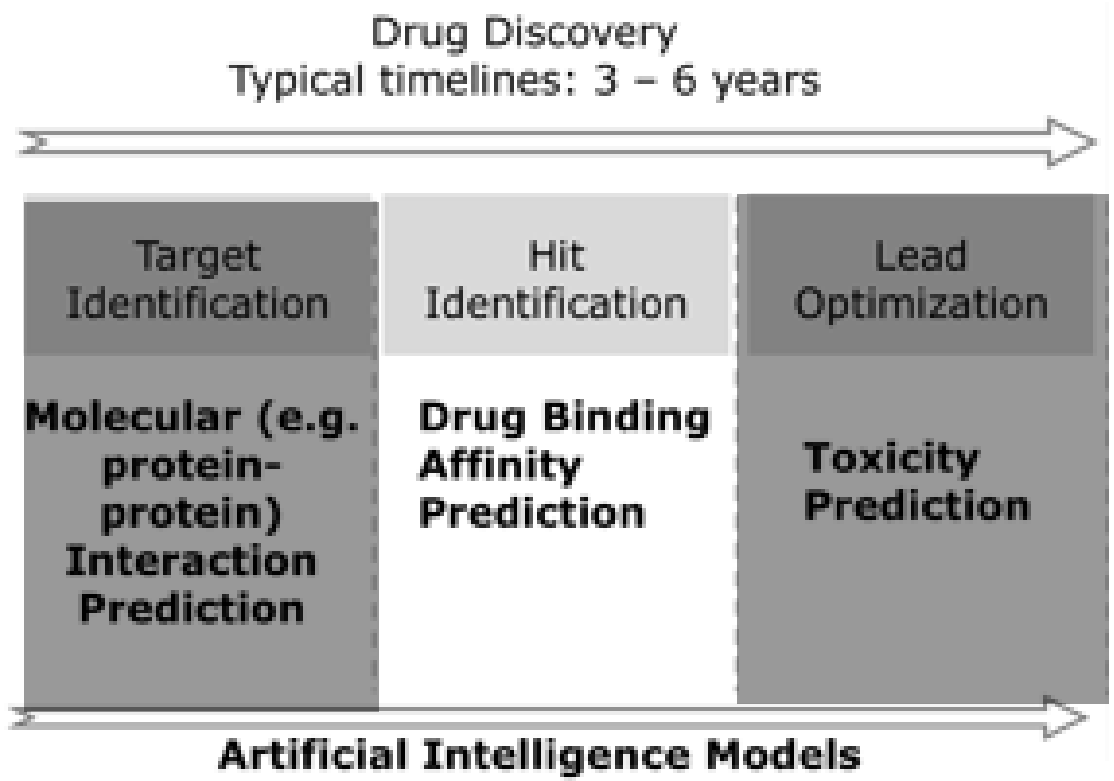
Protein Language Model embeddings and EGNN contributes to the improved model quality estimation performance

Drug discovery task
Target identification

Target identification

- Predicting molecular interactions, including **protein–protein and protein–nucleic acid interactions**, and assessing their quality can potentially help target identification in several ways
- **Equivariant graph neural networks** capture molecular symmetry and are ideal for representing complex structures
- Using **large protein language** models and **evolutionary information** further improves prediction performance
- Challenges remain due to the limited availability of high-quality protein–protein interaction databases
  - Highly accurate methods, such as **AlphaFold3**, can help generate well-curated databases for improved model development

Hit Identification



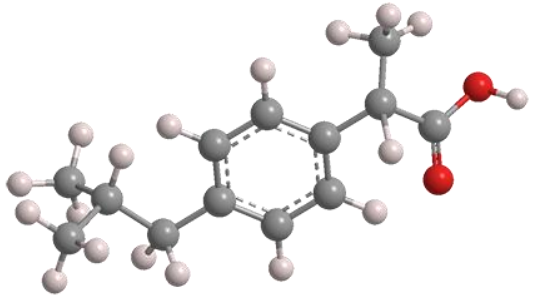
Hit identification: Drug-target binding affinity

- Hit identification is the process of on finding small molecules or compounds that can bind to a biological target such as a protein or nucleic acid to modulate its activity
- In vivo/experimental methods
  - High-throughput binding assays
  - Fragment-based screening
  - Surface plasmon resonance (SPR)

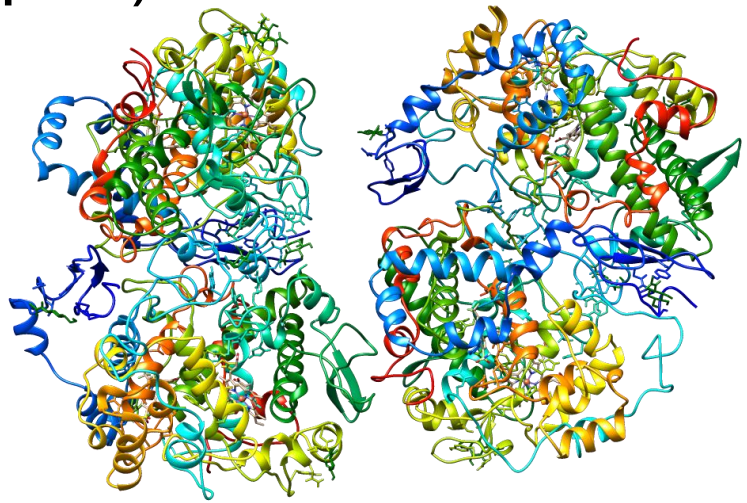
- In silico approaches
  - Drug-target biding affinity prediction (GraphDTA)
  - Drug-target interaction prediction (iNGNN-DTI)

Drug-target binding affinity

- Measures how strongly a drug molecule binds to its target (usually a protein such as enzyme or receptor)



Ibuprofen: C13H18O2

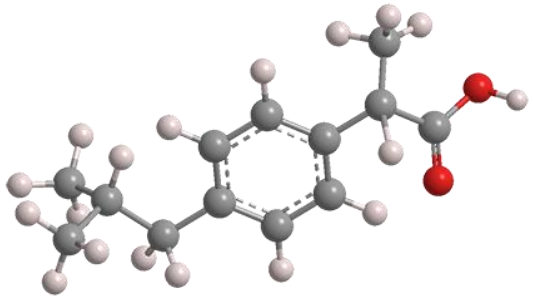


CYCLOOXYGENASE-2
PDB ID: 1CX2

- Higher affinity = stronger binding = more effective drug action
- Helps prioritize best candidate compounds (hits) before lab testing
- Reduces experimental cost and time by avoiding weak binders
- Used in virtual screening, lead optimization, and repurposing.

Drug-target binding affinity prediction: approach

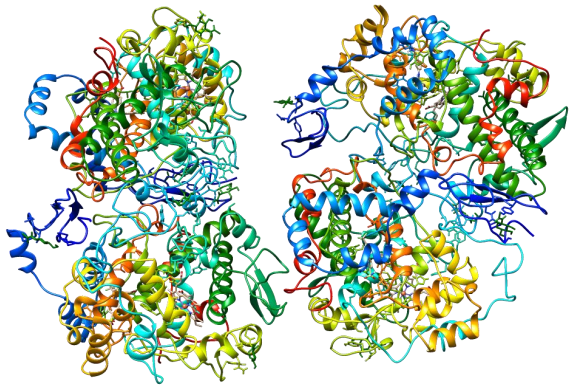
- Drug molecules are represented using SMILE string
- Protein targets are represented either in sequences or 3D-Structures



Ibuprofen: C13H18O2

SMILE

CC(C)Cc1ccc(cc1)[C@@H](C)C(=O)O



CYCLOOXYGENASE-2
PDB ID: 1CX2

Sequence

ANPCCSNPCQNRGECMSTGFDQ…

AI Models

Binding affinity

https://www.rcsb.org/structure/1ACB
SMILE: Simplified Molecular Input Line Entry System

Drug discovery task
Hit identification

Drug-target binding affinity prediction: Dataset

Dataset	Targets	Compounds	Total Records	Binding Metric	Value Range
DAVIS	379 proteins	68 inhibitors	30,056	pKd ($-\log_{10}$ Kd)	5.0 – 10.80
KIBA	467 proteins	52,498 compounds	243,251	KIBA score	–3.10 – 17.8
PDBbind 2020	19,443 proteins	15,477 ligands	19,443	pKd / pKi	0.4 – 15.22
Metz	170 proteins	1,423 ligands	35,259	Ki	4.0 – 11.1
BindingDB (2025)	4,649 targets	252,224 molecules	598,427 (Redundancy removed)	pKd / pKi / IC_{50} / EC_{50}	0-15

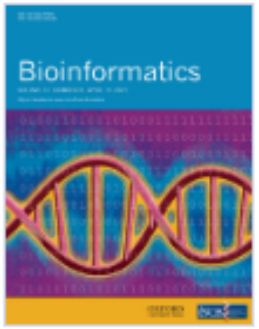
- **Kd:** measure of how easily a complex dissociates
- **Ki:** measure of how effectively an inhibitor binds
- Higher Kd and Ki values = weaker binding, and vice-versa

Drug discovery task
Hit identification

Drug-target binding affinity prediction

Bioinformatics

Issues | Advance articles | Submit ▾ | Alerts | About ▾ | Bioinformatics ▾



Volume 37, Issue 8
March 2021

Article Contents

JOURNAL ARTICLE

GraphDTA: predicting drug–target binding affinity with graph neural networks FREE

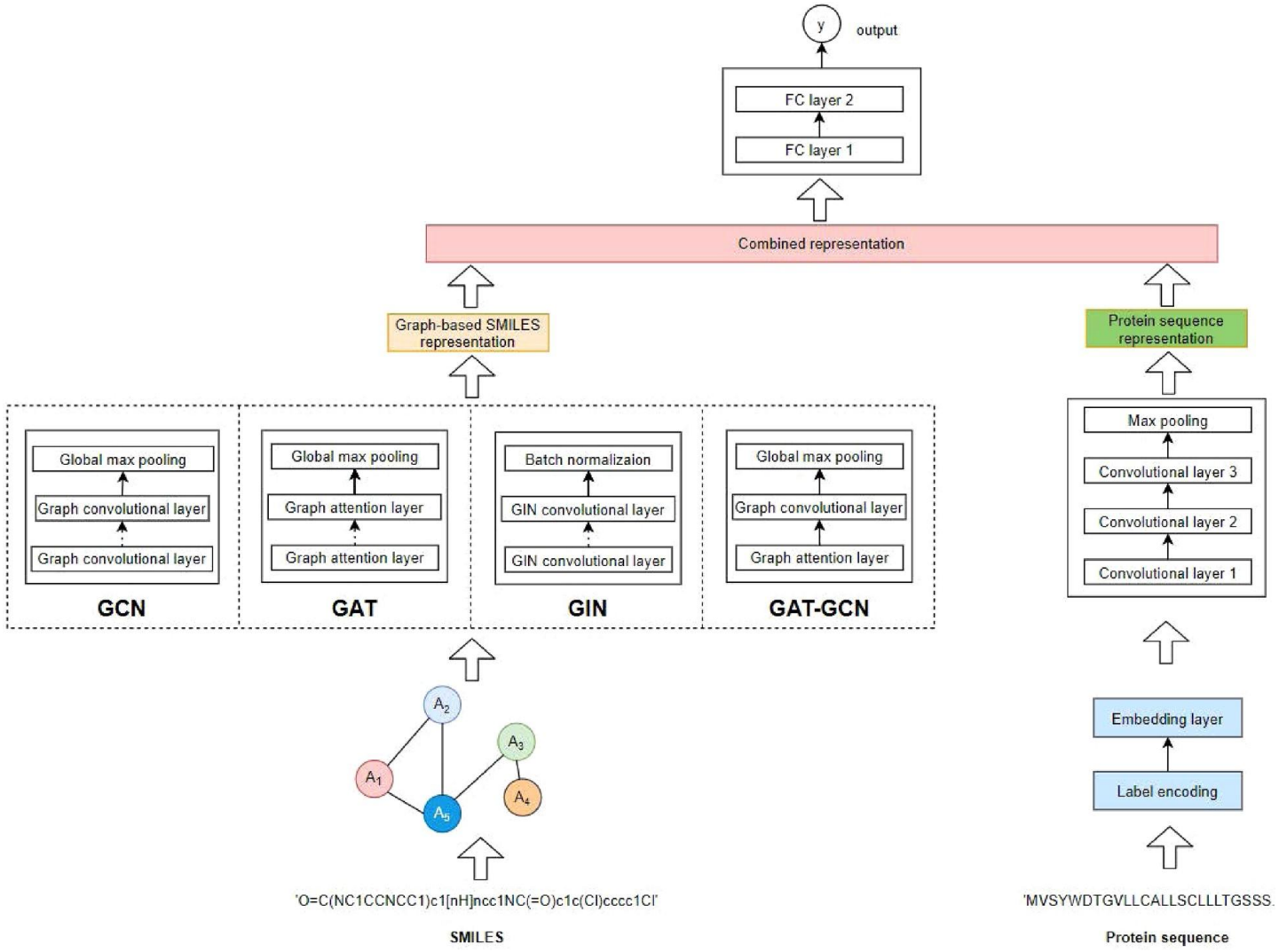
Thin Nguyen ✉, Hang Le, Thomas P Quinn, Tri Nguyen, Thuc Duy Le, Svetha Venkatesh

Bioinformatics, Volume 37, Issue 8, March 2021, Pages 1140–1147,
https://doi.org/10.1093/bioinformatics/btaa921
Published: 24 October 2020 **Article history** ▾

PDF | Split View | Cite | Permissions | Share ▾

Drug discovery task
Hit identification

GraphDTA: Method architecture



- Accepts a drug–target pair as input and predicts their affinity
- The drug SMILES string is converted into a molecular graph and encoded
- The protein sequence is embedded and processed with 1D CNN layers
- The two feature vectors are combined and passed through fully connected layers to generate the affinity score
- Trained on Davis and Kiba datasets

Nguyen et al., 2021

Drug discovery task
Hit identification

GraphDTA: Results

- Concordance Index (CI)
  - Measures ranking consistency between predicted and true affinities
  - Higher CI indicates better predictive ranking (range: 0.5 – 1.0)
- Mean Squared Error (MSE)
  - Lower MSE indicates more accurate predictions
- Pearson Correlation Coefficient (R)
  - Measures linear correlation between predicted and actual values
  - Higher R indicates more accurate predictions
- Spearman Rank Correlation (ρ)
  - Measures monotonic relationship between predicted and actual rankings
  - Higher ρ indicates more accurate predictions

GraphDTA: Results

Prediction performance on the Kiba dataset, sorted by MSE

Method	Protein rep.	Compound rep.	CI	MSE
Baseline models				
DeepDTA (Öztürk *et al.*, 2018)	1D	Pubchem-Sim	0.718	0.571
DeepDTA (Öztürk *et al.*, 2018)	Smith-Waterman	Pubchem-Sim	0.710	0.502
KronRLS (Cichonska *et al.*, 2017, 2018)	Smith-Waterman	Pubchem-Sim	0.782	0.411
SimBoost (He *et al.*, 2017)	Smith-Waterman	Pubchem-Sim	0.836	0.222
DeepDTA (Öztürk *et al.*, 2018)	Smith-Waterman	1D	0.854	0.204
DeepDTA (Öztürk *et al.*, 2018)	1D	1D	0.863	0.194
WideDTA (Öztürk *et al.*, 2019)	1D + PDM	1D + LMCS	*0.875*	*0.179*
Proposed method—**GraphDTA**				
GAT	1D	Graph	0.866	0.179
GIN	1D	Graph	**0.882**	**0.147**
GCN	1D	Graph	**0.889**	**0.139**
GAT_GCN	1D	Graph	**0.891**	**0.139**

- GIN variant performed best across both datasets and metrics
- Graph-based representations (1D + graph) outperform traditional sequence/similarity-based methods

Drug discovery task
Hit identification

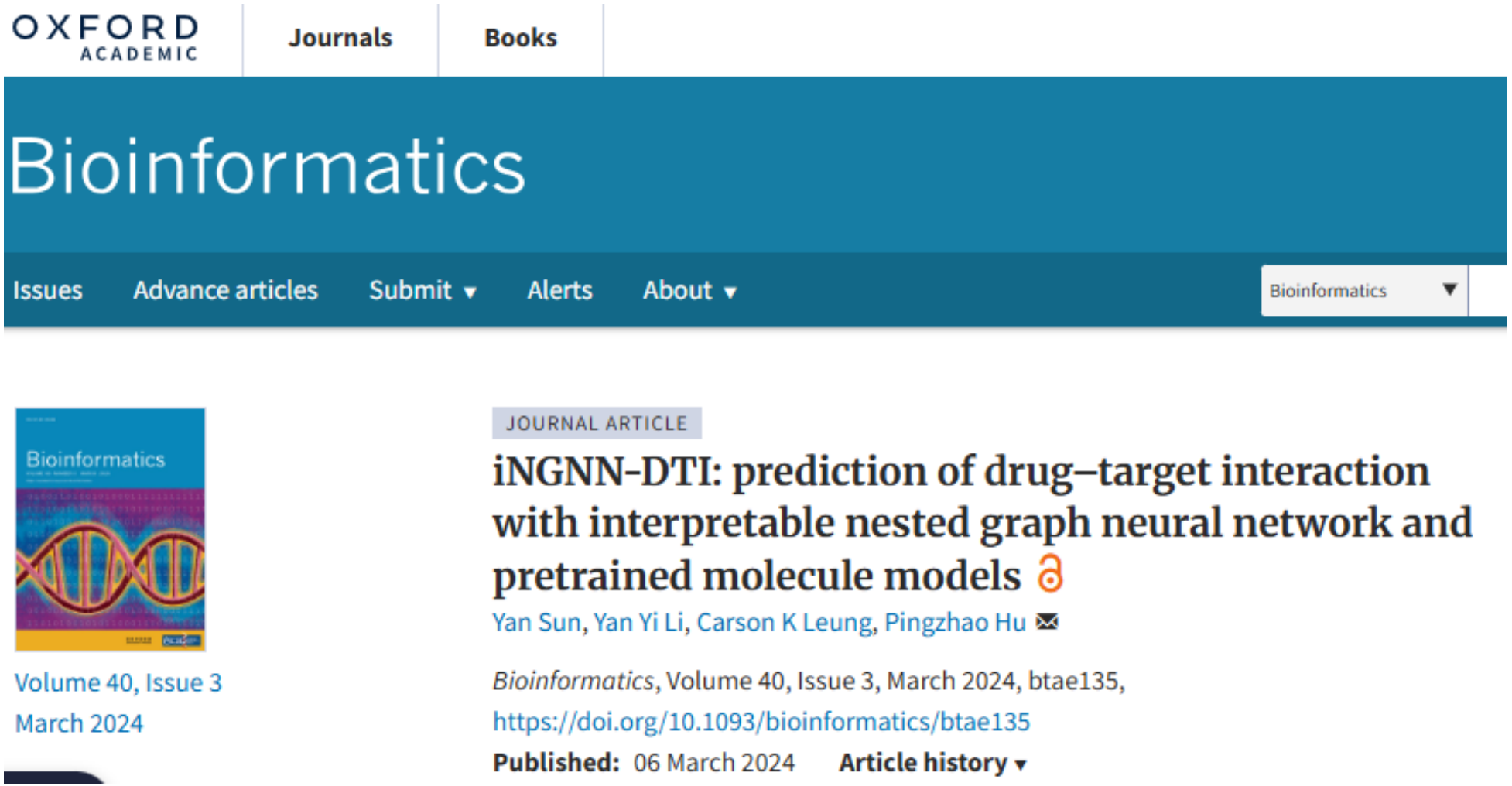
GraphDTA: Conclusion

- GraphDTA tested with GCN, GAT, GIN, and GAT-GCN for drug–target affinity prediction
- Performs strongly on Davis and Kiba datasets across key metric
- Learns molecular descriptor importance without prior knowledge
- Few drugs contribute disproportionately to prediction error
- Future improvements may involve protein graph representations using 3D structural data

Drug-target interaction prediction

- Drug-target interaction prediction is the process of predicting whether a drug and a target protein interact
- Simpler than drug-target binding affinity prediction
- DTI is generally used to identify possible drug–target pairs, and DTA is then applied to estimate and prioritize their binding strengths

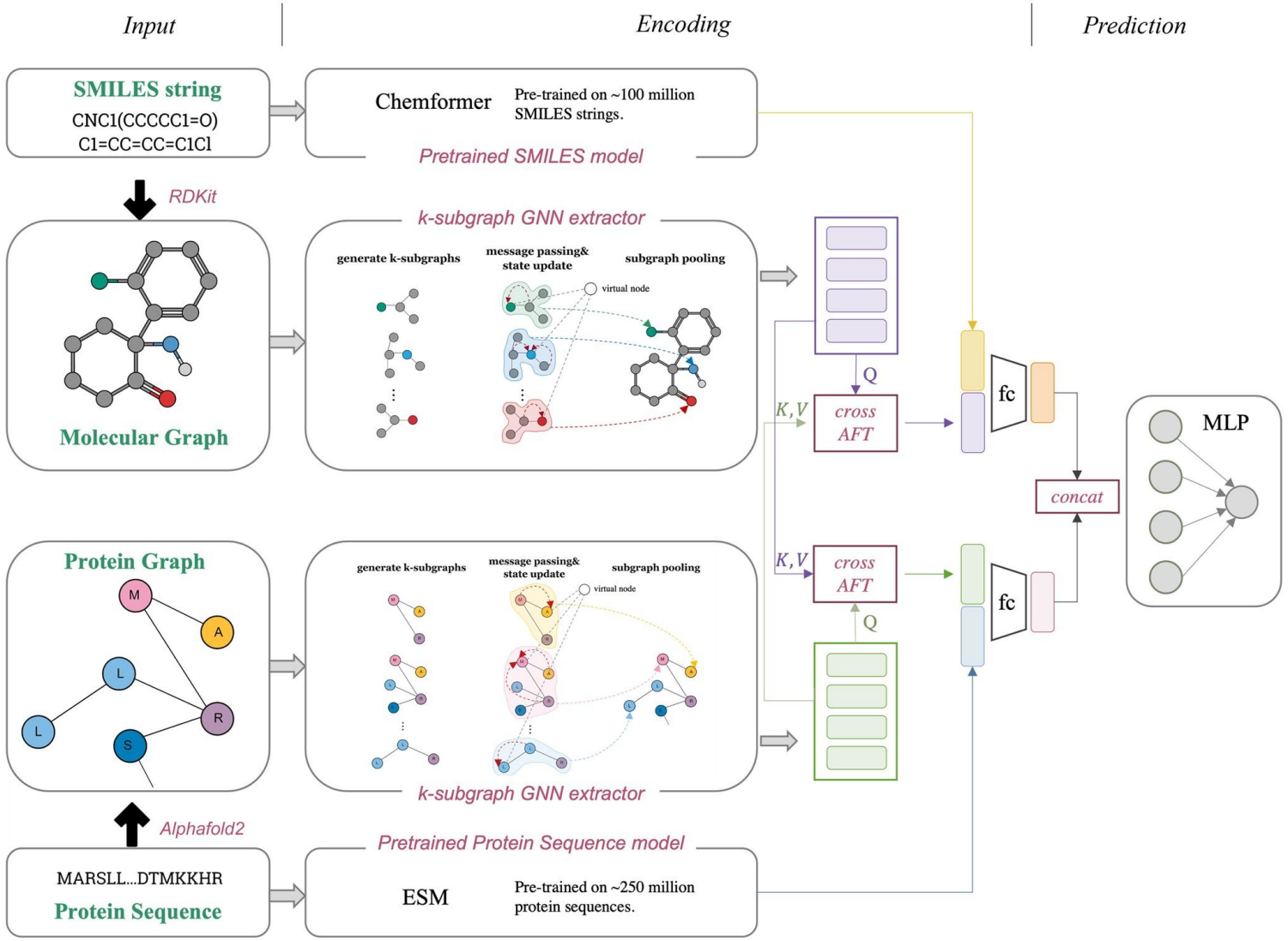
Drug-target interaction prediction



Sun et al., 2024

Drug discovery task
Hit identification

iNGNN-DTI: Method architecture



- iNGNN-DTI converts drug SMILES strings and protein sequences into graph representations.
- A k-subgraph GNN encoder processes the graph data, while Chemformer and ESM encode the SMILES and amino acid sequences.
- The cross-AFT module captures interaction information between drugs and proteins.
- Drug and protein representations are combined with global features from unsupervised models.
- The final concatenated representation is fed to an MLP to predict drug–target interactions.

Drug discovery task
Hit identification

iNGNN-DTI : Results

Model performance comparison.

Method	AUROC	AUPRC	Sensitivity	Specificity
DAVIS dataset				
DeepDTA	0.892 ± 0.0066	0.378 ± 0.0231	0.854 ± 0.0066	0.792 ± 0.0291
Moltrans	0.898 ± 0.0050	0.371 ± 0.0067	0.865 ± 0.0050	0.783 ± 0.0387
ML-DTI	0.910 ± 0.0034	0.381 ± 0.0247	0.895 ± 0.0034	0.795 ± 0.0183
DGraphDTA (Alphafold2)	0.885 ± 0.0099	0.316 ± 0.0447	0.894 ± 0.0099	0.724 ± 0.0467
iNGNN-DTI	**0.931 ± 0.0027**	**0.473 ± 0.0167**	**0.922 ± 0.0155**	**0.802 ± 0.0240**
KIBA dataset				
DeepDTA	0.912 ± 0.0037	0.743 ± 0.0127	0.881 ± 0.0056	0.780 ± 0.0127
Moltrans	0.899 ± 0.0022	0.691 ± 0.0142	0.872 ± 0.0116	0.760 ± 0.0160
ML-DTI	0.909 ± 0.0020	0.727 ± 0.0108	0.878 ± 0.0111	0.779 ± 0.0113
DGraphDTA (Alphafold2)	0.911 ± 0.0004	0.739 ± 0.0043	0.881 ± 0.0183	**0.784 ± 0.0277**
iNGNN-DTI	**0.915 ± 0.0016**	**0.753 ± 0.0071**	**0.888 ± 0.0107**	0.779 ± 0.0146

- For the Davis dataset, a pKd value of at least 7.1 is considered an interaction.
- For the KIBA dataset, a score above 12.1 indicates a positive interaction.
- The model achieved better performance on both the Davis and KIBA datasets

Drug discovery task
Hit identification

Drug-target binding affinity and interaction prediction

- Drug–target binding affinity (DTA) and drug–target interaction (DTI) prediction are related
- DTA is a regression task predicting the strength of binding, while DTI is a classification task predicting whether an interaction occurs.
- DTI is often used to screen candidate drugs, and DTA helps prioritize the strongest binders

Challenges and gaps

- Need to develop ensemble models that effectively combine multiple architectures to capture important features and better learn complex drug–target representations
- Most deep learning–based DTA prediction methods lack interpretability, making it necessary to incorporate explainability to improve trustworthiness and understanding
- Incorporating AlphaFold-predicted protein structures can enhance the use of accurate structural features for improved prediction
- Current models often struggle to generalize due to limited and imbalanced datasets, therefore, there is a need for diverse, high-quality, and well-curated databases to ensure broader applicability.

Xin et. Al. 2024
Amit et. al. 2024

Acknowledgement

Collaborators

- Dr. Debswapna Bhattacharya (Former advisor)
- Dr. Rahmatullah Roche
- Bernerd Moussad
- Sumit Tarafder

Funding sources and support

- NIH
- NIGMS
- NSF
- National AI Research Resource Pilot
- PVAMU Division of R&I

Key references

1. Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., Bodenstein, S. W., Evans, D. A., Hung, C.-C., O'Neill, M., Reiman, D., Tunyasuvunakool, K., Wu, Z., Žemgulytė, A., Arvaniti, E., … Jumper, J. M. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 1–3. https://doi.org/10.1038/s41586-024-07487-w
2. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), Article 7873. https://doi.org/10.1038/s41586-021-03819-2
3. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., Costa, A. dos S., Fazel-Zarandi, M., Sercu, T., Candido, S., & Rives, A. (2022). *Evolutionary-scale prediction of atomic level protein structure with a language model* (p. 2022.07.20.500902). bioRxiv. https://doi.org/10.1101/2022.07.20.500902
4. Nguyen, T., Le, H., Quinn, T. P., Nguyen, T., Le, T. D., & Venkatesh, S. (2021). GraphDTA: Predicting drug–target binding affinity with graph neural networks. *Bioinformatics*, *37*(8), 1140–1147. https://doi.org/10.1093/bioinformatics/btaa921
5. Qureshi, R., Irfan, M., Gondal, T. M., Khan, S., Wu, J., Hadi, M. U., Heymach, J., Le, X., Yan, H., & Alam, T. (2023). AI in drug discovery and its clinical relevance. *Heliyon*, *9*(7), e17575. https://doi.org/10.1016/j.heliyon.2023.e17575
6. Roche, R., Moussad, B., Shuvo, M. H., & Bhattacharya, D. (2023). E(3) equivariant graph neural networks for robust and accurate protein-protein interaction site prediction. *PLOS Computational Biology*, *19*(8), e1011435. https://doi.org/10.1371/journal.pcbi.1011435
7. Roche, R., Moussad, B., Shuvo, M. H., Tarafder, S., & Bhattacharya, D. (2024). EquiPNAS: Improved protein–nucleic acid binding site prediction using protein-language-model-informed equivariant deep graph neural networks. *Nucleic Acids Research*, *52*(5), e27. https://doi.org/10.1093/nar/gkae039
8. Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W. R., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D. T., Silver, D., Kavukcuoglu, K., & Hassabis, D. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, *577*(7792), Article 7792. https://doi.org/10.1038/s41586-019-1923-7
9. Shuvo, M. H., & Bhattacharya, D. (2025). EquiRank: Improved protein-protein interface quality estimation using protein language-model-informed equivariant graph neural networks. *Computational and Structural Biotechnology Journal*, *27*, 160–170. https://doi.org/10.1016/j.csbj.2024.12.015
10. Sun, Y., Li, Y. Y., Leung, C. K., & Hu, P. (2024). iNGNN-DTI: Prediction of drug–target interaction with interpretable nested graph neural network and pretrained molecule models. *Bioinformatics*, *40*(3), btae135. https://doi.org/10.1093/bioinformatics/btae135
11. Zeng, X., Li, S.-J., Lv, S.-Q., Wen, M.-L., & Li, Y. (2024). A comprehensive review of the recent advances on predicting drug-target affinity based on deep learning. *Frontiers in Pharmacology*, *15*, 1375522. https://doi.org/10.3389/fphar.2024.1375522

Thank You!
Questions?

Data-Driven AI Frameworks toward Computational Drug Discovery

---

Md Hossain Shuvo, Ph.D.

mhshuvo@pvamu.edu